Truncation Cross Entropy Loss for Remote Sensing Image Captioning

Xuelong Li, *Fellow, IEEE,* Xueting Zhang, Wei Huang, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Recently, remote sensing image captioning (RSIC) has drawn an increasing attention. In this field, the encoder-decoder based methods have become the mainstream due to their excellent performance. In encoder-decoder framework, the convolutional neural network (CNN) is utilized to encode a remote sensing image into a semantic feature vector, and a sequence model such as long short-term memory (LSTM) is subsequently adopted to generate a content-related caption based on the feature vector. During the traditional training stage, probability of the target word at each time step is forcibly optimized to 1 by Cross Entropy (CE) loss. However, because of the variability and ambiguity of possible image captions, the target word could be replaced by other words like its synonyms, and therefore such optimization strategy would result in over-fitting of the network. In this paper, we explore the over-fitting phenomenon in RSIC caused by CE loss, and correspondingly propose a new Truncation Cross Entropy (TCE) loss aiming to alleviate the over-fitting problem. In order to verify the effectiveness of the proposed approach, extensive comparison experiments are performed on three public remote sensing image captioning datasets, including UCM-captions, Sydney-captions and RSICD. The state-of-the-art result of Sydney-captions and RSICD and the competitive results of UCM-captions achieved by TCE loss demonstrate that the proposed method is beneficial to RSIC.**

***Index Terms*—remote sensing, image captioning, Truncation Cross Entropy loss, over-fitting**

I. INTRODUCTION

HIGH resolution remote sensing images, which have a wide range of applications [1]–[7], can be easily obtained nowadays due to the rapid development of remote sensing technology. How to efficiently mine the relationship between the visual features and semantic information hidden in remote sensing images has been widely concerned. Motivated by natural image captioning [8]–[14], remote sensing image captioning (RSIC) [15] has been explored in the past few years. RSIC, which combines computer vision with natural language processing [16], aims to let machine automatically generate human understandably descriptions from the given remote sensing images.

Benefiting from the technology of deep learning, the methods based on neural encoder-decoder architecture have gradually become a growing trend in the field of remote sensing image captioning. This kind of methods typically contain two stages: image understanding and caption generation. In image understanding phase, the convolutional neural network (CNN) is utilized as a feature extractor to encode the input remote sensing image into a high-level semantic feature vector of fixed dimension, which aggregates the visiual features and objects of the images. Subsequently, in caption generation phase, the logic and syntax relationship of these visual features and objects is decoded into a well-formed sentence by a sequence model, such as long short-term memory (LSTM) [17]. In this way, a lot of concise and meaningful sentences would be generated for the input remote sensing images, which is in line with the logic of human language and is helpful for human cognitive understanding.

(a) Many tennis courts arranged neatly with some plants surrounded.
(b) Many tennis courts arranged neatly with some trees surrounded.
(c) Many tennis courts arranged neatly with some plants surrounded.
(d) Some tennis courts arranged neatly with some plants surrounded.

Fig. 1: The sentences for an image is not unique, and some words can be replaced by other words like their synonyms.

Commonly, during the training stage, the encoder-decoder models are optimized by Cross Entropy (CE) loss [18], which regards the word prediction as a classification task [19]. At each time step, a word is predicted by the decoder network according to the following three components: the feature vector, the previous word and the generated sentence. It is worth mentioning that the probability of the target word at each time step is forcibly optimized to 1 by CE loss. Nevertheless, due to the variability and ambiguity of possible image captions, the target word may be strongly related to other non-target words like its synonyms. For example, as shown in Fig. 1, the given remote sensing image can be described by several sentences of (a)-(d). For (a) and (b), the object words of "plants" and "trees" could be replaced by each other. Similarly, for (c) and (d), "Many" and "Some" are alternative. To improve the output probability of target word, the unreasonable noise in the image would be learned. Thus, such 1-Probability optimization strategy would lead to the over-fitting of the network when the noise is learned.

In order to quantitatively explore the over-fitting caused by CE loss, in this paper, three different forms of CE loss are defined firstly according to the optimized objects, which are

This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U186420461773316, U1801262, and 61871470.

X. Li, X. Zhang, W. Huang and Q. Wang are with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: li@nwpu.edu.cn; zxt@mail.nwpu.edu.cn; hw2hwei@gmail.com; crabwq@gmail.com).

Q. Wang is the corresponding author.

Positive Cross Entropy (PCE) loss, Negative Cross Entropy (NCE) loss, and Positive-Negative Cross Entropy (PNCE) loss. In PCE loss, only the target word regarded as positive sample is adopted to optimize the model, and it is intrinsically equal to the normal CE loss used in RSIC. In contrast, when using NCE loss, only the non-target words treat as negative samples are employed for optimization of the model. In terms of PNCE loss, both the positive and negative samples (*i. e.*, all the words in the vocabulary) would be adopted to optimize the whole model.

Besides, to alleviate the over-fitting problem, a novel Truncation Cross Entropy (TCE) loss for remote sensing image caption generation is first presented in this paper. Different from the conventional CE loss, the proposed TCE loss is a piecewise loss which consists of two parts: a traditional CE loss and a truncation loss. More specifically, during the training stage, an upper limit would be set to $1-\gamma$ to decide which loss would be selected. When the output probability of the current word is lower than $1-\gamma$, the traditional CE loss would be employed to optimize the whole model. And when the output probability of the word exceeds the upper limit of $1-\gamma$, the truncation loss mechanism would be activated immediately, which means the target probability of this word will not be optimized higher than $1-\gamma$. In this way, a margin of γ, which is also quite valuable for the sentence, can be reserved for the non-target words in the vocabulary. Such a fuzzy mechanism can effectively alleviate the over-fitting phenomenon caused by CE loss and enhence the performance of the base model.

Overall, the main contributions of this paper can be summarized as follows:

1) To explore the over-fitting caused by Cross Entropy loss in RSIC, we present three types of loss, including PCE loss, NCE loss and PNCE loss according to the optimized words. And the corresponding comparison experiments are conducted with quantitative analysis.
2) A novel Truncation Cross Entropy (TCE) loss, which aims to alleviate the over-fitting problem caused by CE loss, is first proposed for caption generation of remote sensing images. By researving a probability margin for non-target words, the proposed TCE loss is helpful to generate more flexible and concise descriptions for remote sensing images and further enhence the generalization performance of the whole model.
3) Different CNNs combined with LSTM are applied on three public datasets to verify the effectiveness of the proposed TCE loss. The state-of-the-art result of Sydney-captions and RSICD and the competitive results of UCM-captions, which are achieved by TCE loss, demonstrate that the proposed approach is beneficial for RSIC.

