electronics

MDPI

Article

A Modularized Architecture of Multi-Branch Convolutional Neural Network for Image Captioning

Shan He and Yuanyao Lu *

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; heshan0123@sina.cn

* Correspondence: luyy@ncut.edu.cn

Received: 31 October 2019; Accepted: 25 November 2019; Published: 28 November 2019

Abstract: Image captioning is a comprehensive task in computer vision (CV) and natural language processing (NLP). It can complete conversion from image to text, that is, the algorithm automatically generates corresponding descriptive text according to the input image. In this paper, we present an end-to-end model that takes deep convolutional neural network (CNN) as the encoder and recurrent neural network (RNN) as the decoder. In order to get better image captioning extraction, we propose a highly modularized multi-branch CNN, which could increase accuracy while maintaining the number of hyper-parameters unchanged. This strategy provides a simply designed network consists of parallel sub-modules of the same structure. While traditional CNN goes deeper and wider to increase accuracy, our proposed method is more effective with a simple design, which is easier to optimize for practical application. Experiments are conducted on Flickr8k, Flickr30k and MSCOCO entities. Results demonstrate that our method achieves state of the art performances in terms of caption quality.

Keywords: image captioning; convolutional neural network (CNN); multi-branch expansion; long short-term memory (LSTM)

1. Introduction

As an important source of information, numerous images are digitally stored and transmitted on the Internet. Since most of the human communication relies on natural language, enabling computers to describe the visual world can bring a wide range of possible applications, such as natural human-computer interaction, child education, information retrieval, and assistance for visually impaired people.

Image captioning is a comprehensive task in computer vision (CV) [1] and natural language processing (NLP) [2], which can complete multi-modal conversion from image to text. The algorithm automatically generates corresponding descriptive texts according to an input image. The task can be specifically described as: I represents an input image of a given group (I, S). $S = \{S_1, S_2, \cdots\}$ represents the target word sequence where S_t is the word extracted from the dataset. The training goal of the model is to maximize the likelihood $p(S|I)$ to complete a multimodal mapping from image I to describing sentence S.

As a challenging and meaningful area of artificial intelligence, automatically generating image descriptions are attracting more and more attention. The goal of image captioning is to generate linguistically plausible sentences that are semantically true to the content of the image. Therefore, image description has two basic aspects: visual understanding and language processing [3]. To ensure that the generated sentences are grammatically and semantically correct, CV and NLP techniques should be used to deal with the problems generated by the corresponding modalities and to properly integrate them.

With the booming growth of data scale and computing capability, machine learning based on data and hardware begins to show unique advantages, which directly promote the prosperity of artificial intelligence in various application areas. Under the pioneering research of many outstanding scholars, deep learning with the kernel structure of neural network emerged. It has made extraordinary progress in CV, NLP and speech recognition [4]. In 2012, many participants of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [5] adopted a Convolutional Neural Network (CNN) [6] in their successful and innovative programs, which attracted widespread attention in academia and industry. Deep learning began to become popular and quickly introduced into many other classic CV tasks, including image classification, object recognition and object detection [5]. At the same time, excellent deep learning models in NLP such as Word2Vec [7] and GloVe [8] have been proposed. The performance of other NLP tasks has also been generally improved, such as sentiment analysis, text summarization and machine translation.

Since Alex and his mentor took the first place with the 16.4% top-5 error rate in ILSVRC2012 [9], the dominance of CNN in CV was established. VGGNet [10], developed by the Visual Geometry Group of University of Oxford and Google DeepMind, built a highly scalable neural network by repeatedly stacking 3×3 small convolution kernels and 2×2 maximum pooling layers. This structure reduced training parameters using small convolution kernels. With this advantage, the number of feature maps could be greatly increased, and more information could be extracted. In 2014, the 22-layer Inception V1 model [11] reduced the top-5 error rate to 6.67% in the ILSVRC2014 competition. The structure of GoogLeNet was deeply influenced by the structure of the Network-in-Network model proposed by Lin et al. [12], which was characterized using the Inception module to build small network architecture within the model, so that the convolutional neural network could be deeper and wider and the number of training parameters was reduced. Besides, an additional classifier was used in the network to effectively solve the gradient disappearance problem. ResNet [13] proposed by He et al. introduced identity mapping in the residual network to solve the training problem under deep layers. It was inspired by the idea of the Highway Network [14], which added a direct connection next to the convolution channel. The straight-through channel could accelerate the training of the super-neural network and the accuracy of the model was greatly improved.

Traditional methods of image captioning are retrieval-based methods [15] and template-based methods [16]. Farhadi et al. [3] detected objects to form a triple of scene elements and infer the final text relying on description templates. Devlin et al. [15] discussed and summarized some characteristics of retrieval-based methods and the key issues that needed to be paid attention to in the image description tasks through experiments. They determined the key factors for the excellent description effect of this method. However, these two methods are too dependent on the training corpora and the generated text descriptions are lack diversity.

The early machine learning based image processing method uses some image processing operators to extract the features of the image, which adopts Support Vector Machine (SVM) [17] as a classifier to get the target of the image and then use the obtained target and its attributions as the foundation for a generated sentence. This approach overly relies on the extraction of image features and the rules of generating sentences. The difficulty of training large datasets limits the captioning effect.

