




Hadamard Product Perceptron Attention for Image Captioning

Weitao Jiang¹ · Haifeng Hu¹ 

Accepted: 20 July 2022 / Published online: 28 July 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Recently, great progress has been made in promoting image captioning by improving Transformer structure. As a key component, dot product self-attention can update the representation of each feature vector in the visual encoder and guide the caption decoding process. However, the pairwise interaction in dot product self-attention enables attention weights to be learned at the instance or local level, making it difficult for attention module to obtain global feature representations. Furthermore, self-attention is always implemented in a multi-head fashion, where the calculation of each attention head is independent. It makes the model unable to exploit the complementary information contained in different heads. In this paper, we propose a Hadamard Product Perceptron Attention (HPPA) for image captioning, which introduces a more global feature interaction and incorporates interaction among attention heads to calculate attention results. Feature interaction method based on Hadamard product can integrate multimodal features more effectively than dot product and provide rich feature representation. Therefore, HPPA first utilizes Hadamard product to fuse the input features. Then, it generates a set of attention memory vectors containing global interaction features. The final attention weights are calculated via these vectors dynamically. When the multi-head mechanism is incorporated, the complementary information between different heads can be utilized by HPPA. We further integrate HPPA into Transformer encoder and propose a Hadamard Product Perceptron Transformer (HPPT) as a feature enhancement encoder. Moreover, HPPA and HPPT can be easily applied to existing attention or Transformer based models. Extensive experiments on MSCOCO and Flickr30k datasets demonstrate the effectiveness and generalizability of our proposal.

Keywords Image captioning · Hadamard product · Attention mechanism · Transformer

✉ Haifeng Hu
huhaif@mail.sysu.edu.cn

Weitao Jiang
jiangwt5@mail2.sysu.edu.cn

¹ Sun Yat-Sen University, Guangzhou, China

1 Introduction

Image captioning aims to automatically generate a grammatically correct and informative sentence to describe the given image. Most existing methods follow the encoder-decoder framework [2, 15, 46] with an attention mechanism. In such a framework, a Convolutional Neural Network (CNN) encoder first converts an input image into a set of regional feature vectors, and then a Recurrent Neural Network (RNN) based decoder generates a descriptive caption from these vectors. Meanwhile, the attention guides the decoding process by selecting the attended visual feature. However, CNN [2, 43] cannot further exploit the relationship information existing between visual features. Hence, this framework still suffers from insufficient use of visual features [6, 15].

Recently, with the introduction of Transformer [40], such a framework has been further promoted. As a key component of Transformer, the dot product self-attention can not only update the feature representation of visual features, but also provide visual cues to the language decoder. Based upon the CNN-RNN paradigm, [15] introduced an additional Transformer-based image encoder to refine the image features extracted by the CNN encoder. [25] proposed a full-attentive architecture that leveraged Transformer encoder and decoder to generate captions. Despite the excellent achievements that self-attention and Transformer have achieved, problems still remain.

Firstly, for most existing attention-based methods, the feature interactions in the attention module are obtained through dot product or addition between query and key vectors. That is, the relative importance (attention weights) of each feature vector relative to all other candidate vectors is calculated through region-level pairwise interactions. Consequently, feature interactions only occur in different local regions, and it is difficult to provide a global feature representation in attention mechanism.

Secondly, unlike conventional additive attention, the self-attention mechanism is always implemented in a multi-head fashion. The ‘multi-head’ attention consists of multiple parallel attention heads, each of which has a different projection on its inputs and outputs. Through the multi-head implementation, the performance of self-attention is improved due to the ability to attend to multiple positions simultaneously. However, each head is independent and there is no interactions between heads when calculating attention weights. In other words, the information between different heads is not shared.

As presented in Fig. 1a, when dot product self-attention is adopted to refine the feature representation of object region (denoted by node), each node updates its feature by a weighted sum of other nodes. In this process, the node with smaller attention weight can be ignored. Therefore, the interaction between nodes only occurs in local regions, and merely the nodes at higher layers can fuse more node information and contain global feature representations.

In the VQA field, numerous researches [20, 21] have proved that the feature interaction method based on Hadamard product can model the multimodal interactions more effectively. Inspired by this and to mitigate the issues mentioned above, we propose a Hadamard Product Perceptron Attention (HPPA) for image captioning. As illustrated in Fig. 2, HPPA works different from the additive attention and self-attention. HPPA first models the feature interactions between input vectors through Hadamard product. Then, HPPA calculates the attention memory vectors via multi-layer perceptron, in which each set of vectors contains the global interaction information among all regional features. Finally, the corresponding attention weights are calculated via a linear projection layer with softmax function.

To our knowledge, the superposition of multiple simple distributions can enhance the expression ability of the original distribution, such as Gaussian mixture model [10]. There-

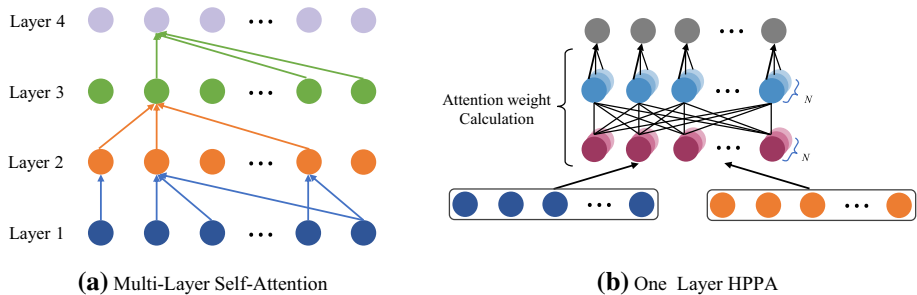


Fig. 1 Feature update process for self-attention and HPPA, where nodes represent the feature vectors of the object regions. **a** The traditional dot-product self-attention updates each feature node through pair-wise interactions between all features. This method only considers the local relationships among nodes (the node connections with low attention weight is omitted in the figure for brevity). **b** The proposed HPPA models the feature interaction between all input nodes, and then utilizes these nodes to adaptively learn the attention map. In this process, each node contains global feature interaction information

fore, when the multi-head mechanism is incorporated into HPPA, the multi-layer perceptron can superpose the attention distribution of all heads and contact all isolated heads. Benefiting from multi-head self-attention, we design a novel all-MLP architecture-based attention module, which is more effective than traditional attention models. As shown in Fig. 1b, compared with self-attention, one HPPA layer can fuse more local features. By employing Hadamard product and multi-layer perceptron to perform global feature interactions and fuse information between all attention heads, each updated node contains informative feature representation.

