Scaling Graph Transformers with Expander Graphs

Danica Sutherland UBC + Amii (she/her)

based on:
Exphormer: Scaling Graph Transformers with Expander Graphs (ICML 2023; arXiv:2303.06147)
Even Sparser Graph Transformers (Spexphormer; not quite on arXiv yet, but soon!)

with:

Hamed Shirzad UBC
Ameya Velingker Google
Balaji Venkatachalam Google → Meta
Ali Sinop Google (Exphormer only)
Honghao Lin CMU (Spexphormer only)
David Woodruff CMU + Google (Spexphormer only)

SFU VCR/AI seminar, June 2024
Learning on graphs

Molecular Graphs

Knowledge Graphs

Social Network Graphs

Road Network Graphs
Message-passing neural networks

- **Features** at each node $x_i$, maybe also each edge $e_{i→j}$

- **Message** from node $j$ to node $i$: $m_{j→i} = f_e(x_j, x_i, e_{j→i})$ for some NN $f_e$

- New features $x_i^+ = f_v(x_i, \sum_j m_{j→i})$ for some NN $f_v$
Problems with message-passing neural nets

- Information only propagates along the graph edges
  - What if there are long-range dependencies?

- Over-smoothing
  - In deeper layers, all node features often end up basically the same as each other

- Over-squashing
Self-attention layers

$Q_2 = W_Q x_2$

$a_{1\to 2} = softmax(K_1^\top Q_2 / \sqrt{d})$

$a_{2\to 2} = softmax(K_2^\top Q_2 / \sqrt{d})$

$a_{3\to 2} = softmax(K_3^\top Q_2 / \sqrt{d})$

Looks like message-passing over the full graph

$x_2^+ = \sum_j m_{j\to 2}$
Positional encodings for sequences

• Problem: the order of the sequence is lost!
• Solution: add features for positional encodings that tell you “where you are”
• “Default” version: trigonometric features

\[
PE_{(pos, 2i)} = sin(pos/10000^{2i/d_{model}}) \\
PE_{(pos, 2i+1)} = cos(pos/10000^{2i/d_{model}})
\]
Positional encodings

• Trigonometric functions: eigenfunctions of Laplacian on Euclidean space

Eigenvalues and eigenfunctions of the Laplacian
Andrew Hassell

The setting
In this talk I will consider the Laplace operator, $\Delta$, on various geometric spaces $M$. On a Euclidean domain,
$$\Delta f = - \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$ 

Examples
• The interval $[0, a]$. Eigenfunctions and eigenvalues are
  $$u_n = \sqrt{\frac{2}{a}} \sin \frac{\pi nx}{a}, \quad \lambda_n = \frac{\pi n}{a}.$$ 

• The torus $T^2_\pi$. Eigenfunctions and eigenvalues are
  $$u = \frac{1}{\pi} e^{i l x} e^{i m y}, \quad \lambda = \sqrt{l^2 + m^2}.$$
Positional encodings

• Trigonometric functions: eigenfunctions of Laplacian on Euclidean space
• What should we do so that a self-attention layer “knows” about the structure of a graph?
  • “Default” choice: eigenvectors of the graph Laplacian $L = D - A$
Figure 3: Examples of eigenvalues $\lambda_i$ and eigenvectors $\phi_i$ for molecular graphs. The low-frequency eigenvectors $\phi_1, \phi_2$ are spread across the graph, while higher frequencies, such as $\phi_{14}, \phi_{15}$ for the left molecule or $\phi_{10}, \phi_{11}$ for the right molecule, often resonate in local structures.
Learned positional encodings (Kreuzer et al. 2021)

- Can be helpful to post-process the positional encodings further
Spectral Attention Networks (SAN) (Kreuzer et al. 2021)

- Self-attention layers with learned positional encodings
- Intersperse with MLPs processing each node feature independently, as in Transformer encoders

Figure 1: The proposed SAN model with the node LPE, a generalization of Transformers to graphs.
Full attention and position encodings help

