symmetry

MDPI

Article

Sequential Dual Attention: Coarse-to-Fine-Grained Hierarchical Generation for Image Captioning

Zhibin Guan [1], Kang Liu [1,*], Yan Ma [1], Xu Qian [1] and Tongkai Ji [1,2]

[1] School of Mechanical Electronic & Information Engineering, China University of Mining & Technology (Beijing), Beijing 100083, China; gzbcumtb@163.com (Z.G.); macumtb@163.com (Y.M.); xuqian@cumtb.edu.cn (X.Q.); jitk@gdeii.com.cn (T.J.)

[2] G-Cloud Technology Corporation, Cloud Computing Center, Chinese Academy of Sciences, Dongguan 523808, China

* Correspondence: Kangliu@cumtb.edu.cn; Tel.: +86-188-0131-1135

Received: 25 October 2018; Accepted: 7 November 2018; Published: 12 November 2018

Abstract: Image caption generation is a fundamental task to build a bridge between image and its description in text, which is drawing increasing interest in artificial intelligence. Images and textual sentences are viewed as two different carriers of information, which are symmetric and unified in the same content of visual scene. The existing image captioning methods rarely consider generating a final description sentence in a coarse-grained to fine-grained way, which is how humans understand the surrounding scenes; and the generated sentence sometimes only describes coarse-grained image content. Therefore, we propose a coarse-to-fine-grained hierarchical generation method for image captioning, named SDA-CFGHG, to address the two problems above. The core of our SDA-CFGHG method is a sequential dual attention that is used to fuse different grained visual information with sequential means. The advantage of our SDA-CFGHG method is that it can achieve image captioning in a coarse-to-fine-grained way and the generated textual sentence can capture details of the raw image to some degree. Moreover, we validate the impressive performance of our method on benchmark datasets—MS COCO, Flickr—with several popular evaluation metrics—CIDEr, SPICE, METEOR, ROUGE-L, and BLEU.

Keywords: image caption generation; sequential dual attention; coarse-to-fine-grained; SDA-CFGHG

1. Introduction

Automatic generation of image captions is drawing increasing interest in artificial intelligence. It is a fundamental task to build a bridge between image and its description in text. The main aim of image caption generation is to make machines generate a textual sentence to accurately depict the content of a given image. Therefore, it is related to two major artificial intelligence fields: computer visual (CV) and natural language processing (NLP). Image caption generation can be applied to different aspects which simultaneously involve visual systems and language systems, such as semantic visual search, visual intelligence in chatting robots, and assisting visual impaired people to perceive the content of surrounding scenes. Generally, it plays a significant role in scene understanding.

The methods used for image captioning can be roughly grouped into three categories: (1) template based methods; (2) retrieval based methods; and (3) artificial neural network (ANN) based methods. The template based methods and retrieval based methods are the early methods that have some unnecessary restrictions. The template based methods, which use a predefined sentence structure to generate final image caption sentence, are more rigid and lack diversity; and the retrieval based methods, which re-use description sentences available from the searched tagged images, cannot generate novel captions. Differently, the ANN based methods try to generate image captions based

on an encoder–decoder framework, which is commonly used in machine translation. The image is first encoded into a fixed-length embedding vector by an encoder; then, this vector is injected into a decoder to generate final textual description that can depict the image content.

Furthermore, the commonly used encoders are based on convolutional neural networks (CNNs) [1], which are effective for object detection and recognition in area of image processing, including VGG [2], ResNet [3], Adaptive Artificial Neural Network (AANN) [4], Spiking Deep Neural Network (SDNN) [5], Probabilistic Neural Network (PNN) [6], and so on [7–9]. For example, CNN and specific image pre-processing steps are combined for facial expression recognition in [7], and CNN is also incorporated as a final classifier to assist the object detection of 2D images [8].

On the other hand, since the decoders used in image captioning are applied to generate textural information, the most classical models used in NLP can be transferred for image captioning, such as recurrent neural networks (RNNs) [10], bi-directional recurrent neural network (Bi-RNN) [11], long short term memory (LSTM) cell [12], gated recurrent unit (GRU) [13], etc. The advantage of ANN based methods is that it can generate novel image captions without a pre-defined rigid sentence template.

Nonetheless, there is still a great distance between machine intelligence and human-being intelligence in image captioning research. For example, humans usually try to get a rough understanding of a given image. After that, the details of the image are perceptive by searching a series of sub-regions of the image. It is how humans understand the surrounding scenes: a coarse-grained to fine-grained way. However, the coarse-grained to fined-grained generation for image captioning is rarely involved in existing ANN-based methods. On the one hand, most existing ANN-based methods focus on a main gist contained in the image and depict it with a similar sentence. On the other hand, the generated sentence sometimes only describes coarse-grained image content.

Therefore, we propose a coarse-to-fine-grained hierarchical generation (SDA-CFGHG) method for image captioning, which can be used to generate a final image description sentence in a coarse-grained to fine-grained way. As opposed to previous ANN-based methods, the SDA-CFGHG method tries to fuse different grained visual information of a given image based on sequential dual attention.

The method we propose has three key parts: (1) image information extraction; (2) sequential dual attention; and (3) language generation model. The image information extraction part focuses on extracting different grained visual information from the raw image, including the global image feature, a set of sub-spatial feature maps, the object features and attribute labels. The sequential dual attention part is applied to fuse different grained visual information above sequentially. The language generation model is used to generate a final fine-grained image description in text.

In the first part of our research, we respectively used the ResNet-152 model and the faster-RCNN model [14] to extract the global image feature, the object features and attribute labels, since the two models used are more accurate and robust than other optional models. It is worth noting that the set of sub-spatial feature maps is also generated by the ResNet-152 model. In the second part, we use "soft" attention as the basic attention mechanism of our proposed sequential dual attention, since it can be embedded into the language model for training directly. In the last part, a stacked two-layer RNN with an LSTM cell is applied to generate final fine-grained image caption sentence, since the LSTM cell can accumulate long-term sequence information to some degree.

Concretely, the contributions of this paper are as follows:

- We propose a new SDA-CFGHG method for image captioning.
- We extracted different grained visual information from the raw image, including the global image feature, a set of sub-spatial feature maps, the object features and attribute labels, all of which are sequentially used to generate final image descriptions in text.
- We propose sequential dual attention to fuse different grained visual information above, which can ensure the generation of an image caption sentence in a coarse-grained to fine-grained way, and the final generated description can describe details of an image.

- Evaluation results show that our approach can achieve impressive performance with several well-known evaluation metrics, including CIDEr [15], SPICE [16], METEOR [17], ROUGE-L [18], and BLEU [19].

