Knowledge-Based Systems 203 (2020) 105920

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning

Xiangqing Shen [a,b], Bing Liu [a,b,c,*], Yong Zhou [a,b], Jiaqi Zhao [a,b], Mingming Liu [d,a,e,*]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China
[b] Mine Digitization Engineering Research Center of Ministry of Education, China,
[c] Institute of Electrics, Chinese Academy of Sciences, Beijing, 100190, China
[d] School of Intelligent Manufacturing, Jiangsu Vocational Institute of Architectural Technology, Xuzhou, Jiangsu 221008, China
[e] School of Mechatronic Engineering, Jiangsu Normal University, Xuzhou, Jiangsu 221008, China

ARTICLE INFO

Article history:
Received 15 January 2020
Received in revised form 9 April 2020
Accepted 13 April 2020
Available online 23 April 2020

Keywords:
Transformer
Variational Autoencoder
Transfer learning
Remote sensing image captioning
Self-attention mechanisms
Convolutional neural network
Reinforcement learning

ABSTRACT

Image captioning, i.e., generating the natural semantic descriptions of given image, is an essential task for machines to understand the content of the image. Remote sensing image captioning is a part of the field. Most of the current remote sensing image captioning models suffered the overfitting problem and failed to utilize the semantic information in images. To this end, we propose a **V**ariational Autoencoder and **R**einforcement Learning based **T**wo-stage **M**ulti-task Learning **M**odel (VRTMM) for the remote sensing image captioning task. In the first stage, we finetune the CNN jointly with the Variational Autoencoder. In the second stage, the Transformer generates the text description using both spatial and semantic features. Reinforcement Learning is then applied to enhance the quality of the generated sentences. Our model surpasses the previous state of the art records by a large margin on all seven scores on Remote Sensing Image Caption Dataset. The experiment result indicates our model is effective on remote sensing image captioning and achieves the new state-of-the-art result.

1. Introduction

Recently, there have been extensive study and analysis on remote sensing images of high resolution, and deep neural networks achieved satisfactory results in scene classification and object detection. Despite the successful application of deep neural networks in the task aforementioned, it should be pointed out that the existing research usually attaches more importance to the image feature of the remote sensing images. Limited work has been done in capturing the semantic meaning and correlations of different objects in remote sensing images, which is also a key issue for the machine to understand the images better.

We focus on the remote sensing image captioning task in this paper, allowing for generating the semantic descriptions by teaching a machine to comprehend the content of the image. In the past years, a little effort has been devoted to the text descriptions of remote sensing images. Liu et al. [1] applied the semantic mining method in the remote sensing image retrieval model. Zhu et al. [2] proposed SAL-LDA (Semantic Allocation Level-Latent Dirichlet allocation), which is a new strategy based on the semantic distribution. Yang et al. [3] modeled underlying relations between features and the context in the given image with the Conditional Random Field (CRF) theory. Wang and Zhou [4] explored a strategy using semantic information to retrieve remote sensing images in the dataset. Chen et al. [5] proposed to use the graph model theory to extract object semantic relations. Li [6] present an object detection-based semantic model by making comparisons between different themes in different categories on the semantic level. There are some limitations to these approaches to fully utilize the image contents and generate the natural fluent text descriptions. Deep neural networks with the encoder–decoder framework have been proven successful in solving natural image captioning tasks. The theory of Reinforcement Learning [7] is also gradually being applied to the image captioning.

Inspired from work on natural image captioning, some research works have been published for remote sensing image captioning. Qu et al. [8] employed an RNN as the decoder of a multi-modal model to describe the content of remote sensing images. Shi and Zou [9] proposed a remote sensing image captioning model, which first leverages a convolutional neural network (CNN). Lu et al. [10] exposed a dataset, Remote Sensing Image Captioning Dataset (RSICD), and performed several experiments on it with different methods to validate their performance, including multi-modal models and attention-based models. Wang et al. [11] measured the representation of images and captions by embedding them to the same semantic space. Zhang et al.

* Corresponding authors.
E-mail addresses: liubing@cumt.edu.cn (B. Liu), zc@cumt.edu.cn (M. Liu).

https://doi.org/10.1016/j.knosys.2020.105920
0950-7051/

[12] introduced the attribute attention mechanism in their model, which can better capture the correspondence between the semantic information and the specific object in the remote sensing image.

However, there are still some limitations on these approaches:

1. Based on the transfer learning theory, the CNN adopted by the models above are pre-trained on the ImageNet dataset to enhance the image feature extraction ability. However, compared with natural images in ImageNet dataset, most remote sensing images lack some salient objects that can attract our attention. Due to the unique "view of God" of remote sensing images, many items are equally important and need to be taken into consideration simultaneously. It may not perform well to directly apply the CNN pre-trained on ImageNet dataset as the encoder of remote sensing image captioning due to the gap between the remote sensing images and natural images. On the other hand, ImageNet dataset is designed for the image classification task. Compared between image classification and image captioning, it is more important for image captioning models to be able to encode complete image information as well as the correlations between the objects in the image.
2. The RNN precludes parallelization within training examples due to its inherently sequential nature [13], making it difficult to train. The Transformer [13], constructed completely using the attention mechanism to model the sequence dependency, thus removing recurrence, has been proven superior to RNN in both feature extraction ability and training efficiency. Zhu et al. [14] utilized the Transformer as the decoder of the natural image captioning model, but few works have been investigated on remote sensing image captioning.
3. The Reinforcement Learning (RL) has achieved great success in natural image captioning by solving the gap between training loss and evaluation metrics. However, how to further enhance the performance of remote sensing image captioning via RL is still under-explored.

The main purpose of this paper is to overcome the above mentioned limitations, and our main contributions and motivations are listed as follows:

1. Introducing VAE to regularize the shared encoder and extract image features more effectively by reconstructing input images. A VAE [15] can be regarded as an autoencoder whose training is regularized to avoid overfitting and ensure that the latent space has good properties to generate some new data. Adding a VAE branch can relieve the overfitting problem caused by the lack of remote sensing images. Furthermore, the reconstruction process in VAE can help CNN pre-trained on ImageNet encodes better representation for the given remote sensing image.
2. Improving the performance of image caption significantly by virtue of low-level and high-level image features simultaneously. Zeiler and Fergus [16] visualized the different layers of CNN and found that high-level features contain more semantic information, while low-level features focus more on details. It will be more effective to take advantage of both high and low features so that they can complement each other.
3. Enhancing the final text description quality by adding self-attention to spatial features. Vaswani et al. [13] introduced the self-attention mechanism and calculated it with vectors named Query, Key, and Value. Query and Key are used to construct the relationships. Value summarizes all relations within and concludes the output containing relations between input and all other words. Since the high level features focus more on semantic information, different spatial features are semantic representations for different areas in the image. Self-attention mechanism can be utilized to achieve better regional semantic representation by extracting more information from more related fields in the image.

