This is the peer reviewd version of the followng article:

Towards Retrieval-Augmented Architectures for Image Captioning / Sarto, S., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R.. - In: ACM TRANSACTIONS ON MULTIMEDIA COMPUTING, COMMUNICATIONS AND APPLICATIONS. - ISSN 1551-6865. - 20:8(2024), pp. 1-22. [10.1145/3663667]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

09/11/2026 07:29

(Article begins on next page)

Towards Retrieval-Augmented Architectures for Image Captioning

SARA SARTO, University of Modena and Reggio Emilia, Italy
MARCELLA CORNIA, University of Modena and Reggio Emilia, Italy
LORENZO BARALDI, University of Modena and Reggio Emilia, Italy
ALESSANDRO NICOLOSI, Leonardo S.p.A., Italy
RITA CUCCHIARA, University of Modena and Reggio Emilia, Italy and IIT-CNR, Italy

The objective of image captioning models is to bridge the gap between the visual and linguistic modalities by generating natural language descriptions that accurately reflect the content of input images. In recent years, researchers have leveraged deep learning-based models and made advances in the extraction of visual features and the design of multimodal connections to tackle this task. This work presents a novel approach towards developing image captioning models that utilize an external *k*NN memory to improve the generation process. Specifically, we propose two model variants that incorporate a knowledge retriever component that is based on visual similarities, a differentiable encoder to represent input images, and a *k*NN-augmented language model to predict tokens based on contextual cues and text retrieved from the external memory. We experimentally validate our approach on COCO and nocaps datasets and demonstrate that incorporating an explicit external memory can significantly enhance the quality of captions, especially with a larger retrieval corpus. This work provides valuable insights into retrieval-augmented captioning models and opens up new avenues for improving image captioning at a larger scale.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Matching**; *Computer vision tasks.*

Additional Key Words and Phrases: Image Captioning, Image Retrieval, Vision-and-Language.

ACM Reference Format:
Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. 2024. Towards Retrieval-Augmented Architectures for Image Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* (2024)

1 INTRODUCTION

Because of the important role it can play in connecting vision and language in multimedia systems [2, 4, 38], image captioning has emerged as a fundamental task at the intersection of Computer Vision, Natural Language Processing, and Multimedia. Image captioning architectures consist of an image encoding component and a language model that produces a coherent sentence in natural language describing the visual content of the input image. Therefore, it is important to focus on developing appropriate connections between the visual and textual modality [76]. In this context, recent innovations include the usage of attentive-like architectures [4, 60], the incorporation of attributes and tags [49, 91], objects [4], or scene graphs [48, 87, 90]. Despite this progress, the task still features some unique open challenges, which range from having a grounded and detailed understanding of the visual input to the selection of visual objects and semantics that are worth mentioning and their

Authors' addresses: Sara Sarto, University of Modena and Reggio Emilia, Modena, Italy, sara.sarto@unimore.it; Marcella Cornia, University of Modena and Reggio Emilia, Modena, Italy, marcella.cornia@unimore.it; Lorenzo Baraldi, University of Modena and Reggio Emilia, Modena, Italy, lorenzo.baraldi@unimore.it; Alessandro Nicolosi, Leonardo S.p.A., Rome, Italy, alessandro.nicolosi@leonardo.com; Rita Cucchiara, University of Modena and Reggio Emilia, Modena, Italy and IIT-CNR, Pisa, Italy, rita.cucchiara@unimore.it.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. *ACM Transactions on Multimedia Computing, Communications and Applications*

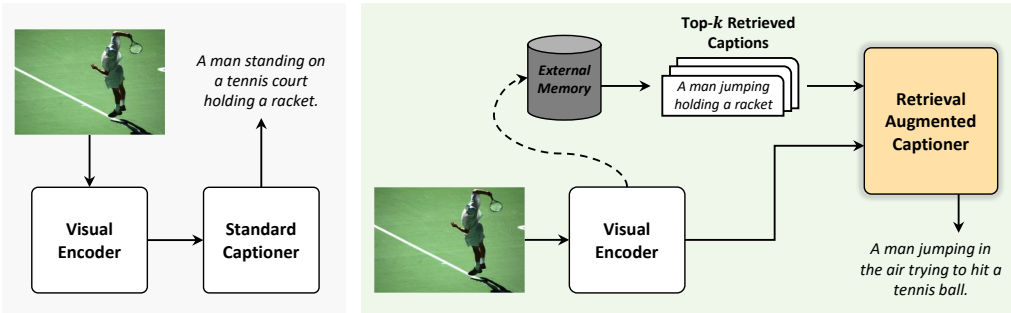

Fig. 1. Comparison between a standard captioner (left) and the proposed retrieval-augmented captioning schema (right), in which an external memory is employed to condition the generation process.

proper translation into a fluent and coherent sentence. To tackle these problems, a recent popular trend has been that of increasing the size of the model [2, 45, 94], an approach which can enhance the ability to memorize information and consistently improve the quality of image descriptions even in few or zero-shot settings. This, however, comes at the cost of increasing the number of learnable parameters and, as a result, the cost of training.

The design of large-scale language models is currently encountering similar issues, where scaling the model to increase its generation capabilities has been the dominant approach so far [11, 97]. An alternative design path is that of separating *language modeling* from *memorization* in the architecture. While the first is an inherent ability that must be learned by the model, the second can also be tackled in a semi-parametric way by presenting the model with relevant examples from the training set, which can be exploited as suggestions during the generation phase. This can be achieved through the insertion of retrieval components [8, 12, 57] that can allow the model to attend textual tokens or hidden states coming from an external memory, rather than relying exclusively on its own activations. This approach reduces the model memorization demands and instead delegates the task to the external memory, which is capable of handling larger-scale data and can be easily accessed through approximate nearest-neighbor searches.

In this paper, we present a thorough analysis on the design of image captioning models which follow the design path outlined above, and which are thus endowed with retrieval components (Fig 1). In particular, we introduce two Transformer-based architectures entirely based on fully-attentive mechanisms that integrate a knowledge retriever component and can provide the language model with appropriate cues from an external memory. The language model then exploits the information retrieved from the external memory to generate textual tokens, taking into consideration both prior context and retrieved textual elements. Following recent modeling trends, we investigate the design of an encoder-decoder protocol – in which the connection with the external memory is based on cross-attention – and that of a decoder-only protocol – in which the language model is conditioned on retrieved items through self-attention. In the former case, a kNN-augmented attention layer is fused with the local context via a learnable gate, while in the latter case pre-processed retrieved captions are provided directly to the self-attention layers of the decoder.

Our experiments and analyses are conducted on the COCO dataset [51], the reference benchmark for standard image captioning, in comparison with Transformer-based baselines that do not leverage additional knowledge retrieved from an external memory. To assess the generalization abilities of our approach, we also validate its performance on the novel object captioning task, using the nocaps dataset [1]. Finally, we investigate the role of the external memory content, by building

different retrieval indexes with image-text pairs from both COCO [51] and Conceptual Captions 3M (CC3M) [70]. The results of our experiments demonstrate that incorporating an external memory can significantly boost the generation process and improve the caption quality, especially with a larger retrieval corpus, in both of our proposed architecture variations.

Contributions. To sum up, the contributions of this paper are as follows:

- We propose a novel framework for image captioning in which the model is augmented with an external knowledge retrieval component. Specifically, our language model makes use of information retrieved from the external memory to generate textual tokens, considering both the previous context and the retrieved text.
- To validate the effectiveness of our retrieval-augmented architecture under different settings, we design two distinct fully-attentive variants, where one feeds the retrieved captions to self-attention layers while the other combines a kNN-augmented attention layer with the local context through a learnable gate and cross-attention.
- We carefully evaluate our model on COCO and nocaps, demonstrating that incorporating an external memory can effectively improve the generation process.
- Finally, we show that using a richer and larger retrieval index, such as one containing textual elements from CC3M, can further improve the quality and semantic richness of generated captions.

This work is an improved and extended version of our conference paper [68]. With respect to this previous work, the proposed solution is extended by introducing a decoder-only architecture in which the retrieved captions are pre-processed and concatenated to the input, rather than utilizing them in cross-attention. Moreover, additional analyses and experiments are conducted, also considering a larger retrieval corpus and showing the effectiveness of both versions of our retrieval-augmented captioning model on novel object captioning.

