Image Caption Report

Gong Chao
2020 Artificial Intelligence
ID: 20307140043
20307140043@fudan.edu.cn

Abstract

This report describes my final project: Image captioning. This task bridges vision and natural language. I used the Flickr30k dataset, which contains more than 30k images and corresponding captions. According to the mainstream encoder-decoder method, I use CNN and RNN to encode and decode respectively, and add attention mechanism in the middle. In my experiments, I randomly printed some images, generated captions and plotted the loss curve to measure the effect of the network. I performed regularization and data augmentation to better perform the task.

1 Introduction

Automatic generation of image captions shows the computer's understanding of images and is a fundamental task of artificial intelligence. For the caption model, in addition to finding out which objects are contained in the image, it also needs to be able to express the relationship between them in natural language. We can do this by the classical encoder-decoder model. In the model, I used resnet50 as the encoder and take its output, the features, as the input of the decoder. I used LSTM as the decoder.In order to better recognize these objects and clarify their relationships, I implemented the attention mechanism in the project, which can capture the most important information in the image and the relationship between these features. In the project, I applied an additive model, Bahdanau attention.

2 Dataset

The dataset we used is the flickr30k dataset, which contains 31783 images with five captions per image. The images include men, women and animals in different places doing different things, and each image is of different sizes.

I used spacy to segment words and added words with higher frequency than the threshold to my dictionary. The way to handle the data is to inherit paddle's Dataset, attach each image's data to its caption, and load it into a Dataloader that can batch.

3 Methodology

I'll describe the approach I used here.

3.1 Encoder

The Encoder encodes the input image with 3 color channels into a smaller image with "learned" channels. This smaller encoded image is a summary representation of all that's useful in the original image. Since we want to encode images, we use Convolutional Neural Networks (CNNs). Here I chose to use the 50 layered Residual Network trained on the ImageNet classification task.

Preprint. Under review.

These models progressively create smaller and smaller representations of the original image, and each subsequent representation is more "learned", with a greater number of channels. The final encoding produced by our ResNet-50 encoder has a size of 7x7 with 2048 channels, i.e., a 2048, 7, 7 size tensor.

3.2 Decoder

The Decoder's job is to look at the encoded image and generate a caption word by word. Since it's generating a sequence, it would need to be a Recurrent Neural Network (RNN). We used an LSTM. In a typical setting without Attention, you could simply average the encoded image across all pixels. You could then feed this, with or without a linear transformation, into the Decoder as its first hidden state and generate the caption. Each predicted word is used to generate the next word.Just as the Fig 1.

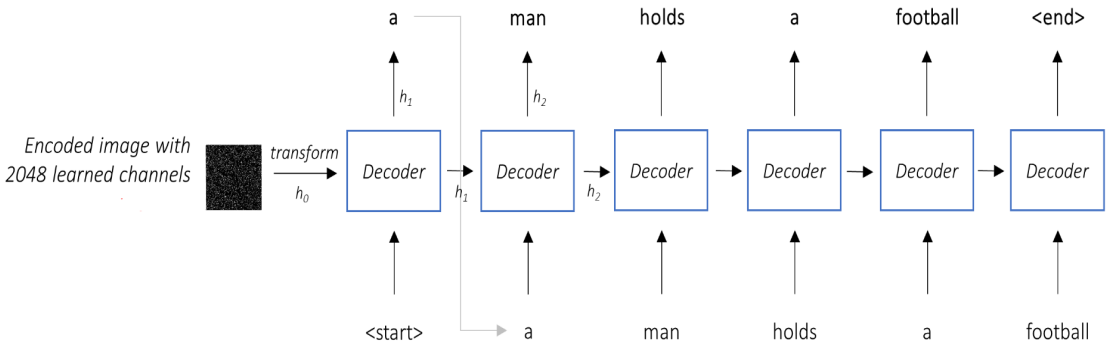

Figure 1: Decoder

In a setting with Attention, we want the Decoder to be able to look at different parts of the image at different points in the sequence. For example, while generating the word "football" in "a man holds a football", the Decoder would know to focus on – you guessed it – the football. It's shown in Fig 2

