CISC7021 - Applied Natural Language

Group 48

Review Report, 2023

1 Introduction

Machine Translation (MT) is a task aimed at automatically translating text from one language to another, playing a crucial role in globalization and cross-cultural communication. With the rapid advancement of information technology and the driving force of globalization, the demand for machine translation has been increasing as it can overcome language barriers and enhance communication efficiency among people.

Before the emergence of Neural Machine Translation (NMT), traditional rule-based machine translation methods, also known as Statistical Machine Translation (SMT), were widely used. SMT employs language models to capture the probability distribution of words or phrases in both the source and target languages. These models estimate the likelihood of generating a specific translation in the context of the source sentence. SMT has achieved success in many translation tasks and has made valuable contributions to the field. However, it also has limitations, such as difficulties in handling long-range dependencies and capturing complex linguistic phenomena.

Neural Machine Translation (NMT) has emerged as the current mainstream approach in machine translation, showcasing significant progress. Unlike traditional rule-based machine translation methods, NMT utilizes neural network models to directly model the mapping relationship between source language sentences and target language sentences, enabling an end-to-end translation process.

NMT directly maps source language sentences to target language sentences using a neural network model. It learns the complex mapping relationship between the source and target languages through end-to-end training, without relying on traditional rule-based or statistical methods. The encoder-decoder structure of NMT consists of an encoder and a decoder. The encoder converts the source sentence into a fixed-length vector representation, capturing its semantics and contextual information. The decoder generates the translation based on the encoder's semantic representation. It predicts each word or character step by step, combining the context information of previous predictions. This process continues until a complete target language sentence is generated. NMT's end-to-end modeling approach enables it to capture semantic and contextual information more effectively, resulting in more accurate and fluent translation.

In general, SMT may encounter difficulties when dealing with long sentences. Since SMT primarily relies on phrase or word-level alignments, the translation quality of SMT may not meet expectations for sentences with long-range dependencies. On the other hand, NMT performs better in handling long sentences. Due to its end-to-end modeling using neural network models, NMT can better capture long-distance semantic relationships, thus improving translation quality.

This project aims to reproduce and evaluate the Gated Recursive Convolutional Neural Network, an NMT model based on an encoder-decoder architecture with gate mechanisms for selective information storage and forgetting. The project aims to improve the model in terms of accuracy, model size, or runtime. Approaches for improvement may include incorporating attention mechanisms, introducing skip connections, and modifying the RNN model structure.

2 Experimental Design

In this section, we will describe the experimental design that we have currently established, along with the rationale behind it and the expected outcomes.

2.1 Dataset

The original dataset used in the paper is a French-English dataset. It was compiled by combining data from multiple sources, including Europarl (61 million words), news commentaries (5.5 million words), the United Nations (4.21 billion words), and two web-crawled corpora (90 million words and 78 million words), resulting in a total of 348 million sentence pairs.We will replicate the training and testing process using the same French-English dataset as mentioned in the original paper. This step aims to validate the results reported in the paper using the identical experimental setup. Following the replication, we will acquire a Chinese-English dataset from the First Conference on Machine Translation (WMT16). This dataset will be used to train and test the model, allowing us to evaluate its generalizability and applicability to Chinese-English translation tasks. By including both the replication of the original dataset and the expansion to a Chinese-English dataset, we aim to assess the model's performance across different language pairs and evaluate its suitability for both French-English and Chinese-English translation tasks.

2.2 Replication

In this step, we will replicate the Gated Recursive Convolutional Neural Network mentioned in the paper. It utilizes a recursive recurrent neural network and introduces gated GRU (Gated Recurrent Unit) units.

Recursive Neural Network is effective in processing the structure and grammatical information of sentences. By recursively combining word and phrase representations, they can capture the hierarchical structure and dependency relationships within sentences, thereby enhancing the understanding of sentence meaning. The structure is illustrated in Figure 1.

