Submitted to International Journal of Multimedia Information Retrieval
Final authenticated publication: https://doi.org/10.1007/s13735-019-00178-7

Learning Visual Features for Relational CBIR

Nicola Messina · Giuseppe Amato · Fabio Carrara · Fabrizio Falchi · Claudio Gennaro

Submitted: 15 April 2019 / Revised: 20 July 2019 / Accepted: 4 September 2019

Abstract Recent works in deep-learning research highlighted remarkable relational reasoning capabilities of some carefully designed architectures. In this work, we employ a relationship-aware deep learning model to extract compact visual features for use as relational image descriptors. In particular, we are interested in Relational Content-Based Image Retrieval (R-CBIR), a task consisting in finding images containing similar inter-object relationships. Inspired by the Relation Networks (RN) employed in Relational Visual Question Answering (R-VQA), we present novel architectures to explicitly capture relational information from images in the form of network activations that can be subsequently extracted and used as visual features. We describe a two-stage Relation Network module (2S-RN), trained on the R-VQA task, able to collect non-aggregated visual features. Then, we propose the Aggregated Visual Features Relation Network (AVF-RN) module, that is able to produce better relationship-aware features by learning the aggregation directly inside the network. We employ an R-CBIR ground-truth built by exploiting scene-graphs similarities available in the CLEVR dataset in order to rank images in a relational fashion. Experiments show that features extracted from our two-stage RN (2S-RN) model provide an improved retrieval performance with respect to standard non-relational methods. Moreover, we demonstrate that the features extracted from the novel AVF-RN can further improve the performance measured on the R-CBIR task, reaching the state-of-the-art on the proposed dataset.

N.Messina, G. Amato, F. Carrara, F. Falchi, C. Gennaro
via G. Moruzzi, 1 - 56124 Pisa, Italy
<name>.<surname>@isti.cnr.it

Keywords CLEVR, Content-Based Image Retrieval, Deep Learning, Relational Reasoning, Relation Networks, Deep Features

1 Introduction

Recent advances in deep-learning technologies brought to light remarkable capabilities of neural networks. In particular, focusing on the computer vision world, one of the aims of deep-learning architectures consists in understanding the content of an image at a high-level of abstraction. In this respect, some specific tasks have been developed in order to test the capabilities of newly proposed architectures to cope with high-level reasoning.

Understanding relationships between entities is considered a difficult task since it requires complex reasoning skills. For this reason, some challenging tasks such as Relational Visual Question Answering (R-VQA) and Visual Relationships Detection (VRD) have been introduced as reference tasks for probing relational capabilities of deep-learning solutions. R-VQA consists of answering questions related to difficult inter-object relationships in an image; on the other hand, VRD tries to recover relationships between couples of objects in the images by coding the information under the form of triplets *subject*, *predicate*, *object*. R-VQA and VRD underlined some of the difficulties that current deep-learning approaches present when it comes to reasoning about relationships between different objects: plain convolutional architectures showed major performances in tasks such as image classification or object recognition; however, they exhibit some limitations in relational contexts.

In this work, we analyze the possibility of applying relational understanding capabilities to the Content-Based Image Retrieval (CBIR) task. More in details, we are interested in the sub-field of Relational-CBIR (R-CBIR) in which the aim is to retrieve images with given relationships among objects.

This study is focused on bringing image retrieval a step further with respect to current approaches, keeping the basic idea untouched. In fact, the similarity between two images is always measured as affinity among some sort of high-level features extracted from the image. Our objective consists in extracting a relationship-aware descriptor able to embed relational information. These descriptors should be easily comparable using standard distance metrics so that they can be used in standard indexing engines. The distance between features should embody the dissimilarity between the respective images in terms of relationships between the objects contained in them.

The key contribution of this work is the introduction of architectures able to learn relational features directly inside the network. These proposed architectures, however, are not trained directly on the R-CBIR task; instead, this work investigates upon the possibility of learning features from networks trained on the task of R-VQA.

The transfer-learning methodology is not a novel approach for CBIR. Standard CBIR features are extracted from architectures trained for example on image classification tasks. Image classification, however, does not require the architecture to learn difficult relational concepts. Hence, as far as R-CBIR is concerned, relational-aware features can be extracted from architectures trained on a task that requires high-level reasoning capabilities, and the R-VQA tasks perfectly fill this need. In fact, we rely on the assumption that architectures that are able to correctly answer questions on complex inter-object relationships have internally learned some relational concepts that can be later extracted and compared.

We perform this study in a fully controlled environment, using the images and scene graphs provided by the CLEVR synthetic dataset. CLEVR is a diagnostic dataset originally designed for the task of R-VQA, and it is composed of 3D rendered scenes made up of simple shapes. Unlike real-world datasets like Visual Genome, it avoids common relational biases. Also, being a highly controlled environment, it is useful to test in fine details the very specific relational capabilities of deep-learning architectures.

In this work, we extend the study published at the CEFRL workshop of ECCV 2018 on 2S-RN [18] in which we discussed the possibility of extracting relationship-aware visual features from an architecture trained on the R-VQA task. 2S-RN is designed in a way that extracted features should be aggregated afterwards, by averaging all the contributions from every objects couple. For this reason, it is possible that the aggregated features are not embedding in an efficient manner all the information needed for fully describing a scene. The novel proposed network Aggregated Visual Features Relation Network (AVF-RN) solves this problem by learning the aggregation directly inside the network. By doing so, we are obliging the network to incorporate as much information as possible inside the aggregated features. Hence, the extracted activations can immediately be used as compact visual features. To sum up, we extend the 2S-RN approach by adding the following contributions:

- we propose the Aggregated Visual Features Relation Network (AVF-RN), a novel architecture that is able to learn aggregated relationship-aware features directly inside the network;
- we train AVF-RN on the R-VQA task on the CLEVR dataset;
- we compare the features extracted from the AVF-RN network with 2S-RN features on the R-CBIR task, using three different CLEVR dataset configurations; we also include as non-relational baseline the CNN features extracted from a simple model trained on multi-label classification on CLEVR scenes.

The rest of the paper is organized as follows. In section 2, we review some of the works belonging to the Relational Learning world, mainly focusing on VRD, R-VQA, and R-CBIR. In section 3, we describe in details the process needed for creating the relational ground truth from CLEVR. In section 4, we describe in details the proposed AVF-RN architecture. In section 5, we describe our experimental setup, we collect the results also considering baseline architectures present in the literature, and we discuss the obtained results. Finally, in section 6, we recap our contribution, and we present future directions for this research.