The remainder of this article is organized as follows: Section II introduces the related works of remote sensing image captioning and Cross Entropy loss. In Section III, we describe the proposed Truncation Cross Entropy loss based method in detail. The experimental results and analysis on three datasets are introuced in Section IV. Finally, conclusions are provided in Section V.

II. Related Work

In this section, the relevant work of remote sensing image captioning and Cross Entropy loss will be briefly introduced.

A. Remote Sensing Image Captioning

Generally, the methods of RSIC can be roughly divided into three categories: retrival based methods, template based methods and encoder-decoder model based methods.

Retrival based methods depend on the retrival results and the matching degree. For example, Wang et al. [20] presented a collective semantic metric learning architecture to describe the image content more diversely. Based on the techonolegy of metric learning, this model maps the dimensions of input images representation and their corresponding captions representation to the same space. By computing the distances between the test image and all collective captions, the caption with the smallest distance would be picked up as the final descriptive sentence. However, since the generated captions are sereached from the existing database, the methods based on retrival are difficult to perform well when facing an input image which has low similarity with all the images in database.

Template based methods aim to define a fixed sentence template with several blanks reserved. Then after extracting features from input images, the detected objects, their attributes and the relationship among them would be correspondingly filled in the reserved blanks. For instance, a Fully Convolutional Networks (FCN) based method was proposed by Shi et al. [21] for better describing the remote sensing images in human languge, where fully convolutional networks [22] are employed to capture the elements with three different levels in a remote sensing image. In addition, a collective of triplets are used to guide the generation of the captions. Nevertheless, for templated based methods, one problem researchers have to consider is that the predefined template has greatly limited the flexibility of the generated captions, thus making the forms of the generated sentences quite rigid.

The encoder-decoder model based methods [23] are end-to-end which mainly contains two stages of encoding and decoding. The purpose of encoding is to represent the input image as a feature vector of fixed dimension, where different kinds of feature extractors, especially CNN, can be employed. For decoding, a sequence model such as RNN [24] or LSTM [17] can be utilized to generate the corresponding caption word by word with the guidance of the feature vector. The methods of this type for RSIC were first proposed in [25] by Qu et al., where a multimodal nueral network model is specially designed to understand the remote sensing images in semantic level. Besides, different CNNs combined with RNN or LSTM are explored to find the best combination for RSIC in this paper with two public datasets released. After that, Lu et al. [26] presented a new public dataset for RSIC, and both the “soft” and “hard” attention are introduced in this paper. After that, attention mechanism is widely used in the field of RSIC. For example, Yuan et al. [27] presented a multi-level attention module concentrating on different spatial positions and different scales. In general, methods based on encoder-

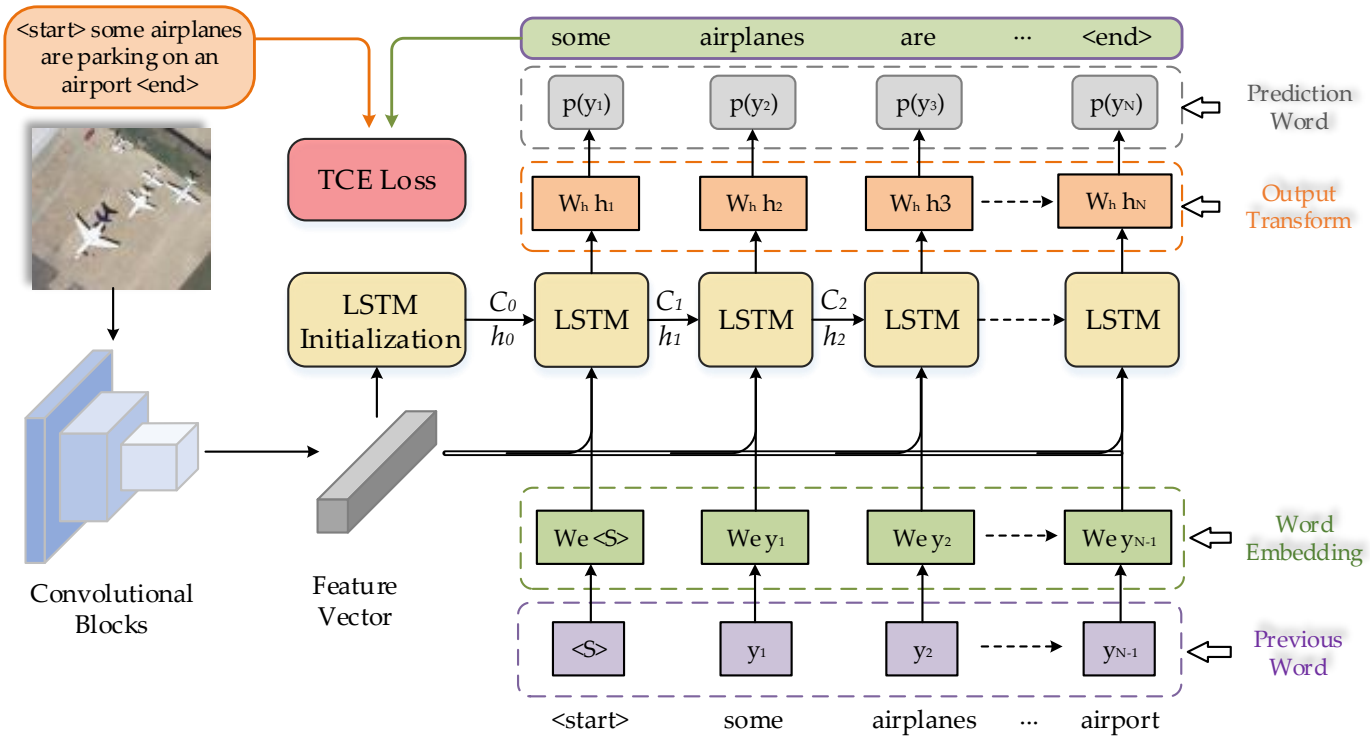


Fig. 2: The framework of the proposed TCE loss based method.

decoder model have become more and more popular because of their excellent performace.

B. Cross Entropy Loss for RSIC

Cross Entropy (CE) is mainly used to indicate the difference information between two probability distributions. Generally, for encoder-decoder based model of RSIC, the procedure of word prediction can be viewed as a classification task, where CE loss is employed for optimization. For decoding part, all words in candidate vocabulary would be assigned a probability after going through a softmax layer. The goal of CE loss is to optimize the probability of the target word to 1 while non-target words to 0, and only one word would be generated at each time step. CE loss used in RSIC can be formulated as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y^{(i)} * \log \hat{y}^{(i)} \tag{1}$$

where $\mathcal{L}_{CE}$ denotes the Cross Entropy loss, $y^{(i)}$ and $\hat{y}^{(i)}$ refer to the ground truth label and its factual output probability, respectively. N represents the total number of classes.

