PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Attention-Aligned Transformer for 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

Recently, attention-based image captioning models, which are expected to ground correct image regions for proper word generations, have achieved remarkable performance. However, some researchers have argued "deviated focus" problem of existing attention mechanisms in determining the effective and influential image features. In this paper, we present $\mathcal{A}^2$ - an attention-aligned Transformer for image captioning, which guides attention learning in a perturbation-based self-supervised manner, without any annotation overhead. Specifically, we add mask operation on image regions through a learnable network to estimate the true function in ultimate description generation. We hypothesize that the necessary image region features, where small disturbance causes a obvious performance degradation, deserve more attention weight. Then, we propose four aligned strategies to use this information to refine attention weight distribution. Under such a pattern, image regions are attended correctly with the output words. Extensive experiments conducted on the MS COCO dataset demonstrate that the proposed $\mathcal{A}^2$ Transformer consistently outperforms baselines in both automatic metrics and human evaluation. Trained models and code for reproducing the experiments are publicly available.

假设扰动更大导致性能下降更多的部分应该值得更多的权重。

1 Introduction

The task of generating a concise textual summary of a given image, known as image captioning, is one of the most challenges that require joint vision and language modeling. Currently, most image captioning algorithms follow an encoder-decoder paradigm in which an RNN-based decoder network is used to predict words according to the image features extracted by the CNN-based encoder network (Vinyals et al. 2015). In particular, the incorporation of attention mechanisms has greatly advanced the performance of image captioning and can be used to provide insights for the inner workings (Xu et al. 2015; Anderson et al. 2018; Huang et al. 2019; Li et al. 2019; Cornia et al. 2020; Pan et al. 2020). It dynamically encodes visual information by weighting more those regions relevant to the current word generation.

However, it is widely questioned whether highly attended image regions have a true correlation on the caption generation. On the one hand, Serrano and Smith (2019) find that erasing the representations accorded high attention weights do not necessarily lead to a significant performance decrease sometimes. On the other hand, Liu et al. (2020) state that most attention-based image captioning models use the hidden state of the current input to attend to the image regions and attention weights are inconsistent with other feature importance metrics (Selvaraju et al. 2019). It further proves that attention mechanisms are incapable of precisely identifying decisive inputs for each prediction (Zhang et al. 2021), also referred to as "deviated focus", which would impair the performance of image content description. As show in Figure 1, at the time step to generate the 5^{th} word, original attention mechanisms focus most on the local "shelf" region, as a result, the incorrect noun "sink" is generated. The unfavorable attended image region also impairs the grounding performance and ruins the model interpretability (Cornia, Baraldi, and Cucchiara 2019).

Figure 1: Illustration of the sequence of attended image regions in generating each word for the description before (blue) and after (red) attention alignment. At each time step, only the top-1 attended image region is shown. The original attended image regions are grounded less accurately, demonstrating the deficiency of previous attention mechanisms.

In this paper, we propose a novel perturbation-based self-supervised attention-aligned method for image captioning, referred to as $\mathcal{A}^2$ Transformer, without any additional annotation overhead. To be specific, we keep applying mask operation to disturb the original attention weights with a learnable network, and evaluate the final performance change of image captioning model, so as to discover which input image regions affect the performance of image captioning model most. In between, we add a regular term, aims to determine the smallest perturbation extents that cause the most prominent degrading description performance. Under this condition, we can find the most informative and necessary image features for the caption prediction, which deserve more attention. Later, we use this supervised information to refine the attention weight distribution. In particular, we design four fusion methods to incorporate the updated attention weights into the original attention weights: i) max pooling, ii) moving average, iii) exponential decay, and iv) gate mechanism. Finally, the image captioning model is optimized based on the modified attention.

It is notable that the aligned attention method is model-agnostic and can be easily incorporated into existing state-of-the-art image captioning models to improve their captioning performances. Extensive experiments are conducted to verify our method's effectiveness on the MS COCO dataset. According to both automatic metrics and human evaluations, the image captioning models equipped with the attention-aligned method can significantly boost performance. More intuitive example can be see in Figure 1, at the time step to generate the 5^{th} word, our method align the attention weight for the new image region *shelf* more and the matched correct word *shelf* is generated correspondingly. We further analyze the correlation between mask perturbation and feature importance metrics as well as investigate when attention weights need to be corrected for various layers.

