FEATURE EXPANSION FOR GRAPH NEURAL NETWORKS

1. Tsinghua University 2. International Digital Economic Academy 3. Carnegie Mellon University 4. Mohamed bin Zayed University of Artificial Intelligence 5. Chinese University of Hong Kong

JIAQI SUN[1], LIN ZHANG[2], GUANGYI CHEN[3,4], PENG XU[5], KUN ZHANG[3,4], YUJIU YANG[1]

SPEAKER: JIAQI SUN

2023/07/27

CONTENT

- Background
- Motivation
- Analysis
- Method
- Experiments
- Conclusions & open questions

BACKGROUND

- Graph neural networks are widely adopted on a wide range of applications, e.g., bioinformatics and social study, especially popular for node classification task.
- Recently, many methods have studied the representations of GNNs from the perspective of optimization goals and spectral graph theory.
- However, there still no perspective could explain the existing problems integratively, e.g., over-smoothing problem and inferior performance for heterophilic graphs.
- GNNs is actually representation learning methods, while its representation space is not studied yet – **feature space.**

MOTIVATION

- We propose to analysis the feature space for existing GNNs with the help of a linear approximation.
  - Reasonableness will be indirectly verified in the post-experiments.

$$H = \overline{\texttt{GNN}}(X, \hat{A}) = \sum_{t=0}^{T-1} \Phi_t(X, \hat{A})\Theta_t$$

- where $\Phi_t(X, \hat{A}) \in \mathbb{R}^{n \times d_t}$ is the non-parametric feature space constructing function that inputs the graph data (e.g., node attributes and graph structure) and outputs a feature subspace, and
- $\Theta \in \mathbb{R}^{d_t \times c}$ is the parameter space to reweight the corresponding feature subspace for each class c, and T is a hyper-parameter of the number of the feature subspaces that the GNN contains.

ANALYSIS

- We can summarize existing spatial and spectral GNN approaches as follows:

Table 1. Feature space and parameters for GNN models (better viewed in color)

	Original formula*	Linear approximation formulations
GCN (Kipf & Welling, 2017)	$H^{(k+1)} = \sigma\left(\hat{A}H^{(k)}W^{(k)}\right)$	$H^{(K)} = \hat{A}^K X \prod_{i=0}^{K-1} W^{(i)}$
GIN (Xu et al., 2018)	$H^{(k+1)} = \sigma\left(\left(\epsilon^{(k)}I + \hat{A}\right)H^{(k)}W_0^{(k)}\right)W_1^{(k)}$	$H^{(K)} = \sum_{t=0}^{K} \hat{A}^k X \sum_{\{q_0,\cdots,q_{K-t-1}\}\subseteq\{\epsilon^{(0)},\cdots,\epsilon^{(K-1)}\}} \prod_i q_i \cdot \prod_{j=0}^{K-1} W_0^{(j)} W_1^{(j)}$
GCNII (Chen et al., 2020)	$H^{(l+1)} = \sigma\left(\left((1-\alpha^{(l)})\hat{A}H^{(l)} + \alpha^{(l)}H^{(0)}\right)\left((1-\beta^{(l)})I + \beta^{(l)}W^{(l)}\right)\right)$	$H^{(K)} = \sum_{l=0}^{K-1} \hat{A}^l X \prod_{i=L-l}^{L-1}(1-\alpha^{(i)})\alpha^{(L-l-1)} \prod_{j=L-l-1}^{L-1} W^{(j)}\} + \hat{A}^K \prod_{h=0}^{K-1}(1-\alpha^{(h)})W^{(h)}$
ARMA (Bianchi et al., 2021)	$H^{(K)} = \sigma(\tilde{L}H^{(K-1)}W_1 + XW_2)$	$H^{(K)} = \sum_{t=0}^{K} \tilde{L} X W_2^t W_1^{K-t}$
APPNP (Klicpera et al., 2019)	$H^{(k+1)} = (1-\alpha)\hat{A}H^{(l)} + \alpha H^{(0)}; H^{(0)} = \sigma(XW_1)W_2$	$H^{(K)} = \sum_{t=0}^{K} (1-\alpha)^t \hat{A}^l H^{(0)} + \sum_{i=0}^{t-1} \alpha(1-\alpha)^i \hat{A}^i H^{(0)} W_1 W_2$
ChebyNet** (Defferrard et al., 2016)	$H = \sum_{k=0}^{K} P_k(\hat{L}) X W^{(k)}$	$H^{(K)} = \sum_{t=0}^{K} P_t(\hat{L}) X W^{(t)}$
GPRGNN (Chien et al., 2021)	$H = \sum_{k=0}^{K} \gamma^{(k)} \hat{L}^k \sigma(XW_1)W_2$	$H^{(K)} = \sum_{t=0}^{K} \hat{L}^t X \gamma^{(t)} W_1 W_2$
BernNet (He et al., 2021)	$H = \sum_{k=0}^{K} \frac{1}{2^K}\binom{K}{k}\gamma^{(k)}(2I-\hat{L})^{K-k}\hat{L}^k\sigma(XW_1)W_2$	$H^{(K)} = \sum_{t=0}^{K} \sum_{j=0}^{t} \frac{1}{2^j}\binom{K}{j}\hat{L}^t \sum_{j=0}^{t} \gamma^{(j)} W_1 W_2$

