# Transformer Reasoning Network for Image-Text Matching and Retrieval

Nicola Messina, Fabrizio Falchi, Andrea Esuli, Giuseppe Amato

*Institute of Information Science and Technologies*

*National Research Council*

Pisa, Italy

{name.surname}@isti.cnr.it

*Abstract*—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 interplaying 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.

## I. INTRODUCTION

Recent advances in deep learning research brought to life interesting tasks and applications which include joint processing of data from different domains. Image-text matching is an interesting task that consists in aligning information coming from visual and textual worlds, in order to benefit of the complementary richness of these two very different domains.

Visuals and texts are two important modalities used by humans to fully understand the real world. While text is already a well-structured description developed by humans in hundreds of years, images are basically nothing but raw matrices of pixels hiding very high-level concepts and structures. If we want to obtain an informative textual description of a visual scene we are required not only to understand what are the salient entities in the image, but we need also to reason about the relationships between the different entities, e.g. "The kid *kicks* the ball". In this respect, it is necessary not only to perceive objects on their own but also understanding spatial and even abstract relationships linking them together.
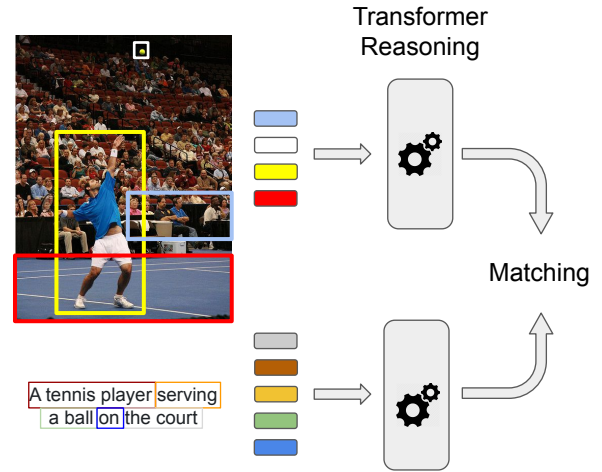


Fig. 1. Overview of the presented architecture. Image and text are seen respectively as sets of image regions and sequences of words, and they are processed using a transformer-based reasoning engine.

This has important implications in many modern AI-powered systems, where perception and reasoning play both important roles. In this work, we concentrate our effort on the cross-modal information retrieval research field, in which we are asked to produce compact yet very informative object descriptions coming from very different domains (visual and textual in this scenario).

Vision and language matching has been extensively studied [1]–[5]. Many works employ standard architectures for processing images and text, such as CNNs-based models for image processing and recurrent networks for language. Usually, in this scenario, the image embeddings are extracted from standard image classification networks, such as ResNet or VGG, by employing the network activations before the classification head. Usually, descriptions extracted from CNN networks trained on classification tasks are able to only capture global summarized features of the image, ignoring important localized details.

Although these networks demonstrated noticeable performances in the image-text matching task, they are not able to infer what an *object* really is. The *objectness* prior is an important feature of the perception system that helps filtering

out irrelevant zones in the images, while focusing the attention on entities of interest. As far as the matching problem is concerned, finding entities of interest inside the image helps in creating a representation that has a level of abstraction comparable with the related text. In fact, a visual object present in an image, such as a dog, can be matched in an almost one-to-one relationship with the nouns *dog*, or *animal* present in the corresponding image caption. Furthermore, the *objectness* prior is the first step towards higher-level abstraction tasks such as reasoning about inter-object relationships.

We are to tackle this important problem with the goal of finding compact cross-modal descriptions of images and texts which can incorporate detailed relational insights of the scene. Compact and informative descriptions are required in the context of large scale retrieval systems, where image and text embeddings can be compared and indexed using a simple similarity function (e.g., cosine similarity) defined on a common embedding space.

Some works have recently tackled the matching problem using a relational approach, trying to reason on substructures of images and texts (regions and words respectively) using attention and self-attention mechanisms [3], [5], [6], or graph networks [7].

In particular [3], [4], [6] try to learn a scoring function $s = \phi(I, C)$ measuring the affinity between an image and a caption, where $I$ is an image, $C$ is the caption and $s$ is a normalized score in the range $[0, 1]$. The problem with this approach is that it is not possible to extract compact features describing images and texts separately. In this setup, if we want to retrieve images related to a given query text, we have to compute all the similarities by means of the $\phi$ function, and then sort the resulting scores in descending order. This is unfeasible if we want to retrieve images from a large database in few milliseconds.

