Tensorized Transformer

- Single-block Attention by Tucker Decomposition
  
  \[ \text{Atten}_{TD}(\mathbf{g}; K, V) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbf{G}_{i,j,m} Q_{i,j} V_{k,m} \]

- Multi-linear Attention by Block-term Decomposition
  
  \[ \text{Multi-linear}(\mathbf{g}; K, V) = \text{SplitComcat} \left( \frac{1}{d} \cdot (T_1 + \ldots + T_d) \right) W^o \]
  where \( T_j = \text{Atten}_{TD}(\mathbf{g}; K, V) \).

Main Ideas

- **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)

Experimental Results

- **Our methods** achieve a more better results 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