Final Project

Yue Zhao, Cindy Lin
UML
COMP 4420 Natural Language Processing

Abstract

In this work, we fine-tuned a pre-trained t5-base model and used it to solve a machine translation task that translates English to German. The goal of our model is to achieve a higher BLEU score than a model without pre-trained components. Our model is trained on a bilingual dataset which has a total of 800 thousand sentence pairs. Experimental results show that our model outperformed the baseline model by 2 BLEU points.

1 Introduction

Natural Language Processing (NLP) is the part of artificial intelligence that focuses on using computers to understand human language which has recently gained increasing attention. It has been applied in many fields such as machine translation, spam detection, information extraction, summarization, medical and question answering systems and so on.

In recent years, with the rise of Deep Learning, people have been trying to apply Deep Learning to NLP and have made many breakthroughs. One of them is the Seq2Seq model which takes a sequence of items and outputs another sequence of items (Sutskever et al., 2014, Cho et al., 2014). For example, in the case of machine translation the input is text or a series of words, and the output is the translated text or translated series of words. Our goal is to show that our model which consists of a fully pre-trained Seq2Seq t5-base network outperforms models that are trained from scratch on English-German machine translation tasks.

2 Background

2.1 Pre-trained T5 models

In this paper, we fine-tuned a pre-trained t5-base model to solve our machine translation task. The pre-trained model we used is a Seq2Seq model which is trained on a large dataset by Google and uses resources that are not usually available to everyone. Hence the pre-trained t5-base model would be beneficial for us to avoid training a model from scratch.

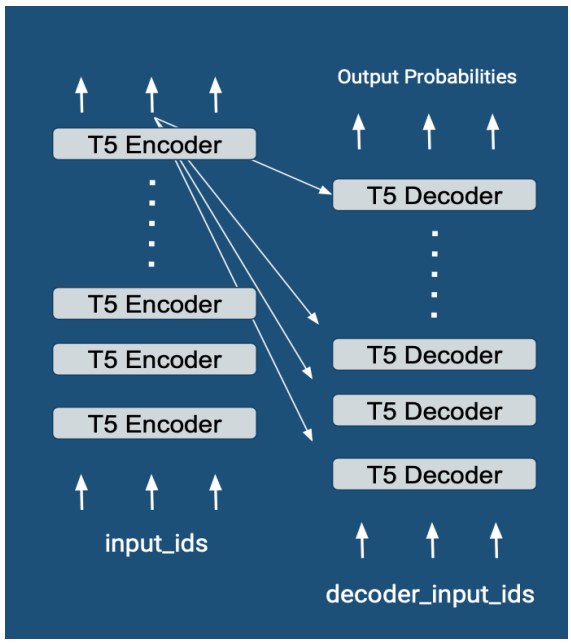

Figure 1: Architecture of T5 model

The purpose of the pre-trained model is to build a bridge between the original task and the target task. It allows us to first pre-train on multiple tasks to obtain universal language representations, and then use a small amount of data on the target task for fine-tuning, so that the fine-tuned model can handle the downstream NLP tasks well. Since pre-trained models have very high accuracy and require less training and implementation time compared to custom-built models, they are getting used more and more often on almost any NLP task.

2.2 Seq2Seq models

The t5-base model is a Seq2Seq model which is built based on encoder decoder architecture.

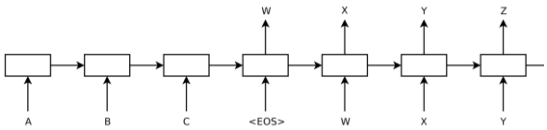

Figure 2: credit:(Sutskever et al., 2014, Cho et al., 2014)

Seq2Seq models break through the traditional fixed-size input problem and have been confirmed that it has a good performance in the application of English-French translation, English-German translation and short question answering tasks.
Deep Neural Networks (DNNs) cannot handle sequence modelling problems due to the limitation of the length of both input and output vectors are predefined. For example, in machine translation where the length of input (i.e. I went to a shop yesterday in English) and output (i.e. Ich war gestern in einem Geschäft in German) could be different, in this case DNNs cannot solve such problems (Sutskever et al., 2014, Cho et al., 2014) [1]. Therefore, it is necessary to propose a new solution to deal with the sequence problem of non-fixed lengths.
A simple Seq2Seq model consists of 3 parts, Encoder-LSTM, Decoder-LSTM, Context. As shown in the figure above, the input and output of this model are different at each time. For example, if the input sequence is “A B C EOS" (EOS=End of Sentence), the Encoder-LSTM will process the input sequence and return the hidden state of the entire input sequence at the last layer. Then Decoder-LSTM predicts the next character of the target sequence step by step according to the hidden state. Then it will return the sequence "W X Y Z EOS" as output. Since the input sequence is processed in reverse order, the model can process long sentences and improve accuracy (Sutskever et al., 2014, Cho et al., 2014) [1].

