remote sensing

MDPI

Letter

LAM: Remote Sensing Image Captioning with Label-Attention Mechanism

Zhengyuan Zhang [1,2,3], Wenhui Diao [1,2], Wenkai Zhang [1,2], Menglong Yan [1,2], Xin Gao [1,2] and Xian Sun [1,2,*]

1 Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China; zhangzhengyuan16@mails.ucas.ac.cn (Z.Z.); whdiao@mail.ie.ac.cn (W.D.); zhangwk@aircas.ac.cn (W.Z.); yanml@aircas.ac.cn (M.Y.); gaxi@mail.ie.ac.cn (X.G.)
2 Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
3 School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: sunxian@mail.ie.ac.cn

Received: 12 September 2019; Accepted: 4 October 2019; Published: 10 October 2019

Abstract: Significant progress has been made in remote sensing image captioning by encoder-decoder frameworks. The conventional attention mechanism is prevalent in this task but still has some drawbacks. The conventional attention mechanism only uses visual information about the remote sensing images without considering using the label information to guide the calculation of attention masks. To this end, a novel attention mechanism, namely Label-Attention Mechanism (LAM), is proposed in this paper. LAM additionally utilizes the label information of high-resolution remote sensing images to generate natural sentences to describe the given images. It is worth noting that, instead of high-level image features, the predicted categories' word embedding vectors are adopted to guide the calculation of attention masks. Representing the content of images in the form of word embedding vectors can filter out redundant image features. In addition, it can also preserve pure and useful information for generating complete sentences. The experimental results from UCM-Captions, Sydney-Captions and RSICD demonstrate that LAM can improve the model's performance for describing high-resolution remote sensing images and obtain better S_m scores compared with other methods. S_m score is a hybrid scoring method derived from the AI Challenge 2017 scoring method. In addition, the validity of LAM is verified by the experiment of using true labels.

Keywords: remote sensing image captioning; remote sensing image; image understanding; semantic understanding

1. Introduction

Image captioning is a task aiming at generating natural language sentences to comprehensively describe the content of images. Compared with other tasks in the image field, such as classification, object detection [1,2], and semantic segmentation [3,4], image captioning can find more information about relationships between objects and scenes in the images. It also allows the images to be displayed in the form of language sentences and intuitively express the content of images. In addition, with the rapid development of remote sensing technology, quantities of remote sensing images can be easily accessed. This convenience stimulates the demand for semantically understanding remote sensing images and the demand of retrieving specific images in a large number of images. At the same time, the increase of quantity of remote sensing images brings more difficulties for managing such big data. Especially when a specific demand, such as searching for "warships in the harbor", needs to be

satisfied. In this situation, both "warship" and "harbor" are needed, and the relationship between these two objects is "in". In this situation, image captioning task can not only detect objects in the image but also describe the relationship between the objects with human language. Therefore, the remote sensing image captioning task can well solve the above problem and provide help for managing remote sensing data. What is more, this challenging task also plays an important role in many fields, such as scene understanding, image retrieval and military intelligence generation [5].

Compared to the template-based models [6–9] and retrieval-based models [10–12], encoder-decoder based models are good at generating length-variable and syntax-variable sentences. Therefore, encoder-decoder based models [13–15] are widely used in image caption tasks. Several works about natural image captioning have been done in recent years. Most of them are based on encoder-decoder framework, which was firstly proposed by J. Mao et al. [16] to generate sentence descriptions to explain the content of images. Later on, the attention-based method [14,17] was prevalent in image captioning task, owing to its effectiveness in adaptively focusing on the regions of interest in input images. This operation can make the extracted image features more salient. In addition to this, this purer image information can be provided for the subsequent decoder for generating more accurate sentences. Thus, the attention-based method is also adopted in the proposed model in this paper.

While attention-based models can achieve better results on natural image captioning datasets, attention-based models do not perform well in remote sensing image captioning without considering the characteristics of remote sensing images. This is because the characteristics of remote sensing images are different from those of natural images, such as the imaging angle of view and the foreground ratio. Remote sensing images are top-view images that cover a wide area and contain a wide variety of objects. In addition, the background in remote sensing images occupies a considerable proportion. All of the above issues increase the difficulty of applying attention mechanism in remote sensing images. Thus, for remote sensing image captioning task, when a particular word needs to be generated at one time step, the area of interest in the image should be focused more accurately by attention-based models.

Several works have been done on remote sensing image captioning task by utilizing high-level image features. This due to the fact that, compared with the middle-level image features, the high-level image features contain more global information, which contributes to generate better sentences for describing the remote sensing images. At the earliest, X. Zhang et al. [18] used the categories of input images to initialize the hidden state of an LSTM at the initial time step. This method replaces the high-level image features by the word embedding vectors and obtains good results. However, the middle-level image features of the input images are abandoned, which contains more detailed information. This detailed information is helpful for generating detailed description of objects or relationships in images. Later on, X. Zhang et al. [19] added an attention mechanism into the remote sensing image caption models. In addition, the high-level image features were also used as the attributes. The image features, highly relevant with the attributes, were used to calculate attention masks. However, the high-level image features used in this method are implicit and not specific. In addition, the high-level image features are directly concatenated with middle-level image features. This immediate concatenation cannot ensure that the global features and local features are fused well.

To solve the above issues, we adopt the predicted labels' word embedding vectors instead of high-level features. Moreover, the label information is introduced into the formation of attention layers instead of direct concatenation to improve the attention effects. In summary, the main contributions of this paper are as follows:

1. A novel attention mechanism, namely Label-Attention Mechanism (LAM), is proposed to guide the calculation of attention masks by using label information in attention models. The label information is instructive to attention masks and make them able to attend regions of interest according to the categories of input images. The label-guided image features expose more

label-related object information and relationships. Thus, complete description for input images can be easily generated.