For this reason, we propose the Transformer Encoder Reasoning Network (TERN), a transformer-based architecture able to map images and texts into the same common space while preserving important relational aspects of both modalities. In doing so, we avoid cross-talking between the two pipelines, so that it remains possible to separately forward the visual and the language pipeline to obtain compact image/caption descriptors.

The general transformer architecture [8] was introduced to process sequential data, like natural languages. However, the encoder part of the transformer has no sequential prior hard-coded in its architecture. Therefore, it is a good candidate for processing also image regions: with the very desirable self-attention mechanism it incorporates, the transformer encoder can be employed to link together different image regions, effectively constructing a powerful visual reasoning pipeline.

Concerning the evaluation of the proposed matching procedure in an information retrieval setup, the *Recall@*K metric is usually employed, where typically $K = \{1, 5, 10\}$. However, in common search engines where the user is searching for related images and not necessarily exact matches, the *Recall@*K evaluation could be too rigid, especially when $K = 1$.

For this reason, we propose to measure the retrieval abilities of the system through a discounted cumulative gain metric with relevance computed exploiting caption similarities, proceeding in a similar way to [2].

The contributions of this paper are:

- We introduce the Transformer Encoder Reasoning Network (TERN), a transformer-based architecture able to map both visual and textual modalities into the same common space, preserving the relational content of both modalities. The learned representations can be used for efficient and scalable multi-modal retrieval.
- We introduce a novel evaluation metric able to capture non-exact search results, by weighting different results through a relevance measure computed on the caption similarities.
- We show that our architecture reaches state-of-the-art performances with respect to other architectures on the introduced metric, for the image retrieval task.

## II. RELATED WORK

In this section, we review some of the previous work related to image-text matching and high-level relational reasoning. Also, we briefly summarize the evaluation metrics available in the literature for the image-caption retrieval task.

### Image-Text matching

Image-text matching is often cast to the problem of inferring a similarity score among an image and a sentence. Usually, one of the common approaches for computing this cross-domain similarity is to project images and texts into a common representation space on which some kind of similarity measure can be defined (e.g.: cosine or dot-product similarities).

Images and sentences are preprocessed by specialized architectures before being merged together at some point in the pipeline.

Concerning image processing, the standard approach consists in using Convolutional Neural Networks (CNNs), usually pretrained on image classification tasks. In particular, [9]–[13] used VGGs, [1], [14]–[16] used ResNets. The problem with these kinds of CNNs is that they usually are able to extract extremely summarized and global descriptions of images. Therefore, a lot of useful fine-grained information needed to reconstruct inter-object relationships for precise image-text alignment is permanently lost.

For this reason, recent works exploited the availability of precomputed region-level features extracted from state-of-the-art object detectors. In particular, following the work by [17], [5], [7] used bottom-up features extracted from Faster-RCNN. The bottom-up attention mechanism resembles the attentive mechanism present in the human visual system, and it is an important feature for filtering out unimportant information. This lays the foundations for a more precise and lightweight reasoning mechanism, downstream of the bottom-up perception module, which should process the resulting image regions in a careful way to obtain an expressive representation of the overall scene.

Concerning sentence processing, many works [1], [4], [5], [7], [16] employ GRU or LSTM recurrent networks to process natural language.

Recently, the transformer architecture [8] achieved state-of-the-art results in many natural language processing tasks, such as next sentence prediction or sentence classification. In particular, the BERT embeddings [18] emphasized the power of the attention mechanism to produce accurate context-aware word descriptions.

Given the enormous flexibility of the transformer encoder architecture, some works [3], [6] tried to apply the attention mechanism of the transformer encoder architecture to process visual inputs and natural language together. The main idea behind visual processing using the transformer encoder is to leverage its self-attention mechanism to link together different image regions in order to catch important inter-object relationships. This is possible since this model is perfectly agnostic on the nature of the vectors given as input, and has no built-in sequential priors.

These latest works were able to achieve state-of-the-art results in caption/image retrieval. However, they cannot separately produce image and caption embeddings; this is a mandatory requirement to produce features that are actually usable in real-world search engines. They model a function $s = \phi(I, C)$ that measures the affinity between an image and a caption, where $I$ is an image, $C$ is the caption and $s$ is a normalized score in the range $[0, 1]$. Following this path, an exhaustive sequential search is needed to rank all the images given a query caption or vice-versa.

