A Simple German-English Translation Task using Transformer

PENG Muyuan, HE Chengping
3030089799, 3030097306
Department of Electrical and Electronic Engineering
University of Hong Kong
pengmy@connect.hku.hk, hecp@connect.hku.hk

Abstract

In recent years, the Transformer model has emerged as a powerful tool for various natural language processing tasks, including machine translation. This paper presents a comprehensive investigation into the performance of different Transformer-based approaches for translation tasks. We begin by providing an overview of the Transformer architecture, highlighting its key components, such as the self-attention mechanism and positional encoding. To evaluate the performance of Transformer-based translation models, we use a three-sentence German-English translation task and give out the training and evaluation of the model.

1 Introduction

1.1 Overview

The field of machine translation has witnessed significant advancements in recent years, fueled by the advent of neural network models. Among these models, the Transformer architecture has emerged as a breakthrough approach, revolutionizing the way translation tasks are approached. The Transformer model, introduced by Vaswani et al. (2017), has demonstrated remarkable performance in capturing long-range dependencies and effectively translating between different languages.

Traditional statistical machine translation methods, such as phrase-based and hierarchical models, often struggled with capturing complex linguistic patterns and handling long-distance word dependencies. In contrast, the Transformer model overcomes these limitations by leveraging the self-attention mechanism, which allows it to attend to different parts of the input sequence during the translation process. This attention mechanism enables the model to capture context and dependencies more effectively, leading to improved translation quality.

The success of the Transformer model has spurred further research and development in the area of neural machine translation. Several variations and extensions of the original Transformer architecture have been proposed, each aiming to enhance translation performance in different ways. These variants include models such as BERT, GPT, and XLNet, which have leveraged the Transformer's core components while incorporating additional techniques such as pretraining and masked language modeling.

1.2 Architecture of Transformer

The Transformer architecture consists of two main components: the encoder and the decoder. The encoder processes the input sequence, typically the source language, while the decoder generates the corresponding output sequence, typically the target language translation. Both the encoder and decoder are composed of multiple layers, each containing sub-layers that perform specific operations.

NeurIPS 2023 Generative AI and Biology Workshop.

At the heart of the Transformer model lies the self-attention mechanism. This mechanism allows the model to weigh the importance of different words in the input sequence, considering their relevance to each other. By attending to different parts of the input sequence, the model can capture dependencies and contextual information more effectively than traditional recurrent neural networks.

In the encoder, self-attention is applied to the input sequence in parallel. Each word in the sequence attends to all other words, allowing the model to learn relationships and dependencies between words regardless of their positions. This parallel processing significantly speeds up training and inference compared to sequential models.

To incorporate positional information into the model, the Transformer uses positional encoding. This technique assigns a unique positional embedding to each word in the input sequence, representing its position relative to other words. This positional encoding enables the model to understand the sequential order of words, which is crucial for tasks such as machine translation.

In addition to self-attention and positional encoding, the Transformer model also employs feed-forward neural networks within each layer. These networks consist of multiple fully connected layers with a non-linear activation function, enabling the model to learn complex patterns and representations from the input sequence.

The decoder component of the Transformer follows a similar structure as the encoder but incorporates an additional attention mechanism. This attention mechanism enables the decoder to attend not only to the input sequence but also to the previously generated output sequence. This attention mechanism is crucial for producing accurate and coherent translations by aligning the source and target languages.

The Transformer architecture also includes residual connections and layer normalization, which help alleviate the vanishing gradient problem and improve the stability and convergence of the model during training. The residual connections allow gradients to flow more easily through the network, facilitating the learning process.

2 Related Works

2.1 Attention Mechanism: Transformer[1]

The paper "Attention is all you need" introduced the Transformer architecture, which uses self-attention mechanisms to process sequences of data, such as natural language. The Transformer achieved state-of-the-art performance in several natural language processing tasks, while being more computationally efficient than previous models that relied on recurrent neural networks. The paper's contributions have greatly influenced the development of modern language models, including BERT and GPT-2.

2.2 ReFormer: An Efficient Transformer[2]

From the survey on Transformer model, it was found that the attention weight map is always sparse. This means that most of the vector-vector dot products make little contribution to the token-mixer. Considering dramatic calculations in Transformer model, the author proposed to use LSH to filter vectors that make little contributions to the model. This paper reduce steps of attention calculation from $O(L^2)$ to $O(L)$.

2.3 Vision Transformer[3]

Vision Transformer, or ViT for short, is a deep learning model that uses a transformer architecture to process visual data. It's a type of neural network that can analyze images and extract important features, allowing it to perform tasks like object recognition and image classification. ViT has shown promising results in computer vision tasks and has the potential to be used in various applications, such as self-driving cars and medical imaging.

3 Methods

In this report, we use a Transformer model with 6 encoder and 6 decoder to implement a two-sentence translation task from German to English.

For the attention mechanism, we choose multihead attention to make it easier for batch proceeding.

Here we use PyTorch architecture to implement the Transformer model.

4 Experiments

There are mainly several parameters including embedding size, dimension of K and V and number of heads. After several temps, we decide the parameters as embedding size is 256, dimension of K and V is 64 and number of heads is 8.

Here is the training result of the network, we can find that the loss will be very low after about 100 epochs.

For the test result, we can find that the two sentences are all well translated. This means that our network is of good effect.

5 Conclusions

In this work, we have successfully implemented the Transformer model in a German-English translation task. The Transformer's unique architecture, which includes attention mechanisms and self-attention layers, has been shown to be highly effective in processing natural language and has produced impressive results in machine translation.

Through the careful tuning of hyperparameters, our implementation of the Transformer model achieved a high level of accuracy and fluency in translating German sentences into English.

Overall, this work demonstrates the power and potential of the Transformer model in advancing the field of machine translation and cross-lingual communication. With continued research and development, the Transformer model could become a key tool in breaking down language barriers and facilitating global communication.

6 Figures

Note that figures related to the project is in the PDF file called 'Figures.pdf'

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient transformer[J]. arXiv preprint arXiv:2001.04451, 2020.

[3] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.