2021

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Annual Reports

AIMH research activities 2021

AIMH lab., CNR-ISTI, Pisa, Italy

ISTI-AR-2021/003

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

AIMH Research activities 2021
AIMH lab.
ISTI-AR-2021/003

Abstract
The Artificial Intelligence for Media and Humanities laboratory (AIMH) has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability.
This report summarize the 2021 activities of the research group.
Artificial intelligence, Computer vision, Multimedia information retrieval, Similarity search, Natural language processing, Digital humanities.

Citation
AIMH lab. *AIMH Research activities 2021*, ISTI Annual Reports 2021/003. DOI: 10.32079/ISTI-AR-2021/003

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

AIMH Research Activities 2021

Nicola Aloia, Giuseppe Amato, Valentina Bartalesi, Filippo Benedetti, Paolo Bolettieri, Donato Cafarelli, Fabio Carrara, Vittore Casarosa, Davide Coccomini, Luca Ciampi, Cesare Concordia, Silvia Corbara, Marco Di Benedetto, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, Gabriele Lagani, Fabio Valerio Massoli, Carlo Meghini, Nicola Messina, Daniele Metilli, Alessio Molinari, Alejandro Moreo, Alessandro Nardi, Andrea Pedrotti, Nicolò Pratelli, Fausto Rabitti, Pasquale Savino, Fabrizio Sebastiani, Gianluca Sperduti, Costantino Thanos, Luca Trupiano, Lucia Vadicamo, Claudio Vairo

Abstract
The Artificial Intelligence for Media and Humanities laboratory (AIMH) has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability. This report summarize the 2021 activities of the research group.

Keywords
Multimedia Information Retrieval – Artificial Intelligence — Computer Vision — Similarity Search – Machine Learning for Text – Text Classification – Transfer learning – Representation Learning

[1] *AIMH Lab, ISTI-CNR, via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy*
*__Corresponding author__: giuseppe.amato@isti.cnr.it

Contents

AIMH
ARTIFICIAL INTELLIGENCE FOR MEDIA AND HUMANITIES

http://aimh.isti.cnr.it

Introduction

The Artificial Intelligence for Media and Humanities laboratory (AIMH) of the Information Science and Technologies Institute "A. Faedo" (ISTI) of the Italian National Research Council (CNR) located in Pisa, has the mission to investigate and advance the state of the art in the Artificial Intelligence field, specifically addressing applications to digital media and digital humanities, and taking also into account issues related to scalability.

The laboratory is composed of four research groups:

AI4Text

The AI4Text is active in the area at the crossroads of machine learning and text analysis; it investigates novel algorithms and methodologies, and novel applications of these to different realms of text analysis. Topics within the above-mentioned area that are actively researched within the group include representation learning for text classification, transfer learning for cross-lingual and cross-domain text classification, sentiment classification, sequence learning for information extraction, text quantification, transductive text classification, cost-sensitive text classification, and applications of the above to domains such as authorship analysis and technology-assisted review. The group consists of Fabrizio Sebastiani (Director of Research), Andrea Esuli (Senior Researcher), Alejandro Moreo (Researcher), Silvia Corbara, Alessio Molinari, Andrea Pedrotti, and Gianluca Sperduti (PhD Students), and is led by Fabrizio Sebastiani.

Humanities

Investigating AI-based solutions to represent, access, archive, and manage tangible and intangible cultural heritage data. This includes solutions based on ontologies, with a special focus on narratives, and solutions based on multimedia content analysis, recognition, and retrieval. The group consists of Carlo Meghini (Director of Research), Valentina Bartalesi, Cesare Concordia (Researchers), Luca Trupiano (Technologist), Daniele Metilli (PhD Student), Nicolò Pratelli (Graduate Fellows), and Costantino Thanos, Vittore Casarosa, Nicola Aloia (Research Associates), and is led by Carlo Meghini.

Large-scale IR

Investigating efficient, effective, and scalable AI-based solutions for searching multimedia content in large datasets of non-annotated data. This includes techniques for multimedia content extraction and representation, scalable access methods for similarity search, multimedia database management. The group consists of Claudio Gennaro, Pasquale Savino (Senior Researchers), Lucia Vadicamo (Researcher), Claudio Vairo (Researchers), Paolo Bolettieri (Technician), Luca Ciampi, Gabriele Lagani (PhD Students), and Fausto Rabitti (Research Associate), and is led by Claudio Gennaro.

Vision and Deep Learning

Investigating novel AI-based solutions to image and video content analysis, understanding, and classification. This includes techniques for detection, recognition (object, pedestrian, face, etc), classification, counting, feature extraction (low- and high-level, relational, cross-media, etc), anomaly detection also considering adversarial machine learning threats. We also have specific AI research fields such as hebbian learning and relational learning. The group consists of Giuseppe Amato (Director of Research), Fabrizio Falchi (Senior Researcher), Marco Di Benedetto, Fabio Carrara (Researchers), Alessandro Nardi (Technician), Fabio Massoli (Post-doc Fellow), Davide Alessandro Coccomini, Gabriele Lagani, Nicola Messina (PhD Students), Donato Cafarelli (Graduate Fellow), and is led by Fabrizio Falchi.

The rest of the report is organized as follows. In Section 1, we summarize the research conducted on our main research fields. In Section 2, we describe the projects in which we were involved during the year. We report the complete list of papers we published in 2021, together with their abstract, in Section 3. The list of theses on which we were involved can be found in Section 4. In Section 6 we highlight the datasets we created and made publicly available during 2021.

1. Research Topics

In the following, we report a list of active research topics and subtopics at AIMH in 2021.

1.1 Artificial Intelligence

1.1.1 Hebbian Learning

Traditional neural networks are trained using gradient descent methods with error backpropagation. Despite the great success of such training algorithms, the neuroscientific community has doubts about the biological plausibility of backpropagation learning schemes, proposing a different learning model known as *Hebbian principle*: "Neurons that fire together wire together". Starting from this simple principle, different Hebbian learning variants have been formulated. These approaches are interesting also from a computer science point of view, because they allow to perform common data analysis operations - such as clustering, Principal Component Analysis (PCA), Independent Component Analysis (ICA), and others - in an online, efficient, and neurally plausible fashion. Taking inspiration from biology, we investigate how Hebbian approaches can be integrated with today's machine learning techniques [31], in order to improve the training process in terms of speed, generalization capabilities, sample efficiency [30].

An even more biologically plausible model of neural computation is based on Spiking Neural Networks (SNNs). In this model, neurons communicate via short pulses called *spikes*. This communication approach is the key towards energy efficiency in the brain. We are using SNNs to accurately simulate real neuronal cultures, in collaboration with neuroscience colleagues, who can produce such cultures in lab. Multi-Electrode Array (MEA) devices can be used to stimulate and record activity from cultured networks, raising the question

of whether such cultures can be trained to perform AI tasks. Our simulations help us understand the optimal parameters a cultured network should have in order to solve a given task, providing insights to guide neuroscientists in the creation of real cultures with the desired properties [32].

1.1.2 Abstract Visual Reasoning

Humans always think by relating distant and abstract concepts through complex analogical reasoning. Therefore, it is interesting to understand to which extent a machine learning algorithm — and a Deep Neural Network in particular — can solve apparently simple yet challenging tasks requiring distant comparisons. Given the importance of images in our world, we are especially interested in tackling *abstract visual reasoning* problems that require this kind of relational intelligence to be correctly solved.

In particular, we tackled apparently trivial visual reasoning tasks, known as the *same-different* tasks. In short, the same-different tasks consist in understanding if two shapes in an image satisfy a certain rule. In the simplest case, the rule is merely that *the two shapes must be equal*; however, the rule is not known a priori and must be internally understood from the provided positive and negative examples. It is a challenging set of tasks for machine learning algorithms. In fact, it is required not to learn specific shape patterns to solve the problem; instead, they require to grow some abstract internal representation that is powerful enough to draw a logical conclusion on a fact hidden in the image (e.g., the shapes in the images are the same even if they are orientated in different ways).

In [44] we probed many state-of-the-art CNNs, to understand if they are able to solve these challenging visual task. In this work we also introduced little variations to the presented architectures to better understand the role of the architectural features in the convergence or generalization abilities. More recently, in the preliminary work presented in [43] we recently proposed to use a Recurrent Transformer network to perform high-level reasoning, using the features extracted by a simple upstream CNN. This architecture seems to defeat many of the previous fully convolutional models by using less free parameters and reaching better data efficiency. Furthermore, the learned attention maps clearly indicate which image patches the model is attending to answer correctly.

1.1.3 Deep Anomaly Detection

Anomalies are met in every scientific field. The term "anomaly" is itself a source of ambiguity since it is usually used to point at both outliers and anomalies. Training deep learning architectures on the task of detecting such events is challenging since they are rarely observed. Moreover, typically we don't know their origin and the construction of a dataset containing such a kind of data is too expensive. For such a reason, unsupervised and semi-supervised techniques are exploited to train neural networks.

In [37], we proposed a novel method named MOCCA, in which we exploit the piece-wise nature of deep learning models to detect anomalies. We tasked the model to minimize the deep features distance among a reference point, the class centroid for anomaly-free images, and the current input. By extracting the deep representations at different depths and combining them, MOCCA improved upon the state-of-the-art considering the one-class classification setting on the task of anomaly detection.

1.1.4 Adversarial Machine Learning

Adversarial machine learning is about attempting to fool models through malicious input. The topic has become very popular with the recent advances on Deep Learning. We studied this topic with a focus on detection of adversarial examples and images in particular, contributing to the field with several publications in the last years. Our research investigated adversarial detection powered by the analysis of the internal activation of deep networks (a.k.a. deep features) collecting encouraging results over the last three years. This year's research activity on the topic included increasing adversarial robustness of a novel deep architecture — Neural Ordinary Differential Equations. Neural Ordinary Differential Equations comprise novel differentiable and learnable models (also referred to as ODE-Nets) whose outputs are defined as the solution of a system of parametric ordinary differential equations. Those models exhibit benefits such as a O(1)-memory consumption and a straight-forward modelling of continuous-time and inhomogeneous data, and when using adaptive ODE solvers, they acquire also other interesting properties, such as input-dependent adaptive computation and the tunability (via a tolerance parameter) of the accuracy-speed trade-off at inference time. We studied their unique properties under an adversarial setting; we analyzed the accuracy-speed trade-off they offer at inference time and how tuning this trade-off affects robustness to strong adversarial attacks [9]. Our findings showed an innate improved robustness of these models against adversarial attacks with respect to standard neural networks.

Given the experience we have in face recognition and cross-resolution in particular, we developed a specific approach for adversarial faces [36].

1.2 AI and Digital Humanities

The AI & DH group at AIMH employs AI-based methods to research, design and experimentally develop innovative tools to support the work of the scholar humanist. These methods hinge on formal ontologies as powerful tools for the design and the implementation of information systems that exhibit intelligent behavior. Formal ontologies are also regarded as the ideal place where computer scientists and humanists can meet and collaborate to co-create innovative applications that can effectively support the work of the latter. The group pursues in particular the notion of formal narrative as a powerful addition to the information space of digital libraries; an ontology for formal narratives has been developed in the last few years and it is currently being enriched through the research carried out by the members of the group and tested through the validation carried out in the context of the Mingei project.

The group is also engaged in the formal representation of literary texts and of the surrounding knowledge, through the HDN project which continues the seminal work that led to the DanteSources application where an ontology-based approach was firstly employed. Finally, through the participation to the ARIADNEplus and the SSHOC projects, the group is actively involved in the making of two fundamental infrastructures in the European landscape, on archaeology and on social sciences & humanities, respectively.

1.3 AI for Text

1.3.1 Learning to quantify

Learning to quantify has to do with training a predictor that estimates the prevalence values of the classes of interest in a sample of unlabelled data. This problem has particular relevance in scenarios characterized by distribution shift (which may itself be caused by either covariate shift or prior probability shift), since standard learning algorithms for training classifiers are based on the IID assumption, which is violated in scenarios characterized by distribution shift. The AI4Text group has carried out active research on learning to quantify since 2010.

One of our recent activities in this direction has involved looking back at past approaches to learning to quantify with a critical eye. In one such study [57] we have reassessed the true merits of "classify and count", the baseline of all quantification studies, due to the fact that, as we have found out, in many published studies this method has been a straw man rather than a baseline, due to lack of or suboptimal parameter optimization. We have proposed new quantification-oriented parameter optimization protocols, and reassessed classify and count, and several other quantification methods, under the new lens that they provide. In another such study [21] we have looked back at past research on sentiment quantification, and found that the different approaches to such a task have been compared inappropriately, due to a faulty experimental protocol. We have thus carried out a complete reassessment of these approaches, this time using a much more robust protocol which involves a much more extensive experimentation; also these results have upturned past conclusions concerning the relative merits of such approaches.

In a different effort [27], we have devised a method that uses quantification in order to measure the bias of a classifier with respect to a sensitive attribute of interest (e.g., race) in scenarios in which the values of this attribute are not known at classifier training time. This method thus allows to address a common application scenario, since organizations often avoid collecting the values of sensitive attributes unless needed. Our quantification-based method is also a privacy-preserving one since, differently from a classification-based one, it does not allow recovering the value of the sensitive attribute (which would be undesirable), but only allows inferences to be carried out at the aggregate level.

Three further activities in which we have engaged are

- the organization of the 1st International Workshop on Learning to Quantify (LQ 2021)[1] [21], which has taken place in November 2021 as an online event;
- the organization of LeQua 2022[2] [26], the first shared task entirely devoted to learning to quantify, which is being organized under the umbrella of the CLEF 2022 conference [3];
- the implementation of QuaPy [51], an open-source, Python-based software library for learning to quantify, that contains implementations of the most important methods, evaluation measures, and evaluation protocols for quantification, as well as datasets frequently used in the quantification community.

1.3.2 Learning to classify text

The supervised approach to text classification (TC) is almost 30 years old; despite this, text classification continues to be an active research topic, due to its central role in a number of text analysis and text management tasks.

One problem we have worked on is how to improve the accuracy of text classification via techniques from representation learning. We have devised a new type of embedded representations, called *word-class embeddings* (WCEs – [53]), that encode correlations between words and classes as learnt from the training set via straightforward matrix computations, and that can be used alongside other embedded representations (such as standard word embeddings), demonstrably improving the accuracy of text classification.

Another problem we have been working on [56] is *cross-lingual TC*, i.e., the task of leveraging training data for a "source" language in order to perform TC in a different, "target" language for which we have little or no training data. In [56, 54] we have extended a previously proposed method for heterogeneous transfer learning (called "Funnelling") to leverage correlations in data that are informative for the TC process; while Funnelling exploit class-class correlations, our "Generalized Funnelling" system also exploits word-class correlations (by employing the above-mentioned WCEs), word-word correlations (for which we employ MUSE embeddings), and correlations between words-in-context, obtained via Multilingual BERT.

In a different effort we have studied *transductive transfer learning* [50], discussed how the term "transduction" has been misused in the transfer learning literature, and proposed a clarification. We have also observed that the above terminology misuse has brought about a literature of misleading experimental comparisons, with inductive transfer learning methods that have been incorrectly compared with transductive transfer learning methods. Our clarification has allowed a reassessment of the field, and of the relative merits of the major, state-of-the-art algorithms for transfer learning in text classification.

[1] https://cikmlq2021.github.io/
[2] https://lequa2022.github.io/
[3] https://clef2022.clef-initiative.eu/

Our more recent efforts address how to improve the accuracy of text classification (and other tasks too) in the presence of misspelled texts. For this we have devised a new type of embedded representations (called *garbled-word embeddings* – [63]) that find their roots in psycholinguistics, and that allow representing in a compact way the distributional semantics of entire classes of equivalence among words, where two words are considered equivalent if they are misspelled variations of each other.

1.3.3 Technology-assisted review

Technology-assisted review (TAR) is the task of supporting the work of human annotators who need to "review" automatically labelled data items, i.e., check the correctness of the labels assigned to these items by automatic classifiers. Since only a subset of such items can be feasibly reviewed, the goal of these algorithms is to exactly identify the items whose review is expected to be cost-effective. We have been working on this task since 2018, proposing TAR *risk minimization* algorithms that attempt to strike an optimal tradeoff between the contrasting goals of minimizing the cost of human intervention and maximizing the accuracy of the resulting labelled data. An aspect of TAR we have worked on more recently is improving the quality of the posterior probabilities that the risk minimization algorithm receives as input by an automated classifier. To this end, we have carried out a thorough study of SLD, an algorithm that, while being the state of the art in this task, had insufficiently been studied. Our study [24] has determined exactly in what conditions SLD can be expected to improve the quality of the posterior probabilities (and hence to be beneficial to the downstream TAR algorithm), and has determined that in other conditions SLD can instead bring about a deterioration of this quality.

1.3.4 Authorship analysis

Authorship analysis has to do with training predictors that infer characteristics of the author of a document of unknown paternity. We have worked on a sub-problem of authorship analysis called *authorship verification*, which consists of training a binary classifier that decides whether a text of disputed paternity is by a candidate author or not. Specifically, we have concentrated on a renowned case study, the so-called *Epistle to Cangrande*, written in medieval Latin apparently by Dante Alighieri, but whose authenticity has been disputed by scholars in the last century. To this end, we have built and made available to the scientific community two datasets of Medieval Latin texts, which we have used for training two separate predictors, one for the first part of the *Epistle* (which has a dedicatory nature) and one for the second part (which is instead a literary essay). The authorship verifiers that we have built indicate, although with different degrees of certainty, that neither the first nor the second part of the *Epistle* are by Dante. These predictions are corroborated by the fact that, once tested according to a leave-one-out experimental protocol on the two datasets, the two predictors exhibit extremely high accuracy [20].

An additional research we have carried out concerning authorship analysis for prose texts written in the Latin language, is the study of how features derived from "syllabic quantity" can impact the accuracy of predictors [19]. Syllabic quantity is an attribute of words, and it is well-known how different Latin authors used different syllabic quantity patterns in their writings. Our study has determined that extracting syllabic quantity from Latin prose texts is beneficial for authorship attribution.

1.4 AI for Mobility Analysis

1.4.1 Language modelling applied to trajectory classification

Mobility information collected by Location-Based Social Networks (e.g., Foursquare) allow modelling mobility at a more abstract and semantically rich level than simple geographical traces. These traces are called multiple-aspect trajectories, and include high level concepts, e.g., going to a theatre and then to an Italian restaurant, in addition to the geographical locations. Multiple-aspect trajectories enable to implement new services that exploit similarity models among users based on these high level concepts, rather than simple match of geographical locations. In this context a collaboration among AIMH, contributing the expertise on language modelling methods, colleagues of the High Performing Computing laboratory, and colleagues of the Universidade Federal de Santa Catarina (Florianópolis, Brazil) led to the development of a novel method for semantic trajectory modelling and classification, Multiple-Aspect tRajectory Classifier (MARC) [39]. MARC uses a trajectory embeddings method derived from the Word2Vec model and then a recurrent neural network to recognize the user who generated it, achieving state-of-the-art results.