2 RELATED WORK

Image Captioning. Image captioning is a wide-ranging task that has witnessed research on visual information extraction, text generation, and semantics incorporation. Before the advent of deep learning, traditional image captioning approaches were based on the generation of simple template sentences, which were later filled by the output of an object detector or an attribute predictor [75, 88]. Throughout the years, numerous deep learning-based techniques have been suggested: from early methods that primarily used CNN-based encoders [30, 74, 77] and RNN-based language models [16, 32] to nowadays models that make use of attentive and Transformer-based architectures [79]. The former are usually enhanced with additive attention mechanisms that can incorporate spatial knowledge, initially from a grid of CNN features [22, 66, 81, 85] and then using image regions extracted with an object detector [4, 36, 82, 96]. To further improve the encoding of objects and their relationships, graph convolution neural networks have been employed as well [87, 90], to integrate semantic and spatial relationships between objects or to encode scene graphs. Transformer-based models, instead, are employed both in the visual encoding stage [17, 54], either applied directly to image patches [23] or to refine features from a visual backbone, as well as in the language modeling part [31, 55]. In this context, following the success of the Transformer model in Natural Language Processing (NLP) tasks [21, 79], several captioning solutions based on fully-attentive mechanisms have been proposed, becoming the dominant choice in image captioning [19, 46, 59].

The emergence of Transformer-based models in this field has also led to the development of effective variations of self-attentive operators [19, 31, 35, 53, 60] and early-fusion approaches [33, 49] based on BERT-like architectures [21] that merge vision and language features using a single

Transformer stream. As for image encoding, a common and recent strategy involves using visual features extracted from large-scale cross-modal architectures [7, 10, 68, 73], such as CLIP [62].

This paper follows the prevalent path of utilizing visual features obtained from large-scale multi-modal models and a fully-attentive language model, and proposes a retrieval-enhanced Transformer-based architecture.

Retrieval-Augmented Architectures. The task of retrieval has been applied for decades in image similarity search, advancing over the years from the use of local descriptors to convolutional encoders, until nowadays solutions [24, 25, 86] based on ViT models [23]. On the other side, as large-scale language models become bigger, they gain the ability to retain more information from their training data which leads to improved performance on a variety of downstream tasks [9, 63]. This suggests that enhancing models with retrieval, thus fostering their memorization capabilities, may lead not only to further improvements but also to savings in model size.

In the past, image [20, 26, 52, 72, 93] and video [71] tagging has been recognized as a successful practice to boost relevance matching for information retrieval. In fact, tagging is a mechanism for assigning a set of text labels (*e.g.* keywords or terms) to an image or a video, and can be treated as anchors to guide the visual-language alignment more explicitly. Some vision-and-language methods used tags as an additional input to boost the final performance [18, 27, 49, 91]. However, predicted tags may be incomplete, inconsistent, and sparse, especially when compared to sentences, longer paragraphs, or entire documents that usually contain more complete information. Recently, the same idea has been applied to language models, gaining significant interest [57]. To integrate knowledge into a language model, this line of work retrieves, from an external memory, items that are related to the input, either from a single modality [8, 29, 40] or from multimodal documents [13, 14, 34], allowing the language model to exploit them and generate more accurate predictions. Approaches such as REALM [29], RAG [44], and RETRO [8] integrate Wikipedia passages and other web-scale sources as external memory to benefit downstream knowledge-intensive tasks as, for example, question answering. Some works train the retrieval model via contrastive learning [39], while others [28] train a single-document retriever by concatenating each retrieved result with the query, to compute the final loss independently. A recent work [34], instead, focuses on multimodal tasks and proposes to incorporate the retrieval scores directly into an attentive fusion module, allowing the gradients to be backpropagated through the retriever component.

In this paper, we draw inspiration from these lines of research and explore the integration of retrieval techniques in image captioning. In our case, gradients are not backpropagated into the external memory, which is crucial to the scalability of our method. To incorporate the corresponding k-most similar retrieved captions with the rest of our system, we perform kNN searches on the extracted visual features, as described in [37]. In this way, the model can use the retrieval mechanism to access all the training data and is also, in principle, not constrained to the data seen during training. kNN searches have also been applied in the model proposed in [22] where retrieved captions are used exclusively during inference to determine the best caption among a set of predicted ones according to the METEOR score [6], without actually contributing to the generation process as in our case. Another related approach is that presented by Wu *et al.* [84], where a gated attention module is introduced to attend to the internal states of a Transformer seen during past training iterations. In our approach, retrieval is employed on an external memory rather than on internal activations, eventually using a single scalar gate to effectively merge retrieved information with input words, instead of using a learned per-head parameter.

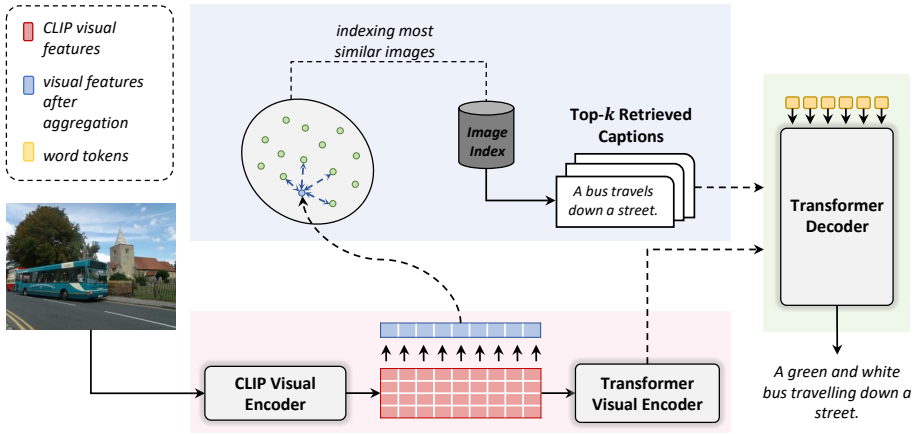

Fig. 2. Schema of the proposed knowledge retriever component (see Fig. 3 for architectural details of the language models). Given an input image, visual features are extracted using a CLIP-based image encoder. These features are then used to retrieve a set of similar textual sentences, starting from the corresponding image representations, that are employed as additional knowledge during the generation of the caption.

3 PROPOSED METHOD

3.1 Preliminaries

Traditionally, image captioning models aim at modeling a distribution $p(y|I)$ over possible natural language descriptions y given an input image I. During the training stage, the model is optimized with a time-wise language modeling objective, usually expressed with a cross-entropy loss. Given an image I and a ground-truth caption $\hat{y}$, in the form of a sequence of tokens, the objective at each time-step t is to predict a probability density over the token dictionary given previous ground-truth tokens $\{\hat{y}_\tau\}_{\tau<t}$ [18, 19]. Depending on the particular tokenization algorithm of choice, tokens might correspond to entire words or to sub-words, and the cross-entropy loss encourages the predicted probability distribution to match the ground-truth token $\hat{y}_t$. An additional fine-tuning stage based on reinforcement learning might also be carried out – in this case, the model is usually asked to generate an entire caption y by relying on its own prediction of previous tokens. The generated caption is then usually matched with ground-truth captions to obtain a reward signal [66].

With the aim of separating the language modeling and memorization capabilities of the captioner, we augment the model with an external memory of textual descriptions (Fig. 2), which will serve as the memory of the model. Under this setting, we can decompose the modeling of the probability distribution $p(y|I)$ into two steps: ① *retrieval* of relevant textual items from the external memory and ② *prediction* of the textual description (or language modeling), conditioned on retrieved items. Firstly, given an image I we retrieve a set of descriptions $\{z_i\}_i$ from the external memory, performing kNN searches in a visual similarity space. Then, we condition our language model on both the input image I and the set of retrieved descriptions $\{z_i\}_i$. From a probabilistic point of view, this amounts to modeling $p(y|\{z_i\}_i, I)$ and marginalizing over the set of retrieved captions.

3.2 External Memory and Knowledge Retrieval

The retrieval of relevant textual items from the external memory aims at modeling $p(z|I)$ given a corpus of image-text pairs and an input query image I. This is done by performing an approximate nearest-neighbor search in the external memory, which we define through an inner product

similarity between image embeddings. The relevance function $f(\cdot,\cdot)$ between the query image and images in the corpus is defined as

$$f(I_1, I_2) = \texttt{Embed}(I_1)^\intercal \texttt{Embed}(I_2), \tag{1}$$

where $\texttt{Embed}(\cdot)$ is a function that maps an image to a vector. After kNN lookup, the external memory can return all the captions of the selected images, which can then be employed as a source of conditioning for the language model.

