Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning

1 Objective

The time complexity of inference step of BERT model is $\mathcal{O}\left(n^3\right)$. The paper has proposed new model to reduce the time complexity to $\mathcal{O}(n^2)$.

During training, BERT uses Masked Language Modeling (MLM) objectives which predicts the original ids of masked out words from the input. Due to the MLM objective, each contextual word representation should be computed by a two-step process:

1. masking a word in the input
2. then feeding the input into BERT

During inference, this process needs to be repeated n times - one for each word of the input of length n. Since a forward pass in BERT takes $\mathcal{O}(n^2)$ time, the time complexity of the inference process for an input sequence of length n becomes $\mathcal{O}(n^3)$. This work has proposed a novel bidirectional language model called Transformer based Text Autoencoder (T-TA) in which the complexity of the inference time reduced to $\mathcal{O}(n^2)$.

2 The Methods

To achieve $\mathcal{O}(n^2)$ inference time complexity, T-TA is designed in such a way that its output at position i, denoted by $\mathbf{o}_i$, is dependant on all but the i-th tokens of the input (with out any input masking), making it possible to compute the contextual representation of all the tokens of the input sequence in one forward pass. More concretely, let $x_1, x_2, \cdots, x_n$ be the input sequence where x_i be the i-th word/token. Let $\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_n$ be the output of T-TA for the above input, i.e.

$$\mathrm{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_n = \textbf{T-TA}(x_1, x_2, \cdots, x_n).$$

such that $\mathbf{o}_i \in R^d$ is the contextual vector representation of the input token x_i. In T-TA, the i-th output vector $\mathbf{o}_i$ depends on $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$ and does not depend on x_i for all $i = 1, 2, \cdots, n$.

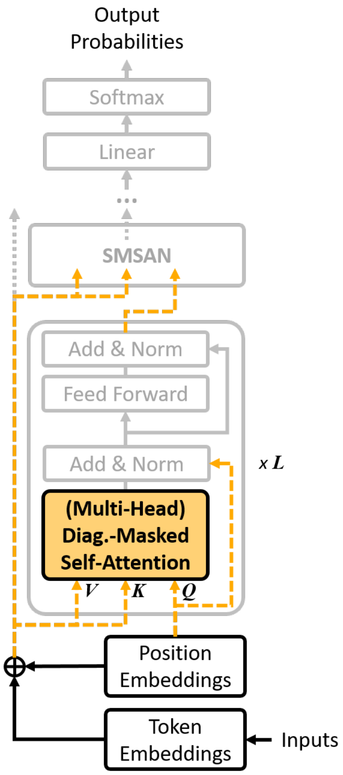

Figure 2: Architecture of our T-TA. Highlighted box and dashed arrows are newly invented in this paper.

Figure 1: Input Isolation

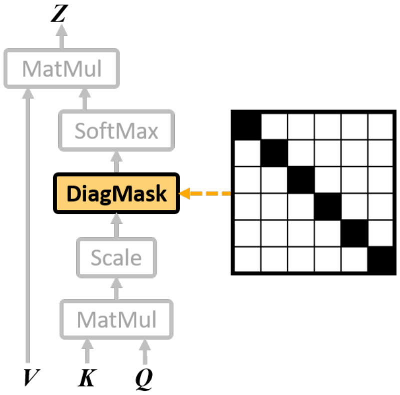

Figure 3: Diagonal masking of the scaled dot-product attention mechanism. Highlighted box and dashed arrow are newly invented in this paper.

Figure 2: Diagonal Masking

The T-TA model have achieved the above modeling goal using two modifications in a typical transformer network:

Input Isolation It isolates the key-value pair (K-V) input of all the layers from the network flow. They are kept fixed to the sum of the token embeddings and the position embeddings. Only query inputs (Q) are updated across the layers. Additionally, only the position embeddings is used as the query input Q of the first layer. It is elaborated in Figure 1.

Diagonal Masking In order to be "self-unknown" during the inference inside the scaled dot-product attention, it uses a diagonal masking as shown in Figure 2.

3 Results

The work has shown that the proposed T-TA performs over six times faster than the BERT-based model in the reranking task and twelve times faster in the semantic similarity task. Furthermore, the T-TA shows competitive or even better accuracy than those of BERT on the above tasks.