3 Machine Translation

3.1 Machine Translation

Machine Translation is a process in which computer software translates text from one language to another with the meaning of the input text preserved without human involvement. Due to its ability to handle large volumes of content and translate near-instantaneously and cost-effectively, it is widely used by institutional users, the military, social networking, etc. However, unlike other NLP problems, the difficulty in machine translation is that a word can have more than one meaning which causes word-sense disambiguation. One way to solve it is to use a pre-trained model, to achieve high accuracy on the downstream task with relatively low data and training time.

3.2 The data set

We are using **WMT14 English-German Translation Data with further preprocessing dataset,** which has 4548885 examples with size of 1.28 GB for training and 2169 examples with size of 0.5 MB.

The dataset includes sentence pairs for English and German.

{ "de": "Im Parlament besteht der Wunsch nach einer Aussprache im Verlauf dieser Sitzungsperiode in den nächsten Tagen.", "en": "You have requested a debate on this subject in the course of the next few days, during this part-session." }

3.3 Hardware

We trained our models on one Google Cloud Platform VM with 1 NVIDIA TESLA V100 GPU and 8 VCPU with 52 GB memory. Each training step took about 0.4 seconds. We trained the t5-base model for a total of 100,000 steps.

3.4 Preprocessing

The first thing that needs to be done in preprocessing is to tokenize both the input dataset (for English) and the target dataset (for German). To ensure the tokenized data can be saved in GPU with only 16 GB memory, we set max_length as 128. With the max length, we can use the largest batch size as 24. Without max length, our batch size needs to be reduced to 4. After tokenization, we took input_ids out from the target dataset and saved it into decoder_input_ids.

3.5 Training

To train our model, we chose to use the pre-trained t5-base Seq2Seq model as it has been pre-trained on a data-rich task and supports both unsupervised and supervised tasks (Raffel et al., 2020). It has been proven to provide excellent prediction results for machine translation with the source language as English. To fine-tune it with our dataset, we initialized the T5Tokenizer with passing of only EOS and PAD tokens since the BOS token is not supported in the t5-base model. And when calculating the loss of the model during

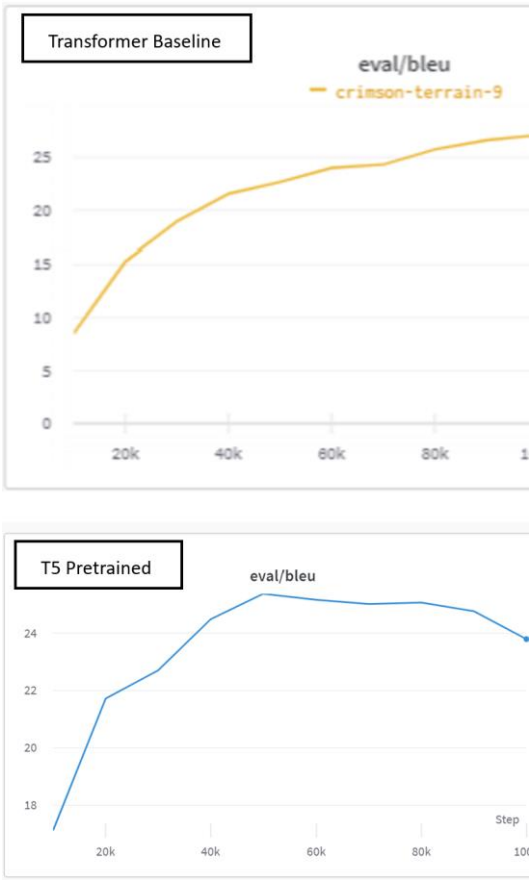

Figure 3: bleu comparison.

training, we used the loss from model's logits directly.

3.6 Analysis

	bleu
Transformer Baseline	29.33
T5-base Pretrained	33.27

Table 1: bleu report.

