CANYU ZHANG

Missing Words Prediction For Long Sentence

Document damage is a common concern faced by researchers. The original paragraph might be meaningless when some key words are missing. Even though the missing word prediction has drawn more and more attention, current models sill can address this problem ideally. For example, current word prediction models always performs poorly on the long paragraph, since it's hard for the model to learning the meaning from long sentences. In this paper, we focus on solving this problems. We utilize a transformer based network to make the prediction for missing words in long paragraphs. To justify the effectiveness of proposed method, we generate a long sentence dataset. Proposed method generates a promising result, which can further help people deal with the missing words in long sentences.

Additional Key Words and Phrases: Word Prediction, Long Paragraph, NLP

ACM Reference Format:
Canyu Zhang. 2022. Missing Words Prediction For Long Sentence. 1, 1 (November 2022), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Word missing is a common problem happens in natural language processing area, which has drawn more and more interests. Though a lot of methods have been proposed for this tasks, they are far from practical usage. Most existing models focus on predicting missing words in short sentences, the prediction ability will be worse on long sentences. However, the long sentences is very common in real life. In order to solve this limitation, we choose to modify the previous transformer based models for long sentence missing parts prediction. Our model can consider the whole sentences as a whole, beating previous models like RNN and LSTM. Since there are few long sentence datasets, we generate a dataset for training and evaluation. To justify the performance of proposed method, we conduct thorough experiments on proposed models and baselines.

2 RELATED WORK

Transformer. Transformer can analysis the different parts from sentences or images with the help of self-attention block. It can consider the whole input as whole, capturing more global information compared with previous models like RNN and LSTM. It is firstly proposed for machine translation, and recently it has been used for image classification, e.g., ViT succeeds in many other image processing tasks, with superior performance over classical CNN-based models. Transformer network has also been widely used in 3D point cloud processing. For instance, Point Transformer constructs self-attention networks for 3D point cloud semantic segmentation and classification. In this paper, we choose transformer based model to solve the word prediction task.

Author's address: Canyu Zhang, UofSC.

XXXX-XXXX/2022/11-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METHOD

Our model is based on previous NLP model BERT, which is a combination of bidirectional transformers, which has been widely used for masked and next sentences prediction. It is pretrained on a large dataset comprising the Toronto Book Corpus and Wikipedia. The base model is similar to the original transformer. And we will fine-tune the pretrained BERT model for our task since training a transformer from the beginning is difficult. The input is the masked long sentences, following the previous work, we use symbol [CLS] to show the beginning of a sentence, [SEP] to separate different sentences and [MASK] to indicate the masked part to be predicted. The model will output the predictions for the masked words. During the training, we use cross entropy loss.

4 EXPERIMENTS

We train our model using Adam as the optimizer (1= 0.9, 2 = 0.999) with the learning rate 0.0001. Training epoch is 100 epochs and the batch size is set to 16. All experiments are implemented on NVIDIA Tesla V100 GPU.

5 EVALUATION

For evaluation, we use cosine similarity to evaluate the performance of different models. In first step, we use Word2Vec to process the words, then we compare the similarity between the word embeddings. The Word2Vec model is trained on Google news dataset. When the mask rate is 10%, the average similarity is 0.8358. When the mask rate is 10%, the average similarity is 0.7632. The model is more likely to make wrong predictions when facing more masked words. For pretrained model without fine-tuning, the average similarity is 0.7214. Those results shows our model can perform better on long sentences compared with the original one.

Here we show an example:

Input: Lutyens [MASK] his final design to the [MASK]of Works in early July, and on 7 July received [MASK] that the [MASK] had been approved by the foreign secretary, Lord Curzon, who was organising the parade. The unveiling of the monument, [MASK] in wood and plaster by the Office of Works, was [MASK] in The Times as a quiet and unofficial ceremony. It [MASK] place on 18 July 1919, the day [MASK] the Victory Parade.Lutyens was not [MASK]. During the parade, 15,000 soldiers and 1,500 officers marched past and [MASK] the Cenotaph.

Prediction: submitted, Office, confirmation, design, painted, described, took, of, present, around.

Ground truth: submitted, Office, confirmation, design, built, described, took, before, invited, saluted.