sensors

MDPI

Article

Attention-Guided Image Captioning through Word Information

Ziwei Tang, Yaohua Yi * and Hao Sheng

School of Printing and Packaging, Wuhan University, Wuhan 430072, China; tangziwei@whu.edu.cn (Z.T.); 2015301750041@whu.edu.cn (H.S.)
* Correspondence: whudcil@whu.edu.cn

Abstract: Image captioning generates written descriptions of an image. In recent image captioning research, attention regions seldom cover all objects, and generated captions may lack the details of objects and may remain far from reality. In this paper, we propose a word guided attention (WGA) method for image captioning. First, WGA extracts word information using the embedded word and memory cell by applying transformation and multiplication. Then, WGA applies word information to the attention results and obtains the attended feature vectors via elementwise multiplication. Finally, we apply WGA with the words from different time steps to obtain previous word guided attention (PW) and current word attention (CW) in the decoder. Experiments on the MSCOCO dataset show that our proposed WGA can achieve competitive performance against state-of-the-art methods, with PW results of a 39.1 Bilingual Evaluation Understudy score (BLEU-4) and a 127.6 Consensus-Based Image Description Evaluation score (CIDEr-D); and CW results of a 39.1 BLEU-4 score and a 127.2 CIDER-D score on a Karpathy test split.

Keywords: image captioning; word level attention; previous word guidance; current word guidance

check for updates

Citation: Tang, Z.; Yi, Y.; Sheng, H. Attention-Guided Image Captioning through Word Information. *Sensors* **2021**, *21*, 7982. https://doi.org/10.3390/s21237982

Received: 8 October 2021
Accepted: 26 November 2021
Published: 30 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Image captioning is synthetic research that spans computer vision and natural language processing to generate natural descriptions of images. In recent years, image captioning has made great progress with the rapid development of classification [1], object detection [2], and machine translation. However, there are many problems, such as object recognition and interactions, and corresponding relations between objects and words, making it a challenging task [3–7].

Inspired by attention mechanisms [8] and sequence–sequence models [9] exploited in machine translation tasks, an encoder–decoder framework [10–14] has been widely used for image captioning. In such a framework, images are encoded to feature vectors by a pretrained image classification model, object detection model, or semantic segmentation model, and then decoded to words via an RNN. Within the RNN network [15], the decoder process is implemented as a sequence to generate words one by one. Until the attention mechanism was proposed, there was little optimization for this framework. The attention mechanism [16], which comes from machine translation, can guide the generation of words by weighting a feature to connect a portion of an image with a word at each time step.

Currently, an attention mechanism is widely applied in image captioning [17] systems. Because the attention directly determines the caption of an image, the inference direction determined by attention modules must be correct. However, attention is concentrative and superficial in general. It is prone to results in the decoder knowing little or mistaking the objects of the image, such as the "*dog*" in Figure 1b. In detail, the decoder may be misled to list nouns into generated sentences simply and ignore the relationships among the objects, for example the relation between "*person*" and "*motorcycle*" in Figure 1a and what the "*dog*" is doing in Figure 1b. Moreover, an attended region only represents one word, which means that the decoder may overlook the details of an object, for instance the word "*little*" depicting the "*girl*" in Figure 1c.



Captions: A person on a motorcycle on a road.	Captions: A dog and a banana on a couch.	Captions: A little girl in a field with a frisbee.
(a)	(b)	(c)

Figure 1. Illustration of some defects that attention may produce without guidance. We present some of them, such as a simplistic logical description (**a**), incorrect object recognition (**b**), and poor details of a target (**c**). The words in red are the incorrect issues.

To address this issue, we propose word guided attention (WGA), which is created from word information, to bring novel specific guidance to the decoder. First, we design new information processing for words with several transformations and activation functions, which is similar to GLU [18]. The information processing method includes memory cell weighting, embedded words, and basic attention. Based on this process, we construct a WGA module in the decoder. Subsequently, we utilize the WGA and propose respective methods for the different step words. Fused with the previous step word, the previous word guided attention (PW) can be achieved. In addition, the current step word constitutes the current word guided attention (CW).

In this paper, we apply self-attention [19] as the basic attention unit for both the encoder and decoder phases. In the encoder, self-attention is used to build relations among the objects by weighting feature vectors extracted from an image. In the decoder, self-attention points out the major objects in an image and plays a guided role in PW/CW. Furthermore, we propose PW to expand the scope of objects describing and intensify the relationships via word-level attention. On the other hand, CW is concentrated on the current saliency region to obtain more detailed content and deeper relations.

We evaluate our method on the MSCOCO dataset and perform quantitative and qualitative analysis. The results show that our WGA is effective. The proposed PW/CW model is superior to other published image caption models. The main contributions of our paper include the following:

- We propose a novel word guided attention module for image captioning to determine the relationships among the attention features of an encoded image.
- We use the WGA with the previous step word and the current step word. With the previous word, the WGA concentrates on covering more objects in the scene and describing the relevance among them. With the current step word, the WGA is devoted to obtaining more details and deeper relation information from the current attention region.

2. Related Work

2.1. Image Captioning

Recent image captioning approaches are based on the framework of an encoder–decoder, which benefits from the development of deep learning and machine translation [8]. For example, an end-to-end CNN-LSTM framework is proposed to encode an image into CNN feature vectors and decode them into a sentence [20]. In [21], high-level semantic information is proposed as a CNN-LSTM framework. In [22], a two-layer LSTM is applied to give attention to a performing stage. Moreover, some complicated information, such as attributes and relationships, is integrated to improve the generated captions covering an image more completely [23–25].