* Without specification, $H^{(0)} = X$.

** $T_k(x)$ denotes Chebyshev polynomial $P_0(x) = 1, P_1(x) = x, P_k(x) = 2xP_{k-1} - P_{k-2}$.

ANALYSIS

Proposition 3.1. *Suppose the feature subspaces are constructed sequentially by* $\{\Phi_t = \hat{A}^t X\}_t$. *As* $i \in \mathbb{Z}$ *increases, the subspace* Φ_{t+i} *gradually tends to be linearly correlated with* Φ_t.

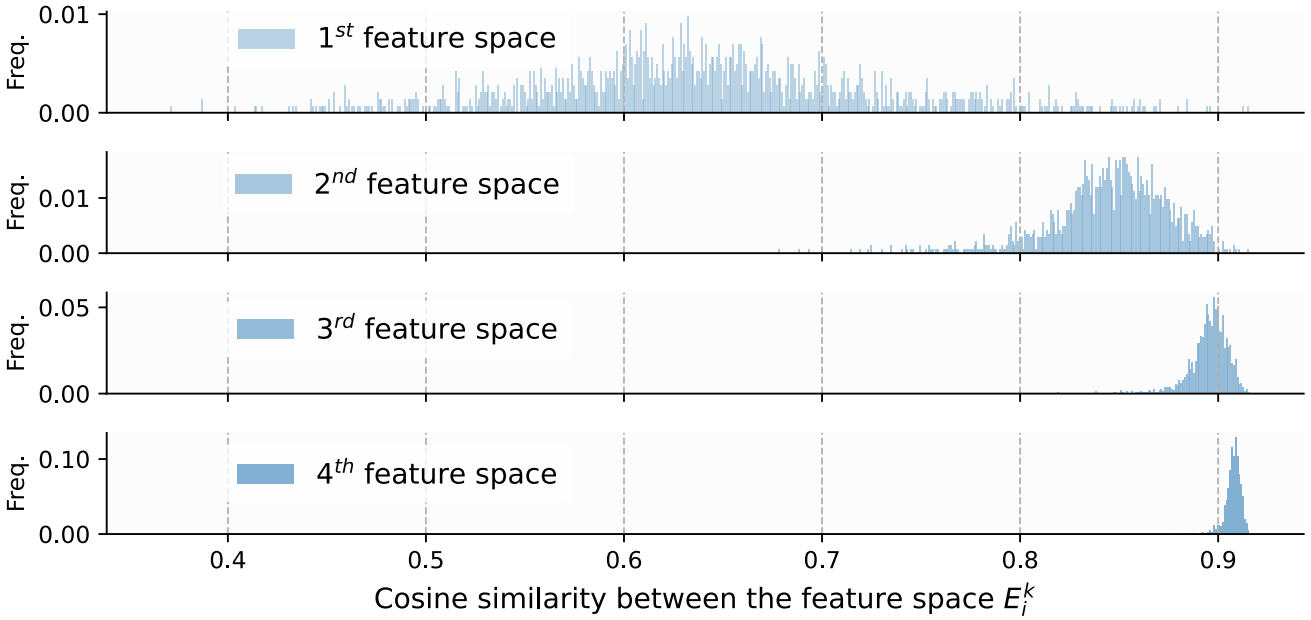

Figure 1. Distribution of the mutual correlation values between the later feature (sub)spaces to the previous total ones.