Remainder The remainder of this paper is organized as follows: In Section 2, we review the related work of image captioning. Then, the details of our proposed approach are introduced in Section 3. Next, the results and discussion are given in Section 4. Finally, we conclude our work with future research in Section 5.

2. Related Work

In this section, we review the related work of image captioning. They can be broadly divided into three categories: template based methods [20–22], retrieval based methods [23–25], and ANN based methods [26–41].

2.1. Template Based Methods

Template based methods are traditional image captioning methods, which need to predefine a rigid sentence structure. Then, the predicted nouns, verbs, and scenes are applied to fill in the syntactic structure to compose a description sentence.

Due to the previous template based methods trying to interpret what happened in an image by a direct representation, which cannot explore the relationship between images and sentences. Authors in [20] proposed a method to compute a score to link an image to its description sentence. The scoring procedure in [20] was built around an intermediate representation between images and sentences, called "meaning" space. The "meaning" space was represented by triplets: <object, action, scene>. The strength of this approach is that it is symmetric: a best description (resp. image) can be searched from a large set for a given image (resp. description).

Furthermore, with the goal of solving the main linguistic constrains of templates: the grammatically correct sentences cannot be generated by language models along. A new approach, proposed by [21], tried to make a tight connection between the sentence generation process and the particular image content. The statistics learning was used to parse large quantities of text data, and the recognition algorithms from CV were applied to detect objects in an image. The strength of this method is that it can produce more relevant sentences for images than previous attempts.

Furthermore, image caption generation is a complex process, which involves three parts: perceiving the visual space; grounding to world knowledge in the language space; and the generation of textual sentence. Hence, authors in [22] proposed a computationally feasible framework to integrate these components together, of which the semantic grounding was obtained from a large textural corpus. The hidden Markov model (HMM) was used to model the generation process of captions. Although the sentences generated by this method are both readable and relevant for given images, there are some fails of the predicted nouns or verbs, since the detected objects may be mistaken.

2.2. Retrieval Based Methods

Due to the produced descriptions of most retrieval based image captioning methods lacking creativity. A holistic data-driven approach was presented by authors in [23], which exploited both the image data and the language descriptions. Given a raw image, it first retrieved existing human-written phrases from a caption database by measuring the visual similarity. In addition, it then tried to generate a relevant sentence by using Integer Linear Programming (ILP). The main contribution of this method is that it systematically incorporated several CV approaches to retrieve visually relevant phrases, but the captions generated by using ILP may result in some mistakes, i.e., the "bike" was mistaken for a "flower".

In addition, a problem of data-driven matching methods is that the final generated captions could be hampered in some cases, such as the pool alignment between images and human-composed

captions. Therefore, authors in [24] proposed a nonparametric density estimation technique to address the problem mentioned. A word frequency model was used to search a smoothed estimate of visual content across multiple captions, rather than depending on a single noisy estimate of visual similarity. The drawback of this method is that objects may be missed due to some generated caption sentences only describing the background of an image.

Furthermore, in order to make clear what is the relationship between images and language sentences. Two variations on text retrieval were presented by authors in [25], including retrieving the entire existing image description, and retrieving bits of phrases based on visual and geometric similarity of objects and scenes. The main strength of these two methods is that it is the first attempt to search the internet for general captioned images.

Although the template based methods and retrieval based methods can generate relevant descriptions of images, they need a predefined rigid language template or cannot generate novel captions for a given image.

2.3. ANN Based Methods

There are various ANN based methods that have been proposed for image caption generation, which rely on the attention mechanism to decide which part of the visual information is important.

In terms of the problem that the language sentence generated by the encoder–decoder framework being common and only weakly coupled to the raw image, authors in [26] proposed an extension of the LSTM, named gLSTM, which can ensure that the generated words are more tightly coupled to the image content. Concretely, the semantic information of target image or retrieved texts was extracted as extra information at each time step to guide the language model generating the related word. Nonetheless, its generated image caption sometimes may be overly guided. In [27], a discriminability loss was incorporated for training, which could allow an image described as being better identified by both machines and humans. The main contribution of this method is that a retrieval model and a caption generator were combined into a collaborative framework.

Authors in [28] proposed two different attention algorithms, named "soft" and "hard" attention, which were respectively applied to two different image captioning frameworks to better comprehend the visual content of an image. "Soft" attention is a parameterized method, which can be used to generate an encoded feature vector by calculating the weights of all input feature vectors. On the other hand, due to the fact that "hard" attention does not rely on all hidden states of the language model, the gradient in it needs to be estimated by using Monte Carlo-based sampling. Therefore, most existing attention-based methods utilize "soft" attention as the basic mechanism, since it can be embedded into a language model for training directly.

In addition, an SCA-CNN model, is proposed by [29] to solve the problem that attention is only applied in the last conv-layer in existing methods. In SCA-CNN, the channel-wise features were intergrated, which can help the language model gain a better understanding of how CNN features evolved in the process of image caption generation. The limitation of this method is that its performance cannot outperform the ensemble models, due to SCA-CNN being a single model. To address problems related to object missing and object misjudgment, authors in [30] proposed a global-local attention (GLA) model, in which the image-level and object-level features were integrated by an attention mechanism. Furthermore, the attentive linear transformation (ALT) method was proposed in [31] to address the limitation that the extracted each region of image contains relevant and irrelevant information. A constant transformation weight and an attention matrix were learned in ALT to help caption models and explore more useful concepts of raw images. In [32], there is the problem that traditional attention models use the image visual information first, which may lead to the loss increased direction with time progressing. The authors in [32] proposed a parallel-stack LSTM (PS-LSTM) model for image captioning, which can ensemble more parameters on a single model. It achieved comparable performance.

Moreover, authors in [33] proposed a novel text-conditional attention based on the gLSTM model, which can interpret image features based on textual content. "Areas of attention" were proposed by [34] to solve the problem that previous attention-based works associated image regions only to the RNN state. It is a trainable system. In addition, in [35], authors proposed a sequence-to-sequence RNN method to make the input image a sequence of detected objects to generate final corresponding captions. A bottom-up and top-down method was proposed by [36] for image captioning and visual question answering (VQA). It combined "soft" and "hard" attention to achieve image caption generation.

All ANN-based methods existing above attempt to depict the main gist of an image with a similar textual sentence, while the fine-grained visual information may not be reflected, due to generated description sometimes only describing coarse-grained image content. Therefore, we propose an SDA-CFGHG method for image captioning that can generate a final image caption sentence in a coarse-grained to fine-grained way, and the generated final description can describe details of an image.

3. Materials and Methods

In this section, we focus on introducing the contributions of this paper, including the image information extraction, the sequential dual attention, and the language generation model. The main purpose of our research is to make the machine perceive an image in a coarse-grained to fine-grained way, which is how humans understand the surrounding scenes. The framework of our sequential dual attention: coarse-to-fine-grained hierarchical generation (SDA-CFGHG) method is given in Figure 1.