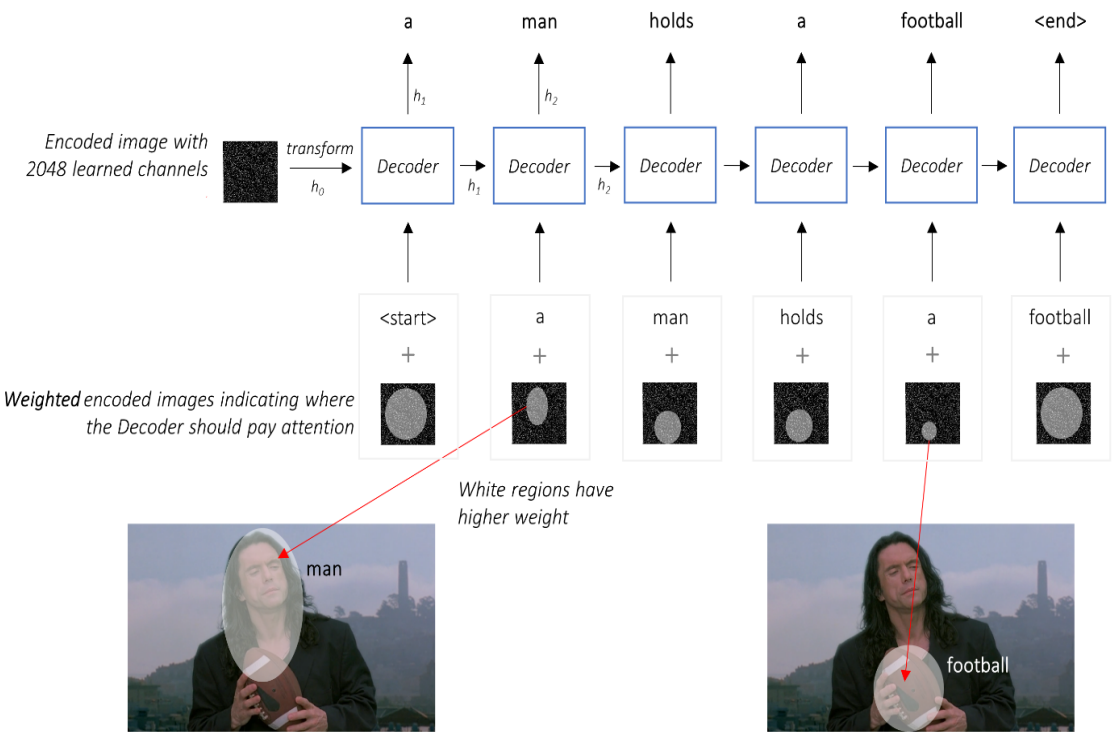

Figure 2: Decoder with attention

We used a linear layer to transform the Decoder's output into a score for each word in the vocabulary. The straightforward – and greedy – option would choose the word with the highest score and use it to predict the next word.

3.3 Attention

Attention mechanism can be used as a resource allocation scheme, which will select more relevant or important information to the task from a large number of candidate information.

In order to select the information relevant to a particular task from N input vectors, we need to introduce a task-relevant representation, called a Query Vector, and calculate the relevance between each input vector and the query vector through a scoring function. Given a task-relevant query vector q, we first compute the Attention Distribution, which is the probability of selecting the nth input vector:

$$\alpha_n = softmax(s(\mathbf{x}_n, \mathbf{q})).$$

Where $s(x, q)$ is the attention scoring function.Here we use the additive model Bahdanau attention:

$$s(\mathbf{X}, \mathbf{q}) = \mathbf{v}^T \tanh(\mathbf{X}\mathbf{W} + \mathbf{q}^T\mathbf{U}).$$

After obtaining the attention distribution, the input vectors can be weighted and averaged to obtain the final representation of the entire sequence:

$$\mathbf{z} = \sum_{n=1}^{N} \alpha_n \mathbf{x}_n.$$

The complete attention mechanism can be seen in the Fig 3.