Our paper is organized as follows: In Section 2, we introduce the related works on natural image captioning and remote sensing image captioning. In Section 3, we explain the methods we proposed for remote sensing image captioning. In Section 4, we report our experimental settings and analyze the experiment results. In Section 5, we make the final conclusion of our paper.

2. Related work

There have been extensive studies and analyses on remote sensing images of high resolution [17,18]. The task of remote sensing image usually stems from the natural image, e.g., image captioning task. There are three different categories of methods for natural image captioning task: retrieval-based methods, template-based methods, and encoder–decoder based methods. The retrieval-based methods [19–21] firstly search in the dataset the image most similar to the input image and obtain the corresponding annotation as the template sentence. The result text description of the given image is then generated using the template sentence. The template-based methods proposed in [22–24] consists of three parts: the predefined sentences with blanks in it, the object detection model, and the relation model. The relation model is used to describe the relations of the objects detected by the object detection model. Then, the blanks in the predefined sentences are properly filled with these objects, such as entities, attributes, and behaviors.

The encoder–decoder based methods are inspired from the sequence-to-sequence neural network proposed in Neural Machine Translation. In order to enable CNN and RNN interact with each other for more information during the text description generation, Mao et al. [25] introduced the multimodal layer in the multimodal recurrent neural work (m-RNN). Vinyals et al. [26] proposed to employ LSTM as the sentence decoder to replace the traditional RNN to alleviate the vanishing gradient problem. Xu et al. [27] firstly proposed to apply attention mechanisms into the image captioning model where spatial features are extracted and input to the LSTM to decode the target sentence. Wu et al. [28] found that high-level semantic features captured by deeper layers of CNN play a more important role than image features maps in boosting the model performance. The importance of semantic level feature in the image captioning task is also investigated in [29–31]. Lu et al. [32] further utilized the attention mechanism to teach model knowing whether it is appropriate to focus on image features or the next word generation. Instead of using traditional grid-like features from a CNN, Anderson et al. [33] proposed to use bottom-up and top-down attention based on a ResNet within a Faster R-CNN framework to extract region-specific features. Traditional image captioning can only describe the object appearing during the training stage. Novel object captioning [34,35] attempted to describe unknown objects using an approach similar to template-based methods. They first generated the caption template with the known objects as the placeholder, and then replaced them with the detected new objects in the second stage. Rennie et al. [7] presented an RL-based self-critical sequence training (SCST) method to improve the performance of image captioning considerably. Zhu et al. [14] demonstrated that the Transformer, proposed by Vaswani et al. [13], also outperforms LSTM in natural image captioning. Image captioning is a multi-model application. Recently, there also

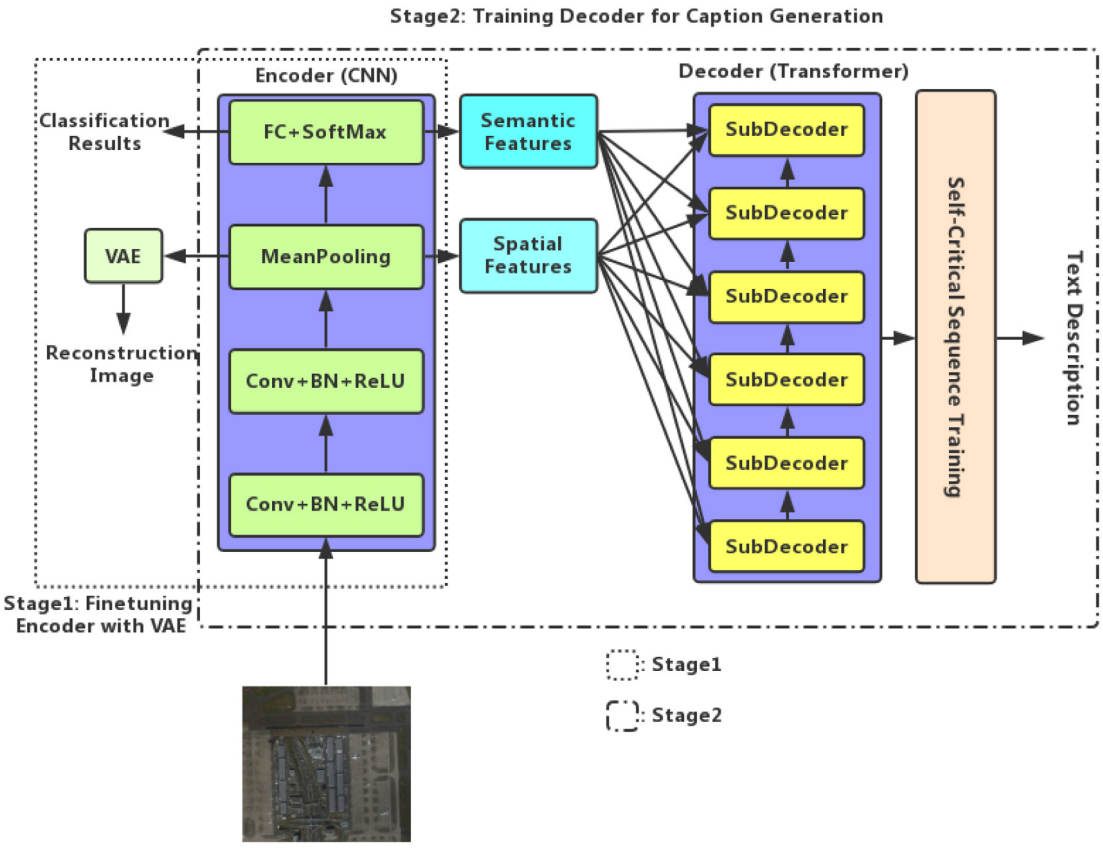

Fig. 1. The structure of VRTMM. "Encoder" is a pre-trained convolutional neural network such as Vgg16. "Decoder" is the transformer.

have been other similar vision and language tasks proposed. Das et al. [36] introduced a new dataset on visual dialog. They also developed a novel two-person chat data-collection protocol and tested a family of models based on encoder–decoder architecture on it. Zhu et al. [37] investigated video question answering and explored the approach similar to template-based methods for a finer understanding of video content. A dual-channel ranking loss was introduced to answer multiple-choice questions. The EmbodiedQA is a task of training an embodied agent to answer textual questions by interacting with a simulated environment to gather necessary visual information. Wu et al. [38] made the agent more generalized to the new scenes by randomly placing some markers when exploring the new environment. This simple baseline can be trained end-to-end and achieved competitive results to the state-of-the-art.

