Retrieval-Augmented Transformer for Image Captioning

Sara Sarto
University of Modena and Reggio Emilia
Modena, Italy
237128@studenti.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

Rita Cucchiara
University of Modena and Reggio Emilia
Modena, Italy
rita.cucchiara@unimore.it

ABSTRACT

Image captioning models aim at connecting Vision and Language by providing natural language descriptions of input images. In the past few years, the task has been tackled by learning parametric models and proposing visual feature extraction advancements or by modeling better multi-modal connections. In this paper, we investigate the development of an image captioning approach with a *k*NN memory, with which knowledge can be retrieved from an external corpus to aid the generation process. Our architecture combines a knowledge retriever based on visual similarities, a differentiable encoder, and a *k*NN-augmented attention layer to predict tokens based on the past context and on text retrieved from the external memory. Experimental results, conducted on the COCO dataset, demonstrate that employing an explicit external memory can aid the generation process and increase caption quality. Our work opens up new avenues for improving image captioning models at larger scale.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Visual content-based indexing and retrieval**; **Matching**; *Computer vision tasks.*

KEYWORDS

image captioning, image retrieval, vision-and-language.

ACM Reference Format:
Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-Augmented Transformer for Image Captioning. In *Proceedings of International Conference on Content-based Multimedia Indexing (CBMI 2022).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Image captioning has recently emerged as an important task at the intersection of Computer Vision, Natural Language Processing, and Multimedia, thanks to the key role it can have to connect Vision and Language in multimedia systems [3, 8, 31]. Recent advances on image captioning, indeed, have demonstrated that fully-attentive architectures can provide high-quality image descriptions, even in few or zero-shot settings [1, 15, 36, 43]. Regardless of this progress, accurately describing the concepts contained in input image is still a considerable challenge. While some recent works have addressed this by increasing the size of the model [1], thus improving its memorization capabilities, this comes at the cost of increasing the number of learnable parameters and, ultimately, the training cost.

The same issue is currently being faced in large-scale language models, where the adoption of retrieval components is being explored as a viable solution [9, 58]. The basic idea behind these models is that of letting the language model attend textual chunks or hidden states retrieved from an external memory, instead of relying solely on its own activations. In this manner, the memorization requirements of the model are relieved in favor of the external memory, which can handle larger-scale data and can easily be accessed via approximate nearest neighbor searches.

In this paper, we investigate the development of a retrieval component for image captioning. We propose a fully-attentive architecture with a knowledge retriever based on approximate *k*NN searches, which can provide the language model with suitable hints from an external memory. Our language model then employs a *k*NN-augmented attention layer to predict tokens based on the past context and on text retrieved from the external memory. To the best of our knowledge, our proposal is the first model to integrate a retrieval-based memory into an image captioning pipeline.

We conduct experiments on the COCO dataset for image captioning, in comparison with a fully-attentive baseline that does not employ an external memory and with an adaptation of the RETRO architecture [9], which has been devised for large-scale language models. Our experiments show that using an external memory can significantly improve the generation quality and that adding a retrieval component to multi-modal models can be a viable solution. Further, we show that our proposed architecture overcomes the RETRO design [9] by a significant margin.

2 RELATED WORK

Image Captioning. Image captioning is a broad topic that has witnessed research on visual information extraction, text generation, and semantics incorporation. Over the years, different approaches have been proposed. Early works relied on CNN-based encoders and RNN-based language models [18, 34, 48, 57]. Nowadays, attentive and Transformer-based architectures [55] are often employed both in the visual encoding stage [14, 40], either applied to image patches directly [19, 54] or to refine features from a visual backbone, and as language models [11, 13, 25]. Regarding language models, in the last few years, large performance improvements have come from

CBMI 2022, Sept. 14-16, 2022, Graz, Austria

increasing the amount of training data, model size, and performing large-scale training [10, 46].

The introduction of Transformer-based models in image captioning has also brought to the development of effective variants of the self-attention operator [16, 22, 25, 27, 39, 43] and to that of vision-and-language early-fusion approaches [26, 36, 65] based on BERT-like architectures [17]. On the image encoding side, a recent paradigm is that of employing visual features extracted from large-scale multi-modal architectures [6, 7, 15, 50] like CLIP [45].

