AMA: Adaptive Memory Augmentation for Enhancing Image Captioning

Shuang Cheng [1,2] Jian Ye [1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China [2]University of Chinese Academy of Sciences

BMVC 2023

Introduction

Motivation

1. Memory-Augmented Image Captioning (MA-IC) has demonstrated significant performance improvements over standard neural image captioning systems. It effectively combines a well-trained captioning model with additional explicit knowledge from a memory bank to enhance captioning accuracy.
2. The k-nearest neighbor algorithm used in MA-IC retrieves the same number of nearest neighbors for each target token, which may lead to prediction errors when the retrieved neighbors contain noise.

Contribution

1. The adaptive memory-augmented (AMA) approach uses a lightweight network to adaptively filter neighbor information for each target token, eliminating the need for extensive training samples.
2. When applied to existing baselines, AMA significantly enhances performance on the COCO benchmark dataset, outperforming models without memory augmentation.
3. Further analysis reveals that this memory mechanism can be seamlessly integrated into existing captioning models, improving performance with minimal extra training, thus showcasing its practicality and efficiency.

Background

Memory-Augmented Image Captioning

A memory-augmented image captioning system consists of two main phases: constructing a memory bank and making predictions based on it.

1. **Memory Bank Construction:** The memory bank is a set of key-value pairs, where each key is a semantic embedding of an image-text sample computed by a mapping function $f(\cdot)$, and the corresponding value is the ground truth word y_t. The memory bank $\mathcal{D} = (\mathcal{K}, \mathcal{V})$ encompasses all key-value pairs constructed from the entire training examples.

$$\mathcal{D} = (\mathcal{K}, \mathcal{V}) = \{(f(x, y_{<t}), y_t) | \forall y_t \in \mathcal{Y}, (x, y) \in (\mathcal{X}, \mathcal{Y})\}$$

2. **Combined Inference:** During the inference stage, the system calculates the context embedding $f(x, y_{<t})$, retrieves k nearest neighbors from the memory bank $\mathcal{D}$, and aggregates the retrieved tokens to form the distribution $P_{MA}(y_t \mid x, y_{<t})$.

$$P_{MA}(y_t \mid x, y_{<t}) \propto \sum_{(k_i, v_i) \in \mathcal{D}} I_{y_t = vi} \exp\left(\frac{-dis(k_i, f(x, y_{<t}))}{T}\right)$$

The final probability is derived as an interpolation of the image captioning model's distribution $P_{IC}(y_t \mid x, y_{<t})$ and $P_{MA}(y_t \mid x, y_{<t})$.

$$P(y_t \mid x, y_{<t}) = \lambda P_{MA}(y_t \mid x, y_{<t}) + (1 - \lambda) P_{IC}(y_t \mid x, y_{<t})$$

where the fixed weight λ balances the two distributions.

Method

Dynamic Memory-Augmented Image Captioning

The traditional MA-IC method presents limitations, such as being vulnerable to noise due to the reliance on k nearest neighbors and employing a fixed weight parameter λ. To address these, we propose 1) adaptively leveraging information from varying numbers of neighbors, and 2) incorporating a learnable network to determine the weights for different target tokens adaptively.

Method

1. **Multiple kNN Classifiers:** We employ information from multiple kNN classifiers based on the current retrieval results. We consider a set of k-values, and $k = 0$ represents the distribution of the IC model. We input the features, constructed using the results of each kNN classifier, into a lightweight network to determine the corresponding interpolation weight.

$$\mathcal{S} = \{0\} \cup \{k \in N \mid \log_2 k \in N, k \leq K\}$$

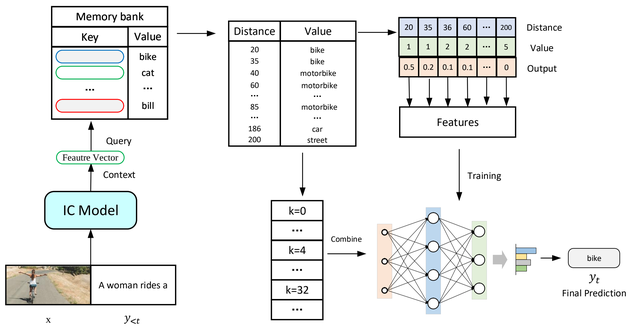

Figure 1. Adaptive Memory-Augmented Image Captioning

2. **Importance Measurement:** We employ a lightweight network to estimate the importance of different distributions. We design three features: distance, count, and output. The distance between q_t and each neighbor, the distribution of target tokens in the retrieval results, and the IC output are all considered. The normalized weights for each available k are computed as:

$$P_\beta(k) = \text{softmax}(f_\beta([d : c : o]))$$

Where $k \in \mathcal{S}$, and $k = 0$ corresponds to the distribution generated directly by the IC model.