To encode input images and build the $\texttt{Embed}(\cdot)$ function, we choose to employ the visual encoder of a CLIP model [62] based on the ResNet [30] architecture. In this kind of architecture, the classical ResNet structure is completed with an attention pooling layer: here, a single learnable query is employed to perform attention over the grid of output features of the last convolutional layer, so to pool the grid of activation in an attention-aware fashion. In our case, we remove the attention pooling layer and adopt a custom aggregation function (*e.g.* average or max) so to have a more fine-grained visual representation and a higher control of the pooling procedure.

Compared to more traditional visual feature backbones (*e.g.* CNNs trained on visual classification) the CLIP backbone has the advantage of having been trained to match image-text pairs, which creates a representation that is inherently multimodal. Empirically, we observed that this representation is more robust with respect to vision-only descriptors, as also reported in recent literature [7, 73].

3.3 Designing Retrieval-Augmented Language Models

Given an external memory from which a set of relevant captions $\{z_i\}_i$ can be extracted, we now discuss the design of a retrieval-augmented language model $p(y|\{z_i\}_i, I)$, which is in charge of predicting the output caption while being conditioned on both the input image and items retrieved from the external memory. Compared to a traditional image captioning model, which only models $p(y|I)$, a retrieval-augmented model must implement a connection between its inherent language modeling capabilities, which in a Transformer-based model take place in self-attention layers, and the sequences of tokens that form the retrieved captions. The framework we employ for caption generation is an encoder-decoder Transformer [79], where the encoder is in charge of processing the input image, while the decoder acts as language model.

3.4 Visual Encoder

The input of the encoder is a sequence of grid feature vectors extracted from the input image (see Sec. 3.2). Each encoder layer is then composed of a self-attention layer and a feed-forward layer, as in the standard Transformer [79].

In particular, all intra-modality interactions between image-level features are modeled via scaled dot-product attention, without using recurrence. Attention operates on three sets of vectors, namely a set of queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$ and values $\boldsymbol{V}$, and takes a weighted sum of value vectors according to a similarity distribution between query and key vectors. In the case of scaled dot-product attention, the operator can be formally defined as

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \tag{2}$$

where $\boldsymbol{Q}$ is a matrix of n_q query vectors, $\boldsymbol{K}$ and $\boldsymbol{V}$ both contain n_k keys and values, all with the same dimensionality, and d is a scaling factor.

Given a set of grid image features $\boldsymbol{X}$ extracted from an input image, attention is used to obtain a permutation invariant encoding of $\boldsymbol{X}$ through the self-attention operations used in the Transformer [79]. In this case, queries, keys, and values are obtained by linearly projecting the input features, and the operator can be defined as

$$\mathcal{S}(\boldsymbol{X}) = \text{Attention}(W_q\boldsymbol{X}, W_k\boldsymbol{X}, W_v\boldsymbol{X}), \tag{3}$$

where W_q, W_k, W_v are matrices of learnable weights. The output of the self-attention operator is a new set of elements $\mathcal{S}(\boldsymbol{X})$, with the same cardinality as $\boldsymbol{X}$, in which each element of $\boldsymbol{X}$ is replaced with a weighted sum of the values, *i.e.* of linear projections of the input (Eq. 2).

The output of the self-attention attention is then applied to a position-wise feed-forward layer composed of two affine transformations with a single non-linearity, which are independently applied to each element of the set. Each of these sub-components (self-attention and position-wise feed-forward) is then encapsulated within a residual connection and a layer norm operation. The complete definition of an encoding layer can be finally written as:

$$\begin{aligned} \boldsymbol{J} &= \text{AddNorm}(\mathcal{S}(\boldsymbol{X})) \\ \tilde{\boldsymbol{X}} &= \text{AddNorm}(\mathcal{F}(\boldsymbol{J})), \end{aligned} \tag{4}$$

where $\mathcal{F}$ indicates a feed-forward layer and AddNorm indicates the composition of a residual connection and of a layer normalization.

Given the aforementioned structure, multiple encoding layers are stacked in sequence, so that the i-th layer consumes the output set computed by layer $i-1$. This amounts to creating multi-level encodings of the relationships between image features, in which higher encoding layers can exploit and refine relationships already identified by previous layers. A stack of N encoding layers will therefore produce an output $\tilde{\mathcal{X}} = \tilde{\boldsymbol{X}}^N$, obtained from the output of the last encoding layer.

3.5 Textual Decoder

The decoder, instead, takes as input the sequence of tokens comprising the ground-truth caption and is asked to predict a left-shifted version of it. The self-attention here is masked so that each token can attend only elements to its left, and the decoder then effectively models an autoregressive generation process. Each layer of the decoder comprises at least one self-attention layer, a cross-attention layer with the encoder output, and one feed-forward layer.

Given an input sequence of vectors $\boldsymbol{Y}$, and outputs from the last encoder layer $\tilde{\mathcal{X}}$, the cross-attention operators connects $\boldsymbol{Y}$ to all elements in $\tilde{\mathcal{X}}$. Formally, the operator is computed using queries from the decoder and keys and values from the encoder:

$$C(\tilde{\boldsymbol{X}}^i, \boldsymbol{Y}) = \text{Attention}(W_q\boldsymbol{Y}, W_k\tilde{\boldsymbol{X}}^i, W_v\tilde{\boldsymbol{X}}^i). \tag{5}$$

As the prediction of a word should only depend on previously predicted words, the decoder layer comprises a masked self-attention operation that connects queries derived from the t-th element of its input sequence $\boldsymbol{Y}$ with keys and values obtained from the left-hand subsequence, *i.e.* $\boldsymbol{Y}_{\leq t}$. Also, the decoder layer contains a position-wise feed-forward layer, and all components are encapsulated within AddNorm operations. The final structure of the decoder layer can be written as:

$$\begin{aligned} \boldsymbol{J} &= \text{AddNorm}(C(\tilde{\mathcal{X}}, \text{AddNorm}(\mathcal{S}_{\text{mask}}(\boldsymbol{Y}))) \\ \tilde{\boldsymbol{Y}} &= \text{AddNorm}(\mathcal{F}(\boldsymbol{J})), \end{aligned} \tag{6}$$

where $\boldsymbol{Y}$ is the input sequence of vectors and $\mathcal{S}_{\text{mask}}$ indicates a masked self-attention over time. Finally, our decoder stacks together multiple decoder layers, helping to refine both the understanding

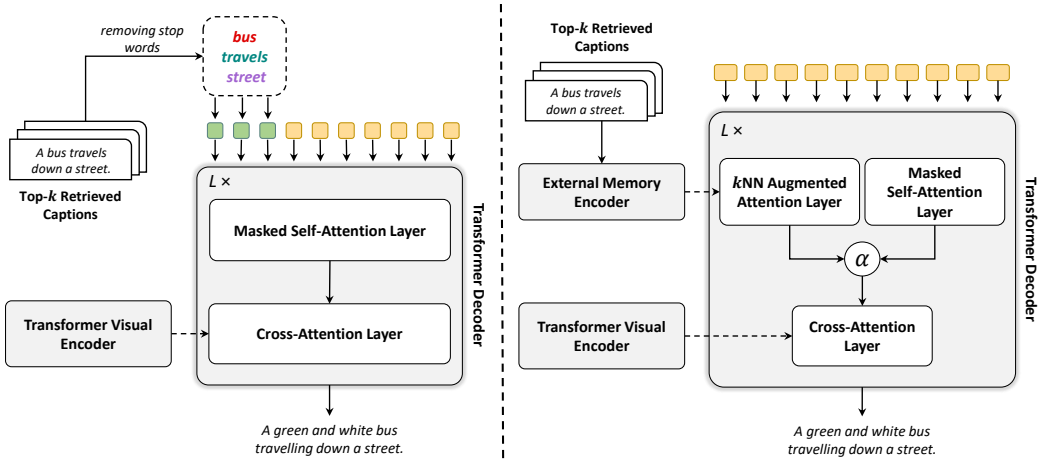

Fig. 3. Architectural schema of the RA-T$^{\mathcal{S}}$ (self-attention-based) and RA-T$^{\mathcal{X}}$ (cross-attention-based) language models. In RA-T$^{\mathcal{S}}$, retrieved captions are employed as prefix of the decoder textual sequence, after removing stop words and duplicate words. In RA-T$^{\mathcal{X}}$, instead, retrieved captions are first passed through a Transformer encoder and then used in a kNN cross-attention layer inside the captioner decoder. The contribution of retrieved captions is regulated by a learnable gating mechanism that combines the output of the kNN cross-attention layer with those of the standard self-attention over the input sequence.

of the textual input and the generation of the next tokens. Overall, the decoder takes as input word vectors, and the t-th element of its output sequence encodes the prediction of a word at time $t+1$, conditioned on $Y_{\leq t}$. After taking a linear projection and a softmax operation, this encodes a probability over tokens in the dictionary.

