

Training Graph Neural Networks with 1000 Layers

Guohao Li^{1,2} Matthias Müller¹ Bernard Ghanem² Vladlen Koltun¹

Abstract

Deep graph neural networks (GNNs) have achieved excellent results on various tasks on increasingly large graph datasets with millions of nodes and edges. However, memory complexity has become a major obstacle when training deep GNNs for practical applications due to the immense number of nodes, edges, and intermediate activations. To improve the scalability of GNNs, prior works propose smart graph sampling or partitioning strategies to train GNNs with a smaller set of nodes or sub-graphs. In this work, we study reversible connections, group convolutions, weight tying, and equilibrium models to advance the memory and parameter efficiency of GNNs. We find that reversible connections in combination with deep network architectures enable the training of overparameterized GNNs that significantly outperform existing methods on multiple datasets. Our models RevGNN-Deep (1001 layers with 80 channels each) and RevGNN-Wide (448 layers with 224 channels each) were both trained on a single commodity GPU and achieve an ROC-AUC of 87.74 ± 0.13 and 88.24 ± 0.15 on the *ogbn-proteins* dataset. To the best of our knowledge, RevGNN-Deep is the deepest GNN in the literature by one order of magnitude.

1. Introduction

Graphs are all around us. Whether we watch a show on Netflix, browse through friends' feeds on Facebook, buy something on Amazon, or look up a researcher on Google, chances are that our actions trigger queries on a large graph. Movie and book databases are often encoded as knowledge graphs for efficient recommendation systems. Social media services rely on social graphs. Shopping platforms leverage product co-purchasing networks to boost sales. Citation indices like Web of Science, Scopus, and Google Scholar

¹Intel Labs ²King Abdullah University of Science and Technology. Correspondence to: Guohao Li <guohao.li@kaust.edu.sa>.

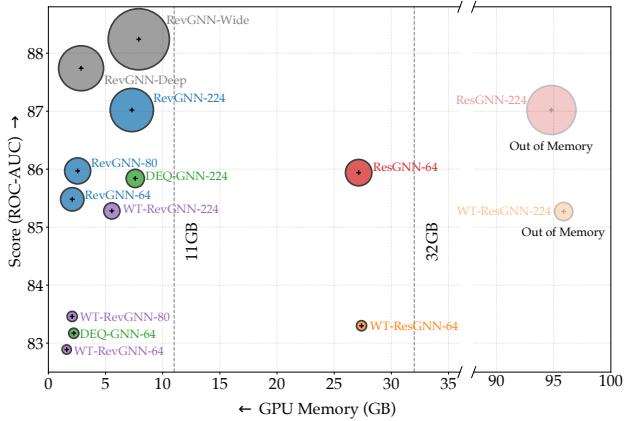


Figure 1. ROC-AUC score vs. GPU memory consumption on the *ogbn-proteins* dataset. We find that deep reversible GNNs are very powerful and outperform existing models by a margin; our best models are RevGNN-Deep and RevGNN-Wide. We also compare reversible connections (RevGNN-x), weight-tying (WT-x), and equilibrium models (DEQ-x) for 112-layer deep GNNs (x denotes the number of channels per layer). Reversible models consistently achieve the same or better performance as the baseline using only a fraction of the memory. Weight-tied and equilibrium models offer a good performance to parameter efficiency trade-off. Datapoint size is proportional to \sqrt{p} , where p is the number of parameters.

construct large citation graphs. Even the Internet itself is in essence a vast graph with billions of nodes and edges. Graphs are also powerful tools for representing 3D data such as point clouds and meshes or biological data such as molecular structures and protein interactions.

One prominent and powerful approach to process such graphs are graph neural networks (GNNs). GNNs have achieved impressive performance on relatively small graph datasets (Yang et al., 2016; Zitnik & Leskovec, 2017; Shchur et al., 2018). Unfortunately, the most interesting and impactful real-world problems rely on very large graphs where limited GPU memory quickly becomes a bottleneck. In order to train GNN models on large graphs, the common practice is to reduce the number of model parameters. This is counterproductive, since processing larger graphs would likely benefit from more parameters. While there is evidence that over-parameterized models generalize better (Neyshabur et al., 2019; Belkin et al., 2019), the relationship between performance and parameters is best illustrated in the context

of language modelling. Recent progress in natural language processing (NLP) has been enabled by a massive increase in parameter counts: GPT (110M) (Radford et al.), BERT (340M) (Devlin et al., 2019), GPT-3 (175B) (Brown et al., 2020), Gshard-M4 (600B) (Lepikhin et al., 2021), and DeepSpeed (1T) (Rasley et al., 2020). More parameters mean deeper or wider networks that consume more memory.

GNNs have shown a lot of promise on recent large-scale graph datasets such as Benchmarking GNNs (Dwivedi et al., 2020), Open Graph Benchmark (OGB) (Hu et al., 2020; 2021) and Microsoft Academic Graph (MAG) (Wang et al., 2020). Recent works (Li et al., 2019; 2020; Chen et al., 2020) have successfully trained deep models with a large number of parameters and achieved state-of-the-art performance. However, these models have large memory footprints and operate at the physical limits of current hardware. In order to apply deeper and wider GNNs with more parameters, we need either different hardware or more efficient architectures that consume less memory.

Existing works try to overcome the memory bottleneck by mini-batch training, *i.e.* sampling a smaller set of nodes (Hamilton et al., 2017; Chen et al., 2018b;a) or partitioning large graphs (Chiang et al., 2019; Zeng et al., 2020) into smaller subgraphs and sampling from those. These approaches have proven successful, but they introduce further hyperparameters that need to be tuned. For instance, if the sampled size of nodes or subgraphs is too small, it may break important structures in the graph. While these methods are a step in the right direction, they do not scale well as the models become deeper or wider, since memory consumption is still dependent on the number of layers. Another approach is efficient propagation via K -power adjacency matrices or graph diffusion matrices (Wu et al., 2019; Klicpera et al., 2019; Bojchevski et al., 2020; Liu et al., 2020; Frasca et al., 2020). However, the propagation schemes of these methods are non-trainable, which may lead to sub-optimality.

Inspired by efficient architectures from computer vision and natural language processing (Gomez et al., 2017; Xie et al., 2017; Bai et al., 2018; 2019), here we investigate several methods to obtain more efficient GNNs that use less memory than conventional architectures while achieving state-of-the-art results (see Figure 1). We explore grouped reversible graph connections in order to reduce the memory complexity with respect to the number of layers from $\mathcal{O}(L)$ to $\mathcal{O}(1)$. In other words, the memory consumption is independent of the depth. This allows us to train very deep, over-parameterized models with constant memory consumption.

We also investigate parameter-efficient architectures. We study deep weight-tied GNNs that have the parameter cost of only a single layer. We also develop a deep graph equilibrium GNN, which is essentially a weight-tied network with infinite depth. We directly solve for the equilibrium point of

this infinite-layer network using a root-finding method. We backpropagate through the equilibrium point using implicit differentiation. Hence, we do not need to store intermediate states and get an infinite-depth network at the memory and parameter cost of a single layer.

Our analysis of these methods shows that deep reversible architectures are the most powerful in terms of achieving state-of-the-art performance on benchmark datasets. This is due to their very large capacity at low memory cost and only slightly increased training time. Weight-tied models offer constant parameter size regardless of depth. However, due to the smaller number of parameters, performance on large datasets suffers and has to be compensated by increasing the width. Finally, graph equilibrium models have the same memory efficiency as reversible models and the same parameter efficiency as weight-tied models. They perform similarly to weight-tied models, and the training time *vs.* performance tradeoff can be further adjusted by tuning the number of iterations in each optimization step.

Our methods can be applied to different GNN operators. In our experiments, we successfully apply them to GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2018), and DeepGCN (Li et al., 2019). We can also combine the proposed techniques with sampling-based methods to further reduce memory and boost performance on some datasets. To the best of our knowledge, we are the first to train a GNN with more than 1000 layers. Our model RevGNN-Deep, outperforms all state-of-the-art approaches on the *ogbn-proteins* dataset (Hu et al., 2020) with an ROC-AUC of 87.74%, while only consuming 2.86 GB of GPU memory during training, one order of magnitude less than the current top performer. We can also trade our memory savings for larger width, pushing performance to new heights. Our RevGNN-Wide achieves an ROC-AUC of 88.24% on the *ogbn-proteins* dataset, ranking first on the leaderboard by a large margin.