Prediction

We eliminate the fixed hyper-parameter λ and the final prediction probability becomes a weighted ensemble of different kNN predictions combined with the output of the IC model:

$$P(y_t \mid x, y_{<t}) = \sum_{k \in \mathcal{S}} P_\beta(k) \cdot P_{k\text{NN}}(y_t \mid x, y_{<t})$$

Where P_{kNN} denotes the k nearest neighbor probability. We then derive the final predicted word as follows:

$$w_t = \text{argmax}(P(y_t \mid x, y_{<t}))$$

Experience

Performance comparison

	Bleu-1	Bleu-4	Meteor	Rouge-L	CIDEr-D	Spice
State-of-the-art models						
M2[2]	80.8	39.1	29.2	58.6	131.2	22.6
RSTNet[6]	81.1	39.3	29.4	58.8	133.3	23.0
LEMON$_{base}$[4]	82.1	40.3	30.2	59.8	133.3	23.3
OFA$_{base}$[5]	**82.5**	**41.0**	**30.9**	**60.2**	**138.2**	**24.2**

Experience

memory-utilized models						
ICMK[1]	81.9	38.4	28.7	58.7	125.5	-
M2+MA[3]	80.9	39.3	29.3	58.7	132.0	22.7
RSTNet+MA[3]	81.2	39.7	29.5	59.0	134.0	23.1
LEMON$_{base}$+MA[3]	82.5	40.5	30.3	60.1	134.7	23.4
OFA$_{base}$+ MA[3]	**82.9**	**41.2**	**31.0**	**60.8**	**138.5**	**24.4**
Our adaptive memory-augmented models						
M2+UMA	80.7	39.2	29.2	58.5	131.7	22.5
RSTNet+UMA	81.0	39.6	29.4	58.8	133.4	23.0
LEMON$_{base}$ +UMA	82.4	40.4	30.2	59.9	133.5	23.3
OFA$_{base}$+ UMA	82.7	41.1	30.9	60.4	138.3	24.2
M2+AMA	81.1	39.8	29.5	58.8	133.4	22.9
RSTNet+AMA	81.6	40.3	29.6	59.3	135.2	23.3
LEMON$_{base}$ +AMA	82.9	40.6	30.4	60.3	136.3	23.5
OFA$_{base}$+ AMA	**83.1**	**41.3**	**31.1**	**61.1**	**138.8**	**24.5**

Table 1. Performance comparison with baseline methods.

Effectiveness of K and hidden size

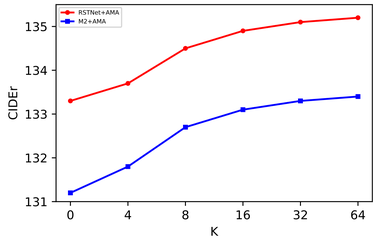

Figure 2. Effect of the number of K.

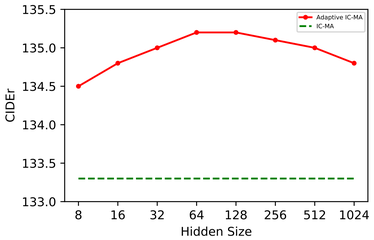

Figure 3. Effect of the hidden size.

Conclusion

Our research introduces an adaptive memory of feedback, employing a lightweight network to efficiently utilize information from various numbers of neighbors. This network, with only thousands of parameters, is easily trained using a validation dataset. We demonstrate its effectiveness in filtering noises and significantly outperforming the MA-IC approach through experiments on the MS COCO benchmark. Further, our method has proven robustness even with low-quality databases.

References

[1] Hui Chen, Guiguang Ding, Zijia Lin, Yuchen Guo, Caifeng Shan, and Jungong Han. Image captioning with memorized knowledge. *Cognitive Computation*, 13(4):807–820, 2021.

[2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[3] Zhengcong Fei. Memory-augmented image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Online*, pages 2–9, 2021.

[4] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022.

[5] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[6] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15465–15474, 2021.