기말 발표

Self-Attention Graph-Pooling (SAGP)

Yeongtak Oh

Seoul National University

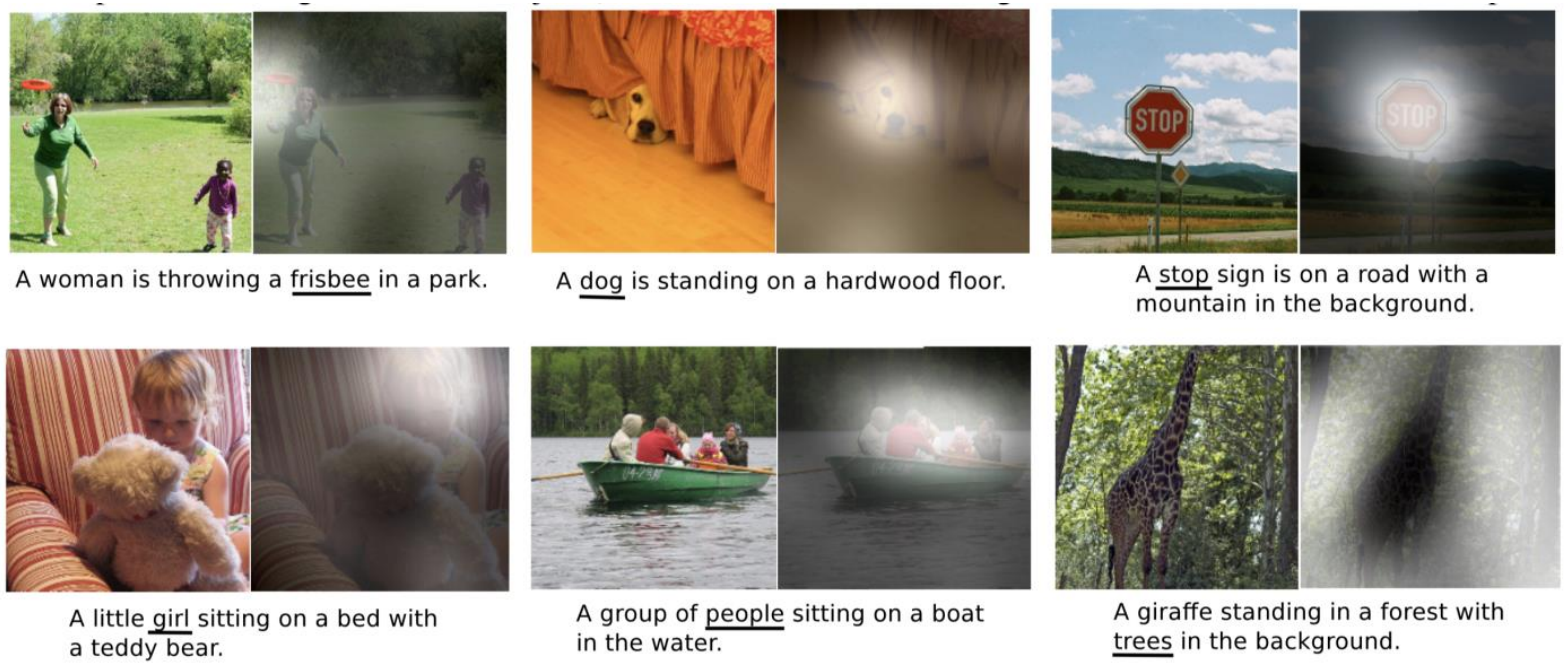
Lee, Junhyun, Inyeop Lee, and Jaewoo Kang. "Self-attention graph pooling.“, ICML 2019, Citation: 42

Self-Attention Graph-Pooling

- Introduction

1) Attention : used for feature selection based on multi-modal relationships of data[1]

2) Self-attention : allows input features to be the criteria for the attention itself[2]



[1]