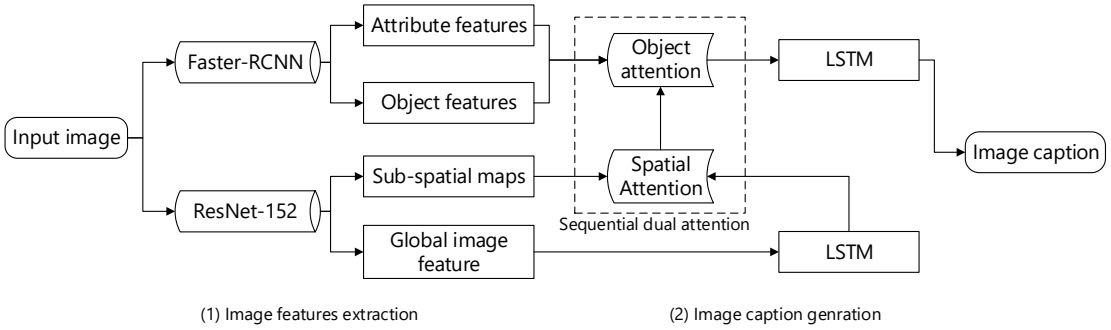


Figure 1. The framework of our proposed method.

The main idea of our method is to integrate different grained visual information in sequential means by using sequential dual attention, which can solve the problem discussed before.

3.1. Image Information Extraction

The visual information of a raw image we used in our research is the global image feature, a set of sub-spatial feature maps, the object features and the attribute features.

Since the global image feature encoded by an encoder is a fixed-length embedding vector, which cannot contain more details of a raw image, we use it as the coarse-grained visual information, which can provide the language generation model, a general perception of the raw image. In our approach, we utilize the ResNet-152 model to extract the global image feature, since the ResNet-152 model can generate accurate semantic features compared with VGG. The ResNet-152 is pre-trained on the large-scale ImageNet classification dataset [42]. In addition, the dimension of the global image feature is equal to 2048, since the size of the last full connection layer of ResNet-152 model is set to 2048 (see Table 1). Therefore, the global image feature of each image is a 2048-dimension vector, denoted as $\mathscr{G}$ (see Equation (1)).

$$\mathscr{G} \in \mathbb{R}^{2048}. \tag{1}$$

Furthermore, it is worth noting that the set of sub-spatial feature is also generated by the ResNet-152 model. Due to the last convolutional map of the ResNet-152 model being directly mapped into a 2048-dimension vector through the last convolutional layer in the previous global feature extraction process, some information from the raw image may be missed.

Table 1. Coarse-grained visual information generated by ResNet-152.

Categories of Information	Optional Values
global feature	fixed-length feature vector: 2048×1
sub-spatial map	size of each random selected map: 14×14

To address this problem, we try to extract more detailed visual information from the last convolutional map of the ResNet-152 model. Concretely, the last convolutional map generated by ResNet-152 model is multi-channel. Thus, we can randomly extract a series of regional feature maps from the last multi-channel feature map through average pool, and generate a set of sub-spatial maps. The size of average pool is set to 14×14 in our research. Finally, since every regional sub-spatial feature map generated is multi-channel as well, we utilize a convolutional layer to map each sub-spatial feature map into a fixed-length 2048-dimension vector for easy calculation. The extracted set of sub-spatial features is denoted as $\mathscr{S}$ (see Equation (2)):

$$\mathscr{S} = \{s_1, s_2, ..., s_C\}, s_j \in \mathbb{R}^{2048}, j \in (1, C). \tag{2}$$

Moreover, we use a faster-RCNN model to obtain the fine-grained visual information of a raw image, including the object features and the attribute labels. A faster-RCNN model is an end-to-end framework designed to localize instances of objects with bounding boxes and identify what certain classes they belong to. Thus, it can be used to simultaneously generate a series of bounding boxes and the corresponding class labels.

Concretely, the size of each image is equal to 224×224 after the resizing and cropping preprocessing operations. Then, each image is injected into the faster-RCNN model to generate a series of bounding boxes and class labels. After that, the feature vectors of the generated bounding boxes of a raw image are used as our object features, since each bounding box contains details of a specific target. On the other hand, the generated class labels are seen as attributes (see Table 2). It is worth noting that the generated textual attribute labels are not composed for the content directly. We utilize only the feature vector of each attribute label rather than the textual attribute labels. For each attribute label, it will be mapped into a fixed-length vector through an embedding layer, just like the words of reference sentence. The size of the embedding layer is set to 1024.

Table 2. Fine-grained visual information generated by faster-RCNN.

Categories of Information	Optional Values
Object features	vector from bounding-box region: (P_x, P_y, P_w, P_h)
attribute labels	man, woman, child, chair, train, dog, bag, cat...

The obtained object features and the attribute features are respectively denoted as set $\mathscr{O}$ (see Equation (3)) and set $\mathscr{A}$ (see Equation (4)):

$$\mathscr{O} = \{o_1, o_2, ..., o_Q\}, o_j \in \mathbb{R}^{2048}, j \in (1, Q), \tag{3}$$

$$\mathscr{A} = \{a_1, a_2, ..., a_Z\}, a_j \in \mathbb{R}^{2048}, j \in (1, Z). \tag{4}$$

Finally, the object features are combined with the attribute features to generate final fine-grained information set, denoted as $\mathscr{V}$ (see Equation (5)):

$$\mathscr{V} = \{v_1, v_2, ..., v_{Q+Z}\}, v_j \in \mathbb{R}^{2048}, j \in (1, Q+Z). \tag{5}$$

3.2. Sequential Dual Attention

The core of our proposed SDA-CFGHG method is sequential dual attention, which is applied to integrate different grained visual information of an image sequentially. The sequential dual attention consists of two parts: spatial attention and object attention (see Figure 2).

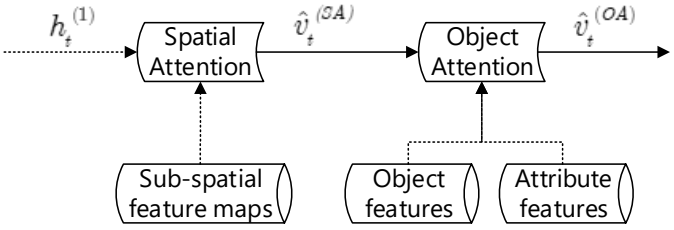


Figure 2. Sequential Dual Attention; the dotted line indicates the input information at each time step t.

