Supplementary Materials: L^2-GCN: Layer-Wise and Learned Efficient Training of Graph Convolutional Networks

Yuning You,* Tianlong Chen,* Zhangyang Wang, Yang Shen
Texas A&M University
{yuning.you,wiwjp619,atlaswang,yshen}@tamu.edu

Appendix

A. Proof of Theorem 5

Let GNN $\mathcal{A}_{Con} = \mathcal{R} \circ \mathcal{L}^{(L)}_{Con} \circ ... \circ \mathcal{L}^{(1)}_{Con}$ be conventionally trained by the optimization formulation (10) in the main text, with the conditions in Theorem 2 holding, i.e. $\mathcal{L}^{(l)}_{Con}, l \in L$ are injective, therefore $\mathcal{A}_{Con}$ is as powerful as WL test, we have:

$$Prob\{\mathcal{R} \circ \mathcal{L}^{(L)}_{Con} \circ ... \circ \mathcal{L}^{(1)}_{Con}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(L)}_{Con} \circ ... \circ \mathcal{L}^{(1)}_{Con}(G_2) | G_1 \ncong G_2\} = C_{WL}.$$

Now we prove that it can also be layer-wise trained by the optimization formulation (11) in the main text to achieve $Prob\{\mathcal{R} \circ \mathcal{L}^{(L)}_{Lay} \circ ... \circ \mathcal{L}^{(1)}_{Lay}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(L)}_{Lay} \circ ... \circ \mathcal{L}^{(1)}_{Lay}(G_2) | G_1 \ncong G_2\} = C_{WL}$.

(a) When training the 1st-layer mapping, we are going to solve the optimization problem as:

$$\mathcal{L}^{(1)}_{Lay} = max_{\mathcal{L}^{(1)}}\ Prob\{\mathcal{R} \circ \mathcal{L}^{(1)}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}(G_2) | G_1 \ncong G_2\}. \tag{1}$$

We can show that $\mathcal{L}^{(1)}_{Lay}$ is injective as follows. *Suppose that the optimal solution $\mathcal{L}^{(1)}_{Lay}$ is not injective.* Since we can conventionally train $\mathcal{A}_{Con}$, we have a feasible injective solution $\mathcal{L}^{(1)}_{Lay}$. For any non-isomorphic graph pairs G_1 and G_2, if layer-wise training has the correct mapping as $\mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_2)$, but conventional training maps wrongly as $\mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_1) = \mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_2)$, due to $\mathcal{R}$ is injective on multiset, for conventional training we have:

$$\begin{aligned}&\{\boldsymbol{x}^{(1)}_{i,G_1,Con} : \boldsymbol{x}^{(1)}_{i,G_1,Con} = \\ &\mathcal{L}^{(1)}_{Con}(\boldsymbol{x}^{(0)}_{i,G_1}, \{\boldsymbol{x}^{(0)}_{j,G_1} : j \in \mathcal{N}_{G_1}(i)\}), i \in N\} \\ = &\{\boldsymbol{x}^{(1)}_{i,G_2,Con} : \boldsymbol{x}^{(1)}_{i,G_2,Con} = \\ &\mathcal{L}^{(1)}_{Con}(\boldsymbol{x}^{(0)}_{i,G_2}, \{\boldsymbol{x}^{(0)}_{j,G_2} : j \in \mathcal{N}_{G_2}(i)\}), i \in N\}.\end{aligned}$$

Therefore there existing a bijective mapping $\phi : \{1, ..., N\} \rightarrow \{1, ..., N\}$ such that:

$$\begin{aligned}&\mathcal{L}^{(1)}_{Con}(\boldsymbol{x}^{(0)}_{i,G_1}, \{\boldsymbol{x}^{(0)}_{j,G_1} : j \in \mathcal{N}_{G_1}(i)\}) \\ &= \mathcal{L}^{(1)}_{Con}(\boldsymbol{x}^{(0)}_{k,G_2}, \{\boldsymbol{x}^{(0)}_{j,G_2} : j \in \mathcal{N}_{G_2}(k)\}), i = \phi(k).\end{aligned}$$