In summary, we investigate several techniques to increase the memory efficiency of GNNs and perform an extensive analysis. We significantly outperform current state-of-the-art methods on several datasets by employing reversible connections to train deeper and wider models. Further, we demonstrate the generality of these techniques by applying them to multiple GNN operators. We release our implementation, which supports PyTorch Geometric (Fey & Lenssen, 2019) and the Deep Graph Library (Wang et al., 2019a).

2. Related Work

GNNs were introduced by Gori et al. (2005) and Scarselli et al. (2008) for learning a graph representation by finding stable states through fixed-point iterations. Bruna et al. (2014) generalized Convolutional Neural Networks (CNNs)

to graphs using the Fourier basis by computing the eigen-decomposition of the graph Laplacian. Follow-up works (Henaff et al., 2015; Defferrard et al., 2016; Kipf & Welling, 2017; Levie et al., 2018; Li et al., 2018) propose different ways to improve and extend spectral GNNs. For instance, Kipf & Welling (2017) simplify spectral graph convolutions by limiting the filters to operate on 1-hop neighborhoods. Instead of defining GNNs from the spectral view, Gilmer et al. (2017) introduce a general framework termed Message Passing Neural Networks (MPNNs) to unify GNNs. MPNNs define convolutions on the graph by propagating messages from spatial neighbors. Many subsequent works (Hamilton et al., 2017; Monti et al., 2017; Niepert et al., 2016; Gao et al., 2018; Xu et al., 2019; Veličković et al., 2018; Wang et al., 2019b) fall into this framework.

Most prior state-of-the-art (SOTA) works are limited to shallow depths due to the over-smoothing (Li et al., 2018) and vanishing gradient problems (Li et al., 2019; 2021) present in GNNs. Current works have explored different approaches to resolve these difficulties in training deep GNNs. The first line of research employs skip connections across layers. Xu et al. (2018) propose jump connections to adaptively select intermediate representations to the last layer. Li et al. (2019) adapt residual and dense connections (He et al., 2016; Huang et al., 2017) and dilated convolutions (Yu & Koltun, 2016) from CNNs to GNNs; they successfully train GNNs with up to 56 layers on 3D point clouds and achieve SOTA results. Li et al. (2020) train a 112-layer residual GNN and achieve SOTA performance on the large-scale *ogbn-proteins* dataset (Hu et al., 2020). Other works also demonstrate that residual connections aid in training deeper GNNs (Gong et al., 2020; Chen et al., 2020; Xu et al., 2021). We push further along the path of deep residual GNNs and train a SOTA model with more than 1000 layers by addressing the memory bottleneck of current approaches.

Researchers have also studied different normalization and regularization techniques such as DropEdge (Rong et al., 2020), DropConnect (Hasanzadeh et al., 2020), PairNorm (Zhao & Akoglu, 2019), weight normalization (Oono & Suzuki, 2019), differentiable group normalization (Zhou et al., 2020), and GraphNorm (Cai et al., 2020). Another line of work (Wu et al., 2019; Klicpera et al., 2019; Bojchevski et al., 2020; Liu et al., 2020; Frasca et al., 2020) proposes an efficient propagation scheme to avoid stacking layers by incorporating multi-hop information into a single GNN layer via K -th power adjacency matrices or personalized PageRank diffusion matrices. However, the propagation scheme of these methods is non-trainable, thus making it difficult to learn hierarchical features and limiting capacity.

Many applications yield huge graphs with millions of nodes and edges. Memory limitations preclude full-batch training in such settings. Specifically, to train an L -layer GNN with

D hidden channels on a graph \mathcal{G} with N nodes and M edges, the memory complexity of storing activations is $\mathcal{O}(LND)$. GraphSAGE (Hamilton et al., 2017) incorporates a recursive node-wise sampling scheme to improve the scalability of GNNs. Instead of training on the whole graph, GraphSAGE uniformly samples a fixed number of neighbors for a batch of nodes. However, the recursive neighborhood leads to exponential memory complexity as the number of GNN layers increases. If B is the batch size of nodes and R is the number of sampled neighbors per node, the memory complexity of GraphSAGE is $\mathcal{O}(R^L BD)$. VR-GCN (Chen et al., 2018b) enhances GraphSAGE by reducing the variance of mini-batch training to increase the convergence rate. However, to reduce the estimation variance, historical activations need to be stored in memory ($\mathcal{O}(LND)$).

To avoid recursive neighborhood expansion, FastGCN (Chen et al., 2018a) performs layer-wise sampling to sub-sample nodes for each layer independently with a degree-based distribution. While this further reduces the memory complexity to $\mathcal{O}(LRBD)$, it leads to sparse connections. To better maintain the graph structure, Cluster-GCN (Chiang et al., 2019) and GraphSAINT (Zeng et al., 2020) propose subgraph-wise sampling methods for mini-batch training with memory complexity $\mathcal{O}(LBD)$. However, even with smart sampling methods, GNNs still do not scale as the number of layers increases (*i.e.* the L dimension). We tackle this issue and study orthogonal approaches that eliminate the L dimension from the memory complexity. Our techniques enable unlimited network depth at no memory cost and can be combined with existing sampling methods.

3. Building Deep GNNs

3.1. Preliminaries

A graph \mathcal{G} is represented by a tuple $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$ is an unordered set of vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges. Let N and M denote the number of vertices and edges, respectively. For convenience, a graph can be equivalently defined as an adjacency matrix $A \in \mathcal{A} \subset \mathbb{R}^{N \times N}$, where $a_{i,j}$ denotes the link relation between vertex v_i and v_j . In some scenarios, vertices and edges are associated with a vertex feature matrix $X \in \mathcal{X} \subset \mathbb{R}^{N \times D}$ and an edge feature matrix $U \in \mathcal{U} \subset \mathbb{R}^{M \times F}$, respectively. We use GNN operators that map the vertex feature matrix X , the adjacency matrix A , and the edge feature matrix U (optional) to a transformed vertex feature matrix X' :

$$f_w : \mathcal{X} \times \mathcal{A} \times \mathcal{U} \mapsto \mathcal{X}', \quad (1)$$

where $f_w(X, A, U)$ is parameterized by learnable parameters w . For simplicity, we assume that the transformed vertex feature matrix X' has the same dimension as the input vertex feature matrix X . We also assume that the

adjacency matrix A is the same for all GNN layers. When the edge feature matrix U is present, it is fed into each layer with its initial values $U^{(0)}$.

3.2. Over-parameterized GNNs

Deep GNNs. Recent works (Li et al., 2019; 2020) show how adding residual connections (He et al., 2016) to vertex features ($X' = f_w(X, A, U) + X$) enables training of deep GNNs that achieve promising results on graph datasets. However, the memory complexity of the activations remains $\mathcal{O}(LND)$, where L is the number of GNN layers, N is the number of vertices, and D is the size of vertex features. Hence, the memory consumption of deep GNNs scales linearly with the number of layers. Since the memory footprint of the network parameters is usually negligible, we focus on memory consumption induced by the activations.

Grouped Reversible GNNs. Inspired by reversible networks (Gomez et al., 2017; Liu et al., 2019; Kitaev et al., 2019) and grouped convolutions (Krizhevsky et al., 2012; Xie et al., 2017), we generalize reversible residual connections to grouped reversible residual connections for GNNs. Specifically, the input vertex feature matrix X is uniformly partitioned into C groups $\langle X_1, X_2, \dots, X_C \rangle$ across the channel dimension, where $X_i \in \mathbb{R}^{N \times \frac{D}{C}}$. A grouped reversible GNN block operates on a group of inputs and produces a group of outputs: $\langle X_1, X_2, \dots, X_C \rangle \mapsto \langle X'_1, X'_2, \dots, X'_C \rangle$.

