XLNet: Generalized Autoregressive Pretraining for Language Understanding

1 Objective

The paper has proposed a novel pre-training method, namely XLNet, which overcomes the following drawbacks of the existing pre-training methods like BERT and GPT:

1. Autoregressive pre-trainings like GPT represent the output at time-step t as a function of the tokens up to time step $t-1$, thus they are unable to model bidirectional context.

2. Since BERT like pre-training corrupts the input with an additional mask token which is not present in fine-tuning data, it suffers from input-output discrepancy.

3. Denoising autoencoding based pre-trainings like BERT, though model the bidirectional context, ignore the dependency between the mask tokens.

XLNet overcomes the above drawbacks by using a generalised autoregressive pre-training objective which maximizes the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order (please refers to Figure 1 for illustration). The objective allows XLNet to model bidirectional context. Moreover, since XLNet does not rely on input corruptions, it does not suffer from input-output discrepancy.

2 Improvement of the XLNet Model by Incorporating Ideas from Transformer-XL

To improve the model, XLNet has incorporated two ideas from Transformer-XL

Relative Positional Encoding Instead of using absolute positional encoding, XLNet has used relative positional encoding like Transformer-XL.

Segment Recurrence In Transformer-XL, instead of discarding the hidden states after the computation of a segment, they are saved in memory. During the computation of the following segments, the self attention is

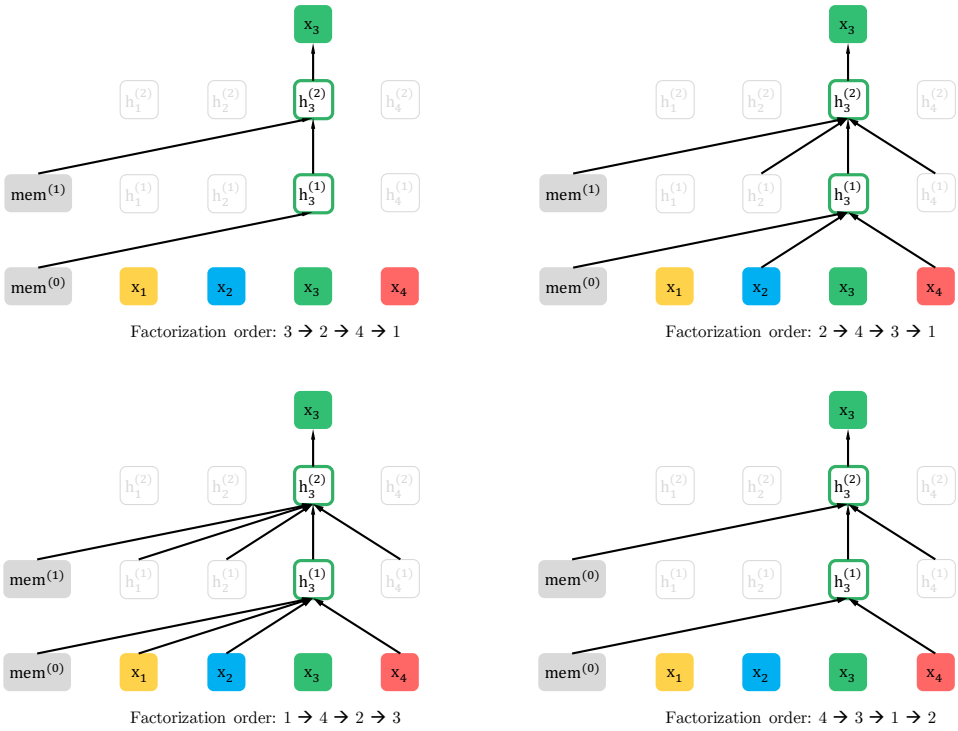

Figure 4: Illustration of the permutation language modeling objective for predicting x_3 given the same input sequence $\mathbf{x}$ but with different factorization orders.

Figure 1: Illustration of bidirectional language modeling using permutation of factorization order.

applied over the hidden states of both the current segment and the memory, thus has an increased context length. Similar mechanism has been used in XLNet.

3 Results

The paper has empirically compared XLNet with BERT pre-training. They have found that XLNet had performed better than BERT on 20 tasks under similar experimental setting.