2. The proposed LAM can provide more precise label information by adopting the predicted labels' word embedding vectors instead of high-level image features. The content of input images can be concisely represented by the predicted labels' word embedding vectors. Without redundant information in high-level image features, better sentences can be easily generated to obtain better scores. What is more, the label information is introduced into the calculation of attention masks instead of directly concatenated with middle-level image features. This way of fusion avoids middle-level image features diluting the label information. Simultaneously, it contributes to guiding the calculation of attention masks.

A series of experiments for remote sensing image captioning is performed on UCM-Captions, Sydney-Captions, and RSICD. The experiments infer that models applied LAM can achieve higher scores on multiple metrics than other methods. What is more, another experiment of using true labels is designed to verify the validity of LAM. More importantly, concise and comprehensive sentences can be easily generated by models applied LAM.

2. Related Works

2.1. Natural Image Captioning (NIC)

With the rapid development of computer vision and natural language processing. Many methods [20–24] have been tried for processing natural image captioning task. There are mainly three ways for generating sentences to describe natural images: template-based methods, retrieval-based methods, and encoder-decoder based methods.

Template-based methods need different pre-defined sentence templates for images with different categories in the datasets [6–9]. The objects in the images can be detected or classified, and their corresponding words will be filled in the blanks in pre-defined templates. In this way, whole sentences for describing natural images can be generated. However, template-based methods cannot be trained end-to-end. More seriously, the types of sentence templates are limited, and the lengths of sentence templates are not variable.

Retrieval-based methods need to retrieve similar images according to the query image in the training dataset [10–12] at first. Then, the retrieved images' corresponding sentences will be regarded as the sentences for describing the query image. Although the styles and content of the retrieved sentences can be different, the retrieved sentences cannot totally correctly describe the content of the query image. Especially, the details in the query image cannot be described in the retrieved sentences. What is more, Retrieval-based models cannot be trained end-to-end, either.

Encoder-decoder based methods are widely used in natural image captioning task and obtain many great results. In general, the encoders are composed of CNN, such as AlexNet [25], VGG [26], Inception [27], and ResNet [28] to extract image features from the given images. The decoders are generally made of RNN, GRU, or LSTM so as to generate sentences for describing the given images. Vinyals et al. [29] propose a neural image captioning (NIC) model at the earliest to generate sentences for describing natural images. In this NIC model, RNN is replaced by LSTM to avoid long term gradient dissipation. However, the high-level image features are only used to initialize LSTM at the first time step. Hence, the influence brought by image features decrease as time flows. What is more, the local image features are neglected and not utilized in this method. This may lead to some details being neglected in the generated sentences. This is because the local image features contain more richer image details than the global image features. Hence, K. Xu et al. [13] introduced attention mechanisms into encoder-decoder models. In this model, the local image features are adopted and attended by attention masks. At each time step, this attention model will focus on different regions in images and give different weights to image features. Then this attended image features, changed at each time step, will be fed into LSTM to generate words to describe the content of images. J. Lu et al. [14] propose

a model, which can adaptively choose image features or word features to help LSTM to predict the next word. If the predicted word is a visual word, the attention masks will focus on the image features more than word features.

2.2. Remote Sensing Image Captioning (RSIC)

B. Qu et al. [30] adopt a multimodal encoder-decoder method to generate sentences for describing images. In this method, CNN is utilized to extract image features, and the image features will then fused with hidden state to predict the word step by step. Finally, the whole sentences can be obtained by concatenating all the predicted words. This is the earliest work for processing remote sensing image captioning task. In addition, two remote sensing image captioning datasets, UCM-Captions and Sydney-Captions, are published for further research. X. Lu et al. [5] adopt an attention-based encoder-decoder model for generating sentences for remote sensing images and achieve better results. This proves that attention mechanism is adaptive in remote sensing image captioning task, although the remote sensing images are different with natural images. In addition, they create and publish a big new remote sensing dataset, RSICD. This new dataset contains more images and more sentence patterns than UCM-Captions and Sydney-Captions. X. Zhang et al. [19] propose a novel attribute attention mechanism for processing remote sensing image captioning. Models applying this attribute attention mechanism can dynamically focus on the image regions according to the attributes of images at each time step. However, this attribute attention utilizes the implicit high-level image features instead of specific words. Thus, the effects brought by attribute attention can be further improved.

3. Method

The whole architecture of the proposed method, based on an encoder-decoder framework, is illustrated in Figure 1, which can be seen as three parts. Firstly, image features can be extracted from the given remote sensing images by CNN. Secondly, the image features and last hidden state will be used to calculate the attention masks in models applied LAM. Image features can be attended by attention masks to improve the salience of key regions in images. Then, by summing up the attended image features, context vectors can be obtained. Finally, sentences will be generated with the help of context vectors by decoders. Models applied LAM can be trained end-to-end. The details are elaborated in the following parts.

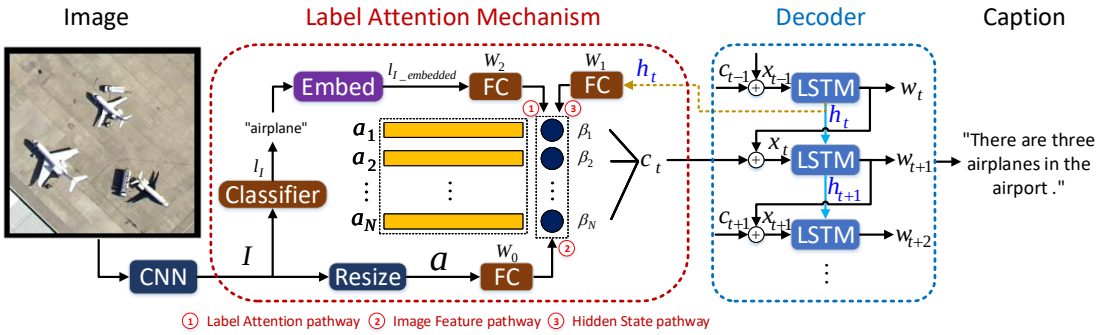

Figure 1. An overview of the proposed framework. It is composed of three parts, a CNN part for extracting image features, a Label-Attention Mechanism part for attending image features, and a decoder (LSTM) for generating captions. FC stands for fully connected layers. The classifier is also made of one fully connected layer. The text labels are embedded by the embedding layer, the same layer for embedding caption words.