Overall, the contributions of this paper are as follows:

- We introduce a simple and effective approach to automatically evaluate the influence of image region features with mask operation, and use it as supervised information to guide the attention alignment;
- We design four fusion strategies to force the attention weight incorporating supervised information, which can be easily applied into existing models to improve the performance of captioning;
- We evaluate attention alignment for image captioning on the MS COCO dataset. The captioning models equipped with our method significantly outperform the ones without it. To improve reproducibility and foster new researches in the field, we publicly release the source code and trained models of all experiments.

2 Background

In this paper, we first introduce the basic framework of Transformer (Vaswani et al. 2017) for image captioning briefly, which has an encoder-decoder structure with stacking layers of attention blocks. Each attention block contains multi-head attention (MHA) and feed-forward networks (FFN). To simplify the optimization, shortcut connection and layer normalization are applied after all the MHA and FFN. Generally, given the image region features $x = \{x_1, x_2, \ldots, x_n\}$, visual encoder projects them to hidden states $h = \{h_1, h_2, \ldots, h_n\}$ in latent space, which further feed into the caption decoder to generate the target sentences $y = \{y_1, y_2, \ldots, y_m\}$.

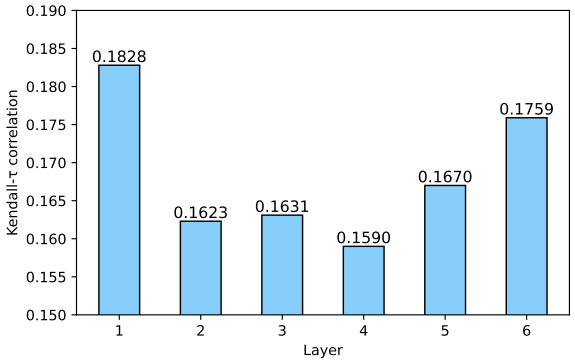

Figure 2: The Kendall-τ correlation between attention weights (α) of image regions and gradient importance metrics (τ) of generated words for different attention layers on the MS COCO validation set.

Multi-head attention, which serves as the core component of the Transformer, enables each prediction to attend overall image region features from different representation subspaces jointly. In practice, hidden states $h = \{h_1, h_2, \ldots, h_n\}$ are projected to keys K and values V with various linear projections. To predict the target word, scaled dot-product attention (Vaswani et al. 2017) is adopted. That is, we first linearly project the hidden state of previous caption decoder layer to the query Q. Then we multiply Q by keys K to obtain an attention weight, which is further used to calculate a sum of values V.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \cdot V, \qquad (1)$$

where d_k corresponds to the dimension of the keys, which is used as scaling factor. Such attention module learns the attended features that consider the pairwise interactions between two features. For MHA, the model, contains several parallel heads, is allowed to attended to diverse information from different representation subspaces. For more advanced improvement, such as mesh-like connectivity and memory module, please refer to (Cornia et al. 2020) in detail. We employ the Transformer of the basic version that performs N = 6 attention layers and employs h = 8 parallel attention heads for each time.

3 Is Current Attention Mechanism in Image Captioning Good Enough?

Attention mechanism plays an essential role in image captioning, which provides an important weight for visual features. However, some researchers have found that the highly attended image regions exist a "deviated focus" problem

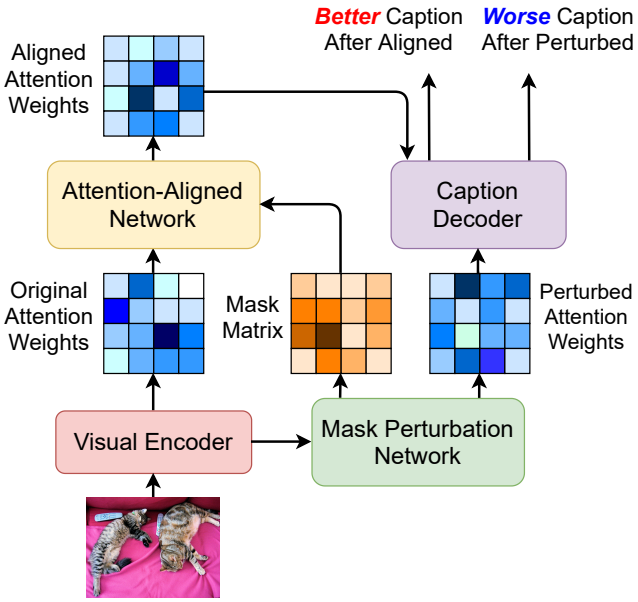

Figure 3: Architecture of the $\mathcal{A}^2$ Transformer. Mask perturbation network is trained to perturb the attention weights of decisive and effective input features to impair the captioning performance. Attention-aligned network targets to look for which input regions are perturbed and enhance the corresponding attention weights.