The input information of the sequential dual attention including a hidden state $h_t^{(1)}$, a set of sub-spatial features maps, the object features and attribute features. The hidden state $h_t^{(1)}$ comes from our language generation model, which can be seen in Section 3.3. They are injected into our dual attention by using sequential means. Here, we try to use sequential means to integrate the above information, since it can allow our language model to understand a given image in a coarse grained to fine grained way, just like humans.

Furthermore, the "soft" attention is the basic attention mechanism of our sequential attention because it can be embedded into our language generation model for training directly. Concretely, for given feature vectors ξ_j and ξ_{base}, the similarity between these two vectors can be measured by using the cosine function, as shown in Equation (6):

$$\mathscr{F}_t(\xi_j, \xi_{base}) = \cos\left(\xi_j, \xi_{base}\right) = \frac{\xi_j \cdot \xi_{base}}{\|\xi_j\| \|\xi_{base}\|}, j \in [1, N]. \tag{6}$$

In addition, the attention weight of each input feature ξ_j at time step t is denoted as α_t, which can be calculated as Equation (6):

$$\alpha_t(\xi_j | \xi_{base}) = \frac{\mathscr{F}_t(\xi_j, \xi_{base})}{\sum_j \mathscr{F}_t(\xi_j, \xi_{base})}, j \in [1, N], \tag{7}$$

where the attention weights satisfies the constraint $\sum_{j=1}^{N} \alpha_t(\xi_j | \xi_{base}) = 1$.

Therefore, in spatial attention, the similarity and attention weight between each sub-spatial map s_j and the hidden state $h_t^{(1)}$ at time step t can be calculated through Equations (6) and (7), respectively. Furthermore, the computational details of the fused feature vector $\hat{v}_t^{(SA)}$ at time step t can be referred to as Equation (8).

$$\hat{v}_t^{(SA)} = \sum_{j=1}^{C} \alpha_t(s_j | h_t^{(1)}) s_j, s_j \in \mathscr{S}. \tag{8}$$

On the other hand, in object attention, the attention weight of each feature vector v_j at time step t can be calculated according to the previously generated fused feature vector $\hat{v}_t^{(SA)}$, denoted as $\alpha_t(v_j|\hat{v}_t^{(SA)})$. In addition, the fused vector $\hat{v}_t^{(OA)}$ in object attention is calculated as Equation (9):

$$\hat{v}_t^{(OA)} = \sum_{j=1}^{Q+Z} \alpha_t(v_j|\hat{v}_t^{(SA)}) v_j, v_j \in \mathcal{V}. \tag{9}$$

Finally, the generated $\hat{v}_t^{(OA)}$ is used in our language generation model for image captioning. From above sequential means, our language generation model can understand a given image sequentially, which avoids the loss of partial fine-grained visual information.

3.3. Language Generation Model

In this subsection, we mainly focus on introducing the details of our language generation model. It is worth noting that the previous sequential dual attention is embedded into our language model for training directly.

A stacked two-layer RNN with LSTM cell is used as our language generation model to achieve image captioning in a coarse-grained to fine-grained way (see Figure 3). The reason why we chose LSTM cell as the basic node of our language model is that it can accumulate long-term sequence information to some degree.

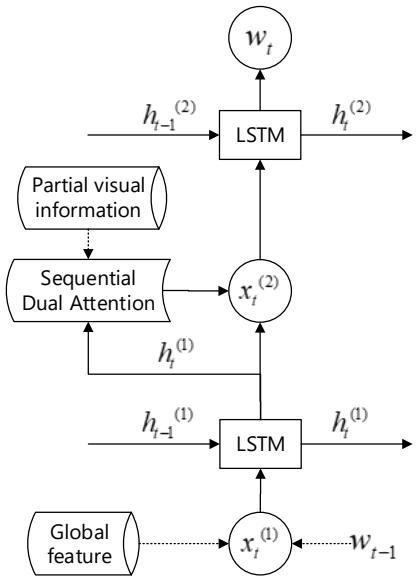


Figure 3. Language Generation Model. The dotted line indicates the input information at each time step t.

As shown in Figure 4, each LSTM cell consists of four significant gates: input gate, forget gate, output gate, and memory gate; and they are respectively denoted as i_t, f_t, o_t, and g_t at time step t. The detailed calculations of these gates are given in Equation (10):

$$\begin{aligned} i_t &= \sigma\left(W_{ix}x_t + W_{ih}h_{t-1} + b_i\right), \\ f_t &= \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right), \\ o_t &= \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + b_o\right), \\ g_t &= \tanh\left(W_{gx}x_t + W_{gh}h_{t-1} + b_g\right), \end{aligned} \tag{10}$$

where x_t indicates the input information of the LSTM cell at time step t, and h_{t-1} is the generated hidden state of the LSTM cell at time step $t-1$. W_* and b_* are the shared parameter, respectively, which should be learned in all time steps. In addition, σ is the sigmoid activation function.

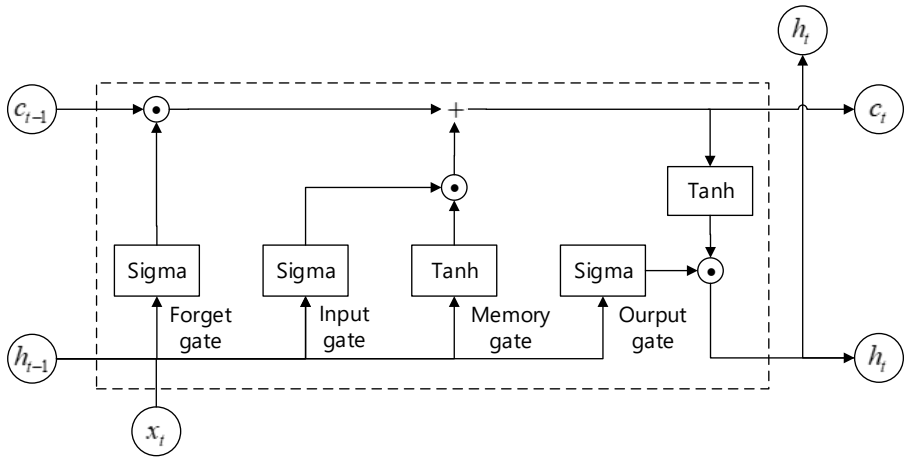


Figure 4. The architecture of long short term memotry (LSTM) cell. "Sigma" and "tanh" represent the sigmoid function and the tanh function, respectively.