We further integrate HPPA with Transformer architecture and propose a Hadamard Product Perceptron Transformer (HPPT). As a relationship modeling module for visual features, HPPT can be directly embedded into most existing image captioning models. Extensive experiments on MSCOCO and Flickr30k datasets verify the effectiveness and generalizability of our proposal. Main contributions of this paper are as follows:

- We propose a Hadamard Product Perceptron Attention (HPPA) for image captioning, which can model the global interactions among region features, and utilize the complementary information between attention heads to calculate attention weights. HPPA can provide the decoder with a informative visual feature representation and boost the model performance.
- We embed HPPA into Transformer and propose a refining image encoder named Hadamard Product Perceptron Transformer (HPPT). As a generic image encoder, HPPT can be utilized to enhance the representation of extracted visual features and improve the accuracy of generated captions.
- We carry out extensive experiments on three classic image captioning methods (Soft-Attention, Up-Down and AoANet) combined with HPPA and HPPT respectively, and the results demonstrate the effectiveness and generalizability of our proposal.

2 Related Work

Image Captioning. As an essential task that links computer vision and natural language processing, image captioning has made significant progress with the development of these two fields, such as image classification [30], recognition [13], cross-modal retrieval [31] and

language model [14, 40]. In general, the methods of image captioning can be divided into three categories: template-based, retrieve-based and neural network-based.

The template-based methods [23, 32] first generate templated image caption with slots, and then fill the slots with the outputs of object detection and attribute prediction. The retrieve-based methods [12, 24] generate sentences by modifying the retrieved captions with similar visual contents. However, limited by retrieved results or manually defined templates, the captions produced by these two categories are insufficient in term of expression and accuracy.

To tackle the issues that remained in traditional methods and inspired by the advancement of deep neural network, current mainstream approaches are mainly neural network-based and adopt the encoder-decoder framework. Vinyals *et al* [43] utilized a CNN to encode the image, and then adopted LSTM to produce the final description. Afterwards, [46] proposed soft and hard attention mechanism, which can be integrated into the language decoder seamlessly, to calculate the most relevant image region features for word prediction. Based upon the soft attention, [4] presented a channel-wise attention mechanism, which not only calculated the spatial attention of image features, but also considered the channel attention of feature maps. [28] introduced a visual sentinel into the soft attention, which can provide textual information to the language decoder when the model does not need visual attention information. Anderson *et al* [2] proposed the bottom-up attention for the visual encoder to extract the salient object region feature for the input image, and presented the top-down attention for the language decoder to select the important region features. Moreover, other attention mechanisms such as: semantic attention [11], multi-modal attention network [44] and multi-head self-attention [40] were proposed by researchers to further boost the model performance.

Transformer Architecture-based Image Captioning. Recently, [40] proposed the Transformer model and showed the effectiveness and powerfulness of self-attention. Later on, some works extended it to image captioning task and state-of-the-art performances have been achieved. [15] introduced Transformer to model the relationships among visual features by a stack of self-attention layers, and further measured the relevance between query and attention result with an attention gate. Similarly, [6] proposed a meshed Transformer model to exploit the multi-level feature representation across multi-modal inputs in the encoder and decoder respectively. At the same time, [25] used the Transformer as image encoder and language decoder separately. It utilized visual and semantic features simultaneously in an entangled manner to exploit the complementary information from different modalities.

Although these attention or Transformer-based methods have made progress in improving the performance of baseline model. They are still restricted by the acquisition of global feature representation and the interaction of attention heads. Recently, some works [20, 21] have found that the feature interaction method based on Hadamard product can obtain a more global feature interaction. Moreover, there are also lots of works [5, 42] that have found that different heads in the multi-head attention mechanism contain complementary information. Inspired by this, we propose a Hadamard Product Perceptron Attention (HPPA) to model global feature representation and introduce attention heads interaction. Based on HPPA, we further propose a Hadamard Product Perceptron Transformer (HPPT) to better model the relationships among object region features and refine their representation.

3 Method

In this section, the details of the proposed Hadamard Product Perceptron Attention (HPPA) and Hadamard Product Perceptron Transformer (HPPT) will be elaborated. The HPPA and HPPT can be embedded in most existing attention-based framework. Therefore, in Sect. 3.1, we introduce the encoder-decoder architecture of image captioning, to better illustrate our proposal. In Sect. 3.2, the structure comparison between traditional additive attention, dot product self-attention, and HPPA will be described. In Sect. 3.3, we introduce the details of HPPT. At last, the combination of HPPA and HPPT with classic models will be presented in Sect. 3.4.

3.1 The Framework of Image Captioning

Current mainstream image captioning models all adopt an encoder-decoder architecture with attention mechanism. Typically, for a given image I , a pre-trained CNN is firstly utilized to generate a set of visual feature vectors $V_r = \{v_1, v_2, \dots, v_N\}$ containing accurate and sufficient information. Then, a LSTM is utilized as a language decoder to generate the complete description sentence $w = \{w_1, w_2, \dots, w_T\}$. The attention mechanism is incorporated into the decoder to dynamically provide more specific visual information for word reasoning. This process can be defined as follows [2, 4, 46]:

$$V_r = CNN(I) \quad (1)$$

$$v_t = Att(h_{t-1}, V_r) \quad (2)$$

$$h_t = RNN(v_t, x_t) \quad (3)$$

$$p_t = Softmax(v_t, h_t) \quad (4)$$

Where h_{t-1} represents the hidden state of the decoder at previous time step. $Att(\cdot)$ and v_t denote attention module and visual attention result respectively. x_t represents the word embedding vector of word w_t . p_t denotes the word probability distribution over the pre-defined vocabulary dictionary. It is worth noting that many different CNNs can be adopted as visual feature extractors, such as VGGNet [39], ResNet [13] and Faster R-CNN [35]. There are also many variants of the RNN decoder as [2, 15, 46].