3.1. Label-Attention Mechanism

Label-Attention Mechanism is also based on spatial attention [13], which is a method of attending regions of interest in the original images. Details of LAM architecture have been illustrated in Figure 1.

There are three pathways for calculating the attention mask β, which are termed as Label-Attention pathway, Image Feature pathway, and Hidden State pathway.

3.1.1. Label-Attention Pathway

Label-Attention pathway is a novel pathway, which is designed for calculating the attention masks in models. This pathway is composed of one classifier, one word embedding layer, and one fully connected layer. First of all, the category of an input image will be predicted by the classifier to obtain the label word l_I. Then processed by a word embedding layer, this label word l_I will be represented as a vector $l_{I\_embedded}$, which contains the label information of the given image. Finally, this embedded vector $l_{I\_embedded}$ will be transformed to a vector, whose dimension is 512, and adaptive for fusing vectors when calculating attention masks β.

3.1.2. Image Feature Pathway

Image Feature pathway is made up of a fully connected layer. The size of extracted image features $\boldsymbol{I}$, sent to this pathway, is $14 * 14$ pixels with 512 channels. Then the image features $\boldsymbol{I}$ will be reshaped from $14 \times 14 \times 512$ to 196×512 (N is 196 in Figure 1). In this way, these 196 pixels in the resized feature map $\boldsymbol{a}$ ($\boldsymbol{a} = \{\boldsymbol{a}_i\}, \mathrm{i} = 1, 2, \ldots, \mathrm{N}$) correspond to 196 regions in the original image. In this way, the later process of attending image features can be operated. Each region will be weighted by a different value. Those regions weighted by a large value will be salient in the feature map and vice versa.

3.1.3. Hidden State Pathway

Hidden State pathway only contains a fully connected layer. Similarly, the hidden state generated from LSTM at last time also needs to be transformed to a vector of 512 dimension by the fully connected layer to be the same as other vectors before sent to calculate attention masks β. It is worth noting that attention masks β are changed with the hidden state of LSTM at each time step.

The conventional attention-based method utilizes the visual and text information to generate sentences. The attention mask, β, is only calculated by image features $\boldsymbol{a}$ and hidden states $\boldsymbol{h}$. In this case, the attention ability of attention layers is weak because the information input to the calculation of attention masks is less and not representative. To improve the effect of attention, an extra parameter item, containing the label information of the given remote sensing image, has been introduced to the formula of calculating attention masks β. Multiplied by weight W_2, the embedded label $l_{I\_embedded}$ will be added to the conventional equation of calculating attention masks β and regarded as the third parameter item. Hence, label information of the given images is introduced to the attention masks β. Finally, by summing up the dot product of β (196×1) and $\boldsymbol{a}$ (196×512), the context vector $\boldsymbol{c}$ (512) can be obtained, which means that the sum of attended image features. The details are as follows:

$$l_{I\_embedded} = \mathrm{embed}(l_I) \tag{1}$$

$$\beta = \varphi(W_x(\sigma(W_0(\boldsymbol{a}) + W_1(\boldsymbol{h}) + W_2(l_{I\_embedded})))) \tag{2}$$

$$\boldsymbol{c} = \sum_{i=0}^{k^2} \beta_i \cdot \boldsymbol{a}_i \tag{3}$$

where l_I is the predicted label. W_0, W_1, and W_2 are trainable parameters. σ is a ReLU layer and φ is a SoftMax layer. In order to merge image feature, hidden state, and embedded label, they will be multiplied by W_0, W_1 and W_2, respectively to be consistent in dimension. $\boldsymbol{a}_i$ is a pixel at the i-th location in one channel.

3.2. Generating Sentences

Experimental results in [5] prove that the scores obtained by LSTM are better than RNN (Recurrent Neural Network). Therefore, the basic LSTM is used as the decoder in this procedure as in other methods [5,14,17]. The specific caption generation process has been showed in Figure 1.

In the training stage, at time step t, the image context vector $\boldsymbol{c}_t$ can be calculated by (3) at first. Next, the t-th predicted word w_t will be projected to an embedding vector $w_{t\_embedded}$. Then, one of the inputs of LSTM, $\boldsymbol{x}_t$, can be obtained by concatenating the image context vector $\boldsymbol{c}_t$ and the word embedding vector $w_{t\_embedded}$. Finally, by combining $\boldsymbol{x}_t$ with current hidden state $\boldsymbol{h}_t$, next hidden state $\boldsymbol{h}_{t+1}$ can be predicted with the help of LSTM. Then at next time step $t+1$, β_{t+1}, $\boldsymbol{c}_{t+1}$ and $\boldsymbol{x}_{t+1}$ can be recomputed to get $\boldsymbol{h}_{t+2}$. This loop will not stop until EOS has been output in the generate sentences. The equations are as follows:

$$w_{t\_embedded} = \text{embed}(w_t) \tag{4}$$

$$\boldsymbol{x}_t = [\boldsymbol{c}_t; E_s \cdot w_{t\_embedded}] \tag{5}$$

$$\boldsymbol{h}_{t+1} = \text{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_t) \tag{6}$$

where [;] is a concatenation operator. E_s is a matrix with the dimension of $\mathbb{R}^{h \times D}$. $\boldsymbol{h}$ is the dimension of embedding space.

The probability of predicted words can be calculated by SoftMax. The loss function of the proposed method is the sum of the negative log likelihood of the predicted correctly words at each step. The functions are as follows:

$$p(w_{t+1}) = softmax(E_o \cdot \boldsymbol{h}_{t+1}) \tag{7}$$

$$\text{loss}(I, S) = -\sum_{t=1}^{N} \log p(w_t) \tag{8}$$

where E_o is also a matrix with the dimension of $\mathbb{R}^{D \times h}$. I stands for the given image and, S is the corresponding sentence.