E.g., GPRGNN,

$$E_i^k = \max_{j=0,\cdots,k-1} \mu(\hat{L}^j X, \hat{L}^k X_{\cdot i}),$$

ANALYSIS

- **Issue 1: Constraint from the weight sharing mechanism.**

  From Table 1. we see that existing GNNs usually share parameter weights between different subspaces. Under linear correlation condition, using the weight-sharing method limits the expressiveness of the feature space.

➢ **Modification 1: Feature Subspaces Flattening**

Theorem 3.2. *(its proof can be found in Appendix A.4) Suppose* $\Phi_a, \Phi_b \in \mathbb{R}^{n\times d}$ *are two linearly correlated feature subspaces, i.e. there exists* $W_a \in \mathbb{R}^{d\times d}$ *such that* $\Phi_a W_a = \Phi_b$*, and suppose a matrix* $B \in \mathbb{R}^{n\times c}$*,* $c << d$*. If* B *can be represented by using both subspaces with a common weight* W_B*, i.e.,* $\gamma_a\Phi_a W_B + \gamma_b\Phi_b W_B = B$ *and* $\gamma_a, \gamma_b \in \mathbb{R}$*, then* B *can always be represented by only one subspace* Φ_a*, i.e.,* $\Phi_a W'_B = B$ *and* $W'_B \in \mathbb{R}^{d\times c}$*.*

Theorem 4.1. *(its proof can be found in Appendix A.5) Suppose* $\Phi_a, \Phi_b \in \mathbb{R}^{n\times d}$ *are two linearly correlated feature subspaces, i.e., there exits* $W_a \in \mathbb{R}^{d\times d}$ *such that* $\Phi_a W_a = \Phi_b$*, and a matrix* $B \in \mathbb{R}^{d\times c}$*,* $c << d$*. If* B *can be expressed by* Φ_a*, i.e.,* $\Phi_a W_B = B$*, then using both subspaces* Φ_a *and* Φ_b *independently, i.e.,* $\Phi_a W_a + \Phi_b W_b = B$*, the optimum is more easily achieved than a weight-sharing style.*

ANALYSIS

- **Issue 2: Constraint from limited dimensionality of node attributes.**
  When the node attributes X have a thin shape, is far from a sufficient feature space, making the regression system strictly over-determined.
  In addition, without any assumption about the feature space construction, there is hardly an exact solution, e.g. heterophilic condition.

➢ **Modification 2: Structural Principal Components**

Theorem 4.2. *(its proof can be found in Appendix A.6) Suppose the dimensionality of the node attributes is much smaller than the number of nodes, i.e., $d << n, X \in \mathbb{R}^{n\times d}$, and a z-truncated SVD of $\hat{L}$, which satisfies $||U_z S_z - \hat{L}||_2 < \epsilon$, where ϵ is a sufficiently small constant. Then the linear system $(\Phi_k, U_z S_z)W'_B = B$ can achieve a miner error than the linear system $\Phi_k W_B = B$.*

$$S = \tilde{Q}\tilde{V}; \hat{A} = QVR^T$$

$$H = \sum_{k=0}^{K} P_k(\hat{L})XW^{(k)} + SW_s.$$

METHOD

$$H = \hat{A}^3 X \prod_{k=0}^{2} W^{(k)} \qquad H = \sum_{k=0}^{2} \gamma_k P_k(\hat{L}) X W^{(0)} \qquad H = \sum_{k=0}^{2} P_k(\hat{L}) X W^{(k)} + S W_s$$