Instead, we are interested in employing two different mapping functions, $I_v = \psi_v(I)$ and $I_c = \psi_t(C)$ which separately project the two modalities into the same common space, without preconditioning one of the modalities to the other.

The most important work that successfully explored this approach is [7]. The authors were able to achieve very good results in caption/image retrieval. They introduced a visual reasoning pipeline built of a Graph Convolution Networks (GCNs) and a GRU to sequentially reason on the different image regions. Furthermore, they impose a sentence reconstruction loss to regularize the training process.

Differently from their work, we leverage on the reasoning power of the transformer encoder, both for the visual and linguistic pipelines.

*High-level reasoning*

Another branch of research is focused on the study of relational reasoning models for high-level understanding. [19] proposed an architecture that separates perception from reasoning. They tackle the problem of Visual Question Answering by introducing 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. Recently, [20], [21] extended the RN for producing compact features for relation-aware image retrieval. However, they did not explore the multi-modal retrieval setup.

Other solutions try to stick more to a symbolic-like way of reasoning. [22], [23] 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.

Recent works employed Graph Convolution Networks (GCNs) to reason about the interconnections between concepts. In particular, [24]–[26] used GCNs to reason on the image regions for image captioning, while [27], [28] used GCN with attention mechanisms to produce the scene graph from plain images.

*Retrieval evaluation metrics*

All the works involved with image-caption matching evaluate their results by measuring how good the system is at retrieving relevant images given a query caption (image-retrieval) and vice-versa (caption-retrieval). In other words, they evaluate their proposed models using a retrieval setup.

Usually the *Recall@*K metric is used [1], [3], [6], [7], [29], where typically $K = \{1, 5, 10\}$. On the other hand, [2] introduced a novel metric able to capture non-exact results by weighting the ranked documents using a caption-based similarity measure.

We embrace the same idea of [2], and we extend it bringing to life an alternative yet powerful evaluation metric. Relaxing the constraints of exact-match similarity search is an important step towards an effective evaluation of real search engines.

## III. REVIEW OF TRANSFORMER ENCODERS (TEs)

Our proposed architecture is based on the well established Transformer Encoder (TE) architecture. Following, we review some of the major strengths of this architecture.

The Transformer Encoder (TE) architecture relies heavily on the concept of self-attention. The basic attention mechanism, as described in [8], is built upon three quantities: queries, keys, and values. The attention mechanism maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed using a softmax activation function over the inner product of the query with the corresponding key. More formally,

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

Where $Q \in \mathbb{R}^{t \times d_k}, K \in \mathbb{R}^{s \times d_k}$ and $V \in \mathbb{R}^{s \times d_v}$; $s$ is the input sequence length and $t$ is the length of the conditioning sequence that drives the attention. The factor $\sqrt{d_k}$ is used to mitigate the vanishing gradient problem of the softmax function in case the inner product assumes too large values.

The self-attention derives trivially from the general attention mechanism when either V, K, and Q are computed from the same input set, i.e., when the set that we use to drive the attention is the same as the input set. In this case, in fact, $t = s$ and the scalar product $QK^T \in \mathbb{R}^{s \times s}$ is a square matrix

that encodes the affinity that each element of the set has with all the others elements of the same set.

In the self-attention case $Q, K, V$ are computed by linear projecting the same input embeddings using three different matrices $W^{d_k \times d_i}, W^{d_k \times d_i}$ and $W^{d_v \times d_i}$, where $d_i$ is the dimensionality of the input embeddings.

The TE output is computed by further processing the vectors produced by the self-attention mechanism. A simple feedforward layer on the $Att(Q, K, V)$ vectors, with a ReLU activation function, is used for this purpose. This simple feedforward layer casts in output a set of features having the same dimensionality of the input set.

We argue that the transformer encoder self-attention mechanism is able to drive a simple but powerful reasoning mechanism able to spot hidden links between the vector entities given in input to the encoder, whatever nature they have. Also, the encoder is designed in a way that multiple instances of the same architecture could be stacked in sequence, producing a deeper reasoning pipeline.