Based on significant advances of encoder–decoder structure in machine translation [18,19], a generative model called Neural Image Caption (NIC) [20] was proposed. Vinyals et al. replaced the encoder Recurrent Neural Network (RNN) by a deep CNN and took an image as input. They improved the initial NIC model in the first MSCOCO competition with image model fine tuning and beam size reduction, which achieved overall improvement on different metrics [21]. With a similar idea, Karpathy et al. introduced a ranking model that combined visual and language modalities through a common, multimodal embedding, where the multimodal RNN model was utilized to generate descriptions of images [22]. Li and her team of Stanford University made an image semantic description system similar to NIC, called Neuraltalk, which used other models to map image regions to sentence segments [23]. Xu et al. [24] first introduced the attention mechanism of human visual system into the image description generation algorithm. Different from the NIC, the model used the feature map of the last convolutional layer of CNN as the image feature. At the decoding stage, the attention mechanism allowed the model to dynamically select particular image region features that needed attention.

Recently, some optimization algorithms of neural networks were related to image captioning. The multi-threaded learning control mechanism could minimize the training time of CNN [19]. Cao et al. [25] presented Bag-LSTM to obtain more text-related image features by feedback propagation. In addition, Yan et al. [26] proposed a hierarchical attention mechanism via using both the global CNN features and the local object features for improved results.

Despite these advancements, the realization of these refined methods has been accompanied with complicated network structures and the growing number of parameters. The captioning models then may lose controllability and effectiveness. In addition, it has limited ability to adapt the network architectures to other datasets and tasks. To address these above issues, we propose a simple end-to-end image captioning model with extended CNN architecture. In order to improve the accuracy of the CNN model, the widely used method is to deepen or widen the network. However, as the number of hyper-parameters increases (like channel and filter size), the difficulty of network design and computational overhead will also increase. Therefore, we propose an extended CNN structure based on residual learning to obtain rich representations of input images. The learned representations of each image are transmitted into language model to generate sentence.

Overall, the main contributions of this paper are:

- We develop a novel encoder–decoder framework to increase the accuracy of predicted sentences for image captioning via focusing on visual features.
- We first propose a new multi-branch CNN model based on residual learning for image captioning. The simple design consists of blocks of the same topology, which could improve the effect of feature extraction while reducing complexity and the number of parameters.
- Through comprehensive experiments, we validate the effectiveness of our model on three benchmark datasets: Flickr8k [27], Flickr30k [28] and MSCOCO [29]. Our method achieves state-of-the-art performance, showing competitive image captioning results on adopted evaluation metrics.
- We perform an extensive analysis of our expanded model, which could be easily used in other tasks because of its modularized architecture.

2. Model

In this paper, we propose a neural framework to generate descriptions from images. Inspired by the encoder–decoder model of machine translation, it is possible to replace the RNN encoding the source text in machine translation with CNN to encode the image, which aims to get a caption of the image. Figure 1 illustrates the overview of our proposed image captioning model. The extended CNN structure we presented works as the encoder. This multi-branch expansion approach improves the accuracy without increasing the complexity of structure and also reduces the number of hyper-parameters. Then the extracted feature vector from CNN is utilized as an input to the RNN decoder to generate captions. The improved network has a large receptive field that is important for learning the complex relationship of object categories.

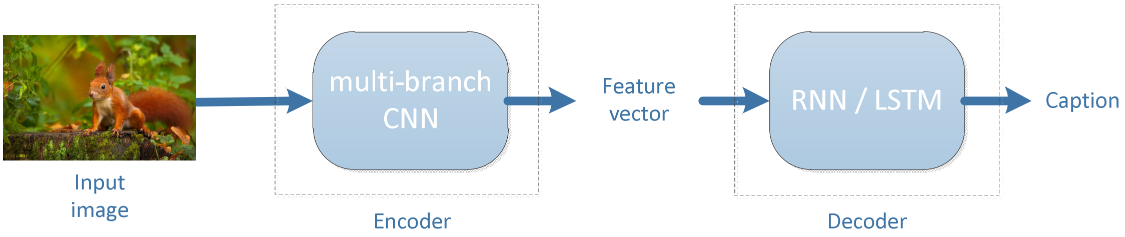


Figure 1. Overview of our captioning model based on encoder–decoder architecture.

2.1. ResNet

Most previous deep convolutional neural networks for image feature extraction are VGG and ResNet. VGGNet explores the relationship between the depth of the convolutional neural network and its performance. By repeatedly stacking 3 × 3 small convolution kernels and 2 × 2 maximum pooling layers, deep CNNs of 16 to 19 layers are successfully constructed. It is generally believed that the deeper the neural network (more complex, more parameters), the stronger the expressive ability. With this basic principle, the CNN classification network has evolved from the AlexNet of 7-layers to the VGGNet of 16 to 19 layers, even the GoogLeNet of 22 layers. However, we later found that after the depth of deep CNN reached a certain extent, the increase of layers did not lead to further improvement in classification performance. On the contrary, it would cause the network convergence to become slower and the classification accuracy of the test dataset to become worse.

The residual learning of ResNet can solve the problem that the accuracy does not increase but decrease when the network reaches a certain depth. The main idea of ResNet is to add a direct connection channel to the network, which is the core concept of the Highway Network. The previous network structure is a non-linear transformation of the performance inputs, while the Highway Network allows retaining a certain proportion of the output of the previous network layer. The idea of ResNet is very similar to that of the Highway Network, which allows the original input information to be directly transmitted to the later layer, as shown in Figure 2.