1.5 Computer Vision

1.5.1 Learning from Virtual Worlds

In the new spring of artificial intelligence, particularly in its sub-field known as machine learning, a significant series of important results have shifted the focus of industrial and research communities toward the generation of valuable data from which learning algorithms can be trained. In the era of big data, the availability of real input examples to train machine learning algorithms is not considered an issue for several applications. However, there is not such an abundance of training data for several other applications. Sometimes, even if data is available, it must be manually revised to make it usable as training data (e.g., by adding annotations, class labels, or visual masks), with a considerable cost. Although a series of annotated datasets are available and successfully used to produce noteworthy academic results and commercially profitable products, there is still a considerable amount of scenarios where laborious human intervention is needed to produce high-quality training sets. An appealing solution is to gather synthetic data from virtual environments resembling the real world, where the labels are *automatically* collected interacting with the graphical engine. However, data coming from virtual worlds cannot be fully exploited due to the

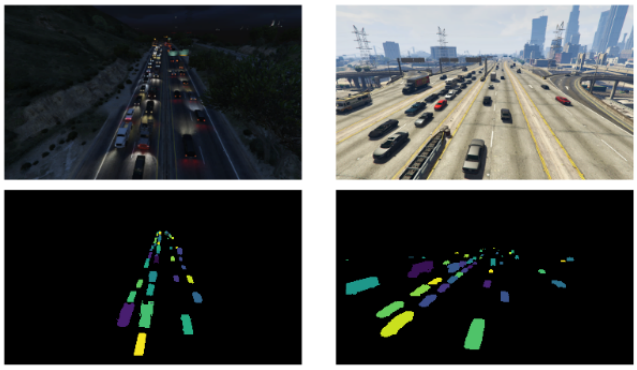

Figure 1. Some examples of images of our *Grand Traffic Auto* dataset, together with the *automatically* generated instance segmentation annotations.

Synthetic-to-Real Domain Shift, i.e., the image appearance difference between the synthetic training data and the real-world ones on which the AI-based algorithm, in the end, shall be used. This domain gap between the two data distributions leads to performance degradation at test time. To mitigate this domain gap, we propose a new methodology [13, 14] to design image-based vehicle density estimators and counting via an *Unsupervised Domain Adaptation (UDA)* technique. In particular, during the training phase, we exploit the supervised learning provided by the synthetic automatically labeled data exploiting the *Grand Traffic Auto (GTA)* dataset, the first collection of images with precise *per-pixel* annotations gathered using the graphical engine of a video game, and, at the same time, we infer some knowledge from the real-world *unlabeled* images. In other words, we tackle the problem of data scarcity from two complementary sides: on the one hand, we exploit the significant variability of the synthetic data, while, on the other hand, we mitigate the domain gap existing between the synthetic and the real-world images in an *unsupervised* fashion. We show some sample of the GTA dataset in Fig. 1.

1.5.2 Visual Counting

The counting task aims to estimate the number of objects instances, like people or vehicles, in still images or video frames. Due to its inter-disciplinary and widespread applicability, it has recently drawn the attention of the scientific community. Current solutions are formulated as supervised deep learning-based problems belonging to one of two main categories: counting by *detection* and counting by *regression*. Detection-based approaches require prior detection of the single instances of objects. On the other hand, regression-based techniques try to establish a direct mapping between the image features and the number of objects in the scene, either directly or via the estimation of a target map, such as a density map(i.e., a continuous-valued). Regression techniques show superior performance in crowded and highly-occluded scenarios but often lose the ability to locate objects precisely.

In [12], we propose a novel solution to improve car counting in parking lots when scaled up with multi-camera setups. We introduce a multi-camera system that combines a CNN-based technique, which can locate and count vehicles present in images belonging to individual cameras, along with a decentralized geometry-based approach that is responsible for aggregating the data gathered from all the devices and estimating the number of cars present in the *entire* parking lot. A remarkable peculiarity of our solution is that it performs the task directly on the *edge* devices, i.e., the smart cameras — vision systems with limited computational capabilities able to capture images, extract information from them, make decisions, and communicate with other devices.

Recently, we also tackled the task of counting cells in microscopy images [10]. We describe more in details this activity in Section 1.7.1. Furthermore, we also introduced an approach to estimate traffic density and counting vehicles in urban scenarios, exploiting synthetic data [13, 14]. We provide a more in-depth description of our solution in Section 1.5.1.

1.5.3 Facial Expression Recognition

Facial expressions play a fundamental role in human communication. Their study, which represents a multidisciplinary subject, embraces a great variety of research fields, e.g., psychology, computer science, among others. Concerning DL, recognizing facial expressions is a task named Facial Expression Recognition (FER). With such an objective, the goal of a learning model is to classify human emotions starting from a facial image of a given subject. Typically, face images are acquired by cameras that have, by nature, different characteristics, such as the output resolution. Moreover, other circumstances might involve cameras placed far from the observed scene, thus obtaining faces with very low resolutions. Therefore, since the FER task might involve analyzing face images that can be acquired with heterogeneous sources, it is plausible to expect that resolution plays a vital role. In such a context, the AIMH group proposes a multi-resolution training approach to solve the FER task ([35], [7], [34]). We grounded our intuition on the observation that, often, face images are acquired at different resolutions. Thus, directly considering such property while training a model can help achieve higher performance on recognizing facial expressions. We show in Fig. 2 an example of the output of our solution.

1.6 Multimedia Information Retrieval

1.6.1 Video Browsing

Video data is the fastest growing data type on the Internet, and because of the proliferation of high-definition video cameras, the volume of video data is exploding. This data explosion in the video area has led to push research on large-scale video retrieval systems that are effective, fast, and easy to use for content search scenarios.

Within this framework, we developed a content-based video retrieval system VISIONE[4], to compete at the Video Browser Showdown (VBS), an international video search competition that evaluates the performance of interactive video

[4] http://visione.isti.cnr.it/

Figure 2. Example of the output of our Facial Expression Recognition system. We recognize human facial expressions by automatically analyzing images.

retrievals systems. The tasks evaluated during the competition are: *Known-Item-Search (KIS)*, *textual KIS* and *Ad-hoc Video Search (AVS)*. The visual KIS task models the situation in which someone wants to find a particular video clip that he has already seen, assuming that it is contained in a specific collection of data. In the textual KIS, the target video clip is no longer visually presented to the participants of the challenge but it is rather described in details by text. This task simulates situations in which a user wants to find a particular video clip, without having seen it before, but knowing the content of the video exactly. For the AVS task, instead, a textual description is provided (e.g. "A person playing guitar outdoors") and participants need to find as many correct examples as possible, i.e. video shots that fit the given description.

Figure 3. VISIONE User Interface

VISIONE can be used to solve both Known-Item and Ad-hoc Video Search tasks as it integrates several content-based analysis and retrieval modules, including a keyword search, a spatial object-based search, a spatial color-based search, and a visual similarity search. The user interface, shown in Figure 3, provides a text box to specify the keywords, and a canvas for sketching objects and colors to be found in the target video.

VISIONE is based on state-of-the-art deep learning approaches for the visual content analysis and exploits highly efficient indexing techniques to ensure scalability. In particular, it uses specifically designed textual encodings for indexing and searching video content. This aspect of our system is crucial: we can exploit the latest text search engine technologies, which nowadays are characterized by high efficiency and scalability, without the need to define a dedicated data structure or even worry about implementation issues.

A detailed description of all the functionalities included in VISIONE and how each of them are implemented is provided in [2]. Moreover, in [1] we presented an analysis of the system retrieval performance, by examining the logs acquired during the VBS 2019 challenge.

1.6.2 Similarity Search

Searching a data set for the most similar objects to a given query is a fundamental task in many branches of computer science, including pattern recognition, computational biology, and multimedia information retrieval, to name but a few. This search paradigm, referred to as *similarity search*, overcomes limitations of traditional *exact-match search* that is neither feasible nor meaningful for complex data (e.g., multimedia data, vectorial data, time-series, etc.). In our research, we mainly focus on *metric search* methods, which are based on the assumption that data objects are represented as elements of a space (D,d) where the metric function d provides a measure of the closeness (i.e. dissimilarity) of the data objects. A proximity query is defined by a query object $q \in D$ and a proximity condition, such as "find all the objects within a threshold distance of q" (*range query*) or "finding the k closest objects to q" (*k-nearest neighbour query*). The exact response to a query is the set of all the data objects that satisfy the considered proximity condition.

Providing an exact response to a proximity query is not feasible if the search space is very large or it has a high intrinsic dimensionality since in such cases, the exact search rarely outperforms a sequential scan (phenomenon known as the *curse of dimensionality*). To overcome this issue, the research community has developed a wide spectrum of techniques for *approximate search*, which have higher efficiency though at the price of some imprecision in the results (e.g. some relevant results might be missing or some ranking errors might occur).

In the past, we developed and proposed various techniques to support approximate similarity search in metric spaces. Many of those techniques exploits the idea of transforming the original data objects into a more tractable space in which we can efficiently perform the search. For example, we proposed several *Permutation-Based Indexing* approaches where data objects are represented as a sequence of identifiers (*permutation*) that can be efficiently indexed and searched (e.g., by using inverted files). In the last years, we also investigated the use of some geometrical properties (namely, the 4-point property and the n-point property) to support metric search. For the class of metric space that satisfy the 4-point

property, called *Supermetric spaces*, we derived a new pruning rule named *Hilbert Exclusion*, which can be used with any indexing mechanism based on hyperplane partitioning in order to determine subset of data that do not need to be exhaustively inspected. Moreover, for the large class of metric spaces meeting the n-point property (notably including Cartesian spaces of any dimension with the Euclidean, Cosine or Quadratic Form distances) we defined the *nSimplex projection* that allows mapping metric objects into a finite-dimensional Euclidean space where upper- and lower- bounds of the actual distance can be computed. Spaces having the n-point property also meet the 4-point property.

In the context of approximate metric search, in 2021, we worked on generalizing the definition of permutations associated to metric objects by introducing the concept of permutations induced by a metric transformation f. It is worth noting that the Permutation-based Indexing approaches have been proved to be particularly suitable for dealing with large data collections. These methods employ a permutation-based representation of the data, which can be efficiently indexed using data structures such as inverted files. In the literature, the definition of the permutation of a metric object was derived by reordering the distances of the object to a set of pivots (reference objects). In [68], we generalized this definition in order to enlarge the class of permutations that can be used by PBI approaches. As a practical outcome, we defined a new type of permutation that is induced by a combination of pivots and the tensor product of several planar projections related to some pivot pairs. The proposed technique produces longer permutations than traditional ones for the same number of object-pivot distance calculations. The advantage is that the use of inverted files built on permutation prefixes leads to greater efficiency in the search phase when longer permutations are used.

During 2021, we further investigated the use of the 4-point and n-point properties for Approximate Nearest Neighbor search. In particular, in [69] we presented an approach that exploits a pivot-based local embedding to refine a set of candidate results of a similarity query. We focused our attention on refining of a set of approximate nearest neighbour results retrieved using a permutation-based search system. However, our approach can be generalized to other types of approximate search provided that they are based on the use of anchor objects (pivots) from which we pre-calculate the distances for other purposes. The core idea of the proposed technique is using the distances between an object and a set of pivots (pre-computed at indexing time) to embed the data objects into a low-dimensional where it is possible to compute upper- and lower-bounds for the actual distance (e.g. using the nSimplex projection). Dissimilarity functions defined upon those bounds are then adopted for re-ranking the candidate objects. The main advantage is that the proposed refining approach does not need to access the original data as done, instead, by the most commonly used refining technique that relies on computing the actual distances between the query and each candidate object. Moreover, in [67], we presented a method to obtain good distance bounds between a query and all the database elements using a minimally-sized representation comprising only two reference object identifiers, and two floating point values, per database object. The two floating point values are the coordinates in a two-dimensional Euclidean space (obtained using the nSimplex projection with $n = 2$ reference objects) where a lower-bound for the actual distance to a query can be efficiently computed. The caveat is that the mapping is local: in other words, each object is mapped using a different mapping. Indeed, each data object is associated with a pair of reference objects that is well-suited to filter that particular object, in cases where it is not relevant to a query. This maximises the probability of excluding that object from a search. At query time, a query is compared with a pool of reference objects which allow its mapping to all the planes used by data objects. Then, for each query/object pair, a lower bound of the actual distance is obtained.

1.6.3 Relational Cross-Modal Visual-Textual Retrieval

In the growing area of computer vision, modern deep-learning architectures are quite good at tasks such as classifying or recognizing objects in images. Recent studies, however, demonstrated the difficulties of such architectures to intrinsically understand a complex scene to catch spatial, temporal and abstract relationships among objects. Motivated by these limitations of the content-based information retrieval methods, we tried to explicitly handle relationships in multi-modal data, using attentive models. Specifically, we addressed the problem of cross-modal visual-textual retrieval, which consists in finding pictures given a natural language description as a query (sentence-to-image retrieval) or vice-versa (image-to-sentence retrieval). This task requires a deep understanding of both intra and inter-modal relationships to be effectively solved. We initially tackled the sentence-to-image retrieval scenario, as it is the more attractive in real world use-cases. We extended the Transformer Encoder Reasoning Network (TERN) [47], a deep relational neural network which is able to match images and sentences in a highly-semantic common space. In particular, we proposed TERAN (Transformer Encoder Reasoning and Alignment Network) [45] which is able to obtain a fine-grained region-word alignment keeping the context into consideration. However, the network is still supervised at a global image-sentence level, and the fine-grained correspondences are automatically discovered. With this constraint during the learning phase, we obtained state-of-the-art results on the Recall@K metrics and on the novel NDCG metric with ROUGE-L and SPICE textual similarities used as relevances. This novel network effectively produces visually pleasant precise region-word alignments, and we also demonstrated how the fine-grained region-word alignment objective improves the retrieval effectiveness of the original TERN cross-modal descriptions. The core of the architecture is constituted of recently introduced deep relational modules called *transformer encoders*, which can spot out hidden intra-object relationships. We showed that this simple pipeline is able to create compact relational cross-modal descriptions that

can be used for efficient similarity search.

After measuring the effectiveness of these cross-modal attentive models, we moved towards efficiency concerns. In particular, in [46] we proposed employing existing sparsification and quantization techniques (scalar quantization and deep permutation) to obtain features suitable for large-scale cross-modal retrieval. These techniques enable the use of off-the-shelf textual search tools for indexing any $\mathscr{R}^n$ vector, allowing the exploitation of well-established and efficient text-based indexes. The results from this research were recently deployed in a real-world use case. Namely, the efficient cross-modal search using the TERN features was integrated into VISIONE, a large-scale video retrieval tool developed by our group (see Section 1.6.1).

Another application of the same attentive technologies has been used in [48]. In this work, we proposed a Transformer-based model for detecting persuasion techniques in memes from social networks. A meme is a photo with some overlayed text, which potentially carries misleading information for political propaganda. We participated to the SemEval 2021 competition, placing at the 4-th position on the public leaderboard.

1.7 Medical Imaging

During this year, we applied our expertise on vision-based AI systems to research and develop healthcare and life science applications in collaboration with the Institute of Neuroscience of the CNR of Pisa. Our activity concerned the automatic analysis of medical images such as cell counting in fluorescence microscopy images and real-time pupillometry in IR images on mice and humans subjects. We describe the activities in detail in the following sections.

1.7.1 Cell Detection and Counting

Counting cells in microscopy images is an essential yet challenging task crucial for the diagnosing of many diseases. Recently, several vision models (mostly based on Convolutional Neural Networks) have been successfully adopted to count cells and other biological structures from microscopy images. However, the performance of these techniques is often measured only considering the counting errors occurring at inference time (i.e., the difference between the predicted and the actual cell numbers), which often leads to masked mistaken estimations. Indeed, counting errors do not take into account *where* the cells have been localized in the images and, consequently, counting models might achieve low values of errors while providing wrong predictions (e.g., a high number of false positives and false negatives). Therefore, it is hard to perform a fair comparison between the different state-of-the-art cell counting approaches to determine which performs best. In [10], we investigate three baseline solutions belonging to the three main counting methodologies — a *segmentation-based* approach, a *localization-based* approach, and a *count-density estimation* approach — that have been successfully exploited for counting several different categories of objects, such as people and vehicles, and that represent the conceptual basis also for the cell counting techniques. We conduct experiments on three public datasets containing different cell types and characterized by distinct peculiarities. In addition to comparing the performance of investigated methods against state-of-the-art cell counters using established counting evaluation metrics, we also measure the ability of the models to localize the counted cells correctly. We show that commonly adopted *counting* metrics (like mean absolute error) do not always agree with the *localization* performance of the tested models, and thus we suggest measuring both whenever possible to facilitate the practitioner in picking the most suitable solution.

1.7.2 Pupillometry

Pupil dynamics alterations have been found in patients affected by a variety of neuropsychiatric conditions, including autism. Studies in mouse models have used pupillometry for phenotypic assessment and as a proxy for arousal. Both in mice and humans, pupillometry is noninvasive and allows for longitudinal experiments supporting temporal specificity; however, its measure requires dedicated setups. In [40], we introduce a convolutional neural network that performs online pupillometry in both mice and humans in a web app format. This solution dramatically simplifies the usage of the tool for the nonspecialist and nontechnical operators. Because a modern web browser is the only software requirement, this choice is of great interest given its easy deployment and setup time reduction. The tested model performances indicate that the tool is sensitive enough to detect both locomotor-induced and stimulus-evoked pupillary changes, and its output is comparable to state-of-the-art commercial devices. Our tool is available at `https://www.pupillometry.it`.

1.8 Quantum Machine Learning

In recent years, Quantum Computing witnessed massive improvements both in terms of resources availability and algorithms development. The ability to harness quantum phenomena to solve computational problems is a long-standing dream that has drawn the scientific community's interest since the late '80s. In this regard, quantum computers might offer new solutions that exploit quantum phenomena such as interference, superposition, and entanglement. Such a characteristic is expected to speed up the computational time and to reduce the requirements for extensive resources, yielding the concepts of *quantum advantage* and *quantum supremacy*. In the last two decades, there has been a strong interest and commitment in the scientific community to develop quantum algorithms to solve Machine Learning problems, giving life to the field of *Quantum Machine Learning*.