## IV. VISUAL-TEXTUAL REASONING USING TRANSFORMER ENCODERS

Our work relies almost entirely on the TE architecture, both for the visual and the textual data pipelines.

The TE takes as input sequences or sets of entities, and it is able to reason upon these entities disregarding their intrinsic nature. In particular, we consider the salient regions in an image as visual entities, and the words present in the caption as language entities.

More formally, the input to our reasoning pipeline is a set $I = \{r_0, r_1, ...r_n\}$ of $n$ image regions representing an image $I$ and a sequence $C = \{w_0, w_1, ...w_m\}$ of $m$ words representing the corresponding caption $C$. Following, we will describe the methodology we adopted to extract $r_i$ from images and $w_j$ from captions.

### Region and Word Features

$I$ and $C$ descriptions come from state-of-the-art visual and textual pretrained networks, Faster-RCNN with Bottom-Up attention and BERT respectively.

Faster-RCNN [30] is a state-of-the-art object detector. It has been used in many downstream tasks requiring salient object regions extracted from images. Therefore, Faster-RCNN is one of the main architectures implementing the human-like visual perception.

[31] introduced bottom-up visual features by training Faster-RCNN with a Resnet-101 backbone on the Visual Genome dataset [32]. Using these features, they were able to reach remarkable results on the two downstream tasks of image captioning and visual question answering.

Therefore, in our work we employ the bottom-up features extracted from every image as image description $I = \{r_0, r_1, ...r_n\}$.

Concerning text processing, we used BERT [18] for extracting word embeddings. BERT already uses a multi-layer transformer encoder to process words in sentences and capture

their functional relationships through the same powerful self-attention mechanism. BERT embeddings are trained on some general natural language processing tasks such as sentence prediction or sentence classification and demonstrated state-of-the-art results in many downstream natural language tasks. BERT embeddings, unlike word2vec [33], capture the context in which each word appears. Therefore, every word embedding carries information about the surrounding context, that could be different from caption to caption. Since the transformer encoder architecture does not embed any sequential prior in its architecture, words are given a sequential order by mixing some positional information into the learned input embeddings. For this reason, the authors in [8] add sine and cosine functions of different frequencies to the input embeddings. This is a simple but effective way to transform a set into a sequence.

### Transformer Encoder Reasoning Network (TERN)

Our reasoning engine is built using a stack of transformer encoder layers. The same reasoning architecture is applied to both the textual and visual pipelines.

The reasoning module continuously operates on sets and sequences of $n$ and $m$ objects respectively for images and captions.

In the end, we need to produce a compact representation of the $n$ processed regions and of the $m$ processed words suitable for the downstream task of image-text matching in a common space with fixed dimensionality. One of the easiest ways to proceed is to pool the elements of the set/sequence using symmetric functions like *sum* or *avg*; on the other hand, [7] try to grow a meaningful aggregated representation inside the hidden state of a recurrent network such as a GRU or an LSTM.

Our method, instead, follows the approach by BERT [18]: we reserve a special token both at the beginning of the regions set and of the words sequence (I-CLS and T-CLS in Figure 2) devoted to carrying global information along the two pipelines. For this reason, we effectively expand the number of image regions to $n + 1$ and the number of words to $m + 1$, with $r_0$ and $w_0$ reserved for this purpose.

Initially, $w_0$ is set to the T-CLS BERT token, while $r_0$, i.e., I-CLS, is a zero vector. At every reasoning step, this information is updated in an attentive manner by the self-attention mechanism of the TEs. In the end, our final image and caption features will be $r_0$ and $w_0$ in output from the last transformer encoder layer.

In the last layers of the TERN architecture, the abstracted representations of the visual and textual pipelines should be comparable. In order to enforce this constraint, we share the weights of the last layers of the TEs before computing the matching loss $L_m$ on the common space.