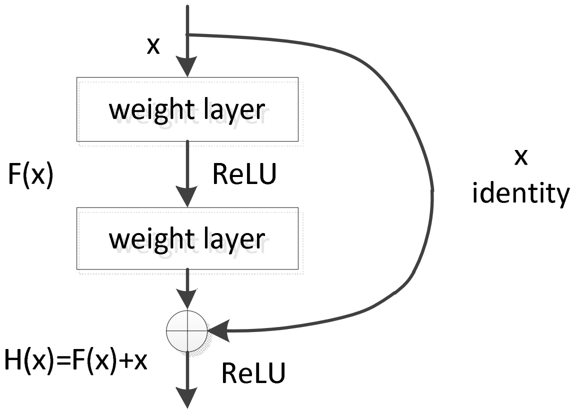


Figure 2. Residual learning block of ResNet.

Figure 2 shows the principle of deep residual learning. Denoting a stacked layers network as H, the output of this network block with x as an input will be $\mathrm{H(x)}$. Generally, CNN such as AlexNet or VGG will directly learn the expression of the parameter function H through training, thereby directly learning $\mathrm{x} \rightarrow \mathrm{H(x)}$. Residual learning is aimed to learn the residual mapping between input and output $\mathrm{F(x)} = \mathrm{H(x)} - \mathrm{x}$. The original mapping H(x) is rewritten into $\mathrm{F(x)} + \mathrm{x}$. Although both forms could approximate the desired functions, it is easier to learn and optimize the residual one.

Traditional convolutional networks or fully connected networks often have the problems of information loss, gradient disappearance and gradient explosion during transmission, which will lead to the deep network unable to train. ResNet solves the problem to a certain extent via bypassing the input information directly to the output to protect the integrity of information. The entire network only needs to learn the difference between the input and output, simplifying the learning objectives and difficulties. The biggest difference between ResNet and VGG is that ResNet has many bypasses that connect the input directly to the back layers. This structure is also called shortcut or skip connection.

2.2. Multi-Branch CNN

In the field of image captioning, the commonly used CNN are VGG and ResNet, which are also widely applied in image classification and image detection. We improved the encoder part of image captioning model by adopting the idea of stacking modules of VGG/ResNet to build a deep neural network. By repeatedly stacking modules of the same topology, we can reduce the selection of hyper-parameters and reduce the risk of over-adapting the hyper-parameters to certain datasets.

In the process of building a neural network, the most straightforward way to improve network performance is to increase the depth and width. Depth refers to the number of layers of the network and width refers to the number of channels per layer. However, these approaches bring some disadvantages: (1) It is prone to overfitting. As the depth and width continue to increase, the parameters that need to be learned are also increasing, and huge parameters are likely to cause overfitting. (2) Simply stacking large convolutional layers consumes a mass of computational resources. (3) It is difficult to optimize the model because the deeper the network is, the easier the gradient disappears.

According to [11], running filters with multiple sizes on the same level makes the network essentially “wider” rather than “deeper.” Unlike traditional convolutional neural networks, the Inception module could set multiple paths. Each path could have different operations and the same operation could also set different kernels, size and stride (see Figure 3).

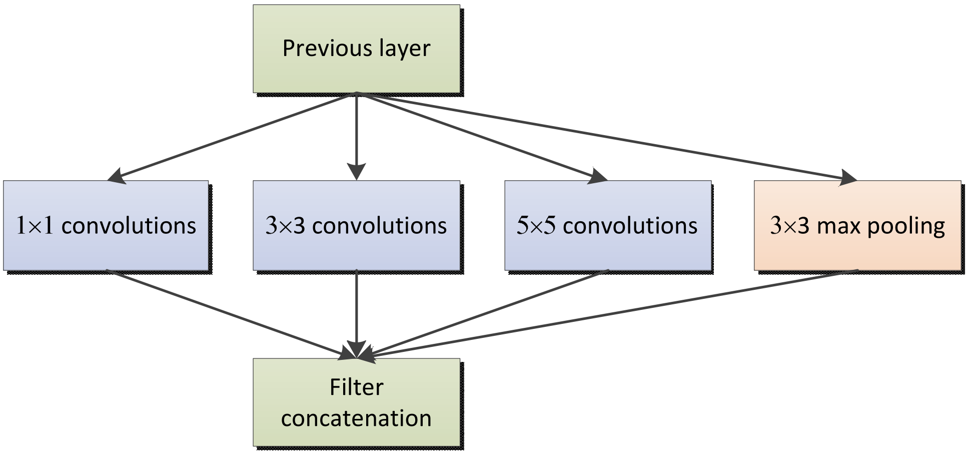
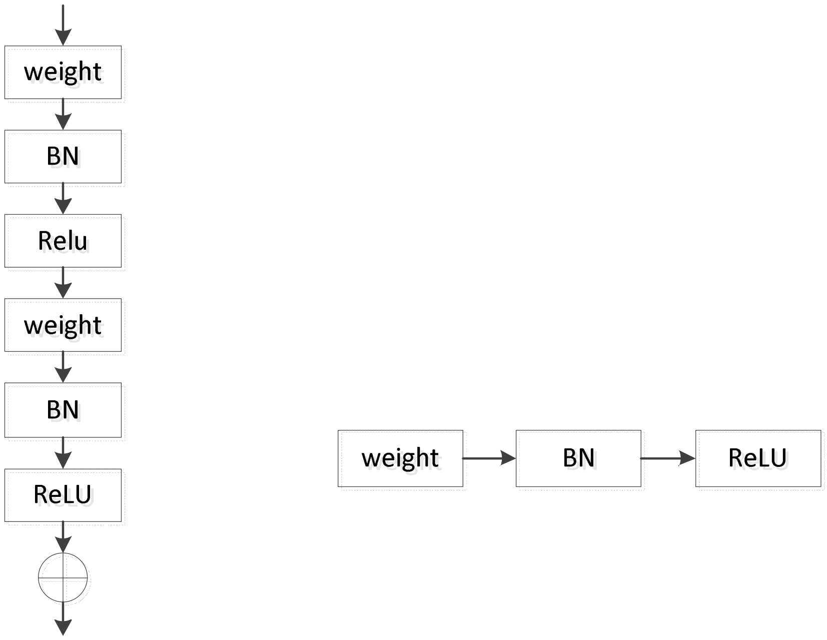