<table>
<thead>
<tr>
<th>Model details</th>
<th>ZINC</th>
<th>PATTERN</th>
<th>CLUSTER</th>
<th>MOLHIV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention</td>
<td>LPE</td>
<td>MAE</td>
<td>% ACC</td>
<td>% ACC</td>
</tr>
<tr>
<td>Sparse</td>
<td>-</td>
<td>0.267 ± 0.032</td>
<td>83.613 ± 0.663</td>
<td>75.683 ± 0.098</td>
</tr>
<tr>
<td>Sparse</td>
<td>Node</td>
<td>0.198 ± 0.004</td>
<td>81.329 ± 2.150</td>
<td>75.738 ± 0.106</td>
</tr>
<tr>
<td>Full</td>
<td>-</td>
<td>0.392 ± 0.055</td>
<td>86.322 ± 0.049</td>
<td>76.447 ± 0.177</td>
</tr>
<tr>
<td>Full</td>
<td>Node</td>
<td>0.157 ± 0.006</td>
<td>86.441 ± 0.040</td>
<td>76.691 ± 0.247</td>
</tr>
</tbody>
</table>

Figure 6: Ablation study on datasets from [15, 21] for the node LPE and full graph attention, with no hyperparameter tuning other than \( \gamma \) taken from Figure 5. For a given dataset, all models use the same hyperparameters, but the hidden dimensions are adjusted to have \( \sim 500k \) learnable parameters. Means and uncertainties are derived from four runs, with different seeds (except MolHIV).
Message passing vs graph transformers

**Message passing networks**
- Update across edges of input graph
- Capture inductive bias from input graph topology
- Efficient computation: $O(N + M)$
- Difficulty with long-range dependencies
- Oversmoothing, oversquashing

**Graph transformers**
- Use global attention
- Computation graph can be different from input graph
- Long-range modeling
- Universal Approximation Theorem
- Knowledge of graph structure
- Loss of inductive bias from graph
- Inefficient computation: $O(N^2)$

*Positional and structural encodings*

*Our work!*
GraphGPS
(Rampášek et al., 2022)

- Framework to combine transformer layers with message-passing layers
- MPNN layers better at preserving graph structure
- Transformer layers better at certain kinds of dependence

Figure 1: Modular GPS graph Transformer, with examples of PE and SE. Task specific layers for node/graph/edge-level predictions, such as pooling or output MLP, are omitted for simplicity.
Sparse attention and Exphormer
Reducing the $N^2$ attention cost

- Lots of work on more-efficient attention mechanisms for sequences
- GraphGPS paper tried to apply some to graphs...but it hurt accuracy a lot
- Need something designed for graphs

Image from “Efficient Transformers: A Survey” (Tay et al.)
Attention graph

• We’ll take an approach of “sparse attention”

• Start by “running” attention along the graph edges (both ways)
  • Adjacent nodes are reasonably likely to “matter” to each other

• Full transformers augment this attention graph with the complete graph
  • Every node connected to every other
  • Allows easy long-range communication
  • Adds too many edges! Makes it slow + memory-hungry
Augmenting attention with virtual nodes

- Add a "virtual" node (or four) that’s connected to every graph node
- Now any node can “talk to” any other node in at most 2 hops
- Also common in MPNNs

- Problem: oversquashing
Expander graphs

• How can we connect far-away parts of the graph without severe oversquashing?

• We want a sparse approximation to the complete graph
  • No too many edges
  • Can get from anywhere to anywhere quickly
  • Have many routes from place to place, to avoid oversquashing

• Mathematical concept called expander graphs
  • Approximate the complete graph (in many senses), with low degree
Expander graphs

**Theorem 4.1.** A $d$-regular $ε$-expander $G$ on $n$ vertices spectrally approximates the complete graph $K_n$ on $n$ vertices.

$$(1 - ε) \frac{1}{n} L_K \preceq \frac{1}{d} L_G \preceq (1 + ε) \frac{1}{n} L_K.$$  

Spectral approximation is known to preserve the cut structure in graphs. As a result, a sparse attention mechanism based on expander edges retains spectral properties of the full attention mechanism: cuts, vertex expansion, and so on.