The overall architecture is shown in Figure 2. If we use only bottom-up features without any spatially related information, the visual reasoning engine is not able to reason about spatial relationships. This is a fairly important aspect to capture since
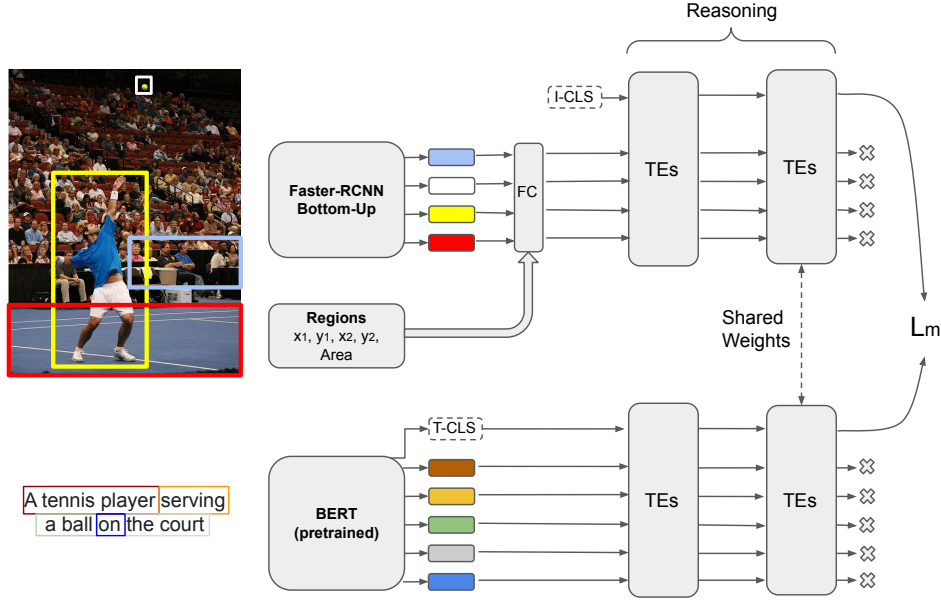
Fig. 2. The proposed TERN architecture. TE stands for Transformer Encoder and its architecture is explained in detail in [8]. Region and words are extracted through a bottom-up attention model based on Faster-RCNN and BERT respectively. BERT already employs positional encoding for representing the sequential nature of words, therefore this step is not reported in the figure. Concerning regions, the extracted bottom-up features are conditioned with the information related to the geometry of the bounding-boxes. This is done thought a simple fully connected stack in the early visual pipeline, before the reasoning steps. $L_m$ is the matching loss, defined as in [1]. The final weight sharing between TE modules guarantees consistent processing of the high-level concepts.

lot of textual descriptions contain spatial indications (e.g. *on top of* or *above*).

In order to include spatial awareness also in the visual reasoning process, we condition the early visual pipeline with the bounding-boxes coordinates. To this aim, we compute the normalized coordinates and the normalized area for each region:

$$c = \left\{ \frac{x_1}{W}, \frac{y_1}{W}, \frac{x_2}{H}, \frac{y_2}{H}, \frac{(x_2 - x_1)(y_2 - y_1)}{WH} \right\} \quad (2)$$

then, we concatenate $c$ with the original bottom-up feature. In the end, we forward this information through a simple Linear-ReLU-Linear stack to obtain the final spatial-aware bottom-up feature.

*Learning*

In order to match images and captions in the same common space, we use a hinge-based triplet ranking loss, focusing the attention on hard negatives, as in [1], [7]. Therefore, we used the following loss function:

$$L_m(i, c) = \max_{c'}[\alpha + S(i, c') - S(i, c)]_+ + \\ \max_{i'}[\alpha + S(i', c) - S(i, c)]_+ \quad (3)$$

where $[x]_+ \equiv max(0, x)$. The hard negatives $i'$ and $c'$ are computed as follows:

$$i' = arg \max_{j \neq i} S(j, c) \\ c' = arg \max_{d \neq c} S(i, d) \quad (4)$$

where $(i, c)$ is a positive pair. $S(i, j)$ is the similarity function between image and caption features. We used the standard cosine similarity as $S(\cdot, \cdot)$. As in [1], the hard negatives are sampled from the mini-batch and not globally, for performance reasons.

## V. EVALUATION METRIC FOR NON-EXACT MATCHING

Many works measure the retrieval abilities of their visual-linguistic matching system by employing the well known *Recall@*K metric. *Recall@*K measures the percentage of queries able to retrieve the correct item among the first k results.

However, in common search engines where the user is searching for related images/captions and not necessarily exact matches, the *Recall@*K evaluation could be too rigid, especially when $K$ is small. In fact, in the scenarios where $K = \{1, 5, 10\}$, we are measuring the ability of the system to retrieve exact results at the top of the ranked list of images/captions. Doing so, we are completely ignoring other relevant but not exact-matching elements retrieved in the first positions. These elements still contribute to good user experience in the context of search engines. The *Recall@*K metric is fully unable to capture this simple yet important aspect.