Researches on remote sensing image captioning are based on the natural image captioning. Qu et al. [8] firstly applied the encoder–decoder framework to remote sensing image captioning, which demonstrated that the encoder–decoder framework is also applicable to remote sensing images. Shi and Zou [9] adopted an object detection based method to replace the LSTM for the sequence generation due to limited training data. In order to leverage the potential of deep neural networks, Lu et al. [10] published Remote Sensing Image Captioning Dataset (RSICD). Wang et al. [11] regarded the caption generation task as a latent semantic embedding task, which can be solved via matrix learning. Zhang et al. [12] introduced the attribute attention mechanism by utilizing the features from SoftMax layer of the CNN so that the detailed correspondence between different parts of images and words can be obtained to improve the model's robust performance.

3. VRTMM

In Section 3.1, we mainly introduce the framework of our model. In Section 3.2, the details of the encoder in our model is presented. In Section 3.3, we first briefly describe the overall architecture of the Transformer and then introduce the modifications we make to adapt the Transformer to the task at hand. In Section 3.4, we introduce the training details during the finetuning procedure.

3.1. Model architecture

Fig. 1 illustrates the overall model architecture. Our model consists of the encoder and the decoder. Due to the mismatch between the text and image information, when the CNN encoder and Transformer decoder are jointly trained, the noise in the initial gradients from the Transformer into the image model corrupt the CNN and will never recover [26]. For this reason, it is not recommended to train the encoder and the decoder simultaneously. There are two sequential stages involved in our framework: encoder finetuning stage and decoder training stage. During the first stage, the encoder is finetuned on a remote sensing image scene classification dataset jointly with the Variational Autoencoder. In the second stage, the decoder is optimized on RSICD dataset using features extracted by the encoder and Reinforcement Learning theory. The parameters of the encoder are fixed for 4/5 of the whole steps in the second stage and finetuned for the remaining steps to avoid the problem of encoder corruption mentioned in [26]. Details of the encoder and the decoder will be described in Sections 3.2 and 3.3.

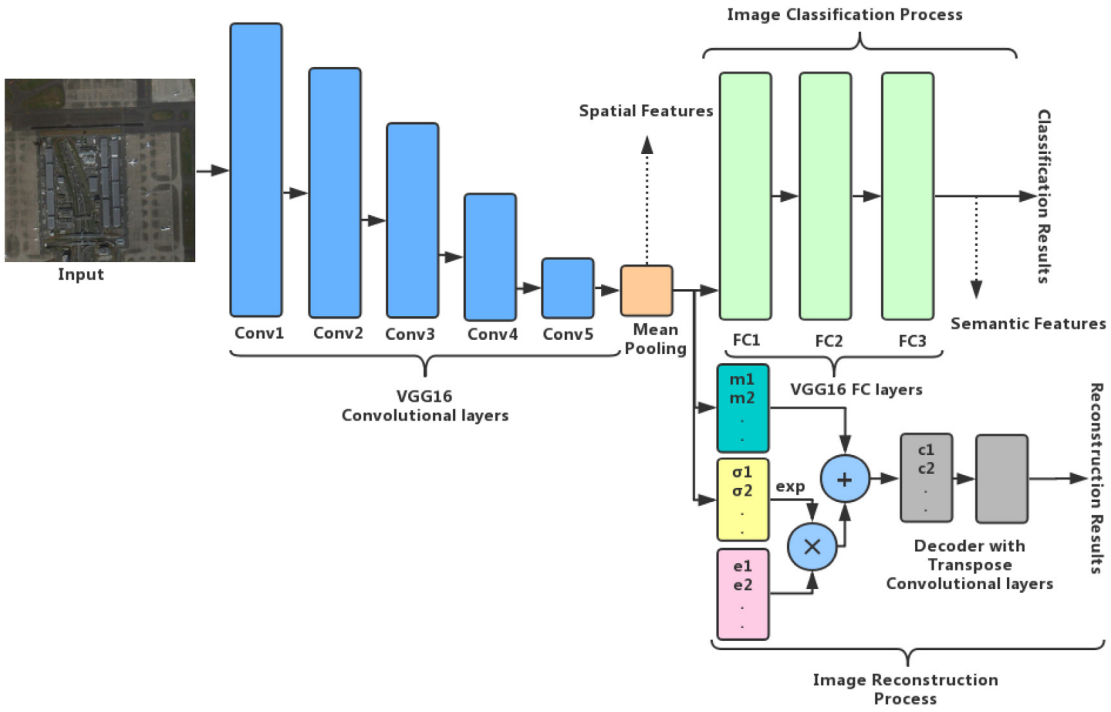

Fig. 2. The detailed structure for the encoder finetuning stage. We take the VGG16 as an example.

3.2. Multi-task encoder finetuning

Fig. 2 illustrates the detailed structure for the encoder finetuning stage. In the image captioning, since the encoder takes an image as the input, the CNN and its different variants are more appropriate for this task. So we replace the encoder part in the Transformer with CNN. In practice, the size of remote sensing datasets for different tasks is usually much smaller than the natural image datasets. The limitation on training data may lead to more severe overfitting problems in the deep neural network during the training procedure. In order to make our model better generalize to the small remote sensing image dataset, we apply multi-task learning to our model. During the experiment, we finetune the CNN on the large scale image classification dataset for remote sensing images. We introduce Variational Autoencoder (VAE) to reconstruct the input images jointly with remote sensing image classification. That is, the model needs to classify the image and reconstruct the input image simultaneously, i.e., they are jointly-trained during the finetuning process of the CNN. Performing a variational inference procedure on this model leads to joint regularization between the VAE and the convolutional neural network classifier, which contributes to avoiding overfitting and poor generalization. In the encoder part of the convolutional neural network, it downsamples the input image with convolutional layers and pooling layers. Then, the input images are reconstructed in the decoder by means of transpose convolution.