**Theorem 4.3.** (Alon, 1986) Suppose $G = (V, E)$ is a $d$-regular $ε$-expander graph on $n$ vertices. Then, for every vertex $v$ and $k \geq 0$, the $k$-hop neighborhood $B(v, r) = \{w \in V : d(v, w) \leq k\}$ has

$$|B(v, r)| \geq \min\{ (1 + c)^k, n \}$$

for some constant $c > 0$ depending on $d, ε$. In particular, we have that $\text{diam}(G) = O_{d, ε}(\log n)$.

**Lemma 4.2.** (Alon, 1986) Let $G = (V, E)$ be a $d$-regular $ε$-expander graph on $n = |V|$ nodes. For any initial distribution $\pi(0) : V \to \mathbb{R}^+$ and any $δ > 0$, $\pi(t)$ satisfies

$$\|\pi(t) - \frac{1}{n} \|_1 \leq δ$$

as long as $t = Ω \left( \frac{1}{ε} \log(n/δ) \right)$.

**Corollary 4.4.** If a sparse attention mechanism on $n$ nodes is a $d$-regular $ε$-expander graph, then stacking $O_{d, ε}(\log n)$ transformer layers models all pairwise node interactions.
Expander graphs turn out to have a slew of applications not only in math but also in computer science and physics. They can be used to create error-correcting codes and to figure out when simulations based on random numbers converge to the reality they are trying to simulate. Neurons can be modeled in a graph that some researchers believe forms an expander, due to the limited space for connections inside the brain. The graphs are also useful to geometers who try to understand how to traverse complicated surfaces, among other problems.
Constructing expander graphs

• How to make an expander graph with max degree $\leq d$:
  • Sample $d/2$ random Hamiltonian cycles, combine these as edges
  • Check (from the eigenvalues of the Laplacian) if it’s an expander
  • If not, try again (happens rarely)
Components of Exphormer

**Original graph**
- Preserve locality from the original graph

**Expander graph**
- Random walk mixing
- Constant degree
- $O(N)$ edges

**Virtual node(s)**
- "Storage sink"
- Short connections between all node pairs

**Exphormer**: Combine all three to form the interaction graph

Use edge features to let the network know which kind of edge it is
Universal approximation

**Theorem E.3.** Suppose $H$ is the attention graph of EXPHORMER (which contains $n$ graph nodes and potentially more virtual nodes), augmented with self loops on all nodes. Suppose $H$ satisfies at least one of the following:

1. $H$ contains at least one node which is connected to all $n$ graph nodes (i.e., at least one virtual node is included).
2. The underlying expander graph of $H$ contains a Hamiltonian path.

Then, it follows that a sparse transformer model, with positional encodings and an attention mechanism following $H$, can universally approximate continuous functions $f : [0, 1]^{d \times n} \to \mathbb{R}^{d \times n}$. That is, for any $1 < p < \infty$ and $\epsilon > 0$, there exists a sparse transformer network $g$, which uses the attention graph $H$ and some positional encodings, such that $\ell^p(f, g) < \epsilon.$
Experimental Results: Various Datasets