III. Methodology

In this section, the proposed Truncation Cross Entropy (TCE) loss is introduced in detail and used to optimize the captioning models obeying the classical encoder-decoder framework for RSIC. As shown in Fig. 2, the workflow of TCE loss based encoder-decoder model mainly consists of three components: (1) Feature Extractor. (2) Caption Generator. (3) Truncation Cross Entropy Loss.

A. Feature Extractor

For an encoder-decoder based method, the goal of image representation is to encode the input images into high-level semantic features, which is a vital part for RSIC. Traditionally, the methods of feature extraction can be generally divided into two categories: handcrafted features and deep learning features. Nowadays, a large number of deep learning models, especially Convolutional Neural Network (CNN) [28]–[31], have shown surprising feature representation in a wide range of image fields, and they also work well in remote sensing image captioning. Different from the methods based on handcrafted features, which require considerable engineering skills and domain knowledge, CNN can automatically learn features from data through a deep-structured neural network. Besides, since many processing layers are generally contained, CNN can learn and obtain more powerful feature representations with multiple levels of abstraction.

Since CNN has achieved excellent visual image representation, it is utilized to extract the features of remote sensing images. For a CNN model, it is usually composed of the stacked convolutional blocks (backbone) and fully-connected layer (classifier). However, the classifier is redundant for RSIC. Therefore, the last fully-connected layer is removed from CNN and the rest CNN backbone is used as feature extractor. To verify that the proposed TCE loss is not limited by CNN models, in this paper, several different kinds of CNN models are used for feature extraction.

Given an RGB remote sensing image of I, the multi-channel semantic feature map with the spatial size of $H \times W$, which is denoted as $F \in \mathbb{R}^{C \times H \times W}$, is extracted from the image by CNN backbone. Here C is the number of feature channels. It

is formulated as:

$$F = CNN_{conv}(I), \tag{2}$$

where CNN_{conv} represents the CNN backbone.

To reduce the model parameters, F is converted into the corresponding feature vector of $v \in \mathbb{R}^C$ by a Global Average Pooling (GAP) layer [32], which is formulated as:

$$v = GAP(F), \tag{3}$$

B. Caption Generator

Generally, the goal of caption generation is to decode the feature vector extracted by CNN into a sequence of words. Here, Long-Short Term Memory (LSTM) [17], which is a widely used sequence generation model, is utilized as a caption generator in our model. Different from the traditional sequence models, since LSTM is able to store long-term memory information, it has a good ability to solve the problem of gradient vanishing.

The structure of LSTM used in this paper is shown in Fig. 3, and it is constructed by several basic blocks stacked together. It is worth noting that the core of LSTM is the cell state, where the transfer process of information from the last cell state c_{t-1} to the current cell state c_t is specially controlled by three gates, *i.e.*, input gate i_t, forget gate f_t, and output gate o_t. At time step t, the update procedure of LSTM can be represented as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \tag{4}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \tag{5}$$

$$\tilde{c_t} = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_i), \tag{6}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}, \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \tag{8}$$

$$h_t = o_t * tanh(c_t), \tag{9}$$

where h_t represents the hidden state of LSTM at time t, which is also the output of LSTM at this time step. $tanh$ and σ sigmoid are respectively the hyperbolic tangent function and sigmoid function. All the W and b are the learnable parameters of weights and bias. In addition, x_t denotes the input of LSTM at time t, and the definition of x_t is formulated by:

$$x_t = concat(v, y_{t-1}), \tag{10}$$

where v and y_{t-1} denote the feature vector extracted by CNN and the output of LSTM at time $t-1$, respectively. x_t is the combination of v and y_{t-1}.

In general, the overall procedure of decoding can be denoted as:

$$h_t = LSTM(x_t), \tag{11}$$

$$y_t = W_{ed} * h_t, \tag{12}$$

where y_t represents the output of an word embedding operation on h_t. At each time step, only one word would be output by LSTM.

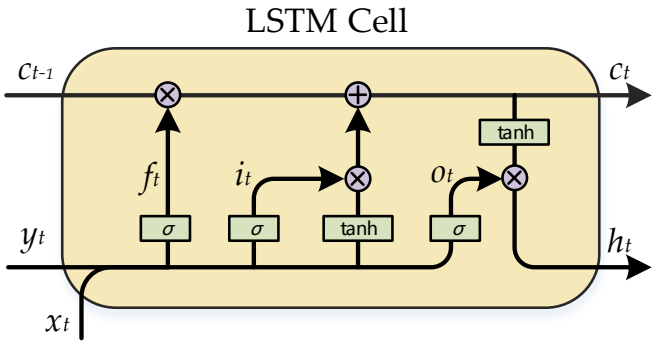


Fig. 3: The structure of LSTM.

C. Optimization Loss

1) ***Loss Defination****:* In image captioning, the word prediction is regarded as a multi-classification task. Therefore, the word probability distribution of the generated sentence and reference sentence can be denoted as $\hat{\boldsymbol{y}} = [\hat{y}_1, ..., \hat{y}_N]$, $\hat{y}_t \in \mathbb{R}^K$, and $\boldsymbol{y} = [y_1, ..., y_N]$, $y_t \in \mathbb{R}^K$. Here N is the sentence length and K is the vocabulary size. For a label sequence of $\boldsymbol{s} = [s_1, ..., s_N]$ where $s_i \in \mathbb{R}^1$ refers to the word number in the vocabulary at time step t. $\hat{\boldsymbol{y}}$ is the one-hot verson of $\boldsymbol{s}$. Generally, for a predicted sentence, the loss utilized for optimization during training can be denoted as:

$$\mathcal{L} = -\sum_{t=1}^{T} \mathcal{L}_t, \tag{13}$$

where t and T represent the current time step and the total time steps, respectively.

2) ***PCE, NCE and PNCE Loss****:* At time step t, the sub-loss of $\mathcal{L}_t$ can be expressed in three different forms of Positive Cross Entropy loss (PCE loss, denoted as $\mathcal{L}_{PCE}$), Negative Cross Entropy loss (NCE loss, denoted as $\mathcal{L}_{NCE}$) and Positive-Negative Cross Entropy loss (PNCE loss), which aim at optimizing target words directly and indirectly. It is worth mentioning that the concepts of NCE loss and PNCE loss are first proposed in this paper.