2.2. Attention Mechanism

The attention mechanism [26], which originates from simulating the human perception system, has been widely employed and has made a great progress for seq-to-seq tasks. In image captioning, attention is an essential part of the model. In [16], a weighted candidate vector is proposed to teach the decoder focusing on the right fields in an image using normalization and SoftMax function. Since then, many studies on attention mechanisms have emerged for image captioning, such as adaptive attention [27], Sca-cnn [28].

In machine translation, we can obtain much inspiration for image captioning. For instance, novel attention achieved from the words is proposed in [29]. In [19], self-attention is proposed and obtains state-of-the-art results.

3. Methods

We first introduce the WGA module. Then we present how the WGA works for different image captioning phases.

3.1. WGA

A basic attention unit $f_{Att}(.)$ provides the weighted feature vectors $\hat{V}$ for queries, keys, and values (denoted as Q, K, and V, respectively) by some operations, as shown in Figure 2a. First, we operated on Q, K, and V to linearize them independently. Then, the similarity weight between Q and K was measured by the dot-product, exerting scale correction and the SoftMax function. Finally, we performed matrix multiplication between the similarity weight and V. Thus, a basic attention unit $\hat{V} = f_{Att}(Q,\ K,\ V)$ can be formulated as

$$sim_{i,j} = \left(W_Q q_i + b^Q\right) \odot \left(W_K k_j + b^K\right), \tag{1}$$

$$\alpha_{i,j} = \frac{sim_{i,j}}{\sqrt{d_{k_j}}}, \tag{2}$$

$$\hat{v}_i = \sum_j \frac{e^{\alpha_{i,j}}}{\sum_j e^{\alpha_{i,j}}} \left(W_V v_j + b^V\right), \tag{3}$$

where $q_i \in Q$, $k_j \in K$, and $v_j \in$ V, W_Q, W_K, $W_V \in \mathbb{R}^{D\times D}$, $b^* \in \mathbb{R}^D$ are the linear transformation, bias of queries, keys, and values respectively, and D is the dimension of a vector. $sim_{i,j}$ denotes a function to calculate the similarity score between q_i and k_j via the dot-product $\odot$, and $\hat{v}_i \in \hat{V}$ is the attended feature vector.

A basic attention unit outputs the preliminarily attended feature vectors, which can direct a language model to generate more nouns and be effective building their relationships. However, it is too compulsive to obtain correlations among irrelevant objects and may ignore some inconspicuous objects.

Therefore, we proposed the WGA module $f_{WGA}(.)$, as shown in Figure 2b, to extract guiding information from a generated word. The WGA module generates a word guiding weighting (WGW) β_i, which is extremely conditioned on the attention unit result $\hat{V}$ and adopts elementwise multiplication for β_i and $\hat{V}$ to output word guided attended feature vectors V' using a residual connection. $V' = f_{WGA}(Q,\ K,\ V,\ X,\ M)$ can be determined by

$$\beta_i = f_{WGW}(\hat{v}_i, W_e x,\ m_{\mathrm{T}}), \tag{4}$$

$$v'_i = W_\lambda \left\{ \left[\hat{v}_i,\ \left(W_\phi(\hat{v}_i \cdot \beta_i) + b^\phi\right)\right] + b^\lambda \right\}, \tag{5}$$

where $v'_i \in V'$, $f_{WGW}(.)$ denotes the process of WGW, [., .] is a function for concatenating two vectors, $x \in X$ defines the embedded word, $m_{\mathrm{T}} \in M$ is a memory cell, and T is the set of time steps. W_λ, $W_\phi \in \mathbb{R}^{D\times D}$, $b^* \in \mathbb{R}^D$ are two linear transformation matrixes and W_e is the word embedding matrix.

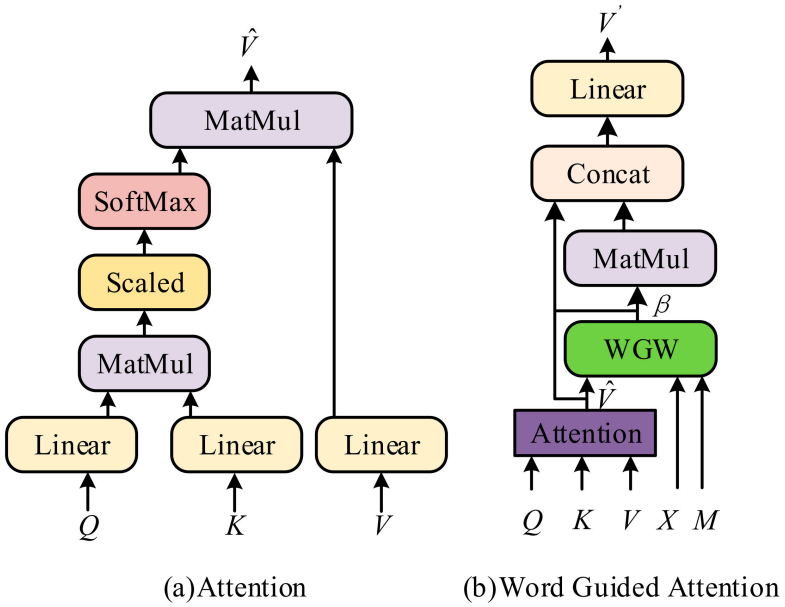


Figure 2. Illustration of a basic attention unit and WGA. (**a**) This unit produces the average weighted feature vectors by linear transformation, scaled dot-product attention, and SoftMax. (**b**) The WGA combines attention with word guided weighting (WGW) via elementwise multiplication.