3.3. Extracting Image Features

All attention-based methods regard CNNs as an encoder to extract image features. Moreover, the extracted image features can be obtained from different convolutional layers of different models. For VGG, conv5_3 layer of VGG16 and conv5_4 layer of VGG19 are often used in the task of image captioning. For a fair comparison, we adopt the same CNN structure as the encoder as other methods [5,14,17], and select conv5_3 (refers to the activations of the 29-th convolution layer in VGG16 [26]) as the extracted image features. Then the extracted image features $\boldsymbol{I}$ will feed into the Label-Attention Mechanism module for calculating attention masks β and context vectors $\boldsymbol{c}$.

The testing stage is similar to the training stage. For input images, CNN is utilized to extract image features, which are provided for predicting the labels of input images and calculating attention layers' attention masks. Attention masks β are calculated by the predicted labels, image features, and hidden states. After the attention process, context vectors can be obtained and sent to LSTM for generating captions.

4. Experiments

In this section, datasets description and multiple evaluation metrics are presented at first. After that, details in the training process are showed. The experimental results and analyses are given at the end.

4.1. Dataset Description

4.1.1. UCM-Captions

UCM-Captions (https://pan.baidu.com/s/1mjPToHq#list/path=%2F) is a remote sensing image caption dataset, which is created based on UC Merced Land Use Dataset [31], initially used for scene classification task. This dataset contains 21 classes, including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. There are 100 images in each category, and the size of each image is $256 * 256$ pixels. In addition, X. Lu et al. [5] supplement five different natural sentences for describing the content of each image. Thus, this new created remote sensing image caption dataset is called UCM-Captions.

4.1.2. Sydney-Captions

Sydney-Captions (https://pan.baidu.com/s/1hujEmcG#list/path=%2F) is based on Sydney Dataset [32], which is also used for scene classification task. This dataset contains 7 classes, including residential, airport, meadow, river, ocean, industrial, and runway. There are totally 613 images in Sydney Dataset, and the size of each image is $500 * 500$ pixels. In addition, X. Lu et al. [5] also supplement five different natural language descriptions for each image in Sydney dataset. This new created remote sensing image caption dataset is called Sydney-Captions.

4.1.3. RSICD

RSICD (https://pan.baidu.com/s/1bp71tE3#list/path=%2F) is a larger remote sensing image caption dataset constructed by X. Lu et al. [5]. All of the images come from satellites or airplanes. There are 30 types of scenes, including airport, bridge, beach, baseball field, open land, commercial, center, church, desert, dense residential, forest, farmland, industrial, mountain, medium residential, meadow, port, pond, parking, park, playground, river, railway station, resort, storage tanks, stadium, sparse residential, square, school, and viaduct. Totally, RSICD contains $10,921$ remote sensing images, each of which is resized to 224×224 pixels and artificially annotated with 5 descriptive sentences. For improving the diversity, several rules are made for annotating the images [5]. Thus, RSICD is a complicated remote sensing image caption dataset and lower scores will be obtained by caption models.

4.1.4. Diversity Analysis

These three remote sensing image captioning datasets, UCM-Captions, Sydney-Captions, and RSICD, are compared in Table 1. "Caption Mean Length" means the mean length of all the captions in one dataset. "Vocab(before)" means the size of the vocabulary build from all the captions in one dataset. "Vocab(after)" means the size of the vocab, in which the words whose word frequency is less than 5 will be removed. It is easy to find that RSICD contains the greatest number of images. The vocabulary size and number of categories of RSICD are also the biggest. Sydney-Captions contains the fewest images. However, its mean length of captions is the longest.

Table 1. Datasets Comparison.

Datasets	Categories	Caption Mean Length	Vocab(before)	Vocab(after)	Count
UCM-Captions	21	11.5	315	298	2100
Sydney-Captions	7	**13.2**	231	179	**613**
RSICD	**30**	11.4	**2695**	**1252**	10,000

4.2. Evaluation Metrics

In this paper, multiple evaluation metrics are proposed to measure the quality of generated sentences. BLEU-n is the most popular evaluation metric in the area of machine translation, which

is built on the n-gram precision [33]. It measures how many words are shared by the generated captions and ground truth captions. In this paper, n is set to 4 as other works usually do. METEOR is a metric by computing unigram precision and recall against all ground truth sentences with some preprocessing on WordNet synonyms and stemmed tokens [34]. ROUGE-L is a metric for measuring the common subsequence with maximum length between the target sentence and the source sentence [33]. CIDEr is a metric designed for evaluating image descriptions by measuring human consensus in image caption [33]. SPICE is an F-score of the matching tuples in the predicted and reference scene graphs [35], which are built from captions. What is more, SPICE is not sensitive to n-gram and increases the diversity of generated sentences.

Since the evaluation metrics for image captioning are so many, the performance of different models cannot be easily judged. Moreover, referring to the Equation (9) in AI challenge 2017 (https://challenger.ai/competition/caption).

$$S_m^* = \frac{1}{4}(BLEU_4 + METEOR + ROUGE_L + CIDEr) \tag{9}$$

Considering the SPICE score, the equation is redesigned and regarded as the overall evaluation metric S_m. S_m will be higher by good performance of image caption models. The equation is as follows:

$$S_m = \frac{1}{5}(BLEU_4 + METEOR + ROUGE_L + CIDEr + SPICE) \tag{10}$$

4.3. Implementations

4.3.1. Preprocessing of Datasets

In this experiment, every dataset is divided into three parts with the default ratio, 80% training set, 10% validation set, and 10% testing set. We have shuffled each dataset with a fixed random seed value to avoid over-fitting or non-convergence in the training phase.

In addition, data augmentation is adopted in this paper. Owing to the size requirements for input images of VGG16 need be 224 × 224 pixels, in the training phase, input images are randomly cropped with 224 × 224 pixels and fed into the encoder. At testing time, input images will be resized to 224 × 224 pixels.