Then, the cumulative information of the LSTM cell, denoted as c_t, can be calculated according to the previous generated c_{t-1} and the output of these four gates. The detailed operation is shown as Equation (11), where $\odot$ denotes element-wise multiplication:

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t. \tag{11}$$

Finally, the hidden state of the LSTM cell at time step t can be calculated through Equation (12):

$$h_t = o_t \odot \tanh(c_t). \tag{12}$$

For easy expression, we use Equation (13) to represent the above operations of Equations (10)–(12),

$$h_t = LSTM(x_t, h_{t-1}). \tag{13}$$

Furthermore, in the first layer of our language generation model, the input information $x_t^{(1)}$ can be calculated through the global image feature $\mathscr{G}$ and the previously generated word w_{t-1}, and the detailed calculation is referred to as Equation (14)

$$x_t^{(1)} = W_\varphi \mathscr{G} + W_w w_{t-1}, \tag{14}$$

where w_{t-1} belongs to the generated textual sentences S, which can be represented as a set of words (see Equation (15)):

$$S = \{w_0, w_1, ..., w_L\}. \tag{15}$$

Therefore, the generated hidden state of the first layer of our language model can be calculated as the following Equation (16):

$$h_t^{(1)} = LSTM(x_t^{(1)}, h_{t-1}^{(1)}). \tag{16}$$

After that, since the input information of the second layer of our language model consists of the hidden state $h_t^{(1)}$ and the fused feature $\hat{v}_t^{(OA)}$, it can be obtained through Equation (17):

$$x_t^{(2)} = W_\phi \hat{v}_t^{(OA)} + W_h h_t^{(1)}. \tag{17}$$

In addition, the output of the second layer is a fused vector, which can be calculated by using Equation (18):

$$h_t^{(2)} = LSTM(x_t^{(2)}, h_{t-1}^{(2)}). \tag{18}$$

Moreover, at time step t, the word w_t is generated based on previously generated words $w_0, ..., w_{t-1}$ and the sequential visual information, which have been integrated sequentially into the fused vector $h_t^{(2)}$. Thus, the generated probability of word w_t can be calculated according to the hidden state $h_t^{(2)}$, as shown in Equation (19):

$$p\left(w_t | w_0, ..., w_{t-1}, \mathscr{G}, \mathscr{O}, \mathscr{A}\right) = softmax\left(W_p h_t^{(2)} + b_p\right). \tag{19}$$

Then, since the probability of final image description is a product of probability of each generated word w_i, the detailed calculation can be represented as Equation (20):

$$p\left(w_0, w_1, ..., w_L\right) = \prod_{t=1}^{L} p\left(w_t | w_0, ..., w_{t-1}, \mathscr{G}, \mathscr{O}, \mathscr{A}\right). \tag{20}$$

The objective function used in this research is the negative cross entropy loss, which is referred to as the following Equation (21):

$$L\left(\theta\right) = -\sum_{t=1}^{L} \log p\left(w_t | w_0, ..., w_{t-1}, \mathscr{G}, \mathscr{O}, \mathscr{A}\right), \tag{21}$$

where θ represents the parameters of our language generation model, including W_* and b_*.

Finally, since the CIDEr score is a commonly used evaluation metric to measure the similarity between generated image caption and the human-written textual sentence, we adopted the Self-Critical Sequence Training (SCST) [43] method to optimize the CIDEr score, which can make final image description closer to human expression. The concrete negative expected reward score is shown as Equation (22):

$$R\left(\theta\right) = -\mathbb{E}_{S \sim p_\theta}\left[r(S)\right], \tag{22}$$

where r represents the CIDEr score function.

It is worth noting that we first used the negative cross entropy (see Equation (21)) to pre-train our language model, and then the SCST method (see Equation (22)) is applied to achieve CIDEr optimization based on the pre-trained language model.

4. Results and Discussion

4.1. Datasets

The datasets we used to evaluate the performance of our methods are the MS COCO [44], Flickr8K [45], and Flickr30K [46]. They are the most popular datasets for evaluating the generated descriptions.

Table 3 shows the detailed comparisons of reference captions on the three datasets above. The MS COCO dataset is a large-scale object detection, segmentation, and captioning dataset. The official version of MS COCO dataset includes 82,783 training images, 40,504 validation images, and 40,775 test images. Since the "Karpathy" split is the commonly used split method for reporting results, as in previous works, we use it to split the official MS COCO dataset to obtain 113,287 training images, 5000 validation images, and 5000 test images. In addition, the Flickr8K dataset is officially split into 6000 images for training, 1000 images for validation images, and 1000 images for testing. Furthermore,

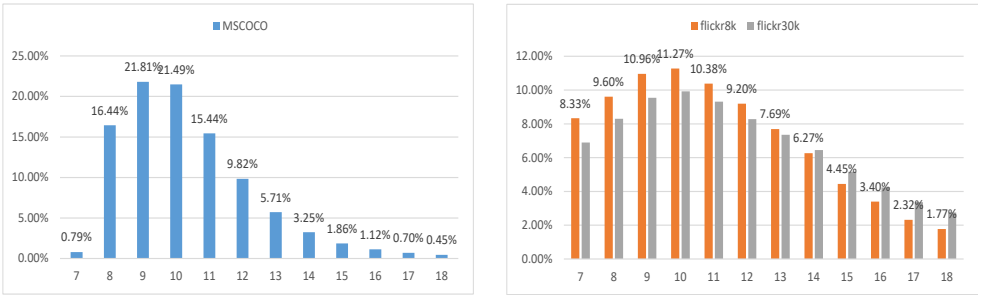
the Flickr30K dataset consists of 31,783 images without an official split, and we split it into 28,000 training images, 1000 validation images, and 1000 images for testing, as in previous works.

Table 3. Comparisons of reference captions on datasets MS COCO, Flickr8K, and Flickr30K.

Datasets	Vocab Size	Max Length	Total Words	Top-10 Words with Higher Occurrences
MS COCO	9486	49	6,421,733	a, on, of, the, in, with, and, is, man, to
Flickr8K	2629	37	422,800	a, in, the, on, is, and, dog, with, man, of
Flickr30K	7648	78	1,892,755	a, in, the, on, and, man, is, of, with, woman

Furthermore, Figure 5 displays the percentage of words in the length of each caption sentence. Since the lengths of the caption sentence are mostly between 7 and 18, only the statistical results in this range are displayed.

For the above datasets, there are five corresponding textual sentences written by humans for each image, which are used as reference captions for training.



(**a**) statistical results of MS COCO captions. (**b**) statistical results of Flickr captions.

Figure 5. Percentage of words in the length of each caption sentence. (**a**) displays the statistical results of reference captions on MS COCO dataset; (**b**) displays the statistical results of reference captions on Flickr8K and Flickr30K datasets. The *x*-axis represents the length of each caption sentence.

4.2. Evaluation Metrics

We evaluate the performance of our method with several well-known metrics that are commonly used for image captioning, including CIDEr, CIDEr [15], SPICE [16], METEOR [17], ROUGE-L [18], and BLEU[19].