In this project, we evaluated both models by the BLEU metric and we did not do human evaluations. As shown in Tabel 1., with the same learning rate (lr = `2e-4`) and close batch size (32 for Transformer Baseline and 24 for t5-base), the bleu of the t5-base pre-trained model can achieve 33.27 while the Transformer Baseline is 29.33. It has improved by more than 13%.

As shown in Fig. 3., the bleu of t5-base pretrained model reaches 25 within only 50k steps, while the Transformer Baseline model took about 80k steps. This proves the pre-trained model's ability to achieve high accuracy on the downstream task with relatively low data and training time. This is due to it already learnt much about the English to German translation in their previous training.

Next, we compared the impact of learning rate to the t5-base pre-trained model with batch size set to be 24. As shown in Fig. 4., the bleu of t5-base pre-trained model for learning rates of 2e-4, 3e-4, and 5e-4 are 33.27, 32.87, and 32.72. There is only a slight improvement when decreasing the learning rate.

Last, we compared the impact of batch size to the t5-base pre-trained model with a learning rate set to be 2e-4. As shown in Fig. 5., the bleu of the t5-base pre-trained model for batch size of 4, 8, 16, and 24 are 0.12, 7.95, 25.37, and 33.27. The t5-base pre-trained model cannot learn the relationship between sentences when batch size is set to be 4. And its performance improved significantly after increasing the batch size. However, due to the GPU limitation, we cannot test it with batch size larger than 24 but based on the trend line, we can see the bleu can be further improved when we use a larger batch size.

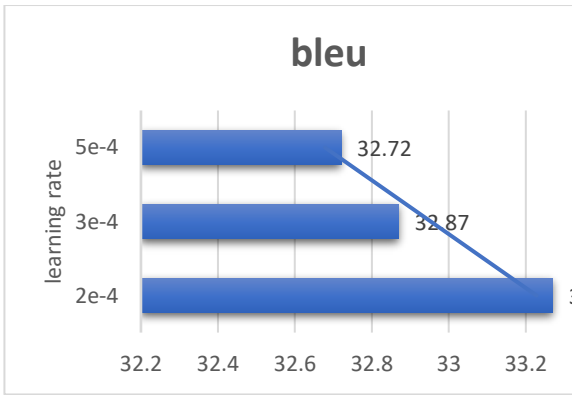

Figure 4: learning rate vs bleu.

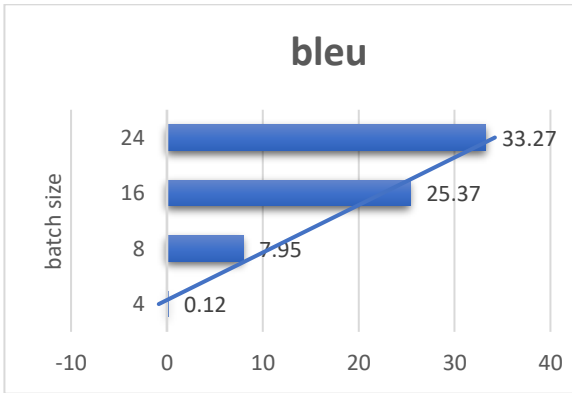

Figure 5: batch size vs bleu.

4 Conclusion and Future Works

In this paper, we compared the performance of the Transformer Baseline and the t5-base pre-trained model in doing machine translation of the WMT14 English-German Translation Data with further preprocessing. We observe significant improvements in bleu, the speed to achieve convergence, and fewer data to be used for training by using the t5-base pre-trained model. We also found that for training t-base pre-trained model, the learning rate has much less impact on the bleu than the batch size.

Due to the time and resource limitations, there were some future works we want to propose. First, we want to get more GPUs to train the t5-base pre-trained model with a larger batch size to see the best bleu we can achieve with it. Second, we encounter an issue when using the bert-base-uncased pre-trained model to replace the encoder in the Transformer Baseline, which causes 0 bleu. We'd like to solve it in future work. And we want to train our model with more datasets and translate other languages.

References

lya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of NIPS, pages 3104–3112.

WMT14 English-German Translation Data with further preprocessing, 2016. https://huggingface.co/datasets/stas/wmt14-en-de-pre-processed.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In Journal of Machine Learning Research 21, pages 1-67, 2020.

Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014). to appear.