For this reason, inspired by the work by [2], we employed a common metric often used in information retrieval applications, the Normalized Discounted Cumulative Gain (NDCG).

The NDCG is able to evaluate the quality of the ranking produced by a certain query by looking at the first $p$ position of the ranked elements list. The premise of NDCG is that highly relevant items appearing lower in a search result list

should be penalized as the graded relevance value is reduced proportionally to the position of the result.

The non-normalized DCG until position $p$ is defined as follows:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \qquad (5)$$

where $rel_i$ is a positive number encoding the affinity that the $i^{th}$ element of the retrieved list has with the query element. The DCG is agnostic upon how the relevance is computed. The $NDCG_p$ is computed by normalizing the $DCG_p$ with respect to the Ideal Discounted Cumulative Gain (IDCG), that is defined as the DCG of the list obtained by sorting all its elements by descending relevance:

$$NDCG_p = \frac{DCG_p}{IDCG_p} \qquad (6)$$

$IDCG_p$ is the best possible ranking. Thanks to this normalization, $NDCG_p$ acquires values in the range $[0, 1]$.

*Computing $rel_i$ values*

We concentrate our attention on image-retrieval, given that is the most common scenario in real-world search engines. Therefore, in our work, we consider a caption as a query, while the retrieved elements are images.

Being a cross-modal retrieval setup, the relevance should be a value obtained from a function operating on an image $I_i$ and a caption $C_j$. In principle, it could be possible to use the $\phi(I_i, C_j)$ learned by [3], [6]. The problem is that $\phi$ is a complex neural network, and $I_i$, $C_j$ are drawn from a dataset of thousands of elements, in the best case. This means that constructing a $N_c \times N_i$ relevance matrix is computationally unfeasible, where $N_c$ is the number of total captions and $N_i$ is the total number of images in the dataset.

Usually, in the considered datasets, images come with a certain number of associated captions. Thus, instead of computing $\phi(I_i, C_j)$, we could think of computing $\tau(\bar{C}_i, C_j)$ instead, where $\bar{C}_i$ is the set of all captions associated to the image $I_i$. With this simple expedient, we could efficiently compute quite large relevance matrices using similarity between captions, that in general are computationally much cheaper.

As a result, for our image-retrieval objective we define $rel_i = \tau(\bar{C}_i, C_j)$, where $C_j$ is the query caption and $\bar{C}_i$ are the captions associated with the $i^{th}$ retrieved image.

In our work we use ROUGE-L [34] and SPICE [35] as functions $\tau$ for computing captions similarities.

## VI. EXPERIMENTS

We train the Transformer Encoder Reasoning Network and we measure its performance on the MS-COCO [36] dataset, by measuring the effectiveness of our approach on the image retrieval task. We compare our results against state-of-the-art approaches on the same dataset, using the introduced metric.

The MS-COCO dataset comes with a total of 123,287 images. Every image has associated a set of 5 human-written captions describing the image.

We follow the splits introduced by [4] and followed by the subsequent works in this field [1], [7], [15]. In particular, 113,287 images are reserved for training, 5000 for validating and 5000 for testing.

At test time, results for both 5k and 1k image sets are reported. In the case of 1k images, the results are computed by performing 5-fold cross-validation on the 5k test split and averaging the outcomes.

We computed caption-caption relevance for the NDCG metric using ROUGE-L [34] and SPICE [35]. These two metrics capture different aspects of the sentences. In particular, ROUGE-L operates on the longest common sub-sequences, while SPICE exploits graphs associated with the syntactic parse trees, and has a certain degree of robustness against synonyms. In this way, SPICE is more sensible to high-level features of the text and semantic dependencies between words and concepts rather than to pure syntactic constructions of the sentences. We set the NDCG parameter $p = 25$ as in [2] in our experiments.

### A. Implementation Details

We employ the BERT model pretrained on the masked language task on English sentences, using the PyTorch implementation by HuggingFace [1]. These pretrained BERT embeddings are 768-D. For the visual pipeline, we used the already available bottom-up features extracted on the MS-COCO dataset. They are freely available on GitHub [2] and they are 2048-D. In the experiments we used the fixed-size descriptors, selecting for each image the features of the top 36 most confident detections. However, our pipeline can work with a variable-length set of regions for each image, by appropriately masking the attention weights in the TE layers.