In $\mathcal{L}_{PCE}$, only the target word, which can be regarded as the positive sample, is treat as optimization objective. It is defined as $\mathcal{L}_{PCE}$:

$$\begin{aligned} \mathcal{L}_{PCE} &= -y_t^{(s_t)} * \log \hat{y}_t^{(s_t)} \\ &= -\log \hat{y}_t^{(s_t)}, \end{aligned} \tag{14}$$

where s_t denotes the number of the target word at time step t. $y_t^{(s_t)}$ and $\hat{y}_t^{(s_t)}$ respectively refer to the output probability and target probability of s_t. It is obviously that the value of $y_t^{(s_t)}$ equals to 1. The goal of PCE loss is to optimize the output probability of the target word to 1. PCE loss intrinsically is the typical CE loss [33] when it is used in RSIC.

In L_{NCE}, only the non-target words at time step t, which can be considered as the negative samples in contrast to the positive sample, are treat as optimization objectives. It is

defined as $\mathcal{L}_{NCE}$:

$$\begin{aligned}\mathcal{L}_{NCE} &= -\frac{1}{N-1}\sum_{i\neq s_t}^{N}(1-y_t^{(i)})*(1-\log\hat{y}_t^{(i)}) \\ &= -\frac{1}{N-1}\sum_{i\neq s_t}^{N}(1-\log\hat{y}_t^{(i)}),\end{aligned} \tag{15}$$

where $y_t^{(i)}$ and $\hat{y}_t^{(i)}$ respectively refer to the output probability and target probability of the i-th word in the vocabulary except s_t. The goal of NCE loss is to optimize the sum of the output probability of all the non-target words to 0.

Although both PCE and NCE loss have the same goal to make the output probability of s_t converge to 1, their optimization objectives and strategies are different. For more comprehensive ablation study, they are composed into a new combination loss of PNCE loss denoted as L_{PNCE}, which is defined as:

$$\begin{aligned}\mathcal{L}_{PNCE} &= \mathcal{L}_{PCE} + \mathcal{L}_{NCE} \\ &= -\log\hat{y}_t^{(s_t)} - \frac{1}{N-1}\sum_{i\neq s_t}^{N}(1-\log\hat{y}_t^{(i)}),\end{aligned} \tag{16}$$

*3) **TCE Loss**:* The aim of Truncation Cross Entropy (TCE) loss is designed to alleviate the over-fitting problem of captioning model for RSIC at training stage. The proposed TCE loss denoted as $\mathcal{L}_{TCE}$ is a piecewise function, which is made up of two components: a common PCE loss and a truncation loss. It is formulated as:

$$\mathcal{L}_{TCE} = \begin{cases} -y_t^{(s_t)} * \log\hat{y}_t^{(s_t)}, & \text{if } y_t^{(s_t)} < 1-\gamma \\ -\log(1-\gamma), & \text{otherwise,} \end{cases} \tag{17}$$

where γ is the value of truncation threshold which reserves a margin for the non-target words during the training stage. In this paper, γ is set to 0, 0.1, 0.2, 0.3 and 0.4, respectively. It is worth mentioning that TCE loss is actually the PCE/CE loss when γ is equal to 0. $y_t^{(s_t)}$ and $\hat{y}_t^{(s_t)}$ respectively refer to the output probability and target probability of s_t. The proposed TCE loss is illustrated in Fig. 4.

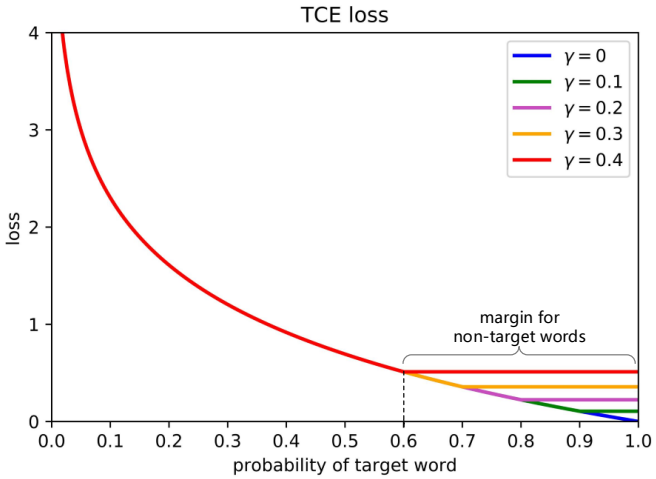


Fig. 4: TCE loss with different values of γ.

1. There are many residential areas near the school.
2. A playground is built next to a white building.
3. Several playgrounds are in the big school.
4. Several playgrounds are in the big school.
5. There are many residential areas near the school.

Fig. 5: An example of an image and its corresponding five captions in RSICD.

TCE loss is developed based on PCE loss. Different from PCE loss, however, an upper limit of $1-\gamma$ is set in TCE loss to determine which loss would be chosen during the optimization procedure. Supposing that the output probability of the target word (*i.e.*, $y_t^{(s_t)}$) exceeds the upper limit of $1-\gamma$, the loss function switches from PCE loss to Truncation loss. Therefore, $y_t^{(s_t)}$ would be optimized to a value of $1-\gamma$ instead of 1, which means a probability margin of γ would be specially reserved for the rest non-target words in the vocabulary except the target word. In this way, the over-fitting phenomenon caused by common CE loss can be effectively alleviated.

IV. Experiments

In this section, experiments are conducted on three datasets to verify the effectiveness and generalization of the proposed method. First, we introduce the experimental datasets and evaluation metrics of RSIC. Then the experiment settings are provided in detail. Following that, the over-fitting phenomenon caused by Cross Entropy loss is discussed, and the ablation experiments of PCE, NCE, PNCE and TCE loss are performed. Finally, our results are compared with some state-of-the-art models with comparative analysis.

A. Datasets

There are three widely used RSIC datasets of different sizes, including Sydney-captions, UCM-captions and RSICD.

1) Sydney-captions: The Sydney-captions dataset is proposed by Qu et al. [25], which is based on Sydney Data Set [34]. It contains 613 images with seven scene categories, including industrial, rivers, residential, meadow, runway, airport and ocean. All the images are collected from Google Earth of Sydney, Australia. Besides, the resolution of each image is 0.5m. For each image, five reference sentences are given to abstract the content from different observers. Totally, there are 237 different words in Sydney-captions dataset. 80% of the images in Sydney-captions are used for training, 10% for validation and the rest 10% for test.