4.3.2. Fine-Tuning

To improve the ability of CNN, which has been pre-trained on ImageNet dataset [36], to extract image features for remote sensing images, the whole model, including the CNN and the classifier, is fine-tuned on UCM-Captions, Sydney-Captions, and RSICD, separately. In addition to this, for a fair comparison, the weights of fine-tuned CNN are shared in all methods in Tables 2–4. In detail, when UCM-Captions is used to train the model, the classifier's output dimension is set to 21. The output dimension for Sydney-Captions is set to 7. While for RSICD, it will be 30.

Table 2. Comparison experiments on UCM-Captions.

Methods	B1	B2	B3	B4	M	R	C	S	S_m
RNNLM [18]	0.7735	0.7119	0.6623	0.6156	0.4198	0.7233	3.1385	0.4677	1.0730
SAT [13]	0.7995	0.7365	0.6792	0.6244	0.4171	0.7441	3.1044	0.4951	1.0770
FC-Att+LSTM [19]	0.8102	0.7330	0.6727	0.6188	0.4280	0.7667	3.3700	0.4867	1.1339
SM-Att+LSTM [19]	0.8115	0.7418	0.6814	0.6296	0.4354	0.7793	3.386	0.4875	1.1435
SAT(LAM)	0.8195	0.7764	0.7485	0.7161	0.4837	0.7908	3.6171	0.5024	1.2219
SAT(LAM-TL)	**0.8208**	**0.7856**	**0.7525**	**0.7229**	**0.4880**	**0.7933**	**3.7088**	**0.5126**	**1.2450**
Adaptive [14]	0.808	0.729	0.665	0.610	0.430	0.766	3.138	0.487	1.0848
Adaptive(LAM)	0.817	0.751	0.699	0.654	0.448	0.787	3.280	0.503	1.1338
Adaptive(LAM-TL)	**0.857**	**0.812**	**0.775**	**0.743**	**0.510**	**0.826**	**3.758**	**0.535**	**1.2734**

4.3.3. Optimization

Adam optimizer has been used in the training stage with alpha 0.8 and beta 0.999. The initial learning rate of the decoder is 4×10^{-4} , while that of the encoder is 1×10^{-5} . All of these learning rates will be annealed by a factor of 0.8 every three epochs.

4.4. Results

In Tables 2–4, the "RNNLM" model stands for the method proposed by X. Zhang et al. [18]. "FC-Att+LSTM" and "SM-Att+LSTM" are the attribute attention models [19]. The method "SAT" stands for the "Show-Attend-and-Tell" model "Models(LAM)" means that models adopt Label-Attention Mechanism instead of the conventional attention mechanism, and "models(LAM-TL)" means that ground-truth labels have only been utilized in the testing phase. B1, B2, B3, B4, M, R, C, S stand for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE, separately. S_m is a mean score of five evaluation metrics. Thus, it is adopted to measure the sentences generated by models.

Table 3. Comparison experiments on Sydney-Captions.

Methods	B1	B2	B3	B4	M	R	C	S	S_m
RNNLM [18]	0.6861	0.6093	0.5465	0.4917	0.3565	0.6470	2.2129	0.3867	0.8188
SAT [13]	0.7391	0.6402	0.5623	0.5248	0.3493	0.6721	2.2015	0.3945	0.8283
FC-Att+LSTM [19]	0.7383	0.6440	0.5701	0.5085	0.3638	0.6689	2.2415	0.3951	0.8355
SM-Att+LSTM [19]	**0.7430**	0.6535	0.5859	0.5181	0.3641	0.6772	2.3402	0.3976	0.8593
SAT(LAM)	0.7405	0.6550	0.5904	0.5304	0.3689	0.6814	2.3519	0.4038	0.8671
SAT(LAM-TL)	0.7425	**0.6570**	**0.5913**	**0.5369**	**0.3700**	**0.6819**	**2.3563**	**0.4048**	**0.8698**
Adaptive [14]	0.7248	0.6301	0.5565	0.4945	0.3609	0.6513	2.1252	0.3973	0.8058
Adaptive(LAM)	0.7323	0.6316	0.5629	0.5074	0.3613	0.6775	2.3455	0.4243	0.8631
Adaptive(LAM-TL)	**0.7365**	**0.6440**	**0.5835**	**0.5348**	**0.3693**	**0.6827**	**2.3513**	**0.4351**	**0.8746**

Table 4. Comparison experiments on RSICD.

Methods	B1	B2	B3	B4	M	R	C	S	S_m
RNNLM [18]	0.6098	0.5078	0.4367	0.3814	0.2936	0.5456	2.4015	0.4259	0.8096
SAT [13]	0.6707	0.5438	0.4550	0.3870	0.3203	0.5724	2.4686	0.4539	0.8403
FC-Att+LSTM [19]	0.6671	0.5511	0.4691	0.4059	0.3225	0.5781	2.5763	0.4673	0.8700
SM-Att+LSTM [19]	0.6699	0.5523	0.4703	0.4068	0.3255	0.5802	2.5738	0.4687	0.8710
SAT(LAM)	0.6753	0.5537	0.4686	0.4026	0.3254	0.5823	2.5850	0.4636	0.8717
SAT(LAM-TL)	**0.6790**	**0.5616**	**0.4782**	**0.4148**	**0.3298**	**0.5914**	**2.6672**	**0.4707**	**0.8946**
Adaptive [14]	0.6621	0.5415	0.4667	0.4015	0.3204	0.5823	2.5808	0.4623	0.8694
Adaptive(LAM)	0.6664	0.5486	0.4676	0.4070	0.3230	0.5843	2.6055	**0.4673**	0.8774
Adaptive(LAM-TL)	**0.6756**	**0.5549**	**0.4714**	**0.4077**	**0.3261**	**0.5848**	**2.6285**	0.4671	**0.8828**

4.4.1. Quantitative Comparison

The results of remote sensing image captioning by different experimental methods are illustrated in Tables 2–4. It is not hard to find that models applied LAM, "SAT(LAM)" and "Adaptive(LAM)", can get higher scores on S_m than conventional attention models, "SAT", and "Adaptive", on UCM-Captions, Sydney-Captions and RSICD. This is because the introduction of label information can lead attention masks focus on regions of interest according to the categories of input images. With salient category-related image features, detailed descriptions of objects and relationships can be easily generated. In other words, the attention layers in models applied LAM are able to be conscious of the categories of input images and "look at" the regions of interest accurately. In this way, more accurate detailed image information can be provided for the subsequent decoder to generate concise and complete sentences.