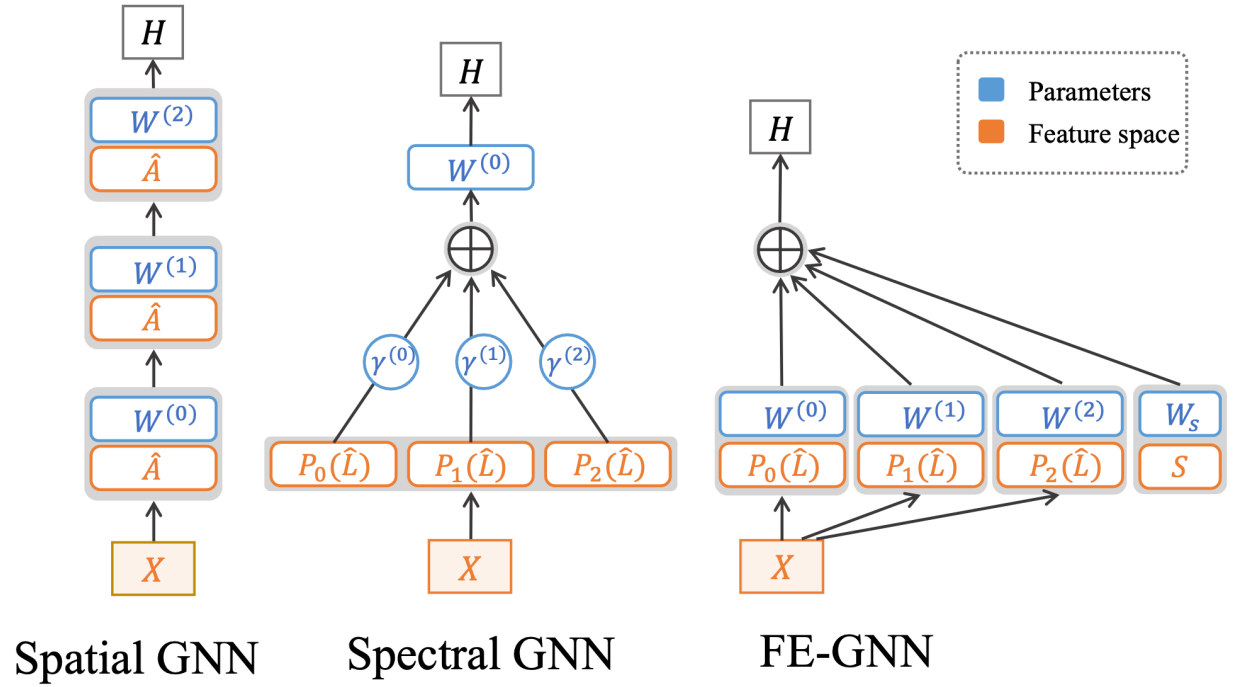

Figure 2. Architecture of our proposal

EXPERIMENTS

- Node classification

Table 2. Overall performance of FE-GNN in node classification

Type	Baseline	Time (ms)	Homophilic graphs					Heterophilic graphs		
			Cora	CiteSeer	PubMed	Computers	Photo	Squirrel	Chameleon	Actor
Spatial	MLP	-	$76.70_{\pm 0.15}$	$76.67_{\pm 0.26}$	$85.11_{\pm 0.26}$	$82.62_{\pm 0.21}$	$84.16_{\pm 0.13}$	$37.86_{\pm 0.39}$	$57.83_{\pm 0.31}$	$38.99_{\pm 0.17}$
	GCN	$17.42_{\pm 1.64}$	$87.69_{\pm 0.40}$	$79.31_{\pm 0.46}$	$86.71_{\pm 0.18}$	$83.24_{\pm 0.11}$	$88.61_{\pm 0.36}$	$47.21_{\pm 0.59}$	$61.85_{\pm 0.38}$	$28.61_{\pm 0.39}$
	GAT	$18.06_{\pm 1.18}$	$88.07_{\pm 0.41}$	$80.80_{\pm 0.26}$	$86.69_{\pm 0.14}$	$82.86_{\pm 0.35}$	$90.84_{\pm 0.32}$	$33.40_{\pm 0.16}$	$51.82_{\pm 1.33}$	$33.48_{\pm 0.35}$
	GraphSAGE	$10.72_{\pm 0.25}$	$87.74_{\pm 0.41}$	$79.20_{\pm 0.42}$	$87.65_{\pm 0.14}$	$87.38_{\pm 0.15}$	$93.59_{\pm 0.13}$	$48.15_{\pm 0.45}$	$62.45_{\pm 0.48}$	$36.39_{\pm 0.35}$
	GCNII	$8.48_{\pm 0.24}$	$87.46_{\pm 0.31}$	$80.76_{\pm 0.30}$	$88.82_{\pm 0.08}$	$84.75_{\pm 0.22}$	$93.21_{\pm 0.25}$	$43.28_{\pm 0.35}$	$61.80_{\pm 0.44}$	$38.61_{\pm 0.26}$
	APPNP	$23.74_{\pm 2.08}$	$87.92_{\pm 0.20}$	$81.42_{\pm 0.26}$	$88.16_{\pm 0.14}$	$85.88_{\pm 0.13}$	$90.40_{\pm 0.34}$	$39.63_{\pm 0.49}$	$59.01_{\pm 0.48}$	$39.90_{\pm 0.25}$
Spectral	ChebyNet	$20.26_{\pm 1.03}$	$87.17_{\pm 0.19}$	$77.97_{\pm 0.36}$	$89.04_{\pm 0.08}$	$87.92_{\pm 0.13}$	$94.58_{\pm 0.11}$	$44.55_{\pm 0.28}$	$64.06_{\pm 0.47}$	$25.55_{\pm 1.67}$
	GPRGNN	$23.55_{\pm 1.26}$	$87.97_{\pm 0.24}$	$78.57_{\pm 0.31}$	$89.11_{\pm 0.08}$	$86.07_{\pm 0.14}$	$93.99_{\pm 0.11}$	$43.66_{\pm 0.22}$	$63.67_{\pm 0.34}$	$36.93_{\pm 0.26}$