2) UCM-captions: The UCM-captions dataset [25] is also proposed in [25], which is based on the UC Merced (UCM) land-use data set [35]. It contains 2100 high resolution remote sensing images with 21 scene categories, including building, beach, airplane, chaparral, forest, harbor, freeway, overpass, intersection, runway, river, agricultural, dense residential, tennis court, sparse residential, golf course, baseball diamond, medium residential, parking lot, mobile home park, and storage tank. All the images are measuring 256 $\times$ 256 pixels with a pixel resolution of 0.3048m. Totally, there are 368 different words in UCM-captions dataset. Similar to Sydney-captions

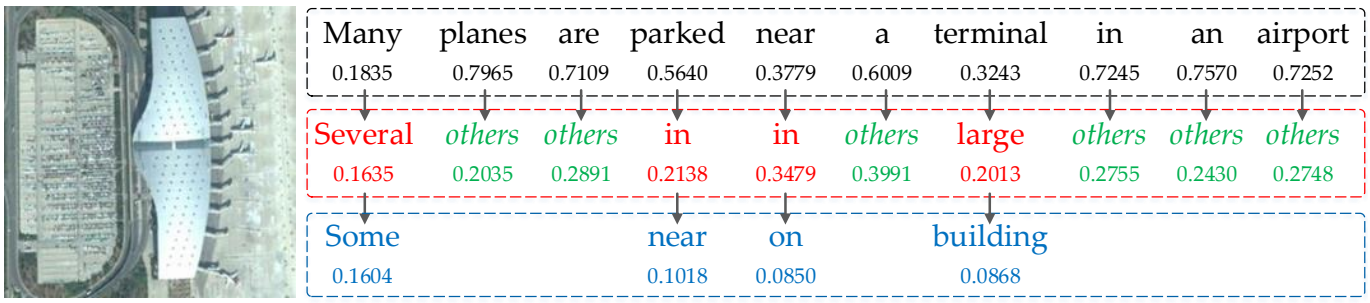


Fig. 6: An example of the probability of the generated top-3 words at each time step. The 1st, 2nd and 3rd row show the words of the **maximal**, the **submaximal** and the **third** probability at each time step. Besides, the ***others*** denotes all the non-maximal words in the vocabulary.

data set, five descriptions are given for each image. And the splitting ratio is the same as Sydney-captions.

3) RSICD: The RSICD dataset is proposed by Lu et al. [26], which consists of 10921 images measuring 224 $\times$ 224 pixels. All the images are collected from MapABC, Baidu Map, Google Earth, and Tianditu with different resolutions. Totally, there are 3323 different words in RSICD dataset. Similar to previous datasets, five reference sentences are provided for each image. It is mentioning that there are repetitions in the five reference sentences and its splitting ratio is the same as Sydney-captions and UCM-captions. There is a captioning example of a remote sensing image is shown in Fig. 5.

B. Evaluation Metrics

In the field of remote sensing image caption generation, four evaluation metrics are commonly used, including BLEU [40], METEOR [41], ROUGE_L [42] and CIDEr [43]. The value range of BLEU, METEOR and ROUGE_L are all from 0 to 1, and the value range of CIDEr is from 0 to 5. For all four metrics, the larger the value is, the better the quality of the generated captions is.

1) BLEU: BiLingual Evaluation Understudy (BLEU) [40] was first utilized to evaluate the quality of machine translation models, and now it is widely used in various sequence generation tasks. By calculating the precision of n-gram of different lengths and performing geometric weighted average, BLEU aims to measure the n-gram coincidence between the generated sequence and the reference sequence. Here, the value of n is set to 1, 2, 3, 4 corresponding to BLEU1, BLEU2, BLEU3 and BLEU4.

2) METEOR: Metric for Evaluation of Translation with Explicit ORdering (METEOR) [41] is measured by computing an alignment between the generated sentence and the reference sentence. Based on a single-precision weighted harmonic mean and single-word recall rate, METEOR takes into account both precision and recall rate, thus it can solve some of the defects inherent in the BLEU standard.

3) ROUGE_L: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [42] was first applied in the field of text summarization. It is similar to BLEU, while the difference is that ROUGE_L concentrates on calculating the recall rate. F-measure based on LCS(Longest Common Subsequence) is adopted in ROUGE_L metric to evaluate the similarity of the reference sentence and the generated sentence.

4) CIDEr: Consensus-based Image Description Evaluation (CIDEr) [43] is a metric for image captioning. In CIDEr, each sentence is translated into a "document" and expressed as a TF-IDF (term frequency inverse document frequency) vector. Then the cosine similarity between the reference sentence and the generated sentence would be calculated by the model. Compared with the above mentioned metrics, CIDEr takes into account the frequence of words in the vocabulary.

C. Experimental Settings

In this paper, all the experiments are built on Pytorch 1.3 of Python 3.7. Four different CNN models pretrained on ImageNet, including AlexNet [44], VGG16 [45], ResNet18 [46], and GoogleNet [47], are employed to explore the efficient CNN feature extractor for RSIC. We remove all layers after the last convolutional layer of CNNs, which are replaced by a global average pooling layer. The output of CNN extractors is a high-dimension feature vector, which is the input of a one-layer LSTM. For AlexNet, the dimension of feature vector is 256, while for the other three CNN extractors, the dimension of the feature vector is 512. During decoder stage, the dimension of both the word embedding and hidden state of LSTM is set to 512 in all the experiments. In order to improve the memory efficiency, the data is processed in batches with the batch size set to 64. During the whole training phase, Adam is utilized to optimize the models. All the models are trained for 50 epochs with the learning rate of 1e-4. For TCE loss, we set the truncation threshold value γ to 0, 0.1, 0.2, 0.3, 0.4, respectively to explore the effect of margin value.

D. Exploring The Efficient CNN Extractor

In order to explore the efficient CNN extractor for RSIC, four classical CNN achitectures are adopted here, including AlexNet, VGG16, ResNet18, and GoogleNet. AlexNet [44] is designed by implements deep convolutional neural network structure for the first time in large-scale image datasets. Compared with AlexNet, VGG16 [45] is improved by replacing the large convolution kernels with several small consecutive convolution kernels, with much deeper layers to learn more complex patterns. GoogleNet [47] has an unique inception

TABLE I: The experimental results with different CNN extractors on UCM-captions dataset.

CNN Extractor	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
AlexNet	0.7567	0.6728	0.6123	0.5626	0.3984	0.6755	2.4720
VGG16	0.7758	0.6965	0.6410	0.5929	0.4275	0.7073	2.6150
ResNet18	0.8079	0.7384	0.6861	0.6397	0.4496	0.7306	2.7688
GoogleNet	**0.8193**	**0.7522**	**0.7007**	**0.6559**	**0.4708**	**0.7483**	**2.8996**

TABLE II: The experimental results of PCE (*i.e.*, CE), NCE and PNCE loss on three datasets.

