Memory-Augmented Image Captioning

Zhengcong Fei[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
feizhengcong@ict.ac.cn

Abstract

Current deep learning-based image captioning systems have been proven to store practical knowledge with their parameters and achieve competitive performances in the public datasets. Nevertheless, their ability to access and precisely manipulate the mastered knowledge is still limited. Besides, providing evidence for decisions and updating memory information are also important yet under explored. Towards this goal, we introduce a memory-augmented method, which extends an existing image caption model by incorporating extra explicit knowledge from a memory bank. Adequate knowledge is recalled according to the similarity distance in the embedding space of history context, and the memory bank can be constructed conveniently from any matched image-text set, *e.g.*, the previous training data. Incorporating such non-parametric memory-augmented method to various captioning baselines, the performance of resulting captioners imporves consistently on the evaluation benchmark. More encouragingly, extensive experiments demonstrate that our approach holds the capbility for efficiently adapting to larger training datasets, by simply transferring the memory bank without any additional training.

1 Introduction

Automatic image captioning, which aims to describe a visual content of a given image, is a core topic in the artificial intelligence area (Bai and An 2018; Fei 2020a). There is a boom in research on image captioning systems due to the advance of deep learning technology, and most existing models adopt encoder-decoder frameworks (Vinyals et al. 2015; Xu et al. 2015; Yao et al. 2017; Anderson et al. 2018; Huang et al. 2019; Fei 2020b). Technically, CNN-based image encoder extracts sufficient and useful visual features from the input image; RNN-based caption decoder builds the semantic part according to the picked visual information and decodes it word by word. The above structures have been shown to learn a substantial amount of in-depth relational knowledge from training data using parameter optimization, without access to external memory. While this development is exciting, such image captioning models do have some drawbacks (Fei 2019; Wang et al. 2020): they cannot easily expand or update their prior memory, and can not straightforwardly provide insight into their current predictions.

To address these issues, numerous hybrid captioning models that combine with the retrieval-based memory mechanism are leveraged, in which outer recalled knowledge can be directly revised and expanded, and its access can be inspected and interpreted (Weston, Chopra, and Bordes 2014). In particular, inspired by the fact that humans benefit from previous similar experiences when taking actions and related examples from training data provide exemplary information when describing a given image, previous works (Poghosyan and Sarukhanyan 2017; Chen et al. 2019; Wang et al. 2020) usually first utilize an image-text matching model to retrieval top-k similar sentence candidates. Then, the target caption will be created under the guide of input image plus these related candidates with a specially designed network. Although current retrieval-based captioning models have achieved promising results, there still have the following weakness: 1) Their performance is limited by the quality of the caption retrieved model. Commonly, retrieval results are less coherent and relevant with the query image than generative models'. Irrelevant retrieved results would even mislead the final caption generation. 2) These models can only make use of individual sentence-level retrieved results, leading to a high variance in the performance (Zhang and Lu 2018). Moreover, the information from very few retrieved results may not be sufficient to enrich the caption decoding. Compared with methods matching based only on image features, our proposed word-level retrieval mechanism considers the complete history information, including given image and previously generated words, which results in a more accurate and comprehensive knowledge application.

In this paper, we introduce a memory augmented approach that equips a trained image caption decoding with considering extra word-level knowledge information, in other words, linearly interpolating the original next word distribution with a top-k matching approximation. The related knowledge in the memory bank is recalled according to the similarity distance in the embedding space of history context (*i.e.*, the image feature and prefix of the caption) and can be drawn from any image-text collection, including the original training data or other extended datasets. We assume that contxts which are close in representation space are more likely to occur the same word. In this manner, our framework does not incorporate additional parameters and allow effective knowledge to be memorized explicitly and interpretably,