After finetuning the CNN, we store the parameters of the CNN and the VAE branch in the encoder finetuning stage is deleted during the decoder training stage. Visualizations of CNN have shown that different levels of information is captured in different layers of the network. High-level features of CNN contain the semantic information of the image while neglecting the details compared to the low-level features. In the experiment, we make use of both semantic information and spatial information. For better controlling different sizes of the spatial feature map, we apply the adaptive pooling before full connected layers. The high-level semantic information is obtained through the full connected layers with additional softmax operation, as Zhang et al. [12] have demonstrated that the output of the softmax layer is superior to that of the fully connected layer. The formula can be written as:

$$f = CNN(I) \tag{1}$$

$$semantic = SoftMax(L(f)) \in R^d \tag{2}$$

$$spatial = MeanPooling(f) = \{V_1, \ldots, V_{n\times n}\}, V_i \in R^d \tag{3}$$

where I is the input remote sensing image, $CNN(\cdot)$ stands for the output final convolutional layer of a CNN, $L(\cdot)$ is the linear layer, *semantic* and *spatial* represent the semantic information and the spatial information respectively, d is the target embedding dimension we define, n $\times$ n is the number of regions in the image and V_i represents a single part of the image. In practice, n is empirically set to 7.

Inspired by the fact that the Transformer is able to model the dependency among words in the sequence, we consider every single part of the spatial features as a single word and apply the self-attention operation, so that different parts of the spatial features can interact with each other, thus integrating more context information. We illustrate the structure of our model during the finetuning process in Fig. 2. More details of the CNN finetuning process are introduced in Section 3.4.

3.3. Decoder training part

The detailed structure for the decoder training stage is illustrated in Fig. 3. Bahdanau et al. [39] firstly applied the attention mechanism into the sequence-to-sequence model and boosted the performance on the Machine Translation task. A number of state-of-the-art models on natural image captioning stem from their work. In most of the image captioning models combining attention mechanisms and encoder–decoder framework [26–28], researchers replace the RNN-like networks in the encoder with CNN for better image features, while the decoder part is left unchanged. Other than previous mainstream methods, we built the decoder based on the Transformer [13], which is better than LSTM in both feature extraction ability and training efficiency.

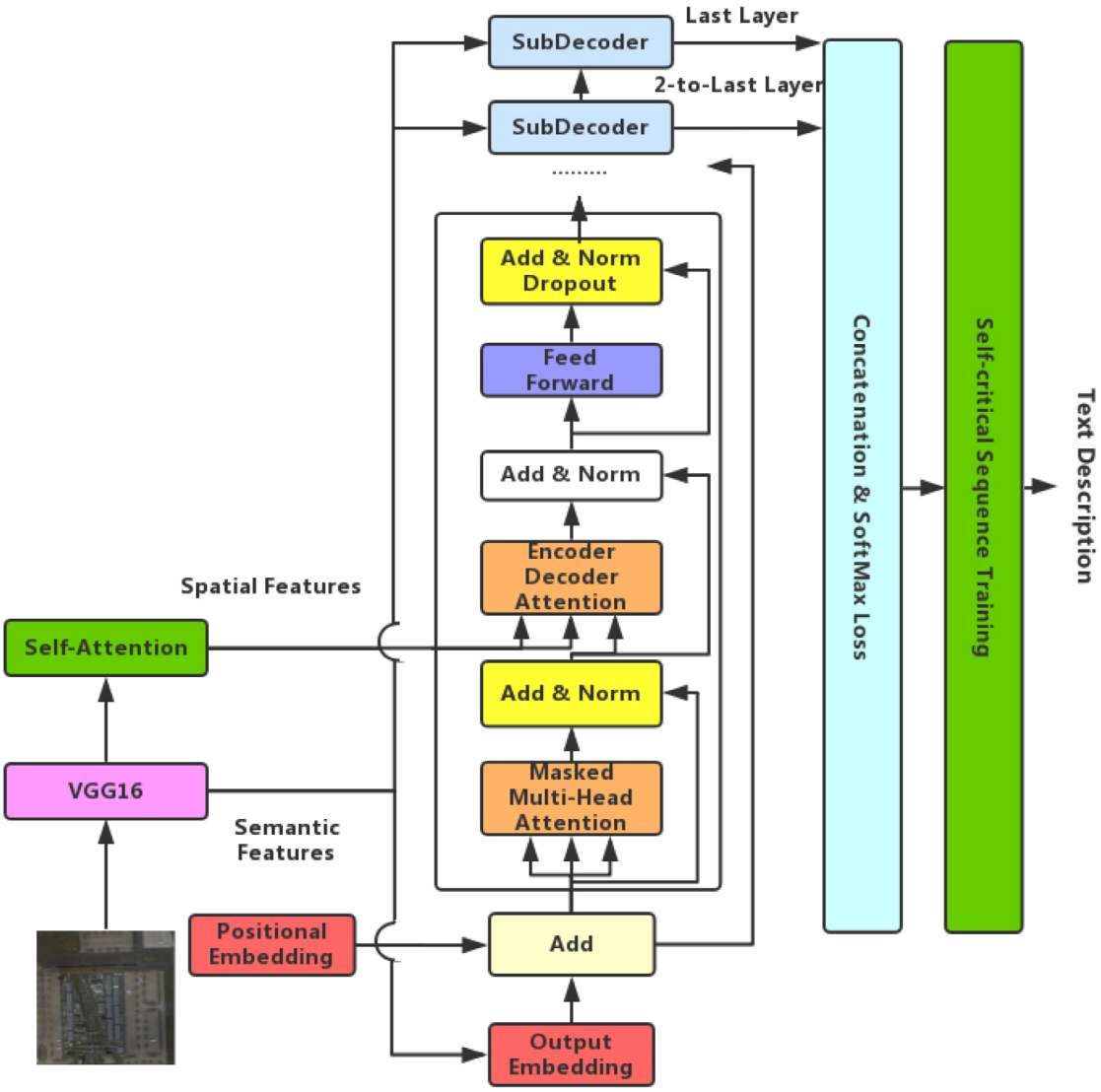

Fig. 3. The detailed structure for the decoder training stage. We only illustrate the detail of the first decoder, and the others are all the same.