Dataset	Loss	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
UCM-captions	$\mathcal{L}_{PCE}$	**0.8193**	**0.7522**	**0.7007**	**0.6559**	**0.4708**	**0.7483**	2.8996
	$\mathcal{L}_{NCE}$	0.8034	0.7327	0.6775	0.6283	0.4586	0.7382	2.8390
	$\mathcal{L}_{PNCE}$	0.8035	0.7442	0.6943	0.6474	0.4617	0.7433	**2.9470**
Sydney-captions	$\mathcal{L}_{PCE}$	**0.7960**	**0.7238**	**0.6496**	**0.5810**	**0.4244**	**0.7034**	2.0180
	$\mathcal{L}_{NCE}$	0.7751	0.6882	0.6066	0.5346	0.4015	0.6837	1.9248
	$\mathcal{L}_{PNCE}$	0.7877	0.7106	0.6512	0.6015	0.4147	0.6920	**2.1999**
RSICD	$\mathcal{L}_{PCE}$	**0.7440**	**0.6183**	**0.5295**	**0.4615**	**0.3421**	**0.6500**	**2.3524**
	$\mathcal{L}_{NCE}$	0.7011	0.5608	0.4710	0.4065	0.3059	0.6068	1.9541
	$\mathcal{L}_{PNCE}$	0.7290	0.6058	0.5203	0.4555	0.3248	0.6438	2.2900

structure consisting of several convolutional kernels of different sizes, and thus is better for multi-scale feature extraction task. For taking advantage of a novel residual block, ResNet [46] is designed to alleviate the problem of gradient disappearance in deep networks with skip-connection operation, which makes the network training faster and more stable. Since the number of images in UCM-captions dataset is moderately, experiments with different CNN extractors are conducted on UCM-captions to find the efficient CNN extractor for RSIC.

The experimental results with four pretrained CNN extractors on UCM-captions dataset are provided in Table I, where the truncation threshold value of γ is fixed to 0, *i.e.*, PCE loss. In the experiments, GoogleNet achieves the best performance in all kinds of evaluation metrics. Taking CIDEr for example, GooleNet has an increase of 0.4276 as compared to AlexNet, 0.2846 as compared to VGG16, and 0.1308 as compared to ResNet18. Thus, it would be used as the feature extractor in all the subsequent experiments. It is probably that GoogleNet can fuse multi-scale features by applying several convolution kernels of different sizes at each layer, which is suitable for feature extraction of remote sensing images

E. Results of PCE, NCE and PNCE Loss

To quantificationally explore the influence of positive and negative sample training strategy for RSIC, we use the proposed three kinds of losses of PCE loss (*i.e.*, CE loss), NCE loss and PNCE loss, to optimize the same encoder-decoder model (GoogleNet + LSTM) on three datasets. The model optimized by PCE loss only uses the positive samples for optimization during training stage. Similarly, all the negative samples are utilized to optimize the model for NCE loss. When it comes to PNCE loss, both the positive and negative samples are adopted for optimization.

The experimental results are shown in Table II. Overall, the encoder-decoder model optimized by PCE loss achieves the best performance on all three datasets, and PNCE loss has the suboptimal results and NCE loss performs worst. Taking RSICD for instance, the BLEU1, BLEU2, BLEU3, BLEU4, METEOR, ROUGE_L and CIDEr scores of the model optimized by PCE loss are respectively 0.7440, 0.6183, 0.5295, 0.4615, 0.3421, 0.6500 and 2.3524, which have increases of 0.0429, 0.0575, 0.0585, 0.055, 0.0362, 0.0432 and 0.3983, respectively, as compared to the model optimized by NCE loss. When NCE is added to PCE loss, *i. e.*, PNCE loss, the scores drop by respectively 0.015, 0.0125, 0.0092, 0.006, 0.0173, 0.0062 and 0.0624 as compared to single PCE loss. Such results show that compared with PCE loss, NCE loss plays a negative role in improving the caption quality.

First of all, it should be noted that the model would be overfitting and further decrease the performance when the output probability of all the words is optimized to 1. Although the training targets of PCE, NCE and PNCE loss are the same to make the output probability of words at each time optimized to 1, but they have different training effects. NCE loss optimizes the negative words at each time step while PCE loss directly optimize the positive words. Limited by the implicit optimization targets, it is understandable that the performance of NCE loss falls behind PCE loss. But when NCE is added to PCE loss, the new PNCE loss would lead to the faster convergence of model and make the model miss the better parameter space.

TABLE III: The experimental results of TCE loss with different margin values on three datasets. When γ is equal to 0, TCE loss is equivalent to PCE/CE loss.

Dataset	Threshold	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
UCM-captions	$\gamma = 0$	0.8193	0.7522	0.7007	0.6559	0.4708	0.7483	2.8996
	$\gamma = 0.1$	0.8210	**0.7622**	**0.7140**	**0.6700**	**0.4775**	0.7567	2.8547
	$\gamma = 0.2$	0.8237	0.7604	0.7089	0.6681	0.4762	**0.7591**	**2.9809**
	$\gamma = 0.3$	0.8057	0.7415	0.6880	0.6387	0.4694	0.7451	2.9038
	$\gamma = 0.4$	**0.8242**	0.7592	0.7059	0.6569	0.4730	0.7551	2.9332
Sydney-captions	$\gamma = 0$	0.7960	0.7238	0.6496	0.5810	0.4244	0.7034	2.0180
	$\gamma = 0.1$	0.7937	0.7304	**0.6717**	**0.6193**	**0.4430**	**0.7130**	**2.4042**
	$\gamma = 0.2$	0.7873	0.7173	0.6512	0.5926	0.4429	0.7039	2.1447
	$\gamma = 0.3$	**0.8067**	**0.7356**	0.6677	0.6082	0.4266	0.7050	2.3607
	$\gamma = 0.4$	0.7820	0.7062	0.6355	0.5757	0.4332	0.6991	2.0517
RSICD	$\gamma = 0$	0.7440	0.6183	0.5295	0.4615	0.3421	0.6500	2.3524
	$\gamma = 0.1$	0.7589	0.6264	0.5319	0.4597	0.3431	0.6565	2.3237
	$\gamma = 0.2$	0.7438	0.6138	0.5212	0.4513	0.3353	0.6448	2.3041
	$\gamma = 0.3$	**0.7608**	**0.6358**	**0.5471**	**0.4791**	0.3425	**0.6687**	**2.4665**
	$\gamma = 0.4$	0.7512	0.6270	0.5398	0.4735	**0.3439**	0.6548	2.4143

TABLE IV: Comparison results of different methods on UCM-captions dataset.