The forward pass is defined as follows:

$$X'_0 = \sum_{i=2}^C X_i \quad (2)$$

$$X'_i = f_{w_i}(X'_{i-1}, A, U) + X_i, \quad i \in \{1, \dots, C\}, \quad (3)$$

where X'_0 is designed for exchanging information across groups. Unlike conventional GNNs, grouped reversible GNNs only need to save the output vertex features of the last GNN block in GPU memory for backpropagation. Therefore, the memory complexity of activations becomes $\mathcal{O}(ND)$, which is independent of the depth of the network. Note that the adjacency matrix A and the edge feature matrix U are not updated during message passing. During the backward pass, only the input vertex features are reconstructed, on the fly, from the output vertex features $\langle X'_1, X'_2, \dots, X'_C \rangle$ for backpropagation:

$$X_i = X'_i - f_{w_i}(X'_{i-1}, A, U), \quad i \in \{2, \dots, C\} \quad (4)$$

$$X'_0 = \sum_{i=2}^C X_i \quad (5)$$

$$X_1 = X'_1 - f_{w_1}(X'_0, A, U). \quad (6)$$

In practice, X_i for $i \in \{2, \dots, C\}$ can be computed in parallel. To reconstruct X_1 , X'_0 needs to be computed in advance. After reconstructing the original input vertex features, gradients can be derived through backpropagation.

Owing to the group processing, the number of parameters reduces as the group size increases. Note that in the special case where the group size $C = 2$, we obtain a similar form to the reversible residual connections proposed for CNNs (Gomez et al., 2017). The definition above is independent of the choice of f_w . However, we find that normalization layers and dropout layers are essential for training deep GNNs. To avoid extra memory usage, normalization layers and dropout layers are embedded into the reversible GNN block. The GNN block f_{w_i} is designed similar to the pre-activation residual GNN block proposed by Li et al. (2020):

$$\hat{X}_i = \text{Dropout}(\text{ReLU}(\text{Norm}(X'_{i-1}))) \quad (7)$$

$$\tilde{X}_i = \text{GraphConv}(\hat{X}_i, A, U). \quad (8)$$

The stochasticity of vanilla dropout layers causes reconstruction errors in the reverse pass. A naive solution would be to store the dropout pattern for all layers. However, the dropout patterns have the same dimension as the activations, which would lead to $\mathcal{O}(LND)$ memory consumption. As an alternative, we adopt a modified dropout layer in which the dropout pattern is shared across layers. Therefore, we only need to store one dropout pattern in every SGD iteration; its memory complexity is independent of the depth: $\mathcal{O}(ND)$. During the reverse pass, the saved dropout pattern is reactivated to reconstruct the input vertex features.

3.3. Parameter-Efficient GNNs

Weight-tied GNNs. Weight-tying is a powerful tool for improving the parameter efficiency of language models (Press & Wolf, 2017; Inan et al., 2017; Bai et al., 2018). We take these works as inspiration to study how weight-tying can improve the parameter efficiency of GNNs. Instead of having different weights for different GNN blocks, the weight-tied GNN shares weights across layers. The GNN model proposed by Scarselli et al. (2008) can be considered the first weight-tied GNN, in which a learned transition function is used to find stable node states by Banach's fixed point theorem (Khamsi & Kirk, 2001). Here we experiment with weight-tied residual GNNs and weight-tied reversible GNNs. For weight-tied residual GNNs, we define:

$$f_w^{(1)} := f_w^{(2)} \dots := f_w^{(L)}, \quad (9)$$

where L is the number of layers. For weight-tied reversible GNNs, weights are shared in a group-wise manner:

$$f_{w_i}^{(1)} := f_{w_i}^{(2)} \dots := f_{w_i}^{(L)}, \quad i \in \{1, \dots, C\} \quad (10)$$

Note that we use the same pre-activation GNN block described in Section 3.2 instead of a contraction map used by Scarselli et al. (2008). Both weight-tied GNNs have explicit layers and are trained with backpropagation. But the weight-tied reversible GNN reconstructs input vertex features on the

fly during backpropagation. Therefore, the memory complexities of the weight-tied residual GNN and weight-tied reversible GNN are $\mathcal{O}(LND)$ and $\mathcal{O}(ND)$, respectively.

Deep Equilibrium GNNs. An alternative way to train weight-tied GNNs with $\mathcal{O}(ND)$ memory consumption is to use implicit differentiation (Scarselli et al., 2008; Bai et al., 2019), which assumes that the network can reach an equilibrium state. We construct a GNN model that is assumed to converge to a fixed point Z^* for any given input:

$$Z^* = f_w^{\text{DEQ}}(Z^*, X, A, U), \quad (11)$$

where the state Z represents the transformed node features. To construct a stable or contractive GNN block, we mimic the design of MDEQ (Bai et al., 2020). We build a GNN block as follows:

$$Z' = \text{GraphConv}(Z_{\text{in}}, A, U) \quad (12)$$

$$Z'' = \text{Norm}(Z' + X) \quad (13)$$

$$Z''' = \text{GraphConv}(\text{Dropout}(\text{ReLU}(Z'')), A, U) \quad (14)$$

$$Z_o = \text{Norm}(\text{ReLU}(Z''' + Z')), \quad (15)$$

where Z_{in} is the input node state, Z_o is the output node state, X serves as the injected input, and Z' forms an internal residual signal to the output Z_o . In practice, X represents the initial node features and Z_{in} is initialized to zeros for the first iteration. Similar to DEQ (Bai et al., 2019), the forward pass of DEQ-GNN is implemented with a root-finding algorithm (*e.g.* Broyden’s method) and the gradients are obtained by implicitly differentiating through the equilibrium node state for the backward pass.

3.4. Analysis of Different Deep GNN Architectures

In the following, we compare the different techniques for building deep GNNs with respect to their performance, memory efficiency and parameter efficiency on the *ogbn-proteins* dataset from the *Open Graph Benchmark* (OGB) (Hu et al., 2020). We use the same GNN operator (Li et al., 2020), hyper-parameters (*e.g.* learning rate, dropout rate, training epoch, *etc.*), and optimizers to make the comparison as fair as possible. We use mini-batch training with random partitioning where graphs are split into 10 parts during training and 5 parts during testing (Li et al., 2020). One subgraph is sampled to form a mini-batch at each step. The reported GPU memory corresponds to the peak GPU memory usage during the first training epoch.

Baseline GNN. We use a recent pre-activation residual GNN (Li et al., 2020), which achieves state-of-the-art performance across several OGB datasets, as our baseline. For simplicity, we use the *max* aggregation function for all the ablated models on the *ogbn-proteins* dataset. We refer to the baseline model as ResGNN. As shown in Figure 2, the

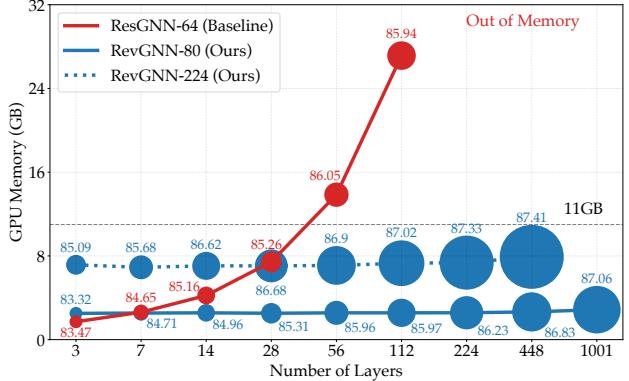


Figure 2. GPU memory consumption vs. number of layers for ResGNN (Li et al., 2020) and RevGNN, our adaptation with reversible connections. RevGNN uses constant memory for deeper networks with more parameters and better performance. We use width 64 and 80 for ResGNN and RevGNN, respectively, to ensure a similar number of parameters per network depth. Datapoints are annotated with the ROC-AUC score of the model and their size is proportional to \sqrt{p} , where p is the number of parameters.

ResGNN with 112 layers and a channel size of 64 achieves 85.94% ROC-AUC on *ogbn-proteins*. However, the memory consumption of ResGNN-64 increases linearly with the number of layers. ResGNN-64 runs out of memory beyond 112 layers, making it impossible to investigate deeper models with current commodity hardware.