In such a context, we posed our contribution [38]. First, we provided a gentle introduction to several basic notions about quantum mechanics, quantum information, and quantum computational models. Finally, we gathered, compared and analyzed the current state-of-the-art concerning Quantum Perceptrons and Quantum Neural Networks by discerning among theoretical formulations, simulations, and implemen-

tations on real quantum devices. The result is a thorough survey on the field of Quantum Neural Networks, which can be used as a guide by beginners, as well as a reference for more experienced practitioners. Moreover, we collected and organized the most relevant papers on this field on a GitHub page[5] allowing the interested readers to easily and quickly browse through the research literature.

1.9 Fighting misinformation

1.9.1 Deep Fake Detection

Deep fake detection is a critical task in the modern society, where increasingly powerful generative methods are used to craft fake images, videos, or fake news through *social bots*. All this ad-hoc generated content is spread on the web usually via social networks, and it is used to propagate misinformation and fake news, with the aim of contaminating public debate. Deep fake images and videos can be used to harm important and strategic public figures. For this reason, it is very important to promptly detect them to stop their diffusion. Although many methods focused on image deep fake detection, in [16] we tackled deep fake detection in videos. The challenge is identifying if there are people having their face replaced or manipulated. In particular, we used a mixed Transformer-Convolutional model to attend the face patches. Differently from current state-of-the-art approaches, we use neither distillation nor ensemble methods, and we obtained remarkable results on the DeepFake Detection Challenge (DFDC) and on FaceForensics++ datasets. In addition to proposing new hybrid architectures to deal with deepfake video detection, alternative approaches to efficient and effective inference were analysed in this study. Indeed, at inference time, the faces from different frames of the video are independently analyzed, grouped and a simple voting algorithm is used to decide if the video shot was altered or not. With the proposed approach it is possible to better manage situations such as the presence of several people in the same video where only one has been manipulated so as to counter false negatives or attacks aimed at deceiving the detector. A Video Deepfake Detector could also be used on a large scale and therefore a short study was also carried out to identify the optimal number of faces to be classified within a video to achieve the best ratio of reliability and scalability of classification.

We have been also involved in research related to detection of deep fake tweets [28]. Despite the critical importance, few works tackled the detection of machine-generated texts on social networks like Twitter or Facebook. With the aim of helping the research in this detection field, in this work we collected the first dataset of real deepfake tweets, TweepFake. It is real in the sense that each deepfake tweet was actually posted on Twitter by social bots. With the aim of showing the challenges that TweepFake poses and providing a solid baseline of detection techniques, we also evaluated 13 different deepfake text detection methods. Some of the detectors exploit text representations as inputs to machine-learning classifiers, others are based on deep learning networks, and others rely on the fine-tuning of transformer-based classifiers. A comprehensive analysis of these techniques showed how the newest and more sophisticated generative methods based on the transformer architecture (e.g., GPT-2) can produce high-quality short texts, difficult to unmask also for expert human annotators. Additionally, the transformer-based language models provide very good word representations for both text representation-based and fine-tuning based detection techniques.

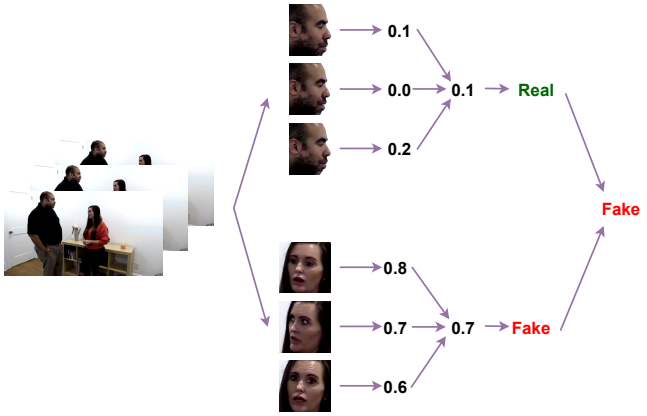

Figure 4. Video deepfake inference strategy. Faces are detected across multiple frames. Then, a voting algorithm decides if at least one of them was manipulated. Image from [16].

1.9.2 Detection of Persuasion Techniques

Social networks play a critical role in our society. Nowadays, most of the ideas, thoughts, and political beliefs are shared through the internet using social platforms like Twitter, Facebook, or Instagram. Although these online services enable information to be spread efficiently and effectively, it is non-trivial to understand if the shared contents are free of subtle meanings altering people's judgment abilities.

In [48] we tackle the problem of recognizing which kind of disinformation technique is used to forge *memes* for a disinformation campaign. Memes are small yet effective units of information able to spread cultural ideas, symbols, or practices and usually exist under the form of pictures, possibly with overlaid text. Memes are created so that they can propagate rapidly and reach a large number of users; for this reason, they are one of the most popular types of content used in an online disinformation campaign In particular, we proposed an architecture based on the well-established Transformer architecture model for processing both the textual and visual inputs from the meme. This architecture, called DVTT (Double Visual Textual Transformer), comprises two full Transformer networks working respectively on images and texts; each of these Transformers is conditioned on the other modality. We consider this task as a multi-label classification problem, where text and/or images from the meme are processed, and proba-

[5] https://github.com/fvmassoli/survey-quantum-computation

bilities of presence of each possible persuasion technique are returned as a result. Our proposed model reached remarkable results on the publicly available leaderboard of the *SemEval 2021 Task 6* challenge[6].

2. Projects & Activities

2.1 EU Projects

AI4EU

In January 2019, the AI4EU consortium was established to build the first European Artificial Intelligence On-Demand Platform and Ecosystem with the support of the European Commission under the H2020 programme. The activities of the AI4EU project include:

- The creation and support of a large European ecosystem spanning the 28 countries to facilitate collaboration between all Europeans actors in AI (scientists, entrepreneurs, SMEs, Industries, funding organizations, citizens...);
- The design of a European AI on-Demand Platform to support this ecosystem and share AI resources produced in European projects, including high-level services, expertise in AI research and innovation, AI components and datasets, high-powered computing resources and access to seed funding for innovative projects using the platform;
- The implementation of industry-led pilots through the AI4EU platform, which demonstrates the capabilities of the platform to enable real applications and foster innovation; Research activities in five key interconnected AI scientific areas (Explainable AI, Physical AI ,Verifiable AI, Collaborative AI, Integrative AI), which arise from the application of AI in real-world scenarios;
- The funding of SMEs and start-ups benefitting from AI resources available on the platform (cascade funding plan of €3M) to solve AI challenges and promote new solutions with AI; The creation of a European Ethical Observatory to ensure that European AI projects adhere to high ethical, legal, and socio-economical standards;
- The production of a comprehensive Strategic Research Innovation Agenda for Europe; The establishment of an AI4EU Foundation that will ensure a handover of the platform in a sustainable structure that supports the European AI community in the long run.

The leader of the AIMH team participating in AI4EU is Giuseppe Amato.

[6] https://propaganda.math.unipd.it/semeval2021task6/index.html

AI4media

Artificial Intelligence for the Society and the Media Industry (AI4Media) is a network of research excellence centres delivering advances in AI technology in the media sector. Funded under H2020-EU.2.1.1., AI4Media started in September 2020 and will end in August 2024.

Motivated by the challenges, risks and opportunities that the wide use of AI brings to media, society and politics, AI4Media aspires to become a centre of excellence and a wide network of researchers across Europe and beyond, with a focus on delivering the next generation of core AI advances to serve the key sector of Media, to make sure that the European values of ethical and trustworthy AI are embedded in future AI deployments, and to reimagine AI as a crucial beneficial enabling technology in the service of Society and Media.

The leader of the AIMH team participating in AI4Media is Fabrizio Sebastiani.

ARIADNEplus

The ARIADNEplus project is the extension of the previous ARIADNE Integrating Activity, which successfully integrated archaeological data infrastructures in Europe, indexing in its registry about 2.000.000 datasets. ARIADNEplus will build on the ARIADNE results, extending and supporting the research community that the previous project created and further developing the relationships with key stakeholders such as the most important European archaeological associations, researchers, heritage professionals, national heritage agencies and so on. The new enlarged partnership of ARIADNEplus covers all of Europe. It now includes leaders in different archaeological domains like palaeoanthropology, bioarchaeology and environmental archaeology as well as other sectors of archaeological sciences, including all periods of human presence from the appearance of hominids to present times. Transnational Activities together with the planned training will further reinforce the presence of ARIADNEplus as a key actor. The technology underlying the project is state-of-art. The ARIADNEplus data infrastructure will be embedded in a cloud that will offer the availability of Virtual Research Environments where data-based archaeological research may be carried out. The project will furthermore develop a Linked Data approach to data discovery. Innovative services will be made available to users, such as visualization, annotation, text mining and geo-temporal data management. Innovative pilots will be developed to test and demonstrate the innovation potential of the ARIADNEplus approach. Fostering innovation will be a key aspect of the project, with dedicated activities led by the project Innovation Manager.

Mingei

The Mingei Project explores the possibilities of representing and making accessible both tangible and intangible aspects of craft as cultural heritage (CH). Heritage Crafts (HCs) involve craft artefacts, materials, and tools and encompass craftsmanship as a form of Intangible Cultural Heritage. Intangible HC dimensions include dexterity, know-how, and skilled use of tools, as well as, tradition, and identity of the communities in which they are, or were, practiced. HCs are part of the history and have impact upon the economy of the areas in which they flourish. The significance and urgency to the preservation of HCs is underscored, as several are threatened with extinction. Despite their cultural significance efforts for HC representation and preservation are scattered geographically and thematically. Mingei provides means to establish HC representations based on digital assets, semantics, existing literature and repositories, as well as, mature digitisation and representation technologies. These representations will capture and preserve tangible and intangible dimensions of HCs. Central to craftsmanship is skill and its transmission from master to apprentice. Mingei captures the motion and tool usage of HC practitioners, from Living Human Treasures and archive documentaries, in order to preserve and illustrate skill and tool manipulation. The represented knowledge will be availed through experiential presentations, using storytelling and educational applications and based on Advanced Reality, Mixed Reality and the Internet. The project has started on December 1, 2019 and will last 3 years.

Multi-Foresee

The main objective of this Action, entitled MULTI-modal Imaging of FOREnsic SciEnce Evidence (MULTI-FORESEE)-tools for Forensic Science[7], is to promote innovative, multi-informative, operationally deployable and commercially exploitable imaging solutions/technology to analyse forensic evidence.

Forensic evidence includes, but not limited to, fingermarks, hair, paint, biofluids, digital evidence, fibers, documents and living individuals. Imaging technologies include optical, mass spectrometric, spectroscopic, chemical, physical and digital forensic techniques complemented by expertise in IT solutions and computational modelling.

Imaging technologies enable multiple physical and chemical information to be captured in one analysis, from one specimen, with information being more easily conveyed and understood for a more rapid exploitation. The enhanced value of the evidence gathered will be conducive to much more informed investigations and judicial decisions thus contributing to both savings to the public purse and to a speedier and stronger criminal justice system.

The Action will use the unique networking and capacity-building capabilities provided by the COST framework to bring together the knowledge and expertise of Academia, Industry and End Users. This synergy is paramount to boost imaging technological developments which are operationally deployable.

The leader of the AIMH team participating in MultiForesee is Giuseppe Amato.

SoBigData

SoBigData++ is a project funded by the European Commission under the H2020 Programme INFRAIA-2019-1, started Jan 1 2020 and ending Dec 31, 2023. SoBigData++ proposes to create the Social Mining and Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. SoBigData plans to open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research. It plans to not only strengthen the existing clusters of excellence in social data mining research, but also create a pan-European, inter-disciplinary community of social data scientists, fostered by extensive training, networking, and innovation activities.

The leader of the AIMH team participating in SoBigData++ is Alejandro Moreo.

SSHOC

Social Sciences & Humanities Open Cloud (SSHOC) is a project funded by the EU framework programme Horizon 2020 and unites 20 partner organisations and their 27 associates in developing the social sciences and humanities area of the European Open Science Cloud (EOSC). SSHOC partners include both developing and fully established European Research Infrastructures from the social sciences and humanities, and the association of European research libraries (LIBER). The goal of the project is to transform the social sciences & humanities data landscape with its disciplinary silos and separate facilities into an integrated, cloud-based network of interconnected data infrastructures. To promote synergies and open science initiatives between disciplines, and accelerate interdisciplinary research and collaboration, these data infrastructures will be supported by the tools and training which allow scholars and researchers to access, process, analyse, enrich and compare data across the boundaries of individual repositories or institutions. SSHOC will continuously moni-

[7] https://multiforesee.com/

tor ongoing developments in the EOSC so as to conform to the necessary technical and other requirements for making the SSHOC services sustainable beyond the duration of the project. Some of the results obtained by the AIMH team involved in SSHOC have been presented in [NN] The leader of the AIMH team participating in SSHOC is Cesare Concordia. https://sshopencloud.eu

2.2 CNR National Virtual Lab on AI

Fabrizio Falchi has coordinated, together with Sara Colantonio, the activities of the National Virtual Lab of CNR on Artificial Intelligence. This initiative connects about 90 groups in 22 research institutes of 6 departments of the whole CNR. The Nationtal Virtual Lab on AI aims at proposing a strategic vision and big and long-term projects.

2.3 National Projects

AI-MAP

AI-MAP is a project funded by Regione Toscana that aims at analyzing digitized historical geographical regional maps using deep learning methods to increase the availability and searchability of the digitized documents. The main objectives of the project is to develop automatic or semi-automatic pipelines for denoising/repairing of the digitized documents, handwritten toponym localization and transcription. The activities are mainly conducted in the context of this project by Fabio Carrara under the scientific coordination of Giuseppe Amato.

AI4CHSites

AI4CHSites is a project funded by Regione Toscana that aims at analyzing visual content from surveillance camera in a touristic scenario. Partners of the project are: Opera della Primaziale Pisana and INERA srl. The activities in the context of this project are mainly conducted by Nicola Messina under the scientific coordination of Fabrizio Falchi.

ADA

In the era of Big Data, manufacturing companies are overwhelmed by a lot of disorganized information: the large amount of digital content that is increasingly available in the manufacturing process makes the retrieval of accurate information a critical issue. In this context, and thanks also to the Industry 4.0 campaign, the Italian manufacturing industries have made a lot of effort to ameliorate their knowledge management system using the most recent technologies, like big data analysis and machine learning methods. In this context, therefore, the main target of the ADA project is to design and develop a platform based on big data analytics systems that allows for the acquisition, organization, and automatic retrieval of information from technical texts and images in the different phases of acquisition, design & development, testing, installation and maintenance of products.

HDN

Hypermedia Dante Network (HDN) is a three year (2020-2023) Italian National Research Project (PRIN) which aims to extend the ontology and tools developed by AIMH team to represent the sources of Dante Alighieri's minor works to the more complex world of the Divine Comedy. In particular, HDN aims to enrich the functionalities of the DanteSources Web application (https://dantesources.dantenetwork.it/) in order to efficiently recover knowledge about the Divine Comedy. Relying on some of the most important scientific institutions for Dante studies, such as the Italian Dante Society of Florence, HDN makes use of specialized skills, essential for the population of ontology and the consequent creation of a complete and reliable knowledge base. Knowledge will be published on the Web as Linked Open Data and will be access through a user-friendly Web application.

IMAGO

The IMAGO (Index Medii Aevi Geographiae Operum) is a three year (2020-2023) Italian National Research Project (PRIN) that aims at creating a knowledge base of the critical editions of Medieval and Humanistic Latin geographical works (VI-XV centuries). Up to now, this knowledge has been collected in many paper books or several databases, making it difficult for scholars to retrieve it easily and to produce a complete overview of these data. The goal of the project is to develop new tools that satisfy the needs of the academic research community, especially for scholars interested in Medieval and Renaissance Humanism geography. Using Semantic Web technologies, AIMH team will develop an ontology providing the terms to represent this knowledge in a machine-readable form. A semi-automatic tool will help the scholars to populate the ontology with the data included in authoritative critical editions. Afterwards, the tool will automatically save the resulting graph into a triple store. On top of this graph, a Web application will be developed, which will allow users to extract and display the information stored in the knowledge base in the form of maps, charts, and tables.

WAC@Lucca

WeAreClouds@Lucca carries out research and development activities in the field of monitoring public places, such as squares and streets, through cameras and microphones with artificial intelligence technologies, in order to collect useful information both for the evaluation of tourist flows and their impact. on the city, both for purposes of automatic identification of particular events of interest for statistical purposes or for security. The project is funded by Fondazione Cassa di Risparmio di Lucca and Comune di Lucca is a partner. Fabrizio Falchi is the scientific coordinator of the project.

NAUSICAA

NAUSICAA - "NAUtical Safety by means of Integrated Computer-Assistance Appliances 4.0" is a project funded by the Tuscany region (CUP D44E20003410009). The project aims at creating a system for medium and large boats in which the conventional control, propulsion, and thrust systems are integrated with a series of latest generation sensors (e.g., lidar systems, cameras, radar, drones) for assistance during the

navigation and mooring phases. In the project, the AIMH researchers are mainly involved in developing techniques for the automatic analysis of video streams from cameras on boats and aerial drones based on artificial intelligence methods. Models and methods will be developed in particular for the recognition and tracking of people and objects in the water (e.g. for rescuing people at sea).

3. Papers

In this section, we report the complete list of paper we published in 2021 organized in four categories: journals, proceedings, magazines, others, and pre-prints.

3.1 Journals

In this section, we report the paper we published (or accepted for publication) in journals during 2021, in alphabetic order of the first author. Our works were published in the following journals:

- **Transactions on Neural Networks and Learning Systems**
  IEEE, IF 10.5: [37]
- **Neural Networks**
  Elsevier, IF 8.0: [30]
- **Neural Computing and Applications**
  Springer, IF 5.6: [31] (Accepted)
- **ACM Transactions on Information Systems**
  ACM Press, IF : 4.8: [24]
- **eNeuro**
  Society of Neuroscience, IF 4.1: [40]
- **Computer Vision and Image Understanding**
  Elsevier, IF 3.9: [36]
- **Pattern Recognition Letters**
  Elsevier, IF 3.8: [44]
- **Data Mining and Knowledge Discvoery**
  SPringer Nature, IF 3.7: [53]
- **ACM Transactions on Multimedia Computing, Communications, and Applications**
  ACM, IF 3.1: [45]
- **PLoS ONE**
  Public Library Science, IF 3.2: [28, 58]
- **Multimedia Tools and Applications**
  Springer, IF 2.9: [23]
- **ACM Transactions on Knowledge Discovery**
  ACM Press, IF : 2.7: [50]
- **Information Systems**
  Elsevier, IF 2.3: [69, 67]
- **ACM Journal on Computing and Cultural Heritage**
  ACM Press, IF : 2.0: [20] (accepted)
- **Digital Scholarship in the Humanities**
  Oxford University Press, IF 0.9: [4, 5]
- **Journal of Imaging**
  MDPI: [1]
- **Heritage (Basel)**
  MDPI: [59]

3.1.1 The visione video search system: exploiting off-the-shelf text search engines for large-scale video retrieval

G. Amato, P. Bolettieri, F. Carrara, F. Debole, F. Falchi, C. Gennaro, L. Vadicamo, C. Vairo In Journal of Imaging, MDPI. [1]

This paper describes in detail VISIONE, a video search system that allows users to search for videos using textual keywords, the occurrence of objects and their spatial relationships, the occurrence of colors and their spatial relationships, and image similarity. These modalities can be combined together to express complex queries and meet users' needs. The peculiarity of our approach is that we encode all information extracted from the keyframes, such as visual deep features, tags, color and object locations, using a convenient textual encoding that is indexed in a single text retrieval engine. This offers great flexibility when results corresponding to various parts of the query (visual, text and locations) need to be merged. In addition, we report an extensive analysis of the retrieval performance of the system, using the query logs generated during the Video Browser Showdown (VBS) 2019 competition. This allowed us to fine-tune the system by choosing the optimal parameters and strategies from those we tested.

