Self-Attention Graph-Pooling (SAGP)

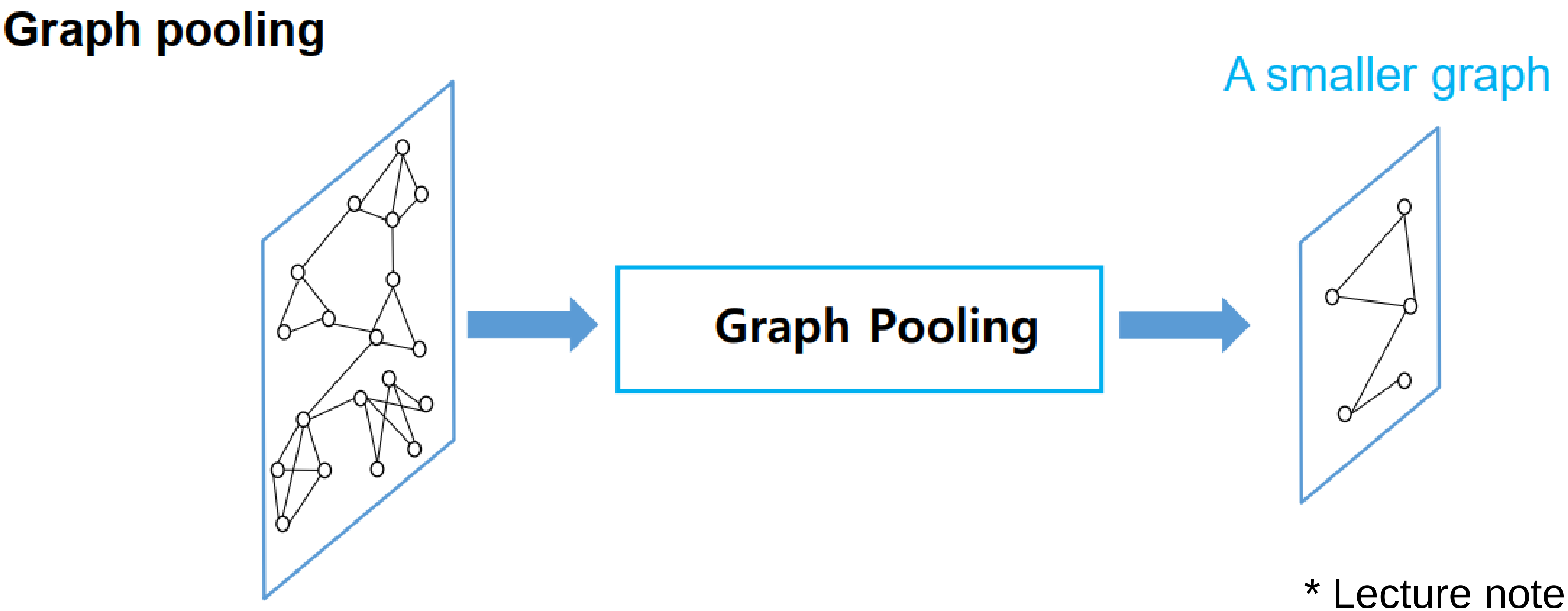
Yeongtak Oh

Seoul National University

Lee, Junhyun, Inyeop Lee, and Jaewoo Kang. "Self-attention graph pooling.“, ICML 2019, Citation: 42

Self-Attention Graph-Pooling

- Introduction



* Lecture note 13, p 15

- Intuition: Down-sample by selecting the most important nodes
- # of nodes: decrease, dimension of graphs: consistent

Self-Attention Graph-Pooling

- Proposed Method

1) Self-Attention Score: Utilizing the graph convolution

- $Z = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}X\Theta_{att}\right)$, $\mathrm{X} \in \mathbb{R}^{N\times F}, \Theta_{\mathrm{att}} \in \mathbb{R}^{F\times 1}$
- $idx = top - rank(Z, [kN]), Z_{mask} = Z_{idx}$