2 Related Work

In this section, we review some of the works related to Relational Learning in particular related to Relational Visual Question Answering (R-VQA) and Visual Relationship Detection (VRD) tasks. Afterward, we review some of the existing approaches to Relational CBIR (R-CBIR).

Visual Relationship Detection (VRD) Recent work has addressed the problem of visual relationships detection

(VRD) in images in the form of triplets (*subject*, *predicate*, *object*), where *subject* and *object* are common objects present in an image, and *predicate* indicates a relationship between them out of a set of possible relationships containing verbs, prepositions, comparatives, etc.

Several datasets are comprised of a large set of visual relationships, such as [11,13,19]. They have opened the way to approaches aimed to detect inter-object relationships in images [13,19,4].

A common approach to VRD employed by many [13, 27,20,29] consists at first in proposing entities using region proposal networks, such as Faster-RCNN [23]. Then, once the entities have been located, a network tries to reason on the relationships occurring between them.

Notwithstanding approaches that solve VRD are able to detect relationships, they usually do not encode the learned information in a compact representation: all possible relationships are combinatorially tested on prediction time.

Relational VQA (R-VQA) R-VQA comes from the basic task of VQA (Visual Question Answering). Plain VQA consists in giving the correct answer to a question asked on a given picture, so it requires connecting together different entities coming from heterogeneous representations (text and visuals).

Some works [31,28] proposed approaches to standard VQA problems on datasets such as VQA [1], DAQUAR [15], COCO-QA [22].

Recently, there is a tendency to conceptually separate VQA and Relational-VQA (R-VQA). In R-VQA, in fact, images contain difficult inter-object relationships, and question are formulated in a way that it is impossible for deep architectures to answer correctly without having understood high-level interactions between the objects in the same image. Some datasets, such as CLEVR [7], RVQA [14], FigureQA [10], move the attention towards this new challenging task.

On the CLEVR dataset, [25] and [21] authors proposed a novel architecture specialized to think in a relational way. They introduced a particular layer called Relation Network (RN), which is specialized in comparing pairs of objects. Objects representations are learned by means of a four-layer CNN, and the question embedding is generated through an LSTM. The overall architecture, composed of CNN, LSTM, and the RN, can be trained fully end-to-end, and it is able to reach superhuman performances. Other solutions [6,8] introduce compositional approaches able to explicitly model the reasoning process by dynamically building a reasoning graph that states which operations must be carried out and in which order to obtain the right answer. These architectures are internally split into two different subcomponents: a *generator network* that produces an execution graph based on the question embeddings, and an *execution network* that executes the graph produced by the generator network taking in input the image features and outputting the answer. Usually, these architectures tend to perform poorly when related to other approaches.

In order to close the performance gap between interpretable architectures and high performing solutions, [16] proposed a set of visual-reasoning primitives that are able to perform complex reasoning tasks in an explicitly interpretable manner.

R-CBIR While standard CBIR captured a lot of attention even before the deep-learning era, R-CBIR involves complex reasoning skills and current deep-learning approaches have shown promising results in this direction.

Nevertheless, in this work, we use the same basic ideas from the standard CBIR methodology; we act only on the features extraction process. We take as reference the work by [26] that introduced RMAC features — one of the state-of-the-art non-relational image descriptors for image instance retrieval. This descriptor encodes and aggregates several regions of the image in a dense and compact global image representation exploiting a pre-trained fully convolutional network for feature map extraction. The aggregated descriptor is obtained by max-pooling the feature map over different regions and scales and summing them together.

As regards the work carried out on R-CBIR, there was some experimentation using both CLEVR and real-world datasets. [9] introduced a CRF model able to ground relationships given in the form of a scene graph to test images for image retrieval purposes. However, this model is not able to produce a compact feature. They employed a simple dataset composed of 5000 images and annotated with objects and their relationships.

More recently, using the Visual Genome dataset, [30] implemented a large scale image retrieval system able to map textual triplets into visual ones (object-subject-relation inferred from the image) projecting them into a common space learned through a modified version of triplet-loss.

The works by [2,18] exploit the graph data associated with every image in order to produce ranking goodness metrics, such as nDCG and Spearman-Rho ranking correlation indexes. Their objective was evaluating the quality of the ranking produced for a given query, keeping into consideration the relational content of every scene.

3 A Relational-CBIR Ground-Truth

In order to evaluate the quality of any relational feature extracted from a relationship-aware system, we compute a specific ground-truth exploiting relational knowledge embedded into graphs (*scene-graphs*).

By carefully choosing a distance function between graphs, we are able to give a good estimation of the relational similarity between scenes. In order to accomplish this task, we need some datasets that include a formal and precise description of relations occurring inside the scene. In this work, we will use the synthetic generated dataset CLEVR [7].

3.1 CLEVR

CLEVR [7] is a synthetic dataset composed of 3D rendered scenes, and it has been designed for the R-VQA task. There are 100k rendered images subdivided among training (70k), validation (15k), and test (15k) sets. The total number of questions is ∼865k again split among training (∼700k), validation (∼150k), and test (∼15k).

The main concept behind CLEVR is the *scene*. A *scene* contains different simple shaped objects with mixtures of colors, materials, and sizes. There are cubes, spheres, and cylinders, each one of which can have a color chosen among eight; they can be big or small, and they can be made of one of two different materials, metal or rubber. The *scene* is fully and uniquely described by a *scene graph*. The scene graph describes in a formal way all the relationships between objects.

The question is formulated under the form of a *functional program*. The answer to a question represented by its functional program on a scene is simply calculated by executing the functional program on the scene graph. Scene graphs are rendered to photo-realistic 3D scenes by using Blender, a free 3D software; instead, functional programs are converted to natural language expressions compiling textual *templates* embedded in the dataset and written in English.

The CLEVR dataset gives us way more control on the learning phase than other datasets present in literature. Information in each sample of the dataset is *complete* and *exclusive*. This means that no common-sense awareness is needed in order to correctly answer the questions. Answers can be given simply understanding the question and reasoning exclusively on the image, without needing external concepts.