Convolutional [4] and fully-attentive language models [41, 60, 64] based on the Transformer paradigm have been used due to the limited representation power and sequential nature of RNN-based language models and thanks to their success in NLP tasks such as machine translation and language understandings [17, 52, 55]. In this paper, we follow the dominant track of employing a Transformer-based language model and propose a fully-attentive architecture augmented with retrieval abilities.

Retrieval-Augmented Approaches. Image retrieval has evolved over time from methods based on local descriptors to convolutional encoders until the use of visual Transformers [20, 21]. Large-scale language models can perfectly memorize parts of their training data [12] and increasing model size predictably improves performance on a wide range of downstream tasks [10, 30, 46]. This suggests that enhancing models with retrieval may lead to further improvements and savings in terms of model size. In this work, we take inspiration from this line of research and investigate the incorporation of retrieval in image captioning. We conduct kNN searches [29] on the extracted visual features to integrate the corresponding k-most similar retrieved captions with the rest of our architecture. In this way, the model can access the entire training dataset through the retrieval mechanism and is also, in principle, not limited to the data seen during training [9]. This idea allows us to go conceptually beyond the traditional Transformer language model in which the benefits of model size and data size are linked.

Retrieval-augmented language models have been recently gaining a lot of attention [23, 28, 32, 35]. Most works tackling this direction have devised novel forms of attention to model the connections with the retrieved chunks of text [58, 62]. Borgeaud *et al.* [9] split input sequences into chunks, which are augmented with the k-nearest neighbors using a chunked cross-attention module to incorporate the retrieved text. Wu *et al.* [58] proposed a gated attention module to attend the internal states of a Transformer, seen during past training iterations. The usage of a learned attention gate is closely related to our formulation, even though in our case retrieval is applied towards an external memory rather than on internal activations, and we employ a single scalar gate instead of a learned per-head parameter.

3 PROPOSED METHOD

The goal of an image captioner is that of modeling a distribution $p(y|I)$ over possible captions y given an input image I. During the pre-training stage, the captioner is trained with a time-wise language modeling objective: given an image I and a ground-truth caption $\hat{y}$ from the training set, the objective is to predict word $\hat{y}_t$ given previous ground-truth words $\{\hat{y}_\tau\}_{\tau<t}$ [15, 16]. During fine-tuning, instead, the model is usually asked to generate an entire caption y without relying on previous ground-truth words. The generated caption is then usually matched with ground-truth captions to obtain a reward signal [48].

Our approach decomposes the probability distribution $p(y|I)$ into two steps: *retrieval* and *prediction*. Firstly, given an image I we retrieve possibly related descriptions $\{z_i\}_i$ from an external memory, thanks to a visual similarity space in which k-NN searches can be carried out. Then, we condition our language model on both the input image I and the set of retrieved descriptions $\{z_i\}_i$, thus effectively modeling $p(y|\{z_i\}_i, I)$ and marginalizing over the set of retrieved captions.

3.1 Knowledge Retriever

Given a corpus of image-text pairs and an input query image I, we model $p(z|I)$ by building a knowledge retrieval component. This performs an approximate k-nearest-neighbor search into the external memory, defined through an inner product similarity between image embeddings, *i.e.*:

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

where $\texttt{Embed}(\cdot)$ is a function that maps an image to a vector. The relevance $f(\cdot,\cdot)$ between the query image and images in the corpus is employed to sort images by decreasing similarity. Then, the knowledge retriever returns all captions associated with the selected images, as a source of conditioning for the language model.

To model the visual embedding function, we employ the visual encoder of one of the CLIP models [45], which have been trained contrastively to match image-text pairs. Empirically, we found this relevance function to be more robust in our scenario when compared to vision-only descriptors, as also reported in recent literature [6, 50]. While the maximum inner product search is carried out employing visual queries and values in Eq. 1, the search is implicitly multimodal as it happens inside a visual-semantic space. In contrast to performing a pure multi-modal search with visual queries and textual keys, however, our strategy is computationally lighter as it does not require to forward through a textual encoder.

Specifically, we select a CLIP ResNet-based [24] visual encoder. In this kind of encoder, the image is processed through a sequence of residual layers, then the grid of activations from the last convolutional layer is fed to an attention pool layer. Here, a single query is built from the global average-pooled feature vector, and all elements of the grid act as keys and values. To get a more fine-grained representation of the image and have a higher control on the pooling strategy, we directly take the grid of features from the last convolutional layer and define the $\texttt{Embed}$ function as an aggregation (*e.g.* average, max) of the features contained in the grid (see Fig.1).