3.1.1 Objective Function

Given an image and the target ground truth caption $w_{1:T}^*$, the model is first optimized by minimizing the cross-entropy loss:

$$L_{XE}(\theta) = - \sum_{t=1}^T \log p_{\theta}(w_t^* | w_{1:t-1}^*) \quad (5)$$

where θ represents the model parameters. Then, following [36] the Self-Critical Sequence Training (SCST) is applied to further finetune the model with CIDEr score reward:

$$L_{RL}(\theta) = -E_{w \sim p_{\theta}}[r(w)] \quad (6)$$

where $r(w)$ denotes the CIDEr score reward of the sampled sentence w . We can compute the approximation of the gradients of $L_{RL}(\theta)$ as:

$$\nabla_{\theta} L_{RL}(\theta) \approx -(r(w^s) - r(\hat{w})) \nabla_{\theta} \log p_{\theta}(w^s) \quad (7)$$

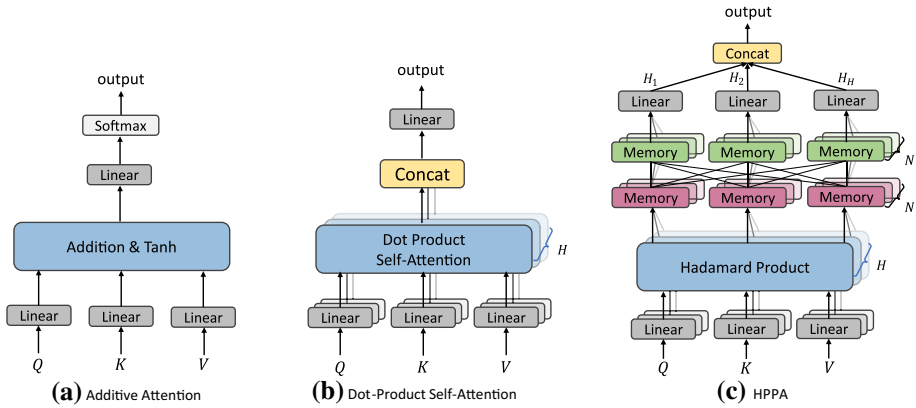


Fig. 2 Comparison of Additive Attention, Dot-product Attention and HPPA. H represents the attention heads in multi-head implementation, and N is the number of object regions

where $w^s = (w_1^s, \dots, w_T^s)$, and w_t^s is the random sampled word at time step t . $r(\hat{w})$ denotes the baseline reward for sentence generated through greedy decoding.

3.2 Hadamard Product Perceptron Attention

Attention-based methods dominate the image captioning community and demonstrate convincing advantages through excellent results. The additive attention and dot product attention are the two most commonly adopted attention mechanisms, as shown in Fig. 2. The former calculates the attention weight by matrix addition between query and key vectors whilst the latter through matrix multiplication (dot product).

We first briefly introduce the details of these two attention mechanisms. Specifically, given the query vector Q and key vectors K , the attention result for additive attention is computed as follows:

$$\alpha = \text{softmax}[W_v^T \tanh(W_k K + W_q Q)] \quad (8)$$

$$\text{Att}_{add-att}(Q, K, V) = \sum_i^N \alpha^i v_i \quad (9)$$

where W_v , W_k and W_q are trainable parameters. N is the number of image regions in V_r and α^i denotes the i -th element in α . For the dot-product attention, the similarity scores between query Q and keys K are calculated first. Then we apply a softmax function to obtain the attention weights W . Finally, the attention result is output through a weighted average over values V according to the weights W :

$$\text{Att}_{dot-att}(Q, K, V) = WV = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (10)$$

where d is a scaling factor (the dimension of Q). The dot-product attention is always implemented in a multi-head fashion, in which each Q , K and V is divided into H slices, and each attention head is responsible for one slice. The attention result for each slice Q_i , K_i , V_i is calculated separately, and the final result is formed by concatenating the outputs of all head:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O \quad (11)$$

$$head_i = Att_{dot-att}(Q_i, K_i, V_i) \quad (12)$$

where W^O is a learnable matrix that projects the concatenation of all heads outputs.

However, both additive attention and dot-product attention only focus on local pairwise interactions between features. In these attention modules, all candidate vectors are isolated, and a more global feature interaction and representation cannot be obtained. Furthermore, for the multi-head dot product attention, each attention head is calculated independently and merely uses late fusion to merge the information from different heads. Therefore, it ignores some complementary information between heads.

To this end, we are committed to improving the attention mechanism so as to promote the model by gaining informative attention features, and propose the Hadamard Product Perceptron Attention (HPPA). As illustrated in Fig. 2, HPPA does not adopt addition or multiplication (dot-product) to calculate the attention weight. Formally, for a given query vector Q and keys K , HPPA first uses Hadamard product to fuse the two input features. Then, a multi-layer perceptron with two hidden layers is utilized to perform the global interaction between all fused features and generate corresponding vector representation. Finally, a linear projection with softmax layer is adopted to compute the dynamic attention weight. The formulation of HPPA can be defined as follows:

$$H = Q \odot K \quad (13)$$

$$M = MLP(H) = \sigma(W_1^T H) W_2 \quad (14)$$

$$Att_{HPPA}(Q, K, V) = softmax(W_3 M) V \quad (15)$$

where \odot represents Hadamard product (element-wise multiplication), $W_1 \in R^{d \times 2d}$, $W_2 \in R^{2d \times d}$ and $W_3 \in R^{d \times 1}$ are trainable parameters. σ denotes ReLU activation function. When the multi-head mechanism is introduced into HPPA, the formulas can be rewritten as follows:

$$H_i = Q_i \odot k_i \quad (16)$$

$$M = [M_1, M_2, \dots, M_H] = \sigma(W_1^T [H_1, H_2, \dots, H_H]) W_2 \quad (17)$$

$$Att_{HPPA}^i(Q_i, K_i, V_i) = softmax(W_3 M_i) V_i \quad (18)$$

In this multi-head implementation, after the Hadamard product of $Q \odot K$, the attention memory matrix M is used to superimpose the results of each head. In this way, the original isolated attention heads are connected, complementary information is utilized, and the attended features is more informative.

3.3 Hadamard Product Perceptron Transformer

The Transformer architecture dominates the natural language processing community through wide success in numerous tasks, such as machine translation [8] and language modeling [7]. Later on, several works that design a Transformer-like encoder [15, 25] introduce Transformer into image captioning task successfully.