Table 1. Comparison of EXPHORMER with baselines on various datasets. Best results are colored in **first**, **second**, and **third**.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR10 Accuracy ↑</th>
<th>MalNet-Tiny Accuracy ↑</th>
<th>MNIST Accuracy ↑</th>
<th>CLUSTER Accuracy ↑</th>
<th>PATTERN Accuracy ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN (Kipf &amp; Welling, 2017)</td>
<td>55.71±0.381</td>
<td>81.0</td>
<td>90.71±0.218</td>
<td>68.50±0.976</td>
<td>71.89±0.334</td>
</tr>
<tr>
<td>GIN (Xu et al., 2018)</td>
<td>55.26±1.527</td>
<td>88.98±0.557</td>
<td>96.49±0.252</td>
<td>64.72±1.553</td>
<td>85.39±0.136</td>
</tr>
<tr>
<td>GAT (Veličković et al., 2018)</td>
<td>64.22±0.455</td>
<td>92.1±0.242</td>
<td>95.54±0.205</td>
<td>70.59±0.447</td>
<td>78.27±0.186</td>
</tr>
<tr>
<td>GatedGCN (Bresson &amp; Laurent, 2017; Dwivedi et al., 2020)</td>
<td>67.31±0.311</td>
<td><strong>92.23±0.65</strong></td>
<td>97.34±0.143</td>
<td>73.84±0.326</td>
<td>85.57±0.088</td>
</tr>
<tr>
<td>PNA (Corso et al., 2020)</td>
<td>70.35±0.63</td>
<td>–</td>
<td><strong>97.94±0.12</strong></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DGN (Beinini et al., 2021)</td>
<td><strong>72.84±0.417</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td><strong>86.68±0.034</strong></td>
</tr>
<tr>
<td>CRaWIL (Toenshoff et al., 2021)</td>
<td>69.01±0.259</td>
<td>–</td>
<td><strong>97.94±0.050</strong></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GIN-AK+ (Zhao et al., 2022b)</td>
<td>72.19±0.13</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td><strong>86.85±0.057</strong></td>
</tr>
<tr>
<td>SAN (Kreuzer et al., 2021)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>76.69±0.65</td>
<td>86.58±0.037</td>
</tr>
<tr>
<td>K-Subgraph SAT (Chen et al., 2022a)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>77.86±0.104</td>
<td><strong>86.85±0.037</strong></td>
</tr>
<tr>
<td>EGT (Hussain et al., 2021)</td>
<td>68.70±0.409</td>
<td>–</td>
<td><strong>98.17±0.087</strong></td>
<td>79.23±0.348</td>
<td>86.82±0.020</td>
</tr>
<tr>
<td>GraphGPS (Rampásek et al., 2022)</td>
<td><strong>72.30±0.356</strong></td>
<td><strong>93.50±0.41</strong></td>
<td>98.05±0.126</td>
<td>78.02±0.180</td>
<td>86.69±0.059</td>
</tr>
<tr>
<td><strong>EXPHORMER (ours)</strong></td>
<td><strong>74.69±0.125</strong></td>
<td><strong>94.02±0.209</strong></td>
<td><strong>98.55±0.039</strong></td>
<td><strong>78.07±0.037</strong></td>
<td><strong>86.74±0.015</strong></td>
</tr>
</tbody>
</table>
Experimental Results: Long-Range Benchmark

Table 3. Comparison of EXPHORMER with baselines from the Long-Range Graph Benchmarks (LRGB, Dwivedi et al., 2022). Best results are colored in first, second, third.