3.2 Retrieval-Augmented Language Model

Having defined a knowledge retrieval strategy that provides captions from similar images of an external memory, we devise a retrieval-augmented language model which can predict a time-wise distribution over the vocabulary while being conditioned on both the input image and the set of retrieved captions.

The bone structure of the model is a vanilla encoder-decoder Transformer [55] in which the encoder is employed to process the

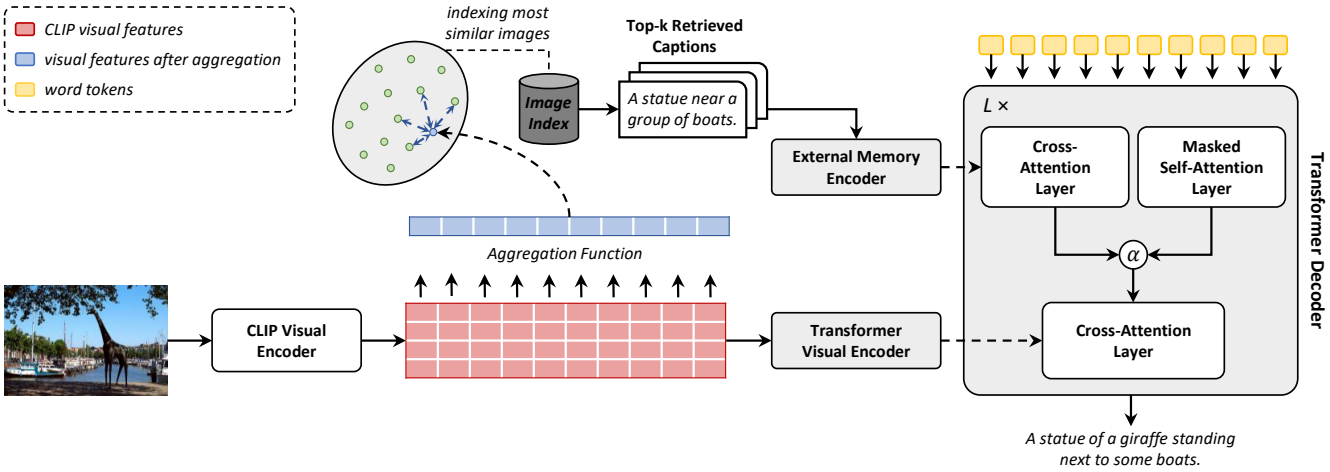

Figure 1: Illustration of our retrieval-augmented Transformer for image captioning.

input image and the decoder acts as a language model. The input of the encoder consists of a flattened sequence of grid feature vectors (as described in Sec. 3.1) which are linearly projected into a vector space. The resulting vectors are then passed through a sequence of encoder layers, each of which does dense self-attention, followed by a feed-forward network (FFN). In the encoder, attention operations are not masked, thus allowing a bidirectional encoding of input image features with complete connectivity. The input text is instead tokenized and embedded into a vector space. The embedding vectors are then passed through a sequence of decoder layers, each of which performs a k*NN-augmented attention*, a cross-attention with the last layer of the encoder, and a feed-forward network. In the decoder, we use a causal attention mask and the token embeddings of the last layer are used to predict the next token.

The kNN-augmented attention layer combines two types of attention: like a normal self-attention layer, it attends the input subsequence encoding the past context. Plus, it performs attention over the set of retrieved captions, thus connecting to the external memory. This is modeled by first encoding all retrieved captions independently through a Transformer encoder, and then performing cross-attention over its outputs. Crucially, the same queries are employed for the self-attention over the input subsequence and for the cross-attention over the encoded retrieved captions.