Then, the WGW $f_{WGW}(.)$ is central to the WGA to obtain the guiding weight information from a word. As shown in Figure 3, WGW first employs a memory cell M and the individual word X to strengthen the influence of generated words for the subsequent sentence content via an activation function and elementwise multiplication. Then, WGW utilizes linear transformation and merges word context information g_{T}^{x} with $\hat{v}_i$ Finally, the SoftMax function is used to obtain the weighting β_i. $\beta_i = f_{WGW}(\hat{v}_i,\ W_e x_{\mathrm{T}},\ m_{\mathrm{T}})$ and can be represented as

$$g_{\mathrm{T}}^{x} = W_e x_{\mathrm{T}} \tanh(m_{\mathrm{T}}), \tag{6}$$

$$\hat{\beta}_i = W_{\hat{\beta}_i} \tanh\Big(\Big[W_{\hat{v}}\hat{v}_i + b^{\hat{v}},\ W_g g_{\mathrm{T}}^{x} + b^{g}\Big]\Big) + b^{\hat{\beta}_i}, \tag{7}$$

$$\beta_i = \frac{e^{\hat{\beta}_i}}{\sum_i e^{\hat{\beta}_i}}, \tag{8}$$

where $W_{\hat{\beta}_i}$, $W_{\hat{v}}$, $W_g \in \mathbb{R}^{D\times D}$, and $b^{*} \in \mathbb{R}^{D}$ are matrixes of linear transformation and corresponding biases. The tanh denotes an activation function.

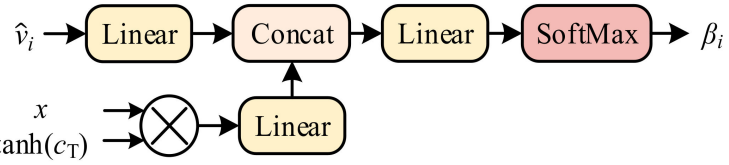


Figure 3. The WGW module in the WGA, where the embedded word and memory cell refine the feature vectors by strengthening the connection with existing words.

3.2. Image Captioning Model

We took advantage of the WGA for image captioning model based on the two-LSTM encoder–decoder framework.

For an image I, CNN feature vectors $A = \{a_1, a_2, \ldots, a_n\}$ are extracted, where $a_n \in \mathbb{R}^{D}$, $n \in N$ and N is the number of feature vectors in the image. In encoder phase *Enc*, we

not only obtained the feature vectors, but also fed A to the basic attention unit (Figure 2a) $f_{Att}(Q,\ K,\ V)$, where Q, K and V are inputs of A. Enc can be formulated as

$$\hat{A}_E = Enc\left(f_{Att},\ A^Q,\ A^K,\ A^V\right), \tag{9}$$

where $\hat{A}_E = \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_n\}$ is the encoded result of an image.

In the decoder, we generated a sequence of captions $\boldsymbol{y}$ with the encoding results $\hat{A}_E$. The two-LSTM framework is composed of a language LSTM $LSTM^{lang}$ and an attention LSTM $LSTM^{Att}$, as shown in Figure 4.

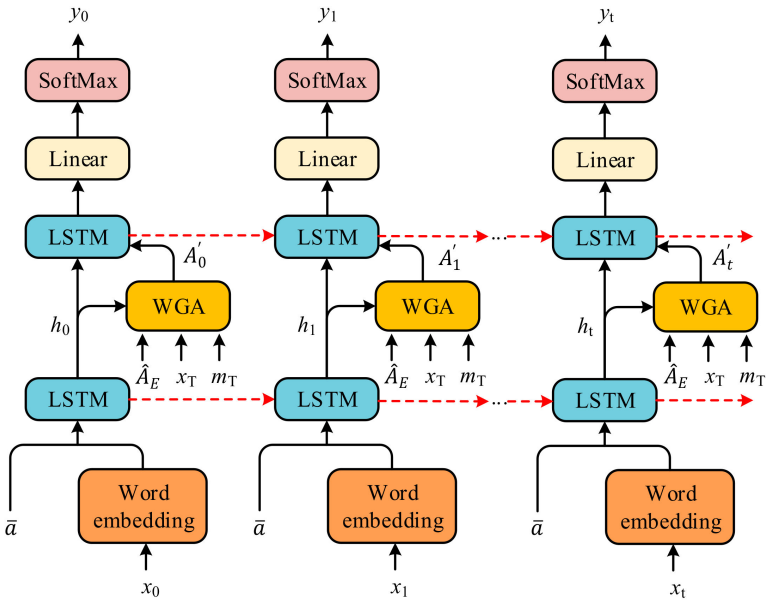


Figure 4. The framework of the decoder, which includes two LSTMs, a WGA module, and a prediction of word probability distribution.