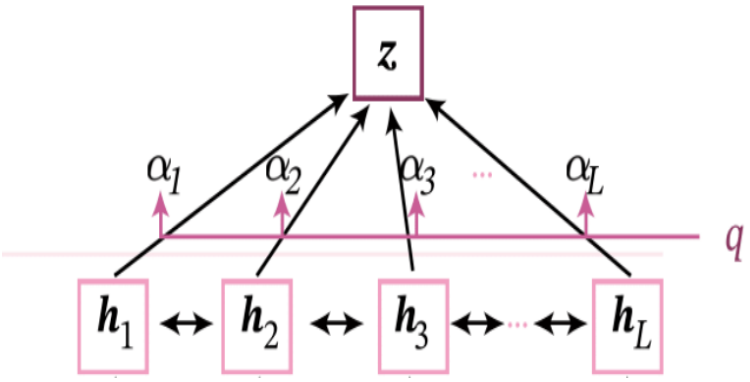

Figure 3: Attention mechanism

This is the full model. Once the Encoder generates the encoded image, we transform the encoding to create the initial hidden state h (and cell state C) for the LSTM Decoder. At each decode step,the encoded image and the previous hidden state is used to generate weights for each pixel in the Attention network. the previously generated word and the weighted average of the encoding are fed to the LSTM Decoder to generate the next word. Then we used Adam optimizer, cross-entropy loss function and mini-batch gradient descent for parameter learning.

3.4 Improvement

The basic model above didn't work very well. To improve the model, I added regularization mechanisms and simple image augmentations. I also tried Beam search but without success, leaving it for future discussion.

4 Results

4.1 The original version

The following two figures are examples from the first model. The model can speak fluent sentences and identify people, but it will mark all people as a man, and even mark animals as humans. In some images, even dogs are marked as men. It cannot accurately recognize the environmental background of the picture, and the sentence pattern is very single.

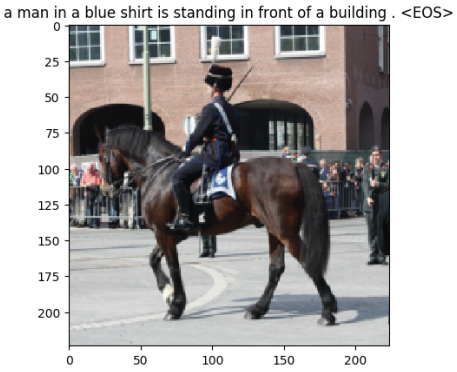

Figure 4: An example from the first model

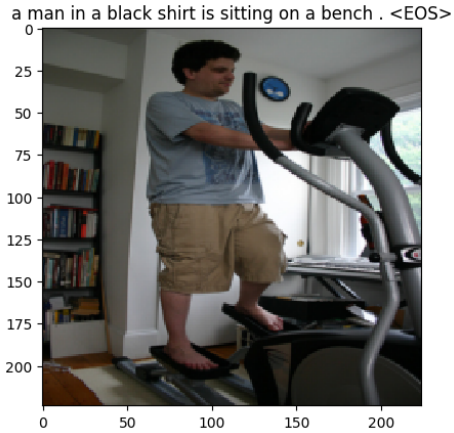

Figure 5: An example from the first model

Fig 6 is the loss curve for the first model. You can see that the loss decreases relatively quickly and finally stabilizes.

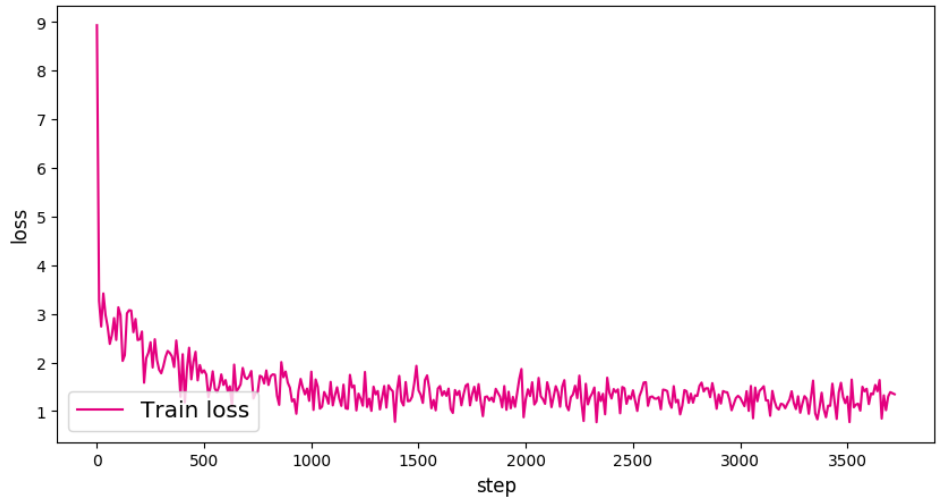

Figure 6: Loss curve for the first model

4.2 The version with regularization

The next two figures is chosen from the model with regularization. It's still not very good, can't distinguish between men and women, animals, complex scenes, the number of people, but the model has learned more new words, such as holding a drink, and is slightly better than the first model. The loss function curve is similar to the previous one and will not be listed.