<table>
<thead>
<tr>
<th>Model</th>
<th>PascalVOC-SP F1 score ↑</th>
<th>COCO-SP F1 score ↑</th>
<th>Peptides-Func AP ↑</th>
<th>Peptides-Struct MAE ↓</th>
<th>PCQM-Contact MRR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN</td>
<td>0.1268 ± 0.0060</td>
<td>0.0841 ± 0.0010</td>
<td>0.5930 ± 0.0023</td>
<td>0.3496 ± 0.0013</td>
<td>0.3234 ± 0.0006</td>
</tr>
<tr>
<td>GINE</td>
<td>0.1265 ± 0.0076</td>
<td>0.1339 ± 0.0044</td>
<td>0.5498 ± 0.0079</td>
<td>0.3547 ± 0.0045</td>
<td>0.3180 ± 0.0027</td>
</tr>
<tr>
<td>GatedGCN</td>
<td>0.2873 ± 0.0219</td>
<td>0.2641 ± 0.0045</td>
<td>0.5864 ± 0.0077</td>
<td>0.3420 ± 0.0013</td>
<td>0.3218 ± 0.0011</td>
</tr>
<tr>
<td>GatedGCN+RWSE</td>
<td>0.2860 ± 0.0085</td>
<td>0.2574 ± 0.0034</td>
<td>0.6069 ± 0.0035</td>
<td>0.3357 ± 0.0006</td>
<td>0.3242 ± 0.0008</td>
</tr>
<tr>
<td>Transformer+LapPE</td>
<td>0.2694 ± 0.0098</td>
<td>0.2618 ± 0.0031</td>
<td>0.6326 ± 0.0126</td>
<td>0.2529 ± 0.0016</td>
<td>0.3174 ± 0.0020</td>
</tr>
<tr>
<td>SAN+LapPE</td>
<td>0.3230 ± 0.0039</td>
<td>0.2592 ± 0.0158*</td>
<td>0.6384 ± 0.0121</td>
<td>0.2683 ± 0.0043</td>
<td>0.3350 ± 0.0003</td>
</tr>
<tr>
<td>SAN+RWSE</td>
<td>0.3216 ± 0.0027</td>
<td>0.2434 ± 0.0156*</td>
<td>0.6439 ± 0.0075</td>
<td>0.2545 ± 0.0012</td>
<td>0.3341 ± 0.0006</td>
</tr>
<tr>
<td>GraphGPS</td>
<td>0.3748 ± 0.0109</td>
<td>0.3412 ± 0.0044</td>
<td>0.6535 ± 0.0041</td>
<td>0.2500 ± 0.0005</td>
<td>0.3337 ± 0.0006</td>
</tr>
<tr>
<td>Exphormer (ours)</td>
<td>0.3975 ± 0.0037</td>
<td>0.3455 ± 0.0009</td>
<td>0.6527 ± 0.0043</td>
<td>0.2481 ± 0.0007</td>
<td>0.3637 ± 0.0020</td>
</tr>
</tbody>
</table>
Experimental Results: Larger Graphs

Table 4. Accuracy of models with different attention mechanisms on transductive graph datasets (numbers in top rows, other than arXiv, are from Chen et al., 2022b). Chen et al. did not report NAGphormer results on this dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>ogbn-arxiv</th>
<th>Computer</th>
<th>Photo</th>
<th>CS</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN</td>
<td>OOM</td>
<td>89.83 ± 0.16</td>
<td>94.86 ± 0.10</td>
<td>94.51 ± 0.15</td>
<td>OOM</td>
</tr>
<tr>
<td>GraphGPS</td>
<td>OOM</td>
<td>OOM</td>
<td>95.06 ± 0.13</td>
<td>93.93 ± 0.12</td>
<td>OOM</td>
</tr>
<tr>
<td>NAGphormer</td>
<td>NA</td>
<td>91.22 ± 0.14</td>
<td>95.49 ± 0.11</td>
<td>95.75 ± 0.09</td>
<td>97.34 ± 0.03</td>
</tr>
<tr>
<td>EXPHORMER</td>
<td>72.44 ± 0.28</td>
<td>91.59 ± 0.31</td>
<td>95.27 ± 0.42</td>
<td>95.77 ± 0.15</td>
<td>97.16 ± 0.13</td>
</tr>
</tbody>
</table>
What if the original graph is too big?
Memory usage

- GPUs only have so much memory available
- Exphormer only adds $\Theta(N)$ extra edges to the original graph

- But what if the original graph already had too many edges to run on?
  - ogbn-proteins has 132,000 nodes (fine) but almost 40,000,000 edges!

- Can easily run out of GPU memory even with just original-graph attention
Which edges matter?