We input the mean pooling visual vector $\bar{a} = \frac{1}{n}\sum \hat{A}_E$ and the t-th time step embedded word $W_e x_t$ to the attention LSTM $LSTM^{Att}$, which can be defined as

$$h_t^{Att},\ m_t^{Att} = LSTM^{Att}\left([W_e x_t,\ \bar{a}],\ h_{t-1}^{Att},\ m_{t-1}^{Att}\right), \tag{10}$$

where $t \in \mathrm{T}$, and $h_{\mathrm{T}}^{Att},\ m_{\mathrm{T}}^{Att} \in \mathbb{R}^D$ are the hidden state and memory cell of the $LSTM^{Att}$ respectively.

To make the WGA produce a marked effect in decoder, we inserted it between the $LSTM^{lang}$ and $LSTM^{Att}$ to guide the language model. As shown in Figure 4, A'_t can be obtained from $f_{WGA}(.)$ fed with $\hat{A}_E$, $W_e x_{\mathrm{T}}$, m_{T} and h_t^{Att}, which is formulated as

$$A'_t = f_{WGA}\left(h_t^{Att},\ \hat{A}_E^K,\ \hat{A}_E^V,\ W_e x_{\mathrm{T}},\ m_{\mathrm{T}}\right), \tag{11}$$

where Q, K, and V are replaced with h_t^{Att} and double $\hat{A}_E$. In addition, $W_e x_{\mathrm{T}}$, m_{T} are discussed after the language LSTM model.

The input to the $LSTM^{lang}$ is the concatenation of the WGA weighted feature vectors A'_t and the current hidden state of the $LSTM^{Att}$. Therefore, the $LSTM^{lang}$ can be presented as

$$h_t^{lang},\ m_t^{lang} = LSTM^{lang}\left(\left[A'_t,\ h_t^{Att}\right],\ h_{t-1}^{lang},\ m_{t-1}^{lang}\right), \tag{12}$$

where $t \in \mathrm{T}$ and $h_{\mathrm{T}}^{lang},\ m_{\mathrm{T}}^{lang} \in \mathbb{R}^D$ are the hidden state and memory cell of the $LSTM^{lang}$, respectively.

Therefore, we can achieve a probability distribution $\boldsymbol{y}_t$ of the caption prediction at time step t:

$$p_\theta(\boldsymbol{y}_t|\boldsymbol{y}_{1:t-1}) = \text{softmax}\Big(W_h h_t^{lang}\Big), \tag{13}$$

where $W_h \in \mathbb{R}^{D\times D}$, $\boldsymbol{y}_{1:\mathrm{T}}$ refers to the generated captions, and SoftMax is the activation function.

As stated earlier, we fed different $W_e x_\mathrm{T},\ m_\mathrm{T}$ into the WGA to realize different generation improvements, as shown in Figure 5.

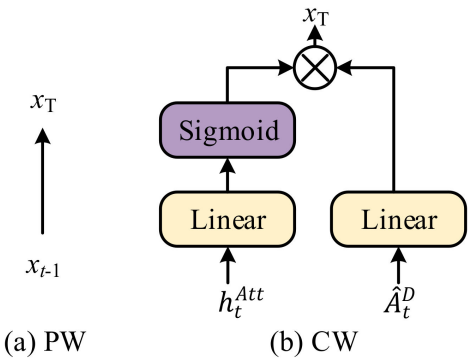


Figure 5. The illustration of PW and CW. (**a**) In PW, word information comes from the previous step and is applied briefly. (**b**) In CW, word information is provided from the current step attention LSTM and refined by linear transformation and sigmoid.

Previous word guided attention. To better describe the entire scene, we made use of previous word information $W_e x_{t-1}$ and memory cell m_{t-1}^{lang} (Figure 5a):

$$W_e x_{t-1},\ m_{t-1}^{lang} \rightarrow W_e x_\mathrm{T},\ m_\mathrm{T}. \tag{14}$$

We believe that the word information from the previous step can protect the logical correctness for generation of the current word. Furthermore, the information is a summary of the previous attention region. It can guide the model to select the correct attention region for the current step, so that some neglected attention regions can be effectively utilized, and the WGA can cover more objects and relations.

Current word guided attention. For a more detailed description of the current attention region, we applied the current memory cell m_t^{Att} of the $LSTM^{Att}$ and simulated generating the current word via h_t^{Att} and gated linear units (Figure 5b):

$$\hat{A}_t^D = f_{Att}\Big(h_t^{Att},\ \hat{A}_E^K,\ \hat{A}_E^V\Big), \tag{15}$$

$$X_c = \Big(W_{Att}\hat{A}_t^D + b^{Att}\Big) \times \delta\Big(W_{gate}h_t^{Att} + b^{gate}\Big), \tag{16}$$

$$X_c,\ m_t^{Att} \rightarrow W_e x_\mathrm{T},\ m_\mathrm{T}, \tag{17}$$

where $\hat{A}_t^D$ means the result of attended $\hat{A}_E$ by basic attention unit $f_{Att}(.)$, $W_{Att},\ W_{gate} \in \mathbb{R}^{D\times D}$, $b^* \in \mathbb{R}^D$, and δ is a sigmoid function. We think that the current word information can help the model focus on the current attention region that is significant by weighting the feature vectors. Thus, the salience region can provide more details that are not only about objects but also deeper relations or status.