* Remind of Graph Convolution

$$h^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}h^{(l+1)}\Theta),\ \Theta \in \mathbb{R}^{F\times F\prime}$$

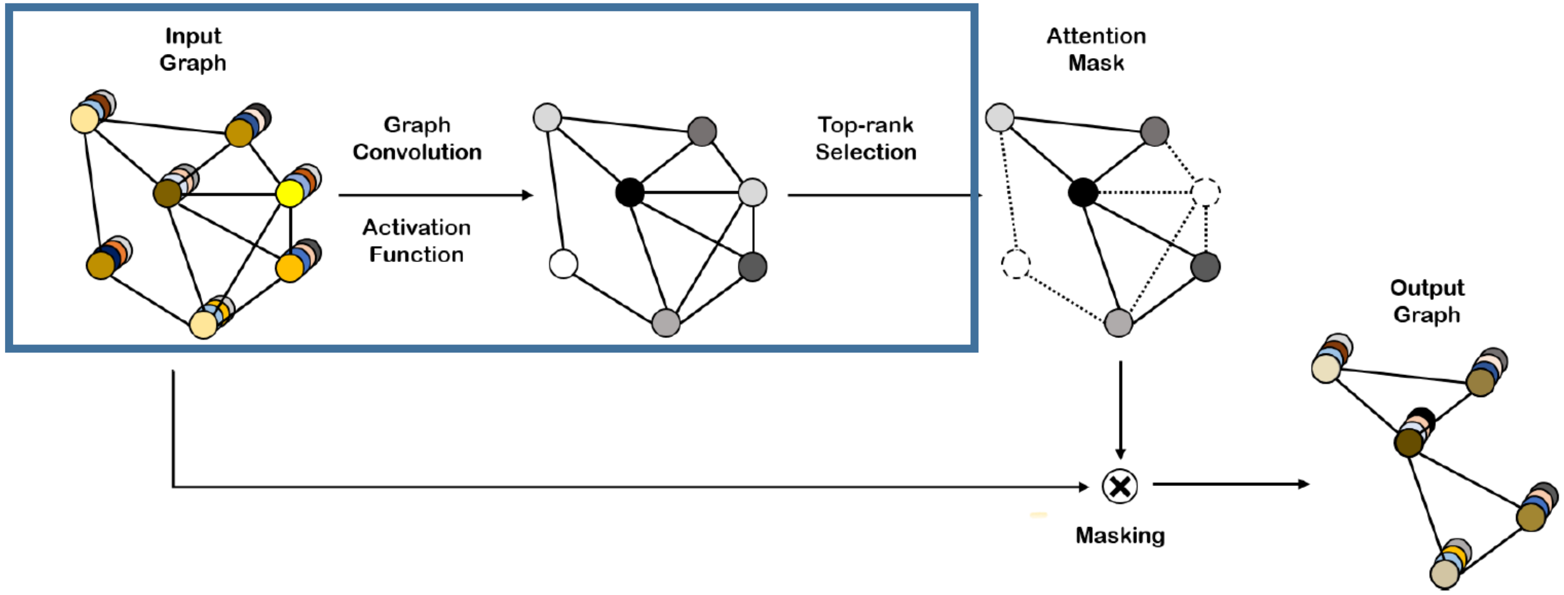


Figure 1. An illustration of the SAGPool layer.