It is noteworthy that the "RNNLM" model gets really the lowest scores in Tables 2–4. This infers that middle-level image features are richer than the high-level image features. Only using the high-level

image features to initialize the RNN is not enough. Both of them should be used together. Thus, both the high-level image features and middle-level image features should be used for improving the performance of generating captions.

Compared with the conventional attention model "SAT", both "FC-ATT+LSTM" and "SM-ATT+LSTM" introduce high-level image features into attention layers in models. The introduction of high-level image features can lead the attention layers to acquire global information of images to focus on regions of interest more accurately. Thus, both "FC-ATT+LSTM" and "SM-ATT+LSTM" can achieve higher scores than "SAT". However, models applied LAM can obtain higher scores than "FC-ATT+LSTM" and "SM-ATT+LSTM". This is owing to the fact that although high-level image features contain label information of images, they are not the specific words and cannot concisely and accurately stand for the content of images. In contrast, in models applied LAM, the predicted categories' word embedding vectors are adopted instead of the high-level image features. In this way, the content of images can be concisely represented by word embedding vectors. Representing the content of input images in the form of word embedding vectors can avoid redundant information in image features. Simultaneously, purer and more useful image features can be remained for generating sentences. In addition, these word embedding vectors are introduced into the equation of calculating attention masks. This way of fusion can lead the attention masks to learn to focus on different key regions in images of different categories. The more details in images are attended by attention masks, the more complete sentences can be generated by models. Therefore, the introduction of specific categories' word embedding vectors is effective in generating sentences for describing the content of images.

The categories of input images are predicted by models applied LAM. Thus, there must exist wrongly predicted categories of input images. To verify the validity of LAM, the experiments using ground-truth labels in the testing phase have been designed on different models. From the experimental results in Tables 2–4, model applied LAM-TL, "SAT(LAM-TL)", and "Adaptive(LAM-TL)" can obtain a little higher scores on S_m than models applied LAM, "SAT(LAM)", and "Adaptive(LAM)". These results can prove that these several comparison experiments can verify the effectiveness of LAM. More significantly, whether the labels are predicted or ground-truth, scores obtained by models applied LAM are much higher than those obtained by conventional attention models.

It is well known that training models with large dataset can avoid over-fit of models. However, for image captioning task, more images may not bring better results by evaluating the generated sentences. From Tables 2–4, it is easy to find that, compared with UCM-Captions, higher scores are obtained by models on Sydney-Captions and RSICD. The scores obtained by models on RSICD are the lowest. This result infers that image caption models are sensitive to the vocabulary size. Larger vocabulary size may bring much difficulties for training the models, even if the models are trained on the dataset with large quantities. What is more, the image caption models are also sensitive to the caption mean length and the quantity of the dataset. For UCM-Captions and Sydney-Captions, their vocabulary size is almost the same as in Table 1. However, the caption mean length of Sydney-Captions is longer than that of UCM-Captions, and the quantities of images are much more than those of Sydney-Captions. The scores obtained by models on Sydney-Captions are lower than models on UCM-Captions.

4.4.2. Qualitative Comparison

The captions, generated by SAT, and SAT(LAM), and the ground-truth are shown with corresponding images in Figure 2. From the point of view of generated sentences, the sentences generated by models applied LAM are more accurate and complete in describing scenes and object attributes. What is more, the generated sentences contain more detail descriptions than those generated by the conventional attention models.