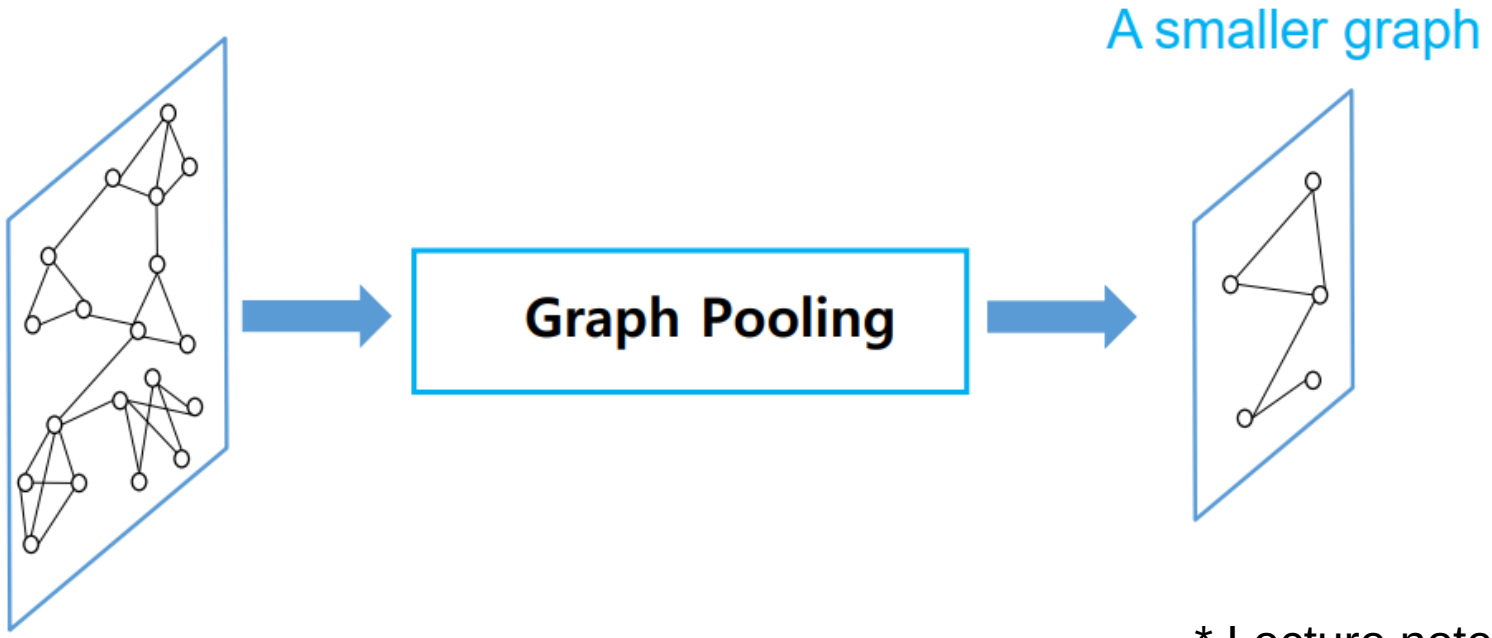
[1] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *ICML*. 2015.

[2] Vaswani, Ashish, et al. "Attention is all you need." *NIPS*. 2017.