CIDEr and SPICE are all human consensus metrics. CIDEr can be used to measure the similarity between generated captions and a set of human-written descriptions. In addition, SPICE is a principled metric, which is applied to evaluate how effectively the generated image sentence recover objects, attributes and the relations between them. METEOR calculates the sentence-level similarity scores based on the harmonic mean of uni-gram recall and precision. ROUGE-L can be used for gisting evaluation. Its score is calculated by measuring the number of overlapping units, such as n-gram, word sequences, and word pairs between generated captions and the human-writing sentences. BLEU is a commonly used metric in machine translation tasks. It is only based on the co-occurrences of n-grams precision.

4.3. Experiment Setting and Results

Just like the traditional method of model parameter acquisition, the experiment of our research was separated into two parts: the training part and the test part. Since our sequential dual attention was based on "soft" attention, it could be embedded into our language model for training directly. In the training part, the ResNet-152 model and the faster-RCNN model were respectively trained to extract the image visual information, including the global image feature, a set of sub-spatial maps, the object features and attribute features. After that, the language model was trained based on the previous generated visual information and the human-written reference sentences. In the test part,

the ResNet-152 model, the faster-RCNN, and our language generation model are used by sequential means to generate final fine-grained description sentences.

In our language model, the cost function is optimized by the RMSProp algorithm, which is the popular method for finding local convergence points. The learning rate is set to 0.0001. In addition, the value of gradient clipping is equal to 0.1, which can be used to avoid gradient explosion. Furthermore, the size of embedding layer in our research is equal to 1024, which is used to reduce the dimension of the input word. In addition, the number of hidden cells (LSTM) in our language model is set to 1024. Furthermore, in our sequential dual attention, the number of attention hidden cells (LSTM) is set to 512. In addition, the Nividia TITAN X (PASCAL) (Santa Clara, CA, USA) is the main graphic processing unit (GPU) we used to accelerate the training process in our search.

Finally, in order to ensure the quality of the final generated textual sentence, we used the beam search method as our description generation strategy in this work. The advantage of the beam search method is that it can automatically select top-K best sentences with higher probabilities at each time step. It is worth noting that the new top-K best sentences are selected based on the old top-K sentences at the previous moment of each time step. In our research, the value of K was set to 3.

4.3.1. Results of Evaluation on Benchmark Datasets

In this part, we display the evaluation results of our proposed SDA-CFGHG method for image captioning on the benchmark datasets, including MS COCO, Flickr8K, and Flickr30K datasets. Since the results of the previous methods for image caption generation have been published publicly in the literature, we can compare our method with these state-of-the-art works directly. The methods used for comparison in this research consist of gLSTM [26], soft attention [28], hard attention [28], Log Bilinear [47], ATT [48], F-G Attention [49], GLA [30], and Topdown [36]. In addition, the evaluation scores were generated by the commonly used coco-caption code.

Table 4 shows the detailed results of our method on MS COCO dataset, in which the "SDA-CFGHG" indicates the evaluation scores of our method. From the comparison results of our SDA-CFGHG method with previous approaches, we can observe that our method has state-of-the-art performance when using the global image feature, a set of sub-spatial maps, the object features and attribute labels by sequential means. Our method can achieve the impressive performance in that the above features can provide different grained visual information of a raw image to our language model and make it understand a raw image in a coarse-grained to fine-grained way.

Table 4. Evaluation results of our method on the MS COCO dataset.

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
gLSTM [26]	35.8	26.4	22.7	—	81.25	—
soft attention [28]	34.4	24.3	23.9	—	—	—
hard attention [28]	35.7	25.0	23.0	—	—	—
Log Bilinear [47]	34.4	24.3	20.0	—	—	—
ATT [48]	40.2	30.4	24.3	—	—	—
F-G Attention [49]	36.2	25.9	24.5	—	—	—
GLA [30]	41.7	31.2	24.9	53.3	96.4	—
Topdown [36]	—	33.4	26.1	54.4	105.4	19.2
SDA-CFGHG	47.7	36.6	28.0	57.2	115.8	21.0

Concretely, although the global image feature is a fixed-length embedding vector, it is enough to provide a coarse-grained understanding for our language generation model. Furthermore, the sub-spatial maps help the language model to decide which parts of the raw image are significant during image captioning. Moreover, while generating each word of final description sentences, the object features and the attribute labels can help our language model perceive the fine-grained information of a raw image.

Furthermore, Tables 5 and 6 display the evaluation scores of our method on Flickr8K and Flickr30K dataset, respectively. Compared to other method, our approach differs in that our language model

uses the proposed sequential dual attention to fuse the extracted image features, sequentially. On the one hand, the spatial attention of the sequential dual attention is first used to fuse the set of sub-spatial maps according to the hidden state $h_t^{(1)}$ that contains partial global feature $\mathscr{G}$, and then to generate a intermediate feature vector $\hat{v}_t^{(SP)}$. On the other hand, the next object attention is applied to integrate the object features and attribute features with the previously generated $\hat{v}_t^{(SP)}$ and then to generate a fused feature vector $\hat{v}_t^{(OA)}$. Therefore, our language generation model can generate final textual sentences in a coarse-grained to fine-grained way, which is how humans perceive the surrounding visual world.

Table 5. Evaluation results of our method on the Flickr8K dataset.

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
gLSTM [26]	31.8	21.6	20.2	—	—
soft attention [28]	29.9	19.5	18.9	—	—
hard attention [28]	31.4	21.3	20.3	—	—
Log-Bilinear [47]	27.7	17.7	17.3	—	
F-G Attention [49]	33.7	23.8	22.6	—	—
GLA [30]	23.9	14.8	16.9	36.2	41.9
SDA-CFGHG	34.2	24.3	23.1	38.6	46.0

Table 6. Evaluation results of our method on the Flickr30K dataset.

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
gLSTM [26]	30.5	20.6	18.0	—	—
soft attention [28]	28.8	19.9	18.5	—	—
hard attention [28]	29.9	19.9	18.5	—	—
Log-Bilinear [47]	25.4	17.1	16.9	—	—
F-G Attention [49]	31.3	21.4	20.0	—	—
GLA [30]	23.2	14.6	16.6	36.2	41.9
SDA-CFGHG	33.4	22.1	20.5	38.4	45.9

4.3.2. Results of Quantitative Analysis

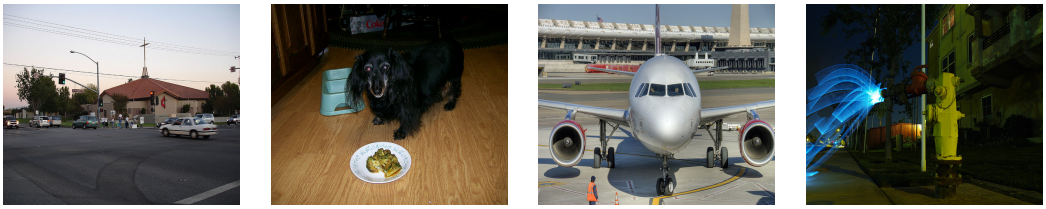
Besides the evaluation scores of our method, we also display some concrete description caption generated by our SDA-CFGHG approach.