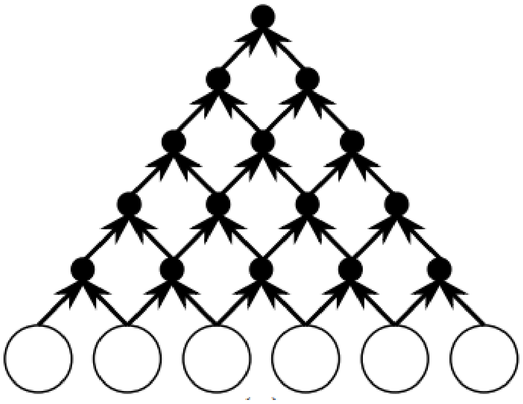

Figure 1: The recursive convolutional neural network

GRU (Gated Recurrent Unit) introduces update gates and reset gates, which enable better capturing of long-term dependencies. The update gate determines whether the current input should be remembered, while the reset gate determines how to utilize the historical information. This enhances the modeling capability of GRU in handling long sequence tasks and helps alleviate the vanishing gradient problem, making training more stable and efficient on long sequences. The structure of GRU is illustrated in Figure 2.

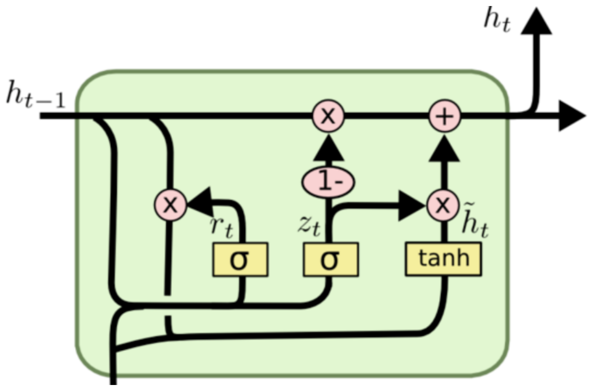

Figure 2: The GRU

2.3 Model Modification and Enhancement

2.3.1 Attention mechanism

The attention mechanism allows the model to focus on relevant parts of the input sequence while ignoring irrelevant parts. This helps the model capture important information and improve overall performance. For long sequence data, the attention mechanism can automatically learn and emphasize key information, reducing the model's reliance on the entire sequence and improving its ability to handle long sequences. By using the attention mechanism, the model can dynamically adjust attention weights based on the relevance of the context, leading to a better understanding of the input sequence. This enhances the model's ability to comprehend semantics and context, thereby improving its performance in natural language processing tasks. The principle of the attention mechanism is illustrated in Figure 3.

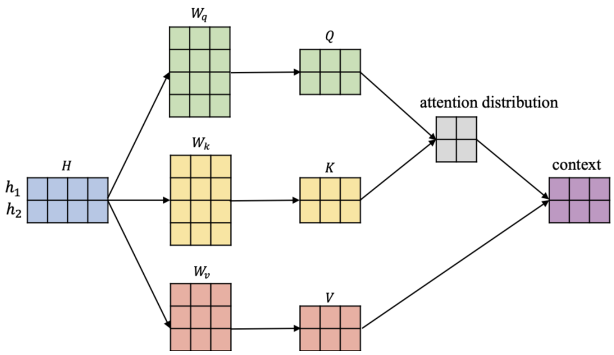

Figure 3: Attention mechanism

Regarding the attention mechanism, we considered conducting ablation experiments from three perspectives: encoder self-attention, decoder self-attention, and encoder-decoder attention, to determine which one is more suitable for the model. Introducing self-attention mechanism in the encoder helps the model capture semantic relationships within the input sentence more effectively. Using self-attention mechanism in the decoder assists the model in better focusing on the generated parts during the target language sentence generation process. Introducing attention mechanism between the encoder and decoder aids the model in aligning and attending to the encoder's outputs more effectively during the decoding process.