(Liu et al. 2020) that holds low relevant to generated words, thus impairs the model performance. To make a deeper analysis about if current attention mechanisms can focus on the decisive and effective image regions, we evaluate the correlation with attention weights and feature importance metrics in image captioning. Practically, we refer (Anjomshoae, Jiang, and Framling 2021; Clark et al. 2019) to apply gradient-based methods to evaluate the importance of each visual representation, *i.e.*, hidden state h_i for the generated word y_t, which is estimated as $\tau_{it} = |\bigtriangledown_{h_i} p(y_t|x)|$.

Experimentally, we train a plain Transformer model on MS COCO dataset as the baseline. All the structure and parameter settings are kept untouched as (Chen et al. 2015). We record the average attention weights of image features over various heads, and the Kendall-τ correlation between attention weights and metrics is presented in Figure 2. We can see that the correlation between attention weights of image features and the corresponding gradient importance metrics is weak, all below 0.2. In between, 0 indicates no relevance, while 1 implies strong concordance. The experimental results show that the highly-attended image features are not always responsible for the word generation, which is also consist with previous studies (Liu et al. 2020).

4 Methodology

In this section, to tackle the inaccurate issue of attention weights, we propose a perturbation-based self-supervised method to enhance the attention learning focused on the effective image regions. The basic architecture is shown in Figure 3. Firstly, we introduce how to discover the important image regions for caption generation, where we design a learnable mask perturbation to destroy the description performance with limited operation on the original attention weights. Based on the performance change, we can automatically evaluate the image regions most effected. Then, we illustrate how to use the supervised information to refine the original attention weights with attention-aligned network. Finally, we describe the entire training and inference procedure in detail.

4.1 Learnable Mask Perturbation

The basic assumption of our design is the fact that under the premise of incorporation the same perturbation, important image regions leads to more performance changes than unimportant ones (Li et al. 2021). Specifically, a little perturbation on influential image features can results in a dramatic changes in final generated words, while greater perturbation on the unimportant ones will not easily change the results. Therefore, we can estimate the importance of image region features by observing how the performance changes as perturbing different parts of the input image features. Inspiring from (Fong and Vedaldi 2017; Fan et al. 2021), we apply a learnable mask to scale the attention weight of each image region, which simulates the process of perturbation.

At the time step to generate t-th word, the learnable mask operation m_t is obtained based on the hidden state h_t^d from the d-th layer of caption decoder as:

$$m_t = \sigma(W^m \cdot h_t^d + b^m), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, W^m and b^m are trainable parameters vary among different attention layers and heads. Correspondingly, the perturbed attention weight α_t^p can be modeled based on the mask matrix as:

$$\alpha_t^p = m_t \cdot \alpha_t + (1 - m_t) \cdot \overline{\alpha}, \tag{3}$$

where $\overline{\alpha}$ is an average vector of attention heads rather than zero to avoid the abnormal effect value (Kim et al. 2021). Qualitatively, a smaller value of mask m_t corresponds to a smaller reservation in original attention weight α_t, in other word, a larger perturbation extent.

Recalling that the mask operation is targeted to make the smallest perturbation in image region features and achieves a most extent of performance degrading. Based on this, we can design the training objective of the mask perturbation network as follows:

$$\mathcal{L}(\theta^m) = -\mathcal{L}_{IC}(\alpha_t^p, \theta) + \lambda||1 - m_t||_2^2, \tag{4}$$

where θ denotes the parameters of the original image captioning model. $\mathcal{L}_{IC}(\alpha_t^p, \theta)$ is the loss of the image captioning model when incorporating the perturbed attention weights α_t^p. $\theta^m = \{W^m, b^m\}$ represents the parameters of the mask perturbation network. The second one serves as a regular term to punish too much mask operation and λ is the balancing factor. As the perturbed attention α_t^p is infected by θ^m, both two term in Equation 4 are parameterized with θ^m. Thus, this loss only optimizes the parameter of mask perturbation network θ^m without accessing to the original image captioning model.

4.2 Attention-Aligned Network

According to the analysis above, our mask perturbation network generate feature importance estimation for each word generation, where the perturbation is quantified according to the mask magnitude. Here, we do not use mask matrix to generate a new attention distribution to replace the original attention weights. Rather, we use it as supervised information. We want the model notices more features that have an influence on output. In this way, some ignored image features with great importance can be discovered by attention learning. In the following, we describe how to exploit the mask matrix to guide the alignment of attention.