As a key component of the Transformer architecture, the multi-head self-attention owns the ability of capturing long-range dependencies of input sequence features and achieves remarkable performance. Since the proposed HPPA and multi-head self-attention have similar functions, we further embed HPPA into the Transformer encoder and propose a Hadamard Product Perceptron Transformer (HPPT) to refine the visual feature representation through global feature interaction.

Typically, HPPT contains two sub-layers (see Fig. 4): a HPPA layer and a Feed-Forward Network (FFN) layer. Furthermore, layer normalization and residual connection are adopted

to the two sub-layers respectively. Given an image, a pre-trained Faster R-CNN [35] is firstly adopted to generate feature vectors V_r . Then, instead of inputting these feature vectors directly into the language decoder, the proposed HPPT is utilized to model the relationships between these vectors, refine and enhance the representation of visual features V_r . The operation of HPPT can be formulated as follows:

$$\tilde{V} = \text{LayerNorm}(V_r + \text{Att}_{\text{HPPA}}(W_Q V_r, W_K V_r, W_V V_r)) \quad (19)$$

$$V' = \text{LayerNorm}(\tilde{V} + \text{FFN}(\tilde{V})) \quad (20)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are linear transformation matrices. Att_{HPPA} represents the multi-head HPPA defined in Eq. 18, in which the interaction between different attention heads are considered. $\text{FFN}(x) = \sigma(xW_{f1} + b_{f1})W_{f2} + b_{f2}$, where x is the input of $\text{FFN}(\cdot)$, σ denotes ReLU, and W_{f1}, W_{f2}, b_{f1} and b_{f2} are all trainable parameters. Through HPPT, the feature vectors are updated $V_r \rightarrow V'$. Note that the HPPT does not change the dimension of the input features, so, similar to Transformer encoder, it can be stacked with L layers.

3.4 Combine HPPA and HPPT with Classic Models

As an attention mechanism, HPPA provides the language decoder with the necessary visual information. As a feature enhancer, the HPPT models the relationships among all region features and aggregates all local information to form the final global feature representation. Neither module is designed for a specific model, hence they can be replaced or inserted into many existing models. In this part, we introduce how to apply HPPA and HPPT to three popular baseline models [2, 15, 46].

3.4.1 Soft-Attention

A typical attention-based image captioning model proposed in [46], which adopts the additive attention and feeds the attended image region features to the LSTM layer for word predicting. In our implementation, we utilize the pre-trained Faster R-CNN as the image encoder rather than the ResNet-101 in [46] for fair comparison.

3.4.2 Up-Down

A widely-used baseline architecture for image captioning presented in [2], which adopts two LSTM layers with additive attention mechanism to generate captions. Up-Down is the first job that leverages the object region image features instead of uniform grid features.

3.4.3 AoANet

AoANet [15] has the similar encoder-decoder framework of Soft-Attention and Up-Down. But in particular, AoANet designs an additional image feature refiner to refine the representation of the encoded object region features, and adopts the 'attention on attention' mechanism to filter out irrelevant attention results for the language decoder.

The structure of Soft-Attention, Up-Down and AoANet and the combination of HPPA are illustrated in Fig. 3. The proposed HPPA can directly replace the existing attention module of these models without changing any model structure.

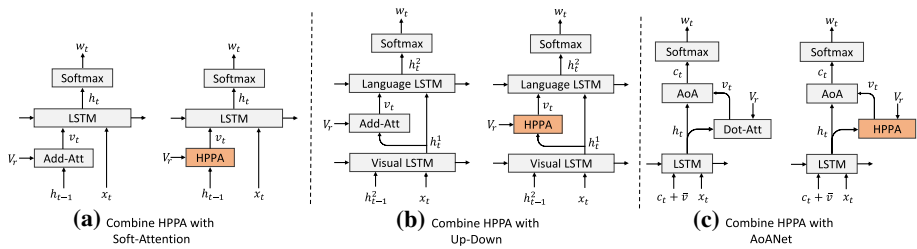


Fig. 3 In the language decoder, the combination of HPPA with three popular attention-based baseline models: **a** Soft-Attention, **b** Up-Down, **c** AoANet

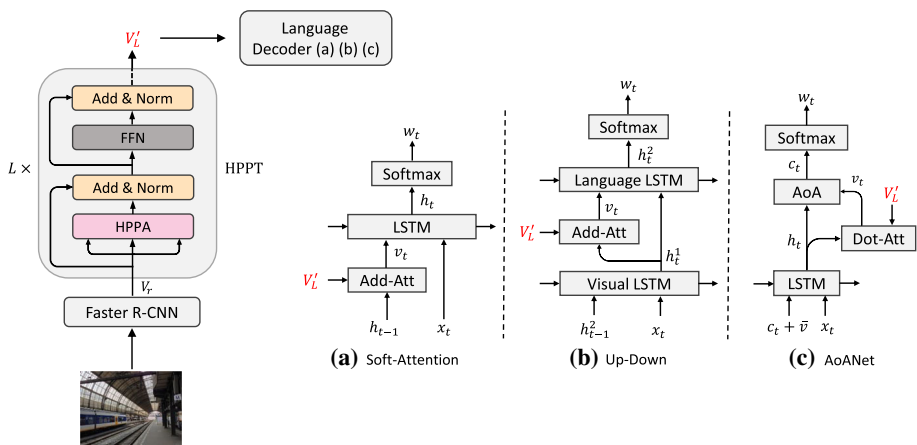


Fig. 4 In the image encoder, the combination of HPPT with three popular attention-based baseline models: **a** Soft-Attention, **b** Up-Down, **c** AoANet

As shown in Fig. 4, Soft-Attention, Up-Down and AoANet can be combined with HPPT seamlessly. For Soft-Attention and Up-Down model, we merely enhance the feature representation of the object region features extracted by Faster R-CNN (from V_r to V_L'). For AoANet, we replace its original feature refinement module with HPPT.

4 Experiments

4.1 Datasets and Settings

All the experiments are carried out on MSCOCO [27] and Flickr30k [34] to verify the effectiveness of our proposal, and the widely used “Karpathy” data split is adopted [19] for fair comparisons with other methods. Specifically, for MSCOCO, it contains 113, 287 images for training, and 5, 000 images respectively for testing and validation. For Flickr30k, it includes 29, 000 training images, and 1, 000 images for validation and testing respectively. Each image in the dataset has 5 ground-truth sentences. For preprocessing of text information, we convert all words in sentences to lowercase, and the words count less than 5 times are removed. Then, we truncate all captions to a maximum length of 16. Finally, two vocabularies with 9, 487 and 7, 000 words will be obtained respectively.