Figure 6 displays the sampled MS COCO images and the corresponding generated textual sentences. For each image, there are three different descriptions that are generated by using different visual information. Concretely, the "only G" indicates that only the global image feature is used while generating this sentence. "G+SP" means that the global image feature and the sub-spatial maps are applied to generated image caption, and the object attention is not activated in this situation. Finally, "G+SP+OA" represents that our proposed sequential dual attention is applied for image captioning.

On the other hand, Figure 7 shows the sampled Flickr images and their corresponding image descriptions. There are some failed image textual sentences generated in our experiment. Figure 7d shows the failed image caption. We can observe that whether using the sequential dual attention or not, the language model can not correctly identify the object obtained in the images. For example, the "woman" in Figure 7d is indirectly detected as "man". The reason is that our method may be confused about such content, that is, gender information cannot be correctly perceived. In future work, we will try to improve the ability of our image captioning model to identify certain significant attribute information, such as age, gender, shape, texture, etc.

It is worth noting that the image description sentence is generated based on the extracted different categories of visual feature vectors, including the global image feature, a series of sub-spatial features, the object features and attribute features. The concrete classes in final textual sentences, such as animals, cars, buildings, etc., are generated directly through our language generation model.

Finally, we can conclude that our approach can make the machine finally generate image captions in a coarse-grained to fine-grained way, which is how humans understand the surrounding visual scene. In addition, the generated sentence may describe details of a given image.



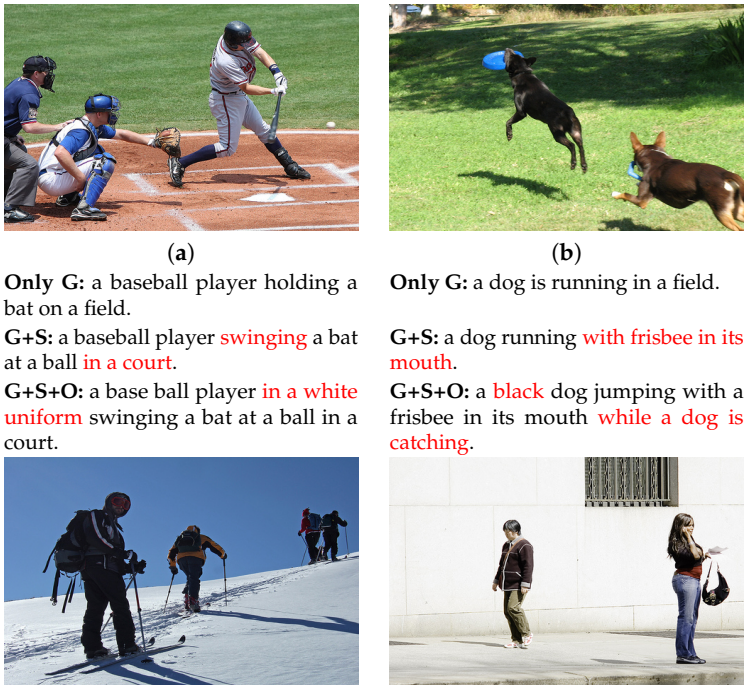
Only G: a car is driving down a street.
G+SA: a car is driving down a street with a traffic light.
G+SA+OA: a car is driving down a street with a group of people in the background.
(**a**)

Only G: a black dog sitting on the floor.
G+SA: a black dog sitting on the floor with a plate of food.
G+SA+OA: a black dog sitting on a wooden floor next to a plate of food.
(**b**)

Only G: a passenger jet sitting on the ground.
G+SA: a large passenger jet sitting on top of an airport tarmac.
G+SA+OA: a large passenger jet sitting on top of an airport tarmac next to a man.
(**c**)

Only G: a fire hydrant sitting on a sidewalk.
G+SA: a fire hydrant sitting in the middle of a sidewalk.
G+SA+OA: a yellow fire hydrant sitting on the side of a road.
(**d**)

Figure 6. Results of image caption generation on the MS COCO dataset. The red color represents the positive example. "G": global image feature; "SA": spatial attention; "OA": object attention. (**a**) Street with cars; (**b**) Black dog; (**c**) Passenger jet; (**d**) Fire hydrant.



(**a**)
Only G: a baseball player holding a bat on a field.
G+S: a baseball player swinging a bat at a ball in a court.
G+S+O: a base ball player in a white uniform swinging a bat at a ball in a court.

(**b**)
Only G: a dog is running in a field.
G+S: a dog running with frisbee in its mouth.
G+S+O: a black dog jumping with a frisbee in its mouth while a dog is catching.

(**c**)
Only G: a group of people going up a mountain.
G+S: a group of people climbing on a snowy slope.
G+S+O: a group of skiers climbing on a snowy slope.

(**d**)
Only G: a man and a woman are standing on a sidewalk.
G+S: a man and woman are walking down the street.
G+S+O: a man and woman are standing on a sidewalk.

Figure 7. Results of image captioning on Flickr30K and Flickr8K. (**a**,**b**) display the results of some Flickr30K images; (**c**,**d**) display the results of some Flickr8K images. The red color represents the positive example, and blue color represents the negative example. "G": global image feature; "SA": spatial attention; "OA": object attention. (**a**) Baseball players; (**b**) Dogs; (**c**) Skiers; (**d**) Pedestrians.

5. Conclusions

In this research, we tried to address a problem in existing ANN-based methods, i.e., that the coarse-grained to fined-grained generation for image caption generation is rarely involved, which is the way humans understand the surrounding scenes.

We propose an SDA-CFGHG method for image caption generation, which can generate the final image caption sentence in a coarse-grained to fine-grained way. The core of our method is sequential dual attention that is used to fuse different grained visual information sequentially. The sequential dual attention consists of spatial-attention and object-attention. In addition, the different grained visual information includes the global image feature, a set of sub-spatial feature maps, the object features and attribute features.

We evaluate the performance of our SDA-CFGHG method with several popular metrics—CIDEr, SPICE, METEOR, ROUGE-L, and BLEU. The experiment results indicate that the method we proposed can understand an image in a coarse-to-fine-grained manner by simulating the way humans perceive visual scenes. In addition, the generated image captions can avoid losing the fine-grained content of the raw image to a certain extent.

However, the object features and attribute features are integrated by an attention mechanism in our research, which may result in the role of the attributes being weakened. Therefore, we will try to separate the middle-level attributes from the raw image and use them to retouch the final image caption in future work.