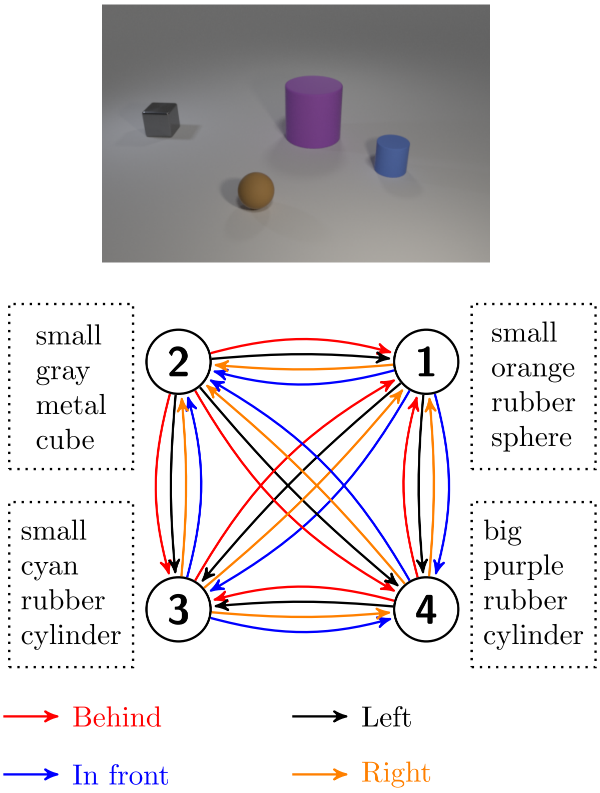


Fig. 1: CLEVR scene with associated scene graph.

3.2 Scene graphs

The best way to formally describe relations inside a scene is by making use of *scene graphs*, already available in CLEVR. More in details, a scene graph contains *nodes*, that account for objects occupying the scene, and *edges*, that describe relations occurring among them. Every node or edge can be assigned a set of attributes that fully describe them. CLEVR includes some specific objects attributes, namely the *color*, the *shape*, the *material* and the *size*, and accounts for the following spatial relationships: *to the left of*, *to the right of*, *in front of*, *behind*.

In Figure 1, we report an example image from CLEVR with the associated scene-graphs. Note that, although CLEVR graph is complete, half of the edges can be removed without losing information, since *to the right of* implies an opposite edge *to the left of* and *in front of* implies an opposite edge *behind*.

3.3 Ground-truth generation

We define a ground-truth for retrieving images with similar relations among objects relying on the similarity between scene graphs. Two scene graphs should be similar if they can depict almost the same relations between the same objects. However, evaluating the similarity between two graphs is not trivial; furthermore, it is often a subjective task, since there are aspects of the graph (e.g., the attributes associated to nodes) that weight differently depending on the specific application.

Although many solutions have been proposed in literature for defining distances between graph-structured data [3], concerning this particular use-case, we decide to employ the *graph edit-distance* (GED), that is an extension of the well-known edit-distance working on strings.

Differently from strings, edit operations on graphs include *delete*, *insert*, and *substitute* for both nodes and edges, for a total of 6 edit operations. The computation of the GED is faced as an optimization problem. Since the GED problem is known to be computationally hard, in this work we employ an approximated version of the GED algorithm. Computational times become easily unworkable on CLEVR scene graphs, even if removing the redundant *behind* and *left* edges. For this reason, we used an implementation based on [24], that is able to perform an efficient approximation of the algorithm. The approximated GED algorithm does not consider the entire span of solutions, but instead, it looks for a tiny subset of edit sequences, obtained by first matching similar nodes using linear assignment and then matching edges on the ruled node pairing.

The node-edge edit costs can be customized on the basis of their attributes. In particular, we use a cost of 1 for nodes-edges insertion/deletion and a cost of 1 if edges do not belong to the same kind of relation. A null cost is applied otherwise. Node substitution cost is driven by a policy that weights equally all attributes. Since in CLEVR there are 4 attributes per node, every attribute substitution costs 0.25.

To clarify GED algorithm functioning using our cost policy, we report an example in Figure 2. This instance of GED computation transforming the upper image into the below one returns a cost of 1.5.

In the light of this, given a query, we compute the ground-truth ranking of the dataset by sorting all scenes using computed GEDs between the scene graph of the query image and the graphs from all the others.

Given an image ranking produced by an arbitrary relationship-aware system, a rank correlation metric is computed against the ground-truth ranking. In this work, we use the *Spearman-Rho* correlation index, that is a common ranking similarity measure often employed in information retrieval scenarios [17].

4 Models

In this section, we describe our architectures tailored to explicit relationship-aware features learning. First of all, we review the basic formulation of the Relation Network (RN) for the sake of comparison with the newly introduced architecture. Then, we describe our proposals, namely the 2-stage Relation Network (2S-RN, previously introduced in [18]) and its extension — the novel AVF-RN architecture. Differently from 2S-RN, AVF-RN performs the aggregation of the visual features directly inside the network.

4.1 RN and 2S-RN overview

The Relation Network (RN) [25] approached the task of R-VQA and obtained remarkable results on the CLEVR dataset. RN modules combine input objects forming all possible pairs and applies a common transformation to them, producing activations aimed to store information about possible relationships among input objects. For the specific task of R-VQA, authors used a four-layer CNN to learn visual object representations, that are then fed to the RN module and combined with the textual embedding of the question produced by an LSTM, conditioning the relationship information on the textual modality. The core of the RN module is given by the following:

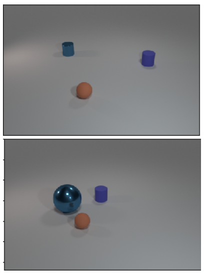
$$r = \sum_{i,j} g_\theta(o_i, o_j, q)\,, \tag{1}$$

where g_θ is a parametric function whose parameters θ can be learned during the training phase. Specifically, it is a multi-layer perceptron (MLP) network. o_i and o_j are the objects forming the pair under consideration, and q is the question embedding vector obtained from the LSTM module. The answer is then predicted by a downstream network f_ϕ followed by a softmax layer that outputs probabilities for every answer:

$$a = softmax(f_\phi(r))\,. \tag{2}$$

Relationship-aware features useful for R-CBIR should be extracted from a stage inside the network still not conditioned to the question. Hence, valid R-CBIR features can be extracted from the original RN module only at the output of the convolutional layer since, after that, questions condition entirely the remaining pipeline.