As the mask value closer to 1 means to keep the original attention weights and make mask operation less, the mask-based attention weights α_t^m based on mask m_t can be designed following (Lu et al. 2021) as:

$$\alpha_t^m = \alpha_t \cdot e^{1-m_t}. \tag{5}$$

In particular, we design four fusion methods to incorporate α_t^m into the original one α_t to obtain the final aligned attention weights α_t^a as follows:

Max Pooling. The most intuitive idea is to replace the original ignored attention with newly highlighted ones:

$$\alpha_t^a = \max(\alpha_t, \alpha_t^m). \tag{6}$$

Moving Average. The mask-based attention weights are linearly added to the original attention weights in the entire process, with a fixed ratio η as:

$$\alpha_t^a = \alpha_t + \eta \cdot \alpha_t^m. \tag{7}$$

Exponential Decay. Inspired by curriculum learning (Bengio et al. 2009), we make the influence of α_t^m to be smaller at the beginning and gradually growing with the training forwards. For simplicity, we utilize exponential decay (Zhou, Wang, and Bilmes 2021) to update ratio of α_t^m as:

$$\alpha_t^a = e^{-\frac{s}{TP}} \cdot \alpha_t + (1 - e^{-\frac{s}{TP}}) \cdot \alpha_t^m, \tag{8}$$

where s is the training step and TP is a temperature factor.

Gating Mechanism. We further employ a learnable gate (Xu et al. 2019) to dynamically control the extent of the supervised information from the mask perturbation network into the aligned attention.

$$\alpha_t^a = g_t \cdot \alpha_t + (1 - g_t) \cdot \alpha_t^m, \tag{9}$$

$$g_t = \sigma(W^g \cdot q_t + b^g), \tag{10}$$

where W^g and b^g are trainable parameters vary among different attention layers and heads, and σ corresponds to sigmoid activation function.

4.3 Training and Inference

For training procedure, we first train the mask perturbation network with loss in Equation 4, then do the optimization process for the attention-aligned network following the loss in Equation 12, and finally optimize the mask perturbation network and image captioning model jointly. Similar to the previous studies (Huang et al. 2019; Cornia et al. 2020), the training of image captioning model is splitted into two stages. In the first stage, with the aligned attention weights α_t^a, the image captioning model is firstly optimized with cross-encropy loss as:

$$\mathcal{L}_{IC}(\theta) = -\sum_{t=1}^{m} \log p(y_t | y_{<t}, x; \alpha_t^a, \theta). \tag{11}$$

In the second stage, the image captioning model is fine-tuned with self-critical reinforcement learning strategy following (Rennie et al. 2017) as:

$$\mathcal{L}_{IC}(\theta) = -\mathbb{E}_{y \sim p(y|x;\alpha_t^a,\theta)}[r(y)], \tag{12}$$

where r denotes the reward function for generated sentence, *e.g.*, CIDEr score (Vedantam, Lawrence Zitnick, and Parikh 2015) in common cases.

During inference procedure, at each time step, given the image and generated sentence, the learned mask perturbation network can determine the most important image regions with mask matrix. Then, the attention-aligned network fuse the original attention weights and mask matrix, and feed aligned attention into caption decoder to make the final decision. More encouragingly, the aligned attention can serve as an effective visual interpretation (Patro, Namboodiri et al. 2020) to qualitative measurement of the captioning model.

5 Experiments

5.1 Experimental Setup

Dataset. All the experiments are conducted on the most popular image captioning dataset MS COCO (Chen et al. 2015). As the largest English dataset, MS COCO contains totally 164,062 images. Each image is equipped with five human-set captions. 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. All the training sentences are preposed by converting them into lower case and dropping the words that occur rarely as (Huang et al. 2019; Cornia et al. 2020). We also evaluate the model on the MS COCO online test server, composed of 40,775 images whose annotations are not made publicly accessible.

Evaluation Metrics. We use five standard automatic evaluation protocol 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), and denoted as B-N, M, R, C and S for simplify. Concretely, BLEU-N indicates the n-gram matching, SPICE is based on scene graph matching, METEOR measures both the precision and recall, and CIDEr considers the n-gram similarity with TF-IDF weights.