3.1.2 Towards a knowledge base of medieval and renaissance geographical Latin works: the IMAGO ontology

V. Bartalesi, D. Metilli, N. Pratelli, P. Pontari In Digital Scholarship in the Humanities. [4]

In this article we present the first achievement of the Index Medii Aevi Geographiae Operum (IMAGO)–Italian National Research Project (2020-23), that is, the ontology we have created in order to formally represent the knowledge about the geographical works written in Middle Ages and Renaissance (6th-15th centuries). The IMAGO ontology is derived from a strict collaboration between the Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR) and the scholars who are involved in the project, who have supported ISTI-CNR in defining a conceptualization of the domain of knowledge. Following the re-use logic, we have selected as reference ontologies the International Committee on Documentation CRM vocabulary and its extension FRBRoo, including its in-progress reformulation, LRMoo. This research is included in a wider project context whose final aim is the creation of a knowledge base (KB) of Latin geographic literature of the Middle Ages and Renaissance Humanism in which the data are formally represented following the Linked Open Data paradigm and using the Semantic Web languages. At the end of the project, this KB will be accessed through a Web application that allows retrieving and consulting the collected data in a user-friendly way for scholars and general users, e.g. tables, maps, CSV files.

3.1.3 A formal representation of the divine comedy's primary sources: The Hypermedia Dante Network ontology

V. Bartalesi, N. Pratelli, C. Meghini, D. Metilli, G. Tomazzoli, L.M.G. Livraghi, M. Zaccarello Digital Scholarship in the Humanities. [5]

Hypermedia Dante Network (HDN) is a 3-year Italian National Research Project, started in 2020, which aims to enrich the functionalities of the DanteSources Digital Library to efficiently represent knowledge about the primary sources of Dante's Comedy. DanteSources allows users to retrieve and visualize the list and the distribution of Dante's primary sources that have been identified by recent commentaries of five of Dante's minor works (i.e. Vita nova, De vulgari eloquentia, Convivio, De Monarchia, and Rime). The digital library is based on a formal ontology expressed in Resource Description Framework Schema (RDFS) language. Based on the DanteSources experience, the HDN project aims to formally represent the primary sources of the Divine Comedy whose identification is based on several commentaries included in the Dartmouth Dante Project corpus. To reach this goal, we restructured and extended the DanteSources ontology to provide a wider and more complete representation of the knowledge concerning the primary sources of the Comedy. In this article, we present the result of this effort, i.e. the HDN ontology. The ontology is expressed in OWL and has as reference ontologies the CIDOC CRM and its extension FRBRoo, including its in-progress reformulation LRMoo. We also briefly describe the semi-automatic tool that will be used by the scholars to populate the ontology.

3.1.4

MedLatinEpi and MedLatinLit: Two datasets for the computational authorship analysis of medieval Latin texts

S. Corbara, A. Moreo, F. Sebastiani, M. Tavoni. In ACM Journal on Computing and Cultural Heritage (accepted) (ACM Press). [20]

We present and make available MedLatinEpi and MedLatinLit, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatinEpi and MedLatinLit consist of 294 and 30 curated texts, respectively, labelled by author; MedLatinEpi texts are of epistolary nature, while MedLatinLit texts consist of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification. Along with the datasets we provide experimental results, obtained on these datasets, for the authorship verification task, i.e., the task of predicting whether a text of unknown authorship was written by a candidate author or not. We also make available the source code of the authorship verification system we have used, thus allowing our experiments to be reproduced, and to be used as baselines, by other researchers. We also describe the application of the above authorship verification system, using these datasets as training data, for investigating the authorship of two medieval epistles whose authorship has been disputed by scholars.

3.1.5

Learning accurate personal protective equipment detection from virtual worlds

M. Di Benedetto, F. Carrara, E. Meloni, G. Amato, F. Falchi, C. Gennaro In Multimedia Tools and Applications, Springer. [23]

Deep learning has achieved impressive results in many machine learning tasks such as image recognition and computer vision. Its applicability to supervised problems is however constrained by the availability of high-quality training data consisting of large numbers of humans annotated examples (e.g. millions). To overcome this problem, recently, the AI world is increasingly exploiting artificially generated images or video sequences using realistic photo rendering engines such as those used in entertainment applications. In this way, large sets of training images can be easily created to train deep learning algorithms. In this paper, we generated photo-realistic synthetic image sets to train deep learning models to recognize the correct use of personal safety equipment (e.g., worker safety helmets, high visibility vests, ear protection devices) during at-risk work activities. Then, we performed the adaptation of the domain to real-world images using a very small set of real-world images. We demonstrated that training with the synthetic training set generated and the use of the domain adaptation phase is an effective solution for applications where no training set is available.

3.1.6

A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment

A. Esuli, A. Molinari, F. Sebastiani. In ACM Transactions on Information Systems (ACM Press). [24]

We critically re-examine the Saerens-Latinne-Decaestecker (SLD) algorithm, a well-known method for estimating class prior probabilities ("priors") and adjusting posterior probabilities ("posteriors") in scenarios characterized by distribution shift, i.e., difference in the distribution of the priors between the training and the unlabelled documents. Given a machine-learned classifier and a set of unlabelled documents for which the classifier has returned posterior probabilities and estimates of the prior probabilities, SLD updates them both in an iterative, mutually recursive way, with the goal of making both more accurate; this is of key importance in downstream tasks such as single-label multiclass classification and cost-sensitive text classification. Since its publication, SLD has become the standard algorithm for improving the quality of the posteriors in the presence of distribution shift, and is still considered a top contender when we need to estimate the priors (a task that has become known as "quantification"). However, its real effectiveness in improving the quality of the posteriors has been questioned. We here present the results of systematic experiments conducted on a large, publicly available dataset, across multiple amounts of distribution shift and multiple learners. Our experiments show that SLD improves the quality of the posterior probabilities and of the estimates of the prior probabilities, but only when the number of classes in the classification scheme is very small and the classifier is calibrated. As the number of classes grows, or as we use non-calibrated classifiers, SLD converges more slowly (and often does not converge at all), performance degrades rapidly, and the impact of SLD on the quality of the prior estimates and of the posteriors becomes negative rather than positive.

3.1.7
TweepFake: About detecting deepfake tweets

T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi In Plos ONE. [28]

The recent advances in language modeling significantly improved the generative capabilities of deep neural models: in 2019 OpenAI released GPT-2, a pre-trained language model that can autonomously generate coherent, non-trivial and human-like text samples. Since then, ever more powerful text generative models have been developed. Adversaries can exploit these tremendous generative capabilities to enhance social bots that will have the ability to write plausible deepfake messages, hoping to contaminate public debate. To prevent this, it is crucial to develop deepfake social media messages detection systems. However, to the best of our knowledge no one has ever addressed the detection of machine-generated texts on social networks like Twitter or Facebook. With the aim of helping the research in this detection field, we collected the first dataset of real deepfake tweets, TweepFake. It is real in the sense that each deepfake tweet was actually posted on Twitter. We collected tweets from a total of 23 bots, imitating 17 human accounts. The bots are based on various generation techniques, i.e., Markov Chains, RNN, RNN+Markov, LSTM, GPT-2. We also randomly selected tweets from the humans imitated by the bots to have an overall balanced dataset of 25,572 tweets (half human and half bots generated). The dataset is publicly available on Kaggle. Lastly, we evaluated 13 deepfake text detection methods (based on various state-of-the-art approaches) to both demonstrate the challenges that Tweepfake poses and create a solid baseline of detection techniques. We hope that TweepFake can offer the opportunity to tackle the deepfake detection on social media messages as well.

3.1.8
Comparing the Performance of Hebbian against Backpropagation Learning using Convolutional Neural Networks

G. Lagani, F. Falchi, C. Gennaro, G. Amato In Neural Computing and Applications. [31] (Accepted)

We explore competitive Hebbian learning strategies to train feature detectors in Convolutional Neural Networks (CNNs), without supervision. We consider variants of the Winner-Takes-All (WTA) strategy explored in previous works, i.e. k-WTA, e-soft-WTA and p-soft-WTA, performing experiments on different object recognition datasets. Results suggest that the Hebbian approaches are effective to train early feature extraction layers, or to re-train higher layers of a pre-trained network, with soft competition generally performing better than other Hebbian approaches explored in this work. Our findings encourage a path of cooperation between neuroscience and computer science towards a deeper investigation of biologically inspired learning principles.

3.1.9
Hebbian semi-supervised learning in a sample efficiency setting

G. Lagani, F. Falchi, C. Gennaro, G. Amato In Neural Networks, Elsevier. [30]

We propose to address the issue of sample efficiency, in Deep Convolutional Neural Networks (DCNN), with a semi-supervised training strategy that combines Hebbian learning with gradient descent: all internal layers (both convolutional and fully connected) are pre-trained using an unsupervised approach based on Hebbian learning, and the last fully connected layer (the classification layer) is trained using Stochastic Gradient Descent (SGD). In fact, as Hebbian learning is an unsupervised learning method, its potential lies in the possibility of training the internal layers of a DCNN without labels. Only the final fully connected layer has to be trained with labeled examples. We performed experiments on various object recognition datasets, in different regimes of sample efficiency, comparing our semi-supervised (Hebbian for internal layers + SGD for the final fully connected layer) approach with end-to-end supervised backprop training, and with semi-supervised learning based on Variational Auto-Encoder (VAE). The results show that, in regimes where the number of available labeled samples is low, our semi-supervised approach outperforms the other approaches in almost all the cases.

3.1.10
MOCCA: Multilayer One-Class Classification for Anomaly Detection

F.V. Massoli, F. Falchi, A. Kantarci, S. Akti, H.K. Ekenel, G. Amato. In IEEE Transactions on Neural Networks and Learning Systems. [37].

Anomalies are ubiquitous in all scientific fields and can express an unexpected event due to incomplete knowledge about the data distribution or an unknown process that suddenly comes into play and distorts the observations. Usually, due to such events' rarity, to train deep learning (DL) models on the anomaly detection (AD) task, scientists only rely on "normal" data, i.e., nonanomalous samples. Thus, letting the neural network infer the distribution beneath the input data. In such a context, we propose a novel framework, named multilayer one-class classification (MOCCA), to train and test DL models on the AD task. Specifically, we applied our approach to autoencoders. A key novelty in our work stems from the explicit optimization of the intermediate representations for the task at hand. Indeed, differently from commonly used approaches that consider a neural network as a single computational block, i.e., using the output of the last layer only, MOCCA explicitly leverages the multilayer structure of deep architectures. Each layer's feature space is optimized for AD during training, while in the test phase, the deep representations extracted from the trained layers are combined to detect anomalies. With MOCCA, we split the training process into two steps. First, the autoencoder is trained on the reconstruction task only. Then, we only retain the encoder tasked with minimizing the L_2 distance between the output representation and a reference point, the anomaly-free training data centroid, at each considered layer. Subsequently, we combine the deep features extracted at the various trained layers of the encoder model to detect anomalies at inference time. To assess the performance of the models trained with MOCCA, we conduct extensive experiments on publicly available datasets, namely CIFAR10, MVTec AD, and ShanghaiTech. We show that our proposed method reaches comparable or superior performance to state-of-the-art approaches available in the literature. Finally, we provide a model analysis to give insights regarding the benefits of our training procedure.

3.1.11
Detection of face recognition adversarial attacks

F.V. Massoli, F. Carrara, G. Amato, F. Falchi In Computer Vision and Image Understanding, Elsevier. [36]

Deep Learning methods have become state-of-the-art for solving tasks such as Face Recognition (FR). Unfortunately, despite their success, it has been pointed out that these learning models are exposed to adversarial inputs – images to which an imperceptible amount of noise for humans is added to maliciously fool a neural network – thus limiting their adoption in sensitive real-world applications. While it is true that an enormous effort has been spent to train robust models against this type of threat, adversarial detection techniques have recently started to draw attention within the scientific community. The advantage of using a detection approach is that it does not require to re-train any model; thus, it can be added to any system. In this context, we present our work on adversarial detection in forensics mainly focused on detecting attacks against FR systems in which the learning model is typically used only as features extractor. Thus, training a more robust classifier might not be enough to counteract the adversarial threats. In this frame, the contribution of our work is four-fold: (i) we test our proposed adversarial detection approach against classification attacks, i.e., adversarial samples crafted to fool an FR neural network acting as a classifier; (ii) using a k-Nearest Neighbor (k-NN) algorithm as a guide, we generate deep features attacks against an FR system based on a neural network acting as features extractor, followed by a similarity-based procedure which returns the query identity; (iii) we use the deep features attacks to fool an FR system on the 1:1 face verification task, and we show their superior effectiveness with respect to classification attacks in evading such type of system; (iv) we use the detectors trained on the classification attacks to detect the deep features attacks, thus showing that such approach is generalizable to different classes of offensives.

3.1.12
MEYE: Web App for Translational and Real-Time Pupillometry

R. Mazziotti, F. Carrara, A. Viglione, L. Lupori, L. Lo Verde, A. Benedetto, G. Ricci, G. Sagona, G. Amato, T. Pizzorusso In eNeuro Vol. 8 Issue 5, Society of Neuroscience. [40]

Pupil dynamics alterations have been found in patients affected by a variety of neuropsychiatric conditions, including autism. Studies in mouse models have used pupillometry for phenotypic assessment and as a proxy for arousal. Both in mice and humans, pupillometry is noninvasive and allows for longitudinal experiments supporting temporal specificity; however, its measure requires dedicated setups. Here, we introduce a convolutional neural network that performs online pupillometry in both mice and humans in a web app format. This solution dramatically simplifies the usage of the tool for the nonspecialist and nontechnical operators. Because a modern web browser is the only software requirement, this choice is of great interest given its easy deployment and setup time reduction. The tested model performances indicate that the tool is sensitive enough to detect both locomotor-induced and stimulus-evoked pupillary changes, and its output is comparable to state-of-the-art commercial devices.

3.1.13
Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders

N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet In ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). [45].

Despite the evolution of deep-learning-based visual-textual processing systems, precise multi-modal matching remains a challenging task. In this work, we tackle the task of cross-modal retrieval through image-sentence matching based on word-region alignments, using supervision only at the global image-sentence level. Specifically, we present a novel approach called Transformer Encoder Reasoning and Alignment Network (TERAN). TERAN enforces a fine-grained match between the underlying components of images and sentences (i.e., image regions and words, respectively) to preserve the informative richness of both modalities. TERAN obtains state-of-the-art results on the image retrieval task on both MS-COCO and Flickr30k datasets. Moreover, on MS-COCO, it also outperforms current approaches on the sentence retrieval task. Focusing on scalable cross-modal information retrieval, TERAN is designed to keep the visual and textual data pipelines well separated. Cross-attention links invalidate any chance to separately extract visual and textual features needed for the online search and the offline indexing steps in large-scale retrieval systems. In this respect, TERAN merges the information from the two domains only during the final alignment phase, immediately before the loss computation. We argue that the fine-grained alignments produced by TERAN pave the way toward the research for effective and efficient methods for large-scale cross-modal information retrieval. We compare the effectiveness of our approach against relevant state-of-the-art methods. On the MS-COCO 1K test set, we obtain an improvement of 5.7% and 3.5% respectively on the image and the sentence retrieval tasks on the Recall@1 metric. The code used for the experiments is publicly available on GitHub at `https://github.com/mesnico/TERAN`.

3.1.14
Solving the same-different task with convolutional neural networks

N. Messina, G. Amato, F. Carrara, C. Gennaro, F. Falchi In Pattern Recognition Letters, Elsevier. [44]

Deep learning demonstrated major abilities in solving many kinds of different real-world problems in computer vision literature. However, they are still strained by simple reasoning tasks that humans consider easy to solve. In this work, we probe current state-of-the-art convolutional neural networks on a difficult set of tasks known as the same-different problems. All the problems require the same prerequisite to be solved correctly: understanding if two random shapes inside the same image are the same or not. With the experiments carried out in this work, we demonstrate that residual connections, and more generally the skip connections, seem to have only a marginal impact on the learning of the proposed problems. In particular, we experiment with DenseNets, and we examine the contribution of residual and recurrent connections in already tested architectures, ResNet-18, and CorNet-S respectively. Our experiments show that older feed-forward networks, AlexNet and VGG, are almost unable to learn the proposed problems, except in some

specific scenarios. We show that recently introduced architectures can converge even in the cases where the important parts of their architecture are removed. We finally carry out some zero-shot generalization tests, and we discover that in these scenarios residual and recurrent connections can have a stronger impact on the overall test accuracy. On four difficult problems from the SVRT dataset, we can reach state-of-the-art results with respect to the previous approaches, obtaining super-human performances on three of the four problems.

3.1.15
Word-class embeddings for multiclass text classification

A. Moreo, A. Esuli, F. Sebastiani. In Data Mining and Knowledge Discovery (Springer Nature). [53]

Pre-trained word embeddings encode general word semantics and lexical regularities of natural language, and have proven useful across many NLP tasks, including word sense disambiguation, machine translation, and sentiment analysis, to name a few. In supervised tasks such as multiclass text classification (the focus of this article) it seems appealing to enhance word representations with ad-hoc embeddings that encode task-specific information. We propose (supervised) word-class embeddings (WCEs), and show that, when concatenated to (unsupervised) pre-trained word embeddings, they substantially facilitate the training of deep-learning models in multiclass classification by topic. We show empirical evidence that WCEs yield a consistent improvement in multiclass classification accuracy, using six popular neural architectures and six widely used and publicly available datasets for multiclass text classification. One further advantage of this method is that it is conceptually simple and straightforward to implement. Our code that implements WCEs is publicly available at https://github.com/AlexMoreo/ word-class-embeddings.

