MPNet: Masked and Permuted Pre-training for Language Understanding

Link: https://arxiv.org/pdf/2004.09297v2.pdf

BERT neglects the dependency among predicted and masked tokens and assumes independence, but in natural language, context has complex dependencies.

XLNet doesnt completely leverage position and suffers from discrepancy.

- PLM is like unscrambling the words, doesnt depend on order of words and understands the sentence. Rightmost 15% tokens as the predicted tokens.
- It predicts sentences with different orders factors in how order of words would affect the meaning and how meaning of a word can be influenced by the words that come before or after it.

Model must have knowledge of the entire sentence, in MLM, it has minimal knowledge like length of sentence, while PLM only knows about preceding tokens and lacks knowledge of full sentence.

Unified View of MLM and PLM

- In MLM there is shuffling and masking, while in PLM there is shuffling and predicting, essentially masking and predicting of a token are the same.
- Training objective of MPNet is sum of all tokens from position t to n, that is logarithm of the probability of predicting a specific token xzt. The probability depends on the tokens that come before it (xz<=c), the masked tokens (Mz>c), and some parameters represented as θ

Proposed Methodology

- Input Tokens and Positions: Permute a sequence, and mask tokens before predicted part, so (x1, x3, x5, M, M, M). becomes non predicted and (x4, x6, x2) becomes predicted, which makes the sequence (1,3,5,4,6,2,4,6,2), non predicted is for bidirectional modelling, understanding relations.
- **Modeling Output Dependency with Two-Stream Self-Attention:** it is difficult to predict tokens because of permuted manner, we can use two stream self attention,

    - Query stream, previous token, positions, and current position, not current token,
    - Content stream, previous tokens, positions and current token.
    - we get more informed predictions, with all the information like context of previous token and content of current token.
    - but doesn't have information of full sentence. for predicting x6, it has info of x1,x3,x5,x4 and position info of all these tokens, but doesnt have info of p6 and p2.
- **Reducing Input Inconsistency with Position Compensation**
    - the attention masks always sees a fixed number of tokens (original sequence length) and helps in predictions.
    - the authors add position compensation by adding masked tokens M6 and M2 and P6 and P2 to ensure query has access to full information.
    - this helps model in generating predictions by considering the entire sentence info and permuted order.

> 💡 An example sentence is "the task is sentence classification" so MPNet = log *P*(sentence | the task is [M] [M]) + log *P*(classifification | the task is sentence [M])

- MLM consider 85% of tokens and 100% of positions,
- PLM considers 85% of unmasked tokens and positions, along with 7.5% of the masked tokens and positions.
- MPNet has minimal discrepancy since it has access to the entire positional information.

During finetuning, no query stream is used, the model is pretrained using weights from RoBERTa large.

Albation Study

- with no positional compensation, it becomes PLM, scores in DS tasks drop by 0.6-2.3
- with no permuted operations, it becomes MLM + output dependency, and scores drop by 0.5-1.7

- with no permuted operations and output dependency it becomes MLM, scores drop by 0.5-3.7