	BernNet	$36.88_{\pm 0.84}$	$87.66_{\pm 0.26}$	$79.34_{\pm 0.32}$	$89.33_{\pm 0.07}$	$88.66_{\pm 0.08}$	$94.03_{\pm 0.08}$	$44.57_{\pm 0.33}$	$63.07_{\pm 0.43}$	$36.89_{\pm 0.30}$
Unified	GNN-LF	$52.77_{\pm 4.50}$	$88.12_{\pm 0.06}$	**$83.66_{\pm 0.06}$**	$87.79_{\pm 0.05}$	$87.63_{\pm 0.05}$	$93.79_{\pm 0.06}$	$39.03_{\pm 0.08}$	$59.84_{\pm 0.09}$	$41.97_{\pm 0.06}$
	GNN-HF	$53.28_{\pm 4.51}$	$88.47_{\pm 0.09}$	$83.56_{\pm 0.10}$	$87.83_{\pm 0.10}$	$86.94_{\pm 0.06}$	$93.89_{\pm 0.10}$	$39.01_{\pm 0.51}$	$63.90_{\pm 0.11}$	**$42.47_{\pm 0.07}$**
	ADA-UGNN	$14.36_{\pm 0.21}$	$88.92_{\pm 0.11}$	$79.34_{\pm 0.09}$	$90.08_{\pm 0.05}$	$89.56_{\pm 0.09}$	$94.66_{\pm 0.07}$	$44.58_{\pm 0.16}$	$59.25_{\pm 0.16}$	$41.38_{\pm 0.12}$
	FE-GNN (C)	$15.8_{\pm 0.11}$	**$89.45_{\pm 0.22}$**	$81.96_{\pm 0.23}$	**$90.27_{\pm 0.49}$**	**$90.79_{\pm 0.08}$**	$95.36_{\pm 0.14}$	$67.82_{\pm 0.26}$	**$73.33_{\pm 0.35}$**	$40.54_{\pm 0.15}$
	FE-GNN (M)	$14.6_{\pm 0.32}$	$89.09_{\pm 0.22}$	$81.76_{\pm 0.23}$	$89.93_{\pm 0.23}$	$90.60_{\pm 0.11}$	**$95.45_{\pm 0.15}$**	**$67.90_{\pm 0.23}$**	$73.26_{\pm 0.38}$	$40.91_{\pm 0.22}$

EXPERIMENTS

- Ablation study

Table 3. Ablation study of 1) flattening feature subspaces

	Cora	CiteSeer	Chameleon	Squirrel
k=2	$89.15\pm_{0.86}$	$81.97\pm_{1.10}$	$73.41\pm_{1.40}$	$67.37\pm_{0.83}$
k=2 (WS)	$87.21\pm_{0.83}$	$78.39\pm_{0.72}$	$72.94\pm_{1.22}$	$66.01\pm_{1.31}$
k=4	$88.56\pm_{2.01}$	$80.19\pm_{0.84}$	$73.27\pm_{1.56}$	$67.40\pm_{0.90}$
k=4 (WS)	$87.28\pm_{1.38}$	$77.72\pm_{0.86}$	$73.23\pm_{1.72}$	$66.43\pm_{1.72}$
k=8	$88.92\pm_{0.88}$	$81.11\pm_{0.89}$	$73.85\pm_{1.52}$	$67.93\pm_{2.04}$
k=8 (WS)	$86.92\pm_{1.66}$	$77.32\pm_{0.33}$	$73.15\pm_{1.83}$	$66.63\pm_{2.38}$
k=16	$88.26\pm_{0.14}$	$80.54\pm_{1.03}$	$73.88\pm_{1.53}$	$67.82\pm_{1.54}$
k=16 (WS)	$87.34\pm_{1.98}$	$78.60\pm_{0.70}$	$72.94\pm_{1.79}$	$66.65\pm_{2.15}$