3.1.16
Lost in transduction: Transductive transfer learning in text classification

A. Moreo, A. Esuli, F. Sebastiani. In ACM Transactions on Knowledge Discovery from Data (ACM Press). [50]

Obtaining high-quality labelled data for training a classifier in a new application domain is often costly. Transfer Learning (a.k.a. "Inductive Transfer") tries to alleviate these costs by transferring, to the "target" domain of interest, knowledge available from a different "source" domain. In transfer learning the lack of labelled information from the target domain is compensated by the availability at training time of a set of unlabelled examples from the target distribution. Transductive Transfer Learning denotes the transfer learning setting in which the only set of target documents that we are interested in classifying is known and available at training time. Although this definition is indeed in line with Vapnik's original definition of "transduction", current terminology in the field is confused. In this article we discuss how the term "transduction" has been misused in the transfer learning literature, and propose a clarification consistent with the original characterization of this term given by Vapnik. We go on to observe that the above terminology misuse has brought about misleading experimental comparisons, with inductive transfer learning methods that have been incorrectly compared with transductive transfer learning methods. We then give empirical evidence that the difference in performance between the inductive version and the transductive version of a transfer learning method can indeed be statistically significant (i.e., that knowing at training time the only data one needs to classify indeed gives an advantage). Our clarification allows a reassessment of the field, and of the relative merits of the major, state-of-the-art algorithms for transfer learning in text classification.

3.1.17
Tweet sentiment quantification: An experimental re-evaluation

A. Moreo, F. Sebastiani. In PLOS ONE, Public Library of Science. [58]

Sentiment quantification is the task of estimating the relative frequency (or "prevalence") of sentiment-related classes (such as Positive, Neutral, Negative) in a sample of unlabelled texts; this is especially important when these texts are tweets, since most sentiment classification endeavours carried out on Twitter data actually have quantification (and not the classification of individual tweets) as their ultimate goal. It is well-known that solving quantification via "classify and count" (i.e., by classifying all unlabelled items via a standard classifier and counting the items that have been assigned to a given class) is suboptimal in terms of accuracy, and that more accurate quantification methods exist. In 2016, Gao and Sebastiani carried out a systematic comparison of quantification methods on the task of tweet sentiment quantification. In hindsight, we observe that the experimental protocol followed in that work was flawed, and that the reported results are thus unreliable. We now re-evaluate those quantification methods (plus a few more modern ones) on the very same datasets, this time following a now consolidated and much more robust experimental protocol, that (even without counting the newly added methods) involves 5775 as many experiments as run in the original study. Our experimentation yields results dramatically different from those obtained by Gao and Sebastiani, and thus provide a different, much more solid understanding of the relative strengths and weaknesses of different sentiment quantification methods.

3.1.18
Representation and Presentation of Culinary Tradition as Cultural Heritage

N. Partarakis, D. Kaplanidi, P. Doulgeraki, E. Karuzaki, A. Petraki, D. Metilli, V. Bartalesi, I. Adami, C. Meghini, X. Zabulis. In Heritage (Basel), MDPI. [59]

This paper presents a knowledge representation framework and provides tools to allow the representation and presentation of the tangible and intangible dimensions of culinary tradition as cultural heritage including the socio-historic context of its evolution. The representation framework adheres to and extends the knowledge representation standards for the Cultural Heritage (CH) domain while providing a widely accessible web-based authoring environment to facilitate the representation activities. In strong collaboration with social sciences and humanities, this work allows the exploitation of ethnographic research outcomes by providing a systematic approach for the representation of culinary tradition in the form of recipes, both in an abstract form for their preservation and in a semantic rep-

resentation of their execution captured on-site during ethnographic research.

3.1.19
Re-ranking via local embeddings: A use case with permutation-based indexing and the nSimplex projection

L. Vadicamo, C. Gennaro, F. Falchi, E. Chávez, R. Connor, G. Amato In Information Systems, Elsevier. [69]

Approximate Nearest Neighbor (ANN) search is a prevalent paradigm for searching intrinsically high dimensional objects in large-scale data sets. Recently, the permutation-based approach for ANN has attracted a lot of interest due to its versatility in being used in the more general class of metric spaces. In this approach, the entire database is ranked by a permutation distance to the query. Typically, permutations allow the efficient selection of a candidate set of results, but typically to achieve high recall or precision this set has to be reviewed using the original metric and data. This can lead to a sizeable percentage of the database being recalled, along with many expensive distance calculations. To reduce the number of metric computations and the number of database elements accessed, we propose here a re-ranking based on a local embedding using the nSimplex projection. The nSimplex projection produces Euclidean vectors from objects in metric spaces which possess the n-point property. The mapping is obtained from the distances to a set of reference objects, and the original metric can be lower bounded and upper bounded by the Euclidean distance of objects sharing the same set of references. Our approach is particularly advantageous for extensive databases or expensive metric function. We reuse the distances computed in the permutations in the first stage, and hence the memory footprint of the index is not increased. An extensive experimental evaluation of our approach is presented, demonstrating excellent results even on a set of hundreds of millions of objects.

3.1.20
Query filtering using two-dimensional local embeddings

L. Vadicamo, R. Connor, E. Chávez In Information Systems, Elsevier [67]

The idea is that each element of the data is associated with a pair of reference objects that is well-suited to filter that particular object, in cases where it is not relevant to a query. This maximises the probability of excluding that object from a search. At query time, a query is compared with a pool of reference objects which allow its mapping to all the planes used by data objects. Then, for each query/object pair, a lower bound of the actual distance is obtained. The technique can be applied to any metric space that possesses the four-point property, therefore including Euclidean, Cosine, Triangular, Jensen-Shannon, and Quadratic Form distances.

Our experiments show that for all the datasets tested, of varying dimensionality, our approach can filter more objects than a standard metric indexing approach. For low dimensional data this does not make a good search mechanism in its own right, as it does not scale with the size of the data: that is, its cost is linear with respect to the data size. However, we also show that it can be added as a post-filter to other mechanisms, increasing efficiency with little extra cost in space or time. For high-dimensional data, we show related approximate techniques which, we believe, give the best known compromise for speeding up the essential sequential scan. The potential uses of our filtering technique include pure GPU searching, taking advantage of the tiny memory footprint of the mapping.*In high dimensional datasets, exact indexes are ineffective for proximity queries, and a sequential scan over the entire dataset is unavoidable. Accepting this, here we present a new approach employing two-dimensional embeddings. Each database element is mapped to the XY plane using the four-point property. The caveat is that the mapping is local: in other words, each object is mapped using a different mapping.*

The idea is that each element of the data is associated with a pair of reference objects that is well-suited to filter that particular object, in cases where it is not relevant to a query. This maximises the probability of excluding that object from a search. At query time, a query is compared with a pool of reference objects which allow its mapping to all the planes used by data objects. Then, for each query/object pair, a lower bound of the actual distance is obtained. The technique can be applied to any metric space that possesses the four-point property, therefore including Euclidean, Cosine, Triangular, Jensen-Shannon, and Quadratic Form distances.

Our experiments show that for all the datasets tested, of varying dimensionality, our approach can filter more objects than a standard metric indexing approach. For low dimensional data this does not make a good search mechanism in its own right, as it does not scale with the size of the data: that is, its cost is linear with respect to the data size. However, we also show that it can be added as a post-filter to other mechanisms, increasing efficiency with little extra cost in space or time. For high-dimensional data, we show related approximate techniques which, we believe, give the best known compromise for speeding up the essential sequential scan. The potential uses of our filtering technique include pure GPU searching, taking advantage of the tiny memory footprint of the mapping.

3.2 Proceedings

In this section, we report the paper we published in alphabetic order of the first author. Our works were presented, and published in the proceedings of the following conferences:

- **CBMI 2021** – International Conference on Content-Based Multimedia Indexing. [46]
- **CIKM 2021** – International Conference on Knowledge Management. ACM: [22]
- **ECIR 2022** – European Conference on Information Retrieval [2, 26] (accepted)
- **ECML-PKDD 2020** (appeared in 2021) – Joint European Conference on Machine Learning and Knowledge Discovery in Databases [62]
- **ICPR 2020** (postponed to 2021) – 25th International Conference on Pattern Recognition. IAPR: [8, 9, 47]
- **IIR 2021** – 11th Italian Information Retrieval Workshop. [63, 18, 55]
- **MMM 2021** – International Conference on Multimedia Modeling . [2]
- **NER 2021** – 10th International IEEE/EMBS Conference on Neural Engineering. [32]
- **SEBD 2021** – Italian Symposium on Advanced Database System. [34, 14]

- **SemEval 2021** – 15th International Workshop on Semantic Evaluation [48]
- **SIGGRAPH 2021 Talks** – 2021 Special Interest Group on Computer Graphics and Interactive Techniques Conference [3]
- **SISAP 2021** – International Conference on Similarity Search and Applications [68]
- **SWODCH 2021** – International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage [66]
- **VISIGRAPP 2021** – 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications [13]
- **VISIGRAPP 2022** – 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications [10, 42] (accepted9)

3.2.1
VISIONE at Video Browser Showdown 2021

G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, C. Vairo. In International Conference on Multimedia Modeling. [2]

This paper presents the second release of VISIONE, a tool for effective video search on large-scale collections. It allows users to search for videos using textual descriptions, keywords, occurrence of objects and their spatial relationships, occurrence of colors and their spatial relationships, and image similarity. One of the main features of our system is that it employs specially designed textual encodings for indexing and searching video content using the mature and scalable Apache Lucene full-text search engine.

3.2.2
NoR-VDPNet++: Efficient Training and Architecture for Deep No-Reference Image Quality Metrics

F. Banterle, A. Artusi, A. Moreo, F. Carrara In SIGGRAPH '21: ACM SIGGRAPH 2021 Talks. [3]

Efficiency and efficacy are two desirable properties of the utmost importance for any evaluation metric having to do with Standard Dynamic Range (SDR) imaging or High Dynamic Range (HDR) imaging. However, these properties are hard to achieve simultaneously. On the one side, metrics like HDR-VDP2.2 are known to mimic the human visual system (HVS) very accurately, but its high computational cost prevents its widespread use in large evaluation campaigns. On the other side, computationally cheaper alternatives like PSNR or MSE fail to capture many of the crucial aspects of the HVS. In this work, we try to get the best of the two worlds: we present NoR-VDPNet++, an improved variant of a previous deep learning-based metric for distilling HDR-VDP2.2 into a convolutional neural network (CNN). In this work, we try to get the best of the two worlds: we present NoR-VDPNet++, an improved version of a deep learning-based metric for distilling HDR-VDP2.2 into a convolutional neural network (CNN).

3.2.3
The IMAGO project: towards a knowledge base of medieval and renaissance geographical works

V. Bartalesi, N. Pratelli In SWODCH 2021 – International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage. [66]

The image of the world created by the Medieval and Renaissance culture was crucial to the development of Western thought in European history. To the best of our knowledge Medieval and Renaissance geographical works have not been studied using digital methods. The three years (2020-2023) Italian National research project IMAGO – Index Medii Aevi Geographiae Operum – aims at providing a systematic overview of this literature using Semantic Web technologies. As the first step to develop tools to support scholars in creating, evolving and consulting a knowledge base (KB) of the geographical works, we created an OWL 2 DL ontology. Following the re-use logic and to maximize the interoperability, we developed the ontology as an extension of two reference ontologies, that is the CIDOC CRM vocabulary and its extension FRBRoo, including its in-progress reformulation, LRMoo. In this paper, we present the project, the ontology and the tool to populate it that we developed. Furthermore, we present a preliminary study to map the works collected in the IMAGO KB and the manuscripts stored in the KB of the Mapping Manuscript Migrations project.

3.2.4
Defending Neural ODE Image Classifiers from Adversarial Attacks with Tolerance Randomization

F. Carrara, R. Caldelli, F. Falchi, G. Amato. In International Conference on Pattern Recognition. [9]

Deep learned models are now largely adopted in different fields, and they generally provide superior performances with respect to classical signal-based approaches. Notwithstanding this, their actual reliability when working in an unprotected environment is far enough to be proven. In this work, we consider a novel deep neural network architecture, named Neural Ordinary Differential Equations (N-ODE), that is getting particular attention due to an attractive property—a test-time tunable trade-off between accuracy and efficiency. This paper analyzes the robustness of N-ODE image classifiers when faced against a strong adversarial attack and how its effectiveness changes when varying such a tunable trade-off. We show that adversarial robustness is increased when the networks operate in different tolerance regimes during test time and training time. On this basis, we propose a novel adversarial detection strategy for N-ODE nets based on the randomization of the adaptive ODE solver tolerance. Our evaluation performed on standard image classification benchmarks shows that our detection technique provides high rejection of adversarial examples while maintaining most of the original samples under white-box attacks and zero-knowledge adversaries.

3.2.5
Combining GANs and autoencoders for efficient anomaly detection

F. Carrara, G. Amato, L. Brombin, F. Falchi, C. Gennaro. In 25th International Conference on Pattern Recognition (ICPR). [8]

In this work, we propose CBiGAN – a novel method for anomaly detection in images, where a consistency constraint is introduced as a regularization term in both the encoder and decoder of a BiGAN. Our model exhibits fairly good modeling power and reconstruction consistency capability. We evaluate the proposed

method on MVTec AD – a real-world benchmark for unsupervised anomaly detection on high-resolution images – and compare against standard baselines and state-of-the-art approaches. Experiments show that the proposed method improves the performance of BiGAN formulations by a large margin and performs comparably to expensive state-of-the-art iterative methods while reducing the computational cost. We also observe that our model is particularly effective in texture-type anomaly detection, as it sets a new state of the art in this category. Our code is available at `https://github.com/fabiocarrara/cbigan-ad/`.

3.2.6
Domain Adaptation for Traffic Density Estimation

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato. In Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021). [13]

Convolutional Neural Networks have produced state-of-the-art results for a multitude of computer vision tasks under supervised learning. However, the crux of these methods is the need for a massive amount of labeled data to guarantee that they generalize well to diverse testing scenarios. In many real-world applications, there is indeed a large domain shift between the distributions of the train (source) and test (target) domains, leading to a significant drop in performance at inference time. Unsupervised Domain Adaptation (UDA) is a class of techniques that aims to mitigate this drawback without the need for labeled data in the target domain. This makes it particularly useful for the tasks in which acquiring new labeled data is very expensive, such as for semantic and instance segmentation. In this work, we propose an end-to-end CNN-based UDA algorithm for traffic density estimation and counting, based on adversarial learning in the output space. The density estimation is one of those tasks requiring per-pixel annotated labels and, therefore, needs a lot of human effort. We conduct experiments considering different types of domain shifts, and we make publicly available two new datasets for the vehicle counting task that were also used for our tests. One of them, the Grand Traffic Auto dataset, is a synthetic collection of images, obtained using the graphical engine of the Grand Theft Auto video game, automatically annotated with precise per-pixel labels. Experiments show a significant improvement using our UDA algorithm compared to the model's performance without domain adaptation.

3.2.7
Traffic Density Estimation via Unsupervised Domain Adaptation

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato. In Italian Symposium on Advanced Database System (SEBD 2021), CEUR Workshop Proceedings. [14]

Monitoring traffic flows in cities is crucial to improve urban mobility, and images are the best sensing modality to perceive and assess the flow of vehicles in large areas. However, current machine learning-based technologies using images hinge on large quantities of annotated data, preventing their scalability to city-scale as new cameras are added to the system. We propose a new methodology to design image-based vehicle density estimators with few labeled data via an unsupervised domain adaptation technique.

3.2.8
Counting or Localizing? Evaluating cell counting and detection in microscopy images.

L. Ciampi, F. Carrara, G. Amato, C. Gennaro. In 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022). [10]

Image-based automatic cell counting is an essential yet challenging task, crucial for the diagnosing of many diseases. Current solutions rely on Convolutional Neural Networks and provide astonishing results. However, their performance is often measured only considering counting errors, which can lead to masked mistaken estimations; a low counting error can be obtained with a high but equal number of false positives and false negatives. Consequently, it is hard to determine which solution truly performs best. In this work, we investigate three general counting approaches that have been successfully adopted in the literature for counting several different categories of objects. Through an experimental evaluation over three public collections of microscopy images containing marked cells, we assess not only their counting performance compared to several state-of-the-art methods but also their ability to correctly localize the counted cells. We show that commonly adopted counting metrics do not always agree with the localization performance of the tested models, and thus we suggest integrating the proposed evaluation protocol when developing novel cell counting solutions.

3.2.9
Reading Songs: A Computational Analysis of Popular Songs lyrics

S. Corbara, A. Molinari In Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021). [18]

There is no doubt that certain songs are so easily liked by the public because of their melody. However, certain songs strike us for their lyrics, either because they convey an important (for us) meaning, or for their captivating sound when sung. Since the year 1958, the Billboard magazine held the special section Hot 100, with a rank of the 100 most popular songs of the week. Exploiting this invaluable source regarding the musical taste of the past decades until our days, we perform an analysis over various aspects of popular songs lyrics, especially focusing on the question: what is the importance of lyrics, when classifying musical artists and genres? We find out that, as we expected, many artists are not immediately recognizable only by their lyrics; some of them, however, and especially if they belong to some specific genres (such as rap), stand out, opening to the possibility of further analysis over their styles and themes.

3.2.10
Learning to quantify: Methods and applications (LQ 2021)

J.J. del Coz, P. González, A. Moreo, F. Sebastiani. In Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021). [22]

Learning to Quantify *(LQ) is the task of training class prevalence estimators via supervised learning. The task of these estimators is to estimate, given an unlabelled set of data items D and a set of classes* $\mathscr{C} = \{c_1, \ldots, c_{|\mathscr{C}|}\}$*, the prevalence (i.e., relative frequency)*

of each class c_i in D. LQ is interesting in all applications of classification in which the final goal is not *determining which class (or classes) individual unlabelled data items belong to, but estimating the distribution of the unlabelled data items across the classes of interest. Example disciplines whose interest in labelling data items is at the aggregate level (rather than at the individual level) are the social sciences, political science, market research, ecological modelling, and epidemiology. While LQ may in principle be solved by classifying each data item in D and counting how many such items have been labelled with c_i, it has been shown that this "classify and count" (CC) method yields suboptimal quantification accuracy. As a result, quantification is now no longer considered a mere byproduct of classification and has evolved as a task of its own. The goal of this workshop is bringing together all researchers interested in methods, algorithms, and evaluation measures and methodologies for LQ, as well as practitioners interested in their practical application to managing large quantities of data.*

3.2.11
LeQua@CLEF2022: Learning to Quantify

A. Esuli, A. Moreo, F. Sebastiani In Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022). (Accepted) [26]

LeQua 2022 is a new lab for the evaluation of methods for "learning to quantify" in textual datasets, i.e., for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents. While these predictions could be easily achieved by first classifying all documents via a text classifier and then counting the numbers of documents assigned to the classes, a growing body of literature has shown this approach to be suboptimal, and has proposed better methods. The goal of this lab is to provide a setting for the comparative evaluation of methods for learning to quantify, both in the binary setting and in the single-label multiclass setting. For each such setting we provide data either in ready-made vector form or in raw document form.

