MPNet: Masked and Permuted Pre-training for Language Understanding

1 Contribution

The paper has proposed MPNet, a novel pre-training method which can be considered as a generalization of both masked language modeling (MLM) and permuted language modeling (PLM). MPNet naturally inherits the advantages of MLM and PLM and avoids their weakness.

2 Motivation

The paper compares MLM and PLM and finds their strengths and weaknesses which, in turn, leads them to introduce a novel pre-training method. MLM and PLM have been compared from two perspectives: the dependency in the predicted (output) tokens and the consistency between pre-training and fine-tuning in the input sentence.

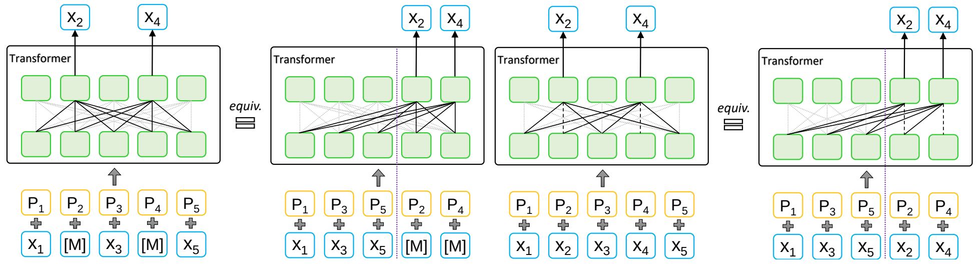
Output Dependency MLM assumes that the masked tokens are independent of each other whereas PLM avoids the independence assumption using the product rule over permuted order.

Input Consistency In MLM, although some tokens are masked, their position information (i.e., the position embeddings) are available to the model to (partially) represent the information of full sentence. However, each predicted token in PLM can only see its preceding tokens in a permuted sentence but does not know the position information of the full sentence.

Thus, PLM is better than MLM in terms of leveraging output dependency while worse in terms of pre-training and fine-tuning consistency. The finding led the authors to introduce MPNet which retains the advantages of both the methods while avoiding their weaknesses.

3 The Approach

To address the issues and inherit the advantages of MLM and PLM, the work first provided a unified view to understand MLM and PLM. The unified view


Transformer
equiv.
x_2
x_4
P_1 P_2 P_3 P_4 P_5
x_1 [M] x_3 [M] x_5
P_1 P_3 P_5 P_2 P_4
x_1 x_3 x_5 [M] [M]
x_1 x_2 x_3 x_4 x_5
x_1 x_3 x_5 x_2 x_4

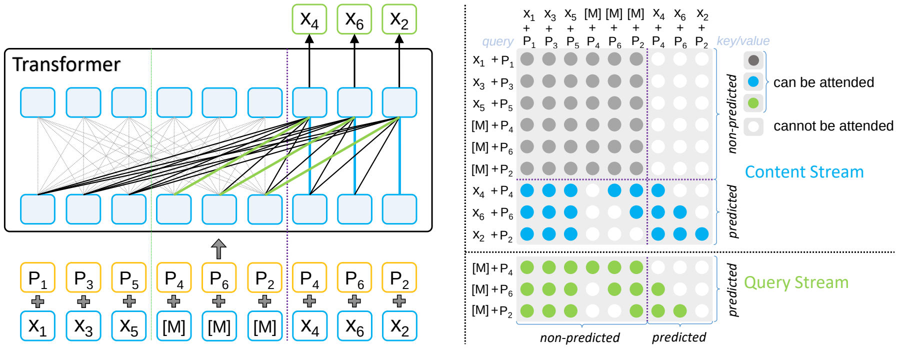
(a) MLM (b) PLM

Figure 1: A unified view of MLM and PLM, where x_i and p_i represent token and position embeddings. The left side in both MLM (a) and PLM (b) are in original order, while the right side in both MLM (a) and PLM (b) are in permuted order and are regarded as the unified view.

Figure 1: The unified view

rearranges and splits the tokens into non-predicted and predicted parts, as illustrated in Figure 1.

From the unified view, the work has proposed MPNet (as shown in Figure 2) which can be considered as a generalization of both MLM and PLM.



(a) (b)

Figure 2: (a) The structure of MPNet. (b) The attention mask of MPNet. The light grey lines in (a) represent the bidirectional self-attention in the non-predicted part $(x_{z_{<=c}}, M_{z_{>c}}) = (x_1, x_5, x_3, [M], [M], [M])$, which correspond to the light grey attention mask in (b). The blue and green mask in (b) represent the attention mask in content and query streams in two-stream self-attention, which correspond to the blue, green and black lines in (a). Since some attention masks in content and query stream are overlapped, we use black lines to denote them in (a). Each row in (b) represents the attention mask for a query position and each column represents a key/value position. The predicted part $x_{z_{>c}} = (x_4, x_6, x_2)$ is predicted by the query stream.

Figure 2: The structure and the attention mask of MPNet.

4 Results

For empirical evaluation MPNet has been pre-trained on a 160 GB dataset and fine-tuned of several down-stream task like GLUE, SQuAD etc. The experimental results show that MPNet performs better than other pre-trained methods like BERT, XLNet, RoBERTa under similar model setting.