Implement Details. We utilize Faster R-CNN (Ren et al. 2015) with ResNet-101 (He et al. 2016) to represent image regions. The feature vector for each region is 2048-dimensional. We employ one-hot vectors and linearly

	Cross-Entropy Loss						CIDEr Score Optimization					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
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
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
DPA (Liu et al. 2020)	-	-	-	-	-	-	-	40.5	29.6	59.2	133.4	23.3
GET (Ji et al. 2021)	-	-	-	-	-	-	81.5	39.5	29.3	58.9	131.6	22.8
$\mathcal{A}^2$ Transformer	78.6	38.2	29.2	58.3	125.0	22.1	81.5	39.8	29.6	59.1	133.9	23.0

Table 1: Performance of $\mathcal{A}^2$ Transformer and other state-of-the-art image captioning models with different evaluation metrics on the MS COCO Karpathy test set. All values are reported as a percentage (%).

project to model input dimensional to represent words. For model structure, we set the dimensionality d of each layer to 512 and the number of heads to 8. We employ a dropout rate of 0.1 after each attention and feed-forward layer. Model is first trained to minimize the negative log-likelihood of the training data following the learning rate scheduling strategy with a warmup equal to 10,00, and then fine-tuned with the CIDEr score using Reinforcement Learning (Rennie et al. 2017) with a fixed learning rate of 5×10^{-6}. We train all models using the Adam optimizer (Kingma and Ba 2014), a batch size of 50 and a beam size of 5. We set the hyperparameter $\eta = 0.1$ in Equation 7 in all experiments.

5.2 Quantitative Analysis

Offline Evaluation. We compare the performance of our attention-aligned approach with those of several recent proposals for image captioning comprehensively. Specifically, we report the results of some competitive models including M2-T, DPA, and GET on the offline MS COCO Karpathy test split. For a fair comparison, the results for each run optimized with both cross-entropy loss and CIDEr score are listed. As presented in Table 1, overall, the proposed $\mathcal{A}^2$ Transformer exhibits better performance than the above models in terms of BLEU-4, METEOR, and CIDEr. In particular, it advances the current state-of-the-art performance on CIDEr by 0.5. Unlike previous attention learning without supervised information, our model yields a prominent improvement by searching the effective image regions with learnable mask perturbation and reasonable guiding strategy. More encouragingly, the additional parameter of the attention-aligned network is negligible, which will not increase much cost to the training and inference process, while gain an obvious improvement.

Online Evaluation. We also evaluate our best variant $\mathcal{A}^2$ Transformer + Gate Mechanism on the official testing set by submitting the ensemble versions, *i.e.*, an average ensemble for four checkpoints trained independently, to the online testing server. The results over official testing images with 5 reference captions (c5) and 40 reference captions (c40) of our approach and the top-performing published works on the leaderboard are reported in Table 2. Note that we do not list pre-training model for fair comparison. As it can be observed, compared to all other popular systems, $\mathcal{A}^2$ Transformer exhibits better performances across most metrics.

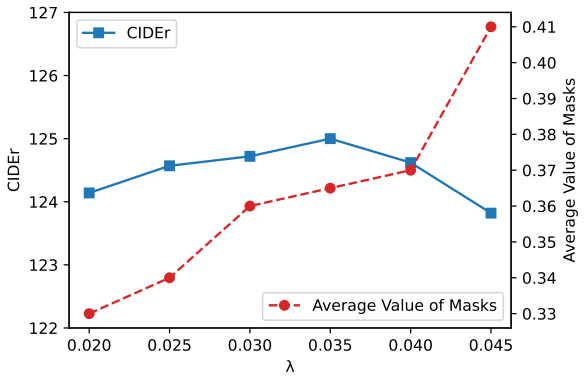

Figure 4: Evaluation metric CIDEr and the average value of generated masks with respect to different hyperparameter λ under gate mechanism on the MS COCO validation set.

5.3 Ablation Study

Effect of Fusion Strategies. As illustrated in Table 3, generally, all of four attention updating methods achieve performance boosting. In between, the moving average show the minimal lifting capacity. Exponential decay, which increase the influence of attention-aligned network after it trained well, provides a more stable training process and better performance. However, the incorporation of hyper-parameter temperature factor S needs more heuristics and increase the uncertainty. Finally, the gate mechanism adaptively controls the ratio between original and updated attention weights, holds more promising for real application.

Effect of Mask Degree. We also analyze the hyperparameter λ in Equation 4, which decides the perturbation degree of mask operation to the original attention weights of image features. We plot the automatic evaluation metric CIDEr changing follows the different of average value of generated masks, by setting different hyperparameter λ, in Figure 4. As we can see, with the increase of mask operation degree, the performance of image captioning model first rise and then fall down. There are optimal parameter options, *i.e.*, $\lambda = 0.035$, under the current scene.