Given the input sequence of tokens $\{w_i\}_i = \{w_0, ..., w_i, ..., w_T\}$ and the set of retrieved captions $\{z_i\}_i = \{z_0, ..., z_i, ..., z_N\}$, the kNN-augmented attention can be written as follows:

$$\tilde{z}_k = \text{MSA}(z_k, z_k) \tag{2}$$

$$\tilde{w}_t^l = \text{MSA}(w_t, \{w_i\}_{i=1}^t) \tag{3}$$

$$\tilde{w}_t^m = \text{MCA}(w_t, \{\tilde{z}_k\}_k), \tag{4}$$

where k indicates a generic item from the set of retrieved captions, t the t-th element of the sequence of tokens, $\{w_i\}_{i=1}^t$ is the sequence of tokens up to the t-th element, $\text{MSA}(x, y)$ indicates a multi-head self-attention with x mapped to query and y mapped to key-values, and $\text{MCA}(x, y)$ a multi-head cross-attention with x as query and y as key-values. The first equation refers to a self-attention between tokens of each retrieved caption, and we drop the dependency to single tokens for readability. The last equation, instead, refers to the cross-attention operation with retrieved captions. Here, given w_t as query, all tokens from all retrieved captions are attended.

The outputs of the self-attention over the input subsequence and that of the cross-attention over the external memory are combined using a learned gate, which allows the model to choose between local context and retrieved captions. Formally,

$$\tilde{w} = \alpha \cdot \tilde{w}^l + (1 - \alpha)\tilde{w}^m, \tag{5}$$

where gate α is learned as the sigmoid of a single scalar parameter.

As it might be noticed, gradients are not backpropagated into the external memory, which is critical to the scalability of our technique. Following a standard practice in image captioning, we first pre-train our language model with a time-wise cross-entropy loss. Then, we fine-tune using the self-critical sequence learning paradigm (SCST), which employs reinforcement learning over sampled sequences. Specifically, we employ the variant proposed in [16] that employs beam search for sampling, and sets the baseline reward equal to the mean of rewards of generated captions inside a beam. We refer the reader to [48] for a comprehensive treatment of the RL-based fine-tuning stage.

4 EXPERIMENTAL EVALUATION

4.1 Setup

Dataset. Following standard image captioning approaches [3, 16, 27], we train and evaluate our model on the COCO dataset [38], thus not relying on large-scale image-text datasets [15]. COCO is composed of more than 120,000 images, each of them associated with 5 human-collected captions. We follow the splits defined by Karpathy *et al.* [31], using 5,000 images for both validation and testing and the rest for training.

Metrics. According to the standard evaluation protocol, we employ the complete set of captioning metrics: BLEU [44], METEOR [5], ROUGE [37], CIDEr [56], and SPICE [2].

Retrieval Index. We build our retrieval index on the COCO training set. During training, to reduce overfitting risks, we avoid retrieving captions that belong to the current training image. We employ

Table 1: Performance of the k nearest-neighbor captions.

	$k=5$						$k=10$						$k=20$						$k=40$					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
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	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)	65.5	14.4	24.8	49.4	77.8	19.1	72.3	22.1	28.5	55.1	96.5	22.6	77.7	30.8	31.9	60.2	114.2	25.4	82.0	39.4	34.9	64.5	130.4	28.0

approximate kNN search rather than exact kNN search because it significantly improves the computational speed of our model. To this aim, we employ the Faiss library [29] and a graph-based HNSW index with 32 links per vertex, which has a size of 6.7 GB. For simplicity, we do not employ any vector transform (*e.g.* PCA) or vector quantization, although they might be employed to reduce the index size and scale to larger datasets.

Implementation Details. To represent images, we employ CLIP-RN50×16 [45] intermediate features. To represent words, of both the input subsequence and retrieved sentences, we use Byte Pair Encoding (BPE) [49] with a vocabulary size of 49,408. We use standard sinusoidal positional encodings [55] to represent word positions. For efficiency, the length of the output token sequence is limited to 40 tokens. Visual features and word tokens are projected into d-dimensional vectors with $d = 384$ and fed to our Transformer-based model, which has $L = 3$ layers in both encoder and decoder with six attention heads. The external memory encoder has 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.

Pre-training with cross-entropy loss is performed using the LAMB optimizer [63] and following the learning rate scheduling strategy of [55] with a warmup equal to 6,000 iterations and a batch size of 1,080. For the CIDEr-based fine-tuning, we adopt the SCST strategy [48] sampling over the $k = 5$ best sequences from a beam-search scheme, using Adam [33] as optimizer, a batch size equal to 80, and a fixed learning rate of 5×10^{-6}.

All experiments are performed by paralyzing training on two Quadro RTX-5000 GPUs, using five gradient accumulation steps during both cross-entropy pre-training and CIDEr optimization. ZeRo memory offloading [47] and mixed-precision [42] are used to accelerate training and save memory.