Self-Attention Graph-Pooling

- Introduction

3) Graph Pooling



* Lecture note 13, p 15

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

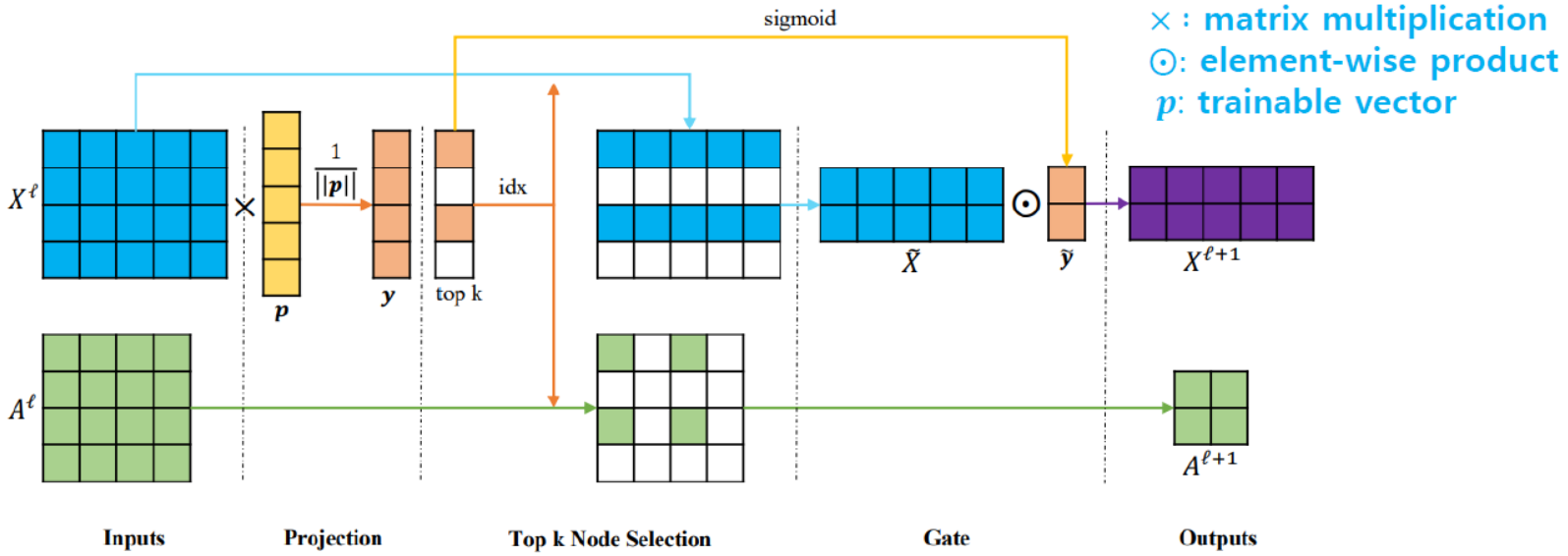
Self-Attention Graph-Pooling

- Proposed Method

1) Graph U-nets: gPool

$$y = softmax(X_p)$$
$$i = y_i > \alpha$$
$$X' = (X \odot y)_i$$
$$A' = A_{i,i}$$

Top-k graph pooling



[3] Gao, Hongyang, and Shuiwang Ji. "Graph u-nets." arXiv preprint arXiv:1905.05178 (2019).

Self-Attention Graph-Pooling

- Proposed Method

1) Self-Attention Graph Pooling

$$y = softmax(GNN(X, A))$$
$$i = y_i > \alpha$$
$$X' = (X \odot y)_i$$
$$A' = A_{i,i}$$

