You've Got the Wrong Number:

Evaluating Deep Learning Training Paradigms Using Handwritten Digit Recognition Data

Giacomo Ignesti(https://orcid.org/0000-0003-2389-3086)[1,2],

Massimo Martinelli(https://orcid.org/0000-0002-5175-5126) [2] and

Davide Moroni(https://orcid.org/0000-0002-5175-5126) [2]

[1] University of Pisa, Pisa, 5123, Italy, [2] Institute of Information Science and Technologies, National Research Council, Pisa, 51264, Italy

Email: giacomo.ignesti@isti.cnr.it, massimo.martinelli@isti.cnr.it, davide.moroni@isti.cnr.it

Abstract

To build more accurate and trustworthy artificial intelligence algorithms in deep learning, it is essential to understand the mechanisms driving classification systems to identify their targets. Typically, post hoc methods provide insights into this process. In this preliminary work, we shift the reconstruction of the class activation map to the training phase to evaluate how the model's performance changes compared to standard classification approaches. The MNIST dataset and its variants, such as Fashion MNIST, consist of well-defined images that facilitate testing this type of training process. Specifically, the classification targets are the only significant content in the images, excluding the background, allowing for a direct comparison of the reconstruction against the input images. To enhance the guidance of the network, we introduce a contrastive loss term to complement the standard classification function, which often uses categorical cross-entropy. By comparing the accuracy and the extracted pattern of the standard approach with the proposed method, we can gain valuable insights into the network's learning process. This approach aims to

improve the interpretability and effectiveness of the model during training, ultimately leading to higher classification accuracy and reliability.

Keywords: Artificial Intelligence, Trustworthy, Reliability, Loss functions, Image Reconstruction

Introduction

Trustworthy Artificial Intelligence (TAI) is one of the scientific community's main objectives, necessitating a multifaceted approach encompassing legal, ethical, and scientific perspectives to ensure these breakthrough technologies' safe development and deployment. Each domain contributes uniquely. Legal frameworks aim to build regulations to adapt Law to the fast growth of AI applications. Ethics proposes guidelines so developers and organisations can create AI systems that respect human rights and societal values, promoting a human-centric approach to technology [3]. The scientific community of mathematics, physics and computer scientists addresses two main questions: what is the system learning, and why is learning that? This paper delves into these latest topics, introducing a new possible paradigm to train deep learning (DL) algorithms. In recent years, the field of Explainable Artificial Intelligence (XAI) [2] has emerged as a crucial approach to understanding the inner workings of AI systems. XAI methods aim to identify the ensemble of processes that define the fully trained algorithm used to solve a specific task. Two primary solutions employed in XAI are post hoc and transparent approaches. Post hoc methods explain why and how trained systems generate outputs given inputs, while transparent approaches concentrate on designing algorithms with more straightforward and precise inner workings. Prominent examples of post hoc methods in deep learning (DL) literature include LIME, SHAP, and GradCAM[1,4], commonly used for post-training model evaluation. In contrast, classical machine learning (ML) algorithms like decision trees or clustering approaches are more interpretable than

DL models. Although XAI is now a staple part of AI research, it is essential to recognise that it is not the only approach to addressing interpretation problems. The scientific community also uses "Reliable AI"[5] to comprehensively identify all the mathematical approaches to solving DL network problems. The main idea is that explainability is just one aspect of developing Plausible AI. Other elements include expressivity, learning, and generalisation. The term "expressivity" refers to the choice of network architecture, while "learning" encompasses the gradient descent and the optimisation issues beyond the choice of the loss function. "Generalisation" includes all the statistical robustness problems of the network. These three and the “Explainability” problem are the key points towards developing a plausible AI. The proposed research begins with testing innovative training approaches to develop more plausible and trustworthy AI systems. The foundation of this study is the MNIST [9] dataset, which consists of 70,000 images of handwritten digits, 60,000 for training and 10,000 for testing. The focal inquiry of this research is: "What features does the network utilise to classify the digits?" To investigate this, the GradCAM algorithm or its adaptations can be employed to identify the specific regions of the image that the network relies on for its predictions. In the case of the MNIST dataset, the network should focus on the digit itself, excluding the background. Given that the images are single-channel grayscale, the network should primarily utilise the grayscale values, mostly avoiding black pixels, to classify the digits. The GradCAM output should ideally align closely with the digit images themselves. Traditionally, network training emphasises optimising hyperparameters, architecture, and loss functions to achieve high accuracy. However, in the MNIST context, evaluating the network's performance is feasible before the training phase. This can be accomplished by implementing a GradCAM-like function during training, providing the network with a reference to ensure its outputs are meaningful and interpretable. The approach outlined here differs from similar methods,