	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
GET	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
$\mathcal{A}^2$ Transformer	82.2	96.4	67.0	91.5	52.4	83.6	40.2	73.8	29.7	39.3	59.5	75.0	132.4	134.7

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

	B-1	B-4	M	R	C	S
Baseline	77.4	37.2	28.7	57.7	120.0	21.8
Max Pooling	77.9	37.7	29.0	58.0	122.8	21.9
Moving Average	77.9	37.6	29.0	57.9	122.3	21.9
Exponential Decay	78.4	37.9	29.0	58.2	124.2	22.0
Gate Mechanism	78.6	38.2	29.2	58.3	125.0	22.1

Table 3: Effect of fusion methods for combined attention weights optimized with cross-entropy loss on validation set.

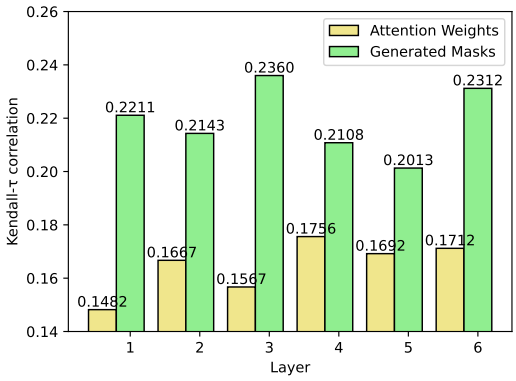

Figure 5: The Kendall-τ correlation between attention weights α and the masks m generated by the mask perturbation model to the gradient importance measures τ on MS COCO validation set.

Accuracy of Mask Operation. To demonstrate the correctness of mask perturbation m, we calculate its correlation with gradient-based importance measures τ (Ross, Hughes, and Doshi-Velez 2017), compared with original attention weights α, the results are shown in Figure 5. It is evident that the generated masks bought out a significant advantage in determining the important image region features, proving the superiority of our self-supervised perturbation network.

5.4 Attention Alignment for Different Layers

We try to explain how our proposed method helps produce better descriptions by investigating which attention weights need to update. Specifically, we dive into the differences between layers, which provide insights into the attention mechanism's inner workings and better understand $\mathcal{A}^2$ Transformer. In practice, we apply Jensen-Shannon Divergence

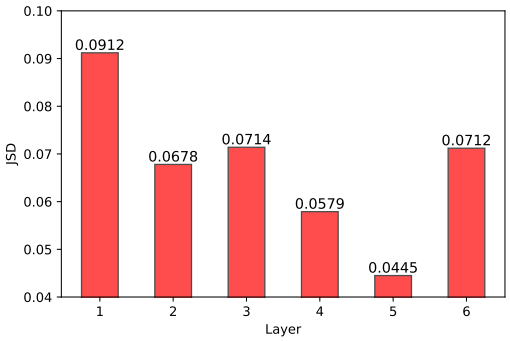

Figure 6: The JSD between attention weights before and after alignment at different layers on MS COCO validation set.

(Fuglede and Topsoe 2004) between attention weights before and after alignment to measure the changing extent as:

$$\mathcal{D}(\alpha, \alpha^a) = \frac{1}{2}\mathrm{KL}[\alpha||\frac{\alpha+\alpha^a}{2}] + \frac{1}{2}\mathrm{KL}[\alpha^a||\frac{\alpha+\alpha^a}{2}], \quad (13)$$

Intuitively, a high JSD score indicates the aligned attention weights are distant from original ones.

Concerning the roles of different attention layers, one natural question is which attention layers are not well-trained in the original image captioning model and have an urgent need to be improved and aligned. Figure 6 depicts the JSD between original and aligned attention weights. We can discover that: i) each attention layer holds an attention change to some degree, ii) high JSD for high layers and low JSD for low layers. These findings prove that different attention layer plays a different role in the image caption generation process. The low layers generally grasp information from various inputs, while the high layers look for some particular elements tied to the final caption generations.

5.5 Qualitative Analysis

Figure 7 showcases several image captioning results from plain Transformer and our $\mathcal{A}^2$ Transformer with gate mechanism, as well as the human-annotated ground truth sentences (GT). Generally, compared with the captions of plain Transformer, which are somewhat relevant to image content and logically correct, our attention-aligned method produces more accurate and richer descriptive sentences by exploiting accurate attention weights for image regions. For example, plain Transformer generates the phrase *wooden bowl* which

GT: two giraffes and another animal in a field
Transformer: three giraffes and other animals in a field
***A*^{2} Transformer:** two giraffes and another animal in a field

GT: a bird in a pot eating a fruit
Transformer: a black bird sitting in a wooden bowl
***A*^{2} Transformer:** a black bird eating in a large pot

GT: a boat with flags and tents is docked next to a grassy bank
Transformer: a boat with a canopy on the water
***A*^{2} Transformer:** a boat with flags sitting on the water

Figure 7: Case studies of original Transformer, plus our $\mathcal{A}^2$ Transformer with gate mechanism, coupled with the corresponding ground truth sentences (GT).