4.2 Quality of Nearest Neighbor Captions

To prove the appropriateness of using nearest neighbor captions, we first investigate their quality with respect to ground-truth captions of a given test image. Specifically, given a sample image from the test set, we retrieve the k nearest captions from the training set according to our relevance function. We then compare the retrieved set with ground-truth captions by computing their mean scores as well as the scores obtained by the retrieved caption with maximum similarity with the ground-truth.

Results are reported in Table 1. As it can be seen, retrieving a relatively limited number of captions (*e.g.* $k = 5$) produces a set of captions that have a significant, although low, overlap with the ground-truth. Increasing the number of retrieved captions degrades mean scores. The maximum (oracle) score, instead, reaches significantly high levels, up to 130.4 CIDEr points when retrieving $k = 40$ captions. The quality reached by the oracle caption for higher k outlines that the quality of the embedding space can still be significantly improved, even though ours is based on state-of-the-art descriptors. On the other side, results confirm that retrieved captions tend to become noisy when increasing k, and that even for small k they do not provide a completely reliable signal. The language model, therefore, will need to selectively copy content from retrieved captions, paying attention to their coherency with the actual image content.

Table 2: Performance of a base Transformer captioner and our retrieval-augmented Transformer, by varying the aggregation function, the number k of retrieved sentences, and the number of layers in the external memory encoder. Results are reported after cross-entropy pre-training.

Aggregation Function	Layers	k	B-1	B-4	M	R	C	S
-	-	-	78.1	38.1	28.5	58.0	121.6	21.8
ℓ_2-norm sum	1	5	78.3	38.6	28.9	58.3	123.1	21.8
ℓ_2-norm sum	1	10	78.5	38.6	28.8	58.3	122.7	22.1
ℓ_2-norm sum	1	20	78.3	38.6	28.9	58.3	123.8	21.9
ℓ_2-norm sum	1	40	78.2	39.1	28.7	57.9	122.8	22.0
max	1	5	78.6	38.6	28.9	58.3	123.6	22.0
max	1	10	78.3	38.5	28.9	58.2	123.8	22.2
max	1	20	78.3	38.6	29.0	58.3	124.0	22.1
max	1	40	78.3	38.3	28.9	58.3	123.6	22.0
mean	1	5	78.6	38.7	29.1	58.5	124.0	22.0
mean	1	10	78.9	38.9	28.9	58.5	124.5	22.1
mean	1	20	78.5	38.6	28.9	58.3	124.2	22.0
mean	1	40	78.4	38.4	28.9	58.3	123.1	22.0
mean	2	10	78.9	38.8	28.9	58.3	124.1	22.0
mean	3	10	78.7	39.2	29.0	58.4	124.3	22.0
mean	2	20	78.3	38.1	28.9	58.1	123.1	22.0
mean	3	20	78.3	38.3	28.8	58.3	122.6	22.1

4.3 Model Ablation and Analysis

Role of Different Aggregation Functions. We then move to our full model, and first analyze the results of different aggregation functions to embed visual features and retrieve the most similar images. Specifically, we consider a standard average pooling over grid features, a max pooling, and a sum of ℓ_2-normalized features followed by an ℓ_2-norm of the result, which has demonstrated to be effective in previous image and video retrieval works [53]. Results are reported in Table 2, after cross-entropy pre-training, in comparison with a standard Transformer-based encoder-decoder model without retrieval. We can first notice that *all configurations with the external memory encoder achieve better performance than the baseline* which obtains 121.6 CIDEr points, thus demonstrating the effectiveness of our retrieval-augmented architecture. When

Table 3: Ablation study results, in comparison with RETRO.

	B-1	B-4	M	R	C	S
Ours w/ RETRO block ($C = 2$) [9]	75.2	34.5	26.2	55.7	108.6	20.2
Ours w/ RETRO block ($C = 6$) [9]	74.8	33.9	26.3	55.8	106.2	19.9
Transformer (w/o external memory)	78.1	38.1	28.5	58.0	121.6	21.8
RA-Transformer (w/o gate)	78.3	38.3	**28.9**	58.1	122.5	21.9
RA-Transformer	**78.9**	**38.9**	**28.9**	**58.5**	**124.5**	**22.1**

comparing the different aggregation functions, the results show that a standard mean of grid features performs generally better than the other considered aggregation functions, also according to a different number k of retrieved captions.