3.6 Retrieval-Augmented Generation

As shown in Fig. 3, we devise two architectural variants for realizing the connection between the decoder self-attention and items retrieved from the external memory, one based on self-attentive connections, termed RA-T$^{\mathcal{S}}$, and one based on cross-attention connections, termed RA-T$^{\mathcal{X}}$.

3.6.1 ***RA-T$^{\mathcal{S}}$.*** Under this configuration, retrieved captions are employed as a prefix of the decoder sequence, so that the self-attention operator can naturally retrieve relevant suggestions coming from the external memory while generating a caption. This might also be seen as a variant of the prompting technique [5]. A naïve concatenation of the retrieved captions would however be computationally intractable with the growth of k; furthermore, the self-attention layer would not have a principled way of distinguishing retrieved and generated tokens. Therefore, we adopt two strategies: we clean the retrieved captions by removing stop words and eliminating duplicate words that appear in more than one caption, so to obtain a set of unique words. Formally, the input of the decoder under this setting can be defined as

$$Y_{\text{RA-T}} = [\texttt{unique}(\{z_i\}_i), Y], \tag{7}$$

where $[\cdot,\cdot]$ indicates concatenation, and $\texttt{unique}(\cdot)$ indicates a function that removes stop words and eliminates duplicates.

To increase the effectiveness of this strategy, we also employ two different learnable segment embeddings [21] to distinguish between retrieved words and generated ones (*i.e.*, Y). Also, as

cleaned words represent an unordered set, we do not apply position embeddings to this segment, so to keep the permutation invariance of the self-attention operator.

3.6.2 ***RA-T^X.*** In this case, an additional cross-attention layer is placed in parallel to the masked self-attention layer of the decoder. Retrieved captions are firstly encoded independently through a bidirectional Transformer encoder to get a refined representation of the tokens, then the aforementioned cross-attention layer performs a cross-attention over the resulting outputs. As the cross-attention layer is placed in parallel to the masked self-attention layer, the same queries are employed for both layers. Formally, given the input sequence of tokens $\boldsymbol{Y}$ and the set of retrieved captions $\boldsymbol{Z} = \{z_i\}_i$, this configuration can be written as follows:

$$\tilde{\boldsymbol{Z}} = \text{Encoder}(\boldsymbol{Z}) \tag{8}$$

$$\tilde{\boldsymbol{L}} = \text{AddNorm}(\mathcal{S}_{\text{mask}}(\boldsymbol{Y})) \tag{9}$$

$$\tilde{\boldsymbol{M}} = \text{AddNorm}(C(\tilde{\boldsymbol{Z}}, \boldsymbol{Y})), \tag{10}$$

where Encoder indicates a Transformer encoder, such as the one employed for visual features encoding. The second equation refers to the self-attention between tokens of the caption. The last equation, instead, refers to the additional cross-attention operation with retrieved captions. Noticeably, in this layer, all tokens from all retrieved captions are attended.

Finally, the outputs coming from the two parallel layers need to be combined. To this aim, we devise a learnable gate, with which the model can regulate the importance of the output coming from the self-attention layer and that coming from the cross-attention layer. Conceptually, this amounts to choosing between the local context encoding and retrieved captions. Formally,

$$\tilde{\boldsymbol{J}} = \alpha \cdot \tilde{\boldsymbol{L}} + (1 - \alpha)\tilde{\boldsymbol{M}}, \tag{11}$$

where α represents the learnable gate. In practice, this is learned as the sigmoid of a single scalar network parameter. The output of this linear combination is then passed to the usual cross-attention with visual features and the feed-forward network to obtain the output sequence.

3.7 Training Protocol

As briefly introduced in Sec. 3.1, at training time the input of the decoder is the ground-truth sentence $\{\text{BOS}, \hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$, and the model is trained with a cross-entropy loss to predict the shifted ground-truth sequence, *i.e.* $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n, \text{EOS}\}$, where BOS and EOS are special tokens to indicate the start and the end of the caption. While during training, the model predicts all output tokens simultaneously, the prediction process at inference happens sequentially. In each step, the model takes the partially decoded sequence as input, then selects the next token by sampling from its output probability distribution, continuing this process until an EOS marker is generated.

Following previous works [4, 65, 66], after a pre-training step using cross-entropy, we further optimize the sequence generation using Reinforcement Learning. Specifically, we implement a variation of the self-critical sequence training method [66], which employs the REINFORCE algorithm on sequences generated through the beam search algorithm [4]. Additionally, unlike the approach in [4, 66], we establish a baseline for the reward using the mean of the rewards rather than relying on greedy decoding. Specifically, given the output of the decoder, we sample the top-k words from the decoder probability distribution at each timestep and always maintain the top-k sequences with the highest probability. We then compute the reward of each sentence y^i and backpropagate with respect to it. The final gradient expression for one sample is thus:

$$\nabla_\theta L(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\left((r(y^i) - b)\nabla_\theta \log p(y^i)\right) \tag{12}$$

where $b = \left(\sum_i r(y^i)\right)/k$ is the baseline, computed as the mean of the rewards obtained by the sampled sequences. To reward the overall quality of the generated caption, we use the image captioning metric which better correlates with human judgment, namely CIDEr [80].

4 EXPERIMENTS

4.1 Experimental Setup

Datasets and Evaluation Protocol. We first analyze the effectiveness of our retrieval-augmented architecture on the COCO dataset [51]. We select this dataset as it is the reference evaluation benchmark for image captioning models [76] as well as the largest dataset with human-collected annotations for the task. In particular, COCO contains more than 120,000 images, each of them manually annotated with five textual captions collected by using Amazon Mechanical Turk. To train and test our solution, we follow the splits defined by Karpathy *et al.* [38], where the training set is composed of 82,783 images while both the validation and test set contain 5,000 images. However, there are also 30,504 images that were originally in the validation set of the original COCO dataset but were left out in this split. As done by previous image captioning [4, 19, 35, 60] and image-text matching [15, 43, 47, 56, 89] literature, we add these images in the training set thus obtaining a total of 113,287 images to train the model. Additionally, we perform experiments on the nocaps dataset [1] which has been introduced for the novel object captioning task where the goal is to effectively describe objects not present in the image-caption pairs used to train the captioning model. The dataset contains 4,500 validation and 10,600 test images from the Open Images V4 dataset [42], where each image has been annotated with 10 human-written captions. Images can be further grouped into three subsets depending on their nearness to COCO (*i.e.* in-domain, near-domain, and out-of-domain images). Specifically, in-domain images only contain objects that are also present in the original COCO dataset, out-of-domain images exclusively contain object classes that are not present in COCO and thus represent the most challenging evaluation set, while near-domain images contain both in-domain and out-of-domain object classes. Under this setting, we train our model on COCO and evaluate on nocaps validation set, submitting generated captions to the nocaps evaluation server[1].

Following captioning literature, we measure the performance of our approach using standard captioning evaluating metrics, namely BLEU [61], METEOR [6], ROUGE [50], CIDEr [80], and SPICE [3]. In particular, BLEU gauges the precision of word n-grams by comparing predicted and ground-truth sentences. As done in previous works, we evaluate our predictions with BLEU using n-grams of length 1 and 4 (referred to as B-1 and B-4, respectively). ROUGE (R) calculates an F-measure with a recall bias, using a technique based on identifying the longest common subsequence. METEOR (M), instead, assesses captions by aligning them to one or more ground-truth sentences, utilizing alignments based on various matches like exact, stem, synonym, and paraphrase between words and phrases. CIDEr (C) computes the average cosine similarity between n-grams present in the generated caption and reference sentences, employing TF-IDF weighting. SPICE (S), finally, prioritizes semantic content over fluency in generated captions by matching tuples from the candidate and reference scene graphs. Empirical evidence suggests that BLEU and ROUGE exhibit weaker correlations with human judgment compared to other metrics [67, 80], yet the standard practice in image captioning literature involves reporting all mentioned metrics. During evaluation, we compare each generated caption with all ground-truth sentences associated with the corresponding image, using the COCO caption evaluation library[2] to obtain the final scores.