Since $\mathcal{L}^{(1)}_{Con}$ is injective, we always have:

$$\begin{aligned}&\boldsymbol{x}^{(0)}_{i,G_1} = \boldsymbol{x}^{(0)}_{k,G_2}, \\ &\{\boldsymbol{x}^{(0)}_{j,G_1} : j \in \mathcal{N}_{G_1}(i) = \{\boldsymbol{x}^{(0)}_{j,G_2} : j \in \mathcal{N}_{G_2}(k)\}, i = \phi(k).\end{aligned}$$

Thus for layer-wise training we have:

$$\begin{aligned}&\boldsymbol{x}^{(1)}_{i,G_1,Lay} = \boldsymbol{x}^{(1)}_{k,G_2,Lay}, \\ &\boldsymbol{x}^{(1)}_{i,G_1,Lay} = \mathcal{L}^{(1)}_{Lay}(\boldsymbol{x}^{(0)}_{i,G_1}, \{\boldsymbol{x}^{(0)}_{j,G_1} : j \in \mathcal{N}_{G_1}(i)\}), \\ &\boldsymbol{x}^{(1)}_{k,G_2,Lay} = \mathcal{L}^{(1)}_{Lay}(\boldsymbol{x}^{(0)}_{k,G_2}, \{\boldsymbol{x}^{(0)}_{j,G_2} : j \in \mathcal{N}_{G_2}(k)\}), i = \phi(k),\end{aligned}$$

which results in:

$$\begin{aligned}&\{\boldsymbol{x}^{(1)}_{i,G_1,Lay} : \boldsymbol{x}^{(1)}_{i,G_1,Lay} = \\ &\mathcal{L}^{(1)}_{Lay}(\boldsymbol{x}^{(0)}_{i,G_1}, \{\boldsymbol{x}^{(0)}_{j,G_1} : j \in \mathcal{N}_{G_1}(i)\}), i \in N\} \\ = &\{\boldsymbol{x}^{(1)}_{i,G_2,Lay} : \boldsymbol{x}^{(1)}_{i,G_2,Lay} = \\ &\mathcal{L}^{(1)}_{Lay}(\boldsymbol{x}^{(0)}_{i,G_2}, \{\boldsymbol{x}^{(0)}_{j,G_2} : j \in \mathcal{N}_{G_2}(i)\}), i \in N\}.\end{aligned}$$

We have $\mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_1) = \mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_2)$, which comes to a contradiction. Hence, we reach that if $\mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_2)$ correctly, then we have $\mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_2)$ correctly. However, not vice versa, it is easily to prove that if $\mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_2)$ correctly, we may have $\mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_1) = \mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_2)$ wrongly. Therefore we have:

$$\begin{aligned}&Prob\{\mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Lay}(G_2) | G_1 \ncong G_2\} \\ &< Prob\{\mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(1)}_{Con}(G_2) | G_1 \ncong G_2\},\end{aligned}$$

which is contradict to (1). Thus, $\mathcal{L}_{Lay}^{(1)}$ is injective.

(b) Assume we have finished training $l-1$ layer-wise mapping $\mathcal{L}_{Lay}^{(l-1)}, ..., \mathcal{L}_{Lay}^{(l)}$ which are injective. When training the lst-layer mapping, we are going to solve the optimization problem as:

$$\mathcal{L}_{Lay}^{(l)} = max_{\mathcal{L}^{(l)}}\ Prob\{\mathcal{R} \circ \mathcal{L}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \neq \mathcal{R} \circ \mathcal{L}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\}. \tag{2}$$

We can show that $\mathcal{L}_{Lay}^{(l)}$ is injective. *Suppose optimal the solution $\mathcal{L}_{Lay}^{(l)}$ is not injective.* Since we can conventionally train $\mathcal{A}_{Con}$, we have a feasible injective solution $\mathcal{L}_{Con}^{(l)}$. For any non-isomorphic graphs G_1, G_2, similar to the induction in (a), we have:

$$\begin{aligned} &Prob\{\mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \neq \mathcal{R} \circ \\ &\mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\} \\ &< Prob\{\mathcal{R} \circ \mathcal{L}_{Con}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \neq \mathcal{R} \\ &\circ \mathcal{L}_{Con}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\}, \end{aligned}$$

which contradicts (2). Thus, $\mathcal{L}_{Lay}^{(l)}$ is injective.