such as those using prototype networks and attribution maps. Usually, prototype networks [12] use a set of representative examples (prototypes) to guide classification, which can sometimes obscure the underlying decision-making process of the model. In contrast, the proposed method leverages GradCAM to visualise the input areas that influence the model's predictions, offering a more intuitive understanding of the model's behaviour. Another similar approach is defined by Concept relevance propagation [8]. This method emphasises the extraction of concept relevance, which aims to translate model outputs into human-understandable explanations by identifying the underlying concepts that influence decisions, providing a more abstract interpretation of model behaviour rather than a direct visualisation of learned and extracted features as in the proposed approach.

One of the primary challenges of this research is ensuring that the GradCAM outputs are accurate and interpretable. If the visualisations do not correlate well with the actual features used for classification, it may lead to misconceptions about the model's decision-making process. However, the benefits of this approach are substantial. Researchers can develop high-accuracy models by integrating interpretability into the training process and providing insights into their operational logic. This can enhance user trust and facilitate the adoption of AI technologies in critical applications where understanding the reasoning behind decisions is essential. In summary, this research aims to bridge the gap between model performance and interpretability, fostering the development of AI systems that are effective and comprehensible to users.

Methods

The partial preliminary method is tested on MNIST and a subset of Fashion MNIST[11]. The data are analysed using the one-channel approach, and the only data transformation is tensorisation and

normalisation to the one-channel MNIST and Fashion MNIST literature values of [0.1307, 0.3081] and [0.2860, 0.3530]. The employed model is a convolutional neural network (CNN). The network consists of three convolutional layers, each followed by batch normalisation, ReLU activation, and max pooling, progressively increasing the number of filters from 32 to 128. A dropout layer is applied before the fully connected layers to mitigate overfitting, which consists of a hidden layer with 256 neurons followed by an output layer with ten neurons corresponding to the digit classes. Additionally, the model incorporates hooks to capture feature maps and gradients from the last convolutional layer, enabling the computation of Class Activation Maps (CAM). During the forward pass, the model processes input images through the convolutional layers, applies dropout, and computes outputs while generating CAMs for each image to visualise the regions contributing to the predictions. The model is tested with two different training functions on a GPU; the functions are the same, with the only difference that one of them incorporates in the loss evaluation function a component of contrastive loss used to evaluate the feature maps. The custom contrastive loss function [7] implemented is designed to optimise the similarity between two input tensors, the GradCAM reconstructed image and the actual input, based on a given label. The forward pass of the losses takes two output tensors and the ground truth label for each image; since the tensor's shape differs in size, the GradCAM image is reconstructed using bicubic interpolation. The Euclidean distance between the two-dimension tensors is then calculated, and a constant small epsilon value [e^{-22}] is added to prevent division by zero. The loss is computed based on the label tensor, where a positive label contributes to the objective based on the squared Euclidean distance. In contrast, a negative label contributes based on the squared clamped distance (margin-distance). The final loss is the mean of the computed values since it’s evaluated on a batch dimension. It is

suggested to check for NaN values in the input tensors and return a zeroed tensor if any are found to avoid invalid operations. The Contrastive Loss LL can be expressed as:

$$¿=\frac{1}{2}\left(y\,d^2+(1-y)\cdot max\left(0,m-d\right)^2\right) \quad (1)$$

Where:

- y is the label (1 for similar pairs, 0 for dissimilar pairs).
- d is the Euclidean distance between the two output tensors.
- m is the margin (a hyperparameter).

The Multi-Class Cross-Entropy Loss LCE can be expressed:

$$LCE=\frac{-1}{N}\sum_{i=1}^{C}\sum_{i=1}^{N} y_{i,c}*\log\left(p_{i,c}\right) \quad (2)$$

One model is trained using only the LCE loss as in classical DL approaches, while the other uses LCE and LL in combination as L_{TOT} = LCE + LL. The optimiser employed is the classical ADAM optimiser with a one over one thousand learning rate. The two obtained models are then evaluated on the test set, and the image that the network generates is used as a final reference. The language of the code implemented is Python, and the module used for DL implementation is PyTORCH. The GPU employed is an NVIDIA QUADRO RTX 5000, and no data parallelism is employed.

Results

In the following, the trained models are labelled based on the dataset used, with an added postfix. Specifically, the postfix 'R' denotes models trained with a combination of contrastive and cross-

entropy loss functions, while the postfix 'C' indicates models trained using only the cross-entropy loss.