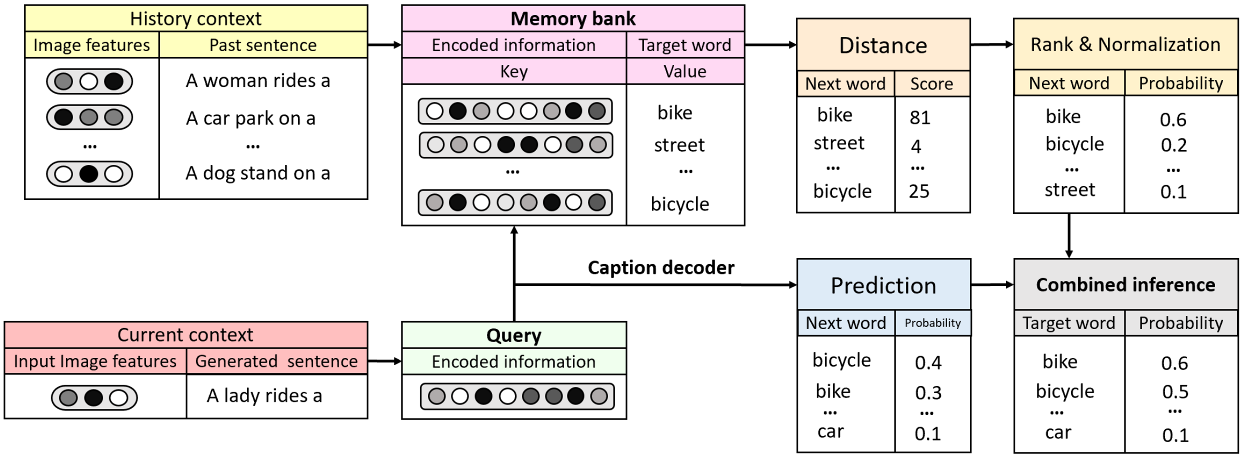


Figure 1: Illustration of our memory augmented caption generation. A memory bank is constructed based on pre-set matched image-text samples, including encoding of its history context, *i.e.*, image features and past sentence (key) and target word (value). During inference, a current context is encoded (query), and the k most similar matches are retrieved from the memory bank. Then, a distribution over the vocabulary is computed with rank and normalization operation. Finally, the distribution is interpolated with the original model's prediction for combined decision. Note that the encoder of the query context is identical to the encoder of the memory bank.

rather than implicitly in the model parameters. To better measure its effects, we conduct an extensive empirical evaluation on the MS COCO benchmark (Chen et al. 2015). Built upon recent stronge captioners with our memory augmentation mechanism showing a prominent improvement over the base when the same training set is employed for modeling the history memory representations. We also demonstrate that our approach holds the capacity for efficiently adapting to larger training datasets, by simply reconstructing the memory bank with the existing image captioning model.

The contributions of this work are as follows:

- We propose a memory-augmented approach which extends the decision of current image captioning model with related knowledge from the no-parametric memory bank. As far as we are concerned, this is the first work to build word-level knowledge from image-text pairs using a trained captioning model, and to use the memory to further enhance the performance of caption generation.
- Extensive experiments demonstrate that captioning models equipped with memory augmented mechanism significantly outperform the ones without it. We also analyze the effect of memory bank scale. More encouragingly, the proposed memory mechanism can be easily incorporated into existing captioning models to improve their performance without additional training.

2 Approach

In this section, we first give a brief review of the implementation for the conventional attention-based encoder-decoder framework in image captioning (Xu et al. 2015; Rennie et al. 2017). This structure is regarded as the state-of-the-art model and will be used as the baseline in this study. Then we introduce the memory augmented mechanism for next word prediction in detail. Finally, we provide a discussion about computational cost as well as other related works.

2.1 Background: Attention-based Encoder-Decoder Paradigm

Overall, two-stage image captioning systems usually consist of an image encoder and a language decoder.

Image Encoder For each input image, a pre-trained Faster-RCNN (Ren et al. 2015) is utilized to detect region-based objects. Here, the top N objects with highest confidence scores are selected, and we denote the corresponding extracted feature vectors as $V = \{v_1, v_2, \ldots, v_N\}$, where $v_n \in \mathbb{R}^{d_v}$, and d_v is the dimension of each feature vector. Note that each feature vector represents a certain aspect of the input image and further serves as a guide for sentence decoder to describe the material visual information.

Caption Decoder During each decoding step t, the sentence decoder takes the word embedding of current input word w_{t-1}, concatenated with the average of extracted image features $\overline{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$ as input to the decoding network as:

$$h_t = f_D(h_{t-1}, [W_e w_{t-1}; \overline{v}]), \tag{1}$$

where $[;]$ is the concatenation operation, W_e denotes the learnable word embedding parameters, and $f_D(\cdot)$ is the decoder network, *e.g.*, LSTM (Hochreiter and Schmidhuber 1997) and Transformer (Vaswani et al. 2017). Next, the output state h_t of the decoding function is utilized as a query to attend to the relevant image regions in the image feature set V and generate the weighted image features, also named as context vector, c_t as:

$$\alpha_t = \text{Softmax}(w_\alpha \text{tanh}(W_h h_t \odot W_V V)), \tag{2}$$

$$c_t = V\alpha_t^T, \tag{3}$$

where w_α, W_h and W_V denote the learnable parameters. $\odot$ denotes the matrix-vector addition, which is calculated by adding the vector to each column of the matrix. Finally, the hidden state h_t and context features c_t are passed to a linear layer together to predict the next word:

$$w_t \sim p_t = \text{Softmax}(W_p[h_t; c_t] + b_p), \tag{4}$$

where W_p and b_p are the learnable parameters. It is worth noticing that some works (Anderson et al. 2018; Yao et al. 2018) also attempt to append more neural network modules, *e.g.*, extra LSTM and GCN, to assist to predict the next word. For training procedure, given a ground-truth description sentence $S^*_{1:T} = \{w^*_1, \ldots, w^*_T\}$ and a captioning model P_{IC} with parameters θ, the optimization objective is to minimize the cross-entropy loss as follows:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log P_{IC}(w^*_t | S^*_{<t}; \theta). \tag{5}$$

Also, as we can see from Equation 3, at each time step t, the context vector c_t contain the past information including image features and generated caption words, which can further be used as the digital certificate, in other words, the key in our memory bank.

2.2 Memory-Augmented Caption Generation

Conventional image captioners try to model a conditional probabilities to each sentence. Specifically, given the entire input image features as well a sequence of previously generated words, denoted as $C_t = \{V, w_1, \ldots, w_{t-1}\}$, image captioniners estimate $P_{IC}(w_t|C_t; \theta)$, the distribution of the next word $w_t \in \mathcal{V}$ over the entire vocabulary for each word in sentence. The caption is created word-by-word. Based on this paradigm, our method equips a trained image captioner with a memory retrieval mechanism, allowing the model access to the most useful explit knowledge easily at each time step, as show in Figure 1.

Memory Bank Construction The memory bank is constructed offline and consists of a set of key-value pairs (k_i, v_i). 1) The key k_i is a representation of the entire caption context C_i computed by an mapping function $f_M(\cdot)$. In actual, for an LSTM–based captioning architecture, $f_M(C_i)$ could result from attended image features or context vector; for a Transformer-based captioning architecture, $f_M(C_i)$ could obtain from an intermediate representation that is output by a self-attention layer. 2) The value v_i is the corresponding ground truth word w_i. For a parallel image-text dataset $\mathcal{D}$, potentially refers to the original training set, the representation can be generated with a single forward pass over each sample by a caption decoder and the total memory bank $(\mathcal{K}, \mathcal{V})$ can be formulated as follows:

$$(\mathcal{K}, \mathcal{V}) = \{(f_M(C_i), w_i) | (C_i, w_i) \in \mathcal{D}\}. \tag{6}$$

Algorithm 1: Memory Augmented Image Caption Generation

```
Input: Memory bank ($\mathcal{K}$, $\mathcal{V}$), trained captioning
       model $P_{IC}(w|C;\theta)$, given image $I$
Output: Descriptive sentence $S = \{w_1, \ldots, w_T\}$
for $t = 1$ to $T$ do
    Specify current context: $C_t = \{I, w_1, \ldots, w_{t-1}\}$;
    Generate original distribution with caption
      decoder: $P_{IC}(w_t|C_t;\theta)$;
    Generate encoded query vector: $q_t = f_M(C_t)$;
    Calculate the distance for each entries in memory
      bank: $d_i = dis(q_t, k_i)$;
    Select top-$k$ candidates, normalize and aggregate
      the corresponding distribution:
      $P_{MA}(w|C_t) \propto \sum_{(k_i,v_i)\in\mathcal{P}} \mathbb{I}_{w=v_i} \text{exp}(-d_i/T))$;
    Combined inference to decide the next word:
      $w_t =$
      $\text{argmax}(\lambda P_{MA}(w|C_t) + (1-\lambda)P_{IC}(w|C_t;\theta))$
end
```

Quick Query According to the previous description, the memory bank contains entries for target word with a corresponding history context in the retrieval data set, which for image captioning can be up to billions of examples. To search over this large memory bank rapidly, we adopt FAISS (Johnson, Douze, and Jegou 2019), an open-source library for fast k-min selection retrieval in high dimensional spaces with GPUs. Specifically, for search acceleration, FAISS clusters the keys and only queries the adjacent cluster centroids; for space usage, FAISS stores the compressed versions of the high-dimension vectors. Preliminary experiments have demonstrated that using L2 distance (Xu, Weinberger, and Chapelle 2012) for FAISS retrieval results in a better performance for image captioning, compared to inner product distance.