Author Contributions: Funding Acquisition, T.J.; Investigation, Z.G., K.L.; Methodology, Z.G. and K.L.; Validation, K.L. and Y.M.; Writing—Original Draft, Z.G.; Writing—Review and Editing, X.Q. and T.J.

Funding: This work was partially funded by the National Key R&G Program of China under Grant (2018YFB1004600), partially funded by the National Key Research and Development Program of China under Grant (2016YFB10005000).

Acknowledgments: The authors are grateful for the constructive advice on the revision of the manuscript from the anonymous reviewers.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. CNN: Single-label to Multi-label. *arXiv* **2014**, arXiv:1406.5726.
2. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arxiv:1409.1556.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
4. Woźniak, M.; Połap, D. Adaptive neuro-heuristic hybrid model for fruit peel defects detection. *Neural Netw.* **2018**, *98*, 16–33. [CrossRef] [PubMed]
5. Kheradpisheh, S.R.; Ganjtabesh, M.; Thorpe, S.J.; Masquelier, T. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Netw.* **2018**, *99*, 56–67. [CrossRef] [PubMed]
6. Woźniak, M.; Połap, D.; Capizzi, G.; Sciuto, G.L.; Kośmider, L.; Frankiewicz, K. Small lung nodules detection based on local variance analysis and probabilistic neural network. *Comput. Methods Programs Biomed.* **2018**, *161*, 173–180. [CrossRef] [PubMed]
7. Lopes, A.T.; de Aguiar, E.; Souza, A.F.D.; Oliveira-Santos, T. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. *Pattern Recognit.* **2017**, *61*, 610–628. [CrossRef]
8. Woźniak, M.; Połap, D. Object detection and recognition via clustered features. *Neurocomputing* **2018**, *320*, 76–84. [CrossRef]
9. Połap, D.; Winnicka, A.; Serwata, K.; Kęsik, K.; Woźniak, M. An Intelligent System for Monitoring Skin Diseases. *Sensors* **2018**, *18*, 2552. [CrossRef] [PubMed]

10. Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
11. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [CrossRef]
12. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. In *Neural Computation*; MIT Press: Cambridge, MA, USA, 1997; pp. 1735–1780.
13. Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
14. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal. *arXiv* **2015**, arXiv:1506.01497.
15. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. *arXiv* **2014**, arXiv:1411.5726.
16. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*; Springer: Cham, Switzerland, 2016; pp. 382–398.
17. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
18. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 Workshop Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
19. Papineni, K.; Roukos, S.; Ward, T. 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.
20. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every Picture Tells a Story: Generating Sentences from Images. In *ECCV 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 15–29.
21. 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 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1601–1608.
22. Yang, Y.; Teo, C.L.; Daumé, H., III; Aloimonos, Y. Corpus-Guided Sentence Generation of Natural Images. In Proceedings of the EMNLP '11 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 27–31 July 2011; pp. 444–454.
23. Kuznetsova, P.; Ordonez, V.; Berg, A.C.; Berg, T.L.; Choi, Y. Collective Generation of Natural Image Descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 8–14 July 2012; pp. 359–368.
24. Mason, R.; Charniak, E. Nonparametric Method for Data-driven Image Captioning. In Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 592–598.
25. Ordonez, V.; Han, X.; Kuznetsova, P.; Kulkarni, G.; Mitchell, M.; Yamaguchi, K.; Stratos, K.; Goyal, A.; Dodge, J.; Mensch, A.; et al. Large Scale Retrieval and Generation of Image Descriptions. *Int. J. Comput. Vis. (IJCV)* **2016**, *119*, 46–59. [CrossRef]
26. Jia, X.; Gavves, E.; Fernando, B.; Tuytelaars, T. Guiding Long-Short Term Memory for Image Caption Generation. *arXiv* **2015**, arXiv:1509.04942.
27. Luo, R.; Price, B.L.; Cohen, S.; Shakhnarovich, G. Discriminability objective for training descriptive captions. *arXiv* **2018**, arXiv:1803.04376.
28. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
29. Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; Li, L. Deep Reinforcement Learning-based Image Captioning with Embedding Reward. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1151–1159.
30. Li, L.; Tang, S.; Zhang, Y.; Deng, L.; Tian, Q. GLA: Global-Local Attention for Image Description. *IEEE Trans. Multimed.* **2018**, *20*, 726–737. [CrossRef]

31. Ye, S.; Han, J.; Liu, N. Attentive Linear Transformation for Image Captioning. *IEEE Trans. Image Process.* **2018**, *27*, 5514–5524. [CrossRef] [PubMed]
32. Zhu, X.; Li, L.; Liu, J.; Li, Z.; Peng, H.; Niu, X. Image captioning with triple-attention and stack parallel LSTM. *Neurocomputing* **2018**, *319*, 55-–65. [CrossRef]
33. Zhou, L.; Xu, C.; Koch, P.A.; Corso, J.J. Watch What You Just Said: Image Caption Generation with Text-Conditional Semantic Attention. *arXiv* **2016**, arXiv:1606.04621.
34. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of Attention for Image Captioning. *arXiv* **2016**, arXiv:1612.01033.
35. Liu, C.; Sun, F.; Wang, C.; Wang, F.; Yuille, A.L. MAT: A Multimodal Attentive Translator for Image Captioning. *arXiv* **2017**, arXiv:1702.05658.
36. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *arXiv* **2017**, arXiv:1707.07998.
37. Wu, C.; Wei, Y.; Chu, X.; Su, F.; Wang, L. Modeling visual and word-conditional semantic attention for image captioning. *Signal Process. Image Commun.* **2018**, *67*, 100-–107. [CrossRef]
38. 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.
39. Wu, Q.; Shen, C.; Liu, L.; Dick, A.; Hengel, A. 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.
40. Wang, L.; Chu, X.; Zhang, W.; Wei, Y.; Sun, W.; Wu, C. Social Image Captioning: Exploring Visual Attention and User Attention. *Sensors* **2018**, *18*, 646. [CrossRef] [PubMed]
41. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the IEEE International Conference on Computer Vision ICCV, Venice, Italy, 22–29 October 2017; pp. 4904–4912.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Curran Associates, Inc.: Lake Tahoe, NV, USA, 2012; pp. 1097–1105.
43. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical Sequence Training for Image Captioning. *arXiv* **2016**, arXiv:1612.00563.
44. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
45. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics. *J. Abbr.* **2013**, *47*, 853–899. [CrossRef]
46. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78.
47. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.
48. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image Captioning with Semantic Attention. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4651–4659.
49. Chang, Y.S. Fine-grained attention for image caption generation. *Multimed. Tools Appl.* **2018**, *77*, 2959–2971. [CrossRef]