- Attention on graphs tends to be kind of sparse, ish
- If we could throw away unimportant edges beforehand, we could afford to do our optimization on the rest

- But how to tell which edges matter if we can’t afford to train?
- Idea: train another network to tell us which edges matter
Attention score estimation

- Idea I: narrow nets are much cheaper in memory than wide nets (lower-dimensional hidden features)
- Idea II: CPU memory is much cheaper than GPU memory

Node characteristics

<table>
<thead>
<tr>
<th>nodes</th>
<th>cores</th>
<th>available memory</th>
<th>CPU</th>
<th>storage</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>64</td>
<td>4000G or 4096000M</td>
<td>2 x AMD Rome 7502 @ 2.50 GHz 128M cache L3</td>
<td>1 x 960G SSD</td>
<td>-</td>
</tr>
<tr>
<td>33</td>
<td>64</td>
<td>2009G or 2057500M</td>
<td>2 x AMD Rome 7532 @ 2.40 GHz 256M cache L3</td>
<td>1 x 960G SSD</td>
<td>-</td>
</tr>
<tr>
<td>159</td>
<td>48</td>
<td>498G or 510000M</td>
<td>2 x AMD Milan 7413 @ 2.65 GHz 128M cache L3</td>
<td>1 x SSD of 3.84 TB</td>
<td>4 x NVidia A100SXM4 (40 GB memory), connected via NVLink</td>
</tr>
</tbody>
</table>

- Train a **narrow** network on **CPU** to tell which edges matter
Attention scores agree across widths

**Figure 2:** Energy distance between the attention scores of various networks to a network of width 64. “Uniform” refers to the baseline placing uniform scores on each neighbor, while “random” refers to the baseline with uniformly distributed logits. The remaining bars refer to networks trained on the appropriately labeled width.

Caveat I: adding layer normalization to value matrices, to make attention scores more comparable
Caveat II: using an annealed temperature schedule to encourage sparser attention
Theory

• Have some partial results about existence of narrow transformers approximating wide ones

• The first attention layer can be well-approximated with only $\log(N)$ width

• Following MLP + attention layers...not totally sure 😅
Choosing the important edges

• Once we know attention scores, how do we decide which to keep?
  • Sample $k$ edges for each node proportionally to the attention score
  • Do different samples per graph, and resample in each epoch

• Theory: approximates attention matrix in spectral norm well
  (extending proof of Achlioptas, Karnin, Liberty, NeurIPS 2013)

• Using reservoir sampling makes it run much faster than default implementation

• Once we have the sparsified graph,
  run a wide transformer (on GPU) on only that sparsified graph
Figure 1: Steps of our method. (a) The attention mechanism for the attention score estimator network combines graph edges with an expander graph and self-loops. The expander graphs are constructed by combining a small number of Hamiltonian cycles – here two, in red and in purple – then confirming the spectral gap is large enough. (b) Self-attention layers in the estimator network use this sparse attention mechanism; its self-attention layers normalize $\mathbf{V}$. (c, d) Attention scores are extracted from this network for each layer, and used to sample, in (e), a sparse directed graph, which becomes the attention graph for the final network (f). This network, with a much larger feature dimension, does not normalize $\mathbf{V}$. 
Advantage: regular graph

- Irregularly-shaped matrix multiplications are slow on GPU

- If each node has exactly $k$ neighbours, we instead can do batched matrix multiplication with consistent shapes (much faster!)
Advantage: neighbourhood sampling

• Can do minibatching by
  • Start with some “core nodes”
  • Include their entire “receptive field” of other nodes that affect them