3.2.12
Assessing pattern recognition performance of neuronal cultures through accurate simulation

G. Lagani, R. Mazziotti, F. Falchi, C. Gennaro, G.M. Cicchini, T. Pizzorusso, F. Cremisi, G. Amato. In 10th International IEEE/EMBS Conference on Neural Engineering (NER). [32]

Previous work has shown that it is possible to train neuronal cultures on Multi-Electrode Arrays (MEAs), to recognize very simple patterns. However, this work was mainly focused to demonstrate that it is possible to induce plasticity in cultures, rather than performing a rigorous assessment of their pattern recognition performance. In this paper, we address this gap by developing a methodology that allows us to assess the performance of neuronal cultures on a learning task. Specifically, we propose a digital model of the real cultured neuronal networks; we identify biologically plausible simulation parameters that allow us to reliably reproduce the behavior of real cultures; we use the simulated culture to perform handwritten digit recognition and rigorously evaluate its performance; we also show that it is possible to find improved simulation parameters for the specific task, which can guide the creation of real cultures.

3.2.13
A multi-resolution training for expression recognition in the wild

F.V. Massoli, D. Cafarelli, G. Amato, F. Falchi In SEBD 2021 – Italian Symposium on Advanced Database System. [34]

Facial expressions play a fundamental role in human communication, and their study, which represents a multidisciplinary subject, embraces a great variety of research fields, e.g., from psychology to computer science, among others. Concerning Deep Learning, the recognition of facial expressions is a task named Facial Expression Recognition (FER). With such an objective, the goal of a learning model is to classify human emotions starting from a facial image of a given subject. Typically, face images are acquired by cameras that have, by nature, different characteristics, such as the output resolution. Moreover, other circumstances might involve cameras placed far from the observed scene, thus obtaining faces with very low resolutions. Therefore, since the FER task might involve analyzing face images that can be acquired with heterogeneous sources, it is plausible to expect that resolution plays a vital role. In such a context, we propose a multi-resolution training approach to solve the FER task. We ground our intuition on the observation that, often, face images are acquired at different resolutions. Thus, directly considering such property while training a model can help achieve higher performance on recognizing facial expressions. To our aim, we use a ResNet-like architecture, equipped with Squeeze-and-Excitation blocks, trained on the Affect-in-the-Wild 2 dataset. Not being available a test set, we conduct tests and model selection by employing the validation set only on which we achieve more than 90% accuracy on classifying the seven expressions that the dataset comprises.

3.2.14
Reinforced Damage Minimization in Critical Events for Self-Driving Vehicless

F. Merola, F. Falchi, C. Gennaro, M. Di Benedetto In 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). [42]

Self-driving systems have recently received massive attention in both academic and industrial contexts, leading to major improvements in standard navigation scenarios typically identified as well-maintained urban routes. Critical events like road accidents or unexpected obstacles, however, require the execution of specific emergency actions that deviate from the ordinary driving behavior and are therefore harder to incorporate in the system. In this context, we propose a system that is specifically built to take control of the vehicle and perform an emergency maneuver in case of a dangerous scenario. The presented architecture is based on a deep reinforcement learning algorithm, trained in a simulated environment and using raw sensory data as input. We evaluate the system's performance on several typical pre-accident scenario and show promising results, with the vehicle being able to consistently perform an avoidance maneuver to nullify or minimize the incoming damage.

3.2.15

Towards Efficient Cross-Modal Visual Textual Retrieval using Transformer-Encoder Deep Features

N. Messina, G. Amato, F. Falchi, C. Gennaro, S. Marchand-Maillet In 2021 International Conference on Content-Based Multimedia Indexing (CBMI). [46]

Cross-modal retrieval is an important functionality in modern search engines, as it increases the user experience by allowing queries and retrieved objects to pertain to different modalities. In this paper, we focus on the image-sentence retrieval task, where the objective is to efficiently find relevant images for a given sentence (image-retrieval) or the relevant sentences for a given image (sentence-retrieval). Computer vision literature reports the best results on the image-sentence matching task using deep neural networks equipped with attention and self-attention mechanisms. They evaluate the matching performance on the retrieval task by performing sequential scans of the whole dataset. This method does not scale well with an increasing amount of images or captions. In this work, we explore different preprocessing techniques to produce sparsified deep multi-modal features extracting them from state-of-the-art deep-learning architectures for image-text matching. Our main objective is to lay down the paths for efficient indexing of complex multi-modal descriptions. We use the recently introduced TERN architecture as an image-sentence features extractor. It is designed for producing fixed-size 1024-d vectors describing whole images and sentences, as well as variable-length sets of 1024-d vectors describing the various building components of the two modalities (image regions and sentence words respectively). All these vectors are enforced by the TERN design to lie into the same common space. Our experiments show interesting preliminary results on the explored methods and suggest further experimentation in this important research direction.

3.2.16

AIMH at SemEval-2021 Task 6: multimodal classification using an ensemble of transformer models

N. Messina, F. Falchi, C. Gennaro, G. Amato. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). [48]

This paper describes the system used by the AIMH Team to approach the SemEval Task 6. We propose an approach that relies on an architecture based on the transformer model to process multimodal content (text and images) in memes. Our architecture, called DVTT (Double Visual Textual Transformer), approaches Subtasks 1 and 3 of Task 6 as multi-label classification problems, where the text and/or images of the meme are processed, and the probabilities of the presence of each possible persuasion technique are returned as a result. DVTT uses two complete networks of transformers that work on text and images that are mutually conditioned. One of the two modalities acts as the main one and the second one intervenes to enrich the first one, thus obtaining two distinct ways of operation. The two transformers outputs are merged by averaging the inferred probabilities for each possible label, and the overall network is trained end-to-end with a binary cross-entropy loss.

3.2.17

Transformer reasoning network for image-text matching and retrieval

N. Messina, F. Falchi, A. Esuli, G. Amato. In 25th International Conference on Pattern Recognition (ICPR). [47]

Image-text matching is an interesting and fascinating task in modern AI research. Despite the evolution of deep-learning-based image and text processing systems, multimodal matching remains a challenging problem. In this work, we consider the problem of accurate image-text matching for the task of multi-modal large-scale information retrieval. State-of-the-art results in image-text matching are achieved by inter-playing image and text features from the two different processing pipelines, usually using mutual attention mechanisms. However, this invalidates any chance to extract separate visual and textual features needed for later indexing steps in large-scale retrieval systems. In this regard, we introduce the Transformer Encoder Reasoning Network (TERN), an architecture built upon one of the modern relationship-aware self-attentive architectures, the Transformer Encoder (TE). This architecture is able to separately reason on the two different modalities and to enforce a final common abstract concept space by sharing the weights of the deeper transformer layers. Thanks to this design, the implemented network is able to produce compact and very rich visual and textual features available for the successive indexing step. Experiments are conducted on the MS-COCO dataset, and we evaluate the results using a discounted cumulative gain metric with relevance computed exploiting caption similarities, in order to assess possibly non-exact but relevant search results. We demonstrate that on this metric we are able to achieve state-of-the-art results in the image retrieval task. Our code is freely available at `https://github.com/mesnico/TERN`.

3.2.18

Generalized funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification (Extended Abstract)

A. Moreo, A. Pedrotti, F. Sebastiani. In Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021). [55]

Funnelling (Fun) is a method for cross-lingual text classification (CLTC) based on a two-tier learning ensemble for heterogeneous transfer learning (HTL). In this ensemble method, 1st-tier classifiers, each working on a different and language-dependent feature space, return a vector of calibrated posterior probabilities (with one dimension for each class) for each document, and the final classification decision is taken by a metaclassifier that uses this vector as its input. In this paper we describe Generalized Funnelling (gFun), a generalization of Fun consisting of a HTL architecture in which 1st-tier components can be arbitrary view-generating functions, i.e., language-dependent functions that each produce a language-independent representation ("view") of the document. We describe an instance of gFun in which the metaclassifier receives as input a vector of calibrated posterior probabilities (as in Fun) aggregated to other embedded representations that embody other types of correlations. We describe preliminary results that we have obtained on a large standard dataset for multilingual multilabel text classification.

3.2.19
QuaPy: A Python-based framework for quantification

A. Moreo, A. Esuli, F. Sebastiani. In Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021). [51]

QuaPy is an open-source framework for performing quantification (a.k.a. supervised prevalence estimation), written in Python. Quantification is the task of training quantifiers via supervised learning, where a quantifier is a predictor that estimates the relative frequencies (a.k.a. prevalence values) of the classes of interest in a sample of unlabelled data. While quantification can be trivially performed by applying a standard classifier to each unlabelled data item and counting how many data items have been assigned to each class, it has been shown that this "classify and count" method is outperformed by methods specifically designed for quantification. QuaPy provides implementations of a number of baseline methods and advanced quantification methods, of routines for quantification-oriented model selection, of several broadly accepted evaluation measures, and of robust evaluation protocols routinely used in the field. QuaPy also makes available datasets commonly used for testing quantifiers, and offers visualization tools for facilitating the analysis and interpretation of the results. The software is open-source and publicly available under a BSD-3 licence via GitHub[8], and can be installed via `pip`[9].

3.2.20
Garbled-word embeddings for jumbled text

G. Sperduti, A. Moreo, F. Sebastiani In Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021). [63]

"Aoccdrnig to a reasrech at Cmabrigde Uinervtisy, it deosn't mtter in waht oredr the ltteers in a wrod are, the olny itmopnrat tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe". We investigate the extent to which this phenomenon applies to computers as well. Our hypothesis is that computers are able to learn distributed word representations that are resilient to character reshuffling, without incurring a significant loss in performance in tasks that use these representations. If our hypothesis is confirmed, this may form the basis for a new and more efficient way of encoding character-based representations of text in deep learning, and one that may prove especially robust to misspellings, or to corruption of text due to OCR. This paper discusses some fundamental psycho-linguistic aspects that lie at the basis of the phenomenon we investigate, and reports on a preliminary proof of concept of the above idea.

3.2.21
On Generalizing Permutation-Based Representations for Approximate Search

L. Vadicamo, C. Gennaro, G. Amato. In 14th International Conference on Similarity Search and Applications. [68]

In the domain of approximate metric search, the Permutation-based Indexing (PBI) approaches have been proved to be particularly suitable for dealing with large data collections. These methods employ a permutation-based representation of the data, which can be efficiently indexed using data structures such as inverted files. In the literature, the definition of the permutation of a metric object was derived by reordering the distances of the object to a set of pivots. In this paper, we aim at generalizing this definition in order to enlarge the class of permutations that can be used by PBI approaches. As a practical outcome, we defined a new type of permutation that is calculated using distances from pairs of pivots. The proposed technique permits us to produce longer permutations than traditional ones for the same number of object-pivot distance calculations. The advantage is that the use of inverted files built on permutation prefixes leads to greater efficiency in the search phase when longer permutations are used.

3.2.22
Automatic Pass Annotation from Soccer Video Streams Based on Object Detection and LSTM

D. Sorano, F. Carrara, P. Cintia, F. Falchi, L. Pappalardo. In 2020 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. [62]

Soccer analytics is attracting increasing interest in academia and industry, thanks to the availability of data that describe all the spatio-temporal events that occur in each match. These events (e.g., passes, shots, fouls) are collected by human operators manually, constituting a considerable cost for data providers in terms of time and economic resources. In this paper, we describe PassNet, a method to recognize the most frequent events in soccer, i.e., passes, from video streams. Our model combines a set of artificial neural networks that perform feature extraction from video streams, object detection to identify the positions of the ball and the players, and classification of frame sequences as passes or not passes. We test PassNet on different scenarios, depending on the similarity of conditions to the match used for training. Our results show good classification results and significant improvement in the accuracy of pass detection with respect to baseline classifiers, even when the match's video conditions of the test and training sets are considerably different. PassNet is the first step towards an automated event annotation system that may break the time and the costs for event annotation, enabling data collections for minor and non-professional divisions, youth leagues and, in general, competitions whose matches are not currently annotated by data providers.

3.3 Magazines

In this section, we report the paper we published in magazines in alphabetic order of the first author.

3.3.1
Report on the 43rd European Conference on Information Retrieval (ECIR 2021)

R. Perego, F. Sebastiani.
In SIGIR Forum (ACM Press). [60].

The 43rd European Conference on Information Retrieval (ECIR 2021 – `https://www.ecir2021.eu/`*), organized under the auspices of the Information Retrieval Specialist Group of the British Computer Society (BCS IRSG), took place between March 28 and April 1, 2021. As sadly customary in these dark times, the conference*

[8] `https://github.com/HLT-ISTI/QuaPy`
[9] `https://pypi.org/project/QuaPy/`

was held entirely online, due to the COVID-19 pandemic. According to the original plans, it should have instead taken place in Lucca, a small Italian town in Tuscany, Italy, which enjoys a beautiful, extremely well-preserved historic centre.

3.4 Editorials

In this section, we report proceedings and books for which we have been editors.

3.4.1
Proceedings of the 43rd European Conference on Information Retrieval Research (ECIR 2021), Volumes I and II

D. Hiemstra, M.F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (eds.), Lecture Notes in Computer Science, Springer Nature, 2021. [29].

3.5 Preprints

In this section, we report the papers published in preprint form on publicly accessible archives, in alphabetic order by first author.

3.5.1
Expression Recognition Analysis in the Wild

D. Cafarelli, F.V. Massoli, F. Falchi, C. Gennaro, G. Amato
arXiv:2101.09231. [7]

Facial Expression Recognition(FER) is one of the most important topic in Human-Computer interactions(HCI). In this work we report details and experimental results about a facial expression recognition method based on state-of-the-art methods. We fine-tuned a SeNet deep learning architecture pre-trained on the well-known VGGFace2 dataset, on the AffWild2 facial expression recognition dataset. The main goal of this work is to define a baseline for a novel method we are going to propose in the near future. This paper is also required by the Affective Behavior Analysis in-the-wild (ABAW) competition in order to evaluate on the test set this approach. The results reported here are on the validation set and are related on the Expression Challenge part (seven basic emotion recognition) of the competition. We will update them as soon as the actual results on the test set will be published on the leaderboard.

3.5.2
Multi-Camera Vehicle Counting Using Edge-AI

L. Ciampi, C. Gennaro, F. Carrara, F. Falchi, C. Vairo, G. Amato
arXiv:2106.02842. [12]

This paper presents a novel solution to automatically count vehicles in a parking lot using images captured by smart cameras. Unlike most of the literature on this task, which focuses on the analysis of single images, this paper proposes the use of multiple visual sources to monitor a wider parking area from different perspectives. The proposed multi-camera system is capable of automatically estimate the number of cars present in the entire parking lot directly on board the edge devices. It comprises an on-device deep learning-based detector that locates and counts the vehicles from the captured images and a decentralized geometric-based approach that can analyze the inter-camera shared areas and merge the data acquired by all the devices. We conduct the experimental evaluation on an extended version of the CNRPark-EXT dataset, a collection of images taken from the parking lot on the campus of the National Research Council (CNR) in Pisa, Italy. We show that our system is robust and takes advantage of the redundant information deriving from the different cameras, improving the overall performance without requiring any extra geometrical information of the monitored scene.

3.5.3
Combining EfficientNet and Vision Transformers for Video Deepfake Detection

D. Coccomini, N. Messina, C. Gennaro, F. Falchi
arXiv:2107.02612. [16]

Deepfakes are the result of digital manipulation to obtain credible videos in order to deceive the viewer. This is done through deep learning techniques based on autoencoders or GANs that become more accessible and accurate year after year, resulting in fake videos that are very difficult to distinguish from real ones. Traditionally, CNN networks have been used to perform deepfake detection, with the best results obtained using methods based on EfficientNet B7. In this study, we combine various types of Vision Transformers with a convolutional EfficientNet B0 used as a feature extractor, obtaining comparable results with some very recent methods that use Vision Transformers. Differently from the state-of-the-art approaches, we use neither distillation nor ensemble methods. The best model achieved an AUC of 0.951 and an F1 score of 88.0%, very close to the state-of-the-art on the DeepFake Detection Challenge (DFDC).

3.5.4
Generative Adversarial Networks for Astronomical Images Generation

D. Coccomini, N. Messina, C. Gennaro, F. Falchi
arXiv:2111.11578. [17]

vide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi Space exploration has always been a source of inspiration for humankind, and thanks to modern telescopes, it is now possible to observe celestial bodies far away from us. With a growing number of real and imaginary images of space available on the web and exploiting modern deep Learning architectures such as Generative Adversarial Networks, it is now possible to generate new representations of space. In this research, using a Lightweight GAN, a dataset of images obtained from the web, and the Galaxy Zoo Dataset, we have generated thousands of new images of celestial bodies, galaxies, and finally, by combining them, a wide view of the universe. The code for reproducing our results is publicly available at this https URL, and the generated images can be explored at this https URL[10].

3.5.5
Syllabic quantity patterns as rhythmic features for latin authorship attribution

S. Corbara, A. Moreo, F. Sebastiani. arXiv:2110.14203 [cs.CL], 2021. [19]

It is well known that, within the Latin production of written text, peculiar metric schemes were followed not only in poetic compositions, but also in many prose works. Such metric patterns were

[10] https://davide-coccomini.github.io/GAN-Universe/

based on so-called syllabic quantity, i.e., on the length of the involved syllables, and there is substantial evidence suggesting that certain authors had a preference for certain metric patterns over others. In this research we investigate the possibility to employ syllabic quantity as a base for deriving rhythmic features for the task of computational authorship attribution of Latin prose texts. We test the impact of these features on the authorship attribution task when combined with other topic-agnostic features. Our experiments, carried out on three different datasets, using two different machine learning methods, show that rhythmic features based on syllabic quantity are beneficial in discriminating among Latin prose authors.