Figure 3. Multiple paths of an Inception module.

The complexity of the Inception architecture makes it difficult to modify and apply to different tasks. If we just simply expand the scale of the structure, the advantages of computing will disappear, and it is difficult to update to new situations while maintaining efficiency.

Therefore, we propose to improve ResNet with the multi-path idea of the Inception network. Figure 4a is a residual block in ResNet, in which BN (Batch Normalization) [30] and ReLU (Rectified Linear Unit) [31] act as activation functions. Figure 4b shows a unit in the residual block. We expand the network structure of Figure 4a by stacking blocks of the same topology in parallel to obtain a new CNN of unit*m_n, as shown in Figure 5, where m is the number of convolutional layers and n is the number of extended paths. Since the topology of each sub-block is the same, the network structure is more concise and modular while maintaining the network capacity.



(**a**) (**b**)

Figure 4. (**a**) A residual block of ResNet; (**b**) a unit in residual block.

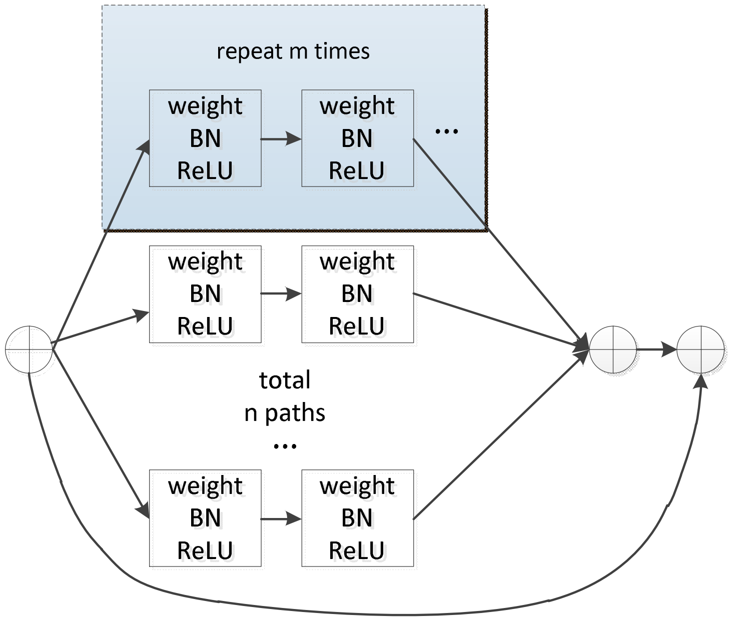


Figure 5. A unit*m_n module of proposed multi-branch convolutional neural network (CNN).

The improved ResNet by extending the residual module is shown in Figure 6. The improved encoder CNN can extract image features through multiple paths, which can reduce the loss of image information of various parts and extract deeper image features.

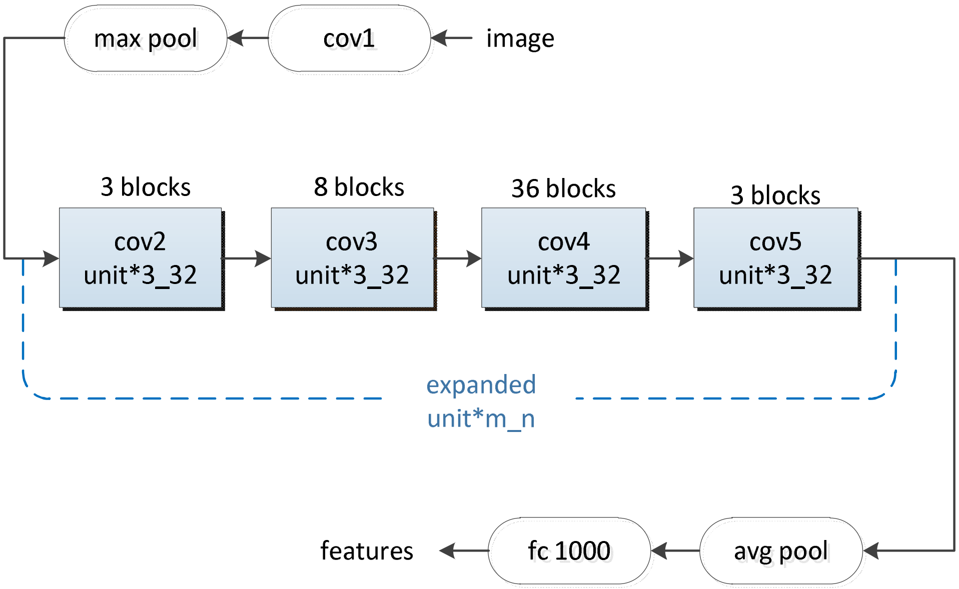


Figure 6. An expanded CNN for image features extraction.