	Transformer wins	Tie	$\mathcal{A}^2$ Transformer wins
Naturalness	25.8	42.0	32.2
Relevance	26.5	45.5	28.0
Richness	20.2	41.4	38.4

Table 4: Results of human evaluation in terms of various metrics. All values are reported as a percentage (%).

is inconsistent with the visual content for the second image, while the words *large pot* in our attention-aligned model depicts more precisely. This again confirms the advantage of capturing accurate attention weights when applying the proposed attention-aligned method.

5.6 Human Evaluation

To better understand the effectiveness of the attention-aligned approach, we conduct a human evaluation to measure the quality of generated captions as (Huang et al. 2019). We randomly select 400 samples from the MS COCO dataset along with human-annotated sentences. We recruit 8 workers to compare the perceptual quality of the caption between our gated $\mathcal{A}^2$ Transformer and original Transformer independently in three 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 4. We can see that $\mathcal{A}^2$ Transformer wins in all metrics than the baseline. In particular, $\mathcal{A}^2$ Transformer with gate mechanism achieves more than 18.2 score in richness. This again confirms that the superiority of attention alignment.

6 Related Works

The attention mechanism is first introduced to augment vanilla recurrent network (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015) in machine translation. For image captioning, Xu et al. (2015) first introduces the visual attention to help the caption decoder focus on the most relevant image regions instead of the whole image; Yao et al. (2017) design an adaptive attention module to decide when to employ the visual attention and Anderson et al. (2018) follows a bottom-up and top-down attention mechanism. Besides, there are numerous other advanced attention mechanisms, *e.g.*, spatial and channel-wise attention (Chen et al. 2017), semantic attention (You et al. 2016) and attention on attention Huang et al. (2019). In recent years, plenty of Transformer-based architectures (Li et al. 2019; Fei 2019; Cornia et al. 2020; Pan et al. 2020; Fei 2021; Yan et al. 2021; Ji et al. 2021) are proposed to replace conventional RNN, achieving new state-of-the-art performances. However, as far as we concerned, improving the attention distribution with self-supervised mask perturbation has never been studied in image captioning task, which push forward our exploration in this paper.

On the other hand, there is plenty of works (Liu et al. 2017; Zhou et al. 2020) found that adding supervision to the attention model is beneficial for the image captioning model. Various approaches have been proposed to improve attention supervision, *e.g.*, referring expression (Liu, Wang, and Yang 2017) grounding visual explanations (Zhou et al. 2020), sparsity regularization (Zhang et al. 2018), and future information (Liu et al. 2020). Unlike them, we never introduce any external knowledge but discover the influential image regions with mask operation. Some works also aim to employ masks as the analytical tools to indicate the importance (Kitada and Iyatomi 2020; Mohankumar et al. 2020), attention head (Fong and Vedaldi 2017), or the contributions of the pixels in the image to the model outputs (Voita et al. 2019). Besides, for other tasks, *e.g.*, visual question answering (Wu and Mooney 2019; Chen et al. 2020), machine translation (He et al. 2019; Lu et al. 2021), and video understanding (Li et al. 2021), similar ideas are also proposed to incorporate mask operation to localize the regions of interest and improve the faithfulness and accuracy of predictions.

7 Conclusion

In this paper, we focus on forcing the image captioning model to attend the import image regions without any extra annotations. To this end, we present $\mathcal{A}^2$ Transformer for effective attention alignment in image captioning. Specifically, a mask perturbation model is applied to automatically discover the decisive and effective image region features based on the description generation changing. Then, we introduce four strategies to combine this supervised information to guide the attention alignment. Extensive experimental results demonstrate that our approaches consistently achieve significant improvements over the state-of-the-art systems. More encouragingly, our work provide valuable reference on self-supervised learning for improved attention in other multi-modal generation frameworks.

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.

Anjomshoae, S.; Jiang, L.; and Framling, K. 2021. Visual explanations for DNNS with contextual importance. In *Proc. ETAAMS*, 83–96. Springer.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proc. ICML*, 41–48.

Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; and Zhuang, Y. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proc. IEEE CVPR*, 10800–10809.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proc. IEEE CVPR*, 5659–5667.

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*.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proc. ACL Workshop*, 276–286.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proc. IEEE CVPR*, 8307–8316.

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

Fan, Z.; Gong, Y.; Liu, D.; Wei, Z.; Wang, S.; Jiao, J.; Duan, N.; Zhang, R.; and Huang, X.-J. 2021. Mask Attention Networks: Rethinking and Strengthen Transformer. In *Proc. NAACL*, 1692–1701.