Table 1 Performance comparisons of embedding HPPA into classic models on MSCOCO Karpathy's test split

Model	B-1	B-2	B-3	B-4	M	R	C	S
Soft-Attention*	76.6	60.8	47.0	36.0	27.0	56.3	111.3	20.1
Soft-Attention + HPPA	77.4	61.9	48.2	37.2	27.3	57.1	114.2	20.3
Up-Down*	77.0	61.1	47.1	36.0	27.4	56.8	113.1	20.6
Up-Down + HPPA	77.5	61.9	48.1	37.3	27.7	57.3	115.5	20.8
AoANet*	77.3	61.6	48.0	37.2	28.3	57.3	118.0	21.3
AoANet + HPPA	77.6	62.0	48.4	37.5	28.3	57.6	119.0	21.5

The best results of the experiment are marked as bold

* denotes that the results comes from our reimplementation

4.2 Implementation Details

The Faster R-CNN [35] pre-trained on Visual Genome [22] and ImageNet [9] is leveraged to detect object regions in image and extract regional features. The dimension of the extract object features is 2,048, and we transform it to 512 dimension before being used by other modules. For our captioning model, in all experiments, we set the dimension of word embedding to 1,024 and the hidden size of LSTM to 512.

For the training stage, the model is first trained for 50 epochs under cross-entropy loss. For all models, the initial learning rate is set to $5e-4$, and the warm-up steps are set to 10,000 when the Transformer-like encoder such as [15] and HPPT is utilized. Then, during CIDEr reward optimization process, the initial learning rate is $1e-5$, and when the CIDEr score does not increase within 3 epochs, it decays by 0.8. As for the inference stage, the beam search decoding strategy with a beam size of 3 is adopted. The following different metrics, BLEU [33], METEOR [3], ROUGE-L [26], CIDEr [41] and SPICE [1], are used to evaluate the performance of our proposal.

4.3 Ablation Study

To verify the effects of HPPA and HPPT for model performance, extensive ablation experiments are designed as follows:

4.3.1 Classic Attention-based Models Combined with HPPA

As mentioned in Sect. 3.4, HPPA is a universal attention module and can be embedded in most existing methods. Therefore, we applied HPPA to three widely-used baseline image captioning models: Soft-Attention [46], Up-Down [2] and AoANet [15]. As shown in Table 1, where the three baseline models are reimplemented under the settings described in Sect. 4.2. From the Table, we can observe that when replacing the original attention module in the baseline models with HPPA, the performance of all baseline models has been obviously improved. The results validate the effectiveness of HPPA compared to additive and dot-product attention mechanism. Especially, the CIDEr scores of Soft-Attention and Up-Down increased by 2.9 and 2.4 respectively. This can be attributed to the advantages of HPPA, which adopts Hadamard product and multi-layer perceptron to capture the global interaction between features and utilize the complementary information between the attention heads.

Table 2 Ablation study about the impact of each component of the proposed HPPT on MSCOCO Karpathy's test split

Model	B-1	B-2	B-3	B-4	M	R	C	S
Baseline	76.7	60.8	47.2	36.4	27.7	56.7	114.2	20.8
+ SA	77.2	61.5	47.8	37.0	28.0	57.1	116.1	21.2
+ Transformer	77.4	61.8	48.0	37.0	28.2	57.2	118.4	21.5
+ HPPA	77.8	62.1	48.4	37.6	28.2	57.5	118.0	21.3
+ HPPT(1-head)	77.1	61.5	47.8	36.9	28.2	57.1	116.9	21.4
+ HPPT(2-head)	77.4	61.7	47.8	36.8	28.2	57.2	117.8	21.4
+ HPPT(4-head)	77.6	62.0	48.3	37.3	28.3	57.3	118.5	21.5
+ HPPT(8-head)	77.9	62.4	48.7	37.7	28.4	57.6	120.1	21.6
+ HPPT(16-head)	77.3	61.7	48.0	37.0	28.3	57.5	117.6	21.2

SA denotes the self-attention

4.3.2 Ablation Study on HPPT

HPPT is mainly adopted as a feature enhancer and is designed based on the Transformer encoder. To quantify the influence of the proposed HPPT, ablation experiments are conducted with the following combinations: (1) Baseline: AoANet without additional refining encoder. (2) SA: Transformer encoder block [40] without the Feed-Forward Network (FFN) layer. (3) Transformer: Complete Transformer encoder block. (4) HPPA: A stack of HPPA module (3 layers). (5) HPPT(H -head): Replace Transformer encoder with HPPT, where HPPT is a multi-head implementation with H attention heads. Not specifically mentioned, all multi-head attentions have 8 heads.

The experimental results are illustrated in Table 2, where SA and Transformer are stacked with 4 layers while HPPA and HPPT are stacked with 3 layers. From Table 2, we can find that all the additional feature enhancers bring obvious performance gain in all evaluation metrics. Without the FFN layer, the proposed HPPA achieves comparable performance with Transformer since it can aggregate more local features. When HPPT is implemented in single head, it has poor performance compared with HPPA. However, as the number of attention heads increased, the performance improvements appeared. That indicates the efficacy of incorporating interaction among attention heads. Meanwhile, HPPT(8-head) surpasses Transformer with 8 heads. We infer that incorporating interaction among attention heads can take advantage of complementary information and bring further performance boost. Moreover, the model with 16-heads does not achieve better performance, probably because of the low-dimensional feature vectors contained in each head, which are difficult to represent complex visual features.