2.3. LSTM

The recurrent neural network is a mature technology in NLP and plays an important role in machine translation and speech recognition. RNN can adequately mine the semantic information in time-series data and use the previous information to assist the current task [32]. In RNN, the output at the current time will remain in the network, and the output of the next time will be determined jointly by combining the input at the next time and the output of current time. RNN is as important in deep learning as convolutional neural network.

However, traditional RNN tends to face the challenge of long-term dependencies. The main reason is that RNN shares parameters in time dimension and repeatedly applies the same operation in the long time series. The deepened network structure is easy to have gradient disappearance or gradient explosion in the process of exploiting backpropagation and gradient descent algorithm. Therefore, the gradient cannot be effectively transmitted to the front network layer after multi-layer propagation during the training process, which makes RNN unable to capture the impact of long-distance information.

To overcome the potential issue of vanishing gradient faced by RNN, Hochreiter et al. [33] improved the RNN with an architecture called Long Short-Term Memory (LSTM). LSTM is a special kind of RNN, which memorizes the information flowing through the network through the memory cell and controls the information updating process through three special "gate" structures. LSTM is easier to learn long-term dependencies than simple recurrent architecture and can effectively capture context information in long sequences. It achieves good application effect in speech recognition, machine translation, image description generation and other application fields. Usually, a recurrent neural network has the form of a chain structure of repeating neural network modules. In ordinary RNN, the repeating modules just simply apply a nonlinear calculation after affine transformation. To replace hidden units in traditional RNN, LSTM introduced well-designed gate units to control information flow. In addition to calculating the hidden state h_t, LSTM maintains a memory cell state m_t, which selectively adds, retains or deletes information via gate units. The LSTM structure is shown in Figure 7.

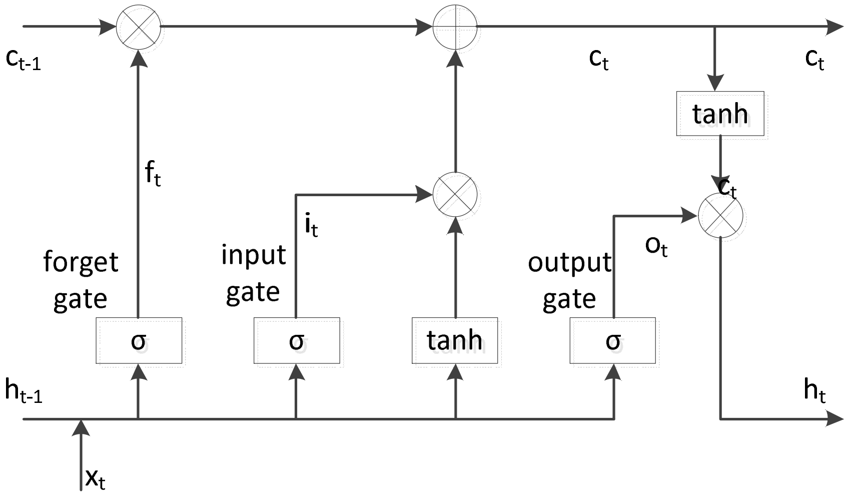


Figure 7. Long short-term memory (LSTM): the memory block controlled by three gates.

The calculation formula for each state update and output during the forward propagation of LSTM are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{3}$$

$$m_t = i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) + f_t \odot m_{t-1} \tag{4}$$

$$h_t = o_t \odot \tanh(m_t) \tag{5}$$

where x_t represents the current input vector at time t, and h_{t-1} and h_t represent the hidden layer states at time $t-1$ and time t, respectively. W_{ix}, W_{ih}, b_i represent the weight and offset of the input gate and the input gate determines how much of the current input is to enter the memory cell. W_{fx}, W_{fh}, b_f represent the weight and offset of the forget gate and the forget gate controls the network to discard useless information. W_{ox}, W_{oh}, b_o respectively represent the weight and offset of the output gate and the output gate determines the current output based on the current input and the information in the latest memory cell. σ represents the sigmoid function, which obtains a probability value between 0 and 1 to control the degree of update of the cell states. $\odot$ indicates element-wise multiplication.

3. Experiments and Result

3.1. Evaluation Metrics

There is no standard evaluation method in image captioning tasks, yet the generation results can be measured by calculating the correspondence between the reference description and the prediction description. Therefore, we use Bilingual Evaluation Understudy (BLEU, B@1, B@2, B@3, B@4) [34], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [35] and Consensus-based Image Description Evaluation (CIDEr) [36] as evaluation metrics. The three evaluation metrics will be presented in detail below.

- BLEU

BLEU (Bilingual Evaluation Understudy) is a quality score metric developed to evaluate the prediction results of machine translation systems. This evaluation method counts the matching n-grams of the candidate translation and the reference text, where n means the number of ordered words in comparison (1-g compares each word and 2-g compares each word pair). It measures the n-gram correlation between prediction descriptions and reference sentences based on the accuracy analysis.