Combined Inference During inference, according to the current history context C_t, including the total input image features and preceding generated subsentence, the image captioner outputs a distribution over the vocabulary $P_{IC}(w|C_t)$ with caption decoder. The model also produces the context vector $f_M(C_t)$ and queries the memory bank with it to retrieve top-k similar pairs $\mathcal{P}$ according to the value of distance function $dis(\cdot,\cdot)$. Next, it computes a normalization distribution over approximates based on a softmax of their negative distances with pre-set temperature T to prevent overfitting, while aggregating over multiple occurances of the same word. Note that items that do not appear in the retrieved targets are set zero probability.

$$P_{MA}(w|C_t) \propto \sum_{(k_i,v_i)\in\mathcal{P}} \mathbb{I}_{w=v_i} \text{exp}(\frac{-dis(k_i, f(C_t))}{T}). \tag{7}$$

Finally, we interpolate the information retrieval distribution P_{MA} with the previous model generation distribution P_{IC}, which is more robust in cases without sufficient recalling, using a balancing parameter λ to decide the final decision

	Cross-Entropy Loss						**CIDEr Score Optimization**					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
State-of-the-art image captioning models												
LSTM-A (Yao et al. 2017)	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
RFNet (Jiang et al. 2018)	76.4	35.8	27.4	56.8	112.5	20.5	79.1	36.5	27.7	57.3	121.9	21.2
Up-Down (Anderson et al. 2018)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
GCN-LSTM (Yao et al. 2018)	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
AoANet (Huang et al. 2019)	77.4	37.2	28.4	57.5	119.8	21.3	80.2	38.9	29.2	58.8	129.8	22.4
M2-T (Cornia et al. 2020)	-	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2	22.6
Retrieval-based hybrid image captioning models												
GST (Jia et al. 2015)	67.0	26.4	22.7	-	81.25	-	-	-	-	-	-	-
Mem-Att(Chen et al. 2018)	-	-	-	-	-	-	75.7	35.0	-	55.7	109.2	-
ICMK(Chen et al. 2019)	-	-	-	-	-	-	81.9	38.4	28.7	58.7	125.5	-
Up-Down + SRT (Wang et al. 2020)	77.1	36.6	28.0	56.9	116.9	21.3	80.3	38.5	28.7	58.4	129.1	22.4
Our memory-augmented image captioning models												
Up-Down†	76.2	36.0	27.2	56.3	113.5	20.1	79.2	36.5	27.7	57.3	120.8	21.2
Up-Down† + MA	77.1	37.1	28.3	57.2	116.3	21.3	80.2	37.5	28.4	58.2	125.4	22.0
AoANet†	77.3	36.9	28.5	57.3	118.5	21.6	80.5	39.1	29.0	58.9	128.9	22.5
AoANet† + MA	78.2	38.0	28.7	57.8	121.0	21.8	81.2	39.5	29.4	59.3	132.0	23.0

Table 1: Performance of our model and other state-of-the-art captioning methods with different evaluation metrics on the MS COCO Karpathy test set. All values are reported as a percentage (%). The original results of these methods from their publications are listed in the top block. The proposed memory augmented approach significantly improves across all the metrics using both cross-entropy loss and CIDEr optimization. † denotes our trained model based on the publicly available source code.

distribution as:

$$P(w_t|C_t) = \lambda P_{MA}(w_t|C_t) + (1-\lambda)P_{IC}(w_t|C_t;\theta). \quad (8)$$

More detailed procedure for predicting the next word at each step is presented in Algorithm 1.

2.3 Discussion

Computational Cost Although the proposed memory-augmented mechanism requires no additional training with an existing image captioning model, it does produce some other computational overheads in theory. The primary cost of buidling the memory bank is generating the keys and values, which requires a single forward across the whole training set and is identical to a fraction of the cost of training for one epoch on the same examples. Once the keys are saved, for the MS COCO dataset, building the cache with 328M entries takes roughly one hour on a single 1080Ti GPU. Considering the fact that the cost of building a large cache grows linearly in the number of entries, our method is almost negligible for the increase in computational burden. During inference, retriving 512 keys from the memory bank results in a around $\times 3$ slower than the base captioner.