Although Zhu et al. [14] have demonstrated the advantages of the Transformer over LSTM in image captioning, we make two modifications to make it accommodate to the remote sensing image dataset:

1. The original Transformer passed the encoded text features to the second sub-layer of each encoder to make the generated sentence and encoded information interact with each other. As shown in Fig. 3, the spatial features are processed in the way the original Transformer does. We regard the vector of every single part in the image as the representation of a word in the source sentence for Machine Translation. At the same time, we regard the semantic features extracted by the CNN as the initial semantic state of the target sentences. At each time step, we first consider the semantic feature as the representation of the first word in the sentence. After going through the first decoder, only the intermediate hidden states of the next word is passed to the next sub-layer, and the semantic feature is only used to complement the information helpful for the next word generation. In this way, the generated words in the sentence collect information from the most relevant part of the spatial features, the semantic feature, and the previous word history with the multi-head attention mechanism. Since we have removed the intermediate hidden start of word zero in the sentence, i.e., the representation of the semantic feature of the image, just like operations the Transformer has performed on the encoded features, we again pass the semantic information to the decoder and make it the initial state of the sentence. The remaining operations in each decoder are the same as the first decoder. Intuitively, this ensures the integrity of the semantic feature, preventing the previous decoder from only allowing a small part relevant feature passed to the next decoder.
2. Devlin et al. [40] compared different feature-based approaches by extracting the feature from one or more layers in the Transformer without fine-tuning. The experiment result shows that the concatenation of the last four hidden layer representation performs better than the weighted sum of last four hidden layers. As is shown in the right part of Fig. 3, we concatenate the representations of the last two layers. To alleviate the overfitting issue, we also add the dropout layer and layer normalization [41]. The final representation of the feature is then projected to the target dimension with a linear layer.

3.4. Two-stage training

3.4.1. Encoder finetuning stage

We first finetune the CNN with VAE branch on the large scale remote sensing image classification dataset. The CNN has been pre-trained on the ImageNet dataset. Starting from the adaptive pooling layer of the CNN, we add the VAE branch. In the VAE decoder part, we first reduce the input to a low dimensional space (half to represent mean and half to represent std). Then we draw a sample from the Gaussian distribution with the given mean and std. The sample is decoded to the input image dimensions following the mirroring architecture of the encoder, i.e., the modules before the adaptive pooling layer.

Our loss function during the finetuning process is formulated as follows:

$$\mathbf{L} = \mathbf{L}_{Softmax} + 0.1 * \mathbf{L}_{L2} + 0.1 * \mathbf{L}_{KL} \tag{4}$$

In Eq. (4), $\mathbf{L}_{Softmax}$ is a Softmax activation plus a Cross-Entropy loss. The CNN will predict the probability over the whole classes for the given image:

$$\mathbf{L}_{Softmax} = -log\left(\frac{e^{s_p}}{\sum_j^C e^{s_j}}\right) \tag{5}$$

where s_p is the score CNN predicts for the positive class.

The second term $\mathbf{L}_{L2}$ in Eq. (4) is the L2 loss on the VAE branch image reconstruction result I_{pred} to match the given input image I_{input}:

$$\mathbf{L}_{\mathrm{L2}} = \left\| I_{input} - I_{pred} \right\|_2^2 \tag{6}$$

The last term $\mathbf{L}_{KL}$ is the KL divergence, which is a standard VAE penalty term [15,42] describing differences between the estimated normal distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ and a prior distribution $\mathcal{N}\left(0, 1\right)$. The closed-form representation can be written as:

$$\mathbf{L}_{\mathrm{KL}} = \frac{1}{N}\sum \mu^2 + \sigma^2 - \log\sigma^2 - 1 \tag{7}$$

We adopt a Vgg16 [43] pre-trained on the ImageNet dataset as the encoder and the size of spatial features in the CNN to 7×7. Considering the different angles of remote sensing images, We apply a random rotation operation on input images during the training process for data augmentation. Adam [44] is chosen as the optimizer. Both the initial learning rate and the coefficient of L2 regularization are set to 0.0001. The learning rate is multiplied by 0.7 when the loss on validation set does not decrease after 2 epochs. Early-stopping is applied if there is no promotion on the validation accuracy for 5 epochs.

3.4.2. Decoder training stage

In practice, the number of layers of the Transformer can be set by balancing the efficiency and the accuracy. The spatial and semantic features extracted by the CNN are passed to the Transformer using the method mentioned above. The loss function without the reinforcement learning can be written as:

$$\begin{aligned} l &= -\frac{1}{N}\sum_{n=1}^{N} \log p\left(W^{(n)} | semantic^{(n)}, \mathrm{spatial}^{(n)}\right) + \lambda_\theta \cdot \|\theta\|_2^2 \\ &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{L^{(n)}} \log p_i\left(W_i^{(n)}\right) + \lambda_\theta \cdot \|\theta\|_2^2 \end{aligned} \tag{8}$$

where N is the total number of images in the dataset, $L^{(n)}$ is the length of nth training sample's caption, $\lambda_\theta \cdot \|\theta\|_2^2$ represents a regularization term. The whole model is trained end-to-end, and the flow-chart of our model can be seen in Fig. 3.

The number of decoders is set to 6. We use the same hyper-parameter settings in [13] for both the model and the optimizer. The technique present in [26] is adopted, that is, finetune the parameters of CNN after training about the 4/5 of the whole steps to address the mismatch between text and image information.

After optimizing the parameters of the CNN and the Transformer, SCST is used to further improve the quality of the sentence. Combined with the RL method, both word and image features in our model can be seen as the external "environment", and the Transformer can be viewed as an "agent", which interacts with the external "environment". The "action" standing for the prediction of the next word is determined by the policy p_θ that is defined by the parameters of the Transformer, θ. As in Reference [45], after each action, the agent (the Transformer) updates its internal "state" (parameters of the Transformer). The "reward", for instance, using the CIDEr score of the generated sentence, denoted by γ in Eq. (9), is computed by the evaluation metric by making comparisons between the generated sentence and the ground-truth sentence upon the prediction of the end-of-sequence (EOS) token. The negative expected reward minimized by our model can be written as:

$$L(\theta) = -\mathbb{E}_{W^s \sim p_\theta}\left[r\left(W^s\right)\right] \tag{9}$$

where $W^s = \left(W_1^s, \ldots, W_T^s\right)$ and W_t^s is the word sampled from the model at the time step t. Since we take advantage of the method introduced in [7], e.t., self-critical sequence training (SCST), during the Reinforcement Learning period, we refer the reader to [7] for details of SCST and final evaluation metric we used in our model is the CIDEr score suggest in [7].