Reversible GNN. Our reversible GNN enables training of very deep networks with constant memory footprint, as illustrated in Figure 2 (RevGNN). We use a group size of 2; since grouping reduces the number of parameters, RevGNN-80 has roughly the same number of parameters as ResGNN-64 for the same number of layers. While the baseline model ResGNN-64 cannot go beyond 112 layers due to memory limitations, RevGNN-80 can go to more than 1000 layers without additional memory cost and achieves much better accuracy (87.06% ROC-AUC). We can invest the saved GPU memory to increase the network width and train a higher-capacity model. This model is not only deep (448 layers) but also wide (224 channels) and further improves performance (87.41% ROC-AUC).

Weight-tied GNN. We compare the weight-tied ResGNN (WT-ResGNN) and its reversible counterpart RevGNN (WT-RevGNN) in Figure 3. Both models have approximately the same number of parameters and achieve a similar accuracy while the reversible model uses significantly less GPU memory. Since the parameters are shared across GNN layers, the number of parameters stays constant as the number of layers increases. For example, with 112 layers, WT-RevGNN-224 has only 337k parameters, while RevGNN-224 has 17.1M parameters. However, training time and GPU memory consumption are similar while the performance of the weight-

tied model is worse (85.28% vs. 87.41%) due to diminishing returns after more than 7 layers. Similar to the results in Figure 2, this is evidence of a clear correlation between the number of parameters and performance.

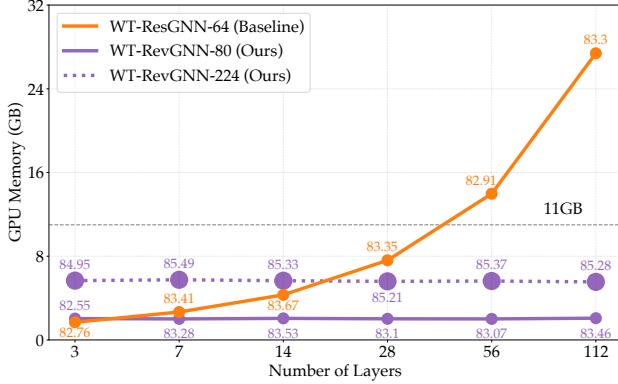


Figure 3. GPU memory consumption vs. number of layers for WT-ResGNN and WT-RevGNN. For a given filter size, the number of parameters is constant regardless of network depth. Our reversible architecture has a constant memory footprint regardless of network depth, while yielding similar accuracy. Datapoints are annotated with the ROC-AUC score of the model and their size is proportional to \sqrt{p} , where p is the number of parameters.

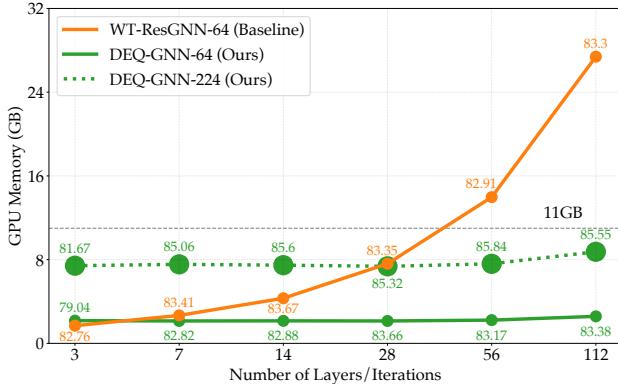


Figure 4. GPU memory vs. number of layers/iterations for ResGNN and DEQ-GNN. Datapoints are annotated with the ROC-AUC score of the model and their size is proportional to \sqrt{p} , where p is the number of parameters.

Equilibrium GNN. Equilibrium networks implicitly model a weight-tied network with infinite depth. As a result, they only have the parameter and memory footprint of a single layer, but the expressiveness of a very deep network. The initial node features X are used as input injection and the initial node states Z are set to zero. We implement our DEQ-GNN based on the original DEQ implementation for CNNs (Bai et al., 2019). Broyden’s root-finding method is used (Broyden, 1965) in both the forward pass and the backward pass to find the equilibrium node states and approximate the inverse Jacobian. The Broyden iterations

terminate when the norm of the objective is smaller than a tolerance ϵ or a maximum iteration threshold is reached. ϵ is set to $10^{-6} \cdot \sqrt{BD}$ and $2 \times 10^{-10} \cdot \sqrt{BD}$ for the forward pass and the backward pass respectively, where B is the number of nodes in the sampled subgraph and D is the channel size. The iteration thresholds in the forward pass and the backward pass are set to the same value. We examine different iteration thresholds for DEQ-GNN with a channel size of 64 (DEQ-GNN-64) and a channel size of 224 (DEQ-GNN-224) in Figure 4. It can be seen that the iteration threshold is crucial for good performance, since it affects the convergence to the equilibrium. We find that DEQ-GNN-64 performs similarly to WT-RevGNN-80 with nearly the same number of parameters and memory consumption. The wider DEQ-GNN-224 model reaches 85.84% ROC-AUC, which is comparable to the 112-layer ResGNN-64, with only around 28% memory footprint and 23% parameters (Mem: 7.60GB vs. 27.1GB, Params: 537k vs. 2.37M). DEQ-GNN-224 slightly outperforms WT-RevGNN-224 with only about 60% of the training time.

Discussion. Reversible networks emerge as the most promising approach for building deep GNNs that achieve state-of-the-art performance. They consume much less memory than the baseline while achieving comparable performance with the same number of parameters (see Figure 1). However, in contrast to the baseline method, we are able to train very deep networks with more parameters and much better performance without running out of memory. While it is possible to go to arbitrary depths with constant memory consumption, the training time increases. In order to reduce the number of parameters and inference time, it is possible to increase the group size. However, we find that group sizes larger than 2 do not lead to a performance increase in practice and may even degrade model performance. We conjecture that this is due to the increased receptive field of early layers and the smaller number of parameters per memory. We provide an ablation study in the appendix.

The weight-tied network limits the number of parameters to a single layer regardless of the effective depth. We find that going deeper with tied weights boosts performance, but returns eventually diminish (see Figure 3). While parameters stay constant, going deeper still increases memory consumption, unless the reversible GNN block is used. An extension of the weight-tied network is the graph equilibrium model. It represents an infinite-depth weight-tied network and uses fixed-point iterations to solve for the optimal weights. This allows for much wider channel size with the same amount of memory. DEQ-GNNs are faster to train, but have more hyper-parameters to tune. Please refer to the appendix for a training time ablation. We find that pretraining is not necessary, but can further improve the results. The results reported in this paper are without pretraining for fair comparison. While weight-tied GNNs and equilibrium GNNs

are not able to achieve state-of-the-art performance, they are very parameter-efficient, which may be relevant for applications on embedded devices where model size matters.

4. Over-parameterized Deep GNNs

We conduct experiments on several datasets from the *Open Graph Benchmark* (OGB) (Hu et al., 2020). We first show state-of-the art results on several datasets with the proposed deep reversible GNN. We then apply the reversible GNN framework to different GNN operators on the *ogbn-arxiv* dataset. To show how mini-batch sampling can further aid the training of deep GNNs, we compare full-batch and mini-batch training of reversible GNNs on the *ogbn-products* dataset. The data splits and evaluation metrics on all datasets follow the OGB evaluation protocol. Mean and standard deviation are obtained across 10 trials. All the ablated models use the same hyper-parameters (*e.g.* learning rate, dropout rate, training epoch, *etc.*) and optimizers as the baseline models. The implementation of all the reversible models is based on PyTorch (Paszke et al., 2019) and supports both PyTorch Geometric (PyG) (Fey & Lenssen, 2019) and Deep Graph Library (DGL) (Wang et al., 2019a) frameworks.