Self-attention graph pooling

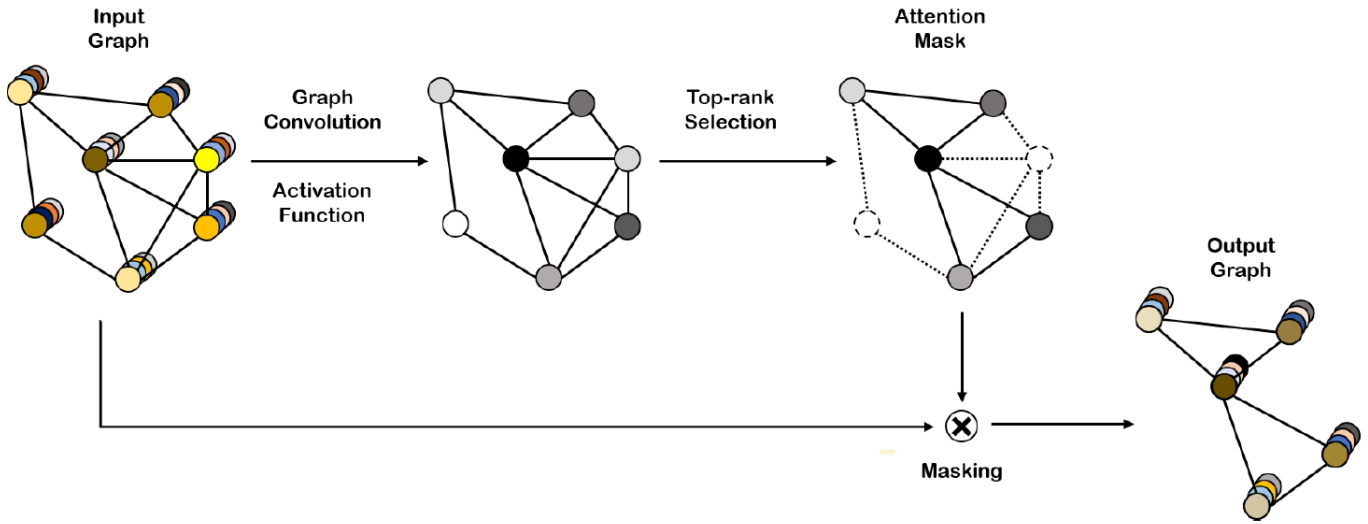


Figure 1. An illustration of the SAGPool layer.