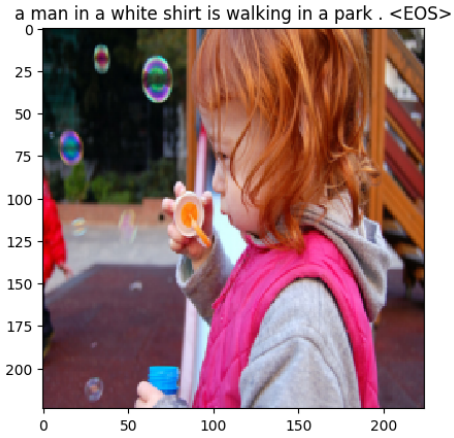

Figure 7: An example from the second model

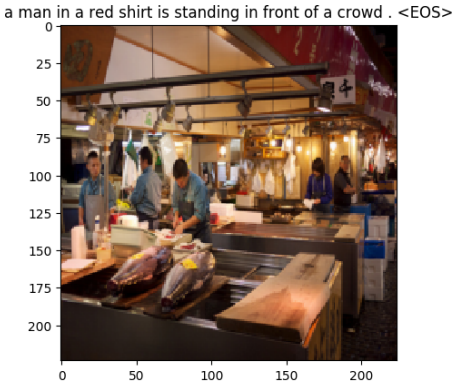

Figure 8: An example from the second model

4.3 The version with image enhancement

The following two images are selected with the image augmentation model. As you can see, the model learns more new words this time, but it is still unable to accurately identify the image scene. The sentence pattern is still very simple.

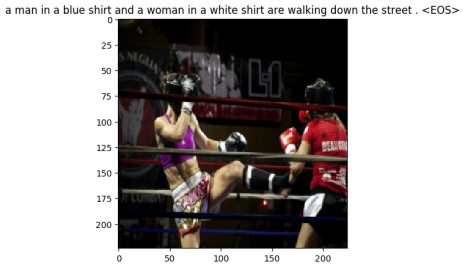

Figure 9: An example from the third model

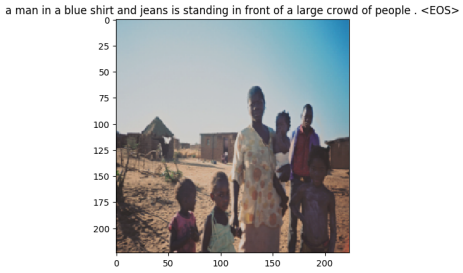

Figure 10: An example from the third model

5 Conclusion

To sum up, I implemented a neural network model based on the image caption of show,attend and tell. The model uses resnet50 as the encoder, LSTM as the decoder, and Bahdanau attention as the attention mechanism. The model was trained on flickr30k dataset. It can master a certain number of new words and speak fluent sentences, and identify the presence of people in the scene. However, due to overfitting and imbalanced samples, the model cannot accurately distinguish the gender, animal, and scene. Regularization and image augmentation did not solve this problem. It is hoped that the image caption task can be effectively completed after learning more knowledge.

The first time I did this kind of slightly larger project, I found myself a lot of shortcomings, but I have been on the way to improve myself.

References

[1] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 2048–2057.

[2] Vinícius Veríssimo, Cecília Silva, Vitor Hanael, Caio Moraes, Rostand Costa, Tiago Maritan, Manuella Aschoff, and Thaís Gaudêncio. 2019. A study on the use of sequence-to-sequence neural networks for automatic translation of brazilian portuguese to LIBRAS. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (WebMedia '19). Association for Computing Machinery, New York, NY, USA, 101–108. https://doi.org/10.1145/3323503.3360292

[3] Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. Multiple object recognition with visual attention. arXiv:1412.7755 [cs.LG], December 2014.