A Tensorized Transformer For Language Modeling

Introduction

Background
At present, pre-training models play an essential role in many neural language processing tasks. However, Transformer, and its variant model BERT, limit the effective deployment of the model to limited resources. Our methods include:

- The compression of large pre-training language model has been an essential problem in NLP research.
- There are some compression methods only study the compression of embedding layers and some methods cannot be integrated into the model after compression.

Research Questions
- To linearly represent a self-attention by a group of basic vectors.
- To compress multi-head attention in Transformer.
- After compressing, it can be directly integrated into the encoder and decoder framework of Transformer.

Our Methods
- Low-rank decomposition
- Parameters sharing

- Using Tucker decomposition formulation is to construct Single-block attention.
- Using Block-term decomposition + Parameters sharing formulation is to construct multi-head mechanisms (Multi-linear attention).

Tensored Transformer

- Single-block Attention by Tucker Decomposition
  \[
  \text{Atten}_{TD}(Q, K, V) = \sum_{j=1}^{R} \sum_{m=1}^{M} \sum_{d=1}^{D} G_{j,m} Q_j \circ K_j \circ V_m
  \]

- Multi-linear Attention by Block-term decomposition
  \[
  \text{MultiLinear}(Q, Q', K', V') = \text{SplitConcat} \left( \left( \frac{1}{R} \sum_{i=1}^{R} \text{Atten}_{TD}(Q, K_j, V_j) \right) W^0 \right)
  \]

where \( T_j = \text{Atten}_{TD}(Q_j, Q'[i], K'[j], V'[k]) \)

Experimental Results

- Single-block attention can reconstruct the self-attention function by the summing over the tensor according to the second index.

\[
\text{Attention}(Q, k, V)_{i, j, m} = \sum_{i=1}^{N} \text{Atten}_{TD}(Q, K, V)_{i, j, m}
\]

- Our method is mainly experimented on three language model datasets, PTB, WikiText-103, and One-Billion, respectively. The lower the PPL, the better the model.
- Our methods achieve a more accurate result with fewer parameters.

Conclusion

- Providing a novel self-attention method, namely Multi-linear attention.
- Combining two compression ideas, parameters sharing and low-rank decomposition.
- Achieving higher compression ratio and better experimental results in language modeling.

Main References