2.3.2 Skip Connections

Skip connections provide several benefits in neural networks for natural language processing (NLP). They facilitate feature reuse by directly connecting shallow and deep layers, allowing the shallow layers to access the outputs of deep layers and leverage the higher-level features learned by them. This enhances the model's expressive power, generalization ability, and reduces the risk of overfitting. Skip connections also help alleviate the challenges of gradient propagation in deep neural networks. Deep networks often face issues such as vanishing or exploding gradients during training. By providing a direct gradient path, skip connections enable faster gradient propagation to the shallow layers, thus accelerating the training process. The principle of the skip connection is illustrated in Figure 3.

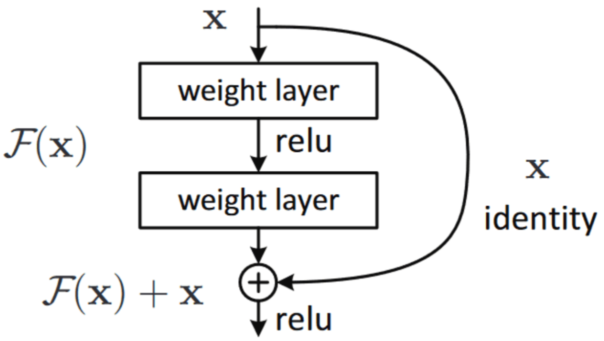

Figure 4: Skip connection

We plan to conduct experiments to select the most suitable approach for adding skip connections in the Gated Recurrent Unit (GRU). We will experiment with adding skip connections between the input and output of the GRU and between different time steps of the GRU.Adding skip connections between the input and output of the GRU allows for more direct information propagation between the recurrent units, promoting the flow of gradients and features. On the other hand, adding skip connections between different time steps of the GRU facilitates information propagation across the temporal dimension, enabling the model to capture long-term dependencies.By conducting these experiments, we aim to determine which approach better enhances the GRU model's performance.

2.3.3 Variants of GRU

We plan to experiment with different variants of GRU to address the translation model's performance. For instance, when encountering a low BLEU score, we can consider using stacked GRU, which involves stacking multiple GRU layers to form a deep network and enhance the model's expressive power and learning capacity. On the other hand, when facing low computational efficiency, we can explore efficient GRU approaches such as matrix decomposition and low-rank approximation, and analyze their impact on GRU performance and computational efficiency.

2.3.4 Depthwise Separable Convolution

By replacing convolutional layers with depthwise separable convolutions in the model, the number of parameters can be reduced, thereby improving the model's computational efficiency. Depthwise separable convolution first applies independent convolutions to each input channel, known as depthwise convolution. Then, pointwise convolution is applied to the convolutional results of each channel, using a 1x1 convolutional kernel to fuse information between channels. Ultimately, by combining these two steps, the final convolutional output is obtained. The structure is shown in Figure 5.

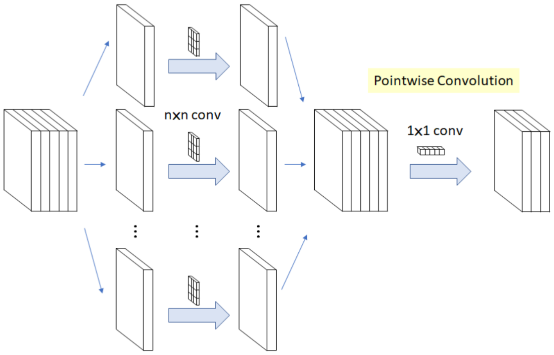

Figure 5: Depthwise Separable Convolution

2.3.5 Model Architecture

It is within our consideration to replace the model architecture. The original paper used a recursive neural network (RNN) as the main structure of the model, following an encoder-decoder pattern. We are also considering other encoder-decoder based models such as U-Net and V-Net.

2.4 Model Evaluation

We will conduct testing using both the original paper's French-English dataset and a new Chinese-English dataset. While the original paper employed BLEU as the evaluation metric, the authors expressed some dissatisfaction with it. Therefore, we will utilize BLEU, TER, METEOR, as well as subjective ratings for a comprehensive evaluation of the model.