4.3.3 Comparison Between Transformer and HPPT

Transformer-based methods usually utilize the Transformer block in a stacked fashion, which causes model sensitive to the number of layers. We first examine the impact of the number of layers when apply Transformer and HPPT to Baseline model, the results are presented in Table 3. With the number of layers L increasing incremenatally from 1 to 5, the performance of Transformer and HPPT shares similar trends. That is, the CIDEr score first improves significantly, and slightly decreases afterwards. Moreover, stacking more layers will not

Table 3 Ablation study about the impact of layers L when applying Transformer and HPPT to Baseline model on MSCOCO Karpathy's test split

Layers	Model	B-1	B-2	B-3	B-4	M	R	C	S
1	Transformer	77.1	61.6	47.6	36.3	27.7	56.8	115.4	21.4
	HPPT	77.1	61.3	47.4	36.4	28.1	57.0	117.2	21.4
2	Transformer	77.2	61.6	47.8	36.6	28.1	57.1	116.2	21.5
	HPPT	77.5	62.1	48.4	37.5	28.4	57.5	119.3	21.5
3	Transformer	77.0	61.5	47.7	36.6	28.2	57.1	117.4	21.4
	HPPT	77.9	62.4	48.7	37.7	28.4	57.6	120.1	21.6
4	Transformer	77.4	61.8	48.0	37.0	28.2	57.2	118.4	21.5
	HPPT	77.5	62.1	48.4	37.3	28.3	57.5	119.5	21.6
5	Transformer	77.6	62.0	48.2	37.2	28.4	57.5	118.2	21.5
	HPPT	77.7	62.2	48.5	37.4	28.3	57.5	119.0	21.7

The best results of the experiment are marked as bold

Table 4 Performance comparisons of embedding HPPT into classic models on MSCOCO Karpathy's test split

Model	B-1	B-2	B-3	B-4	M	R	C	S
Soft-Attention*	76.6	60.8	47.0	36.0	27.0	56.3	111.3	20.1
Soft-Attention + HPPT	78.0	62.6	48.6	37.3	27.7	57.4	117.1	21.0
Up-Down*	77.0	61.1	47.1	36.0	27.4	56.8	113.1	20.6
Up-Down + HPPT	77.4	61.8	48.2	37.5	28.2	57.4	118.4	21.5
AoANet*	77.3	61.6	48.0	37.2	28.3	57.3	118.0	21.3
AoANet + HPPT	77.9	62.4	48.7	37.7	28.4	57.6	120.1	21.6

The best results of the experiment are marked as bold

* denotes that the result comes from our reimplementaion

bring further positive effect. Instead, it will increase complexity and parameters size, as well as reduce performance. Meanwhile, we can see that HPPT achieves better result with fewer layers, especially 1-layer HPPT performances on par with 3-layer Transformer. HPPT reaches the best result when the number of layers is 3, while Transformer requires to stack 4 to 5 layers. The comparison demonstrates the effectiveness of HPPT, which updates the visual feature representation by modeling the global interaction between object regions. If not specifically mentioned, all subsequent experiments will use 3 layers of HPPT.

4.3.4 Classic Models Combined with HPPT

Since HPPT is a general feature enhancer that does not change the dimension of input visual features, and in order to verify the capability of applying HPPT as an additional visual encoder to classic models, we embed HPPT into three widely-used baseline image captioning models: Soft-Attention [46], Up-Down [2] and AoANet [15]. In Table 4, the three baseline models are reimplemented under the settings described in Sect. 4.2. We can observe that applying HPPT to Soft-Attention and Up-Down can bring a remarkable improvement in all evaluation metrics, and the CIDEr score increases by 5.8 and 5.3 respectively. As for AoANet + HPPT, we just replace the original feature refining module with HPPT, which still surpasses the

Table 5 Ablation study about different refining encoders

Refining Encoder	B-1	B-2	B-3	B-4	M	R	C	S
FC + ReLU	76.7	60.8	47.2	36.4	27.7	56.7	114.2	20.8
3 × Transformer	77.0	61.5	47.7	36.6	28.2	57.1	117.4	21.4
3 × AoA	77.1	61.3	47.6	36.8	28.1	57.1	117.2	21.2
6 × AoA	77.3	61.6	48.0	37.2	28.3	57.3	118.0	21.3
3 × HPPT	77.9	62.4	48.7	37.7	28.4	57.6	120.1	21.6

The best results of the experiment are marked as bold

Table 6 Performance comparison with state-of-the-art methods on MSCOCO Karpathy's test split under XE loss. - denotes that the metric is not provided

Model	B-1	B-2	B-3	B-4	M	R	C	S
NIC [43]	–	–	–	29.6	25.2	52.6	94.0	–
SCST [36]	–	–	–	30.0	25.9	53.4	99.4	–
LSTM-A [50]	75.4	–	–	35.2	26.9	55.8	108.8	20.0
Up-Down [2]	77.2	–	–	36.2	27.0	56.4	113.5	20.3
RFNet [17]	76.4	60.4	46.6	35.8	27.4	56.8	112.5	20.5
GCN-LSTM [49]	77.3	–	–	36.8	27.9	57.0	116.3	20.9
ETA [25]	77.3	–	–	37.1	28.2	57.1	117.9	21.4
AoANet [15]	77.4	–	–	37.2	28.4	57.5	119.8	21.3
ETN [37]	77.9	62.5	48.9	38.0	–	57.7	120.0	21.2
Transformer [38]	76.1	59.9	45.2	34.0	27.6	56.2	113.2	21.0
Ours	77.9	62.4	48.7	37.7	28.4	57.6	120.1	21.6

The best results of the experiment are marked as bold

AoANet. These results indicate the powerfulness and generalizability of HPPT in feature enhancement.

4.3.5 Comparison of Different Refining Encoders

In order to verify the effect of the refining encoder with different network structure, comparative experiments with same baseline model are carried out. In Table 5, FC + ReLU represents that the refining encoder consists of a linear transformation layer with ReLU activation function. 3 × Transformer denotes a stack of 3 Transformer encoder layers. AoA represents the refining module in AoANet. As summarized in Table 5, we can find that 3-layer HPPT outperform all other methods. Transformer and AoA bring obvious performance gain over FC + ReLU, while HPPT achieves better result than both of them. It is worth noting that 3-layer HPPT surpasses 6-layer AoA, which demonstrates the advantage of global feature representation as well as the leverage of complementary information between attention heads.