Self-Attention Graph-Pooling

- Proposed Method

2) Graph Pooling

- $X' = X_{idx},\ X_{out} = X' \odot Z_{mask},\ A_{out} = A_{idx,idx}$

Node-wise feature matrix

Element-wise Product

Indexed-Adjacency matrix

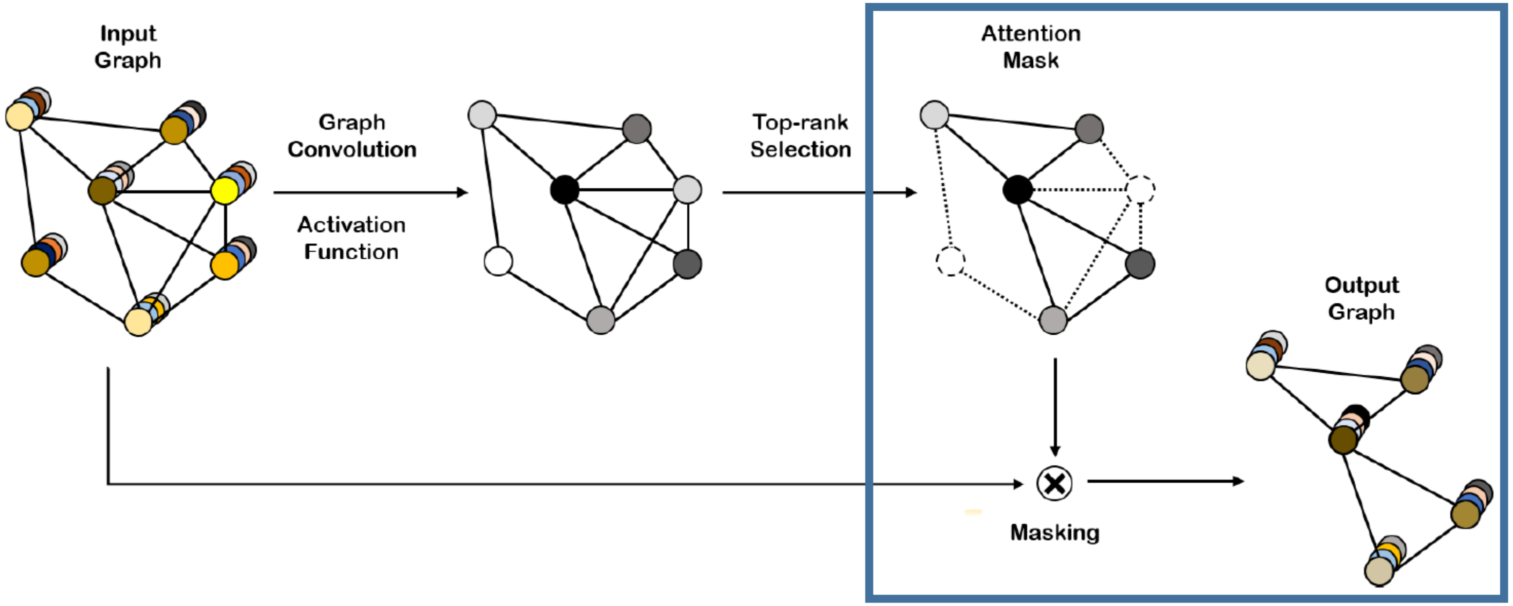
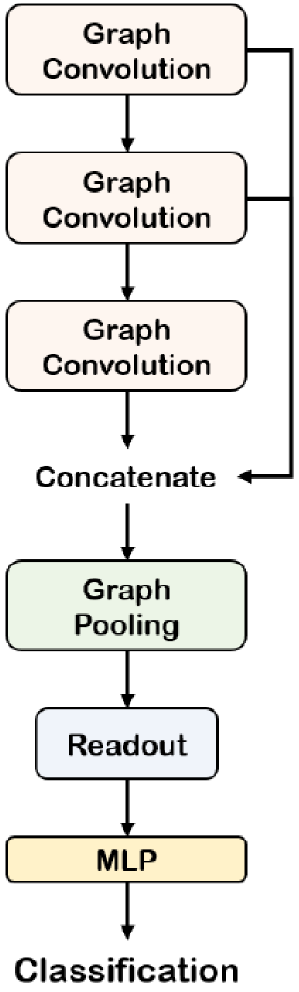


Figure 1. An illustration of the SAGPool layer.