Number of Layers and Retrieved Captions. We also evaluate the effect of changing the number k of retrieved sentences and the number of layers in the external memory encoder. In particular, we evaluate the captioning results employing $k = 5, 10, 20, 40$ retrieved elements. From Table 2, it can be seen that $k = 10$ and $k = 20$ generally lead to the best results according to almost all evaluation metrics. Additionally, we compare the results using a different number of Transformer layers (*i.e.* 1, 2, and 3) in the external memory encoder. The architecture with a single encoding layer obtains better performance than those with an increased number of parameters. Overall, the best performance is obtained by the model with the mean as aggregation function, $k = 10$ retrieved captions, and a single Transformer layer in the external memory encoder, with a CIDEr score equal to 124.5 points. This configuration is used in all experiments reported in the rest of the paper.

Retrieval-Enhanced Transformer (RETRO) Baseline. To evaluate the effectiveness of our retrieval strategy, we devise a variant of our model in which we replace the cross-attention between input tokens and retrieved captions with the chunked cross-attention mechanism proposed in [9]. We report the results in Table 3 by varying the chunk size C (*i.e.* $C = 2$ and $C = 6$). Also in this case, the results are obtained after pre-training with cross-entropy loss. When comparing the results by using different chunk sizes, it can be noticed that increasing the chunk size leads to decreasing the final results. Overall, the performance of our retrieval-augmented Transformer with multi-head cross-attention mechanism (*i.e.* RA-Transformer in the Table) is consistently better than that obtained by the version with chunked cross-attention.

Role of External Memory and Learned Gate. We finally evaluate the effectiveness of both the external memory encoder and learned gate to modulate the contribution of retrieved captions. To do this, as shown in Table 3, we design two different baselines. In the first model, we remove the external memory encoder from our architecture thus obtaining a vanilla Transformer-based encoder-decoder model. In the second baseline, instead, we only remove the learned gating mechanism. In this case, masked self-attention and cross-attention between input tokens and retrieved captions are performed in sequence, and the result is fed to the cross-attention with visual features.

As it can be seen, even just by introducing the external memory encoder in the architecture we can obtain enhanced results, with an improvement of 0.9 CIDEr points (*i.e.* 121.6 vs 122.5). The final results are further improved by the introduction of the learned gate α, with an improvement of 2 CIDEr points with respect to the architecture without gate and 2.9 CIDEr points compared to the vanilla Transformer model without external memory.

Table 4: Comparison with state-of-the-art models on the Karpathy-test split.

	B-1	B-4	M	R	C	S
Up-Down [3]	79.8	36.3	27.7	56.9	120.1	21.4
ORT [25]	80.5	38.6	28.7	58.4	128.3	22.6
GCN-LSTM [61]	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [59]	81.0	39.0	28.4	58.9	129.1	22.2
MT [51]	80.8	38.9	28.8	58.7	129.6	22.3
AoANet [27]	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [16]	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN [43]	80.8	39.5	29.5	59.2	132.0	23.4
X-Transformer [43]	80.9	39.7	29.5	59.1	132.8	23.4
DPA [39]	80.3	40.5	29.6	59.2	133.4	23.3
DLCT [41]	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet [64]	81.8	40.1	**29.8**	59.5	135.6	23.3
Transformer (w/o external memory)	81.9	39.7	29.6	59.4	135.3	23.6
RA-Transformer	**82.4**	**40.5**	**29.8**	**59.8**	**136.5**	**23.8**

Role of Approximated kNN Search. We also test when employing exact kNN search, in place of the HNSW index. Removing the approximation in the retrieval phase brings to an increase of 0.3 CIDEr points on the COCO test set, and no significant improvement on all other metrics, thus confirming that approximate searches provide a convenient efficacy-efficiency balance when employing external memories.

Training and Inference Time Analysis. Training with cross-entropy takes around 24 hours for the model without external memory and around 30 hours for our complete model, while fine-tuning with reinforcement learning employs four and five days for a a standard encoder-decoder model and for our retrieval-augmented Transformer, respectively. In fact, with respect to a basic encoder-decoder captioner, our model requires running approximate kNN searches and introduces additional attention layers and operations. At decoding time, this entails a 26% increase in execution times (from 0.170 seconds to 0.215 seconds to decode a single caption)[1].