4.4 Comparison with State-of-the-art

Tables 6 and 7 summarize the performance comparisons between our proposed method (AoANet + HPPT) and some state-of-the-art methods. The compared models include NIC

Table 7 Performance comparison with state-of-the-art methods on MSCOCO Karpathy's test split under CIDEr reward optimization

Model	B-1	B-2	B-3	B-4	M	R	C	S
NIC [43]	–	–	–	31.9	25.5	54.3	106.3	–
SCST [36]	–	–	–	34.2	26.7	55.7	114.0	–
LSTM-A [50]	78.6	–	–	35.5	27.3	56.8	118.3	20.8
Up-Down [2]	79.8	–	–	36.3	27.7	56.9	120.1	21.4
RFNet [17]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [49]	80.5	–	–	38.2	28.5	58.3	127.6	22.0
SGAE [48]	80.8	–	–	38.4	28.4	58.6	127.8	22.1
AoANet [15]	80.2	–	–	38.9	29.2	58.8	129.8	22.4
ETN [37]	80.6	65.3	51.1	39.2	–	58.9	128.9	22.6
Transformer [38]	80.2	64.8	50.5	38.6	28.8	58.5	128.3	22.6
M^2 Transformer [6]	80.8	–	–	39.1	29.2	58.6	131.2	22.6
Ours	80.9	65.5	51.3	39.3	29.1	58.9	130.5	22.8

The best results of the experiment are marked as bold

– denotes that the metric is not provided

[43], SCST [36], LSTM-A [50], Up-Down [2], RFNet [17], GCN-LSTM [49], SGAE [48], ETA [25], AoANet [15], ETN [37], Transformer [38] and M^2 Transformer [6]. NIC introduces an encoder-decoder (CNN-LSTM) framework for image captioning. SCST first utilizes the reinforcement learning approach to directly optimize the pre-trained model. LSTM-A leverages the semantic attribute to boost the model performance. Up-Down adopts two attention modules and two LSTM layers. RFNet, which utilizes multiple RNNs to fuse the information from multiple CNNs. GCN-LSTM and SGAE outperform previous methods by leveraging extra Graph Convolutional Network (GCN) to exploit the rich object relationship features in the image. ETA, AoANet, Transformer and M^2 Transformer are all Transformer-based image captioning models. ETN edits the image caption generated by AoANet to obtain a final caption rather than from scratch.

For the XE loss training stage, the comparison results on Karpathy's test split are illustrated in Table 6. It can be observed that our proposal outperforms all the compared methods in most evaluation metrics, which demonstrates the effectiveness of our proposal. Moreover, our model still obtains the best results when comparing with other GCN-based and Transformer-based methods. Especially, the architecture of our model is similar to AoANet, both of which adopt additional refining encoder. This confirms that exploiting better image feature representation through HPPT can further boost model performance. Table 7 presents the performances of the CIDEr reward optimized methods. From the table, we can find that the proposed model consistently exhibits better performances and achieves the highest scores in most metrics. It is worth noting that our model performs on par with ETN under XE loss while surpasses it under CIDEr reward. We infer that this discrepancy may be caused by the editing network of ETN, which refines the generated caption from a pre-trained baseline (AoANet). However, the editing network is more efficient when the caption predicted by the baseline is of general quality. The results indicate the importance of global feature representation and exploiting complementary information between attention heads.

To further validate the effectiveness of the proposed model, experiment is conducted on Flickr30k [34] dataset. In Table 8, we report the performance comparison results with some

Table 8 Performance comparison with state-of-the-art methods on Flickr30k Karpathy's test split under XE loss.

Model	B-1	B-2	B-3	B-4	M	R	C	S
SCA-CNN [4]	66.2	46.8	32.5	22.3	19.5	44.9	44.7	–
Adaptive [28]	67.7	49.4	35.4	25.1	20.4	–	53.1	–
NBT [29]	69.0	–	–	27.1	21.7	–	57.5	–
GVD [52]	69.9	–	–	27.3	22.5	–	62.3	16.5
Full-GC [51]	69.8	–	–	29.1	22.7	–	63.5	–
SF [18]	64.7	45.6	32.0	22.4	19.7	44.9	46.7	13.6
BCAN [16]	69.8	51.9	37.8	27.4	21.2	48.8	58.3	–
NSRL [45]	70.5	53.5	39.9	29.5	22.7	50.0	64.6	16.7
CapNet [47]	70.7	53.7	39.9	29.6	22.7	50.6	65.9	–
Ours	72.4	55.3	41.5	31.0	22.9	50.5	66.1	17.3

The best results of the experiment are marked as bold
 - denotes that the metric is not provided



Up-Down:
 a group of bikes parked next to each other.
Up-Down + HPPT:
 a man standing in front of a bunch of bikes.



Up-Down:
 a couple of cars that are parked next to each other.
Up-Down + HPPT:
 a city bus is driving down the street.



Up-Down:
 a group of people sitting around a table eating food.
Up-Down + HPPT:
 a man and woman sitting at a table with food.



Up-Down:
 a train is parked at a train station.
Up-Down + HPPT:
 a blue and white train pulling into a train station.



AoANet:
 a man with a hat holding a frisbee.
AoANet + HPPT:
 a man wearing a hat holding a frisbee in a parking lot.



AoANet:
 a herd of giraffes standing next to each other.
AoANet + HPPT:
 a group of giraffes standing in a zoo.



AoANet:
 a group of young men playing a video game.
AoANet + HPPT:
 a group of men playing a video game in a living room.



AoANet:
 a group of people standing on a sidewalk at night.
AoANet + HPPT:
 a crowd of people standing in front of a train at night.

Fig. 5 Image captioning examples of Up-Down, UP-Down+HPPT, AoANet and AoANet+HPPT

state-of-the-art methods. The compared models include SCA-CNN [4], Adaptive [28], NBT [29], GVD [52], Full-GC [51], SF [18], BCAN [16], NSRL [45] and CapNet [47]. As can be observed, our proposal surpasses all the compared methods, and only performs inferiorly to CapNet on the ROUGE metric. The leading result demonstrates the effectiveness of our proposal on multiple datasets.

4.5 Qualitative Analysis

Some image captioning examples generated by Up-Down, Up-Down + HPPT, AoANet and AoANet+ HPPT are presented in Fig. 5 to perform qualitative analysis of our proposal. For

the results in the first row, the caption generated by Up-Down lacks important object ‘man’ in the image and does not recognize the correct action ‘driving down’ in the first and second examples respectively. In the third and fourth examples, Up-Down+HPPT produces more accurate and detailed descriptions, such as ‘a man and woman’ instead of ‘a group of people’, and ‘a blue and white train’. As for the examples in the second row, we can observe that the captions generated by AoANet are consistent with the content of the image. Nevertheless, when it is combined with HPPT, description with a global scene and detailed expression can be produced, such as ‘parking lot’, ‘zoo’ and ‘living room’. In the fourth example, our model generates more specific caption like ‘in front of a train’ rather than ‘sidewalk’. These results indicate that the global interaction between region features can help the model recognize background or scene information accurately, and leveraging the complementary information in different attention heads can help the model consider detailed information.