[1] https://eval.ai/web/challenges/challenge-page/355/overview
[2] https://github.com/tylin/coco-caption

Retrieval Index. We build distinct versions of the retrieval index using image-text pairs from both COCO [51] and Conceptual Captions 3M (CC3M) [70]. Specifically, when building the index on the COCO dataset, we only consider image-text pairs from the training set and, during training, we do not retrieve captions associated to the current training image to mitigate the risks of overfitting. When instead building the index on CC3M, we consider all image-text pairs available in the training set thus having an index of around 3.1M different elements. In this case, instead of using the original captions that come from noisy alt-text tags, we employ a recently proposed large-scale captioning model to produce a cleaner set of textual sentences. Specifically, we use the BLIP model [45], in its ViT-L/14 version pre-trained on 129M image-text pairs and finetuned on the COCO dataset, to generate a new caption for each of the 3.1M images of the CC3M training set.

To improve the computational efficiency of our model, in our experiments, we utilize approximate kNN search instead of exact kNN search. Specifically, we use the Faiss library [37] and employ a graph-based Hierarchical Navigable Small Worlds (HNSW) index with 32 links per vertex. While vector transform techniques such as PCA or vector quantization can be employed to reduce the index size and scale to larger datasets, we do not use them for simplicity.

Implementation Details. In our model, we utilize intermediate features from CLIP-RN50×16 [62] to represent images. We instead use Byte Pair Encoding (BPE) [69] with a vocabulary size of 49,408 to encode words from both the input subsequence and retrieved sentences. Standard sinusoidal positional encodings [79] are employed to represent word positions. To improve efficiency, we limit the length of the output token sequence to 40 word tokens. We project visual features and word tokens into d-dimensional vectors with $d = 384$ and pass them as input to our Transformer-based model. The model consists of $L = 3$ layers in both the encoder and decoder, with six attention heads. In the RA-T$^{\mathcal{S}}$ version, we employ the NLTK NLP toolkit[3] to remove stop words. In the RA-T$^{\mathcal{X}}$, instead, the external memory encoder is composed of a single Transformer layer with the same number of heads and dimensionality as the rest of the model. The gate α is initialized to zero at the beginning of the training.

Training Details. As outlined in Sec. 3.7, to train our solution, we employ the standard two-stage training protocol typically used by almost all captioning models: in the first stage, we optimize the model with a time-wise cross-entropy loss, while in the second stage we perform finetuning using reinforcement learning with the CIDEr score as reward. During cross-entropy pre-training, we employ LAMB [92] as optimizer with the learning rate scheduling strategy outlined in [79], using a warmup of 6,000 iterations and a batch size equal to 1,080. When instead finetuning with reinforcement learning, we use the Adam optimizer [41] with a batch size of 80 and a fixed learning rate equal to 5×10^{-6}. During both CIDEr-based finetuning and sampling of predicted captions, we employ beam search with a beam size equal to 5.

All experiments are conducted by training the models on two NVIDIA Quadro RTX-5000 GPUs, using five gradient accumulation steps for both training phases. To accelerate the training process and save memory, we use ZeRo memory offloading [64] and mixed-precision [58]. Overall, training the model with cross-entropy typically requires approximately 24 hours for the model that does not utilize the external memory, and around 30 hours for both variations of our complete model. Instead, finetuning with CIDEr-based optimization takes four and five days for a standard Transformer-based model and our retrieval-augmented architecture, respectively.

4.2 Quality of Nearest Neighbor Captions

To confirm that nearest neighbor captions are a suitable source of additional knowledge and that can be thus employed to improve the final performance, we first need to evaluate their relevance with

[3]https://www.nltk.org/

Table 1. Performance of the k nearest-neighbor captions using distinct retrieval indexes.

	$k = 5$						$k = 10$					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
index: *COCO*												
mean score	49.4	10.6	1.70	36.1	44.1	12.0	49.2	10.4	16.8	35.9	43.1	11.8
max score (Oracle)	65.5	14.4	24.8	49.4	77.8	19.1	72.3	22.1	28.5	55.1	96.5	22.6
index: *CC3M (original)*												
mean score	25.5	0.3	9.3	20.4	13.5	5.8	25.3	0.3	9.2	20.2	13.0	5.6
max score (Oracle)	40.9	1.4	15.2	31.9	32.1	11.6	46.5	2.6	17.4	36.0	40.4	13.8
index: *CC3M*												
mean score	59.2	10.7	20.9	44.2	61.4	13.8	58.7	10.4	20.6	43.8	59.8	13.6
max score (Oracle)	74.1	24.9	28.6	56.4	101.8	21.2	78.5	31.4	31.3	60.3	116.6	23.8
	$k = 20$						$k = 40$					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
index: *COCO*												
mean score	48.9	10.2	16.7	35.7	42.1	11.6	48.6	10.0	16.5	35.4	41.1	11.4
max score (Oracle)	77.7	30.8	31.9	60.2	114.2	25.4	82.0	39.4	34.9	64.5	130.4	28.0
index: *CC3M (original)*												
mean score	25.2	0.3	9.1	20.1	12.6	5.5	25.0	0.3	9.0	20.0	12.2	5.3
max score (Oracle)	51.8	4.2	19.3	39.9	49.3	16.1	56.6	6.9	21.3	43.7	58.6	18.1
index: *CC3M*												
mean score	58.2	9.9	20.3	43.4	58.1	13.3	57.6	9.5	20.0	43.0	56.3	13.0
max score (Oracle)	81.9	37.7	33.7	63.6	129.4	25.9	85.1	43.6	36.0	66.7	141.8	27.8

respect to the ground-truth captions. To do this, given an image from the test set, we retrieve the k nearest captions from one of the created nearest neighbor indexes using our relevance function to compare visual elements (*i.e.* in this experiment we use a standard average pooling to aggregate image features). Then, we measure the similarity between retrieved and ground-truth captions by calculating the mean captioning scores and the score of the retrieved caption with the highest similarity to the ground-truth. The latter can be considered as an upper-bound score, where an oracle evaluator is used to select the best caption among the k nearest ones. We perform this analysis using three different retrieval indexes: one containing image-text pairs from the COCO dataset and the others containing elements from the CC3M dataset, either using the original CC3M textual descriptions (*i.e. CC3M (original)*) or the textual sentences predicted by the BLIP model [45].

The results are presented in Table 1 as the number k of retrieved sentences varies. As it can be noticed, retrieving a limited number of captions (*e.g.* $k = 5$) leads, for all retrieval indexes, to a set of captions that only partially correlates with the ground-truth. On the other hand, increasing the number of retrieved captions slightly degrades the performance, with a decrease in the mean CIDEr score from 44.1 to 41.1 when using the index containing COCO elements. The maximum (oracle) score, instead, shows considerably higher results. Specifically, it reaches up to 130.4 and 141.8 CIDEr points, respectively using the COCO and CC3M indexes, when retrieving a large number of captions (*i.e.* $k = 40$). The worst results, in terms of both mean and maximum scores, are obtained with the retrieval index with the original CC3M corpus which leads to 58.6 CIDEr points in terms of maximum score using $k = 40$. These results can be explained by the quality of CC3M textual sentences which are crawled from the web and, although semantically richer, have a substantially different style from the human-annotated captions contained in the COCO dataset.

Table 2. Performance analysis of the two versions of our retrieval-augmented Transformer, by varying the number of retrieved sentences k and the aggregation function used. Results are reported after cross-entropy pre-training using the COCO index.

