**A pluggable framework of multi-modal pre-trained model**

Loc Nguyen

Loc Nguyen’s Academic Network, Vietnam

Email: ng\_phloc@yahoo.com

Homepage: www.locnguyen.net

**Abstract**

**Keywords:**

**1. Introduction**

**2. Related works**

**3. Methodology**

**4. Experimental results and discussions**

**5. Conclusions**

**Appendices**

**A1. Training attention**

Attention is calculated based on scaled product of query matrix *Q*, key matrix *K*, and value matrix *V* in order to make effects on value matrix *V* specifying real sequence by probabilities and moreover, these probabilities are calculated by matching query matrix *Q* specifying query sequence and key matrix *K* specifying key sequence, which is similar to searching mechanism. These probabilities are also based on soft-max function, which implies weights too. Moreover, attention focuses on all tokens of sequence, which improves meaningful context of sentence in NLP. Given matrices *Q*, *K*, and *V*, attention of *Q*, *K*, and *V* is specified as follows (Vaswani, et al., 2017):

Recall that attention or self-attention is calculated based on scaled product of query matrix *Q*, key matrix *K*, and value matrix *V* along with support of soft-max function as follows:

Where matrices *Q*, *K*, and *V* are products of sequence matrix ***X*** = (***x***1, ***x***2,…, ***x****m*)*T* and query weight matrix *WQ*, key weight matrix *WK*, value weight matrix *WV* as follows:

Dimensions of matrices *Q*, *K*, and *V* are *m* x *dk*, *m* x *dk*, and *m* x *dv*, respectively. They are specified as follows:

Where,

Dimensions of weight matrices *WQ*, *WK*, and *WV* are *dm* x *dk*, *dm* x *dk*, and *dm* x *dv*, respectively. They are specified as follows:

It is easy to recognize that training attention is to learn weight matrices *WQ*, *WK*, and *WV* from data ***X*** = (***x***1, ***x***2,…, ***x****m*)*T* by stochastic gradient descent (SGD) algorithm associated backpropagation algorithm. However, the application of SGD is not too easy because of occurrence of soft-max function and so, we need to mention product *QKT* and probability matrix as follows:

Where,

Therefore,

Note,

Each probability *pij*, which is weight indeed, is calculated by soft-max function as follows:

Probability matrix is totally determined:

Where,

Self-attention of *Q*, *K*, and *V* is totally determined as follows:

Where,

Note, is the *j*th column vector of value matrix *V*. Recall that training attention is to learn weight matrices *WQ*, *WK*, and *WV* from data ***X*** = (***x***1, ***x***2,…, ***x****m*)*T* by SGD algorithm associated backpropagation algorithm. Let *A* be computed attention from weight matrices and input data ***X*** and let *A*’ be real output. Of course, *A*’ is also real attention without computation.

Squared error *ε*(*WQ*, *WK*, and *WV*) between computed attention *A* and real attention *A*’ is:

**References**

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, & S. Vishwanathan (Ed.), *Advances in Neural Information Processing Systems (NIPS 2017).* *30.* Long Beach: NeurIPS. Retrieved from https://arxiv.org/abs/1706.03762