3.5.6
LeQua@CLEF2022: Learning to quantify

A. Esuli, A. Moreo, F. Sebastiani
arXiv:2111.11249 [cs.LG], 2021. [25]

LeQua 2022 is a new lab for the evaluation of methods for "learning to quantify" in textual datasets, i.e., for training predictors of the relative frequencies of the classes of interest in sets of unlabelled textual documents. While these predictions could be easily achieved by first classifying all documents via a text classifier and then counting the numbers of documents assigned to the classes, a growing body of litera- ture has shown this approach to be suboptimal, and has proposed better methods. The goal of this lab is to provide a setting for the comparative evaluation of methods for learning to quantify, both in the binary setting and in the single-label multiclass setting. For each such setting we provide data either in ready-made vector form or in raw document form.

3.5.7
Measuring fairness under unawareness via quantification

A. Fabris, A. Esuli, A. Moreo, F. Sebastiani. arXiv:2109.08549 [cs.CY], 2021. [27]

Models trained by means of supervised learning are increasingly deployed in high-stakes domains, and, when their predictions inform decisions about people, they inevitably impact (positively or negatively) on their lives. As a consequence, those in charge of developing these models must carefully evaluate their impact on different groups of people and ensure that sensitive demographic attributes, such as race or sex, do not result in unfair treatment for members of specific groups. For doing this, awareness of demographic at- tributes on the part of those evaluating model impacts is fundamental. Un- fortunately, the collection of these attributes is often in conflict with industry practices and legislation on data minimization and privacy. For this reason, it may be hard to measure the group fairness of trained models, even from within the companies developing them. In this work, we tackle the problem of measuring group fairness under unawareness of sensitive attributes, by using techniques from quantification, a supervised learning task concerned with di- rectly providing group-level prevalence estimates (rather than individual-level class labels). We identify five important factors that complicate the estimation of fairness under unawareness and formalize them into five different experimen- tal protocols under which we assess the effectiveness of different estimators of group fairness. We also consider the problem of potential model misuse to infer sensitive attributes at an individual level, and demonstrate that quantification approaches are suitable for decoupling the (desirable) objective of measuring group fairness from the (undesirable) objective of inferring sensitive attributes of individuals.

3.5.8
MAFER: a Multi-resolution Approach to Facial Expression Recognition

F.V. Massoli, D. Cafarelli, C. Gennaro, G. Amato, F. Falchi
arXiv:2105.02481. [35]

Emotions play a central role in the social life of every human being, and their study, which represents a multidisciplinary subject, embraces a great variety of research fields. Especially concerning the latter, the analysis of facial expressions represents a very active research area due to its relevance to human-computer interaction applications. In such a context, Facial Expression Recognition (FER) is the task of recognizing expressions on human faces. Typically, face images are acquired by cameras that have, by nature, different characteristics, such as the output resolution. It has been already shown in the literature that Deep Learning models applied to face recognition experience a degradation in their performance when tested against multi-resolution scenarios. Since the FER task involves analyzing face images that can be acquired with heterogeneous sources, thus involving images with different quality, it is plausible to expect that resolution plays an important role in such a case too. Stemming from such a hypothesis, we prove the benefits of multi-resolution training for models tasked with recognizing facial expressions. Hence, we propose a two-step learning procedure, named MAFER, to train DCNNs to empower them to generate robust predictions across a wide range of resolutions. A relevant feature of MAFER is that it is task-agnostic, i.e., it can be used complementarily to other objective-related techniques. To assess the effectiveness of the proposed approach, we performed an extensive experimental campaign on publicly available datasets: , , and . For a multi-resolution context, we observe that with our approach, learning models improve upon the current SotA while reporting comparable results in fix-resolution contexts. Finally, we analyze the performance of our models and observe the higher discrimination power of deep features generated from them.

3.5.9
A Leap among Entanglement and Neural Networks: A Quantum Survey

F.V. Massoli, L. Vadicamo, G. Amato, F. Falchi
arXiv:2107.03313. [38]

In recent years, Quantum Computing witnessed massive improvements both in terms of resources availability and algorithms development. The ability to harness quantum phenomena to solve computational problems is a long-standing dream that has drawn the scientific community's interest since the late '80s. In such a context, we pose our contribution. First, we introduce basic concepts related to quantum computations, and then we explain the core functionalities of technologies that implement the Gate Model and Adiabatic Quantum Computing paradigms. Finally, we gather, compare and analyze the current state-of-the-art concerning Quantum Perceptrons and Quantum Neural Networks implementations.

3.5.10
Recurrent Vision Transformer for Solving Visual Reasoning Problems

N. Messina, G. Amato, F. Carrara, C. Gennaro, F. Falchi
arXiv:2111.14576. [43]

Although convolutional neural networks (CNNs) showed remarkable results in many vision tasks, they are still strained by simple yet challenging visual reasoning problems. Inspired by the recent success of the Transformer network in computer vision, in this paper, we introduce the Recurrent Vision Transformer (RViT) model. Thanks to the impact of recurrent connections and spatial attention in reasoning tasks, this network achieves competitive results on the same-different visual reasoning problems from the SVRT dataset. The weight-sharing both in spatial and depth dimensions regularizes the model, allowing it to learn using far fewer free parameters, using only 28k training samples. A comprehensive ablation study confirms the importance of a hybrid CNN + Transformer architecture and the role of the feedback connections, which iteratively refine the internal representation until a stable prediction is obtained. In the end, this study can lay the basis for a deeper understanding of the role of attention and recurrent connections for solving visual abstract reasoning tasks.

3.5.11
QuaPy: A Python-based framework for quantification

A. Moreo, A. Esuli, F. Sebastiani.
arXiv:2106.11057 [cs.LG], 2021. [52]

QuaPy is an open-source framework for performing quantification (a.k.a. supervised prevalence estimation), written in Python. Quantification is the task of training quantifiers via supervised learning, where a quantifier is a predictor that estimates the relative frequencies (a.k.a. prevalence values) of the classes of interest in a sample of unlabelled data. While quantification can be trivially performed by applying a standard classifier to each unlabelled data item and counting how many data items have been assigned to each class, it has been shown that this "classify and count" method is outperformed by methods specifically designed for quantification. QuaPy provides implementations of a number of baseline methods and advanced quantification methods, of routines for quantification-oriented model selection, of several broadly accepted evaluation measures, and of robust evaluation protocols routinely used in the field. QuaPy also makes available datasets commonly used for testing quantifiers, and offers visualization tools for facilitating the analysis and interpretation of the results. The software is open-source and publicly available under a BSD-3 licence via GitHub[11]*, and can be installed via* `pip`[12].

3.5.12
Generalized funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification

A. Moreo, A. Pedrotti, F. Sebastiani. arXiv:2110.14764 [cs.CL], 2021. [54]

Funnelling (Fun) is a recently proposed method for cross-lingual text classification (CLTC) based on a two-tier learning ensemble for heterogeneous transfer learning (HTL). In this ensemble method, 1st-tier classifiers, each working on a different and language-dependent feature space, return a vector of calibrated posterior probabilities (with one dimension for each class) for each document, and the final classification decision is taken by a metaclassifier that uses this vector as its input. The metaclassifier can thus exploit class-class correlations, and this (among other things) gives Fun an edge over CLTC systems in which these correlations cannot be brought to bear. In this paper we describe Generalized Funnelling (gFun), a generalization of Fun consisting of an HTL architecture in which 1st-tier components can be arbitrary view-generating functions, i.e., language-dependent functions that each produce a language-independent representation ("view") of the document. We describe an instance of gFun in which the metaclassifier receives as input a vector of calibrated posterior probabilities (as in Fun) aggregated to other embedded representations that embody other types of correlations, such as word-class correlations (as encoded by Word-Class Embeddings), word- word correlations (as encoded by Multilingual Unsupervised or Supervised Embeddings), and word-context correlations (as encoded by multilingual BERT). We show that this instance of gFun substantially improves over Fun and over state-of-the-art baselines, by reporting experimental results obtained on two large, standard datasets for multilingual multilabel text classification. Our code that implements gFun is publicly available.

4. Dissertations

4.1 MSc Dissertations

4.1.1
Developing and Experimenting Approaches for Facial Expression Recognition in Images from Surveillance Cameras

D. Cafarelli, MSc in Computer Engineering, University of Pisa, 2021 [6]. Advisors: F. Falchi, F.V. Massoli, G. Amato, C. Gennaro

A large number of surveillance cameras are available in many cities. However, their use for automatic visual analysis is very limited. In this thesis, we focus on particular in non-security-oriented analysis such as sentiment through facial expression recognition. Other information that could be useful, especially in combination with the sentiment is age and gender.

4.1.2
Design and Development of Transformer-based Methods for Video Deepfake Detection

D.A. Coccomini, MSc in Artificial Intelligence and Data Engineering, 2021 [15]. Advisors: F. Falchi, C. Gennaro, and N. Messina.

This thesis is about deep learning methods based on self-attention and their application in Computer Vision. In particular, the thesis focuses on TimeSformers, a variant of the basic Transformers architecture specifically designed for video understanding. The main con- tribution of this thesis is the application of this architecture and the development of other alternative architectures based on Vision

[11] https://github.com/HLT-ISTI/QuaPy
[12] https://pypi.org/project/QuaPy/

Trans- formers and EfficientNet for the detection of deepfakes (i.e., synthetic videos where a person in an image or video is swapped with another person's likeness) and to the more general task of anomaly detection. First, the problem of the identification of anomalous situations in videos of surveillance cameras is tackled, and a comparison between various architectures and versions of the transformers is given, study- ing their effectiveness in this particular context, which has never been faced before through the use of this innovative deep learning method. Second, experiments are conducted on the deepfake detection task developing several hybrid architectures of Vision Transformers and convolutional networks, achieving near state-of-the-art performance with networks trained from scratch and partially based on fine-tuning of pretrained networks.

4.1.3
Design and development of AI-based video surveillance applications for low-power embedded systems

A. Liuzzi, MSc in Computer Engineering, University of Pisa, 2021 [33]. Advisors: C. Gennaro, G. Amato, F. Falchi.

Protection and security are two fundamental aspects of human life. Current technologies, through the use of artificial intelligence systems, can be a valid tool to pursue these goals. One use of these technologies could involve the integration of facial recognition systems within those of video surveillance, commonly used in various fields. Facial recognition is an artificial intelligence technique that, using biometric characteristics, detects, tracks, identifies or verifies human faces within images or videos, captured using a digital camera. This thesis work is part of this research area. In particular, the goal of the study is to develop an application of facial recognition that can be run on an embedded system with low power consumption. Embedded systems, in fact, thanks to their small size, low power requirements and computing power, are suitable for such purposes. They are, therefore, easily integrated into video surveillance equipment such as cameras or battery-powered robots equipped with self-control. The development of the application, moreover, is focused on the optimization of the described process achieved through the parallel analysis performed on several faces. The whole system has been implemented through the use of the Python language, the Pytorch framework and the NVIDIA Jetson Nano.

4.1.4
Designing, implementing and experimenting a deep reinforcement learning approach in a simulated environment for autonomous driving in critical scenarios

F. Merola, MSc in Computer Engineering, University of Pisa, 2021 [41]. Advisors: F. Falchi, M. Di Benedetto, C. Gennaro.

Autonomous driving has the potential to drastically change mobility and transport by bringing considerable improvements to safety and comfort on the road. This prospective, together with the recent advancements in

elds such as sensors, computer vision and artificial intelligence, attracted a lot of attention from both the research community and industry. Massive steps forward have already been made and several works have demonstrated that it is possible to rely on sensors to obtain a comprehensive understanding of the surrounding environment and use it to automate the driving task in most standard scenarios. However, to date, there is no self-driving system that can be considered fully autonomous, and this is mostly due to the existence of some particular circumstances that have proven to be really difficult to handle because of factors such as adverse weather or sudden and unexpected events. In this context, the presented work proposes a potential solution to critical preaccident scenarios in a simulated environment, focusing on damage minimization in high-risk circumstances. The novelty of the approach, besides shifting the attention to specific, highly interesting scenarios, is in the method used for the learning process, reinforcement learning, that has seen few applications in driving contexts to date. Shortly, reinforcement learning is a paradigm based on the "learning from interaction" concept, which involves an online training where the model gets to interact with the environment under the guidance of a user-defined reward function, that needs to be maximized. The system was designed and developed around Double Deep Q-learning (Double DQN), a model-free deep reinforcement learning algorithm, using as input the data provided by a frontal RGB camera mounted on the vehicle. The reward function was designed taking into account multiple factors such as speed, collision damage and distance covered, keeping in mind that the vehicle should learn to minimize damage in critical situations while still being able to drive normally in non-critical ones. Finally, several tests were carried out on the simulation platform CARLA, where the system achieved good results by managing to discern high-risk situations from normal ones, acting accordingly most of the time. Despite this, some generalization and driving stability issues were evidenced and analyzed. The thesis ends by proposing some possible future steps that could be taken to make the technology more concrete and efficient, including a transfer learning approach to adapt the model for operation in the real world.

4.1.5
Design and implementation of a deep learning system for knowledge graph analysis

F. Minutella, MSc in Artificial Intelligence and Data Engineering, University of Pisa, 2021 [49]. Advisors: F. Falchi, P. Manghi, M. De Bonis, N. Messina.

Nowadays a lot of data is in the form of Knowledge Graphs, i.e. a set of nodes and relationships between them. Many companies exclude relationships or don't use them to their full potential in order to convert naturally graph-like data into tabular data so that it can be organized in the usual databases and analyzed using simple, familiar processes. This conversion process has the advantage of simplification but brings with it a loss of information that cannot always be ignored. After a review of techniques aimed at performing different tasks on graph data types, some of these were used in the analysis of the data provided by OpenAIRE. OpenAIRE is a platform to support Open Science in Europe and it provides a Research Graph, which is a graph composed of scienti

c resources linked to their authors, where they have been published, and the keywords in them. For the analysis of the Research Graph, it has been used a metapath approach in order to allow the analysis of a heterogeneous graph by transforming it into a series of homogeneous graphs. Such graphs are simpler to be analyzed and they allow to focus the analysis on a single type of element of the graph. A framework was developed to analyze the Research Graph and to highlight the anomalies in the dataset. The framework integrates the metapath approach and a neural network to perform

Node Classi

cation and Node Embedding, and the results were compared with the methods of Graph Neural Networks in the literature. The result of our work is a method that can leverage the node attributes and graph metapaths to perform Node Classi

cation or Node Embedding by identifying the most signi

cant information. The result of the work presented in this thesis is a framework that is scalable, easy to understand and fast. Moreover, it performs better than other unsupervised methods available in the literature.

4.1.6
Design and Implementation of a Framework for Reinforcement Learning for self-driving cars

R. Rabitti, MSc in Computer Engineering, University of Pisa, 2021 [61]. Advisors: G. Amato, M. Di Benedetto, F. Falchi, C. Gennaro.

In this work, we aim to apply Artificial Intelligence techniques, based on the Machine Learning approach, to develop automated driving tools for preventing traffic accidents or, at least, minimizing the resulting damage. In particular, in this project we created a framework designed for testing Reinforcement Learning and Deep Reinforcement Learning algorithms in simulated driving scenarios. We tested multiple configurations of Deep Q-Learning agents to solve the same autonomous driving task, consisting in avoiding efficiently an obstacle on our driving path, without stopping the car if not necessary. These testing driving scenarios were reproduced using CARLA simulation environment, an open-source simulation environment, running on top of Unreal 4 engine, designed to assist autonomous driving research by enacting and visualizing the driving simulations episodes. We present our simulations results, in ascending level of agent configuration complexity, resulting in progressively better performance in the autonomous driving task.

4.1.7
Development of a park monitoring system for smart camera networks

M. Taibi, MSc in Computer Engineering, University of Pisa, 2021 [64]. Advisors: C. Gennaro, A. Giuseppe, F. Falchi.

Nowadays, we are surrounded by video surveillance cameras in public and private spaces. These vision systems are usually smart, i.e. with computational capabilities that allow the development of various applications. A very common application of video surveillance cameras is the Park Monitoring using Deep Learning techniques to detect the vehicles in a parking area. In particular, in this thesis work, we adopt the vehicles counting approach based on deep learning to monitor a parking lot. This approach is able to globally count the cars in the parking area without requiring any information about parking lot locations. Although deep learning is particularly effective, a factor of complexity is represented by challenging situations, due for example to the presence of shadows, variation of light and weather conditions, inter-object occlusions. To overcome these problems, it is possible to use a pair of cameras that monitor the parking lot with different perspectives and different angles of views, or multiple adjacent cameras to cover a wide area. This introduces problems related to the automatic merge of the knowledge extracted from single cameras because their fields of views partially overlap. In this thesis work, we addressed the cars counting problem and developed a solution to count cars from a video stream, using frames captured by multiple cameras. The solution combines deep learning techniques for Object Detection with a geometry-based approach, to find a homography transformation between two adjacent cameras and identify auto- matically the cars in their overlapping area. In particular, each camera uses the Region Based Convolutional Neural Network Mask R-CNN to detect cars in its park portion. To test the solution, we used the existing CNRPark-EXT dataset, composed of images taken by nine smart cameras located in the campus of the National Research Council (CNR) in Pisa and covering challenging scenes. We propose two different approaches to implement the solution: a centralized solution, in which the system is composed of some cameras that send the captured frames to a central server, which sequentially performs cars detection on them and merging those results, or a decentralized solution, in which the system is composed of some smart cameras that communicate with each other and perform the detection and merge tasks on-board. We finally show the results obtained and how a real multi-camera car counting system could be implemented.

4.2 BSc Dissertations

4.2.1
La quantità sillabica nella computational authorship attribution per testi latini

G. Canapa, BSc in Digital Humanities, University of Pisa, 2021 Advisors: V. Casarosa, S. Corbara, F. Sebastiani

5. Datasets

5.0.1 GTA - Grand Traffic Auto Dataset

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato. [13]

The GTA dataset is a vast collection of about 15,000 synthetic images of urban traffic scenes gathered using the highly photo-realistic graphical engine of the GTA V – Grand Theft Auto V video game. About half of them concern urban city areas, while the remaining involve sub-urban areas and highways. To generate this dataset, we designed a framework that automatically and precisely annotates the vehicles present in the scene with per-pixel annotations. To the best of our knowledge, this is the first instance segmentation synthetic dataset of city traffic scenarios. The dataset is freely available at `http://aimh.isti.cnr.it/grand_traffic_auto/`.

5.0.2 NDISPark - Night and Day Instance Segmented Park Dataset

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato. [13]

The NDISPark dataset is a manually annotated dataset of cars in parking lots, consisting of about 250 images. Images are taken under different weather, viewing angles, and occlusions, describing most of the problematic situations that we can find in a real scenario. NDISPark is precisely annotated with instance segmentation labels. Furthermore, it is worth noting that images are taken during the day and the night, showing utterly different lighting conditions. The dataset is freely available at `http://aimh.isti.cnr.it/ndispark/`.