Concerning the reasoning steps, we used a stack of 4 non-shared TE layers for visual reasoning. We found the best results when fine-tuning the BERT pretrained model, and we did not introduce any further non-shared TE layers for the language pipeline.

We used 2 final TE layers with weights shared among the visual and textual pipelines. All the TEs feed-forward layers are 2048-dimensional and the dropout is set to 0.1. Weight sharing in the last TE layers is possible if the input vectors from both visual and textual pipelines share the same number of dimensions. For this reason, before entering the last shared-weight TEs, both the visual and textual vectors are linearly projected to a 1024-D space, which is also the dimensionality of the final common space, as in [1].

We trained for 30 epochs using Adam optimizer with a learning rate of 0.00002. The $\alpha$ parameter of the hinge-based triplet ranking loss is set to 0.2, as in [1], [7].

We used a batch size of 90, instead of 128 as in previous works, due to hardware limitations.

---

[1]https://github.com/huggingface/transformers
[2]https://github.com/peteanderson80/bottom-up-attention

**Query**: A large jetliner sitting on top of an airport runway.



**Query**: An eating area with a table and a few chairs.

Fig. 3. Example of image retrieval results for a couple of query captions. The green marked images represent the exact-matching elements. These are incorrect results for the Recall@1 metric (and for the first query even for Recall@5). However, in the very first positions, we find non-matching yet relevant images. These are common examples where NDCG really succeed over the Recall@K metric.

| Model | Recall@K | | | NDCG | |
| | K=1 | K=5 | K=10 | ROUGE-L | SPICE |
|---|---|---|---|---|---|
| *1K Test Set* | | | | | |
| VSE0 [1] | 43.7 | 79.4 | 89.7 | 0.702 | 0.616 |
| VSE++ [1] | 52.0 | 84.3 | 92.0 | 0.712 | 0.617 |
| VSRN [7] | 60.8 | 88.4 | 94.1 | 0.723 | 0.620 |
| TERN (Ours) | 51.9 | 85.6 | 93.6 | **0.725** | **0.653** |
| *5K Test Set* | | | | | |
| VSE0 [1] | 22.0 | 50.2 | 64.2 | 0.633 | 0.549 |
| VSE++ [1] | 30.3 | 59.4 | 72.4 | 0.656 | 0.577 |
| VSRN [7] | 37.9 | 68.5 | 79.4 | **0.676** | 0.596 |
| TERN (Ours) | 28.7 | 59.7 | 72.7 | 0.6645 | **0.600** |

TABLE I
IMAGE RETRIEVAL RESULTS ON THE MS-COCO DATASET, ON 1K AND 5K TEST SETS, FOR BOTH THE RECALL@K AND THE INTRODUCED NDCG METRICS.

## B. Results and Discussion

We report the results obtained on the MS-COCO dataset on both 5k and 1k image test sets, and we compare them against the state-of-the-art on the image retrieval task.

For VSRN [7] and VSE [1] we used the original code and pretrained models provided by the authors, updating the evaluation protocol by including the NDCG metric.

Concerning VSRN, in the original paper the results are given for an ensemble of two independently trained models. In our case, we did not consider model ensembling. For this reason, we evaluated VSRN using the best snapshot among the two provided by the authors.

Results are reported in Table I. For the sake of completeness, we report also the values for the Recall@K metric.

Our TERN architecture is able to reach top performance on the NDCG metric with the SPICE-based relevance. Due to the high-level abstraction nature of the SPICE metric, this result confirms the ability of our system to understand complex patterns and abstract concepts both in the visual and textual inputs. We obtain the best gap on the 1K test set, where we improve the current state-of-the-art by 5.3%.

Concerning the NDCG metric with the ROUGE-L computed relevance, our TERN architecture is able to perform slightly worse than VSRN. Overall, the gap between VSRN and our TERN architecture is very subtle, confirming the ability of those architectures to be comparable when we focus on the syntactic and less abstract features of the language.

Despite VSRN performing better in terms of Recall@K, we demonstrated through the NDCG metric that our architecture is better at finding non-exact matching yet relevant elements in the top $p$ positions of the ranked images list. This is a very important result for real-world search engines, where users often query the system for relevant but non-exact matching images.