		RA-T$^{\mathcal{S}}$						**RA-T$^{\mathcal{X}}$**					
Aggregation Function	k	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
ℓ_2-norm sum	5	78.2	38.0	28.5	58.0	122.2	21.6	78.3	38.6	28.9	58.3	123.1	21.8
ℓ_2-norm sum	10	77.9	37.5	28.4	57.6	121.5	21.7	78.5	38.6	28.8	58.3	122.7	22.1
ℓ_2-norm sum	20	77.9	38.0	28.7	57.9	123.0	21.7	78.3	38.6	28.9	58.3	123.8	21.9
ℓ_2-norm sum	40	78.5	38.1	28.6	58.0	122.9	21.8	78.2	39.1	28.7	57.9	122.8	22.0
max	5	78.5	38.2	28.8	58.3	122.7	21.8	78.6	38.6	28.9	58.3	123.6	22.0
max	10	78.2	38.1	28.4	58.0	123.1	21.6	78.3	38.5	28.9	58.2	123.8	22.2
max	20	78.4	38.3	28.6	58.4	123.1	21.7	78.3	38.6	29.0	58.3	124.0	22.1
max	40	79.0	38.3	28.6	58.4	122.9	21.9	78.3	38.3	28.9	58.3	123.6	22.0
mean	5	78.7	38.3	28.7	58.1	123.3	22.0	78.6	38.7	29.1	58.5	124.0	22.0
mean	10	78.5	38.4	28.9	58.1	**123.7**	21.9	78.9	38.9	28.9	58.5	**124.5**	22.1
mean	20	78.6	38.3	28.6	58.2	123.4	21.8	78.5	38.6	28.9	58.3	124.2	22.0
mean	40	78.4	37.8	28.7	58.3	122.9	21.0	78.4	38.4	28.9	58.3	123.1	22.0

Although our embedding space is built on top of state-of-the-art descriptors, the high quality achieved by the oracle captions for higher values of k indicates that there is still significant room for improvement in the quality of the embedding space. It is worth noting, also, that the quality of the captions plays a crucial role in determining the quality of the embedding space, and it is not only dependent on the size of the retrieval index, as demonstrated by the results using original CC3M captions. Therefore, even a smaller index with high-quality captions can potentially result in a better embedding space than a larger one with lower-quality descriptions.

4.3 Model Ablation and Analysis

Role of Different Aggregation Functions and Number of Retrieved Captions. We then examine the outcomes of various aggregation functions for aggregating visual features and retrieving the most similar images. In particular, we experiment with three different aggregation strategies: a standard average pooling over spatial features, a max pooling, and a sum of ℓ_2-normalized features followed by an ℓ_2-norm of the result, which has shown promising results in previous works, especially in the field of image and video retrieval [78]. We report the performance analysis in Table 2, showing the results after cross-entropy pre-training. In this case, we only employ the retrieval index built on COCO and compare the results of the architecture based on self-attentive connections (*i.e.* RA-T$^{\mathcal{S}}$) and those of the solution based on cross-attention (*i.e.* RA-T$^{\mathcal{X}}$). According to the results, a standard mean of grid features performs better than other aggregation functions, such as max pooling and sum of ℓ_2-normalized features for both model variants, regardless of the number k of retrieved sentences. Specifically, while ℓ_2-norm and max-based aggregation functions respectively lead to 123.0 and 123.1 in terms of CIDEr for the RA-T$^{\mathcal{S}}$ models and to 123.8 and 124.0 for the RA-T$^{\mathcal{X}}$ version, using a standard average pooling leads an improvement of 0.6 and 0.5 CIDEr points for RA-T$^{\mathcal{S}}$ and RA-T$^{\mathcal{X}}$ respectively.

We also evaluate the effectiveness of our retrieval strategy by varying the number of retrieved captions, using $k = 5, 10, 20, 40$. Results are reported also in this case in Table 2. As it can be noticed, using a number of retrieved captions equal to 10 and 20 generally leads to the best performance in terms of all considered evaluation metrics. Overall, the best results are obtained using the mean as

Table 3. Ablation study results of the two proposed model variants in comparison with a standard Transformer without retrieval. Results are reported after cross-entropy pre-training using the COCO index. For each model, we show the absolute improvement with respect to the baseline (*i.e.* Transformer (w/o external memory).

	B-1	B-4	M	R	C	S
Transformer (w/o external memory)	78.1	38.1	28.5	58.0	121.6	21.8
RA-T$^{\mathcal{S}}$ (w/ full sentences)	78.4	38.3	28.7	58.0	122.7	21.8
RA-T$^{\mathcal{S}}$	**78.5** (+0.4)	**38.4** (+0.3)	**28.9** (+0.4)	**58.1** (+0.1)	**123.7** (+2.1)	**21.9** (+0.1)
Transformer (w/o external memory)	78.1	38.1	28.5	58.0	121.6	21.8
RA-T$^{\mathcal{X}}$ (w/o gate)	78.3	38.3	**28.9**	58.1	122.5	21.9
RA-T$^{\mathcal{X}}$ (w/o stop words)	78.7	38.6	28.8	58.3	124.0	22.1
RA-T$^{\mathcal{X}}$	**78.9** (+0.8)	**38.9** (+0.8)	**28.9** (+0.4)	**58.5** (+0.5)	**124.5** (+2.9)	**22.1** (+0.3)

Table 4. Comparison with state-of-the-art models on the Karpathy-test split. Overall best results are underlined.

	B-1	B-4	M	R	C	S
Up-Down [4]	79.8	36.3	27.7	56.9	120.1	21.4
ORT [31]	80.5	38.6	28.7	58.4	128.3	22.6
GCN-LSTM [90]	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [87]	81.0	39.0	28.4	58.9	129.1	22.2
AoANet [35]	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [19]	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN [60]	80.8	39.5	29.5	59.2	132.0	23.4
X-Transformer [60]	80.9	39.7	29.5	59.1	132.8	23.4
DPA [53]	80.3	40.5	29.6	59.2	133.4	23.3
DLCT [55]	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [95]	81.8	40.1	29.8	59.5	135.6	23.3
DIFNet [83]	81.7	40.0	29.7	59.4	136.2	23.2
Transformer (w/o external memory)	81.9	39.7	29.6	59.4	135.3	**23.6**
RA-T$^{\mathcal{S}}$ (index: *COCO*)	82.0	40.1	29.6	59.4	136.4	23.2
RA-T$^{\mathcal{S}}$ (index: *CC3M*)	**82.5**	**40.8**	**29.7**	**59.8**	**136.7**	**23.6**
Transformer (w/o external memory)	81.9	39.7	29.6	59.4	135.3	23.6
RA-T$^{\mathcal{X}}$ (index: *COCO*)	**82.4**	40.5	29.8	**59.8**	136.5	23.8
RA-T$^{\mathcal{X}}$ (index: *CC3M*)	82.2	**41.0**	**30.0**	**59.8**	**136.7**	**23.9**

aggregation function and $k = 10$ retrieved captions, for both proposed model variants. Specifically, our RA-T$^{\mathcal{S}}$ achieves 123.7 points in terms of CIDEr, while the cross-attention-based model (*i.e.* RA-T$^{\mathcal{X}}$) reaches 124.5 CIDEr points. Therefore, we employ this configuration (*i.e.* $k = 10$ and the mean as aggregation function) for both model variations in all experimental analyses shown in the rest of the section.

Role of External Memory and Other Architectural Details. In Table 3, we first analyze the effectiveness of our retrieval-augmented solution compared to a standard encoder-decoder Transformer architecture without retrieval, with the same dimensionality and number of layers as our complete model. As it can be seen, both model configurations achieve significantly better results than the Transformer baseline, with an increase of 2.1 and 2.9 points in terms of CIDEr score

Table 5. Performances on nocaps validation set. Overall best results are underlined.

	Near		Out		Overall	
	C	S	C	S	C	S
NBT [1]	61.2	9.9	62.4	8.9	60.2	9.5
Up-Down [1]	73.6	11.3	66.4	9.7	73.1	11.1
$\mathcal{M}^2$ Transformer [19]	75.4	11.7	69.4	10.0	75.0	11.4
Transformer (w/o external memory)	87.4	12.7	66.7	10.8	85.3	12.5
RA-T$^{\mathcal{S}}$ (index: *COCO*)	88.2	12.5	68.6	10.6	86.3	12.3
RA-T$^{\mathcal{S}}$ (index: *CC3M*)	**89.3**	**13.0**	**69.5**	**11.0**	**86.8**	**12.7**
Transformer (w/o external memory)	87.4	12.7	66.7	10.8	85.3	12.5
RA-T$^{\mathcal{X}}$ (index: *COCO*)	88.5	12.8	**68.6**	11.0	86.3	12.6
RA-T$^{\mathcal{X}}$ (index: *CC3M*)	**89.4**	**13.1**	**68.6**	**11.1**	**87.0**	**12.8**

respectively for RA-T$^{\mathcal{S}}$ and RA-T$^{\mathcal{X}}$ models. These results confirm the appropriateness of employing an external memory to effectively improve the quality of generated captions.