• Compared to other minibatching strategies on graphs, guarantees that minibatching doesn’t bias the process
Figure 4: Figure (a) shows a very simple synthetic graph where each node has a binary classification task of determining whether there exists a node of the opposite color in the graph. This task requires learning long-range dependencies. Figure (b) shows a natural clustering of the graph, this clustering would mean no node can do its task if models are trained only on the clusters. Figure (c) shows a neighbor sampling starting from the green node, where random sampling fails to select the important edge that bridges the between the two different colored nodes. Figure (d) shows a random subset sampling strategy that a subset of nodes are likely to be selected from the same color only, making it impossible to reason about the task. (e) shows attention scores between the nodes if trained with an attention based network. Dashed lines have near zero attention scores, and thicker the lines the attention score is higher. Knowing these attention scores will mean each node with just one directional edge can do the task perfectly. The attention edges are shown in (f). In case two nodes are equally informative selecting either of them leads to the correct result.
Table 1: Comparison of our model with other GNNs on five homophilic and three heterogeneous datasets (shown in the left and right tables, respectively). The reported metric is ROC-AUC for the Minesweeper and Tolokers datasets, and accuracy for all others.

<table>
<thead>
<tr>
<th>Model</th>
<th>Computer</th>
<th>Photo</th>
<th>CS</th>
<th>Physics</th>
<th>ogbn-arxiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN</td>
<td>89.65 ± 0.52</td>
<td>92.70 ± 0.20</td>
<td>92.92 ± 0.12</td>
<td>96.18 ± 0.07</td>
<td>71.74 ± 0.29</td>
</tr>
<tr>
<td>GraphSAGE</td>
<td>91.20 ± 0.29</td>
<td>94.59 ± 0.14</td>
<td>93.91 ± 0.13</td>
<td>96.49 ± 0.06</td>
<td>71.49 ± 0.27</td>
</tr>
<tr>
<td>GAT</td>
<td>90.78 ± 0.13</td>
<td>93.87 ± 0.11</td>
<td>93.61 ± 0.14</td>
<td>96.17 ± 0.08</td>
<td>72.01 ± 0.20</td>
</tr>
<tr>
<td>GraphSAINT</td>
<td>90.22 ± 0.15</td>
<td>91.72 ± 0.13</td>
<td>94.41 ± 0.09</td>
<td>96.43 ± 0.05</td>
<td>68.50 ± 0.23</td>
</tr>
<tr>
<td>NodeFormer</td>
<td>86.98 ± 0.62</td>
<td>93.46 ± 0.35</td>
<td>95.64 ± 0.22</td>
<td>96.45 ± 0.28</td>
<td>59.90 ± 0.42</td>
</tr>
<tr>
<td>GraphGPS</td>
<td>91.19 ± 0.54</td>
<td>95.06 ± 0.13</td>
<td>93.93 ± 0.12</td>
<td>97.12 ± 0.19</td>
<td>70.92 ± 0.04</td>
</tr>
<tr>
<td>GOAT</td>
<td>90.96 ± 0.90</td>
<td>92.96 ± 1.48</td>
<td>94.21 ± 0.38</td>
<td>96.24 ± 0.24</td>
<td>72.41 ± 0.40</td>
</tr>
<tr>
<td>ExpFormer+GCN</td>
<td>91.59 ± 0.31</td>
<td>95.27 ± 0.42</td>
<td>95.77 ± 0.15</td>
<td>97.16 ± 0.13</td>
<td>72.44 ± 0.28</td>
</tr>
<tr>
<td>ExpFormer</td>
<td>91.16 ± 0.26</td>
<td>95.36 ± 0.17</td>
<td>95.19 ± 0.26</td>
<td>96.40 ± 0.20</td>
<td>71.27 ± 0.27</td>
</tr>
<tr>
<td>SpecFormer</td>
<td>90.93 ± 0.08</td>
<td>95.24 ± 0.12</td>
<td>95.00 ± 0.15</td>
<td>96.70 ± 0.05</td>
<td>70.82 ± 0.24</td>
</tr>
<tr>
<td>SpecFormer</td>
<td>90.93 ± 0.08</td>
<td>95.24 ± 0.12</td>
<td>95.00 ± 0.15</td>
<td>96.70 ± 0.05</td>
<td>70.82 ± 0.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Actor</th>
<th>Minesweeper</th>
<th>Tolokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloGNN</td>
<td>36.4 ± 1.6</td>
<td>51.08 ± 1.23</td>
<td>73.39 ± 1.17</td>
</tr>
<tr>
<td>SGC</td>
<td>27.0 ± 0.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GCN</td>
<td>33.23 ± 1.16</td>
<td>89.75 ± 0.52</td>
<td>83.64 ± 0.67</td>
</tr>
<tr>
<td>GraphGPS</td>
<td>-</td>
<td>90.63 ± 0.67</td>
<td>83.71 ± 0.48</td>
</tr>
<tr>
<td>NAGFormer</td>
<td>-</td>
<td>84.19 ± 0.66</td>
<td>78.32 ± 0.95</td>
</tr>
<tr>
<td>NodeFormer</td>
<td>36.9 ± 1.0</td>
<td>86.71 ± 0.88</td>
<td>78.10 ± 1.03</td>
</tr>
<tr>
<td>GAT</td>
<td>-</td>
<td>81.09 ± 1.02</td>
<td>83.11 ± 1.04</td>
</tr>
<tr>
<td>ExpFormer+GAT</td>
<td>38.68 ± 0.38</td>
<td>90.74 ± 0.53</td>
<td>83.77 ± 0.78</td>
</tr>
<tr>
<td>ExpFormer</td>
<td>39.01 ± 0.69</td>
<td>92.26 ± 0.56</td>
<td>83.53 ± 0.28</td>
</tr>
<tr>
<td>SpecFormer</td>
<td>38.44 ± 0.42</td>
<td>90.72 ± 0.06</td>
<td>83.15 ± 0.12</td>
</tr>
</tbody>
</table>