Self-Attention Graph-Pooling

- Proposed Method

1) Self-Attention Score: Utilizing the graph convolution

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

* Remind of Graph Convolution

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

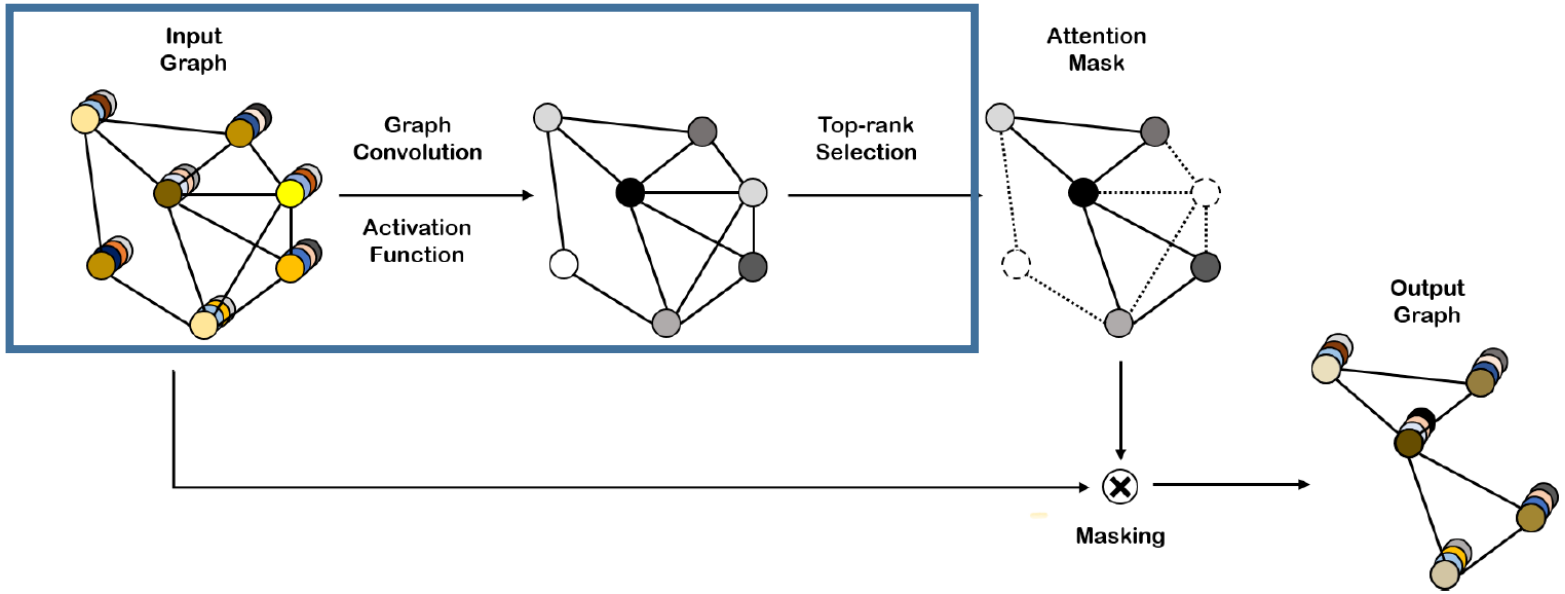


Figure 1. An illustration of the SAGPool layer.