4.1. State-of-the-art Results

We briefly describe two variants of our RevGNN with reversible connections, which reach new SOTA results on the *ogbn-proteins* dataset of the OGB leaderboard (Hu et al., 2020) (Table 1) at the time of submission. RevGNN-Deep has 1001 layers and a channel size of 80. It outperforms the previous SOTA method UniMP+CEF by 0.83% ROC-AUC, while using only 10.5% of the GPU memory for training. RevGNN-Wide uses 448 GNN layers and 224 hidden channels and significantly outperforms UniMP+CEF by 1.33% ROC-AUC with about 29% of the GPU memory. RevGNN-Deep uses the same training setting as mentioned in Section 3.4. The RevGNN-Wide uses a larger dropout rate of 0.2 to prevent overfitting. To boost the performance, the results of RevGNN-Deep and RevGNN-Wide are obtained using multi-view inferences with 10 views on larger subgraphs with a partition size of 3. We perform the inferences on a NVIDIA RTX A6000 (48GB). Please refer to the appendix for the details of multi-view inference. Larger and deeper models incur a cost in terms of training and inference time. RevGNN-Deep and RevGNN-Wide take 13.5 days and 17.1 days, respectively, to train for 2000 epochs on a single NVIDIA V100. Nonetheless, it is affordable for accuracy-critical applications in scientific research such as predicting protein structures (Senior et al., 2020). We demonstrate that it is possible to train huge over-parameterized GNN models on a single GPU. The RevGNN-Wide model has 68.47 million parameters, which is about half the size of the GPT language model (Radford et al.). We believe that this is

Table 1. Results on the *ogbn-proteins* dataset compared to SOTA. RevGNN-Deep has 1001 layers with 80 channels each. It achieves SOTA performance with minimal GPU memory for training. RevGNN-Wide has 448 layers with 224 channels each. It achieves the best accuracy while consuming a moderate amount of GPU memory.

| Model | ROC-AUC \uparrow | Mem \downarrow | Params |
|-----------------------------|-------------------------|------------------|--------|
| GCN (Kipf & Welling) | 72.51 ± 0.35 | 4.68 | 96.9k |
| GraphSAGE (Hamilton et al.) | 77.68 ± 0.20 | 3.12 | 193k |
| DeeperGCN (Li et al.) | 86.16 ± 0.16 | 27.1 | 2.37M |
| UniMP (Shi et al.) | 86.42 ± 0.08 | 27.2 | 1.91M |
| GAT (Veličković et al.) | 86.82 ± 0.21 | 6.74 | 2.48M |
| UniMP+CEF (Shi et al.) | 86.91 ± 0.18 | 27.2 | 1.96M |
| Ours (RevGNN-Deep) | 87.74 ± 0.13 | 2.86 | 20.03M |
| Ours (RevGNN-Wide) | 88.24 ± 0.15 | 7.91 | 68.47M |

an important step forward in developing over-parameterized GNNs for graphs.

Our RevGNNs also achieve new SOTA results on the *ogbn-arxiv* dataset (see Table 2). RevGCN-Deep uses 28 GCN (Kipf & Welling, 2017) layers with 128 channels each and achieves an accuracy of 73.01%, while only using 1.84GB of GPU memory. RevGAT-Wide uses 5 attention layers with 3 heads and 356 channels for each head and outperforms the current top performer UniMP_v2 (Shi et al., 2020) (74.05% *vs.* 73.97%) while using about a third of the memory (8.49GB *vs.* 25.0GB). RevGAT-SelfKD uses self-knowledge distillation (Zhang et al., 2019) with 5 attention layers, 3 heads, and 256 channels each. The teacher models achieve an accuracy of 74.02%. After training with distillation, the student models set a new SOTA with 74.26% test accuracy. Please refer to the appendix for more details.

4.2. Application to Different GNN Operators

The proposed techniques are generic and can in principle be applied to any SOTA GNN to further boost the performance with deeper and wider architectures while saving GPU memory. We show this for the example of reversible GNNs and build RevGNNs with different SOTA GNN operators: GAT (Veličković et al., 2018), GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), and ResGEN (Li et al., 2020). We compare them to their non-reversible residual counterparts on the *ogbn-arxiv* in Table 3. Since *ogbn-arxiv* is much smaller than the *ogbn-proteins*, we are able to run all experiments with full-batch training and report statistics across 10 training runs. We observe that all of the RevGNNs consistently outperform their vanilla residual counterparts with the same channel size. The RevGNNs use less memory due to reversible connections and fewer parameters due to grouping. We increase the channel size of RevGNNs to roughly match the number of parameters of the

Table 2. Results on the *ogbn-arxiv* dataset compared to SOTA. RevGCN-Deep has 28 layers with 128 channels each. It achieves SOTA performance with minimal GPU memory. RevGAT-Wide has 5 layers with 1068 channels each. RevGAT-SelfKD denotes the student models with 5 layers and 768 channels. It achieves the best accuracy while consuming a moderate amount of GPU memory.

| Model | ACC \uparrow | Mem \downarrow | Params |
|-----------------------------|------------------------------------|------------------|--------|
| GraphSAGE (Hamilton et al.) | 71.49 ± 0.27 | 1.99 | 219k |
| GCN (Kipf & Welling) | 71.74 ± 0.29 | 1.90 | 143k |
| DAGNN (Liu et al.) | 72.09 ± 0.25 | 2.40 | 43.9k |
| DeeperGCN (Li et al.) | 72.32 ± 0.27 | 21.6 | 491k |
| GCNII (Chen et al.) | 72.74 ± 0.16 | 17.0 | 2.15M |
| GAT (Veličković et al.) | 73.91 ± 0.12 | 5.52 | 1.44M |
| UniMP_v2 (Shi et al.) | 73.97 ± 0.15 | 25.0 | 687k |
| Ours (RevGCN-Deep) | 73.01 ± 0.31 | 1.84 | 262k |
| Ours (RevGAT-Wide) | 74.05 ± 0.11 | 8.49 | 3.88M |
| Ours (RevGAT-SelfKD) | 74.26 ± 0.17 | 6.60 | 2.10M |

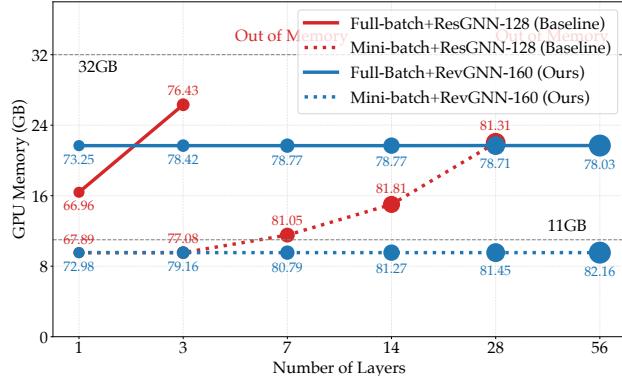


Figure 5. GPU memory consumption vs. number of layers for ResGNN (Li et al., 2020) and RevGNN with full-batch and mini-batch training on the *ogbn-products*. RevGNN uses constant memory for deeper networks with more parameters and better performance. We use a width of 128 and 160 for ResGNN and RevGNN, respectively, to ensure a similar number of parameters per network depth. Datapoints are annotated with the accuracy of the model and their size is proportional to \sqrt{p} , where p is the number of parameters.

corresponding ResGNNs, thus increasing the performance gap further. For instance, the RevGCN with 28 layers and 180 channels reduces the memory footprint by more than 75% while improving the accuracy by 0.76%, as compared to the ResGCN with 28 layers and 128 channels. Utilizing label propagation, the RevGAT with 5 layers and 1068 channels (3 attention heads with 356 channels for each head) achieves SOTA results on *ogbn-arxiv*.