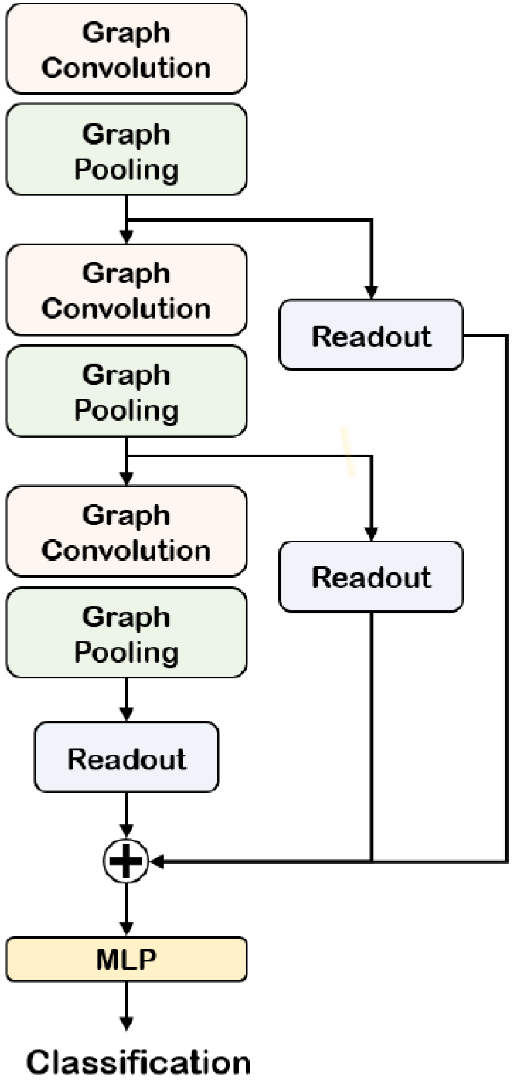
Self-Attention Graph-Pooling

- Proposed method
  - Global Pooling as, $Pool_g$
  - Hierarchical Pooling as, $Pool_h$
  - Readout Layer: Aggregates node features