3.3. Training and Objectives

Training with cross-entropy loss. We first trained the image captioning model using cross-entropy loss L_{XE}:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log\big(p_\theta(\boldsymbol{y}_t^*|\boldsymbol{y}_{1:t-1}^*)\big), \tag{18}$$

where $\boldsymbol{y}^*_{1:\mathrm{T}}$ refers to the ground truth of captions.

Optimize using the CIDEr-D score. Then, we followed the approach of self-critical sequence training (SCST) [30] to optimize the model:

$$L_{RL}(\theta) = -\mathbf{E}_{\boldsymbol{y}_{1:\,\mathrm{T}} \sim p_\theta}[r(\boldsymbol{y}_{1:\,\mathrm{T}})], \tag{19}$$

where the reward $r(.)$ is calculated by the metrics of Consensus-Based Image Description Evaluation score (CIDEr-D) [31]. The gradient is defined as

$$\nabla_\theta L_{RL}(\theta) \approx -(r(\boldsymbol{y}^s_{1:\,\mathrm{T}}) - r(\hat{\boldsymbol{y}}_{1:\,\mathrm{T}}))\nabla_\theta \log p_\theta(\boldsymbol{y}^s_{1:\,\mathrm{T}}), \tag{20}$$

where $\boldsymbol{y}^s$ is the result from the sampled probability, and $\hat{\boldsymbol{y}}$ is the result of the greedy algorithm.

4. Experiments

4.1. Dataset

The proposed method was implemented on the MSCOCO [32] dataset. The MSCOCO dataset contains 123,287 images, including 82,783 training images and 40,775 validation images, with 5 captions for each. The Karpathy split [33] was adopted to obtain 113,287 images for training, 5000 images for validation, and 5000 images for testing. We collected words occurring more than 4 times in all sentences from the MSCOCO dataset and obtained a dictionary of 10,369 words. The metric we used to evaluate our method was Bilingual Evaluation Understudy score (BLEU-*N*) [34], which can be calculated as

$$\mathrm{BLEU} - N = \min\left(1,\ e^{1-\frac{r}{c}}\right) \cdot e^{\frac{1}{N}\sum_{n=1}^{N} \log p_n}, \tag{21}$$

where r and c denote ground truth captions and generated captions, respectively, and p_n is the *n-gram* precision. In addition, we also adopted CIDEr-D [31], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [35], Semantic Propositional Image Caption Evaluation (SPICE) [36], and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [37] to evaluate our method. These values calculate the similarity between the generated caption and the ground truth caption, and higher values represent better results. We computed this metrics with the public code (https://github.com/tylin/coco-caption, accessed on 25 September 2021) from the MSCOCO dataset.

4.2. Implementation Details

The Faster R-CNN model [2] trained on ImageNet [38] and Visual Genome [39] was exploited to extract bottom-up feature vectors from images. These vectors have 2048 dimensions and were transformed to 1024-D vectors to match the hidden size of the LSTM in the decoder. In the training phase with cross-entropy loss, we adopted an initial learning rate of 4e–4 decaying 0.8 every 2 epochs, and the ADAM optimizer was employed over a total of 30 epochs. For training with CIDEr-D score optimization, we set the initial learning rate to 4e–5 and decayed it by 50% when the performance of the validation split was never better in another 20 epochs. In addition, we set the image batch size to 10 during training, and the beam size was 2 while testing.

4.3. Quantitative Analysis

To validate the performance of our method, we gathered some results based on the Karpathy split test from other methods. These methods are based on certain well-known frameworks or improved attention, including LSTM [20], which encodes an image into CNN features and decodes them to a series of words using LSTM; SCST [30], which proposes a sequence train with evaluation metrics using reinforcement learning; Adaptive-Attention [27], which proposes an adaptive attention model with a visual sentinel; RFNet [40], which proposes a novel recurrent fusion network to exploit multiple information from encoders; UpDown [22], which applies two LSTMs to weight bottom-up

image feature; research [41,42] that contributes new attention using semantic-enhanced image features and spatiotemporal memory, respectively; research [43] that provides special decoding phase improved by a ruminant mechanism; research [44] that leverages object attributes to structure linguistically-aware attention for the lack of high-level understanding; VRAtt-Soft [45], which proposes novel visual relationship attention via contextualized embedding for individual regions; and research [46] that extends the caption model by incorporating extra explicit knowledge from a memory bank. The results shown as percentage are presented in Tables 1 and 2.

Table 1. The image captioning results of our method and others on the MSCOCO Karpathy test split with cross-entropy loss.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
LSTM [20]	-	-	-	29.6	25.2	52.6	94.0	-
SCST [30]	-	-	-	30.0	25.9	53.4	99.4	-
Adaptive-Attention [27]	73.4	56.6	41.8	30.4	25.7	-	102.9	-
RFNet [40]	76.4	60.4	46.6	35.8	27.4	56.5	112.5	20.5
UpDown [22]	77.2	-	-	36.2	27.0	56.4	113.5	20.3
Att2in+RD [43]	-	-	-	34.3	26.4	55.2	106.1	19.7
UpDown+STAM [41]	77.4	61.5	47.6	36.5	27.4	56.8	114.4	20.5
Ours: PW	77.4	61.5	47.7	36.8	28.1	57.3	117.0	21.2
Ours: CW	77.2	61.5	47.8	36.9	28.0	57.2	117.4	21.1