Related Memory-based Models The idea of memory augmentation was inspired by the advances in the memory network (Weston, Chopra, and Bordes 2014; Xiong, Merity, and Socher 2016). These models equip neural networks with an external memory module that can be accessed and manipulated via some trainable operations. The memory idea has been utilized in image captioning task in recent years. The early pioneering work (Jia et al. 2015) introduces an extension of LSTM to stay on track and better describe the image content without unrelated phrase. (Chen et al. 2018; Poghosyan and Sarukhanyan 2017) stores the visual and semantic knowledge in the past into memories and generate a global feature to improve the attention model. (Chen et al. 2019) further introduces a selective reading mechanism to retrieve past knowledge information. In these cases, the contribution of the memory is to provide temporary variables to assist caption decoding. In contrast, our work uses memory to store knowledge. The memory in these works could be considered to be notes, while the memory in our work is more like a regularized dictionary. On the other hand, (Wang et al. 2020) introduces a recall mechanism, which includes a recall unit to retrieval words for image and a semantic guide and slot to use the recalled words. The text-retrieval module is sentence-level and identical to solve the image-text matching task, which searches the sentences only based on image content. Comparatively, our retrieval mechanism is no need to extra training, includes all useful context information to make retrieval more accurately and can be adapted to any other dataset directly. Simlar retrieval-based works in other areas include (Khandelwal et al. 2019; Guu et al. 2020; Lewis et al. 2020). Instead of only drawing text embedding, we forcus on cross-modality knowledge constructing and usage. This means that the representation of key and query must incorporate both image and txt information.

	B-1		B-2		B-3		B-4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Up-Down	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
AoANet	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M2-T	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
Mem-Att	75.5	92.7	59.2	85.2	45.5	75.4	34.8	64.8	27.2	36.7	55.8	71.4	106.9	106.7
ICMK	80.8	95.3	64.3	89.0	49.5	79.8	37.5	69.7	28.0	36.9	57.9	73.0	118.9	121.5
AoANet + MA	81.7	96.2	66.5	90.9	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4

Table 2: Leaderboard of different image captioning methods on the online MS COCO test server.

3 Experiments

3.1 Experimental Setup

Dataset We utilize the most popular image captioning dataset MSCOCO (Chen et al. 2015) to evaluate the performance of our proposed method. As the largest English image caption dataset, MSCOCO contains 164,062 images. Each image is annotated with five human captions. Considering the annotations for the official testing set are not provided, in this paper, we follow the common practice as Karpathy splits (Karpathy and Fei-Fei 2015) for validation of model hyperparameters and offline evaluation. This split contains 113,287 images for training and 5,000 respectively for validation and test. We also pre-process all training sentences by converting them into lower case and dropping the words that occur rarely as (Huang et al. 2019).

Evaluation Metrics For quantitative performance evaluation, we use five standard automatic evaluation metrics simultaneously, namely BLEU-N (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016), denoted as B-N, M, R, C and S for simplify. Concretely, BLEU-N indicates the n-gram precision of the generated caption, METEOR measures both the precision and recall, and CIDEr considers the n-gram similarity with TF-IDF weights.

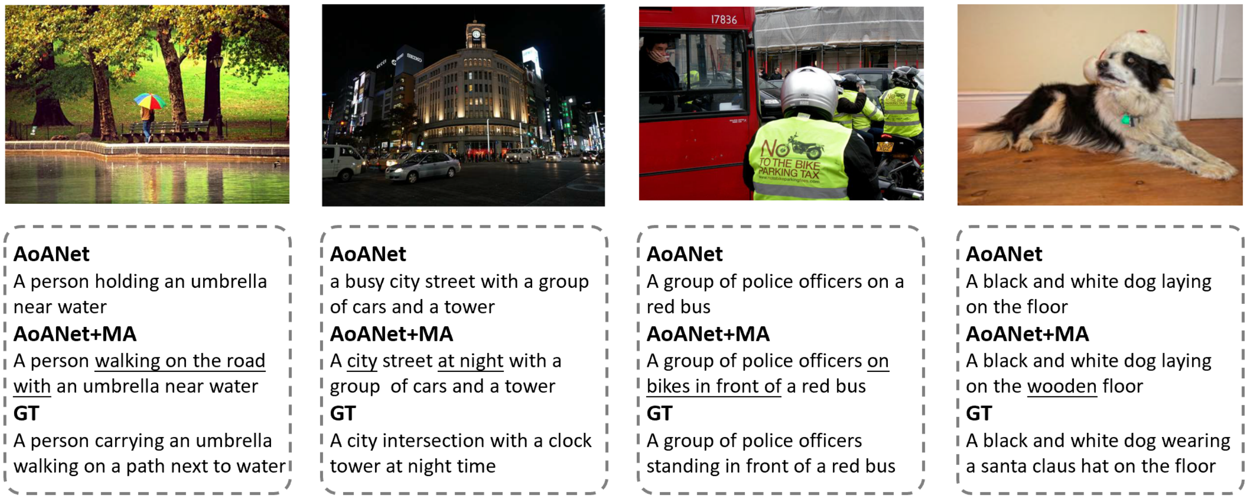
Baselines We equip the proposed memory augmented method with different state-of-the-art methods, including Up-Down (Anderson et al. 2018) and AoANet (Huang et al. 2019). Since the memory mechanism makes no changes to the baseline, we take the exact architecture and optimization described by the corresponding paper. Both image captioning are first trained to minimize the negative log-likelihood of the training data and then fine-tuned with the CIDEr score using Reinforcement Learning (Rennie et al. 2017). All the region's visual features are extracted with Faster-RCNN on the backbone of ResNet-101. The specific variations will be discussed in the following sections.