The closer the prediction description and the reference description are, the higher the BLEU score is. In BLEU@N (N = 1, 2, 3, 4), N represents the length of the n-gram. The longer the n-gram, the more difficult it is to match. If only calculate BLEU for N = 1, it is easy to obtain a relatively high evaluation on a description of low quality. Thus, we calculated four levels of BLEU scores in the actual comparison.

- METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is a more rigorous evaluation metric than BLEU. It is initially used to evaluate machine translation effects and can also be applied to image captioning tasks. METEOR evaluates a description by computing a score based on explicit word-to-word matches between the generated description and a reference sentence. This method obtains the final score by calculating unigram precision and unigram recall, then combining the precision and recall by a harmonic-mean to get Fmean. METEOR can solve the inconsistency problem between the accuracy scores and the generated results of BLEU and has high correlation with human judgments than BLEU. It has greatly improved the correlation with human judgements at the segment level and sentence level.

- CIDEr

CIDEr (Consensus-based Image Description Evaluation) is an automatic evaluation metric that designed specifically for image captioning. It measures the consistency of image captioning by calculating the TF-IDF (Term Frequency-Inverse Document Frequency) weight of each n-gram. CIDEr is a combination of BLEU and vector space models. It treats each sentence as a document and then calculates the average cosine similarity between the candidate sentence and the reference sentence. The advantage of CIDEr is that different n-grams have different weights depending on TF-IDF, because more common n-grams in the entire corpus contain a smaller amount of information. The main evaluation point of image captioning is to see if the model has captured key information, thus it is necessary to reduce the weight of n-grams with less importance.

There are some other evaluation metrics we did not present in our experiments because of some limitations. ROUGE (Recall Oriented Understudy for Gisting Evaluation) [37] from text summarization tends to reward long sentences with high recall. Unfortunately, it has been shown to correlate weakly with human judgment. SPICE (Semantic Propositional Image Caption Evaluation) [38] evaluates the similarity of scene graphs constructed from the candidate and reference sentence. It correlates well with human judgments, but fails to capture the syntactic structure of a sentence and it is not widely used in classic image captioning method.

3.2. Datasets

We performed experiments on the following datasets.

Flickr8k, Flickr30k and MSCOCO are classic datasets of image captioning field. The Flickr8k dataset is a popular dataset composed of 8000 images in total collected from Flickr. Each image in the dataset is accompanied with five reference captions annotated by humans. Similar to Flickr8k, the Flickr30k dataset contains 31,000 images collected from Flickr, together with five reference sentences provided by human annotators. MSCOCO contains images of nature and common objects. Images of MSCOCO are in more complicated background with a large number of objects of small size. Each image of MSCOCO2014 has been annotated with five to seven descriptive sentences that are relatively visual and unbiased.

These datasets were divided into a training set, a validation set and a test set, respectively. The statistics of datasets are shown in Table 1:

Table 1. Datasets for image captioning evaluation.

Dataset	Size		
	Train	**Val**	**Test**
Flickr8k	6000	1000	1000
Flickr30k	29,000	1000	1000
MSCOCO	82,783	40,504	40,775

In order to avoid overfitting, we explored using a pre-trained encoder CNN on our model. Convolutional neural network requires large-scale label data to train a large number of parameters. The advantages of the convolutional neural network are not reflected well with a small database. In this case, we considered pretraining CNN on the large-scale dataset ImageNet to complete the corresponding tasks. ImageNet is a large-scale visualization database for visual object recognition software research and is widely used in deep learning. The pre-trained CNN on ImageNet has strong generalization ability and can be utilized in many research fields such as image classification, object location and object detection.

3.3. Implementation Details

All the experiments were conducted on a server embedded with NVIDIA RTX 2080Ti GPU and installed with the Ubuntu 16.04 operating system. We first trained the encoder CNN on ImageNet to avoid overfitting. The input image from ImageNet was resized to 224 × 224 randomly according to the scale and aspect ratio augmentation. Down-sampling of conv3, 4, and 5 was implemented by stride-2 convolutions in the 3 × 3 layer of the first block in each stage. Batch normalization was performed right after the convolutions and ReLU was performed right after each BN, except for the output of the block where ReLU was performed after the adding to the shortcut. Gradient descent was an effective approach to optimize the algorithm and minimize the cost of the function. BDG (Batch Gradient Descent) calculates all instances for each iteration to achieve global optimum, but it can take a long time for large amounts of data. Thus, we used SDG (Stochastic Gradient Descent) [39] with a mini-batch size of 256 for our large-scale dataset. In SDG, a few instances are selected randomly instead of the whole dataset for each iteration, which makes the update and learning much faster. The initial learning rate was set to 0.1. The momentum of the stochastic gradient descent was set to 0.9 and the weight decay was set to 0.001. The weights were initialized as in [40].

We trained the language model following the work of NIC. For Flickr8k, mini-batch size was set to 16, and for Flickr30k and MSCOCO, mini-batch size was set to 64. The learning rate was initialized with 0.0001 for Flickr8k and Flickr30k and 0.0005 for MSCOCO. The feature vector of the input image extracted from pre-trained CNN was linearly transformed to match the input dimension of LSTM. The dimension of the hidden layer of LSTM was set to 512 and the embedding dimensions of words and images were also set to 512.

