Challenger 2: Brain Network Transformer

Jianpeng Chen

CS, Virginia Tech

Strengths

- **Data: The insight of brain net data**

1. Fully connected graph.
2. Each row in connection profile implies edge information.

- **Models: simpler transformer, but better performance**

1. Cut position encoding
2. Cut edge weight.

- **Theoretical support**

Beyond previous works, the authors firstly try to theoretically prove the superior of orthonormal initialization of cluster centers.

Further discussions

❑ Experimental results

1. The MHSA contributes more but it is discussed less than OCRead.

Remark: 1) connection as initial node feature, no position embeddings. 2) no edge weight attention.

Table 2: Performance comparison AUROC (%) with different readout functions.

Readout	Dataset: ABIDE			Dataset: ABCD		
	SAN	Graphormer	VanillaTF	SAN	Graphormer	VanillaTF
MEAN	63.7±2.4	50.1±1.1	73.4±1.4	88.5±0.9	87.6±1.3	91.3±0.7
MAX	61.9±2.5	54.5±3.6	75.6±1.4	87.4±1.1	81.6±0.8	94.4±0.6
SUM	62.0±2.3	54.1±1.3	70.3±1.6	84.2±0.8	71.5±0.9	91.6±0.6
SortPooling	68.7±2.3	51.3±2.2	72.4±1.3	84.6±1.1	86.7±1.0	89.9±0.6
DiffPool	57.4±5.2	50.5±4.7	62.9±7.3	78.1±1.5	70.0±1.9	83.9±1.3
CONCAT	**71.3±2.1**	63.5±3.7	76.4±1.2	90.1±1.2	89.0±1.4	94.3±0.7
OCREAD	70.6±2.4	**64.9±2.7**	**80.2±1.0**	**91.2±0.7**	**90.2±0.7**	**96.2±0.4**

Further discussions

❑ Experimental results

1. The MHSA contributes more but it is discussed less than OCRead.

❑ Fully connected X & Attention Mechanism

2. Input X is a fully connected graph where each entry denotes the relation/similarity between a pair of nodes, which can be seen as a kind of attention. BrainNet, which conducts transformer on such X, is partly essentially an *l+1*-hop attention aggregation process. From this view, there could be **a discussion about 1-hop and *l+1*-hop aggregation**.

Multi-hop attention

$$\boldsymbol{Z}^l = (\|_{m=1}^{M} \boldsymbol{h}^{l,m}) \boldsymbol{W}_{\mathcal{O}}^l, \boldsymbol{h}^{l,m} = \boxed{\mathrm{Softmax}\left(\frac{\boldsymbol{W}_{\mathcal{Q}}^{l,m}\boldsymbol{Z}^{l-1}(\boldsymbol{W}_{\mathcal{K}}^{l,m}\boldsymbol{Z}^{l-1})^\top}{\sqrt{d_{\mathcal{K}}^{l,m}}}\right)} \boldsymbol{W}_{\mathcal{V}}^{l,m}\boldsymbol{Z}^{l-1}, \quad (1)$$

Original (1-hop) attention

where $\boldsymbol{Z}^0 = \boxed{\boldsymbol{X}}$, $\|$ is the concatenation operator, M is the number of heads, l is the layer index, $\boldsymbol{W}_{\mathcal{O}}^l, \boldsymbol{W}_{\mathcal{Q}}^{l,m}, \boldsymbol{W}_{\mathcal{K}}^{l,m}, \boldsymbol{W}_{\mathcal{V}}^{l,m}$ are learnable model parameters, and $d_{\mathcal{K}}^{l,m}$ is the first dimension of $\boldsymbol{W}_{\mathcal{K}}^{l,m}$.

Further discussions

❑ Experimental results

1. The MHSA contributes more but it is discussed less than OCRead.

❑ Fully connected X & Attention Mechanism

2. Input X is a fully connected graph where each entry denotes the relation/similarity between a pair of nodes, which can be seen as a kind of attention. BrainNet, which conducts transformer on such X, is partly essentially an $l+1$-hop attention aggregation process. From this view, there could be a discussion about 1-hop and $l+1$-hop aggregation.

❑ Selected Tasks

3. The selected binary classification tasks are relatively simple. More complex tasks can better demonstrate the effectiveness.

Thanks!

Jianpeng Chen

CS, Virginia Tech

10/23/2023