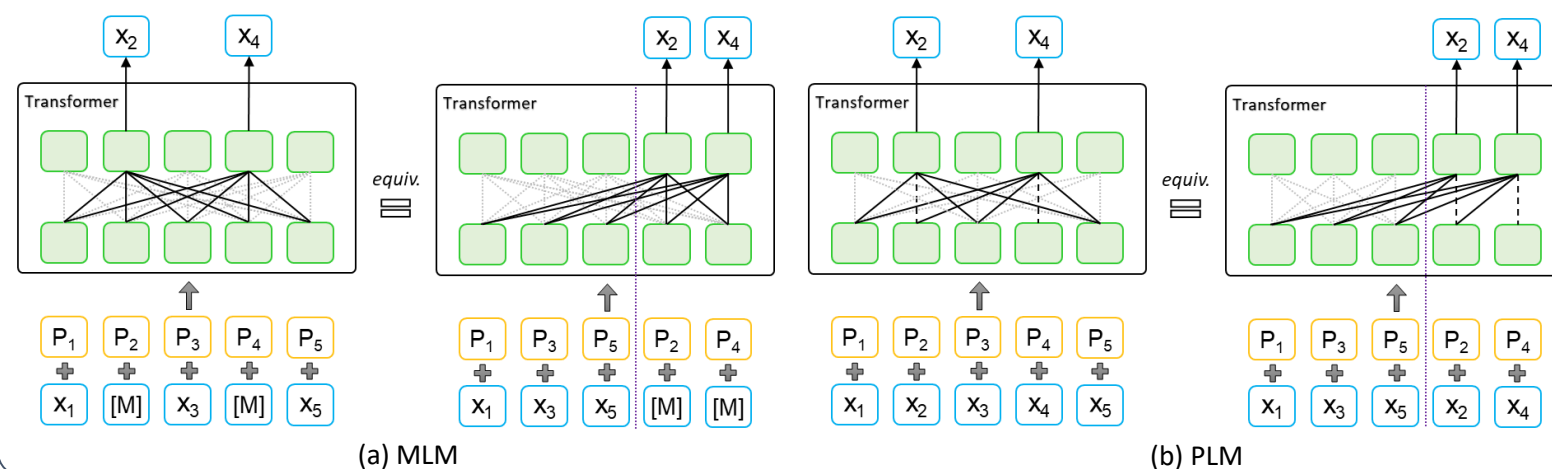




## Motivation

- BERT** (Masked Language Model, **MLM**) allows model to see the full input sentence (*input consistency*), but ignores the dependency between the masked/predicted tokens (*output dependency*).
- XLNet** (Permuted Language Model, **PLM**) leverages dependency between masked/predicted tokens (*output dependency*) but can only see its preceding tokens rather than the full sentence (*input consistency*).



## Method

### MPNet – Inherit advantages of MLM and PLM

- Autoregressive prediction** (avoid the limitation in BERT)
  - Each predicted token condition on previous predicted tokens to ensure *output dependency*
- Position compensation** (avoid the limitation in XLNet)
  - Each predicted token can see full position information to ensure *input consistency*