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
FC-Att+LSTM [36]	0.8135	0.7502	0.6849	0.6352	0.4173	0.7504	2.9958
SM-Att+LSTM [36]	0.8154	0.7575	0.6936	0.6458	0.4240	0.7632	3.1864
Soft Attention [26]	0.7454	0.6545	0.5855	0.5250	0.3886	0.7237	2.6124
Hard Attention [26]	0.8157	0.7312	0.6702	0.6182	0.4263	0.7698	2.9947
GCN Based Multi-Level Attention [27]	**0.8330**	**0.7712**	**0.7154**	0.6623	0.4371	0.7763	3.1684
sound-a-a [37]	0.7484	0.6837	0.6310	0.5896	0.3623	0.6579	2.7281
RTRMN(statistical) [38]	0.8028	0.7322	0.6821	0.6393	0.4258	0.7726	3.1270
VAA [39]	0.8192	0.7511	0.6927	0.6387	0.4380	**0.7824**	**3.3946**
The Proposed Method	0.8210	0.7622	0.7140	**0.6700**	**0.4775**	0.7567	2.8547

F. Results of Truncation Cross Entropy (TCE) Loss

In order to slow down the speed towards overfitting during the training stage and further search for the parameters with more generalization performance, we use Truncation Cross Entropy (TCE) loss to optimize the same encoder-decoder based model (GoogleNet + LSTM) on three datasets. Different from PCE, NCE and PNCE loss mentioned above, the output probability of the target word would not be optimized to 1 by TCE loss, which means a probability margin could be reserved for the rest non-target words related except the target word. The truncation threshold denoted as γ is set to 0, 0.1 0.2 0.3 and 0.4, respectively, to further explore the influence of probability margin in this subsection.

Table III shows the comparative results of PCE loss and TCE loss with different margin values on three datasets. Compared with absolute PCE loss, when adding Margin Max Operation (TCE loss), the performance of RSIC obtains a comprehensive enhancement. Similarly, taking RSICD for example, the BLEU1, BLEU2, BLEU3, BLEU4, METEOR, ROUGE_L and CIDEr scores are respectively 0.7608, 0.6358, 0.5471, 0.4791, 0.3425, 0.6687 and 2.4665 under the circumstance that γ is set to 0.3. Compared with PCE loss optimized model (namely, when γ is equal to 0 in Table III), there are wide increases of 0.0168, 0.0175, 0.0176, 0.0176, 0.0004, 0.0187 and 0.1141, respectively. Besides, it is worth mentioning that the optimal value of γ is 0.1 for both UCM-captions and Sydney-captions, and it increases to 0.3 when using the largest dataset of RSICD. Overall, according to the results, it can be observed that by setting a truncation threshold, the model optimized by TCE loss achieves better performance than PCE loss, and the best threshold varies with the datasets.

It is notable that the best optimal margin value of γ increases with the size of a dataset, more specifically, the size of its vocabulary. Since RSICD contains more words (the total number of words is 1187) than the other two datasets, the same features or objects in RSICD can be described with more synonymous words compared with them in the other

TABLE V: Comparison results of different methods on Sydney-captions dataset.

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
FC-Att+LSTM [36]	0.8076	0.7160	0.6276	0.5544	0.4099	0.7114	2.2033
SM-Att+LSTM [36]	0.8143	0.7351	0.6586	0.5806	0.4111	0.7195	2.3021
GoogleNet Soft Attention [26]	0.7322	0.6674	0.6223	0.5820	0.3942	0.7127	**2.4993**
GoogleNet Hard Attention [26]	0.7591	0.6610	0.5889	0.5258	0.3898	0.7189	2.1819
GCN Based Multi-Level Attention [27]	**0.8233**	**0.7548**	0.6587	0.6003	0.4202	**0.7237**	2.3110
sound-a-a [37]	0.7093	0.6228	0.5393	0.4602	0.3121	0.5974	1.7477
VAA [39]	0.7431	0.6646	0.6029	0.5495	0.3930	0.6999	2.4073
The Proposed Method	0.7937	0.7304	**0.6717**	**0.6193**	**0.4430**	0.7130	2.4042

TABLE VI: Comparison results of different methods on RSICD dataset.

	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE_L	CIDEr
FC-Att+LSTM [36]	0.7459	0.6250	0.5338	0.4574	0.3395	0.6333	2.3664
SM-Att+LSTM [36]	0.7571	0.6336	0.5385	0.4612	0.3513	0.6458	2.3563
Soft Attention [26]	0.6753	0.5308	0.4333	0.3617	0.3255	0.6109	1.9643
Hard Attention [26]	0.6669	0.5182	0.4164	0.3407	0.3201	0.6084	1.7925
GCN Based Multi-Level Attention [27]	0.7597	**0.6421**	**0.5517**	0.4623	**0.3543**	0.6563	2.3614
sound-a-a [37]	0.6196	0.4819	0.3902	0.3195	0.2733	0.5143	1.6386
RTRMN(statistical) [38]	0.6102	0.4514	0.3535	0.2859	0.2751	0.5452	1.4820
The Proposed Method	**0.7608**	0.6358	0.5471	**0.4791**	0.3425	**0.6687**	**2.4665**

two datasets. Therefore, the problem of over-fitting between the interchangeable words during the training stage is more serious for it. Further, more probability margin of each word need to be reserved for its relative words in RSICD, which means that the optimal truncation threshold of γ should be bigger. In contrast, the number of words in UCM-captions and Sydney-captions is respectively 368 and 237, thus a small value of γ can perform well for them.

In summary, compared with the pure PCE loss, by reserving a suitable probability margin for the relative words to replace the target word, TCE loss is rather effective to alleviate the over-fitting for all three datasets during optimization process.

G. Comparison with State-of-the-Art Methods

To verify the effectiveness of the proposed method, comparison experiments are conducted on several state-of-the-art approaches, including FC-Att+LSTM [36], SM-Att+LSTM [36], Soft Attention [26], Hard Attention [26], GCN Based Multi-Level Attention [27], sound-a-a [37], RTRMN(statistical) [38], and VAA [39]. All of these methods are built on encoder-decoder framework. Both FC-Att+LSTM and SM-Att+LSTM are proposed in [36], where the multiple attributes, which are extracted from the high-level semantic features of remote sensing images by attention mechanism, are utilized to obtain better caption quality based on the basic model. The main difference of these two models is that their high-level attributes are the output of different layers of CNN (the last fully connected layer for FC-Att+LSTM while the softmax layer for SM-Att+LSTM). Soft Attention and Hard Attention are introduced in [26]. In particular, Soft Attention is a deterministic method where a weight is given to decide which part of the image should be paid more attention. Differently, Hard Attention is a stochastic method, where a sampling strategy is employed to concentrate on different parts of the image and then reinforcement learning is adopted to better improve the performance of the model. GCN Based Multi-Level Attention is proposed by [27]. Typically, a multilevel attention module is presented for better performance of RSIC, and the position-adaptive and scale-adaptive image representations can be learned by this model. The method of sound-a-a is proposed [37], where a sound mechanism is introduced as an active attention to improve the quality of caption generation. RTRMN(statistical) is proposed in [38]. This model aims to overcome the drawback of long-term information dilution in RNN, and a topic word strategy is presented to fully utilize the given five reference captions. VAA is proposed in [39], where a novel Visual Aligning Attention model is presented to address the problem of not explicitly training the attention layers in encoder-decoder model.