4.3. Full-batch vs. Mini-batch Training

Our method is orthogonal to existing sampling-based approaches, which also reduce memory consumption. Hence,

Table 3. Results with different GNN operators on the *ogbn-arxiv*. All GAT models use label propagation. #L and #Ch denote the number of layers and channels respectively. *Baselines* are in italic.

| Model | #L | #Ch | ACC \uparrow | Mem \downarrow | Params |
|----------------|----|------|------------------------------------|------------------|--------|
| <i>ResGCN</i> | 28 | 128 | 72.46 ± 0.29 | 11.15 | 491k |
| RevGCN | 28 | 128 | 73.01 ± 0.31 | 1.84 | 262k |
| RevGCN | 28 | 180 | 73.22 ± 0.19 | 2.73 | 500k |
| <i>ResSAGE</i> | 28 | 128 | 72.46 ± 0.29 | 8.93 | 950k |
| RevSAGE | 28 | 128 | 72.69 ± 0.23 | 1.17 | 491k |
| RevSAGE | 28 | 180 | 72.73 ± 0.10 | 1.57 | 953k |
| <i>ResGEN</i> | 28 | 128 | 72.32 ± 0.27 | 21.63 | 491k |
| RevGEN | 28 | 128 | 72.34 ± 0.18 | 4.08 | 262k |
| RevGEN | 28 | 180 | 72.93 ± 0.10 | 5.67 | 500k |
| <i>ResGAT</i> | 5 | 768 | 73.76 ± 0.13 | 9.96 | 3.87M |
| RevGAT | 5 | 768 | 74.02 ± 0.18 | 6.30 | 2.10M |
| RevGAT | 5 | 1068 | 74.05 ± 0.11 | 8.49 | 3.88M |

we can use our techniques in conjunction with mini-batch training to further optimize memory. We conduct an ablation study on the *ogbn-products* dataset (Hu et al., 2020) with full-batch training and a simple random-clustering mini-batch training for ResGNNs and RevGNNs. The test results are reported in Figure 5 with full-batch inference. Compared to full-batch training, we find that mini-batch training further reduces the memory consumption of RevGNN to 44% and improves its accuracy from 78.77% to 82.16%.

4.4. Analysis of Complexities

We have discussed the memory complexity of full-batch GNNs, GraphSAGE (Hamilton et al., 2017), VR-GCN (Chen et al., 2018b), FastGCN (Chen et al., 2018a), Cluster-GCN (Chiang et al., 2019) and GraphSAINT (Zeng et al., 2020) in the related work section, and the memory complexity of our Reversible GNN, Weight-tied GNN and DEQ-GNN in the methodology section. We summarize the theoretical memory complexity in Table 4, where L is the number of layers of the GNN, D is the size of hidden channels, N is of the number of nodes in the graph, B is the batch size of nodes and R is the number of sampled neighbors of each node. K is the maximum Broyden iterations for equilibrium GNNs. We only discuss the memory complexity for storing intermediate node features in each layer since the memory footprint of the network parameters is negligible. All prior works suffer from memory consumption with respect to the number of layers, while the memory consumption of our methods is independent of the depth. Our methods can also be combined with mini-batch sampling methods to further reduce the memory complexity with respect to the number of nodes. We also include the parameter complexity and time complexity in Table 4. Note that although the

Table 4. Comparison of complexities. L is the number of layers, D is the number of hidden channels, N is the number of nodes, B is the batch size of nodes and R is the number of sampled neighbors of each node. K is the maximum Broyden iterations.

| Method | Memory | Params | Time |
|-------------------------------|-----------------------|---------------------|-----------------------------------------------|
| Full-batch GNN | $\mathcal{O}(LND)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| GraphSAGE | $\mathcal{O}(R^L BD)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(R^L ND^2)$ |
| VR-GCN | $\mathcal{O}(LND)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2 + R^L ND^2)$ |
| FastGCN | $\mathcal{O}(LRBD)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(RLND^2)$ |
| Cluster-GCN | $\mathcal{O}(LBD)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| GraphSAINT | $\mathcal{O}(LBD)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| Weight-tied GNN | $\mathcal{O}(LND)$ | $\mathcal{O}(D^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| RevGNN | $\mathcal{O}(ND)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| WT-RevGNN | $\mathcal{O}(ND)$ | $\mathcal{O}(D^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| DEQ-GNN | $\mathcal{O}(ND)$ | $\mathcal{O}(D^2)$ | $\mathcal{O}(K \ A\ _0 D + KND^2)$ |
| RevGNN + Subgraph Sampling | $\mathcal{O}(BD)$ | $\mathcal{O}(LD^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| WT-RevGNN + Subgraph Sampling | $\mathcal{O}(BD)$ | $\mathcal{O}(D^2)$ | $\mathcal{O}(L \ A\ _0 D + LND^2)$ |
| DEQ-GNN + Subgraph Sampling | $\mathcal{O}(BD)$ | $\mathcal{O}(D^2)$ | $\mathcal{O}(K \ A\ _0 D + KND^2)$ |

time complexity of RevGNN is the same as for the vanilla GNNs, it has a larger constant term. This is due to the input that needs to be reconstructed during the backward pass. Both the memory complexity and parameter complexity of WT-RevGNN and DEQ-GNN is independent of L .

4.5. More Ablation Studies

Comparison with SGC & SIGN on Ogbn-products. We compare RevGNN with SGC (Wu et al., 2019) and SIGN (Frasca et al., 2020) on ogbn-products. The test accuracies (%) of SGC, SIGN and RevGNN on *ogbn-products* are: 74.87 ± 0.25 (SGC), 77.60 ± 0.13 (SIGN) and 82.16 ± 0.15 (RevGNN) respectively. The results of SGC and SIGN are obtained from the SIGN paper (Frasca et al., 2020). We train the RevGNN with 56 layers and 160 hidden channels in a random-clustering mini-batch training fashion. We find that RevGNN outperforms SGC and SIGN by a large margin.

Results of RevGNN on Ogbg-ppa. We trained ResGNN and RevGNN with 28 layers and similar number of parameters ($\sim 2.3M$) on the *ogbg-ppa* dataset. Both achieve around 77% test accuracy, but RevGNN uses only 16% GPU memory compared to ResGNN.

Comparison of Normalizations on Ogbg-molhiv. We conduct an ablation study on ogbg-molhiv for comparing using BatchNorm (Ioffe & Szegedy, 2015) or GraphNorm (Cai et al., 2020) for RevGNN on graph property prediction. We train RevGNN with 14 layers and 256 hidden channels. Dropout layers with a rate of 0.3 are used. Learnable SoftMax aggregation functions (Li et al., 2020) are used for message aggregation. The RevGNN with GraphNorm achieves 78.62 ROC-AUC which outperforms the RevGNN with BatchNorm (77.82 ROC-AUC) slightly.

5. Conclusion and Discussion

This work addresses a fundamental bottleneck of current deep GNNs: their high GPU memory consumption. We investigate several techniques to reduce the memory complexity with respect to network depth (L) from $\mathcal{O}(L)$ to $\mathcal{O}(1)$. In particular, reversible GNNs enable networks that are an order of magnitude deeper than current SOTA models. Since there is no memory cost associated with depth, the width can be increased for additional representational power. As a result, we can train over-parameterized networks that significantly outperform current models on standard benchmarks while consuming less memory.

However, this comes at an additional cost in training time. While this cost is less than 40%, it can extend the training time by several days on very large datasets. In addition, current GNNs are usually trained for thousands of epochs to achieve SOTA performance. Reducing the training time of GNNs is an interesting avenue for future research. Please refer to the appendix for more details about the opportunities for future work and further discussion on topics like gradient checkpointing, model parallelism, etc. In the appendix, we also provide all the experimental details and further ablation studies.

Acknowledgments

The authors thank Shaojie Bai, researchers at Intel ISL, the reviewers, and area chairs for their helpful suggestions. This project is partially funded by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research under Award No. OSR-CRG2019-4033.