The models MNISTR and MNISTC show similar performance: Table 1 shows the training accuracy is around 98% for the R model and 99% for the C model. On the testing set, the pattern is similar; the R model reaches 98% accuracy while the C model reaches 99% accuracy.

	MNISTR	**MNISTC**
Test Accuracy	98.44%	99.17%
True predictions	9844	9917
False predictions	156	83
True Low confidence	119	30

Table 1 Results of MNISTR and MNISTC models on the test set.

As shown in Table 2, the training time is different: while the R model stops early, at epoch 9, at 18.88 hours of training, the classification model stops at epoch 13, at 23.49 hours.

Both models employ the same number of parameters as exposed in the following table.

Number of parameters:	390.86 K
FLOPS:	16.53 M
MACs:	7.75 M

Table 2 Model information shared between the two models.

The fundamental difference is the quality of the explanations given in which, theoretically, the R model surpasses the C model.

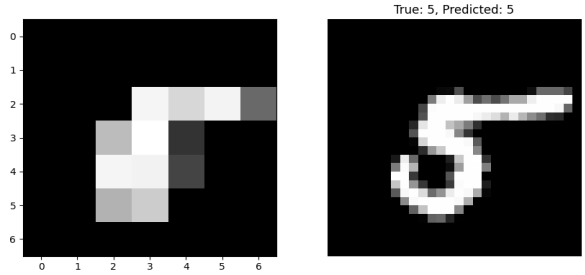

Figure 1 The digits five on the right and the reconstructed map on the left at epochs 1 of the R model, MNIST.

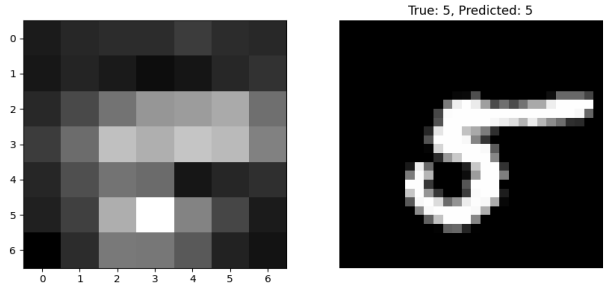

Figure 2 The digits five of MNIST on the right and the reconstructed map on the left at epochs 1 of the C model.

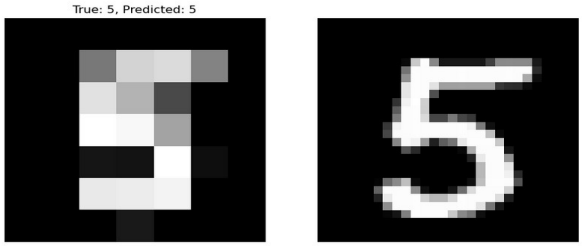

Figure 3 The digits five of MNIST on the right and the reconstructed map on the left at epochs 9 of the R model.

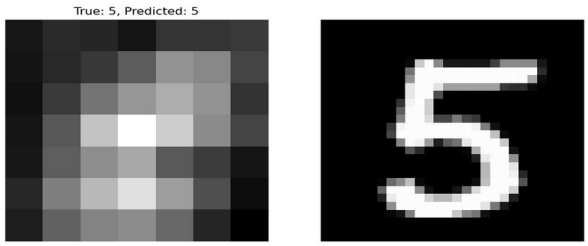

Figure 4 The digit five of MNIST on the right and the reconstructed map on the left at epoch 1 of the R model.

The model R can detect the shape of the five much better in the first epoch and progressively improve during the nine epochs (Figures 1 and 3). In contrast, the C model more accurately delineates something else (Figures 2 and 4). The results of FASHIONMINSTR and FASHIONMNISTC follow a similar pattern. Naturally, the accuracy on the test set, considered a subset of 15%, is lower, around 65% on the test set for the R model and 64% for the C model.

In both cases, the reconstructed images (Figures 5 and 6) are of lower quality.

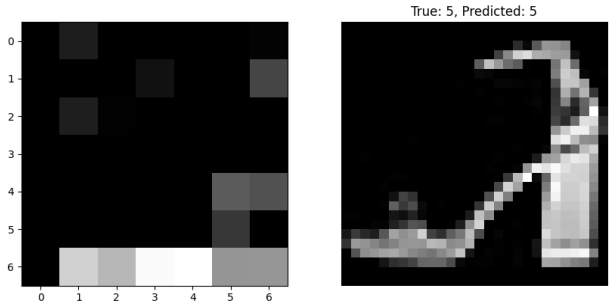

Figure 5 A shoe of Fashion MNIST on the right and the reconstructed map on the left at epochs 1 of the R model.