We also evaluate the role of other architectural details as the removal of stop words from retrieved elements and the effect of the learned gate. In particular, we compare the RA-T$^{\mathcal{S}}$ model with a variant that takes as input the full retrieved sentences instead of removing stop words. From the results, we can notice that the model using the full retrieved captions achieves slightly worse performance, while still being better than the vanilla Transformer model. In particular, the RA-T$^{\mathcal{S}}$ model achieves an improvement of 1.0 CIDEr points (*i.e.* 122.7 vs 123.7) compared to the model taking as input entire retrieved sentences, thus confirming the appropriateness of using the set of words as input for this model variant. Regarding the RA-T$^{\mathcal{X}}$ architecture, we compare it with a model without the learned gating mechanism, where masked self-attention and cross-attention between input tokens and retrieved captions are performed in sequence and a baseline that takes the same input as the RA-T$^{\mathcal{S}}$ counterpart (*i.e.* the clean set of words appearing in the retrieved items without stop words) while maintaining the structure of the RA-T$^{\mathcal{X}}$ version. Also in this case, the results confirm the effectiveness of the architectural choices made, with an improvement of 2 CIDEr points (*i.e.* 122.5 vs 124.5) and 0.5 CIDEr points (*i.e.* 124.0 vs 124.5) respectively compared to the model without learnable gate and the baseline taking as input the captions without stop words.

4.4 Comparison to the State of the Art

Results on COCO. In Table 4 we report the results on the standard Karpathy test split after CIDEr-based finetuning, comparing our model performance with that of different state-of-the-art captioning models. Although several architectures pre-trained on large-scale datasets and then finetuned on COCO have recently been proposed [33, 49, 94], in this analysis we only consider captioning models trained exclusively on the COCO dataset. Specifically, we compare against methods with language models based on LSTMs such as Up-Down [4], eventually enhanced with spatial and scene graphs like GCN-LSTM [90] and SGAE [87] or self-attentive mechanisms such as AoANet [35], X-LAN [60], and DPA [53]. Moreover, we consider captioning architectures entirely based on the standard Transformer model such as ORT [31], $\mathcal{M}^2$ Transformer [19], X-Transformer [60], and RSTNet [95], even combining visual features from multiple backbones as in the case of DLCT [55] and DIFNet [83].

Results of both versions of our complete retrieval-augmented architecture are reported using both COCO and CC3M retrieval indexes and compared with those of a standard Transformer-based model without the retrieval component. As it can be seen, the efficacy of the kNN-augmented

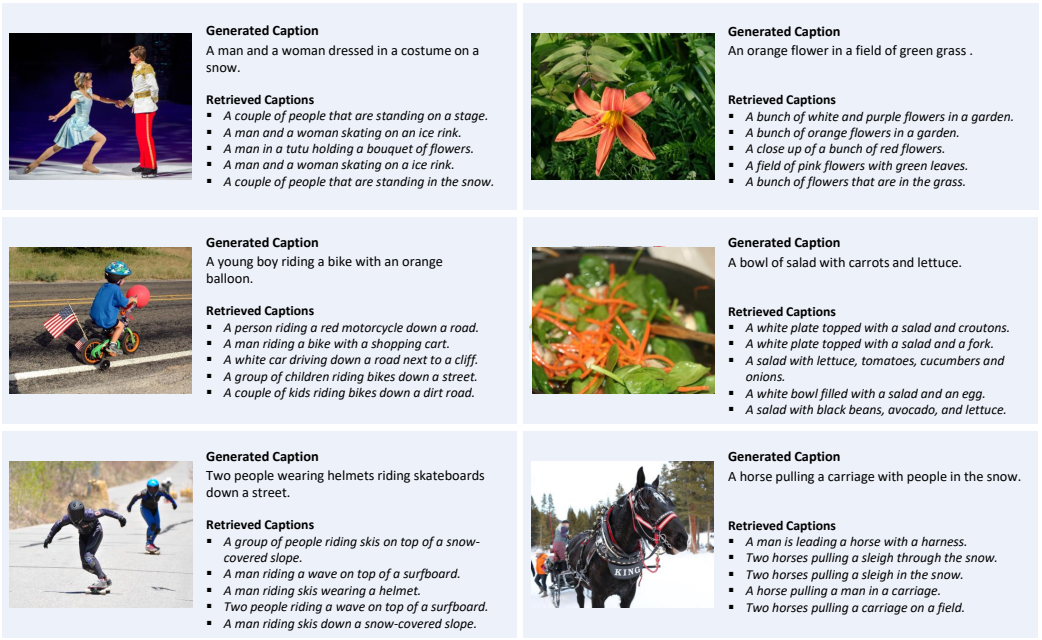

Fig. 4. Generated captions on sample images from the COCO dataset, along with five retrieved captions.

language model is confirmed even after finetuning with CIDEr-based optimization, with an increase of 1.1 and 1.2 CIDEr points respectively comparing RA-T$^{\mathcal{S}}$ and RA-T$^{\mathcal{X}}$ with COCO retrieval index to the standard Transformer-based architecture. The use of a larger index such as the one containing a cleaned version of CC3M captions can further boost the performance, leading to an overall CIDEr score of 136.7 for both model variations. It can also be noticed that, while after cross-entropy pre-training the RA-T$^{\mathcal{X}}$ version slightly outperforms the RA-T$^{\mathcal{S}}$ model, after reinforcement learning finetuning the two model variants perform comparably, thus demonstrating that both architectures can be a valid solution for incorporating external knowledge. Furthermore, we observe that the proposed retrieved-augmented model achieves promising and competitive performance compared to other state-of-the-art methods, and surpasses them in terms of all evaluation metrics.

Results on nocaps. In Table 5, we extend our analysis on the nocaps dataset, using RA-T$^{\mathcal{S}}$ and RA-T$^{\mathcal{X}}$ after finetuning with CIDEr optimization on the COCO dataset. Also in this case, we compare our results against a standard Transformer model and employ both versions of our retrieval index (*i.e.* the one containing COCO captions and the other composed of CC3M sentences predicted by the BLIP model). The effectiveness of the proposed retrieval-augmented strategy is confirmed also in this setting, with an improvement of 1.5 and 1.7 on the entire validation set, respectively for the self- and cross-attention model variants with CC3M index. The contribution of a larger retrieval index becomes more evident, especially on near-domain and out-of-domain image-text pairs, which contain visual concepts outside of the COCO dataset and thus can benefit from a larger and semantically richer set of retrievable items. In fact, the CIDEr score on out-of-domain images is equal to 69.5 for the RA-T$^{\mathcal{S}}$ model with CC3M index compared to 68.6 achieved by the same version of the model augmented with COCO retrieval index.

Qualitative Results. Finally, in Fig. 4 and 5 we report sample captions generated by our model on images respectively from COCO and nocaps. While in the former we show examples of captions

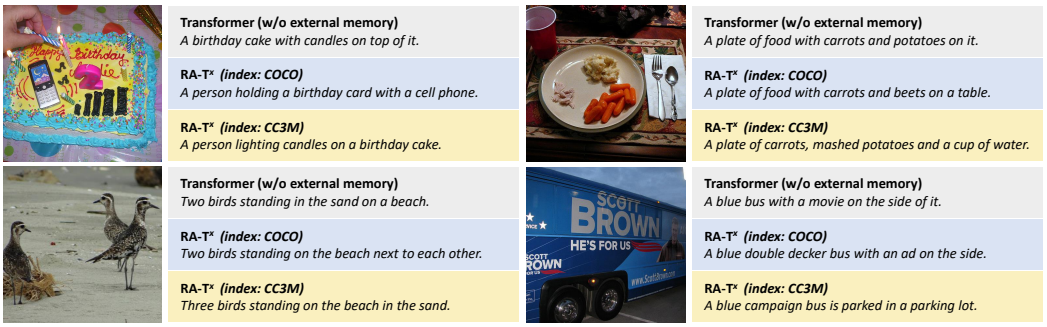

Fig. 5. Qualitative results on sample images from nocaps, comparing captions generated by our model with those generated by a standard Transformer without retrieval.

retrieved from the external memory, in the latter we compare our results generated with both COCO and CC3M indexes to those generated by a Transformer-based model without the retrieval component. As it can be seen, retrieving captions related to the input image can help the language model generate more relevant and accurate captions by providing it with additional contextual information. For example, in the top-right example of Fig. 4, we can observe that the caption generated by our model has highly comparable content to the retrieved sentences (*i.e.* "orange flower" and "green grass"), providing evidence of the efficacy of our retrieval-based approach. When instead comparing our generated captions with those of a standard Transformer model without retrieval (Fig. 5), we can observe that our results are generally more coherent with the visual content of input images and semantically richer, especially when using the retrieval index containing CC3M elements. For example, in the top-right image of Fig. 5, the baseline model without external memory correctly recognizes the "carrots" and "potataoes" on a "plate". However, the caption predicted by our RA-T$^{\mathcal{X}}$ model with CC3M index is more detailed and complete, also describing the presence of a "cup of water" and recognizing that the potatoes are "mashed". Similar observations can be made for the other examples, further confirming from a qualitative point of view the effectiveness of our retrieval-augmented solution for image captioning.