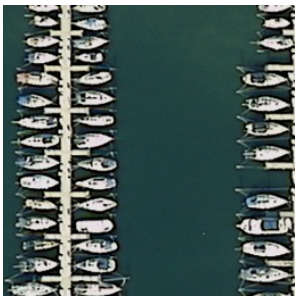
1) Results on UCM-captions: Table IV shows the comparison results between the aforementioned methods and the proposed CNN-LSTM model with TCE loss on UCM-captions dataset. It can be seen that the results of the proposed method are the best among all the methods in terms of BLEU4 and METEOR. According to BLEU1, BLEU2, BLEU3, ROUGE_L and CIDEr, the performance of the proposed method just slightly fall behind the other state-of-the-art approaches.

2) Results on Sydney-captions: The comparative results on Sydney-captions dataset are shown in Table V. It can be observed that the proposed method performs best in terms of BLEU3, BLEU4 and METEOR. As for BLEU1, BLEU2,

(a) There are two tennis courts arranged neatly and surrounded by some plants.

(b) There are two white storage tanks on the ground.

(c) An overpass go across the roads with some cars on the roads.



(d) Many boats docked in lines at the harbor and the water is deep blue.

(e) A residential area with many houses arranged neatly and divided into rectangles by some roads.

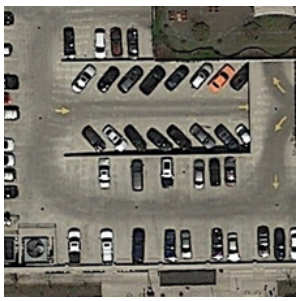
(f) Many buildings and some green trees are in a commercial area.

(g) Many airplanes are parked in an airport.

(h) A baseball field is near several buildings and some green trees.

(i) Some buildings and green trees are in two sides of a river with a bridge.

(j) Some green trees and several buildings are around a circle building.



(k) Many cars are parked in a parking lot.

(l) Many buildings and green trees are around a football field.

Fig. 7: Examples of test images and the corresponding generated captions.

ROUGE_L and CIDEr, our method achieves a competitive performance compared with other state-of-the-art approaches.

3) Results on RSICD: As the biggest RSIC dataset, RSICD contains much more images than the other two datasets with complex scenarios, and thus its result is more robust and convincing. The comparative results are shown in Table VI. The highest score for each metric is marked in bold. It can be observed that compared with all the existing sate-of-the-art approaches, the proposed method takes the first place in BLEU1, BLUE4, ROUGE_L and CIDEr, and competitive results in BLEU2, BLEU3 and METEOR. Such experimental results completely prove the effectiveness and superiority of our approach.

4) Advantages of Our Method: According to the analysis of methods and comparative experimental results on three datasets, it can be found that compared with the other sate-of-the-art methods, our approach has two obvious advantages, which can be summarized as follows:

Without adding extra modules or parameters. Different from the other comparison methods, the proposed method aims to improve the encoder-decoder based model from the aspect of training optimization, which means the model of our method is the basic CNN combined with LSTM, without adding any extra modules. In other words, there are no extra parameters embedded into our model, so the computational complexity and training cost of TCE loss is almost equal to the basic model, which operates much faster than the aforementioned methods.

Getting better performance. From Table IV to Table VI, the comparative results show that the proposed method achieves superior performance on RSIC datasets. Typically, when the number of n-gram for BLEU metric (from BLEU1

to BLEU4) increases, our method gets relatively better performance compared with others. It is probably caused by that the constraints on each word in phrase would be relaxed during the training stage, which makes our method more flexible for multiple consecutive words. It is beneficial for the open caption from different observers. Some examples of test images and the corresponding generated captions are shown in Fig. 6 and Fig. 7.

V. Conclusion

In this paper, we first quantificationally explore the overfitting phenomena in remote sensing image captioning (RSIC) caused by Cross Entropy loss. To deal with this problem, a novel Truncation Cross Entropy (TCE) loss is specially proposed for RSIC. By setting a truncation threshold, the output probability of the target word would not be optimized to 1 by TCE loss, and thus a probability margin can be reserved for the rest words in the vocabulary. Based on a classic encoder-decoder model (CNN plus LSTM), comparative experiments are conducted on three widely used remote sensing image captioning datasets, including UCM-captions, Sydney-captions and RSICD. The superior performance of the experimental results compared with other state-of-the-art methods shows that the proposed TCE loss is rather effective for RSIC.

References

[1] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 3017–3020.

[2] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2018.

[3] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 864–868, 2019.

[4] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, 2020.

[5] R. Pires de Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sensing*, vol. 12, no. 1, p. 86, 2020.

[6] Q. Wang, Z. Yuan, Q. Du, and X. Li, "Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 3–13, 2018.

[7] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 911–923, 2018.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[10] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[11] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 516–12 526.

[12] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[13] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," *arXiv preprint arXiv:2003.00387*, 2020.

[14] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," *arXiv preprint arXiv:2004.00390*, 2020.

[15] X. Zhang, Q. Wang, S. Chen, and X. Li, "Multi-scale cropping mechanism for remote sensing image captioning," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 10 039–10 042.

[16] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[19] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 2015.

[20] B. Wang, X. Lu, X. Zheng, and X. Li, "Semantic descriptions of high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 8, pp. 1274–1278, 2019.

[21] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?" *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[23] W. Huang, Q. Wang, and X. Li, "Denoising-based multiscale feature fusion for remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2020.

[24] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[25] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2016, pp. 1–5.

[26] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2017.

[27] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, 2019.

[28] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *International Journal of Remote Sensing*, vol. 37, no. 10, pp. 2149–2167, 2016.

[29] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2016.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[34] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2014.

[35] Y. Yang and S. Newsam, "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*. ACM, 2010, pp. 270–279.

[36] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

[37] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.
[38] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 256–270, 2020.
[39] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "Vaa: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137 355–137 364, 2019.
[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
[41] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.
[42] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[43] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

Xuelong Li (M'02-SM'07-F'12) is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China.

Xueting Zhang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. She is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. Her research mainly focuses remote sensing image proccessing.

Wei Huang received the B.E. degree in control theory and engineering from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and remote sensing.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.