Table 2. The image captioning results of our method and others on the MSCOCO Karpathy test split after CIDEr-D score optimization.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
LSTM [20]	-	-	-	31.9	25.5	54.3	106.3	-
SCST [30]	-	-	-	34.2	26.7	55.7	114.0	-
RFNet [40]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
UpDown [22]	79.8	-	-	36.3	27.7	56.9	120.1	21.4
Cai et al. [42]	80.0	64.3	49.6	37.5	28.2	58.2	126.0	21.8
UpDown+RD [43]	80.0	-	-	37.8	28.2	57.9	125.3	-
UpDown+STAM [41]	80.2	64.4	49.7	37.7	28.2	58.1	125.9	21.7
UpDown+LAT [44]	80.4	-	-	37.7	28.4	58.3	127.1	22.0
VRAtt-Soft [45]	80.2	63.3	48.7	37.3	28.4	61.4	121.8	21.8
UpDown+MA [46]	80.2	-	-	37.5	28.4	58.2	125.4	22.0
Ours: PW	80.4	65.1	50.8	39.1	28.7	58.7	127.6	22.2
Ours: CW	80.6	65.2	50.9	39.1	28.7	58.8	127.2	22.1

We report the performance of the methods with cross-entropy loss in Table 1, and it can be seen that our PW is superior to other methods in all metrics. CW can also achieve the same performance compared with others and is slightly better than PW. In addition, we present the comparison among methods trained with cross-entropy loss and optimized via CIDEr-D score optimization in Table 2. The results demonstrate that CW achieves the optimal performance among all methods, and PW is second best. Furthermore, we collect four models with different initial training parameters to perform ensemble evaluation and the comparison is in Table 3. Our method obtains satisfactory results in contrast to others.

Table 3. The performance of our ensemble model.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
SCST [30]	35.4	27.1	56.6	117.5	-
RFNet [40]	37.9	28.3	58.3	125.7	21.7
Yan et al. [47]	38.4	27.8	57.9	121.6	21.5
Ours: PW	39.6	28.7	59.1	128.3	22.2
Ours: CW	39.8	28.8	59.4	128.3	22.2

4.4. Qualitative Analysis

We reported some examples of images and corresponding captions gathered from our PW, CW, the baseline, and ground truth. Note that we reimplemented the UpDown [22] model sharing the parameters of our models as a baseline. From Table 4, where we mark the improvements in blue, we found that the baseline rigidly describes the prominent objects without exact relations between objects and a detailed depiction of objects. Moreover, our models were superior in two ways: a) Our models focus on the whole image and obtain nearly all components in an image. For the first example, the baseline just recognizes "*women*" and a shallow relation but ignores others such as wine and glasses. In contrast, the captions from PW contain more objects and count more correctly, including "*wine*" and "*glasses*", and CW also contains the background of "*room*". The other three examples can confirm this conclusion. b) Our models can obtain the connections between objects in greater scope and depth with PW and CW. As seen in the first example, PW achieves objects of "*two women*" and "*wine glasses*" and then prefers to build the relationship between "*women*" and connect "*women*" with "*wine glasses*" by "*holding*". Meanwhile, CW can guide the model in another direction. To use the same example, CW determines the relationship among "*women*", "*wine*", and "*glasses*" and then describes it with "*drinking*". In other examples, we can find the same conclusion. PW and CW have these superiorities because they can guide the model to distribute attention for different purposes. As we can find in Table 4, PW and CW are experts in building relations between objects due to the self-attention and basic WGA. PW can determine how the model covers the objects in an image, and CW can more deeply assess the details of the current attention region, which we will show in Section 4.5.

Table 4. Samples of image captions generated by our PW, CW, and baseline as well as ground truths.

Image	Captions
	**Baseline:** A couple of women standing next to each other. **Our PW:** Two women standing next to each other holding wine glasses. **Our CW:** Two women drinking wine in a room. GT1: Two young women are sharing a bottle of wine. GT2: Two female friends posing with a bottle of wine. GT3: Two women posing for a photo with drinks in hand.
	**Baseline:** A group of people walking down a street. **Our PW:** A group of people standing in the street with an umbrella. **Our CW:** A group of people standing under an umbrella. GT1: Several people standing on a sidewalk under an umbrella. GT2: Some people standing on a dark street with an umbrella. GT3: Some people standing on a dark street with an umbrella.
	**Baseline:** A close up of a horse in a field. **Our PW:** A white horse standing in the grass in a field. **Our CW:** A white horse grazing in a field of grass. GT1: A horse eating grass in a green field. GT2: A while horse bending down eating grass. GT3: A tall black and white horse standing on a lush green field.
	**Baseline:** A group of people on skis in the snow. **Our PW:** A group of people riding skis down a snow covered slope. **Our CW:** Two men are skiing down a snow covered slope. GT1: Two cross country skiers heading onto the trail. GT2: Two guys cross country ski in a race. GT3: Skiers on their skis ride on the slope while others watch.

4.5. Ablative Studies

To quantify the influence of our WGA models, we compared PW and CW against other methods with the same training phase. First, the UpDown method was defined as the baseline model, which adopted two LSTMs and attention to generate captions. Second, we employed the metrics of B@1, B@4, ROUGE-L, and CIDEr-D to evaluate the models trained after CIDEr-D score optimization. Finally, we refer to self-attention as **self-att**, the encoder phase as ***Enc***, and the decoder phase as ***Dec*** in Table 5.