For this reason, the two-stage pipeline [18] was proposed in order to decouple visual relationships processing (*first-stage*) from the question elaboration (*second-stage*) so that the activations from a layer in the first stage can be employed as visual relationship-aware features. The 2S-RN considers all possible relationships



Steps	Cost
1. Substitute node **small-cyan-metal-cylinder** with **big-cyan-metal-sphere** (change 2 attributes)	0.5
2. Substitute edge **small-cyan-metal-cylinder behind small-blue-rubber-cylinder** with **big-cyan-metal-sphere in front of small-blue-rubber-cylinder**	1.0

Fig. 2: GED computation example.

between objects $g_\theta(o_i, o_j)$ in the image. The function g_θ is called *first-stage* of the RN. The output from this stage is a representation of the relationships between objects in the image not conditioned on the question. Then, the obtained relational representations $r_{i,j} = g_\theta(o_i, o_j)$ are combined with the query embedding q as follows:

$$r = \sum_{i,j} h_\psi(r_{i,j}, q) = \sum_{i,j} h_\psi(g_\theta(o_i, o_j), q)\,, \tag{3}$$

where h_ψ is the *second-stage* implemented as a multilayer perceptron network with parameters ψ. Using this solution, the 2S-RN constrains the network to learn relational concepts without considering the questions, at least during the first stage, before the $h_\psi(\cdot)$ function evaluation. Hence, the 2S-RN architecture enables relationship-aware features extraction from the output of any layer of the $g_\theta(\cdot)$ function.

4.1.1 Detailed Configuration

Both the RN and the 2S-RN architectures are trained on the R-VQA task on the CLEVR dataset.

Concerning the RN network, we use the very same setup described by the authors. In particular, the CNN is composed of 4 convolutional layers each with 24 kernels, ReLU non-linearities, and batch normalization; both g_θ and f_ϕ are composed by 256-dimensional fully-connected layers, with ReLU non-linearities after every layer, with four and two layers respectively. The final linear layer with 28 units produces logits for a softmax layer over the answers vocabulary; finally, the learning rate follows an exponential step increasing policy, that doubles it every 20 epochs, from 5e-6 up to 5e-4. Features are extracted directly at the end of the CNN and are aggregated using global average pooling.

2S-RN follows a very similar setup to the one of the original RN. Differently from the RN, g_θ and the novel h_ψ are both composed by 2 fully-connected layers. In this case, features are extracted at the end of the g_θ layer, immediately before the question concatenation. Detailed architectures are shown in Figures 3a and 3b.

Both RN and 2S-RN reaches very high performances when trained on CLEVR R-VQA: they obtain 93,6% and 93,8% accuracy on the test set respectively.

4.2 Aggregated Visual Features Relation Network (AVF-RN)

The 2S-RN approach is able to extract the relational content from the visual pipeline before it is conditioned by the question embedding. Nevertheless, features extracted from the 2S-RN are still not aggregated and contain all the descriptions from every couple of objects. Hence, standard 2S-RN features are aggregated only during the extraction process by simply averaging them iterating through all the couples.

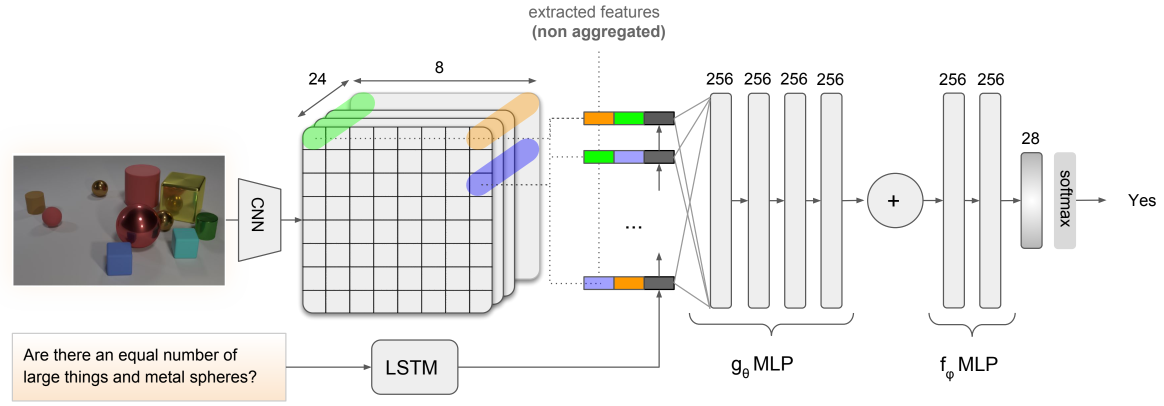
Our contribution consists in learning the feature directly inside the network. To this aim, we slightly changed the 2S-RN architecture in order to aggregate all the object couples before inserting the question embedding into the pipeline. Hence, AVF-RN network can be described by the following equation:

$$r = q, h_\psi \sum_{i,j} r_{i,j} = q, h_\psi \sum_{i,j} g_\theta(o_i, o_j)\,, \tag{4}$$

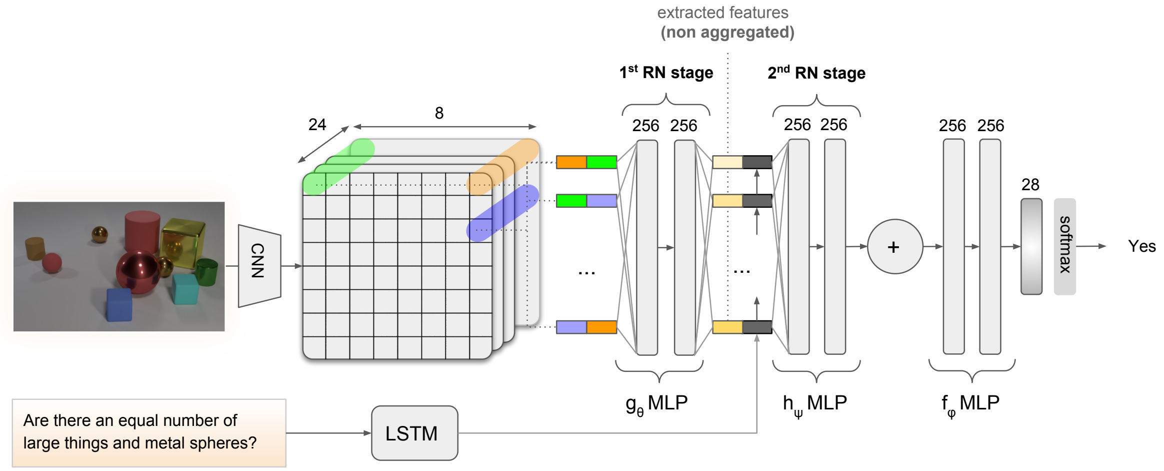
with the same naming conventions used for 2S-RN. However, differently from 2S-RN, h_ψ is not evaluated for every couple; instead, it is evaluated once, on the already aggregated visual features. For this reason, the h_ψ role changes with respect to the 2S-RN case. In AVF-RN the purpose of h_ψ is to process the already aggregated visual feature, while in 2S-RN it processes textual and visual features from every couple of objects.