Implement Details Since our approach aims to incorporate extra knowledge to assist captioning decision and is augmentative to the existing models, we keep the inner structure of the baseline untouched and preserve the original settings. Following (Anderson et al. 2018), the keys used for knowledge retrieval are the 1024-dimensional representations copied from context vectors. We perform a single forward pass over the total training set with the trained captioning model, in order to create the keys and values. A FAISS index is then created using 1.5M randomly sampled keys to learn 2K cluster centroids, and keys are quantized to 64-bytes. During inference, we query the memory with k = 512 most similar entries, and the index looks up 32 cluster centroids while searching for the next word candidates. The tempreture T is set to 100 and the balancing parameter λ is selected based on the CIDEr score on the validation set.

3.2 Quantitative Analysis

Offline Evalaution In Table 1, we report the performance comparisons between state-of-the-art models, conventional retrieval-based models, and baselines incorporated with our memory augmented (MA) method on the offline MS COCO Karpathy test split. For a fair comparison, we report the results for each run optimized with both cross-entropy loss and CIDEr score. Note that the memory bank is constructed with the same data used to train the baseline. In general, all baselines equipped with our memory-augmented approach receive significant performance gains overall metrics and outperforms the popular retrieval-based methods under the same structure. More encouragingly, based on the AoANet, which are the previous state-of-the-arts on MS COCO datasets, our approach sets a new comparative performance, achieving 132.0 CIDEr score and makes the absolute improvement over the baseline AoANet by 3.1%, demonstrating the effectiveness and the compatibility of our proposed approach. On the other hand, to fully verify the generalizability of our memory mechanism for image captioning, we include two variants of our approach by plugging combined decisions into both LSTM-based and Transformer-based encoder-decoder structure. Table 1 also witnesses continuous performance boosting and illustrates the advantage of exploiting adequate knowledge via memory bank for image captioning again.

Online Evaluation In addition, following the common practice (Huang et al. 2019; Wang et al. 2020), we also evaluate our best variants AoANet + MA on the official testing set by submitting the ensemble versions, *i.e.*, an average ensemble for four checkpoints trained independently, to



AoANet
A person holding an umbrella near water
AoANet+MA
A person walking on the road with an umbrella near water
GT
A person carrying an umbrella walking on a path next to water

AoANet
a busy city street with a group of cars and a tower
AoANet+MA
A city street at night with a group of cars and a tower
GT
A city intersection with a clock tower at night time

AoANet
A group of police officers on a red bus
AoANet+MA
A group of police officers on bikes in front of a red bus
GT
A group of police officers standing in front of a red bus

AoANet
A black and white dog laying on the floor
AoANet+MA
A black and white dog laying on the wooden floor
GT
A black and white dog wearing a santa claus hat on the floor

Figure 2: Case studies of baseline AoANet, plus our memory augmented method (AoANet + MA), coupled with the corresponding ground truth sentences (GT).

the online testing server. The results over official testing images with 5 reference captions (c5) and 40 reference captions (c40) of our approach, the top-performing published works, and other memory-based methods on the leaderboard are reported in Table 2. The results clearly show that compared to all the other popular captioning systems, our AoANet + MA exhibit better performances cross over most metrics.

3.3 Qualitative Analysis

Figure 2 showcases several image captioning results of AoANet and our memory-augmented approach AoANet + MA, coupled with human-annotated ground truth sentences (GT). Generally, compared with the captions of AoANet, which are somewhat relevant to image content and logically correct, our memory augmenting method produces more accurate and rich descriptive sentences by exploiting extra explicit knowledge. What's more, the baseline has errors when generating some captions. In contrast, our method corrects these mistakes in a human-like format. For example, AoANet generates the phrase of "on a red bus" that is inconsistent with the visual content and common sense for the third image, while "on bikes in front of a red bus" in our AoANet depicts the visual content more precise. This again confirms the advantage of capturing related knowledge and applies it during inference via our memory-enhancing module. On the other hand, we also observe that some examples include factual knowledge and near-duplicate sentences from the training set. In these cases, assigning train and test instances similar representations appears to be a more common issue than implicitly memorizing the next word with model parameters.