Table 5. Ablation study of 2) structural principal components

	Cora	CiteSeer	PubMed	Squirrel	Chameleon
FE-GNN(C)	$\mathbf{89.45}_{\pm\mathbf{0.22}}$	$\mathbf{81.96}_{\pm\mathbf{0.23}}$	$89.87_{\pm0.49}$	$67.82_{\pm0.26}$	$\mathbf{73.33}_{\pm\mathbf{0.35}}$
FE-GNN(M)	$89.09_{\pm0.22}$	$81.76_{\pm0.23}$	$89.93_{\pm0.23}$	$\mathbf{67.90}_{\pm\mathbf{0.23}}$	$73.26_{\pm0.38}$
w/o norm	$86.23_{\pm1.43}$	$79.32_{\pm0.59}$	$\mathbf{90.27}_{\pm\mathbf{0.49}}$	$64.70_{\pm1.10}$	$68.25_{\pm1.64}$
w/o S	$89.20_{\pm0.93}$	$81.95_{\pm0.87}$	$89.76_{\pm0.46}$	$43.21_{\pm0.99}$	$61.54_{\pm1.52}$
w/o $P_k(\hat{L})X_{k>0}$	$71.10_{\pm1.72}$	$74.38_{\pm1.01}$	$86.61_{\pm0.54}$	$67.90_{\pm0.96}$	$73.35_{\pm1.21}$
w/o $P_k(\hat{L})X_{k=0}$	$84.70_{\pm1.05}$	$58.60_{\pm2.19}$	$85.84_{\pm0.45}$	$65.75_{\pm0.63}$	$72.61_{\pm1.60}$

