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 \( f(x) \), and the corresponding value is the ground truth word \( y \). The memory bank \( D = (X, Y) \) encompasses all key-value pairs constructed from the entire training examples.

\[ D = (X, Y) = \{(f(x_1), y_1), \ldots, (f(x_n), y_n)\} \in (X, Y)^n \]

2. Combined Inference: During the inference stage, the system calculates the context embedding for \( f(x) \), retrieves \( k \) nearest neighbors from the memory bank \( D \), and aggregates the retrieved tokens to form the distribution \( P(y|x; \phi) \).

\[ P(y|x; \phi) = \sum_{(k)} \exp^{-d(x, y_k)/\beta} \]

The final probability is derived as an interpolation of the image captioning model’s distribution \( P(y|x; \phi_s) \) and \( P(y|x; \phi_u) \).

\[ P(y|x; \phi) = \lambda P(y|x; \phi_s) + (1 - \lambda) P(y|x; \phi_u) \]

where the weight \( \lambda \) 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 nearest neighbors and employing a fixed weight parameter \( \lambda \). 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.

1. Multiple kNN Classifiers: We employ information from multiple \( k \) 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:

\[ S = \{0\} \cup \{(k, \lambda_k) \in [N], k \leq K\} \]

where \( k \in S \) and \( k = 0 \) corresponds to the distribution generated directly by the IC model.

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

\[ P(y|x; \phi_s) = \sum_{k} P(y|x; \phi_k) + P(y|x; \phi_u) \]

where \( \phi_{MA} \) denotes the \( k \) nearest neighbor probability. We then derive the final predicted word as follows:

\[ y\hat{} = \arg\max_{y} P(y|x; \phi_{MA}) \]

Performance comparison

Table 1. Performance comparison with baseline methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>RSTNet+AMA</th>
<th>M2+AMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC-MA</td>
<td>81.3</td>
<td>43.8</td>
</tr>
<tr>
<td>Adaptive IC-MA</td>
<td>81.6</td>
<td>43.9</td>
</tr>
<tr>
<td>IC-MA</td>
<td>82.3</td>
<td>43.6</td>
</tr>
<tr>
<td>Adaptive IC-MA</td>
<td>82.5</td>
<td>43.7</td>
</tr>
</tbody>
</table>

Experience

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