Fei, Z. 2021. Memory-Augmented Image Captioning. In *Proc. AAAI*, volume 35, 1317–1324.

Fei, Z.-c. 2019. Fast image caption generation with position alignment. *arXiv preprint arXiv:1912.06365*.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proc. IEEE ICCV*, 3429–3437.

Fuglede, B.; and Topsoe, F. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *Proc. ISIT*, 31. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, 770–778.

He, S.; Tu, Z.; Wang, X.; Wang, L.; Lyu, M. R.; and Shi, S. 2019. Towards Understanding Neural Machine Translation with Word Importance. In *Proc. EMNLP*.

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

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proc. AAAI*, volume 35, 1655–1663.

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

Kim, J.; Kim, S.; Kim, S. T.; and Ro, Y. M. 2021. Robust Perturbation for Visual Explanation: Cross-checking Mask Optimization to Avoid Class Distortion. *IEEE Transactions on Image Processing*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitada, S.; and Iyatomi, H. 2020. Attention Meets Perturbations: Robust and Interpretable Attention with Adversarial Training. *arXiv e-prints*, arXiv–2009.

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.

Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled Transformer for Image Captioning. In *Proc. IEEE ICCV*, 8928–8937.

Li, Z.; Wang, W.; Li, Z.; Huang, Y.; and Sato, Y. 2021. Towards visually explaining video understanding networks with perturbation. In *Proc. IEEE CVPR*, 1120–1129.

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

Liu, C.; Mao, J.; Sha, F.; and Yuille, A. 2017. Attention correctness in neural image captioning. In *Proc. AAAI*, volume 31.

Liu, F.; Ren, X.; Wu, X.; Ge, S.; Fan, W.; Zou, Y.; and Sun, X. 2020. Prophet Attention: Predicting Attention with Future Attention for Improved Image Captioning. In *Proc. NIPS*, 1–12.

Liu, J.; Wang, L.; and Yang, M.-H. 2017. Referring expression generation and comprehension via attributes. In *Proc. IEEE ICCV*, 4856–4864.

Lu, Y.; Zeng, J.; Zhang, J.; Wu, S.; and Li, M. 2021. Attention Calibration for Transformer in Neural Machine Translation. In *Proc. ACL*, 1288–1298.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Mohankumar, A. K.; Nema, P.; Narasimhan, S.; Khapra, M. M.; Srinivasan, B. V.; and Ravindran, B. 2020. Towards Transparent and Explainable Attention Models. In *Proc. ACL*, 4206–4216.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *Proc. IEEE CVPR*, 10971–10980.

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.

Patro, B.; Namboodiri, V.; et al. 2020. Explanation vs attention: A two-player game to obtain attention for vqa. In *Proc. AAAI*, volume 34, 11848–11855.

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.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proc. IJCAI*, 2662–2670.

Selvaraju, R. R.; Lee, S.; Shen, Y.; Jin, H.; Ghosh, S.; Heck, L.; Batra, D.; and Parikh, D. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proc. IEEE ICCV*, 2591–2600.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proc. ACL*, 2931–2951.

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.

Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; and Titov, I. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Wu, J.; and Mooney, R. 2019. Self-Critical Reasoning for Robust Visual Question Answering. *Proc. NIPS*, 32: 8604–8614.

Xu, C.; Ji, J.; Zhang, M.; and Zhang, X. 2019. Attention-gated LSTM for Image Captioning. In *Proc. ICUSAI*, 172–177. IEEE.

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.

Yan, X.; Fei, Z.; Li, Z.; Wang, S.; Huang, Q.; and Tian, Q. 2021. Semi-Autoregressive Image Captioning. In *Proc. ACM MM*, 2708–2716.

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.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proc. IEEE CVPR*, 4651–4659.

Zhang, J.; Zhao, Y.; Li, H.; and Zong, C. 2018. Attention with sparsity regularization for neural machine translation and summarization. *Proc. TASLP*, 27(3): 507–518.

Zhang, W.; Shi, H.; Tang, S.; Xiao, J.; Yu, Q.; and Zhuang, Y. 2021. Consensus graph representation learning for better grounded image captioning. In *Proc. AAAI*, 3394–3402.

Zhou, T.; Wang, S.; and Bilmes, J. 2021. Curriculum Learning by Optimizing Learning Dynamics. In *Proc. ICAIS*, 433–441. PMLR.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More grounded image captioning by distilling image-text matching model. In *Proc. IEEE CVPR*, 4777–4786.