Group-attention Based Neural Machine Translation

Renyu Hu¹, Hao Xu¹, Yang Xiao¹, Chenjun Wu¹ and Haiyang Jia¹,² *,

¹College of Computer Science and Technology, Jilin University Changchun 130012, China
²Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

*Corresponding author e-mail: jiahy@jlu.edu.cn

Abstract. Machine translation is a classic problem in natural language process (NLP). Recent years, the encoder and decoder through an attention mechanism has become a trend. Google proposed a new simple network architecture, the Transformer using attention mechanisms only rather than CNN or RNN in 2017. However, it may lose some important information (e.g., grammatical component, etc) when using attention mechanism for whole sentence. We propose a new brand model based on transformer using Group attention layers and group position embedding. It absorbs the features of Group-CNN combines the algorithm in computer vision (CV) and NLP. The model not only pays more attention to the ingredients (e.g., subject, predicate and adverbial, etc), but also enhances the connection of phrase. It outperforms SofA Transformer in using more syntactic information.

1. Introduction

In 2013, paper, Recurrent Continuous Translation Models [3], brought a encoder-decoder model. The model uses CNN to transform source text into a intermediate vector, then through recurrent neural network (RNN) into target text. After this sequence-to-sequence (seq2seq), encoder-decoder architecture was brought into public eyes. In following years, more models based on this architecture tried to be more efficient in decoding or encoding (e.g., GRU, LSTM, etc). As far as 2017, Facebook Artificial Intelligence Research (FAIR) brought a brand-new CNN approach for NMT that achieved accuracy at nine times the speed of recurrent neural systems. The model advanced the whole encoder-decoder networks by using CNN. It marked the NMT based on deep learning networks obtained a new high level.

However, just few months later, Google proposed a new simple network architecture, the Transformer using attention mechanisms only rather than CNN or RNN. Although the Transformer can be trained faster, they might lose some information in the structure of the sentence. In figure 1, all the previous target words are required to predict the current target word \( y_j \) in Transformer in contrast to the RNN, CNN networks. The ingredients (e.g., subject, predicate and adverbial, etc) and the connection of phrases may be ignored by just using attention mechanism for all the words.

We combine the Group-CNN in CV with the Transformer to build new model called Group-attention Based Neural Machine Translation (GANMT). It absorbs the features of Group-CNN combines the algorithm in computer vision (CV) and NLP. The model not only pays more attention to...
the ingredients (e.g., subject, predicate and adverbial, etc), but also enhances the connection of phrase. It outperforms SofA Transformer in using more syntactic information.

Figure 1. [4] Illustration of the decoding procedure under different neural architectures.

2. Background

2.1. Attention mechanism [5]
The seq2seq models is normally composed of an encoder-decoder architecture. The encoder processes the source material and encodes it into a fix-length vector. This vector is used to be a presentation of the input. The decoder is then initialized with this context vector, using which it starts generating the transformed output.

An obvious disadvantage of architecture is it doesn’t fully fit longer sequences. Often is has forgotten the earlier parts of the sequence once it has processed the entire the sequence. The attention mechanism was born to resolve this problem.

Unlike the fixed context vector used for all the decoder time steps in case of the traditional Seq2Seq models, here in case of Attention, we calculate a context vector for each step by computing attention weights every time. As the name, attention mechanism aims to pay different attention to the words.

2.2. Transformer [1]
Encoder-Decoder is a very common model framework for deep learning, especially in NMT. The whole structure proposed by Google called Transformer maps the source text sequence of $X=(x_1,...x_n)$ to fixed-length context $Z=(z_1,...,z_n)$. Then focusing on the sequence of $Z$, decoder can generate target text sequence of $Y=(y_1,...,y_n)$.

Figure 2. [1] Mechanism of the Transformer.
3. Model Architecture
In this part, we will show the advanced feature in detail. Our main improvements lie in two parts: group position embedding and group-attention layers.

3.1. Group position embedding
There are three steps in this process: (1) Self-position encoding; (2) Group-position encoding; (3) Add up.

   (1) Self-position encoding
   In the paper “Attention is all you need” [1], they use sin and cos functions to mark different positions:

   \[
   PE(pos, 2i) = \sin(pos / 10000^{d_{\text{model}}/d_{\text{model}}}) \\
   PE(pos, 2i + 1) = \cos(pos / 10000^{d_{\text{model}}/d_{\text{model}}})
   \]

   The positional encodings have the same dimension \( d_{\text{model}} \) as the embeddings, so that the two can be summed. We chose the same function not only to allow the model to make full use of relative positions but also to compare under the same conditions.

   (2) Group-position encoding
   Number the label of position. We use the following formula to calculate the label:

   \[
   S / n_{\text{group}} + 1
   \]

   Where \( S \) means the sequence number, \( n_{\text{group}} \) is the number of components like subject, predicate and adverbial. We generally define \( n_{\text{group}} \) as 3 for the sentence usually includes three parts of subject, predicate and adverbial. Then, use the previous part’s formula of sine and cosine function to do the group-position encoding.

   (3) Add up
   To combine the individual feature and position feature together, you can easily add them two up using:

   \[
   w \cdot PE_1 + (1 - w) \cdot PE_2
   \]

   Where \( w \) you can train from a very small value around 0.

3.2. Group-attention layers
Two sub-layers in my group-attention layers:
- Self-attention layers
- Group-attention layers

   We initialize the whole network using the same method in [1]. Group-attention layers are the core of the network. We are inspired by the concept in CV of group-CNN. Just like group-position embedding, we equally use the position information to deferential each word. We should also define a value \( n \) to mark the count of words in a group.

   Etc: Listening to music is so relaxed especially rap music (define \( n = 2 \))

   (1) Double \( n \) to 4, then let listening, to, music in group 1, is, so, relaxed, especially in group 2, rap, music in group 3.

   (2) In each group, let self-attention run.

4. Conclusion
We propose a new model improving transformer by Group attention layers and group position embedding. It absorbs the features of Group-CNN [2] combines the algorithm in computer vision (CV) and NLP. The model not only pays more attention to the ingredients (e.g., subject, predicate and adverbial, etc), but also enhances the connection of phrase. It outperforms SofA Transformer in using more syntactic information.
Acknowledgements
This paper is supported by National Natural Science Foundation of China under Grant Nos. 61502198, 61472161, 61402195, 61103091 and the Science and Technology Development Plan of Jilin Province under Grant No. 20160520099JH. 20190302117GX, 20180101334JC; Innovation Capacity Construction Project of Jilin Province Development and Reform Commission (2019C053-3).

Contact author: Haiyang Jia, jiahy@jlu.edu.cn, Jilin University, 2699 Ave. Qianjin, Changchun, 130012, P.R.China.

References
[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. Attention Is All You Need. In NIPS, 2017.
[2] Ting Zhang, Guo-Jun Qi, Bin Xiao and Jingdong Wang. Interleaved Group Convolutions for Deep Neural Networks. In ICCV, 2017.
[3] Nal Kalchbrenner and Phil Blunsom. Recurrent Continuous Translation Models. In EMNLP, 2013.
[4] Biao Zhang, Deyi Xiong and Jinsong Su. Accelerating Neural Transformer via an Average Attention Network. In ACL, 2018.
[5] Harshall Lamba. Intuitive Understanding of Attention Mechanism in Deep Learning. In towardsdatascience.com.