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

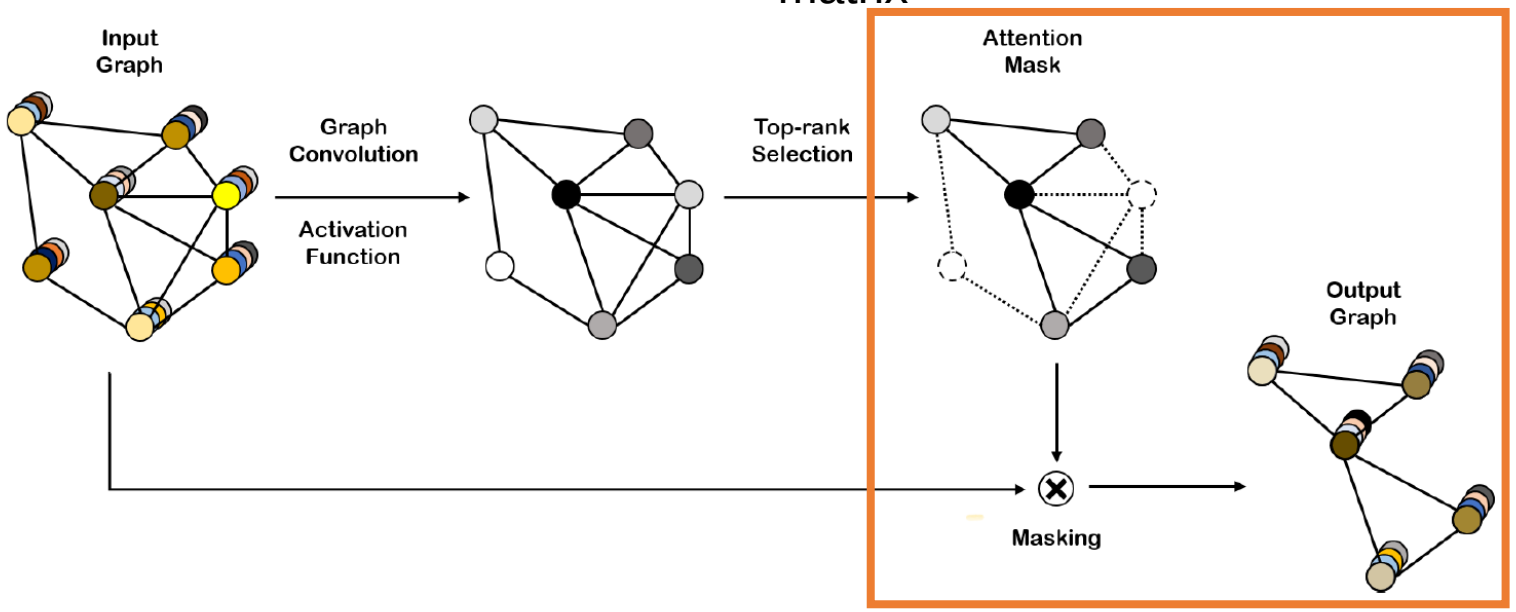
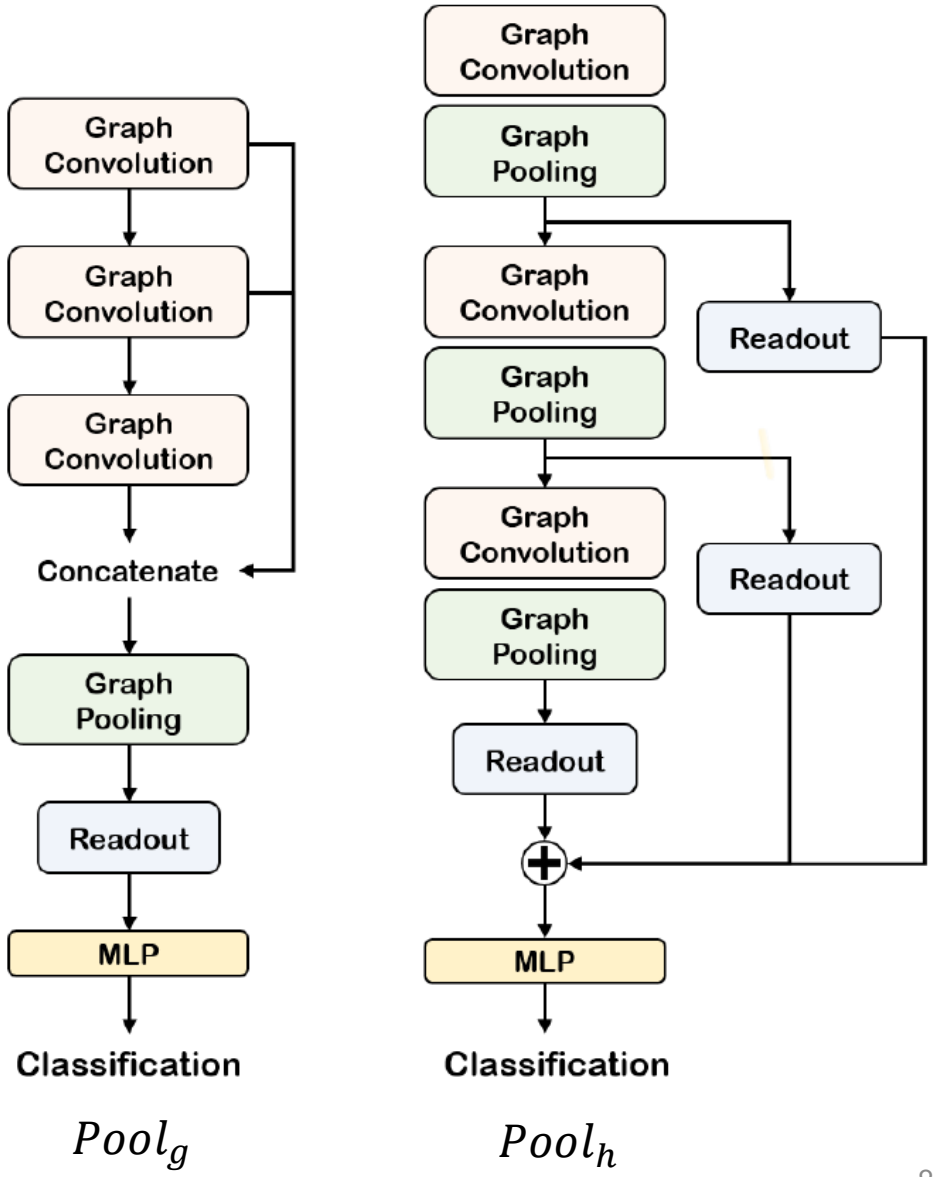


Figure 1. An illustration of the SAGPool layer.