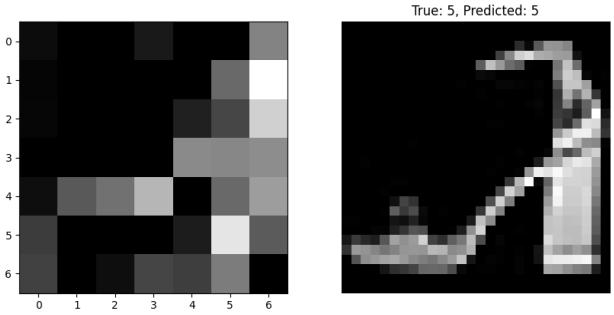

Figure 6 A shoe of Fashion MNIST on the right and the reconstructed map on the left at epochs 1 of the C model.

Discussion

Model accuracy on MNIST has been consistently high since the Convolutional Neural Networks (CNNs) advent. In contrast, the impressive accuracy achieved in initial studies may not be groundbreaking, but it is nonetheless promising, as illustrated in Table 1. Using contrastive loss reconstruction approaches has yielded exciting results, with the network effectively identifying the shapes of the desired digits, thereby enhancing reconstruction over successive epochs and seemingly mitigating overfitting without significantly impairing generalisation. However, specific images classified with low confidence or misclassified by the model can still pose interpretative challenges for an average observer (Figures 7 and 8). Preliminary results from Fashion MNIST highlight the necessity for more refined methodologies, either in model architecture or in evaluating complexity regarding the target class. To further enhance model performance, it is crucial to investigate the dimensions of the feature maps relative to the input images and the target sign [6]. Understanding how the feature maps evolve throughout the network can provide insights into the model's ability to capture essential patterns and shapes relevant to classification tasks. Exploring different optimiser functions, such as RMSprop and SGD, and potential learning rate schedulers could significantly impact convergence speed and accuracy.

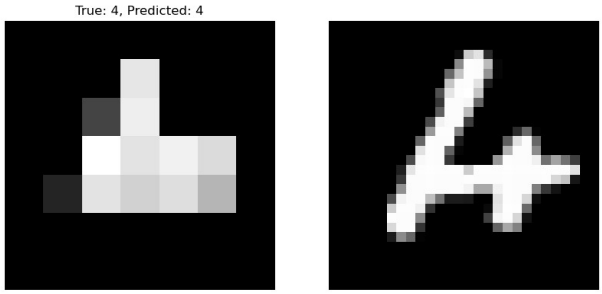

Figure 7 Digits four of MNIST on the right and the reconstructed map on the left in the test set of the C model.

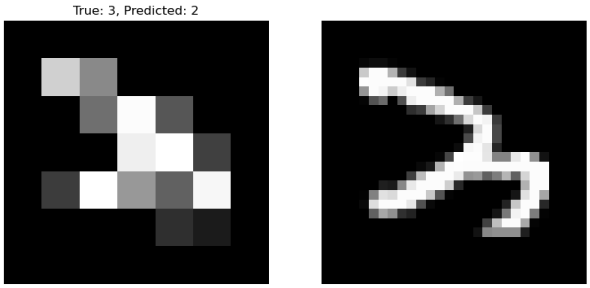

Figure 8 Digits three of MNIST on the right and the reconstructed map on the left in the test of the R model.

The choice of optimiser and its configuration can influence how effectively the model learns from the data. At the same time, a well-tuned scheduler can help maintain optimal learning rates throughout training, preventing issues like overshooting minima. Moreover, considering alternative reconstruction loss functions, such as triplet loss or variational loss, may provide further improvements in capturing the nuances of the data distribution. By systematically evaluating these dimensions, feature map sizes, optimisation strategies, and reconstruction losses, the research can refine its approach to achieve higher accuracy and robustness in both MNIST and more complex datasets like Fashion MNIST.

Conclusion

This work connects the training and explanation steps in developing plausible AI. To comprehensively evaluate the proposed approaches, it is essential to explore further the relationship between these steps, expressivity, and optimisation. Testing different and more advanced backbones for feature extraction, particularly those utilising residual connections[10], is recommended. This strategy will enhance the network's ability to determine whether reconstruction contrastive loss improves its interpretation of classification problems. Minimising information loss in the reconstructed portions of the image is crucial. Future research will focus on a detailed analysis of the Fashion MNIST dataset and validating the developed approach using medical images.