Figure 3: Comparing memory usage of our model with the ExpFormer with expander degree 30.
<table>
<thead>
<tr>
<th>Model</th>
<th>ogbn-proteins</th>
<th>Amazon2M</th>
<th>Pokec</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>72.04 ± 0.48</td>
<td>63.46 ± 0.10</td>
<td>60.15 ± 0.03</td>
</tr>
<tr>
<td>GCN</td>
<td>72.51 ± 0.35</td>
<td>83.90 ± 0.10</td>
<td>62.31 ± 1.13</td>
</tr>
<tr>
<td>SGC</td>
<td>70.31 ± 0.23</td>
<td>81.21 ± 0.12</td>
<td>52.03 ± 0.84</td>
</tr>
<tr>
<td>GCN-NSAMPLER</td>
<td>73.51 ± 1.31</td>
<td>83.84 ± 0.42</td>
<td>63.75 ± 0.77</td>
</tr>
<tr>
<td>GAT-NSAMPLER</td>
<td>74.63 ± 1.24</td>
<td>85.17 ± 0.32</td>
<td>62.32 ± 0.65</td>
</tr>
<tr>
<td>SIGN</td>
<td>71.24 ± 0.46</td>
<td>80.98 ± 0.31</td>
<td>68.01 ± 0.25</td>
</tr>
<tr>
<td>NodeFormer</td>
<td>77.45 ± 1.15</td>
<td>87.85 ± 0.24</td>
<td>70.32 ± 0.45</td>
</tr>
<tr>
<td>SGFormer</td>
<td>79.53 ± 0.38</td>
<td>89.09 ± 0.10</td>
<td>73.76 ± 0.24</td>
</tr>
<tr>
<td>SpeXPhormer</td>
<td><strong>80.66 ± 0.21</strong></td>
<td><strong>90.32 ± 0.01</strong></td>
<td><strong>74.73 ± 0.04</strong></td>
</tr>
</tbody>
</table>

**Table 2:** Comparative results on large graph datasets, with ROC-AUC reported for the ogbn-proteins dataset and accuracy for all others.
Thanks!

- Graph transformers: like regular transformers, but on graphs
- Full pairwise attention way too expensive for large graphs

**Exphormers: Sparse Transformers for Graphs**

- Exphormer
  - Augment attention graph with expanders + maybe virtual nodes
- Spexphormer
  - Sparsify the augmented graph with a pilot network