Table 5. Ablation studies for our method. All results are obtained after CIDEr-D score optimization.

Model	BLEU-1	BLEU-4	ROUGE-L	CIDEr-D
Baseline	79.4	36.7	57.6	122.7
+self-att(*Dec*)	80.0	37.6	58.0	124.7
+self-att(*Enc*+*Dec*)	79.9	38.4	58.4	125.8
Full: PW	80.4	39.1	58.7	127.6
Full: CW	80.6	39.1	58.8	127.2

Effect of self-attention. To evaluate the influence of self-attention, we successively extended the baseline with self-attention on the decoder and encoder. In the decoder, self-attention was located between the two-layer LSTM and became the backbone to build residual construction. From Table 5, we observe that replacing the original attention with self-attention brings benefits and improves the B@4 and CIDEr-D scores of the baseline by 0.9 and 2.0, respectively. In the encoder phase, we utilized self-attention to highlight the principal parts of the image. In Table 5, we can easily conclude that the weighted feature representations give the model effective impact. The results show that further self-attention improved the B@4 and CIDEr-D scores by 0.8 and 1.1, respectively.

Effect of word guided attention. We made further efforts to conduct experiments to test performance of the PW and CW modules. These two were designed following self-attention and constitute the WGA model during the decoding phase. In Table 5, we obtain a B@4 increase of 0.7 and a CIDEr-D increase of 1.8 for PW. On the other hand, CW ameliorated B@4 and CIDEr-D scores of **self-att(*Enc*+*Dec*)** by 0.7 and 1.4, respectively. Unfortunately, the B@4 and CIDEr-D scores were 38.8 and 126.2, respectively, when combining PW with CW by contacting themselves. We think that PW and CW guide the model in distinct directions and mistake the inference results. Even so, the basic word guided module still works, which is the reason why it is always better than **self-att(*Enc*+*Dec*)**.

To qualitatively present the influence of WGA, we visualized the sentences generated by the ablated models. Referring to Table 4, we present the ablation results in Table 6. As we can see, the sentence of an image was increasingly abundant from the baseline to PW/CW. The captions of PW/CW also confirmed their model characteristics. For example, PW and CW obtained different styles of improvement compared with the caption of **self-att(*Enc*+*Dec*)** in the last example. PW added a "*skis*" object and built the connection between "people" and "skis", described as "*riding skis down*". CW replaced "*A group of people*" with "*Two men*" for more detailed information and deep captioning.

Table 6. Visualization of the generated captions of the ablated models, where the colored words are the improvements from the previous caption.

Image	Captions
	**Baseline:** A couple of women standing next to each other. **+self-att(Dec):** A couple of women standing next to each other. **+self-att(Enc+Dec):** Two women are holding wine glasses in a room. **Our PW:** Two women standing next to each other holding wine glasses. **Our CW:** Two women drinking wine in a room.
	**Baseline:** A group of people walking down a street **+self-att(Dec):** A group of people standing in the street. **+self-att(Enc+Dec):** A group of people standing with an umbrella. **Our PW:** A group of people standing in the street with an umbrella. **Our CW:** A group of people standing under an umbrella.
	**Baseline:** A close up of a horse in a field. **+self-att(Dec):** A horse standing in a field. **+self-att(Enc+Dec):** A horse in the grass in a field. **Our PW:** A white horse standing in the grass in a field. **Our CW:** A white horse grazing in a field of grass.
	**Baseline:** A group of people on skis in the snow. **+self-att(Dec):** A man riding skis in the snow. **+self-att(Enc+Dec):** A group of people skiing down a snow covered slope. **Our PW:** A group of people riding skis down a snow covered slope. **Our CW:** Two men are skiing down a snow covered slope.

5. Conclusions

In this paper, we propose a novel attention guided by word information (WGA) for image captioning, which is aimed at extracting more valuable information from the images. The proposed attention contains a novel word guiding weighting (WGW), which is built upon the extended word information, and a residual structure. Therefore, the WGA can provide the various semantic information to address the lack of objects and image details for the captioning model. After that, we propose different applications of WGA for the decoder and obtain the previous word guided attention (PW) and current word attention (CW) with different timing. We demonstrate that PW can expand the insight of the model to cover more objects in the image, and CW can focus on the current region to extract further information. More remarkably, we achieve competitive performance against other methods, and experimental results conclude that our proposed method has stability and universality.

In the future, we will explore how to fuse word information in the encoder to guide the captioning model. For the PW, the key is to find out the breakpoint in which the word information can be embedded. On the other hand, how to simulate word information in the current time step to construct CW is an issue. In addition, the gap between the image feature and word information needs to be solved.

Author Contributions: Conceptualization, Z.T.; methodology, Z.T.; writing—original draft preparation, Z.T.; data curation, H.S.; writing—review and editing, Z.T.; project administration, Y.Y.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number No. 2021YFB220620.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://cocodataset.org/ (accessed on 25 September 2021).