The architecture has been designed so that each function g_θ, h_ψ and f_ϕ can be customized with any number of fully-connected layers with any number of neurons each. More in details:

- g_θ comprises the n layers before the aggregation operation;
- h_ψ comprises the m layers between the aggregation and the question insertion;



(a) Relation Netowrk (RN) architecture.



(b) Two-stage Relation Network (2S-RN) architecture.

Fig. 3: Detailed RN and 2S-RN architectures with layers configuration.

- f_ϕ comprises the k layers after question insertion; they are aimed at processing the joint visual aggregated features and the textual ones to obtain the information needed to predict the answer.

The overall architecture is reported in Figure 4.

4.2.1 Detailed configuration and hyper-parameters tuning

In the case of RN and 2S-RN, the concatenation of the question with all the couples works as a simple but quite effective attention mechanism. The novel AVF-RN model, instead, introduces the question embedding after the aggregation. We gain in feature relational expressiveness but, on the other hand, the attention effect is lost. For this reason, we obtain an overall less accuracy with respect to the RN and the 2S-RN architectures. There are several hyper-parameters that should be tuned and an extensive search is not feasible. Among the hyper-parameters, the most important ones are the number of fully-connected layers for every function g_θ, h_ψ, and f_ϕ, namely n, m, k, and the output size for all of these layers. We try to stick, wherever possible, to successful configurations observed when training the RN and the 2S-RN architectures. In Table 1 we collect some of the hyper-parameters experimentation we performed on this architecture, together with the reached accuracy on the CLEVR R-VQA task.

The best result is obtained using weighted-sum as aggregation, with weights learned during training, one layer of h_ψ and three layers of f_ϕ. The aggregation is positioned after the 4th fully connected layer of g_θ, while the question is inserted after a single fully-connected layer of h_ψ.

The 4th layer of g_θ is larger in order to augment the expressiveness of the aggregated feature. In order to speed up convergence, we initialize the weights for the CNN and the first two fully-connected layers of g_θ with the weights coming from the respective layers of the 2S-RN architecture (they are the only ones to maintain

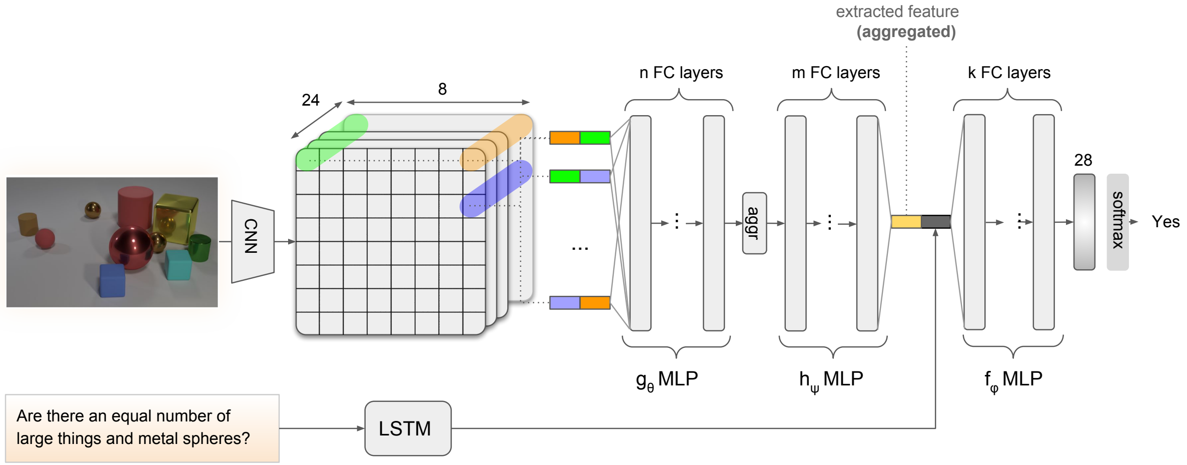


Fig. 4: AVF-RN architecture overview. The number of fully-connected layers is fully customizable, as well as the aggregation function.

Table 1: The accuracy values of different fully-connected layer configurations for every function g_θ, h_ψ and f_ϕ. Each configuration includes the output size for every fully-connected layer.

g_θ config.	h_ψ config.	f_ϕ config.	Aggr. Type	Accuracy(%)
256, 256, 512	256	256, 256	sum	53.8
256, 256, 256, 512	256	256, 256	sum	53.2
256, 256, 256, 256	256	256 ,256, 256	sum	54.0
256, 256, 256, 512	256	256, 256, 256	sum	54.2
256, 256, 256, 1024	-	512 1024	weighted-sum	55.7
256, 256, 256, 512	**256**	**256, 256, 256**	**weighted-sum**	**64.5**

the same role and the same interface with respect to the AVF-RN).

Even if the reached accuracy is quite far from the performance reached by the RN and the 2S-RN architectures, this result is enough for learning relationship-aware visual features.

5 Experimental Setup

In this section, we compare all the different architectures explained in Section 4 on the task of R-CBIR. We use a standard CBIR metric for comparing our results, namely the Spearman-Rho metric. As a baseline, we choose the ranking obtained with one of the state-of-the-art non-relational image descriptors for image instance retrieval, namely the RMAC descriptor [26].

Also, as a non-relational baseline, we train a simple architecture on a multi-classification task, where the objective consists in correctly classifying all the objects inside every CLEVR scene. This simple architecture consists of the CNN already used in the original RN architecture and 2 fully-connected layers with ReLU non-linearities for use as multi-label classifier. Similarly to the basic RN architecture, features are extracted by average-pooling the CNN activations. We call this architecture *Multi-label CNN*.

All the architectures are trained on the clevr train-set; however, features are always extracted on the test-set in order to evaluate the generalization capabilities of the system. All the architectures are trained on an RTX 2080Ti, with a batch size of 640. During experiments, we observed that the training time was almost the same for all the RN-derived architectures. We trained for about 300 epochs. We then picked the model having the highest validation accuracy among all the training epochs.