Figure 3 shows an example of image retrieval using features computed through our TERN architecture. The reported examples show two typical situations in which the NDCG evaluation succeeds over the Recall@K.

## VII. CONCLUSIONS

In this work, we addressed the problem of image-text matching in the context of efficient multi-modal information retrieval. We argued that many state-of-the-art methods do not extract compact features separately for images and text. This is a problem if we want to employ these features in the subsequent indexing stage for efficient and scalable cross-modal information retrieval.

To this aim, we developed a relationship-aware architecture based on the Transformer Encoder (TE) architecture, exploiting self-attention mechanisms, to reason about the spatial and abstract relationships between elements in the image and in the text separately. Perception and reasoning stages are well identifiable and isolated. The final weight sharing between TE modules guarantees consistent processing of the high-level concepts.

In the vision of employing this architecture for efficient multi-modal information retrieval in real-world search engines, we measured our results using an NDCG metric assessing

possibly non-exact but relevant search results. The relevance among images and captions has been evaluated by employing similarity measures defined over captions, `ROUGE-L` and `SPICE` respectively. We demonstrated that our relation-aware approach for reasoning and matching visual and textual concepts achieved state-of-the-art results with respect to current multi-modal matching architectures on the proposed retrieval metric, for the task of image retrieval.

In the near future, we manage to enforce some reconstruction constraints for better shaping the common space, like reconstructing the sentences from the visual features, as in [7], or recovering the image regions from the captions. Also, major interest should be given to the optimization objective. In particular, it would be interesting to attenuate the very aggressive behavior of the hinge-based triplet ranking loss for better appreciating non-exact matches at training time.

## REFERENCES

[1] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *BMVC 2018*. BMVA Press, 2018, p. 12.

[2] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, "Picture it in your mind: generating high level visual representations from textual descriptions," *Inf. Retr. J.*, vol. 21, no. 2-3, pp. 208–229, 2018.

[3] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS 2019*, 2019, pp. 13–23.

[4] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR 2015*. IEEE Computer Society, 2015, pp. 3128–3137.

[5] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11208. Springer, 2018, pp. 212–228.

[6] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *CoRR*, vol. abs/2001.07966, 2020.

[7] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV 2019*. IEEE, 2019, pp. 4653–4661.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS 2017*, 2017, pp. 5998–6008.

[9] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR 2015*. IEEE Computer Society, 2015, pp. 4437–4446.

[10] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[11] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 261–277.

[12] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *CVPR 2017*. IEEE Computer Society, 2017, pp. 7254–7262.

[13] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *CVPR 2017*. IEEE Computer Society, 2017, pp. 1855–1865.

[14] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, 2017, pp. 4127–4136.

[15] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 7181–7189.

[16] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 6163–6171.

[17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017.

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[19] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," pp. 4967–4976, 2017.

[20] N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro, "Learning visual features for relational cbir," *International Journal of Multimedia Information Retrieval*, Sep 2019. [Online]. Available: https://doi.org/10.1007/s13735-019-00178-7

[21] ——, "Learning relationship-aware visual features," in *ECCV 2018 Workshops*, ser. Lecture Notes in Computer Science, L. Leal-Taixé and S. Roth, Eds., vol. 11132. Springer, 2018, pp. 486–501. [Online]. Available: https://doi.org/10.1007/978-3-030-11018-5_40

[22] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11218. Springer, 2018, pp. 711–727.

[25] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR 2019*. Computer Vision Foundation / IEEE, 2019, pp. 10 685–10 694.

[26] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2117–2130, 2019.

[27] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11205. Springer, 2018, pp. 690–706.

[28] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: An efficient subgraph-based framework for scene graph generation," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11205. Springer, 2018, pp. 346–363.

[29] K. Lee, H. Palangi, X. Chen, H. Hu, and J. Gao, "Learning visual relation priors for image-text matching and image captioning with neural scene graph generators," *CoRR*, vol. abs/1909.09953, 2019.

[30] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.

[31] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 6077–6086.

[32] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *CoRR*, vol. abs/1602.07332, 2016.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013*, 2013.

[34] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[35] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV 2016*, ser. Lecture Notes in Computer Science, vol. 9909. Springer, 2016, pp. 382–398.

[36] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.