We trained all sets of weights using stochastic gradient descent with a fixed learning rate. All weights were randomly initialized except for the CNN weights, which were left unchanged since changing them led to a negative impact. The maximum iteration period was set to 25 epochs. The BN layer was introduced to the decoder part of the model to speed up the convergence of model training. After the loss fell below two and stabilized, we add the CNN part in to train together. The caption generation process was ceased until a predefined max length of generated sentence was reached. Training the model took 70 h total on 2 GPUs.

3.4. Results

We compared the proposed method with state-of-the-art image captioning models. (1) Deep VS [22], NIC [20] and m-RNN [41] are end-to-end multimodal networks that adopt pre-trained CNN like VGG or ResNet as an encoder and RNN as a language model. (2) Soft-attention and hard attention [24] introduced two alternative attention-based mechanisms to the image caption generator. Soft deterministic attention was trained by standard back-propagation methods and hard stochastic attention was trained by maximizing a variational lower bound. (3) A spatial attention model was able to extract spatial image features and an adaptive attention mechanism [42] could use visual sentinel instead of a single hidden state to provide a fallback option to the decoder. (4) SCA-CNN [43] incorporated spatial attention and channel-wise attention in CNN to implement every feature entry in the multi-layer feature maps.

We empirically found our multi-branch model performed better, as illustrated in Tables 2–4. Table 2 shows results on the Flickr8k dataset. In Table 3, we present the results of the same experiments on Flickr30k Entities, adding spatial and adaptive attention mechanisms for comparison. Table 4 shows the performance of our method on MSCOCO.

Table 2. Performances compared with the state-of-the-art on Flickr8k dataset.

Flickr8k					
Method	**B@1**	**B@2**	**B@3**	**B@4**	**METEOR**
Deep VS [19]	57.9	38.3	24.5	16.0	-
NIC [20]	63	41	27	-	-
Soft-Attention	67.0	44.8	29.9	19.5	18.9
Hard-Attention	67.0	44.8	29.9	19.5	18.9
SCA- CNN-VGG	65.5	46.6	32.6	22.8	21.6
SCA-CNN-ResNet	68.2	49.6	35.9	25.8	22.4
Ours	69.9	50.6	36.8	26.5	23.0

Table 3. Performances compared with the state-of-the-art on Flickr30k dataset.

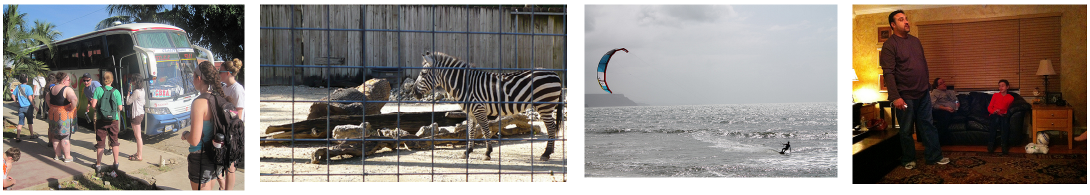
Flickr30k						
Method	**B@1**	**B@2**	**B@3**	**B@4**	**METEOR**	**CIDEr**
Deep VS	57.3	36.9	24.0	15.7	-	24.7
NIC	66.3	42.3	27.7	17.3	-	-
m-RNN	60	41	28	19	-	-
Soft-Attention	66.7	43.3	28.8	19.1	18.5	-
Hard-Attention	66.9	43.9	29.6	19.9	18.5	-
SCA-CNN-VGG	64.6	45.3	31.7	21.8	18.8	-
SCA-CNN-ResNet	66.2	46.8	32.5	22.3	19.5	-
Spatial	64.4	46.2	32.7	23.1	20.2	49.3
Adaptive	67.6	49.4	35.4	25.1	20.4	53.1
Ours	69.1	50.6	36.3	26.0	20.9	55.4

Table 4. Performances compared with the state-of-the-art on MSCOCO dataset.

MSCOCO						
Method	**B@1**	**B@2**	**B@3**	**B@4**	**METEOR**	**CIDEr**
Deep VS	62.5	45.0	32.1	23.0	-	66.0
NIC	66.6	46.1	32.9	24.6	-	-
m-RNN	67	49	35	25	-	-
Soft-Attention	70.7	49.2	34.4	24.3	23.9	-
Hard-Attention	71.8	50.4	35.7	25.0	23.0	-
SCA-CNN-VGG	70.5	53.3	39.7	29.8	24.2	-
SCA-CNN-ResNet	71.9	54.8	41.1	31.1	25.0	-
Spatial	73.4	56.6	41.8	30.4	25.7	102.9
Adaptive	74.2	58.0	43.9	33.2	26.6	108.5
Ours	76.3	59.8	44.9	34.2	27.5	111.3

We notice that the proposed method outperformed all compared approaches in terms of evaluation metrics, testifying that we are capable of generating image captions effectively. The outperformance of our method was due to the fact that our advanced CNN architecture extracted image features completely and effectively, which reveals that our model is more efficient to extract complete visual-semantic information.

We provided some qualitative examples in Figure 8 for a better understanding of our model. The average inference time for each image is about 240 ms for each image. The first three images are successful and the last one shows failure examples. Even in cases where the model produces inaccurate sentences, our generated results focused on reasonable regions of images. For instance, the inaccurate generated sentence of the fourth image may have been due to the deviation of object recognition. At the testing time, we set maximum length of a generated caption to 20 words. The object recognition was influenced by many factors in complicated scenes and images which are difficult to be fully described, which needs to be further improved in future research.