The average training speed was about 25 minutes per epoch. Instead, extracting all the features from the whole test set required only about 1 minute. Questions are not needed at extraction time, so the entire architecture is considerably lighter.

We use three different setups for evaluating the results:

1. **CLEVR-Full** - We use the entire CLEVR test set. Any image can be selected as query and any image could be eligible for being retrieved.

Table 2: Spearman-Rho correlation index for existing methods and our novel AVF-RN features. We report the 95% confidence intervals for the mean over 500 queries.

	CLEVR Full	CLEVR Filtered Queries	CLEVR Subset
RMAC [5]	−0.15±0.02	0.02±0.02	0.09±0.01
Multi-label CNN	0.05±0.05	0.64±0.04	0.18±0.04
RN [25]	0.04±0.05	0.64±0.03	0.20±0.03
2S-RN [18]	0.15±0.04	0.65±0.02	0.26±0.02
AVF-RN (ours)	**0.28±0.04**	**0.72±0.02**	**0.34±0.02**

2. **CLEVR-Filtered-Queries** - We select as queries only the images containing at most N objects, while any image remains eligible for being retrieved.
3. **CLEVR-Subset** - We filter the entire CLEVR test set with images containing at most N objects. Hence, both queries and retrieved images contain at most N objects.

CLEVR-Full is the same scenario used for evaluating 2S-RN performances in [18]. However, the approximated GED algorithm we employ presents some notable differences with the exact version when graphs have a large number of nodes. For this reason, during experimentation, we explore also the simpler scenarios *CLEVR-Filtered-Queries* and *CLEVR-Subset.*

CLEVR comes with rendered images containing no more than 10 objects. In our experiments we set N equal to 5.

Table 2 reports values of Spearman-Rho correlation index for all the experiments on all the three versions of the CLEVR datasets. Spearman-Rho correlations are relative to the ground-truth generated as explained in 3.3 and obtained by ranking images using the approximated version of the GED algorithm. The Spearman-Rho correlation index is evaluated over multiple rankings, generated using 500 query images, in order to produce statistically meaningful results.

5.1 Discussion

The new AVF-RN features reaches the state-of-the-art on the R-CBIR task, defeating both non-relational baseline methods and the RN and 2S-RN relationship-aware techniques. It is worth noting the almost zero performance gap between convolutional features extracted from the RN and the multi-label CNN networks. The results tell us that the simple global average pooling of the last feature maps of the CNN is not able to catch significant relational content, even in the case of a downstream RN network.

On the *CLEVR-Full* scenario, our AVF-RN features obtain an almost doubled Spearman-Rho value with respect to the 2S-RN one. This suggests that the novel AVF-RN architecture is able to correctly order complex relevant scenes in terms of their relational content. However, due to the approximation introduced by *ApproxGED* in case of large number of objects, it is difficult to strongly confirm this claim in this scenario.

On the other hand, in the *CLEVR-Filtered-Queries* scenario, the images with few objects are privileged by the ground-truth. Hence, standard approaches like RMAC or simple CNN features behave quite well since they can exploit their capability of retrieving images having a similar number of objects with respect to the query. Besides counting, they are in any case unable to catch intrinsic inter-object relationships. Instead, these details are well captured by AVF-RN and 2S-RN features. However, the aggregation learned inside the network in AVF-RN obliges the layers after the aggregation to learn compact and smart scene descriptions. Consequently, AVF-RN captures more detailed scene-information with respect to the simple posterior aggregation performed for the 2S-RN feature.

Similarly, in the *CLEVR-Subset* scenario, all the retrieved images are forced to contain a small number of objects, hence the basic recognition abilities by CNN features do not capture the finest relational details. In this case, since all the images contain few objects, the only way to obtain remarkable results is by understanding the intrinsic relational content of the scene. This explains why there is a great improvement of AVF-RN features over standard methods.

Even if it is quite difficult to give an objective evaluation of the proposed methodology by only looking at the first 10 most relevant images, visual evaluation reported in Figure 5 is useful for giving a qualitative feedback and an intuition beyond statistics. We collect these visual results from the challenging *Full CLEVR* experiment. In particular, we can see that RMAC features always try to find the very same objects as the query, in any position inside the image. Similarly, multi-label CNN features seem very noisy.

It appears that 2S-RN and AVF-RN, instead, are interpreting the scene from an high-level perspective by finding all the images having a big object (better