4.4 Comparison to the State of the Art

We compare the results of our model with those of several recent image captioning models trained exclusively on the COCO dataset. In our analysis, we include methods with LSTM-based language models and attention over image regions such as Up-Down [3], GCN-LSTM [61], SGAE [59], and MT [51], eventually enhanced with self-attention mechanisms such as AoANet [27], X-LAN [43], and DPA [39], and captioning architectures entirely based on fully-attentive mechanisms such as ORT [25], $\mathcal{M}^2$ Transformer [16], X-Transformer [43], DLCT [41], and RSTNet [64].

Table 4 shows the results on the Karpathy-test split after finetuning with CIDEr optimization. We report the results of our complete retrieval-augmented Transformer and of the model trained without retrieval. As it can be seen, the effectiveness of the kNN-augmented

[1]Execution times have been measured on a machine with Intel(R) Xeon(R) W-2235 CPU and Quadro RTX 5000 GPU, with a mini-batch composed of a single image and running kNN searches on CPU.

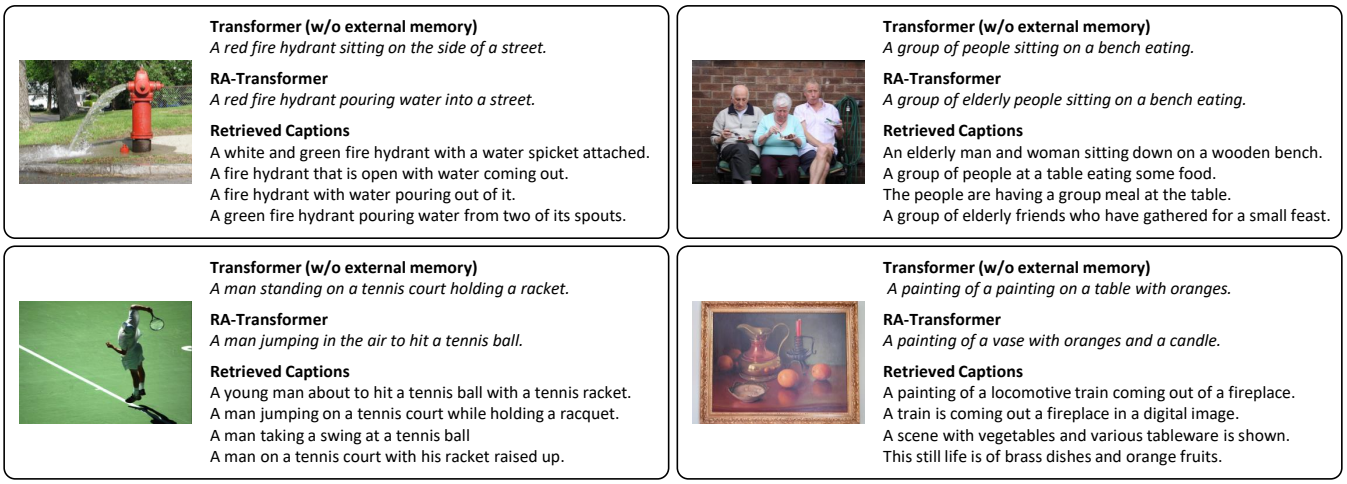

Figure 2: Qualitative results of our model, with and without the use of the external memory, with sample captions retrieved during the generation. In the bottom-right example, we show a failure case of the retrieval component.

attention layer is confirmed also when training with reinforcement learning, with an improvement of 1.2 CIDEr points compared to the standard Transformer-based version. Additionally, we can notice that our model obtains competitive performance compared to other state-of-the-art approaches, surpassing them according to all evaluation metrics.

4.5 Qualitative Results

Finally, in Fig. 2 we show some qualitative results generated by our model and those generated by a Transformer-based model without external memory. For each image, we also report sample captions retrieved from the external memory. As it can be seen, the majority of retrieved captions adequately match the visual content of input images and can provide a helpful external source during the generation process to improve the final results. For example, in the top-right example, we can notice that the predicted caption has very similar content to that of the retrieved sentences (*i.e.* "a group of elderly people"), while the model without external memory fails to generate a detailed description. Similarly, in the bottom-left example, the generated caption contains several words that also appear in the retrieved textual items (*e.g.* "a man jumping" and "a tennis ball"). This further demonstrates the effectiveness of our approach from a qualitative point of view.