Self-Attention Graph-Pooling

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

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



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)	$\mathbf{76.45} \pm 0.97$	$\mathbf{71.86} \pm 0.97$	$\mathbf{67.45} \pm 1.11$	$\mathbf{67.86} \pm 1.41$	$\mathbf{61.73} \pm 0.76$

Larger graph

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 1. Statistics of data sets.

Data set	Number of Graphs	Number of Classes	Avg. # of Nodes per Graph	Avg. # of Edges per Graph
D&D	1178	2	284.32	715.66
PROTEINS	1113	2	39.06	72.82
NCI1	4110	2	29.87	32.30
NCI109	4127	2	29.68	32.13
FRANKENSTEIN	4337	2	16.90	17.88

Hyperparameter	Range
Learning rate	1e-2, 5e-2, 1e-3, 5e-3, 1e-4, 5e-4
Hidden size	16, 32, 64, 128
Weight decay (L2 regularization)	1e-2, 1e-3, 1e-4, 1e-5
Pooling ratio	1/2, 1/4

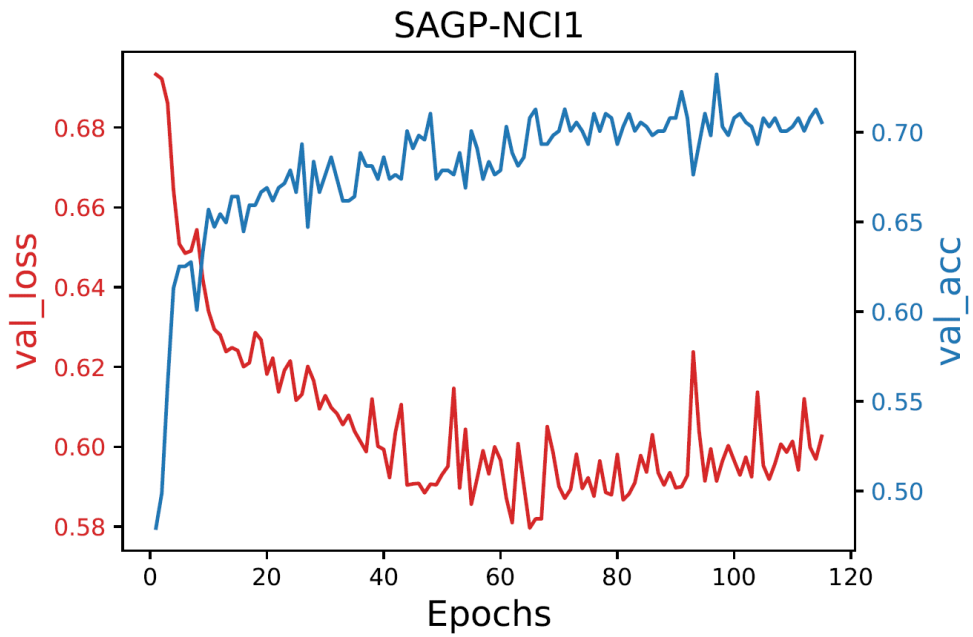
Self-Attention Graph-Pooling

- 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

Graph Convolution	D&D	PROTEINS
SAGPool_h	76.45 ± 0.97	71.86 ± 0.97
$\text{SAGPool}_{h,\text{Cheb}}$	75.82 ± 0.79	71.98 ± 0.93
$\text{SAGPool}_{h,\text{SAGE}}$	76.28 ± 1.06	71.93 ± 0.82
$\text{SAGPool}_{h,\text{GAT}}$	75.49 ± 0.93	71.98 ± 1.01
$\text{SAGPool}_{h,\text{augmentation}}$	77.07 ± 0.82	71.82 ± 0.81
$\text{SAGPool}_{h,\text{serial,2layers}}$	76.68 ± 0.96	72.17 ± 0.87
$\text{SAGPool}_{h,\text{parallel},M=2}$	75.79 ± 0.96	72.05 ± 0.43
$\text{SAGPool}_{h,\text{parallel},M=4}$	76.77 ± 0.61	71.66 ± 0.98