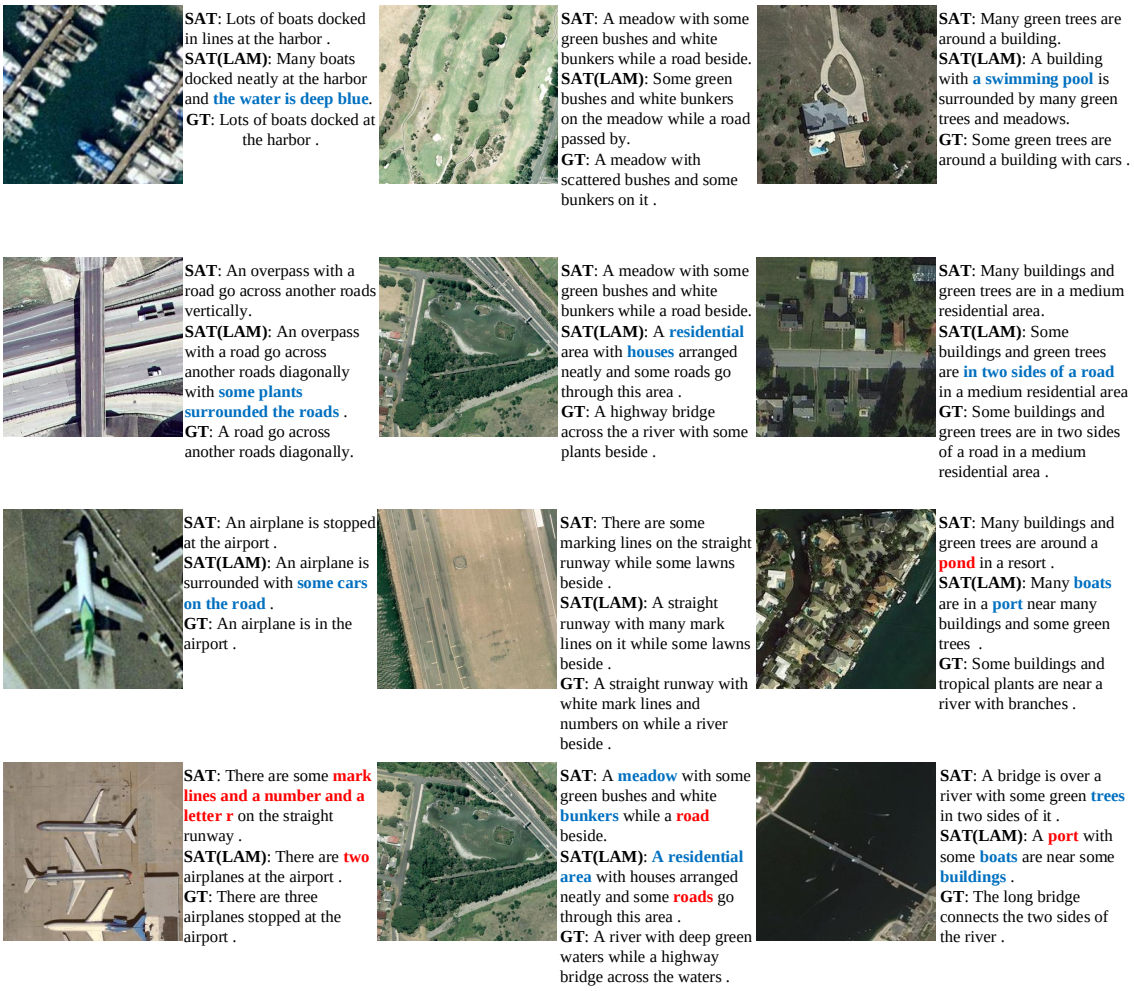

Figure 2. The first column is UCM-Captions, the second column is Sydney-Captions, and the last column is RSICD. The last row shows some negative results. The right side of the images are captions generated by SAT and SAT(LAM), and ground truth, separately. The blue words describe the scene in images but does not exist in ground truth captions. The scene represented by the red words does not exist in the images.

Especially the blue parts in generated sentences by SAT(LAM) can describe more contents than SAT. For the first image in the first column, the extra description of water is generated by SAT(LAM). What is more, the water's color attribute is also described in the generated sentence. For the second image in the first column, compared with the subject overpass, the object plants can also be detected and described in the generated sentence. In addition, the location of the described objects is also described. It is similar to the second image in the third column. For the third image in the first column, the cars, which are not the main objects in the airport, can also be described. It is similar to the second image in the second column, and the first image and the third image in the third column. For the last row, some negative results are shown. SAT(LAM) can generate the sentences to describe the airport. However, the number of airplanes is wrongly predicted. This may be due to the fact that the captions contain numbers are limited. In addition, models based on classification CNNs are not sensitive to numbers. If the encoder is based on detection networks, the performance may be improved. For the second image and the third image, SAT(LAM) fails to predict the "bridge" word in the generated sentence. For humans, the second image is hard to identify whether it is a road or a bridge. However, for the third image, SAT(LAM) indeed neglects the bridge in the generated sentence. It can be inferred that the category of the third image may be predicted wrongly. This wrong category "port" leads SAT(LAM) to find boats in the third image and ignore the bridge.

Above all, the results of these experiments infer that LAM is indeed effective. Higher scores can be obtained by models applied LAM. More complete and accurate sentences can be generated. Moreover, LAM can be utilized widely in models under the encoder-decoder framework.

4.4.3. Parameter Analysis

In order to evaluate the generated sentences based on different CNNs features, the experiments based on different CNN architectures are conducted in this section. In this paper, VGG11, VGG16, and VGG19 [26] are adopted for evaluating the influence brought by different CNNs. The comparison experiments on UCM-captions, Sydney-captions, and RSICD are shown in Tables 5–7, separately.

It is interesting to find that, for UCM-Captions and Sydney-Captions, SAT model based on VGG11 obtains the highest scores, and SAT(LAM) model based on VGG16 gets the highest scores. For RSICD, SAT model based on VGG16 obtains the highest scores, and SAT(LAM) model based on VGG19 gets the highest scores. These results infer that if LAM is applied on models, deeper CNN backbones are needed to obtain higher scores. This can be explained by the fact that, in models applied LAM, the CNN features are not only sent to the encoder but also provided for the classifier to classify the categories of the given images. More tasks need to be performed, thus, CNN backbones with higher ability of discrimination are needed. In addition, RSICD is the most complicated remote sensing image caption dataset. Therefore, compared with UCM-Captions and Sydney-Captions, deeper CNN backbones are needed to extract more discriminative image features.

Table 5. Comparison experiments on UCM-Captions based on different CNNs.

Methods	CNN	B1	B2	B3	B4	M	R	C	S	S_m
SAT [13]	VGG11	0.7834	0.7248	0.6781	0.6163	**0.4190**	0.7402	**3.1406**	0.4728	**1.0778**
	VGG16	**0.7995**	**0.7365**	**0.6792**	**0.6244**	0.4171	0.7441	3.1044	**0.4951**	1.0770
	VGG19	0.7830	0.7172	0.6620	0.6116	0.4109	**0.7497**	3.1221	0.4684	1.0725
SAT(LAM)	VGG11	0.7809	0.7129	0.6598	0.6139	0.4235	0.7424	3.2512	0.4625	1.0987
	VGG16	**0.8195**	**0.7764**	**0.7485**	**0.7161**	**0.4837**	**0.7908**	**3.6171**	**0.5024**	**1.2219**
	VGG19	0.7876	0.7275	0.6785	0.6339	0.4294	0.7424	3.2584	0.4635	1.1055

Table 6. Comparison experiments on Sydney-Captions based on different CNNs.