A group of people are standing on a sidewalk next to a bus.
Inference time: 254.96ms

A zebra is standing in a fenced enclosure with a fence.
Inference time: 233.69ms

A person is parasailing in the ocean on a sunny.
Inference time: 220.62ms

Three man are playing a game with remote controllers.
Inference time: 235.4ms

Figure 8. The sample images and generation sentence results with inference times.

Image captioning is a complex and high-level task. Our proposed multi-branch CNN architecture has a better performance in extracting advanced image features without increasing the depth of the network. However, the feature extraction via CNN still would lose some important information of images. It could not be applied well to some domains that require high precision for image features, such as medical image analysis, which requires the extraction of subtle visual features. Therefore, how to ensure that CNN loses less useful information and discards inferences is the issue we need to study in the next stage.

4. Conclusions

In this work, we have presented a novel deep learning framework for generating captions. We refined the encoder CNN to extract image features by expanding network structure via a simple design. A more effective RNN, LSTM, with visual feature vectors as input was adopted as the language model to generate corresponding descriptive captions. With a modularized structure, the multi-branch CNN has achieved superior performance than the state-of-the-art methods. Experimental results, conducted on Flickr8k, Flickr30k and MSCOCO, validate the effectiveness of our method in terms of accuracy and optimization. We compared our experimental results with previously published state-of-the-art results on three important evaluation metrics: BLEU, METEOR and CIDEr, as it is 15.06% higher than Soft-Attention method on METEOR metric conducted on MSCOCO. Besides, the proposed approach, with its advantages of simplicity and extendibility, could be applied to other CV problems. In the future, we plan to apply our model to other datasets and useful applications, such as video captioning.

Author Contributions: Conceptualization, Y.L. and S.H.; methodology, Y.L. and S.H.; formal analysis, Y.L. and S.H.; project administration, Y.L.; resources, Y.L.; supervision, Y.L.; validation, S.H.; visualization, S.H.; writing—original draft, S.H.; writing—review & editing, Y.L. and S.H.

Funding: This research was funded by the National Natural Science Foundation of China (61571013), by the Beijing Natural Science Foundation of China (4143061), by the Science and Technology Development Program of Beijing Municipal Education Commission (KM201710009003) and by the GreatWall Scholar Reserved Talent Program of North China University of Technology (NCUT2017XN018013).

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

References

1. Oliver, N.M.; Rosario, B.; Pentland, A.P. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 831–843. [CrossRef]
2. Cambria, E.; White, B. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48–57. [CrossRef]
3. 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 Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 15–29.
4. Deng, L. Deep learning: From speech recognition to language and multimodal processing. *APSIPA Trans. Signal Inf. Process.* **2016**, *5*. [CrossRef]
5. 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.* **2014**, *115*, 211–252. [CrossRef]
6. Mishkin, D.; Sergievskiy, N.; Matas, J. Systematic evaluation of CNN advances on the ImageNet. *Comput. Vision Image Understanding* **2017**, *161*, 11–19. [CrossRef]
7. Goldberg, Y.; Levy, O. Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.
8. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors forword representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
11. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
12. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
13. He, K.; Zhang, X.; Ren, S.; Jian, S. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Highway networks. *arXiv* **2015**, arXiv:1505.00387.
15. Devlin, J.; Cheng, H.; Fang, H.; Gupta, S.; Deng, L.; He, X.; Zweig, G.; Mitchell, M. Language Models for Image Captioning: The Quirks and What Works. *arXiv* **2015**, arXiv:1505.01809.
16. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Intelligence, M. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef] [PubMed]
17. Kotsia, I.; Pitas, I. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Trans. Image Process.* **2006**, *16*, 172–187. [CrossRef] [PubMed]
18. Cho, K.; Merrienboer, B.V.; Gulcehre, C.; Bougares, F.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
19. Połap, D.; Woźniak, M.; Wei, W.; Damaševičius, R. Multi-threaded learning control mechanism for neural networks. *Future Gener. Comput. Syst.* **2018**, *87*, 16–34. [CrossRef]
20. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.

21. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef]
22. Karpathy, A.; Fei-Fei, L.F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
23. Karpathy, A.; Joulin, A.; Fei-Fei, L.F. Deep fragment embeddings for bidirectional image sentence mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 8–13 December 2014; pp. 1889–1897.
24. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; 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, Lille, France, 6–11 July 2015; pp. 2048–2057.
25. Cao, P.; Yang, Z.; Liang, S.; Liang, Y.; Yang, M.Q.; Guan, R. Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory. *Neural Process. Lett.* **2019**, *50*, 1–17. [CrossRef]
26. Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image Captioning Based on a Hierarchical Attention Mechanism and Policy Gradient Optimization. *arXiv* **2018**, arXiv:1811.05253.
27. Rashtchian, C.; Young, P.; Hodosh, M.; Hockenmaier, J. Collecting image annotations using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010; pp. 139–147.
28. 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. [CrossRef]
29. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
30. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
31. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
32. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
33. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
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. Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
35. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 29 June 2005; pp. 65–72.
36. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
37. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of summaries. In *Text Summarization Branches Out*; 2004; pp. 74–81.
38. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. *Eur. Conf. Comput. Vis.* **2016**, *11*, 382–398.
39. Lecun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015.
41. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv* **2014**, arXiv:1412.6632.

42. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
43. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.