5.0.3 PNN - A Multi-Rater Benchmark for Perineuronal Nets Detection and Counting in Fluorescence Microscopy Images

L. Ciampi, F. Carrara, V. Totaro, R. Mazziotti, L. Lupori, C. Santiago, G. Amato, T. Pizzorusso, C. Gennaro. [11]

PNN is a dataset of fluorescence microscopy images of mice brain slices stained against perineuronal nets (PNNs). PNNs are dot-annotated by experts for evaluating cell detection and counting. The dataset is composed of two subsets: a large single-rater subset (PNN-SR) and a smaller multi-rater subset (PNN-MR). The annotation procedure in PNN-MR has been performed by seven different raters, and all the annotations of each rater are reported. This enables modeling the confidence of a given detection using the agreeement between raters on that sample as training signal. The dataset is freely available at `https://doi.org/10.5281/zenodo.5567032`.

5.0.4 TweepFake - Twitter deep Fake text Dataset

T. Fagni, A. Martella, F. Falchi, M. Gambini, M. Tesconi. [65]

Social media have always been the perfect vehicle to manipulate and alter public opinion through bots, i.e. agents that behave as human users by liking, re-posting and publishing multimedia content which can be real or machine-generated. In the latter case, the spreading of deep-fakes - potentially deceptive images, video, audio or text autonomously generated by a deep neural network - in social media, have been sowing mistrust, hate and deceits at the expense of the people. As far as it concerns deep-fake texts, the great improvement over their generation has been obtained by the language models (RNN, LSTM, GPT-2, GROVER, CLTR, OPTIMUS, GPT-3): several studies have shown that humans are capable of detecting those deep-fake texts as machine-generated with a detection rate around the chance value. Even though examples of deep-fake messages can already be found in social media (as our dataset shows), there is still no episode of misuse on them; however, the language models' generative capability deeply worries: it is, therefore, necessary to quickly raise shields against this threat as well. Some deep-fake text detection techniques have already been investigated, but there is still a lack of knowledge on how those state-of-the-art deepfake text detection techniques perform in a "real-social-media-setting", in which the text generation method is unknown and the text content is often short (especially on Twitter). A dataset of deep-fake social media messages is required to start the research. Unfortunately, at the best of our knowledge, no one has ever created a properly labelled social media dataset containing only human and deep-fake messages (thus excluding cheap-fake texts that employ simple generative techniques as gap-filling and search-and-replace methods) that can already be found on social media timelines. Focusing on Twitter, we have collected human and deepfake tweets to support the research on deepfake social media text detection in a "real-setting". [13]

[13] `https://www.kaggle.com/mtesconi/twitter-deep-fake-text`

6. Code

6.0.1 Domain Adaptation for Traffic Density Estimation

L. Ciampi, C. Santiago, J.P. Costeira, C. Gennaro, G. Amato [13]

Code for replicating the experiments in [13]. The provided code trains a CNN-based traffic density estimator exploiting GTA, a synthetic collection of urban scenarios suitable for the car counting task. To mitigate the Synthetic2Real Domain Shift, i.e., the image appearance difference between the virtual and the real worlds, it employs an Unsupervised Domain Adaptation training strategy. `https://ciampluca.github.io/unsupervised_counting/`

6.0.2 Counting or Localizing? Evaluating cell counting and detection in microscopy images

L. Ciampi, F. Carrara, G. Amato, C. Gennaro [10]

Code for replicating the experiments in [10]. The provided code trains three commonly adopted cell counting approaches based on detection, density estimation, and segmentation. Through an experimental evaluation over some public collection of microscopy images containing marked cells we show that these counting strategies do not always agree with counting and localization performance. `https://github.com/ciampluca/counting_perineuronal_nets/tree/visapp-counting-cells`

6.0.3 mEye: A Deep Learning Tool for Pupillometry

R. Mazziotti, F. Carrara, A. Viglione, L. Lupori, L. Lo Verde, A. Benedetto, G. Ricci, G. Sagona, G. Amato, T. Pizzorusso [40]

Web app version of the mEye pupillometry tool shipping our trained models.[14] *The tool provides basic pupillometry on pre-recorded videos and live video streams with the web browser as the only requirement. We offer also the code for Python-based workflow at* `https://github.com/fabiocarrara/meye`.

6.0.4 QuaPy: A Python-Based Framework for Quantification

A. Moreo, A. Esuli, F. Sebastiani [51]

QuaPy is an open-source framework for performing quantification (a.k.a. supervised prevalence estimation), written in Python. Quantification is the task of training quantifiers via supervised learning, where a quantifier is a predictor that estimates the relative frequencies (a.k.a. prevalence values) of the classes of interest in a sample of unlabelled data. While quantification can be trivially performed by applying a standard classifier to each unlabelled data item and counting how many data items have been assigned to each class, it has been shown that this "classify and count" method is outperformed by methods specifically designed for quantification. QuaPy provides implementations of a number of baseline methods and advanced quantification methods, of routines for quantification-oriented model selection, of several broadly accepted evaluation measures, and of robust evaluation protocols routinely used in the field. QuaPy also makes available datasets commonly used for testing

[14] Currently hosted at `https://www.pupillometry.it`

quantifiers, and offers visualization tools for facilitating the analysis and interpretation of the results. The software is open-source and publicly available under a BSD-3 licence via GitHub , and can be installed via pip2. [15]

7. Awards

7.1 Best Paper Awards

SAC 2021 The paper "Heterogeneous document embeddings for cross-lingual text classification" (Alejandro Moreo, Andrea Pedrotti, Fabrizio Sebastiani) has won the Best Short Paper Award at the ACM Symposium on Applied Computing (SAC 2021), Gwangju, KR.

IIE 2021

The paper "Garbled-word embeddings for jumbled text" (Gianluca Sperduti, Alejandro Moreo, Fabrizio Sebastiani) has won the Best Short Paper Award at the 11th Italian Information Retrieval Workshop (IIR 2021), Bari, IT.

References

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. The visione video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. *Journal of Imaging*, 7(5):76, 2021.

[2] Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. Visione at video browser showdown 2021. In *International Conference on Multimedia Modeling*, pages 473–478. Springer, 2021.

[3] Francesco Banterle, Alessandro Artusi, Alejandro Moreo, and Fabio Carrara. Nor-vdpnet++: Efficient training and architecture for deep no-reference image quality metrics. In *ACM SIGGRAPH 2021 Talks*, pages 1–2. 2021.

[4] Valentina Bartalesi, Daniele Metilli, Nicolò Pratelli, and Paolo Pontari. Towards a knowledge base of medieval and renaissance geographical latin works: the imago ontology. *Digital Scholarship in the Humanities*, 2021.

[5] Valentina Bartalesi, Nicolò Pratelli, Carlo Meghini, Daniele Metilli, Gaia Tomazzoli, Leyla Maria Gabriella Livraghi, and Michelangelo Zaccarello. A formal representation of the divine comedy's primary sources: The hypermedia dante network ontology. *Digital Scholarship in the Humanities*, 36, 2021.

[6] Donato Cafarelli. Developing and experimenting approaches for facial expression recognition in images from surveillance cameras. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[7] Donato Cafarelli, Fabio Valerio Massoli, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Expression recognition analysis in the wild. *arXiv preprint arXiv:2101.09231*, 2021.

[8] Fabio Carrara, Giuseppe Amato, Luca Brombin, Fabrizio Falchi, and Claudio Gennaro. Combining gans and autoencoders for efficient anomaly detection. In *25th International Conference on Pattern Recognition (ICPR)*, pages 3939–3946. IEEE, 2021.

[9] Fabio Carrara, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato. Defending neural ode image classifiers from adversarial attacks with tolerance randomization. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 425–438. Springer, 2021.

[10] Luca Ciampi, Fabio Carrara, Giuseppe Amato, and Claudio Gennaro. Counting or localizing? evaluating cell counting and detection in microscopy images. In *17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022)*, 2022. accepted.

[11] Luca Ciampi, Fabio Carrara, Valentino Totaro, Raffaele Mazziotti, Leonardo Lupori, Carlos Santiago, Giuseppe Amato, Tommaso Pizzorusso, and Claudio Gennaro. A Multi-Rater Benchmark for Perineuronal Nets Detection and Counting in Fluorescence Microscopy Images, October 2021.

[12] Luca Ciampi, Claudio Gennaro, Fabio Carrara, Fabrizio Falchi, Claudio Vairo, and Giuseppe Amato. Multi-camera vehicle counting using edge-ai. *arXiv preprint arXiv:2106.02842*, 2021.

[13] Luca Ciampi, Carlos Santiago, Joao Costeira, Claudio Gennaro, and Giuseppe Amato. Domain adaptation for traffic density estimation. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2021.

[14] Luca Ciampi, Carlos Santiago, João Paulo Costeira, Claudio Gennaro, and Giuseppe Amato. Traffic density estimation via unsupervised domain adaptation. In *Proceedings of the 29th Italian Symposium on Advanced Database Systems, SEBD 2021, Pizzo Calabro (VV), Italy, September 5-9, 2021*, volume 2994 of *CEUR Workshop Proceedings*, pages 442–449. CEUR-WS.org, 2021.

[15] Daivde Alessandro Coccomini. Design and development of transformer-based methods for video deepfake detection. Master's thesis, MSc in Artificial Intelligence and Data Engineering, University of Pisa, 2021.

[16] Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. *arXiv preprint arXiv:2107.02612*, 2021.

[15] https://github.com/HLT-ISTI/QuaPy

[17] Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Generative adversarial networks for astronomical images generation. *arXiv preprint arXiv:2111.11578*, 2021.

[18] Silvia Corbara and Alessio Molinari. Reading songs: A computational analysis of popular songs lyrics. In *Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021)*, 2021.

[19] Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. Syllabic quantity patterns as rhythmic features for latin authorship attribution. arXiv 2110.14203, 2021.

[20] Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. MedLatinEpi and MedLatinLit: Two datasets for the computational authorship analysis of medieval Latin texts. *ACM Journal of Computing and Cultural Heritage*, 2022. Forthcoming.

[21] Juan José del Coz, Pablo González, Alejandro Moreo, and Fabrizio Sebastiani. Learning to quantify: Methods and applications (lq 2021), 2021.

[22] Juan José del Coz, Pablo González, Alejandro Moreo, and Fabrizio Sebastiani. Learning to quantify: Methods and applications (LQ 2021). In *Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021)*, pages 4874–4875, Gold Coast, AU, 2021.

[23] Marco Di Benedetto, Fabio Carrara, Enrico Meloni, Giuseppe Amato, Fabrizio Falchi, and Claudio Gennaro. Learning accurate personal protective equipment detection from virtual worlds. *Multimedia Tools and Applications*, 80(15):23241–23253, 2021.

[24] Andrea Esuli, Alessio Molinari, and Fabrizio Sebastiani. A critical reassessment of the Saerens-Latinne-Decaestecker algorithm for posterior probability adjustment. *ACM Transactions on Information Systems*, 39(2):Article 19, 2021.

[25] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. LeQua@CLEF2022: Learning to Quantify, 2021. arXiv:2111.11249 [cs.LG].

[26] Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. LeQua@CLEF2022: Learning to Quantify. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022)*, Stavanger, NO, 2022. Forthcoming.

[27] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. Measuring fairness under unawareness via quantification, 2021. arXiv:2109.08549 [cs.CY].

[28] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[29] Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors. *Proceedings of the 43rd European Conference on Information Retrieval Research (ECIR 2021)*. Springer Nature, Cham, CH, 2021. Lecture Notes in Computer Science, Volumes 12656 and 12657.

[30] Gabriele Lagani, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Hebbian semi-supervised learning in a sample efficiency setting. *Neural Networks*, 143:719–731, 2021.

[31] Gabriele Lagani, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Comparing the performance of hebbian against backpropagation learning using convolutional neural networks. *Neural Computing and Applications*, Accepted.

[32] Gabriele Lagani, Raffaele Mazziotti, Fabrizio Falchi, Claudio Gennaro, Guido Marco Cicchini, Tommaso Pizzorusso, Federico Cremisi, and Giuseppe Amato. Assessing pattern recognition performance of neuronal cultures through accurate simulation. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 726–729. IEEE, 2021.

[33] Antonio Liuzzi. Design and development of ai-based video surveillance applications for low-power embedded systems. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[34] Fabio Valerio Massoli, Donato Cafarelli, Giuseppe Amato, and Fabrizio Falchi. A multi-resolution training for expression recognition in the wild. In *SEBD 2021 - Italian Symposium on Advanced Database Systems*, pages 427–433, 2021.

[35] Fabio Valerio Massoli, Donato Cafarelli, Claudio Gennaro, Giuseppe Amato, and Fabrizio Falchi. Mafer: a multi-resolution approach to facial expression recognition. *arXiv preprint arXiv:2105.02481*, 2021.

[36] Fabio Valerio Massoli, Fabio Carrara, Giuseppe Amato, and Fabrizio Falchi. Detection of face recognition adversarial attacks. *Computer Vision and Image Understanding*, 202:103103, 2021.

[37] Fabio Valerio Massoli, Fabrizio Falchi, Alperen Kantarci, Şeymanur Akti, Hazim Kemal Ekenel, and Giuseppe Amato. Mocca: Multilayer one-class classification for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[38] Fabio Valerio Massoli, Lucia Vadicamo, Giuseppe Amato, and Fabrizio Falchi. A leap among entanglement and neural networks: A quantum survey. *arXiv preprint arXiv:2107.03313*, 2021.

[39] Lucas May Petry, Camila Leite Da Silva, Andrea Esuli, Chiara Renso, and Vania Bogorny. Marc: a robust method for multiple-aspect trajectory classification via space, time, and semantic embeddings. *International Journal of Geographical Information Science*, 34(7):1428–1450, 2020.

[40] Raffaele Mazziotti, Fabio Carrara, Aurelia Viglione, Leonardo Lupori, Luca Lo Verde, Alessandro Benedetto, Giulia Ricci, Giulia Sagona, Giuseppe Amato, and Tommaso Pizzorusso. Meye: Web app for translational and real-time pupillometry. *eNeuro*, 8(5), 2021.

[41] Francesco Merola. Designing, implementing and experimenting a deep reinforcement learning approach in a simulated environment for autonomous driving in critical scenarios. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[42] Francesco Merola, Fabrizio Falchi, Claudio Gennaro, and Marco Di Benedetto. Reinforced damage minimization in critical events for self-driving vehicless. In *17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2022. accepted.

[43] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Recurrent vision transformer for solving visual reasoning problems. *arXiv preprint arXiv:2111.14576*, 2021.

[44] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters*, 143:75–80, 2021.

[45] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021.

[46] Nicola Messina, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features. In *18th International Conference on Content-Based Multimedia Indexing, CBMI 2021, Lille, France, June 28-30, 2021*, pages 1–6. IEEE, 2021.

[47] Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. Transformer reasoning network for image-text matching and retrieval. In *25th International Conference on Pattern Recognition (ICPR)*, pages 5222–5229. IEEE, 2021.

[48] Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Aimh at semeval-2021 task 6: multimodal classification using an ensemble of transformer models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1020–1026, 2021.

[49] Filippo Minutella. Design and implementation of a deep learning system for knowledge graph analysis. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[50] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Lost in transduction: Transductive transfer learning in text classification. *ACM Transactions on Knowledge Discovery from Data*, 16(1):13:1–13:21, 2021.

[51] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. QuaPy: A Python-based framework for quantification. In *Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021)*, pages 4534–4543, Gold Coast, AU, 2021.

[52] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. QuaPy: A Python-based framework for quantification, 2021. arXiv:2106.11057 [cs.LG].

[53] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Word-class embeddings for multiclass text classification. *Data Mining and Knowledge Discovery*, 353(3):911–963, 2021.

[54] Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. Generalized funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification, 2021. arXiv:2110.14764 [cs.CL].

[55] Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. Generalized funnelling: Ensemble learning and heterogeneous document embeddings for cross-lingual text classification (extended abstract). In *Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021)*, Bari, IT, 2021.

[56] Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. Heterogeneous document embeddings for cross-lingual text classification. In *Proceedings of the 36th ACM Symposium on Applied Computing (SAC 2021)*, pages 685–688, Gwangju, KR, 2021.

[57] Alejandro Moreo and Fabrizio Sebastiani. Re-assessing the "classify and count" quantification method. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021)*, volume II, pages 75–91, Lucca, IT, 2021.

[58] Alejandro Moreo and Fabrizio Sebastiani. Tweet sentiment quantification: An experimental re-evaluation. *PLoS ONE*, 2022. Forthcoming.

[59] Partarakis N., Kaplanidi D., Doulgeraki P., Karuzaki E., Petraki A., Metilli D., Bartalesi V., Adami I., Meghini C., and Zabulis X. Representation and presentation of culinary tradition as cultural heritage. *Heritage (Basel) Online*, 4:612–640, 2021.

[60] Raffaele Perego and Fabrizio Sebastiani. Report on the 43rd European Conference on Information Retrieval (ECIR 2021). *ACM SIGIR Forum*, 55(1), 2021.

[61] Ruggero Rabitti. Design and implementation of a framework for reinforcement learning for self-driving cars. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[62] Danilo Sorano, Fabio Carrara, Paolo Cintia, Fabrizio Falchi, and Luca Pappalardo. Automatic pass annotation from soccer video streams based on object detection and lstm. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 475–490, Cham, 2021. Springer International Publishing.

[63] Gianluca Sperduti, Alejandro Moreo, and Fabrizio Sebastiani. Garbled-word embeddings for jumbled text. In *Proceedings of the 11th Italian Information Retrieval Workshop (IIR 2021)*, 2021.

[64] Maria Taibi. Development of a park monitoring system for smart camera network. Master's thesis, MSc in Computer Engineering, University of Pisa, 2021.

[65] Fabrizio Falchi Margherita Gambini Maurizio Tesconi Tiziano Fagni, Antonio Martella. Tweep fake – twitter deep fake text dataset. `https://www.kaggle.com/mtesconi/twitter-deep-fake-text`.

[66] Bartalesi Lenzi V. and Pratelli N. The imago project: towards a knowledge base of medieval and renaissance geographical works. In *SWODCH 2021 - International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage, Bolzano, 20-21/09/2021*, 2021.

[67] Lucia Vadicamo, Richard Connor, and Edgar Chávez. Query filtering using two-dimensional local embeddings. *Information Systems*, 101:101808, 2021.

[68] Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato. On generalizing permutation-based representations for approximate search. In *International Conference on Similarity Search and Applications*, pages 66–80. Springer, 2021.

[69] Lucia Vadicamo, Claudio Gennaro, Fabrizio Falchi, Edgar Chávez, Richard Connor, and Giuseppe Amato. Re-ranking via local embeddings: A use case with permutation-based indexing and the nsimplex projection. *Information Systems*, 95:101506, 2021.