In the bottom-right, we instead show an example in which the knowledge retriever partially fails to return textual sentences that effectively describe the visual content of the input image. In fact, while the image contains a painting with a vase, oranges, and other objects, some of the retrieved sentences refer to a painting with a locomotive. This confirms that additional efforts can be done to improve the quality of the retrieval embedding space and that, at the same time, the model maintains the ability to rely on input visual features when the quality of the retrieved captions is poor.

5 CONCLUSION

In this paper, we have presented a retrieval-augmented Transformer for image captioning that integrates kNN-augmented attention layers to generate word tokens based on textual sentences retrieved from an external memory. This enables the model to access an external corpus of textual items during the generation process thus improving the quality of generated captions. Experimental results conducted on the COCO dataset demonstrate the effectiveness of equipping a captioning architecture with retrieval abilities, opening up further research in this direction.

ACKNOWLEDGMENTS

We thank CINECA, the Italian Supercomputing Center, for providing computational resources. This work has been supported by "Fondazione di Modena" 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] 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. *arXiv preprint arXiv:2204.14198* (2022).
[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.
[3] 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 *CVPR*.
[4] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *CVPR*.
[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*.
[6] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis. In *CVPR Workshops*.
[7] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. CaMEL: Mean Teacher Learning for Image Captioning. In *ICPR*.
[8] Roberto Bigazzi, Federico Landi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2020. Explore and Explain: Self-supervised Navigation and Recounting. In *ICPR*.
[9] 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. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426* (2021).
[10] 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 *NeurIPS*.
[11] Marco Cagrandi, Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2021. Learning to Select: A Fully Attentive Approach for Novel Object Captioning. In *ICMR*.
[12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*.
[13] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2020. SMArT: Training Shallow Memory-aware Transformers for Robotic Explainability. In *ICRA*.
[14] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2021. Explaining transformer-based image captioning models: An empirical analysis. *AI Communications* (2021), 1–19.
[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. 2022. Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. *arXiv preprint arXiv:2111.12727* (2022).
[16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *CVPR*.
[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
[18] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.
[19] 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 *ICLR*.
[20] Shiv Ram Dubey, Satish Kumar Singh, and Wei-Ta Chu. 2021. Vision Transformer Hashing for Image Retrieval. In *ICME*.
[21] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. 2021. Training Vision Transformers for Image Retrieval. *arXiv preprint arXiv:2102.05644* (2021).
[22] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and Geometry-Aware Self-Attention Network for Image Captioning. In *CVPR*.
[23] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *ICML*.
[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[25] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image Captioning: Transforming Objects into Words. In *NeurIPS*.
[26] 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 *CVPR*.
[27] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on Attention for Image Captioning. In *ICCV*.
[28] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *EACL*.
[29] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Trans. Big Data* 7, 3 (2019), 535–547.
[30] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
[31] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
[32] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *ICLR*.
[33] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
[34] Federico Landi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2021. Working Memory Connections for LSTM. *Neural Networks* 144 (2021), 334–341.
[35] 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 *NeurIPS*.
[36] 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 *ECCV*.
[37] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshops*.
[38] 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 *ECCV*.
[39] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. 2020. Prophet Attention: Predicting Attention with Future Attention. In *NeurIPS*.
[40] 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).
[41] 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 *AAAI*.
[42] 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 *ICLR*.
[43] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*.
[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.
[45] 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 *ICML*.
[46] 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.
[47] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory optimizations Toward Training Trillion Parameter Models. In *SC*.
[48] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.
[49] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.
[50] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How Much Can CLIP Benefit Vision-and-Language Tasks? *arXiv preprint arXiv:2107.06383* (2021).
[51] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving Image Captioning with Better Use of Captions. In *ACL*.
[52] Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. 2019. Augmenting Self-Attention with Persistent Memory. *arXiv preprint arXiv:1907.01470* (2019).
[53] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations. In *ICLR*.
[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.
[55] 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 *NeurIPS*.
[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.
[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.
[58] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing Transformers. In *ICLR*.
[59] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*.
[60] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2019. Learning to Collocate Neural Modules for Image Captioning. In *ICCV*.
[61] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *ECCV*.
[62] Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive Semiparametric Language Models. *TACL* 9 (2021), 362–373.
[63] 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 *ICLR*.
[64] 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 *CVPR*.
[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.