5 CONCLUSION

In this paper, we have presented a novel framework for image captioning that is augmented with an external memory from which additional knowledge can be retrieved to help the generation process. The COCO and nocaps datasets were used to carry out experimental testing, which showed that adding retrieval capabilities to a captioning architecture can result in high-quality textual descriptions. This finding suggests that there is potential for additional research in this area.

ACKNOWLEDGMENTS

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This work has been conducted under a research grant co-funded by Leonardo S.p.A. and supported by the PNRR-M4C2 (PE00000013) project "FAIR - Future Artificial Intelligence Research", funded by the European Commission, and by the PRIN project "CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content" (CUP B87G22000460001), co-funded by the Italian Ministry of University and Research.

REFERENCES

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[5] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441* (2022).

[6] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[7] Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[8] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the International Conference on Machine Learning*.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[10] Davide Caffagni, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. SynthCap: Augmenting Transformers with Synthetic Data for Image Captioning. In *Proceedings of the International Conference on Image Analysis and Processing*.

[11] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The (R)Evolution of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2402.12451* (2024).

[12] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[13] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[14] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. *arXiv preprint arXiv:2209.14491* (2022).

[15] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-modal Graph Matching Network for Image-text Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4 (2022), 1–23.

[16] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[17] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Explaining Transformer-based Image Captioning Models: An Empirical Analysis. *AI Communications* 35, 2 (2022), 111–129.

[18] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. 2023. Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets. *International Journal of Computer Vision* (2023), 1–20.

[19] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[20] Chaoran Cui, Jialie Shen, Jun Ma, and Tao Lian. 2015. Social Tag Relevance Estimation via Ranking-Oriented Neighbour Voting. In *Proceedings of the ACM International Conference on Multimedia*.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[22] Tiago do Carmo Nogueira, Cássio Dener Noronha Vinhal, Gélson da Cruz Júnior, and Matheus Rudolfo Diedrich Ullmann. 2020. Reference-based model using multimodal gated recurrent units for image captioning. *Multimedia Tools and Applications* 79 (2020), 30615–30635.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

[24] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. 2021. Vision Transformer Hashing for Image Retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo*.

[25] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. 2021. Training Vision Transformers for Image Retrieval. *arXiv preprint arXiv:2102.05644* (2021).

[26] Bruno Fruchard, Sylvain Malacria, Géry Casiez, and Stéphane Huot. 2023. User Preference and Performance using Tagging and Browsing for Image Labeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

[27] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2012. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. *IEEE Transactions on Image Processing* 22, 1 (2012), 363–376.

[28] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning*.

[29] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the International Conference on Machine Learning*.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[31] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *Advances in Neural Information Processing Systems*.

[32] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[33] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling Up Vision-Language Pre-Training for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[34] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[35] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on Attention for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[36] Weitao Jiang, Weixuan Wang, and Haifeng Hu. 2021. Bi-Directional Co-Attention Network for Image Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 4 (2021), 1–20.

[37] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[38] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[39] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[40] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *Proceedings of the International Conference on Learning Representations*.

[41] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*.

[42] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* (2018).

[43] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*.

[44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.

[45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine*

Learning.

[46] Jingyu Li, Zhendong Mao, Hao Li, Weidong Chen, and Yongdong Zhang. 2023. Exploring Visual Relationships via Transformer-based Graphs for Enhanced Image Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications* (2023).

[47] Wenhui Li, Xinqi Su, Dan Song, Lanjun Wang, Kun Zhang, and An-An Liu. 2023. Towards Deconfounded Image-Text Matching with Causal Inference. In *Proceedings of the ACM International Conference on Multimedia*.

[48] Xiangyang Li and Shuqiang Jiang. 2019. Know More Say Less: Image Captioning Based on Scene Graphs. *IEEE Transactions on Multimedia* 21, 8 (2019), 2117–2130.

[49] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*.

[50] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*.

[52] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag Ranking. In *Proceedings of the International Conference on World Wide Web*.

[53] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. 2020. Prophet Attention: Predicting Attention with Future Attention. In *Advances in Neural Information Processing Systems*.

[54] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. CPTR: Full Transformer Network for Image Captioning. *arXiv preprint arXiv:2101.10804* (2021).

[55] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-Level Collaborative Transformer for Image Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[56] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. In *Proceedings of the International Conference on Content-based Multimedia Indexing*.

[57] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. *arXiv preprint arXiv:2302.07842* (2023).

[58] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed Precision Training. In *Proceedings of the International Conference on Learning Representations*.

[59] Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. 2023. Bottom-up and Top-down Object Inference Networks for Image Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 5 (2023), 1–18.

[60] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-Linear Attention Networks for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[61] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*.

[63] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1, 8 (2019), 9.

[64] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*.

[65] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations*.

[66] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[67] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[68] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-Augmented Transformer for Image Captioning. In *Proceedings of the International Conference on Content-based Multimedia Indexing*.

[69] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[71] J. Shen, M. Wang, and T. Chua. 2016. Accurate online video tagging via probabilistic hybrid modeling. *Multimedia Systems* 22 (2016), 99–113.

[72] Jialie Shen, Meng Wang, Shuicheng Yan, and Xian-Sheng Hua. 2011. Multimedia Tagging: Past, Present and Future. In *Proceedings of the ACM International Conference on Multimedia*.

[73] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How Much Can CLIP Benefit Vision-and-Language Tasks?. In *Proceedings of the International Conference on Learning Representations*.

[74] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

[75] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[76] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 539–559.

[77] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[78] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations. In *Proceedings of the International Conference on Learning Representations*.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[80] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[81] Anqi Wang, Haifeng Hu, and Liang Yang. 2018. Image Captioning with Affective Guiding and Selective Attention. *ACM Transactions on Multimedia Computing, Communications and Applications* 14, 3 (2018), 1–15.

[82] Haiyang Wei, Zhixin Li, Feicheng Huang, Canlong Zhang, Huifang Ma, and Zhongzhi Shi. 2021. Integrating Scene Semantic Knowledge into Image Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications* 17, 2 (2021), 1–22.

[83] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. 2022. DIFNet: Boosting Visual Information Flow for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[84] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing Transformers. In *Proceedings of the International Conference on Learning Representations*.

[85] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*.

[86] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-Modal Transformer with Global-Local Alignment for Composed Query Image Retrieval. *IEEE Transactions on Multimedia* (2023), 1–13.

[87] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[88] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2T: Image parsing to text description. *Proc. IEEE* (2010).

[89] Tao Yao, Yiru Li, Ying Li, Yingying Zhu, Gang Wang, and Jun Yue. 2023. Cross-Modal Semantically Augmented Network for Image-Text Matching. *ACM Transactions on Multimedia Computing, Communications and Applications* (2023).

[90] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision*.

[91] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[92] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *Proceedings of the International Conference on Learning Representations*.

[93] Zheng-Jun Zha, Meng Wang, Jialie Shen, and Tat-Seng Chua. 2012. Text Mining in Multimedia. *Mining Text Data* (2012), 361–384.

[94] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[95] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[96] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. 2021. Exploring Pairwise Relationships Adaptively From Linguistic Context in Image Captioning. *IEEE Transactions on Multimedia* 24 (2021), 3101–3113.

[97] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* (2023).