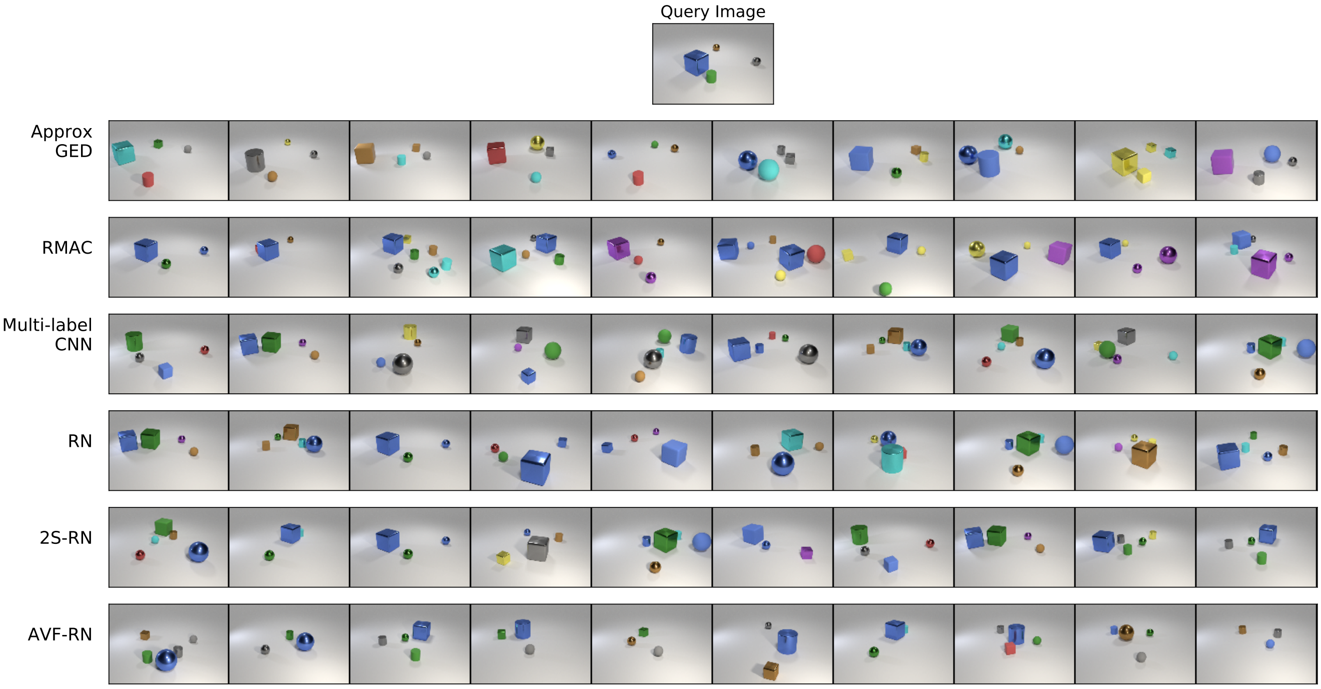


Fig. 5: Most relevant images for the proposed query from Full CLEVR experiment, using both non-relational approaches (RMAC, Multi-Label CNN) and relational ones (RN, 2S-RN, AVF-RN). The first row belongs to the ground-truth generated as explained in section 3.3.

if a metallic blue cube) surrounded by other smaller objects.

On our website `rcbir.org` you can find an interactive browsing system for exploring the R-CBIR results from the proposed methods for different query images.

5.2 Success/Failure Analysis

In Figure 6 we report simple cases of success and failure of the top-performing method AVF-RN against the two baselines RMAC and multi-label CNN. We assume the result as successful if our AVF-RN features can retrieve more ground truth images with respect to the baselines; otherwise, the experiment is considered failed for the examined query. For the sake of simplicity, we analyze only the top 10 results.

It can be noticed that successful retrieved images (Figures 6a and 6b) are well approximating the ground-truth scene graphs. This is because AVF-RN features exhibit some scene-wide image understanding that is not tailored to the features of single objects. On the other hand, RMAC features are quite good at catching the key visual features of the single objects, such as their size, but they have troubles to focus the attention on the global scene arrangement.

Failure cases (Figures 6c and 6d) demonstrate that AVF-RN features cannot always catch the relational content of the scene. In particular, in the failure example of Figure 6c, the AVF-RN features seem to be always triggered by a yellow object, that perhaps is a not so important characteristic if considering the whole scene arrangement.

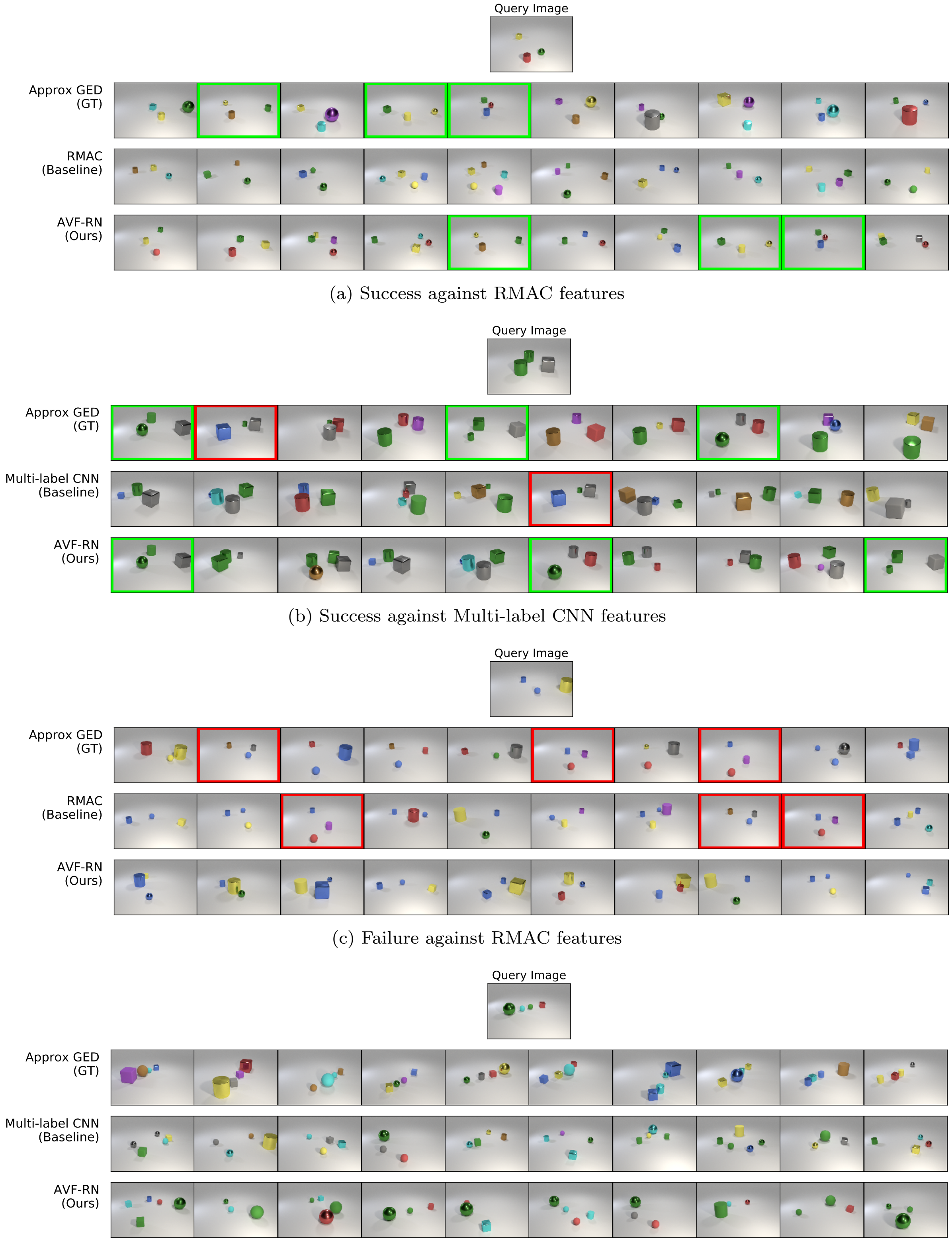
Instead, Figure 6d demonstrates that is difficult to catch objects arranged in precise configurations (in this case, placed on the same line). In this example, both the multi-label CNN baseline and our AVF-RN features fail.