$$s = \frac{1}{N}\sum_{i=1}^{N} x_i \,||\, \max_{i=1}^{N} x_i$$



$Pool_g$



$Pool_h$

Self-Attention Graph-Pooling

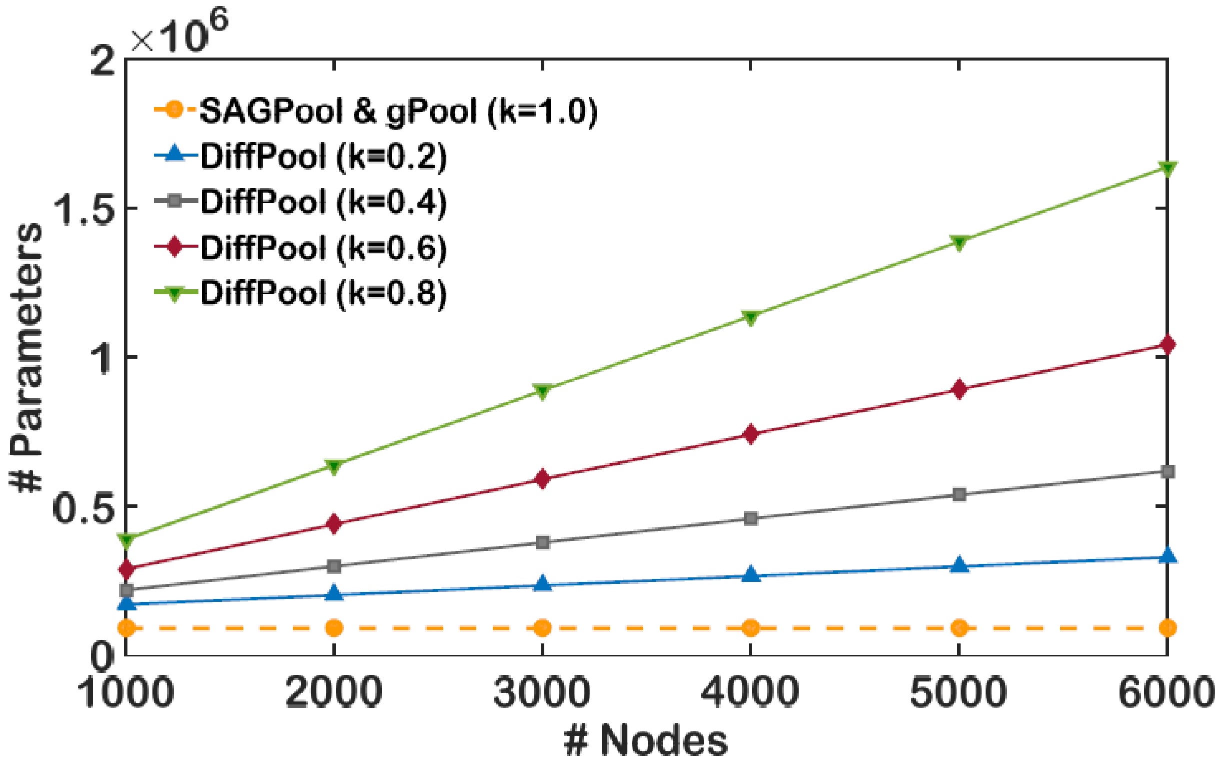
- Results
- 1) SOTA of Both on $Pool_g$ and $Pool_h$ architecture
- 2) $Pool_g$ for smaller graph, $Pool_h$ for a large number of nodes

Table 3. Average accuracy and standard deviation of the 20 random seeds. The subscript g (e.g. $POOL_g$) denotes the global pooling architecture and the subscript h (e.g. $POOL_h$) denotes the hierarchical pooling architecture.

Models	**D&D**	**PROTEINS**	**NCI1**	**NCI109**	**FRANKENSTEIN**
Set2Set$_g$	71.27 ± 0.84	66.06 ± 1.66	68.55 ± 1.92	69.78 ± 1.16	61.92 ± 0.73
SortPool$_g$	72.53 ± 1.19	66.72 ± 3.56	73.82 ± 0.96	74.02 ± 1.18	60.61 ± 0.77
SAGPool$_g$ (Ours)	$\mathbf{76.19} \pm 0.94$	$\mathbf{70.04} \pm 1.47$	$\mathbf{74.18} \pm 1.20$	$\mathbf{74.06} \pm 0.78$	$\mathbf{62.57} \pm 0.60$
DiffPool$_h$	66.95 ± 2.41	68.20 ± 2.02	62.32 ± 1.90	61.98 ± 1.98	60.60 ± 1.62
gPool$_h$	75.01 ± 0.86	71.10 ± 0.90	67.02 ± 2.25	66.12 ± 1.60	61.46 ± 0.84
SAGPool$_h$ (Ours)	$\mathbf{76.45} \pm 0.97$	$\mathbf{71.86} \pm 0.97$	$\mathbf{67.45} \pm 1.11$	$\mathbf{67.86} \pm 1.41$	$\mathbf{61.73} \pm 0.76$

Self-Attention Graph-Pooling

- Results
- 3) Same as $O(|V| + |E|)$ of gPool, but $O(|V|^2)$ of DiffPool
- 4) Consistent number of parameters regardless of the input



Conclusions and Remarks

- Conclusion:

1) Applying the concept of self-attention into a graph pooling
2) Showed a reasonable complexity, and end-to-end representation learning
3) Possible to expand with many variants(e.g. with SAGE, GAT)

- Future works:

Learnable pooling ratio, optimal cluster size, effects of multiple attention mask

Thank you for Listening