3.4 Ablation Study

Training Data	50%		75%		100%	
Memory Data	100%		100%		100%	
Metrics	B-4	C	B-4	C	B-4	C
Up-Down	35.2	112.6	35.7	113.0	36.0	113.5
Up-Down + MA	36.8	115.3	36.9	115.8	37.1	116.3
AoANet	36.2	116.0	36.7	118.0	36.9	118.5
AoANet + MA	37.8	120.5	38.0	121.2	38.0	121.0

Table 3: Evaluation of baseline and our memory-augmented method by training on a subset of MS COCO training split while constructing the memory bank with trained captioning model on the whole training split.

To fully exam the impact of the proposed memory augmented method, we conduct an ablation study by comparing against a set of other ablated models with various settings.

Effect of Memory Data Size The previous section has demonstrated that recalling similar and useful knowledge from the memory bank can significantly improve the image captioning performance. This raises the question: can the memory directly transfer from data that is larger and not trained on? To answer this question, we further estimate the performance of captioner equipped with our proposed method for a subset of the training data while keeping the memory bank construction complete, where $x\%$ denotes the percentage of the total data that is used for training. All these subsets of the training samples are selected randomly. According to the results shown in Table 3, conforming to our common sense, the conventional image captioning model trained on 100% data apparently outperforms the identical captioner trained on smaller data. Concretely, boosting the CIDEr score from 116.0 to 118.5 on the AoANet baseline, validates the advantage of big training data size. On the other

k	B-4	M	R	C	S
0	36.9	28.5	57.3	118.5	21.6
1	37.1	28.5	57.4	119.4	21.6
2	37.3	28.3	57.4	119.8	21.6
8	37.5	28.5	57.5	120.3	21.8
64	37.5	28.6	57.6	120.5	21.7
256	37.8	28.7	57.8	121.0	21.8
512	38.0	28.7	57.8	121.0	21.8

Table 4: Effect of the number of retrieved knowledge entries per next word of AoANet + MA on MS COCO validation set. Recalling more entries from the memory bank monotonically improves the captioning performance.

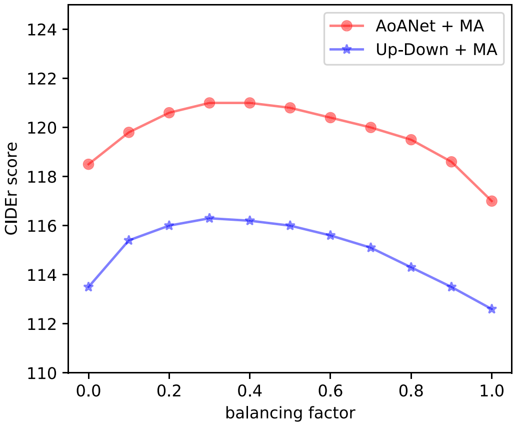


Figure 3: Evaluation results of different balancing factor λ values on MS COCO validation set for different baselines.

hand, adding relevant knowledge retrieval over the total examples to the model trained on the total size benefits the performance a lot; *i.e.*, retrieving related knowledge explicitly from the data set outperforms training on it implicitly sometimes. This phenomenon reveals that rather than training captioning models on ever-larger datasets, we can utilize smaller datasets to learn generic representations and augment them with memory bank for fast adapting.

Effect of Number of Retrieved Knowledge Entries per Query For memory augmenting, the number of retrieved knowledge entries k, is fixed to 512. Here, we investigate the effect of the number of related knowledge entries per query, and the results are listed in Table 4. We can observe that: 1) by incorporating the memory augmented method, all of the evaluation scores are increased, demonstrating the effectiveness of the memory mechanism. 2) The captioning performance monotonically improves as more related knowledge entries are recalled, and suggests that even larger improvements may be possible with a higher value of recalled number k. 3) To make a trade-off between accuracy and speed, a small number of retrieved entries per query, that is, k = 8, is enough to achieve a competitive performance.

	AoANet wins	Tie	AoANet + MA wins
Naturalness	24.8	44.0	31.2
Relevance	26.3	46.7	27.0
Richness	21.4	40.2	38.4

Table 5: Results of human evaluation in terms of various metrics. All values are reported as a percentage (%).

Effect of Balancing Parameter In Equation 8, we integrate a parameter λ to interpolate between the base model distribution P_{IC} and the distribution P_{MA} from memory augmented module over the dataset. In this section, we tuned our system on the training set and used an enumeration search on baseline Up-Down + MA and AoANet + MA to determine the optimal parameter. The CIDEr scores were measured under different parameter values, as shown in Figure 3. We can observe that when $\lambda \sim 0.3$, both model achieves the highest metric values. As λ goes down from 0.3 to 0 or goes up from 0.3 to 1, the CIDEr score drops moderately. Therefore, we set the λ = 0.3 by default in this paper.