5 Conclusion

In this paper, we propose a Hadamard Product Perceptron Attention (HPPA) for image captioning. HPPA aims to address the problems of traditional additive attention and dot product self-attention, including the lack of global feature interaction and the independent calculation of attention heads in the multi-head mechanism. Through Hadamard product and multi-layer perceptron, HPPA can model the global interaction among input features and leverage the complementary information between attention heads. By embedding HPPA into Transformer architecture, we further propose Hadamard Product Perceptron Transformer (HPPT) as a refining encoder to enhance the feature representation. Moreover, since HPPA and HPPT can be utilized by most attention-based models flexibly, we apply them to three classical models. Extensive comparative experiments results on MSCOCO and Flickr30k datasets validate the effectiveness and generalizability of our proposal.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 62076262, Grant 61673402, Grant 60802069 and Grant 61273270.

References

1. Anderson P, Fernando B, Johnson M, Gould S (2016) Spice: Semantic propositional image caption evaluation. In: *Proceedings of the European Conference on Computer Vision*, Springer, pp 382–398
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 6077–6086
3. Banerjee S, Lavie A (2005) Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp 65–72
4. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 5659–5667
5. Clark K, Khandelwal U, Levy O, Manning CD (2019) What does bert look at? an analysis of bert’s attention. *arXiv preprint [arXiv:1906.04341](https://arxiv.org/abs/1906.04341)*
6. Cornia M, Stefanini M, Baraldi L, Cucchiara R (2020) Meshed-memory transformer for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp10578–10587
7. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R (2019) Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint [arXiv:1901.02860](https://arxiv.org/abs/1901.02860)*

8. Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser Ł (2018) Universal transformers. arXiv preprint [arXiv:1807.03819](https://arxiv.org/abs/1807.03819)
9. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255
10. Friedman N, Russell S (1997) Image segmentation in video sequences: A probabilistic approach. In: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, pp 175–181
11. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L (2017) Semantic compositional networks for visual captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5630–5639
12. Gupta A, Verma Y, Jawahar CV (2012) Choosing linguistics over vision to describe images. In: Twenty-Sixth AAAI Conference on Artificial Intelligence
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
14. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
15. Huang L, Wang W, Chen J, Wei X-Y (2019) Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4634–4643
16. Jiang W, Wang W, Haifeng H (2021) Bi-directional co-attention network for image captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17(4):1–20
17. Jiang W, Ma L, Jiang Y-G, Liu W, Zhang T (2018) Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 499–515
18. Kalimuthu M, Mogadala A, Mosbach M, Klakow D (2021) Fusion models for improved image captioning. In: International Conference on Pattern Recognition, Springer, pp 381–395
19. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3128–3137
20. Kim J-H, Lee S-W, Kwak D, Heo M-O, Kim J, Ha J-W, Zhang B-T (2016) Multimodal residual learning for visual qa. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, pp 29:361–369
21. Kim J-H, On K-W, Lim W, Kim J, Ha J-W, Zhang B-T (2016) Hadamard product for low-rank bilinear pooling. arXiv preprint [arXiv:1610.04325](https://arxiv.org/abs/1610.04325)
22. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vision* 123(1):32–73
23. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: Understanding and generating simple image descriptions. *IEEE Trans Pattern Anal Mach Intell* 35(12):2891–2903
24. Kuznetsova P, Ordonez V, Berg TL, Choi Y (2014) Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics* 2:351–362
25. Li G, Zhu L, Liu P, Yang Y (2019) Entangled transformer for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp 8928–8937
26. Lin C-Y (2004) Rouge: A package for automatic evaluation of summaries. In: the Workshop on Text Summarization Branches Out, pp 74–81
27. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Lawrence ZC (2014) Microsoft coco: common objects in context. In: Proceedings of the European Conference on Computer Vision, Springer, pp 740–755
28. Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 375–383
29. Lu J, Yang J, Batra D, Parikh D (2018) Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7219–7228
30. Meng M, Lan M, Jun Yu, Jigang W, Tao D (2019) Constrained discriminative projection learning for image classification. *IEEE Trans Image Process* 29:186–198
31. Meng M, Wang H, Jun Yu, Chen H, Jigang W (2020) Asymmetric supervised consistent and specific hashing for cross-modal retrieval. *IEEE Trans Image Process* 30:986–1000
32. Mitchell M, Dodge J, Goyal A, Yamaguchi K, Stratos K, Han X, Mensch A, Berg A, Berg T, Daumé III H (2012) Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp 747–756
33. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp 311–318

34. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641–2649
35. Shaoqing R, Kaiming H, Ross G, Jian S (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, pp 28:91–99
36. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7008–7024
37. Sammani F, Melas-Kyriazi L (2020) Show, edit and tell: A framework for editing image captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4808–4816
38. Sharma P, Ding N, Goodman S, Soric R (2018) Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2556–2565
39. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
40. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the Conference on Advances in Neural Information Processing Systems, pp 30:5998–6008
41. Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4566–4575
42. Vig J (2019) A multiscale visualization of attention in the transformer model. arXiv preprint [arXiv:1906.05714](https://arxiv.org/abs/1906.05714)
43. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3156–3164
44. Wang J, Tang J, Luo J (2020) Multimodal attention with image text spatial relationship for ocr-based image captioning. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 4337–4345
45. Wang X, Ma L, Fu Y, Xue X (2021) Neural symbolic representation learning for image captioning. In: Proceedings of the 2021 International Conference on Multimedia Retrieval, pp 312–321
46. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp 2048–2057
47. Yang L, Wang H, Tang P, Li Q (2021) Captionnet: A tailor-made recurrent neural network for generating image descriptions. *IEEE Trans Multimedia* 23:835–845
48. Yang X, Tang K, Zhang H, Cai J (2019) Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 10685–10694
49. Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning. In: Proceedings of the European conference on computer vision (ECCV), pp 684–699
50. Yao T, Pan Y, Li Y, Qiu Z, Mei T (2017) Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4894–4902
51. Zhong Y, Wang L, Chen J, Yu D, Li Y (2020) Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision, Springer, pp 211–229
52. Zhou L, Kalantidis Y, Chen X, Corso JJ, Rohrbach M (2019) Grounded video description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6578–6587

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.