Self-Attention Graph-Pooling

- Reproducing Results
- AIDS, DD, PROTEINS, NCI1 dataset에 대해 reproducing experiment를 수행
- Double-axis로 val_loss 및 val_acc 값을 plot

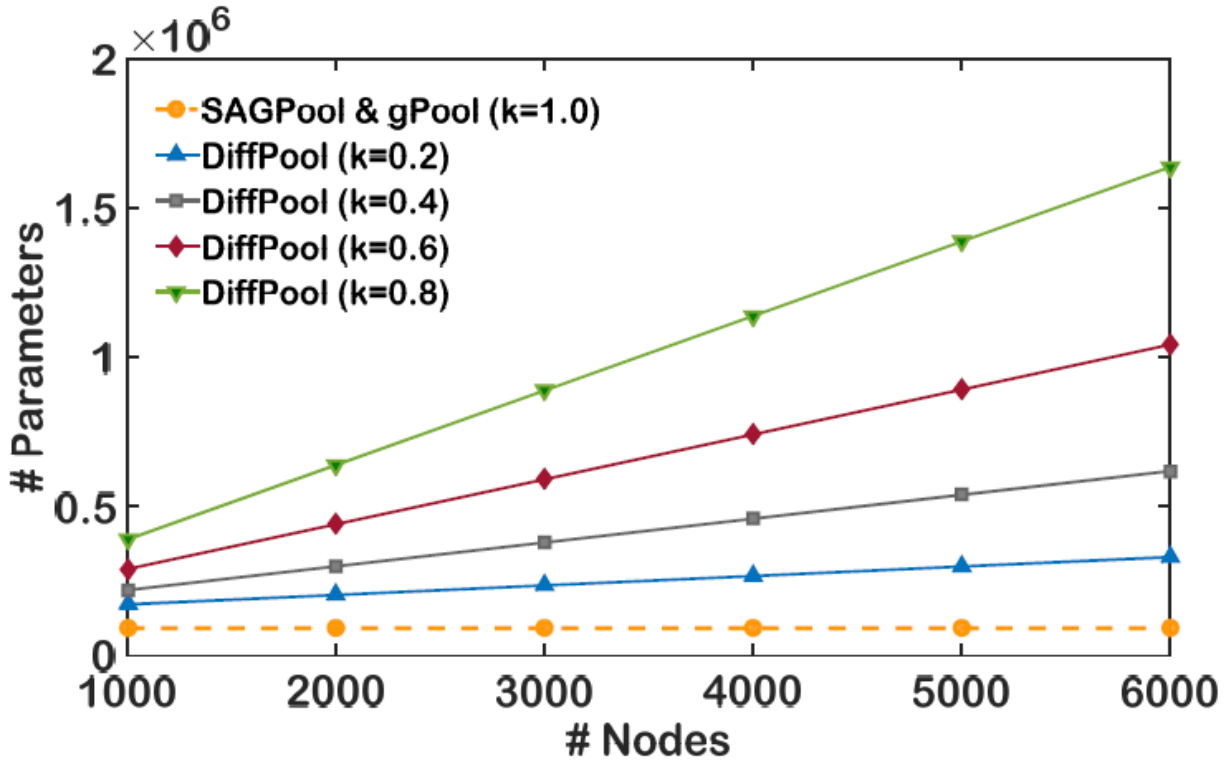


Datasets: https://chrsmrrs.github.io/datasets/docs/datasets/

Reproducing Code: https://github.com/oyt9306/SAGP_reproducing

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, multiple attention mask

Appendix

- Additional References about Graph Pooling:

1) Diehl, Frederik. "Edge contraction pooling for graph neural networks." arXiv preprint arXiv:1905.10990 (2019).

2) Ranjan, Ekagra, Soumya Sanyal, and Partha Pratim Talukdar. "ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations." arXiv preprint arXiv:1911.07979 (2019).

Thank you for Listening