3.5 Human Evaluation

To better understand the effectiveness of the memory augmented method, we also conduct a human evaluation to measure the quality of generated captions. We randomly select 200 samples from the MS COCO datasets along with human-annotated sentences. We recruit 8 workers to compare the perceptual quality of the caption between our memory-based approach and baselines independently in four aspects: naturalness, which indicates the grammaticality and fluency; relevance, which indicates the connection with the given image content; richness, which measures the amount of significant information contained in the sentence. The results are shown in Table 5. We can see that our memory-augmented approach wins in all metrics than baselines. In particular, AoANet + MA achieves more than 17.0 score in richness. This again confirms that the proposed memory augmented method holds the superiority to provide more accurate and abundant descriptions.

4 Conclusion

Inspired that similar context are more likely to be created with same word, in this paper, we introduce a simple and effective memory-augmented method for image captioning, which exploits explicit knowledge and helps to improve trained captioners by directly querying the memory bank at inference time. In particular, the construction of a retrieval-based memory bank is non-parametric as well as no need further training. Extensive experiments conducted on the MS COCO benchmark prove that our memory-augmented mechanism can effectively utilize history information to consistently improve the generated caption quality. More remarkably, the proposed memory augmented mechanism is compatible with any captioning model that can produce fixed-size context representations.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proc. ECCV*, 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proc. IEEE CVPR*, 6077–6080.

Bai, S.; and An, S. 2018. A survey on automatic image caption generation. *Neurocomputing* 311: 291–304.

Chen, H.; Ding, G.; Lin, Z.; Guo, Y.; and Han, J. 2018. Attend to knowledge: memory-enhanced attention network for image captioning. In *Proc. ICBICS*, 161–171. Springer.

Chen, H.; Ding, G.; Lin, Z.; Guo, Y.; Shan, C.; and Han, J. 2019. Image captioning with memorized knowledge. *Cognitive Computation* 1–14.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *Proc. IEEE CVPR*, 10578–10587.

Fei, Z. 2020a. Actor-critic sequence generation for relative difference captioning. In *Proc. ICMR*.

Fei, Z. 2020b. Iterative Back Modification for Faster Image Captioning. In *Proc. ACM Multimedia*, 3182–3190.

Fei, Z.-c. 2019. Fast Image Caption Generation with Position Alignment. *arXiv preprint arXiv:1912.06365* .

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909* .

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proc. IEEE ICCV*, 4634–4643.

Jia, X.; Gavves, E.; Fernando, B.; and Tuytelaars, T. 2015. Guiding the long-short term memory model for image caption generation. In *Proc. IEEE ICCV*, 2407–2415.

Jiang, W.; Ma, L.; Jiang, Y.-G.; Liu, W.; and Zhang, T. 2018. Recurrent fusion network for image captioning. In *Proc. ECCV*, 499–515.

Johnson, J.; Douze, M.; and Jegou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Trans. on Big Data* .

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE CVPR*, 3128–3137.

Khandelwal, U.; Levy, O.; Jurafsky, D.; Zettlemoyer, L.; and Lewis, M. 2019. Generalization through Memorization: Nearest Neighbor Language Models. In *Proc. ICLR*.

Lavie, A.; and Agarwal, A. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. ACL Workshop*, 228–231.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401* .

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of summaries. 74–81.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, 311–318.

Poghosyan, A.; and Sarukhanyan, H. 2017. Short-term memory with read-only unit in neural image caption generator. In *Proc. IEEE CSIT*, 162–167. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NIPS*, 91–99.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *Proc. IEEE CVPR*, 1179–1195.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proc. NIPS*, 5998–6008.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proc. IEEE CVPR*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proc. IEEE CVPR*, 3156–3164.

Wang, L.; Bai, Z.; Zhang, Y.; and Lu, H. 2020. Show, Recall, and Tell: Image Captioning with Recall Mechanism. In *Proc. AAAI*, 12176–12183.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916* .

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *Proc. ICML*, 2397–2406.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. ICML*, 2048–2057.

Xu, Z.; Weinberger, K. Q.; and Chapelle, O. 2012. Distance metric learning for kernel machines. *arXiv preprint arXiv:1208.3422* .

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring Visual Relationship for Image Captioning. In *Proc. ECCV*, 684–699.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *Proc. IEEE CVPR*, 4894–4902.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proc. ECCV*, 686–701.