FUNDING

This work was not supported by any external funding sources. No additional grants or financial support were obtained to carry out or direct this research.

CONFLICT OF INTEREST

The authors of this work declare that they have no conflicts of interest.

COMPLIANCE WITH ETHICAL STANDARDS

This work does not involve any studies with human or animal subjects and complies with all relevant ethical standards.

Bibliography

Journal Paper

1. R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, & S. Lapuschkin (2023). "From attribution maps to human-understandable explanations through concept relevance propagation." *Nature Machine Intelligence*, 5(9), 1006-1019. https://doi.org/10.48550/arXiv.2206.03208

2. F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, & S. Rinzivillo (2023). "Benchmarking and survey of explanation methods for black box models." *Data Mining and Knowledge Discovery*, 37(5), 1719-1778. https://doi.org/10.48550/arXiv.2404.16903

3. M. Brundage, et al. (2020). "Toward trustworthy AI development: mechanisms for supporting verifiable claims." arXiv preprint arXiv:2004.07213. https://doi.org/10.48550/arXiv.2004.07213

4. A. Chattopadhay, et al. (2018). "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 10.1109/WACV.2018.00097

5. C. Chen, O. Li, C. Tao, A.J. Barnett, & J. Su (2019). "This Looks Like That: Deep Learning for Interpretable Image Recognition." *Advances in Neural Information Processing Systems* , 32. https://doi.org/10.48550/arXiv.1806.10574

6. K. He, X. Zhang, S. Ren, & J. Sun (2016). "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778). 10.1109/CVPR.2016.90

7. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, ... & D. Krishnan (2020). "Supervised contrastive learning." *Advances in Neural Information Processing Systems*, 33, 18661-18673. https://doi.org/10.48550/arXiv.2004.11362

8. G. Kutyonik (2024). "The Mathematics of Reliable Artificial Intelligence." *SIAM News*, July/August.

9. Y. LeCun, et al. (1998). "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11), 2278-2324. 10.1109/5.726791

10. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, ... & B. Guo (2021). "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012-10022). 10.1109/ICCV48922.2021.00986

11. R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, & D. Batra (2020). "Grad-CAM: visual explanations from deep networks via gradient-based localization." *International Journal of Computer Vision*, 128(2), 336-359. https://doi.org/10.1109/ICCV.2017.74

12. H. Xiao, K. Rasul, & R. Vollgraf (2017). "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms." arXiv preprint arXiv:1708.07747. https://doi.org/10.48550/arXiv.1708.07747

BIOGRAPHY

Giacomo Ignesti is a PhD student in the National PhD in Artificial Intelligence at the University of Pisa; he is also a Fellow Graduate at the Signals & Images Lab, ISTI-CNR in Pisa, Italy. He graduated in Biomedical Engineering from Sapienza University of Rome and a master's in data science and Machine Learning in Precision Medicine from Padua University. His PhD research focuses on the safe development of AI applications in healthcare, particularly in signal and image processing. Currently, he is involved in PNRR

projects examining the effects of DMT on senior citizens and has previously worked on Telemedicine/Telehealth applications.

Davide Moroni (M.Sc. in Mathematics, University of Pisa, 2001; Dipl. from Scuola Normale Superiore, 2002; Ph.D. in Mathematics, University of Rome La Sapienza, 2006) is a Senior Researcher and Head of the Signals and Images Lab at ISTI, National Research Council, Pisa. He chairs the MUSCLE working group of the European Consortium for Informatics and Mathematics and, since 2018, the TC16 of the International Association for Pattern Recognition (IAPR). His research focuses on geometric modeling, topological data analysis, image processing, computer vision, and medical imaging.

Massimo Martinelli is member of the Signals & Images research laboratory at the Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Italy, Pisa, since 1987. Head of the 'Artificial Intelligence Technologies and Frameworks Area' at SI-Lab since 2017. He was member of the W3C Multimedia Semantics Incubator Group (2006–2007). He is currently leading the CNR-ISTI team in the projects TiAssisto (Tuscany Region), Barilla Agrosat Plus (industrial), Cloud Pathology (industrial), and of the Scientific Agreements with the UO Otolaryngology, audiology and phoniatrics (UNIPI), and with the Italian Mountain Medicine Society. Topic Editor of 'Machine Learning and Biomedical Sensors' of the Sensors Journal. His main scientific interests include Computer Vision, Deep Learning, Decision Support Systems.