References

- Bai, S., Kolter, J. Z., and Koltun, V. Trellis networks for sequence modeling. In *International Conference on Learning Representations*, 2018.
- Bai, S., Kolter, J. Z., and Koltun, V. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.
- Bai, S., Koltun, V., and Kolter, J. Z. Multiscale deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2020.
- Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 2019.
- Bojchevski, A., Klicpera, J., Perozzi, B., Kapoor, A., Blais, M., Różemberczki, B., Lukasik, M., and Günnemann, S. Scaling graph neural networks with approximate pagerank. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2464–2473, 2020.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.
- Broyden, C. G. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965.
- Bruna, J., Zaremba, W., Szlam, A., and Lecun, Y. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations*, 2014.
- Cai, T., Luo, S., Xu, K., He, D., Liu, T.-y., and Wang, L. Graphnorm: A principled approach to accelerating graph neural network training. *arXiv preprint arXiv:2009.03294*, 2020.
- Chen, J., Ma, T., and Xiao, C. Fastgcn: Fast learning with graph convolutional networks via importance sampling. In *International Conference on Learning Representations*, 2018a.
- Chen, J., Zhu, J., and Song, L. Stochastic training of graph convolutional networks with variance reduction. In *International Conference on Machine Learning*, pp. 941–949, 2018b.
- Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1725–1735, 2020.
- Chiang, W.-L., Liu, X., Si, S., Li, Y., Bengio, S., and Hsieh, C.-J. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 257–266, 2019.
- Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pp. 3844–3852, 2016.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4171–4186, 2019.
- Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., and Bresson, X. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982*, 2020.
- Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- Frasca, F., Rossi, E., Eynard, D., Chamberlain, B., Bronstein, M., and Monti, F. Sign: Scalable inception graph neural networks. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020.
- Gao, H., Wang, Z., and Ji, S. Large-scale learnable graph convolutional networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1416–1424, 2018.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- Gomez, A. N., Ren, M., Urtasun, R., and Grosse, R. B. The reversible residual network: Backpropagation without storing activations. In *Advances in Neural Information Processing Systems*, 2017.

- Gong, S., Bahri, M., Bronstein, M. M., and Zafeiriou, S. Geometrically principled connections in graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11415–11424, 2020.
- Gori, M., Monfardini, G., and Scarselli, F. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pp. 729–734. IEEE, 2005.
- Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.
- Hasanzadeh, A., Hajiramezanali, E., Boluki, S., Zhou, M., Duffield, N., Narayanan, K., and Qian, X. Bayesian graph neural networks with adaptive connection sampling. In *International Conference on Machine Learning*, pp. 4094–4104. PMLR, 2020.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Henaff, M., Bruna, J., and LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.
- Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.
- Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., and Leskovec, J. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Inan, H., Khosravi, K., and Socher, R. Tying word vectors and word classifiers: A loss framework for language modeling. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Khamsi, M. A. and Kirk, W. A. *An introduction to metric spaces and fixed point theory*, volume 53. John Wiley & Sons, 2001.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
- Kitaev, N., Kaiser, L., and Levskaya, A. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2019.
- Klicpera, J., Bojchevski, A., and Günnemann, S. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.
- Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.
- Levie, R., Monti, F., Bresson, X., and Bronstein, M. M. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, 2018.
- Li, G., Müller, M., Thabet, A., and Ghanem, B. Deepgcns: Can gcns go as deep as cnns? In *The IEEE International Conference on Computer Vision*, 2019.
- Li, G., Xiong, C., Thabet, A., and Ghanem, B. Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*, 2020.
- Li, G., Müller, M., Qian, G., Perez, I. C. D., Abualshour, A., Thabet, A. K., and Ghanem, B. Deepgcns: Making gcns go as deep as cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- Li, Q., Han, Z., and Wu, X. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Li, R., Wang, S., Zhu, F., and Huang, J. Adaptive graph convolutional neural networks. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. *Advances in Neural Information Processing Systems*, 2019.
- Liu, M., Gao, H., and Ji, S. Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 338–348, 2020.

- Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5115–5124, 2017.
- Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.
- Niepert, M., Ahmed, M., and Kutzkov, K. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pp. 2014–2023, 2016.
- Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2019.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.
- Press, O. and Wolf, L. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, 2017.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training.
- Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. Deep-speed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020.
- Rong, Y., Huang, W., Xu, T., and Huang, J. Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
- Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop, NeurIPS*, 2018.
- Shi, Y., Huang, Z., Wang, W., Zhong, H., Feng, S., and Sun, Y. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*, 2020.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- Wang, K., Shen, I., Huang, C., Wu, C.-H., Dong, Y., and Kanakia, A. Microsoft academic graph: when experts are not enough. *Quantitative Science Studies*, 1(1):396–413, February 2020.
- Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J., and Zhang, Z. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019a.
- Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019b.
- Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. PMLR, 2019.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.
- Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.
- Xu, K., Zhang, M., Jegelka, S., and Kawaguchi, K. Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. *arXiv preprint arXiv:2105.04550*, 2021.

Yang, Z., Cohen, W., and Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Yu, F. and Koltun, V. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., and Prasanna, V. GraphSAINT: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., and Ma, K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

Zhao, L. and Akoglu, L. Pairnorm: Tackling oversmoothing in gnns. In *International Conference on Learning Representations*, 2019.

Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., and Hu, X. Towards deeper graph neural networks with differentiable group normalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Zitnik, M. and Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.

A. Grouping of RevGNN

Grouped convolution is an effective way to reduce the parameter complexity in CNNs. We provide an ablation study to show how grouping reduces the number of parameters of RevGNNs. We conduct experiments on the *ogbn-proteins* dataset with different group sizes and report the results in Table 5. The number of hidden channels for all of these models is set to 224. We find that a larger group size reduces the number of parameters. As the group size increases from 2 to 4, the number of parameters reduces by more than 30%. The performance of models with 3 to 56 layers decreases slightly. The 112-layer networks achieve the same performance while the model with group size 4 uses only around 67% parameters compared to the model with group size 2. However, we observe that the GPU memory usage increases from 7.30 GB to 11.05 GB as the group size increases from 2 to 4 with our current implementation. We conjecture that this is due to our inefficient implementation. Optimizing our code for larger group sizes and conducting a more rigorous analysis is an interesting avenue for future investigation.

Table 5. Ablation of the group size of group reversible GNNs on the *ogbn-protein* dataset. L is the number of layers. The number of hidden channels is 224 for all the models.

| #L | Group=2 | | Group=4 | |
|-----|---------|--------------------|---------|--------------------|
| | Params | ROC-AUC \uparrow | Params | ROC-AUC \uparrow |
| 3 | 490k | 85.09 | 339k | 84.86 |
| 7 | 1.1M | 85.68 | 750k | 85.25 |
| 14 | 2.2M | 86.62 | 1.5M | 85.79 |
| 28 | 4.3M | 86.68 | 2.9M | 86.30 |
| 56 | 8.6M | 86.90 | 5.8M | 86.76 |
| 112 | 17.2M | 87.02 | 11.5M | 87.09 |

B. Experimental Details and More Ablations

B.1. Datasets and Frameworks

We conduct experiments on three OGB datasets (Hu et al., 2020) including *ogbn-proteins*, *ogbn-arxiv* and *ogbn-products*. We follow the standard data splits and evaluation protocol of OGB 1.2.4. Please refer to the OGB website¹ for more details. Our code implementation relies on the deep learning framework Pytorch 1.6.0. We use Pytorch Geometric 1.6.1 (Fey & Lenssen, 2019) for all experiments except for the experiments with GATs where we use DGL 0.5.3 (Wang et al., 2019a). The reversible module is implemented based on MemCNN (?). The deep equilibrium module is implemented based on DEQ (Bai et al., 2019).

¹<https://ogb.stanford.edu/>

Table 6. Results on the *ogbn-proteins* dataset for various 112-layer networks. Note that DEQ-GNN always has only a single layer that approximates an infinitely deep network. Each network is trained on one V100 GPU with 32GB of memory. The column *Mem* reports the GPU memory in GB, *Params* reports the number of model parameters, and *Time* reports the training time in days. *Baselines* are in italic.

| Model | #Ch | ROC-AUC \uparrow | Mem \downarrow | Params | Time \downarrow |
|------------------|-----|--------------------|------------------|--------|-------------------|
| <i>ResGNN</i> | 64 | 85.94 | 27.1 | 2.37M | 1.3 |
| <i>ResGNN</i> | 224 | - | OOM | 28.4M | - |
| <i>WT-ResGNN</i> | 64 | 83.30 | 27.4 | 51.2k | 1.2 |
| <i>WT-ResGNN</i> | 224 | - | OOM | 537k | - |
| DEQ-GNN | 64 | 83.17 | 2.22 | 51.3k | 1.3 |
| DEQ-GNN | 224 | 85.84 | 7.60 | 537k | 2.9 |
| RevGNN | 64 | 85.48 | 2.09 | 1.46M | 1.8 |
| RevGNN | 80 | 85.97 | 2.56 | 2.25M | 2.2 |
| RevGNN | 224 | 87.02 | 7.30 | 17.1M | 4.9 |
| WT-RevGNN | 64 | 82.89 | 1.60 | 35.0k | 1.7 |
| WT-RevGNN | 80 | 83.46 | 2.08 | 51.4k | 2.0 |
| WT-RevGNN | 224 | 85.28 | 5.55 | 337k | 4.8 |

B.2. Hyperparameters and Experimental Settings

We describe all important hyperparameters and training settings that were not mentioned in the main paper for reproducibility. The settings are slightly different for each dataset.