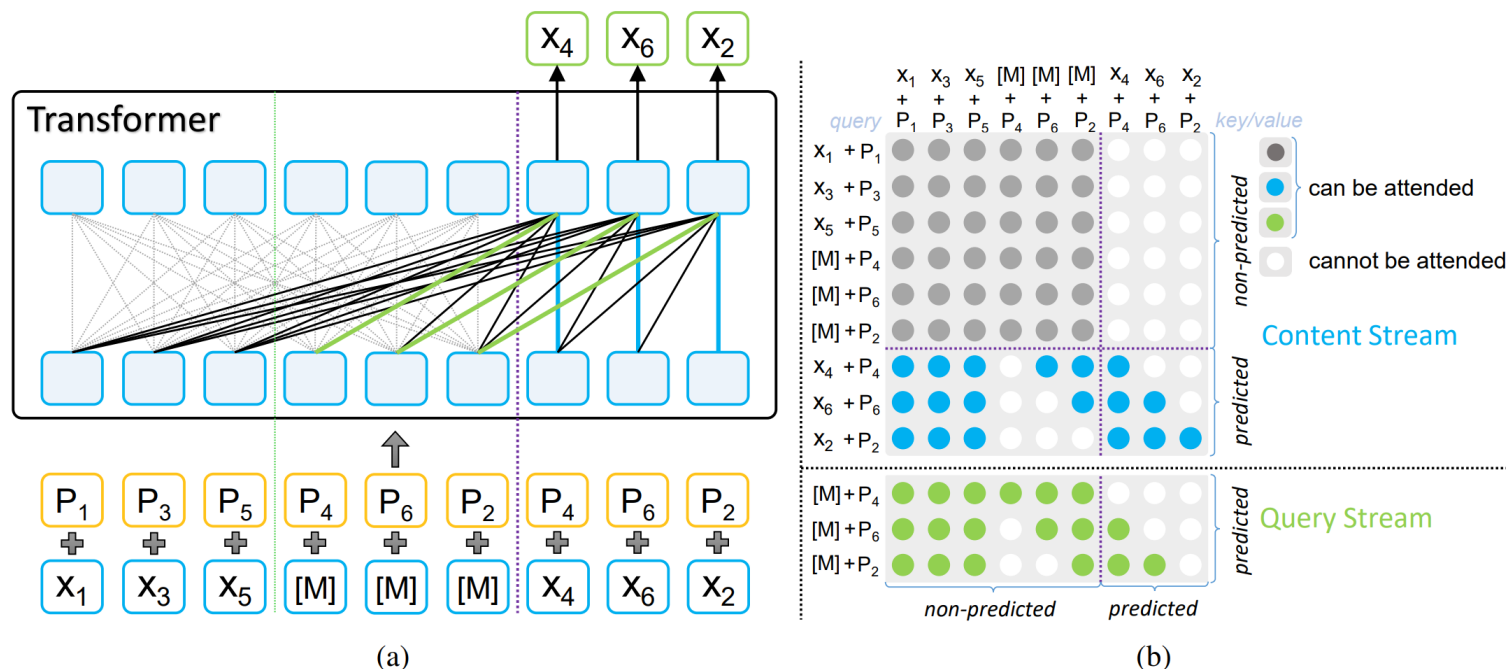


Fig 1. Architecture of MPNet

## Analysis

### Case Study (Factorization)

| Objective   | Modeling                                                                                                                           |
|-------------|------------------------------------------------------------------------------------------------------------------------------------|
| MLM (BERT)  | $\log P(\text{sentence}   \text{the task is } [M] [M]) + \log P(\text{classification}   \text{the task is } [M] [M])$              |
| PLM (XLNet) | $\log P(\text{sentence}   \text{the task is } ) + \log P(\text{classification}   \text{the task is } \text{sentence})$             |
| MPNet       | $\log P(\text{sentence}   \text{the task is } [M] [M]) + \log P(\text{classification}   \text{the task is } \text{sentence } [M])$ |

### Comparisons between MPNet and MLM/PLM (Tokens and Positions)

| Method      | Formulation                                                                      | # Tokens | # Positions |
|-------------|----------------------------------------------------------------------------------|----------|-------------|
| BERT (MLM)  | $E_{z \in Z_n} \sum_{t=c+1}^n \log P(x_{z_t}   x_{z \leq c}, M_{z > c}; \theta)$ | 85%      | 100%        |
| XLNet (PLM) | $E_{z \in Z_n} \sum_{t=c+1}^n \log P(x_{z_t}   x_{z \leq t}; \theta)$            | 92.5%    | 92.5%       |
| MPNet       | $E_{z \in Z_n} \sum_{t=c+1}^n \log P(x_{z_t}   x_{z \leq t}, M_{z > c}; \theta)$ | 92.5%    | 100%        |

## Experiments

- Setting:** Pre-train on English Wikipedia, BooksCorpus, OpenWebText, Stories (160GB) with a batch size of 8192 for 500K Steps. The mask ratio is set as 15%.
- GLUE**

|                              | MNLI | QNLI | QQP  | RTE  | SST  | MRPC | CoLA | STS  | Avg  |
|------------------------------|------|------|------|------|------|------|------|------|------|
| Single model on dev set      |      |      |      |      |      |      |      |      |      |
| BERT (Devlin et al., 2019)   | 84.5 | 91.7 | 91.3 | 68.6 | 93.2 | 87.3 | 58.9 | 89.5 | 83.1 |
| XLNet (Yang et al., 2019)    | 86.8 | 91.7 | 91.4 | 74.0 | 94.7 | 88.2 | 60.2 | 89.5 | 84.5 |
| RoBERTa (Liu et al., 2019a)  | 87.6 | 92.8 | 91.9 | 78.7 | 94.8 | 90.2 | 63.6 | 91.2 | 86.4 |
| MPNet                        | 88.5 | 93.3 | 91.9 | 85.8 | 95.4 | 91.8 | 65.0 | 91.1 | 87.9 |
| Single model on test set     |      |      |      |      |      |      |      |      |      |
| BERT (Devlin et al., 2019)   | 84.6 | 90.5 | 89.2 | 66.4 | 93.5 | 84.8 | 52.1 | 87.1 | 79.9 |
| ELECTRA (Clark et al., 2020) | 88.5 | 93.1 | 89.5 | 75.2 | 96.0 | 88.1 | 64.6 | 91.0 | 85.8 |
| MPNet                        | 88.5 | 93.1 | 89.9 | 81.0 | 96.0 | 89.1 | 64.0 | 90.7 | 86.5 |

### SQuAD

| Method      | SQuADv1.1 (dev) | SQuADv2.0 (dev) | SQuADv2.0 (test) |
|-------------|-----------------|-----------------|------------------|
| BERT [2]    | 80.8/88.5       | 73.7/76.3       | 73.1/76.2        |
| XLNet [5]   | 81.3/-          | 80.2/-          | -/-              |
| RoBERTa [7] | 84.6/91.5       | 80.5/83.7       | -/-              |
| MPNet       | 86.9/92.7       | 82.7/85.7       | 82.8/85.8        |

### Ablation Study

| Model Setting                             | SQuADv1.1 | SQuADv2.0 | MNLI | SST-2 |
|-------------------------------------------|-----------|-----------|------|-------|
| MPNet                                     | 85.0/91.4 | 80.5/83.3 | 86.2 | 94.0  |
| – position compensation (= PLM)           | 83.0/89.9 | 78.5/81.0 | 85.6 | 93.4  |
| – permutation (= MLM + output dependency) | 84.1/90.6 | 79.2/81.8 | 85.7 | 93.5  |
| – permutation & output dependency (= MLM) | 82.0/89.5 | 76.8/79.8 | 85.6 | 93.3  |

Code: <https://github.com/Microsoft/MPNet>