4. Experiments

4.1. Dataset

The finetuning process for the encoder is performed on NWPU-RESISC45 dataset [46]. NWPU-RESISC45 dataset is a public available dataset on the REmote Sensing Image Scene Classification (RESISC) task. It contains 31,500 images and 45 scene classes. For each class, there are 700 images in it. We conduct the image captioning experiment on RSICD dataset [10], the largest remote sensing image captioning dataset so far. RSICD dataset includes 10,921 remote sensing images with 224×224 sizes in different resolutions. Some typical images selected from the RSICD dataset can be seen in Fig. 4.

4.2. Metrics and baselines

Researches have proposed several evaluation metrics to judge whether a description generated by a machine is good or not. The most commonly used metric in the image captioning task is the BLEU score [47]. It calculates the precision of word n-grams between the ground-truth sentence and the generated sentence. ROUGE [48] also gets the evaluation metric by comparing the reference and the generated sentence, but it focuses on the recall. METEOR [49] uses the matching degree including synonym matching to calculate the harmonic mean F-measure and returns the highest score to indicate the quality. More recently, the organizers of the MS COCO Captioning challenge propose CIDEr [50] similar to the BLEU score, which additionally uses a Term Frequency–Inverse Document Frequency (TF-IDF) weighting in n-gram so that the influence of high-frequency words and non-keywords can be reduced.

In order to make the comparison between our proposed model and other models, we perform our experiment with several available models, including the multimodal method proposed in [8],

(a) (b)

Fig. 4. (a) Images selected form RSICD; (b) Five captions of the image. (1) An airport with dark brown and light brown ground in it. (2) Some white planes in the airport while with some dark buildings besides. (3) Some sparse light green meadow inside while with some dark brown ground besides. (4) Some square areas divide into black lines inside. (5) Some planes are parked in an airport dispersedly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Results of the vanilla Vgg16 and Vgg16 plus VAE on the NWPU-RESISC45 Dataset.

	vanilla-VGG16	VAE-VGG16
Classification Acc.	0.9402	**0.9506**

the model using basic attention mechanisms in [10], CSMLF introduced in [11], and the attribute attention model proposed in [12]. Among all the models used for comparison with our model, the attribute attention model achieved the previous state-of-the-art results on RSICD. Further analysis and comparison of generated captions between the attribute attention model and VRTMM will be discussed in Section 4.3.1.

4.3. Experiment settings and results

We adopt a pre-trained VGG16 as our encoder. In order to achieve a good balance between the Softmax loss term and the VAE loss term in Eq. (5), we empirically set the hyper-parameter weight of each term to 0.1. All remote sensing images are cropped into 224×224 before being input to the model. In practice, all the experiments, including the finetuning encoder process and the image captioning training process, are done on a server with 1 Nvidia Tesla P100 graphics card under Ubuntu 18.04. In order to get better captions, the beam search algorithm is applied during the inference period. Tables 1 and 2 report our experiment results on the NWPU-RESISC45 dataset the RSICD dataset respectively.

4.3.1. Results analysis

The experimental results from Table 1 show that adding the VAE branch after the adaptive pooling layer can improve the classification accuracy about 0.01, which validates the effectiveness of the VAE branch. From Table 2, we can conclude deep neural networks (with or without attention mechanisms) perform better than traditional models. The attribute attention model [12] achieves the state-of-the-art results with the newly designed attention mechanisms.

Our model with SCST generates much better results than the attribute attention model. The improvement on BLEU-1∼4, METEOR, and ROUGH is nearly 0.04. CIDEr metric of our model is 2.7930, more than 0.4 higher than the previous state-of-the-art result 2.3563.

From the results shown in Table 2, our model outperforms all the baselines by a large margin, including the previous state-of-the-art attribute attention model, which has validated the effectiveness of our model. The captions generated by our model and the attribute attention model for the same images are illustrated in Fig. 5 for comparison. We can see from Fig. 5 that for most of the images selected from RSICD dataset, VRTMM is able to generate captions of higher quality. We can get text descriptions of relatively more complex grammar structures, such as Fig. 5g. The model can also describe some important attributes of the object in the scene, including the amount (like *four* in Fig. 5f), the color (like *green* in Fig. 5b), the shapes (like *circle* in Fig. 5h, *curved* in Fig. 5i).

For some of the images, the text descriptions of VRTMM are the same (Fig. 5f) or slightly better (*around* more accurate than *near* in Fig. 5e) than the attribute attention model. In some cases, both of them achieved appropriate results but focusing on different objects in the scene (*trees* and *road* in Fig. 5d). Meanwhile, VRTMM is able to generate captions of significantly higher quality in these aspects:

1. VRTMM avoids describing the wrong color, such as *yellow* in Fig. 5a.
2. VRTMM gives more detailed descriptions for the same image. In Fig. 5b and Fig. 5g, VRTMM successfully recognized

Table 2
Results of the baseline and VRTMM on the RSICD Dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH	CIDEr
CSMLF [11]	0.5759	0.3859	0.2832	0.2217	0.2128	0.4455	0.5297
Multimodal [8]	0.6378	0.4756	0.4004	0.3006	0.2905	0.5333	2.2536
Attention [10]	0.7336	0.6129	0.5190	0.4402	0.3549	0.6419	2.2486
AttrAttention [12]	0.7571	0.6336	0.5385	0.4612	0.3513	0.6458	2.3563
VRTMM+SCST	**0.7934**	**0.6794**	**0.5878**	**0.5113**	**0.3726**	**0.6797**	**2.7930**

Fig. 5. The comparison of captions generated by the attribute attention model (denoted by "A") and VRTMM (denote by "V") . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

buildings in the images while attribute attention model failed.

3. The attribute attention model describes objects not appearing in the image. There are no buildings in Fig. 5h, but

Fig. 6. Some examples of failures generated by VRTMM on RSICD dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the word "buildings" is in the generated sentences of the attribute attention model.

4. The attribute attention model misrecognizes some objects in the image. For example, the commercial area is misrecognized as the residential area in Fig. 5c. In Fig. 5h, the attribute attention model mistook a circle center building for a church.
5. The attribute attention model completely fails to describe the image while VRTMM does, such as Fig. 5i.