Ogbn-proteins. The node features are initialized through aggregating connected edge features by a *sum* aggregation at the first layer. We use random partitioning for mini-batch training. The number of partitions is set to 10 for training and 5 for validation for all the ablated models. One subgraph is sampled at each SGD step. One layer normalization is used in the GNN block. Dropout with a rate of 0.1 is used for each layer. We use *max* as the message aggregator. Each model is trained for 2000 epochs using the Adam optimizer with a learning rate of 0.001.

Ablations on Ogbn-proteins. A detailed comparison of ResGNN, Weight-tied ResGNN, DEQ-GNN, RevGNN and Weight-tied RevGNN is shown in Table 6. Except for the DEQ-GNN, all the other models have an explicit depth of 112 layers. The reversible connections reduce the memory consumption significantly and enable training of wider RevGNNs. A 112-layer RevGNN achieves the best performance (87.02 ROC-AUC) among the compared models. DEQ-GNN with 64 channels and WT-RevGNN with 80 channels have a similar number of parameters and memory consumption and also perform similarly. However, training DEQ-GNN is significantly faster than training WT-RevGNN (1.3 days vs. 2 days).

Table 7. Ablations for multi-view inference with RevGNN-Deep and RevGNN-Wide on the ogbn-proteins dataset. *L*, *Ch*, *Views* and *Parts* denote the numbers of layers, channels, views and parts respectively. Doing inference with more views and less parts is favorable.

| Model | #L | #Ch | #Views | #Parts | ROC-AUC ↑ |
|-------------|------|-----|--------|--------|------------------------------------|
| RevGNN-Deep | 1001 | 80 | 1 | 3 | 87.29 ± 0.16 |
| RevGNN-Deep | 1001 | 80 | 5 | 3 | 87.68 ± 0.13 |
| RevGNN-Deep | 1001 | 80 | 10 | 3 | 87.74 ± 0.13 |
| RevGNN-Wide | 448 | 224 | 1 | 3 | 87.84 ± 0.21 |
| RevGNN-Wide | 448 | 224 | 5 | 3 | 88.20 ± 0.16 |
| RevGNN-Wide | 448 | 224 | 10 | 3 | 88.24 ± 0.15 |
| RevGNN-Wide | 448 | 224 | 1 | 3 | 87.84 ± 0.21 |
| RevGNN-Wide | 448 | 224 | 1 | 5 | 87.62 ± 0.18 |
| RevGNN-Wide | 448 | 224 | 1 | 10 | 87.23 ± 0.22 |

Multi-view Inference on Ogbn-proteins. To further improve the evaluation results, we propose multi-view inference which reduces the negative effects of random partitioning and noisy neighbors. During different inference passes, each vertex will see a different set of neighbors. We refer to this as multi-view inference and implement it by partitioning the graphs into different subgraphs in each inference pass. In Table 7, we find that performing inference with more views yields better results. We observe a substantial improvement with increasing number of views for both the RevGNN-Deep and RevGNN-Wide models. The results increase by about 0.4% in terms of ROC-AUC going from 1 view to 10 views. We also observe that a smaller number of partitions is favorable for evaluation. To reduce memory cost, automatic mixed precision² by NVIDIA is used for inference.

Ogbn-arxiv. The directed graph is converted into an undirected graph and self-loops are added. We use the full-batch setting for both training and testing. For the GCN (Kipf & Welling, 2017), SAGE (Hamilton et al., 2017) and GEN (Li et al., 2020) models, batch normalization and dropout with a rate of 0.5 is applied to each layer and the Adam optimizer with a learning rate of 0.001 is used to train the models for 2000 epochs. The GAT-based (Veličković et al., 2018) models are implemented based on the OGB leaderboard submission *GAT + norm. adj. + label reuse*³. The RevGAT models with self-knowledge distillation are implemented based on the submission *GAT + label reuse + self KD*⁴. The teacher models and student models have the same

²<https://developer.nvidia.com/automatic-mixed-precision>

³<https://github.com/Espylapiza/dgl/tree/master/examples/pytorch/ogb/ogbn-arxiv>

⁴<https://github.com/ShunliRen/dgl/tree/master/examples/pytorch/ogb/ogbn-arxiv>

architecture. A knowledge distillation loss is added to the student model to minimize the Kullback–Leibler divergence between the teacher’s predictions and the student’s predictions during training. Please refer to the Github repositories for more details about the implementation.

Ogbn-products. Self-loops are added to the graph. We compare RevGNNs with full-batch training and mini-batch training. For mini-batch training, the graph is randomly partitioned into 10 subgraphs and one subgraph is sampled at each SGD step. We use full-batch testing in both scenarios. Batch normalization and dropout with a rate of 0.5 are used for each GNN block. The model is trained using the Adam optimizer with a learning rate of 0.001 for 1000 epochs.

B.3. GPU Memory Measurement

In all the experiments, the GPU memory usage is measured as the peak GPU memory during the first training epoch. Note that the measured GPU memory is larger than the GPU memory for storing node features due to the intermediate computation and network parameters. We consider the peak GPU memory usage as a practical metric since it is the bottleneck for training neural networks. As is common practice, we use `torch.cuda.max_memory_allocated()` for the memory measurement. However, note that the measured peak GPU memory obtained using `torch.cuda.max_memory_allocated()` is usually smaller than the actual peak GPU memory obtained with NVIDIA-SMI.

B.4. Correlation of Model Predictions

We perform a correlation analysis on model predictions of RevGNN, Weight-tied RevGNN and DEQ-GNN. The pearson correlations of RevGNN with 1000 layers, WT-RevGNN-224 with 7 layers and DEQ-GNN-224 with 56 iterations are: 0.8571 (RevGNN vs. WT-RevGNN), 0.8565 (RevGNN vs. DEQ-GNN) and 0.8948 (WT-RevGNN vs. DEQ-GNN).

C. More Discussion

C.1. Gradient Checkpointing and Model Parallelism

By saving checkpoints every \sqrt{L} steps, gradient checkpointing can achieve a memory complexity of $\mathcal{O}(\sqrt{L}ND)$, which is still higher than the memory complexity $\mathcal{O}(ND)$ of RevGNN. For 112-layer models with 64 hidden channels, ResGNN with gradient checkpointing consumes 2.5X the memory compared to RevGNN (5.22 G vs. 2.09 G) while reaching similar performance on *ogbn-proteins* with similar training time. Model parallelism is orthogonal to our approach. It would be interesting to investigate model parallelism to make RevGNN even wider with multiple GPUs.

C.2. Going Deeper and Datasets

In our experiments, we find that going deeper is very effective on *ogbn-proteins*. It would probably be beneficial to pre-train overparameterized GNNs on larger-scale protein datasets and then apply the pre-trained models to scientific applications such as drug discovery, protein structure prediction and gene-disease associations. For the other datasets such as *ogbn-products* and *ogbn-arxiv*, we observe less improvement when going very deep. It is still unclear what kind of datasets benefit more from depth and overparameterization. Investigating the relationship between overparameterization and factors such as dataset size, graph modality and graph learning task is an important direction of future work to better understand when overparameterized models are beneficial. We also anticipate that overparameterized GNNs will be a promising solution to even larger datasets such as OGB-LSC ([Hu et al., 2021](#)).