Methods	CNN	B1	B2	B3	B4	M	R	C	S	S_m
SAT [13]	VGG11	0.7226	**0.6495**	**0.5888**	**0.5329**	0.3524	**0.6882**	**2.2567**	**0.4082**	**0.8477**
	VGG16	**0.7391**	0.6402	0.5623	0.5248	0.3493	0.6721	2.2015	0.3945	0.8283
	VGG19	0.7170	0.6371	0.5765	0.5268	**0.3565**	0.6536	2.2477	0.4026	0.8374
SAT(LAM)	VGG11	0.7197	0.6394	0.5772	0.5211	0.3626	0.6597	2.3337	0.4021	0.8558
	VGG16	**0.7405**	**0.6550**	**0.5904**	**0.5304**	**0.3689**	**0.6814**	**2.3519**	**0.4038**	**0.8671**
	VGG19	0.7084	0.6249	0.5616	0.5085	0.3663	0.6586	2.3088	0.4123	0.8509

Table 7. Comparison experiments on RSICD based on different CNNs.

Methods	CNN	B1	B2	B3	B4	M	R	C	S	S_m
SAT [13]	VGG11	0.6623	0.5390	0.4576	**0.3946**	0.3114	0.5711	2.4325	**0.4574**	0.8334
	VGG16	0.6707	0.5438	0.4550	0.3870	**0.3203**	**0.5724**	**2.4686**	0.4539	**0.8403**
	VGG19	**0.6756**	**0.5514**	**0.4605**	0.3889	0.3192	0.5687	2.4681	0.4537	0.8397
SAT(LAM)	VGG11	0.6681	0.5453	0.4644	0.3931	0.3246	0.5805	2.5801	0.4585	0.8674
	VGG16	0.6753	0.5537	0.4686	0.4026	0.3254	0.5823	2.5850	0.4636	0.8717
	VGG19	**0.6780**	**0.5620**	**0.4773**	**0.4118**	**0.3267**	**0.5870**	**2.6662**	**0.4697**	**0.8923**

5. Conclusions

In this paper, a novel Label-Attention Mechanism is proposed for remote sensing image captioning task. This proposed LAM is able to solve the implicit label information problem. What is more, it also avoids directly concatenating global information with local features. In models applied LAM, the predicted labels' word embedding vectors are adopted as global information instead of high-level features. These word embedding vectors are more explicit and specific. Moreover, the word embedding

vectors are introduced into the formation of attention layers to improve the attention effects. Simply, LAM can utilize label information of the given image to calculate the attention mask better and make the attended image features more salient. More importantly, models applied LAM can obtain better performance, achieve higher scores, and generate more comprehensive and concise sentences in remote sensing image captioning.

Author Contributions: Z.Z. designed the method. Z.Z. implemented the method and wrote the paper. W.D., W.Z., M.Y., X.G., and X.S. contributed to the supervision of the work, analysis of the method, and paper writing.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 41701508.

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

References

1. Zhang, C.; Zou, T.; Wang, Z. A Fast Target Detection Algorithm for High Resolution SAR Imagery. *J. Remote Sens.* **2005**, *9*, 45–49.
2. Wang, S.; Gao, X.; Sun, H.; Zheng, X.; Sun, X. An Aircraft Detection Method Based on Convolutional Neural Networks in High-Resolution SAR Images. *J. Radars* **2017**, *6*, 195–203. [CrossRef]
3. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Hong, J.; Sun, J.; Zhang, Y.; Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
4. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction From Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
5. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
6. Ordonez, V.; Kulkarni, G.; Berg, T.L. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; pp. 1143–1151.
7. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [CrossRef]
8. Sun, C.; Gan, C.; Nevatia, R. Automatic Concept Discovery from Parallel Text and Visual Corpora. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2596–2604.
9. Gong, Y.; Wang, L.; Hodosh, M.; Hockenmaier, J.; Lazebnik, S. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 529–545.
10. 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.* **2016**, *119*, 46–59. [CrossRef]
11. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D.A. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 15–29.
12. 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, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1601–1608.
13. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; 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 Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
14. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250. [CrossRef]

15. Chen, S.; Zhao, Q. Boosted Attention: Leveraging Human Attention for Image Captioning. In Proceedings of the 15th European Conference on Computer Vision—ECCV2018, Munich, Germany, 8–14 September 2018; pp. 72–88. [CrossRef]
16. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Yuille, A. Explain Images with Multimodal Recurrent Neural Networks. *arXiv* **2014**, arXiv:1410.1090.
17. Aneja, J.; Deshpande, A.; Schwing, A.G. Convolutional Image Captioning. *arXiv* **2017**, arXiv:1711.09151.
18. Zhang, X.; Li, X.; An, J.; Gao, L.; Hou, B.; Li, C. Natural language description of remote sensing images based on deep learning. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2017, Fort Worth, TX, USA, 23–28 July 2017; pp. 4798–4801. [CrossRef]
19. Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. Description Generation for Remote Sensing Images Using Attribute Attention Mechanism. *Remote Sens.* **2019**, *11*, 612. [CrossRef]
20. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *arXiv* **2015**, arXiv:1412.6632.
21. Karpathy, A.; Feifei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
22. Chen, X.; Zitnick, C. Learning a Recurrent Visual Representation for Image Caption Generation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
23. Fang, H.; Gupta, S.; Iandola, F.N.; Srivastava, R.K.; Deng, L.; Dollar, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. From captions to visual concepts and back. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1473–1482.
24. Karpathy, A.; Joulin, A.; Li, F. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1889–1897.
25. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 141, pp. 1097–1105.
26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
27. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WI, USA, 26 June–1 July 2016; pp. 770–778.
29. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
30. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems, Kunming, China, 6–8 July 2016; pp. 1–5.
31. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
32. Zhang, F.; Du, B.; Zhang, L.; Sensing, R. Saliency-Guided Unsupervised Feature Learning for Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**. *53*, 2175–2184. [CrossRef]
33. Li, L.; Tang, S.; Deng, L.; Zhang, Y.; Tian, Q. Image Caption with Global-Local Attention. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4133–4139.
34. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 21–26 July 2017; pp. 5263–5271. [CrossRef]

35. Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; Cottrell, G.W. Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 21–26 July 2017; pp. 7378–7387. [CrossRef]
36. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.S. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]