Acknowledgments: The numerical calculations in this paper were conducted on the supercomputing system in the Supercomputing Center of Wuhan University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
3. Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Baby talk: Understanding and generating simple image descriptions. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1601–1608.
4. Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
5. Yang, Y.; Teo, C.L.; Daumé, H.; Aloimonos, Y. Corpus-guided sentence generation of natural images. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 444–454.
6. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; Daumé, H. Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 747–756.
7. Zhang, L.; Zhang, Y.; Zhao, X.; Zou, Z. Image captioning via proximal policy optimization. *Image Vis. Comput.* **2021**, *108*, 104126. [CrossRef]
8. Cho, K.; Merrienboer, B.v.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv Preprint* **2014**, arXiv:1406.1078.
9. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2, Cambridge, MA, USA, 8–13 December 2014; MIT Press: Montreal, QC, Canada, 2014; pp. 3104–3112.
10. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.W.; Salakhutdinov, R.R. Review networks for caption generation. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Curran Associates Inc.: Barcelona, Spain, 2016; pp. 2369–2377.
11. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic Compositional Networks for Visual Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1141–1150.
12. Chen, Y.; Wang, S.; Zhang, W.; Huang, Q. *Less Is More: Picking Informative Frames for Video Captioning*; Springer International Publishing: Cham, Switzerland, 2018; pp. 367–384.
13. Gan, C.; Gan, Z.; He, X.; Gao, J.; Deng, L. StyleNet: Generating Attractive Visual Captions with Styles. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 955–964.
14. Li, R.; Liang, H.; Shi, Y.; Feng, F.; Wang, X. Dual-CNN: A Convolutional language decoder for paragraph image captioning. *Neurocomputing* **2020**, *396*, 92–101. [CrossRef]
15. Wang, H.; Wang, H.; Xu, K. Evolutionary recurrent neural network for image captioning. *Neurocomputing* **2020**, *401*, 249–256. [CrossRef]
16. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37, Lille, France, 7–9 July 2015; JMLR.org: Brookline, MA, USA, 2015; pp. 2048–2057.
17. Xiao, F.; Gong, X.; Zhang, Y.; Shen, Y.; Li, J.; Gao, X. DAA: Dual LSTMs with adaptive attention for image captioning. *Neurocomputing* **2019**, *364*, 322–329. [CrossRef]
18. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language Modeling with Gated Convolutional Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 933–941.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Long Beach, CA, USA, 2017; pp. 6000–6010.

20. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A Neural Image Caption Generator. In Proceedings of the Name of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
21. Wu, Q.; Shen, C.; Liu, L.; Dick, A.; Hengel, A.V.D. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 203–212.
22. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
23. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10677–10686.
24. Yao, T.; Pan, Y.; Li, Y.; Mei, T. *Exploring Visual Relationship for Image Captioning*; Springer International Publishing: Berlin, Germany, 2018; pp. 711–727.
25. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4904–4912.
26. Corbetta, M.; Shulman, G.L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **2002**, *3*, 201–215. [CrossRef] [PubMed]
27. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250.
28. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6298–6306.
29. Wu, L.; Tian, F.; Zhao, L.; Lai, J.; Liu, T. *Word Attention for Sequence to Sequence Text Understanding*; AAAI: Palo AltoO, CA, USA, 2018.
30. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1179–1195.
31. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
32. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the 13th European Conference on Computer Vision, ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Springer: Zurich, Switzerland, 2014; pp. 740–755.
33. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
34. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
35. Satanjeev, B. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*; ACL: Ann Arbor, MI, USA, 2005; pp. 228–231.
36. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. *SPICE:* Semantic propositional image caption evaluation. In Proceedings of the 21st ACM Conference on Computer and Communications Security, CCS 2014, Scottsdale, AZ, USA, 3–7 November 2014; Springer: Scottsdale, AZ, USA, 2016; pp. 382–398.
37. Lin, C.-Y. *ROUGE: A Package for Automatic Evaluation of Summaries*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
38. Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai, L.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
39. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
40. Jiang, W.; Ma, L.; Jiang, Y.-G.; Liu, W.; Zhang, T. *Recurrent Fusion Network for Image Captioning*; Springer International Publishing: Cham, Sweitzerland; Berlin, Germany, 2018; pp. 510–526.
41. Ji, J.; Xu, C.; Zhang, X.; Wang, B.; Song, X. Spatio-Temporal Memory Attention for Image Captioning. *IEEE Trans. Image Process.* **2020**, *29*, 7615–7628. [CrossRef]
42. Cai, W.; Liu, Q. Image captioning with semantic-enhanced features and extremely hard negative examples. *Neurocomputing* **2020**, *413*, 31–40. [CrossRef]
43. Guo, L.; Liu, J.; Lu, S.; Lu, H. Show, Tell, and Polish: Ruminant Decoding for Image Captioning. *IEEE Trans. Multimed.* **2020**, *22*, 2149–2162. [CrossRef]
44. Kv, G.; Nambiar, A.; Srinivas, K.S.; Mittal, A. Linguistically-aware attention for reducing the semantic gap in vision-language tasks. *Pattern Recognit.* **2021**, *112*, 107812. [CrossRef]

45. Zhang, Z.; Wu, Q.; Wang, Y.; Chen, F. Exploring region relationships implicitly: Image captioning with visual relationship attention. *Image Vis. Comput.* **2021**, *109*, 104146. [CrossRef]
46. Fei, Z. Memory-Augmented Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; pp. 1317–1324.
47. Yan, C.; Hao, Y.; Li, L.; Yin, J.; Liu, A.; Mao, Z.; Chen, Z.; Gao, X. Task-Adaptive Attention for Image Captioning. In *IEEE Transactions on Circuits and Systems for Video Technology (Early Access)*; IEEE: Piscataway, NJ, USA, 2021; p. 1.