Despite the high quality of the captions for most of the images, there are also some examples of failures illustrated in Fig. 6, which is discussed below:

1. Some objects in the generated caption are not in the image. There are no trees in Fig. 6a-b, but the word "trees" is in the final descriptions. It can be caused by the high frequency of some words in the training data. For example, trees often appear in the remote sensing images, so VRTMM tends to generate sentences with "trees" whether or not there are trees in the image.
2. Some objects in the image are not in the generated caption. In Fig. 6c, VRTMM fails to recognize the building in the left part of the image. It remains a problem how to depict minor objects appearing in the edges of the image.
3. Misrecognition. Fig. 6d-h are all examples of misrecognition. Many factors contribute to this problem, such as

Table 3
Ablation study of VRTMM on the RSICD dataset.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH	CIDEr
AttrAttention [10]	0.7571	0.6336	0.5385	0.4612	0.3513	0.6458	2.3563
VRTMM	**0.7813**	**0.6721**	**0.5645**	**0.5123**	**0.3737**	**0.6713**	**2.7150**
VRTMM-VAE	0.7610	0.6383	0.5431	0.4671	0.3622	0.6499	2.4300
VRTMM-VAE-SA	0.7547	0.6342	0.5410	0.4661	0.3588	0.6475	2.4240

sharing the same color (Fig. 6d-e), sharing the similar appearance (Fig. 6f-h). It is still a challenge left open by the existing work on remote sensing images, which is called the small interclass dissimilarity problem [46]. Enabling the model reason the appropriate result with the external knowledge and common sense may help get over around this problem. For example, people can easily acknowledge that in Fig. 6e, the forests usually do not have such a regular shape.

4. Counting errors. Two baseball fields are in Fig. 6i, but VRTMM mistakes 2 for 4. More work needs to be done to make the machine describe the image well and do counting correctly at the same time.

4.3.2. Ablation study

In this section, we investigate the effects of different methods we use in our model. We denote '-' in Table 3 as the exclusion of the corresponding part in the model. In Table 3, we can see that our model still outperforms the attribute attention by a large margin without SCST, verifying the effectiveness of VRTMM. At the same time, when excluding the variational autoencoder finetuning process from our model, all seven scores drop a lot, demonstrating the importance of variational autoencoder finetuning. Meanwhile, our model still outperforms the previous SOTA. We denote 'SA' in Table 3 as the self-attention mechanism applied to the spatial features. The last line of Table 3 records the result only using the modified Transformer with spatial and semantic features directly passed to it. Our model still outperforms the attribute attention model on six metrics, which indicates the modified Transformer has a better ability comparing with LSTM in dependency modeling and feature extraction for sequences.

5. Conclusions

In this paper, we propose a new model for remote sensing image captioning based on the variational autoencoder and the encoder–decoder architecture. We first finetune the CNN with the variational autoencoder branch on the remote sensing image scene classification dataset. The finetuned CNN is then employed to extract both semantic and spatial features of the images. After the self-attention operation on spatial features, both semantic and spatial features are passed to the modified Transformer to generate the final text descriptions of the image. Our model achieves the state-of-the-art result on RSICD dataset. In the future, we will try to utilize VAE to generate fake data to further alleviate overfitting in remote sensing image captioning.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xiangqing Shen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Bing Liu:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Yong Zhou:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Jiaqi Zhao:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Mingming Liu:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61801198), Natural Science Foundation of Jiangsu Province, China (BK20180174), Fundamental Research Funds for the Central Universities, China (2017XKQY082), National Natural Science Foundation of China (61806206), Natural Science Foundation of Jiangsu Province, China (BK20180639). The authors would like to thank the anonymous reviewers and the associate editor for their valuable comments.

References

[1] T. Liu, P. Li, L. Zhang, X. Chen, A remote sensing image retrieval model based on semantic mining, Geomatics Inf. Sci. Wuhan Univ. 34 (2009) 684–687, http://dx.doi.org/10.1042/BSR20080061.
[2] Q.Q. Zhu, Y.F. Zhong, L.P. Zhang, Multi-feature probability topic scene classifier for high spatial resolution remote sensing imagery, in: 2014 IEEE International Geoscience and Remote Sensing Symposium, Igarss, 2014, http://dx.doi.org/10.1109/Igarss.2014.6947071.
[3] J. Yang, Z. Jiang, Q. Zhou, H. Zhang, J. Shi, Remote sensing image semantic labeling based on conditional random field, 36 (2015) 3069–3081, http://dx.doi.org/10.7527/S1000-6893.2014.0356.
[4] J. Wang, H. Zhou, Research on key technologies of remote sensing image data retrieval based on semantics, Comput. Digit. Eng. 40 (2012) 48–50.
[5] K. Chen, Z. Zhou, J. Guo, D. Zhang, X. Sun, Semantic scene understanding oriented high resolution remote sensing image change information analysis, in: Proceedings of the Annual Conference on High Resolution Earth Observation, Beijing, China, 2013, pp. 1–12.
[6] Y. Li, Target Detection Method of High Resolution Remote Sensing Image Based on Semantic Model, Graduate University of Chinese Academy of Sciences, Beijing, China, 2012.
[7] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: 30th IEEE Conference on Computer Vision and Pattern Recognition, Cvpr 2017, 2017, pp. 1179–1195, http://dx.doi.org/10.1109/Cvpr.2017.131.
[8] B. Qu, X.L. Li, D.C. Tao, X.Q. Lu, Deep semantic understanding of high resolution remote sensing image, in: 2016 International Conference on Computer, Information and Telecommunication Systems, Cits, 2016, pp. 124–128.

[9] Z.W. Shi, Z.X. Zou, Can a machine generate humanlike language descriptions for a remote sensing image?, IEEE Trans. Geosci. Remote Sens. 55 (6) (2017) 3623–3634, http://dx.doi.org/10.1109/Tgrs.2017.2677464.

[10] X.X. Lu, B.Q. Wang, X.T. Zheng, X.L. Li, Exploring models and data for remote sensing image caption generation, IEEE Trans. Geosci. Remote Sens. 56 (4) (2018) 2183–2195, http://dx.doi.org/10.1109/Tgrs.2017.2776321.

[11] B. Wang, X. Lu, X. Zheng, X. Li, Semantic descriptions of high-resolution remote sensing images, IEEE Geosci. Remote Sens. Lett. (2019) 1–5, http://dx.doi.org/10.1109/LGRS.2019.2893772.