6 Conclusions

State-of-the-art methods for relational reasoning evaluate their capabilities on some challenging tasks such as R-VQA (Relational Content-Based Image Retrieval) and VRD (Visual Relationships Detection).

In this work, we defined the sub-task of R-CBIR in which retrieved images should be similar to the query in terms of relationships among objects. This was motivated by the fact that current image retrieval systems, performing traditional CBIR, are not able to infer relations among the query and the retrieved images.

Given the novelty of the proposed task, we had to generate a relational benchmark. To this aim, we employed CLEVR, a synthetic and unbiased dataset originally developed for the task of R-VQA. In particular, we



(a) Success against RMAC features

(b) Success against Multi-label CNN features

(c) Failure against RMAC features

(d) Failure against Multi-label CNN features

Fig. 6: Success (a)(b) and failure (c)(d) cases for AVF-RN compared to the baselines, RMAC (a)(c) and Multi-Label CNN (b)(d). Matches among GT and AVF-RN are marked in green, while matches among GT and the baselines in red.

compared scene graphs using a graph distance metric called Graph Edit Distance (GED), in order to define a relational-aware concept of distance between CLEVR scenes.

Distance evaluation among graphs, however, presented some degrees of freedom. In fact, the employed GED distance must be initialized with some cost parameters. Costs have been set to values that were not able to advantage any object attribute over the others, in order to produce the fairest configuration.

We described the 2S-RN approach and, afterwards, we proposed an extension to the 2S-RN module, called Aggregated Visual Features Relation Network (AVF-RN). This modification aims at aggregating the visual features directly inside the network. We proved that features from our AVF-RN are able to encode in a compact representation the relationships between objects in the image, outperforming some baseline non-relational methods as well as the 2S-RN relational features.

Although the AVF-RN system lacks the native attention mechanism that both RN and 2S-RN use when they concatenate the question with all the objects couples, this method can successfully learn compact relational features.

We noticed that, despite the encouraging performances measured with the introduced metrics, our approach generates results of difficult interpretation when images have a high number of objects. This is probably due to the fact that having many objects implies too many relationships that are difficult to track by the human eye. Also, the proposed architectures must be trained on VQA datasets, since the relationships between objects in the image are learned by answering questions. In this regard, the need for a VQA training dataset is overall a strong constraint that should be relaxed in future works.

Next steps in this ongoing research include the possibility of learning features by training architectures directly on the R-CBIR task, by using metric learning approaches such as siamese-learning methods. Also, it would be interesting studying how the performance of the models changes when using real-world datasets such as Visual Genome [11] or Open Images [12].

Acknowledgments

This work was partially supported by Automatic Data and documents Analysis to enhance human-based processes (ADA), CUP CIPE D55F17000290009, and by the AI4EU project, funded by the EC (H2020 - Contract n. 825619). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. CoRR **abs/1505.00468** (2015). URL `http://arxiv.org/abs/1505.00468`
2. Belilovsky, E., Blaschko, M.B., Kiros, J.R., Urtasun, R., Zemel, R.: Joint embeddings of scene graphs and images. ICLR (2017)
3. Cai, H., Zheng, V.W., Chang, K.C.: A comprehensive survey of graph embedding: Problems, techniques and applications. CoRR **abs/1709.07604** (2017). URL `http://arxiv.org/abs/1709.07604`
4. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3298–3308. IEEE (2017)
5. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. arXiv preprint arXiv:1610.07940 (2016)
6. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
7. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning (2017)
8. Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Inferring and executing programs for visual reasoning. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
9. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3668–3678 (2015)
10. Kahou, S.E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., Bengio, Y.: Figureqa: An annotated figure dataset for visual reasoning. CoRR **abs/1710.07300** (2017). URL `http://arxiv.org/abs/1710.07300`
11. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016). URL `https://arxiv.org/abs/1602.07332`
12. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., Ferrari, V.: The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. CoRR **abs/1811.00982** (2018). URL `http://arxiv.org/abs/1811.00982`
13. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (2016)
14. Lu, P., Ji, L., Zhang, W., Duan, N., Zhou, M., Wang, J.: R-vqa: Learning visual relation facts with semantic attention for visual question answering. In: SIGKDD 2018 (2018)
15. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (eds.) Advances in Neural Information Processing Systems 27, pp. 1682–1690. Curran Associates, Inc. (2014)

16. Mascharka, D., Tran, P., Soklaski, R., Majumdar, A.: Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
17. Melucci, M.: On rank correlation in information retrieval evaluation. SIGIR Forum **41**(1), 18–33 (2007). DOI 10.1145/1273221.1273223
18. Messina, N., Amato, G., Carrara, F., Falchi, F., Gennaro, C.: Learning relationship-aware visual features. In: L. Leal-Taixé, S. Roth (eds.) Computer Vision – ECCV 2018 Workshops, pp. 486–501. Springer International Publishing, Cham (2019)
19. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: ICCV 2017- International Conference on Computer Vision 2017. Venice, Italy (2017). URL `https://hal.archives-ouvertes.fr/hal-01576035`
20. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. CoRR **abs/1811.10696** (2018). URL `http://arxiv.org/abs/1811.10696`
21. Raposo, D., Santoro, A., Barrett, D.G.T., Pascanu, R., Lillicrap, T.P., Battaglia, P.W.: Discovering objects and their relations from entangled scene representations. CoRR **abs/1702.05068** (2017). URL `http://arxiv.org/abs/1702.05068`
22. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (eds.) Advances in Neural Information Processing Systems 28, pp. 2953–2961. Curran Associates, Inc. (2015)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (eds.) Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
24. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. Image and Vision Computing **27**(7), 950 – 959 (2009). DOI https://doi.org/10.1016/j.imavis.2008.04.004. 7th IAPR-TC15 Workshop on Graph-based Representations (GbR 2007)
25. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 4967–4976. Curran Associates, Inc. (2017)
26. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
27. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph R-CNN for scene graph generation. CoRR **abs/1808.00191** (2018). URL `http://arxiv.org/abs/1808.00191`
28. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. CoRR **abs/1511.02274** (2015). URL `http://arxiv.org/abs/1511.02274`
29. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. CoRR **abs/1809.07041** (2018). URL `http://arxiv.org/abs/1809.07041`
30. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A.M., Elhoseiny, M.: Large-scale visual relationship understanding. CoRR **abs/1804.10660** (2018). URL `http://arxiv.org/abs/1804.10660`
31. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. CoRR **abs/1512.02167** (2015). URL `http://arxiv.org/abs/1512.02167`