With (a) and (b), we have the result: through layer-wise trained by the optimization formulation (11) in the main text, we have injective layered mappings $\mathcal{L}_{Lay}^{(l)}, l \in L$. With Theorem 2, we come to the conclusion that $\mathcal{A}_{Lay} = \mathcal{R} \circ \mathcal{L}_{Lay}^{(L)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}$ is as powerful as WL test, i.e. $Prob\{\mathcal{R} \circ \mathcal{L}_{Lay}^{(L)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \neq \mathcal{R} \circ \mathcal{L}_{Lay}^{(L)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\} = C_{WL}$, which finishes the proof.

B. Proof of Theorem 6

Let's denote a layer-wise trained GNN as $\mathcal{A}_{Lay} = \mathcal{R} \circ \mathcal{L}_{Lay}^{(L)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}$, whose layer mapping $\mathcal{L}_{Lay}^{(l)}: \mathbb{R}^D \times \mathbb{M}^D \rightarrow \mathbb{R}^D$ can distinguish $\boldsymbol{x}_i^{(l-1)}$, i.e. $\mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_i^{(l-1)}, \{\boldsymbol{x}_j^{(l-1)} : j \in \mathcal{N}_i\}) \neq \mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_k^{(l-1)}, \{\boldsymbol{x}_j^{(l-1)} : j \in \mathcal{N}_k\})$ if $\boldsymbol{x}_i^{(l-1)} \neq \boldsymbol{x}_j^{(l-1)}$. We show that for any two non-isomorphic graphs G_1, G_2, if $l-1$ layer network $\mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}$ can successfully distinguishes them as:

$$\mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \neq \mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2). \tag{3}$$

Then l layer network $\mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}$ also can distinguish them as:

$$\begin{aligned} &\mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \\ &\neq \mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2). \end{aligned} \tag{4}$$

Suppose (4) *does not hold*, since $\mathcal{R}$ is injective on multiset, the same as the proof in Theorem 5, there exists a bijective mapping $\phi : \{1, ..., N\} \rightarrow \{1, ..., N\}$ such that:

$$\begin{aligned} &\mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_{i,G_1}^{(l-1)}, \{\boldsymbol{x}_{j,G_1}^{(l-1)} : j \in \mathcal{N}_{G_1}(i)\}) \\ &= \mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_{k,G_2}^{(l-1)}, \{\boldsymbol{x}_{j,G_2}^{(l-1)} : j \in \mathcal{N}_{G_2}(k)\}), i = \phi(k). \end{aligned}$$

Since the layer mapping can distinguish $\boldsymbol{x}_i^{(l-1)}$, resulting that if $\mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_i^{(l-1)}, \{\boldsymbol{x}_j^{(l-1)} : j \in \mathcal{N}(i)\}) = \mathcal{L}_{Lay}^{(l)}(\boldsymbol{x}_k^{(l-1)}, \{\boldsymbol{x}_j^{(l-1)} : j \in \mathcal{N}(k)\})$, we have $\boldsymbol{x}_i^{(l-1)} = \boldsymbol{x}_k^{(l-1)}$. Therefore we have:

$$\boldsymbol{x}_{i,G_1}^{(l-1)} = \boldsymbol{x}_{k,G_2}^{(l-1)}, i = \phi(k).$$

Due to the injectivity of $\mathcal{R}$, here comes the result:

$$\mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) = \mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2).$$

which is contradict to (3). Thus, (4) holds, and we have the conclusion: for any two non-isomorphic graphs G_1, G_2, if $l-1$ layer network can successfully distinguishes them, then l layer network also can distinguish them, which results in:

$$\begin{aligned} &\text{Prob}\{\mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \\ &\neq \mathcal{R} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\} \\ &\leq \text{Prob}\{\mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_1) \\ &\neq \mathcal{R} \circ \mathcal{L}_{Lay}^{(l)} \circ \mathcal{L}_{Lay}^{(l-1)} \circ ... \circ \mathcal{L}_{Lay}^{(1)}(G_2) | G_1 \ncong G_2\}. \end{aligned} \tag{5}$$

C. Dataset Statistic

Dataset statistic is shown in Table 1.

Table 1: Datasets Statistics.

Dataset	Nodes	Edges	Features	Classes
Cora	2780	13264	1433	7
PubMed	19717	108365	500	3
PPI	56944	818716	50	121
Reddit	232965	11606919	602	41
Amazon-670K	643474	1000746	100	32
Amazon-3M	2460406	48396681	100	38