[12] X.R. Zhang, X. Wang, X. Tang, H.Y. Zhou, C. Li, Description generation for remote sensing images using attribute attention mechanism, Remote Sens. 11 (6) (2019) http://dx.doi.org/10.3390/rs11060612.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 5998–6008, URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[14] X. Zhu, L. Li, J. Liu, H. Peng, X. Niu, Captioning transformer with stacked attention modules, Appl. Sci. 8 (5) (2018) 739, http://dx.doi.org/10.3390/app8050739.

[15] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Y. Bengio, Y. LeCun (Eds.), ICLR, 2014.

[16] M. Zeiler, R. Fergus, Visualizing and understanding convolutional neural networks, in: ECCV 2014, Part I, in: LNCS, vol. 8689, 2013, http://dx.doi.org/10.1007/978-3-319-10590-1_53.

[17] E. Basaeed, H. Bhaskar, M. Al-Mualla, Supervised remote sensing image segmentation using boosted convolutional neural networks, Knowl.-Based Syst. 99 (2016) 19–27, http://dx.doi.org/10.1016/j.knosys.2016.01.028.

[18] S.K. Mylonas, D.G. Stavrakoudis, J.B. Theocharis, GeneSIS: A GA-based fuzzy segmentation algorithm for remote sensing images, Knowl.-Based Syst. 54 (2013) 86–102, http://dx.doi.org/10.1016/j.knosys.2013.07.018.

[19] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, S. Lazebnik, Improving image-sentence embeddings using large weakly annotated photo collections, in: Computer Vision – ECCV 2014, Springer International Publishing, 2014, pp. 529–545, http://dx.doi.org/10.1007/978-3-319-10593-2_35.

[20] C. Sun, C. Gan, R. Nevatia, Automatic concept discovery from parallel text and visual corpora, in: 2015 IEEE International Conference on Computer Vision, Iccv, 2015, pp. 2596–2604, http://dx.doi.org/10.1109/Iccv.2015.298.

[21] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, J. Artificial Intelligence Res. 47 (2013) 853–899, http://dx.doi.org/10.1613/jair.3994.

[22] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: Computer Vision – ECCV 2010, Springer Berlin Heidelberg, 2010, pp. 15–29, http://dx.doi.org/10.1007/978-3-642-15561-1_2.

[23] S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, Y. Choi, Composing simple image descriptions using web-scale n-grams, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 220–228, URL: http://dl.acm.org/citation.cfm?id=2018936.2018962.

[24] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: Understanding and generating simple image descriptions, in: CVPR 2011, IEEE, 2011, http://dx.doi.org/10.1109/cvpr.2011.5995466.

[25] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Deep captioning with multimodal recurrent neural networks (m-RNN), in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1412.6632.

[26] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 652–663, http://dx.doi.org/10.1109/Tpami.2016.2587640.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 37, PMLR, Lille, France, 2015, pp. 2048–2057, URL: http://proceedings.mlr.press/v37/xuc15.html.

[28] Q. Wu, C.H. Shen, L.Q. Liu, A. Dick, A. van den Hengel, What value do explicit high level concepts have in vision to language problems? in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Cvpr, 2016, pp. 203–212, http://dx.doi.org/10.1109/Cvpr.2016.29.

[29] Q.Z. You, H.L. Jin, Z.W. Wang, C. Fang, J.B. Luo, Image captioning with semantic attention, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Cvpr, 2016, pp. 4651–4659, http://dx.doi.org/10.1109/Cvpr.2016.503.

[30] T. Yao, Y.W. Pan, Y.H. Li, Z.F. Qiu, T. Mei, Boosting image captioning with attributes, in: 2017 IEEE International Conference on Computer Vision, Iccv, 2017, pp. 4904–4912, http://dx.doi.org/10.1109/Iccv.2017.524.

[31] Q. Wu, C.H. Shen, P. Wang, A. Dick, A. van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1367–1381, http://dx.doi.org/10.1109/Tpami.2017.2708709.

[32] J.S. Lu, C.M. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: 30th IEEE Conference on Computer Vision and Pattern Recognition, Cvpr 2017, 2017, pp. 3242–3250, http://dx.doi.org/10.1109/Cvpr.2017.345.

[33] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018, http://dx.doi.org/10.1109/cvpr.2018.00636.

[34] Y. Wu, L. Zhu, L. Jiang, Y. Yang, Decoupled novel object captioner, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM Press, 2018, http://dx.doi.org/10.1145/3240508.3240640.

[35] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, Y. Yang, Cascaded revision network for novel object captioning, IEEE Trans. Circuits Syst. Video Technol. (2020) 1, http://dx.doi.org/10.1109/tcsvt.2020.2965966.

[36] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. Moura, D. Parikh, D. Batra, Visual Dialog, 2017, pp. 1080–1089, http://dx.doi.org/10.1109/CVPR.2017.121.

[37] L. Zhu, Z. Xu, Y. Yang, A.G. Hauptmann, Uncovering the temporal context for video question answering, Int. J. Comput. Vis. 124 (3) (2017) 409–421, http://dx.doi.org/10.1007/s11263-017-1033-7.

[38] Y. Wu, L. Jiang, Y. Yang, Revisiting EmbodiedQA: A simple baseline and beyond, IEEE Trans. Image Process. 29 (2020) 3984–3992, http://dx.doi.org/10.1109/tip.2020.2967584.

[39] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1409.0473.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL: https://www.aclweb.org/anthology/N19-1423.

[41] L.J. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, CoRR abs/1607.06450 URL: arXiv:1607.06450.

[42] C. Doersch, Tutorial on variational autoencoders, 2016, CoRR abs/1606.05908 URL: arXiv:1606.05908.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1409.1556.

[44] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1412.6980.

[45] M. Ranzato, S. Chopra, M. Auli, W. Zaremba, Sequence level training with recurrent neural networks, in: Y. Bengio, Y. LeCun (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings, 2016, URL: http://arxiv.org/abs/1511.06732.

[46] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proc. IEEE 105 (10) (2017) 1865–1883, http://dx.doi.org/10.1109/jproc.2017.2675998.

[47] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2001, http://dx.doi.org/10.3115/1073083.1073135.

[48] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81, URL: https://www.aclweb.org/anthology/W04-1013.

[49] S. Banerjee, A. Lavie, 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, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72, URL: https://www.aclweb.org/anthology/W05-0909.

[50] R. Vedantam, C. Zitnick, D. Parikh, CIDER: Consensus-based image description evaluation, 2015, pp. 4566–4575, http://dx.doi.org/10.1109/CVPR.2015.7299087.