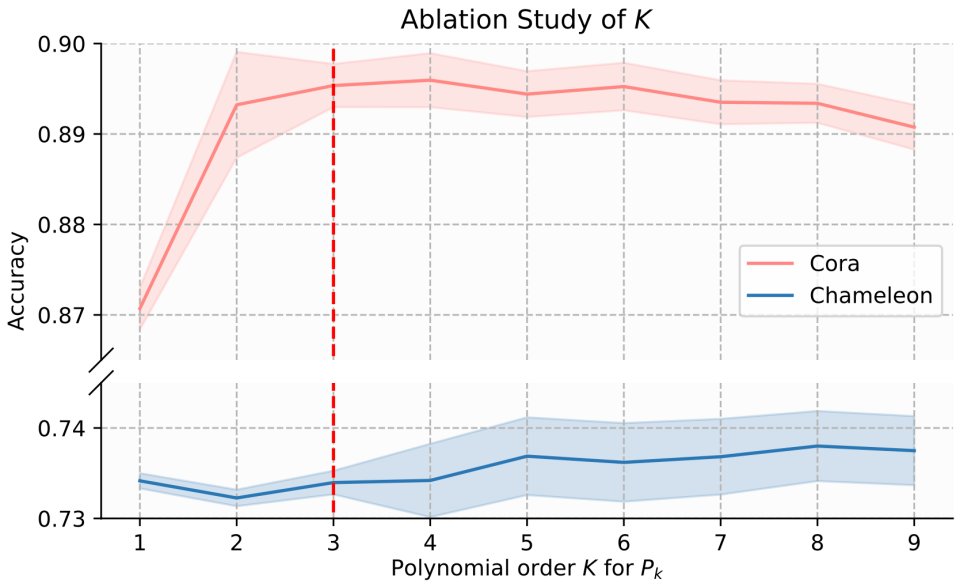

Figure 4. Ablation study of K

EXPERIMENTS

- SVD analysis

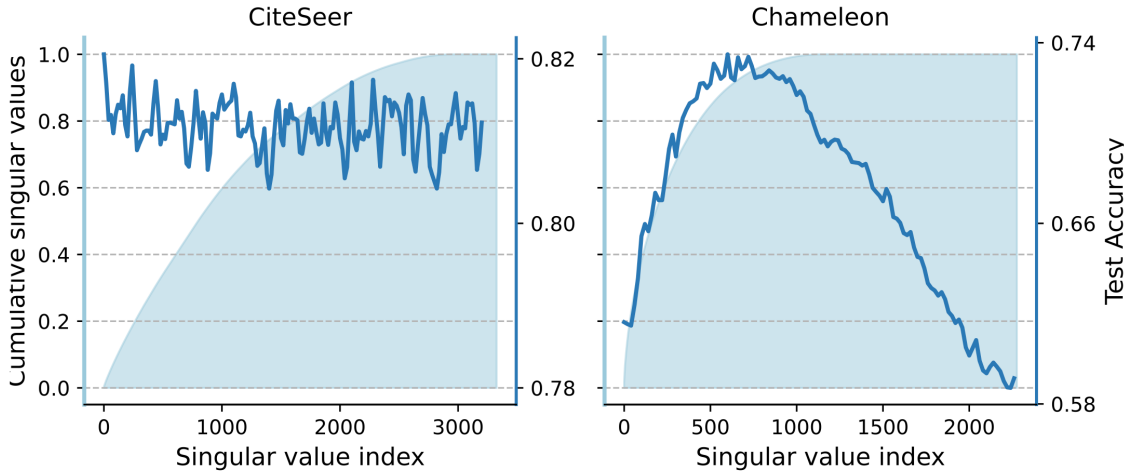

Figure 6. Sensitivity study of truncated SVD

Table 4. Time consumption of SVD

	Cora	CiteSeer	Chameleon	Squirrel	Actor
Training time (ms)	4000.01 ± 52.23	4103.46 ± 133.77	2818.21 ± 81.87	6096.43 ± 403.95	6074.39 ± 547.34
SVD time (ms)	3.88 ± 0.08	8.76 ± 0.05	61.80 ± 0.24	432.43 ± 0.70	3.99 ± 0.02
# of epochs	252	252	252	271	252
SVD time / Training time (%)	0.097	0.21	2.2	7.1	0.066

EXPERIMENTS

- Efficiency check

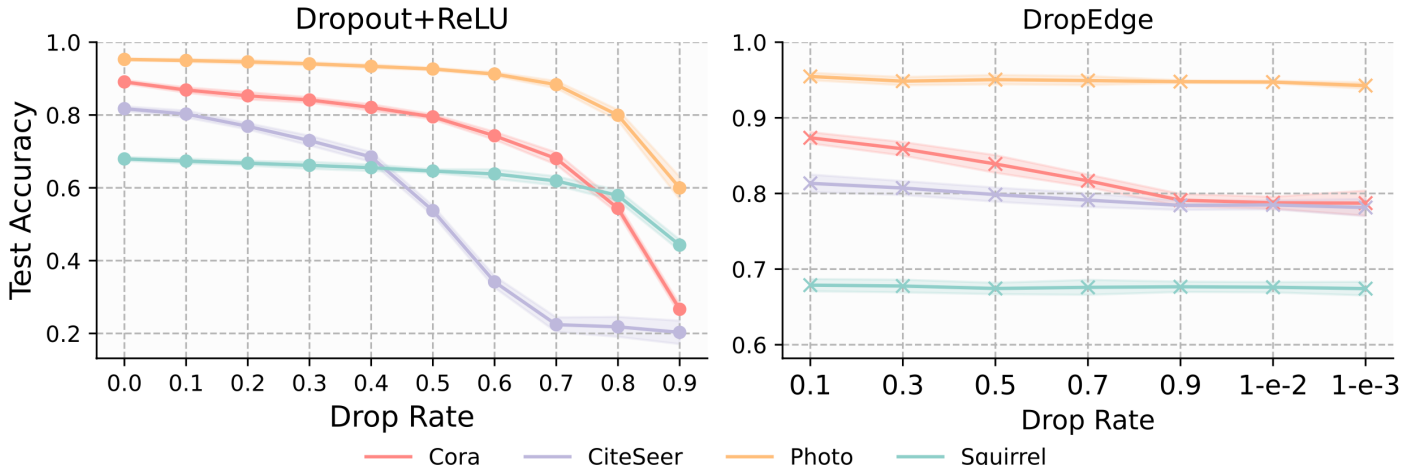

Figure 3. Analysis of the deep artifices on FE-GNN.

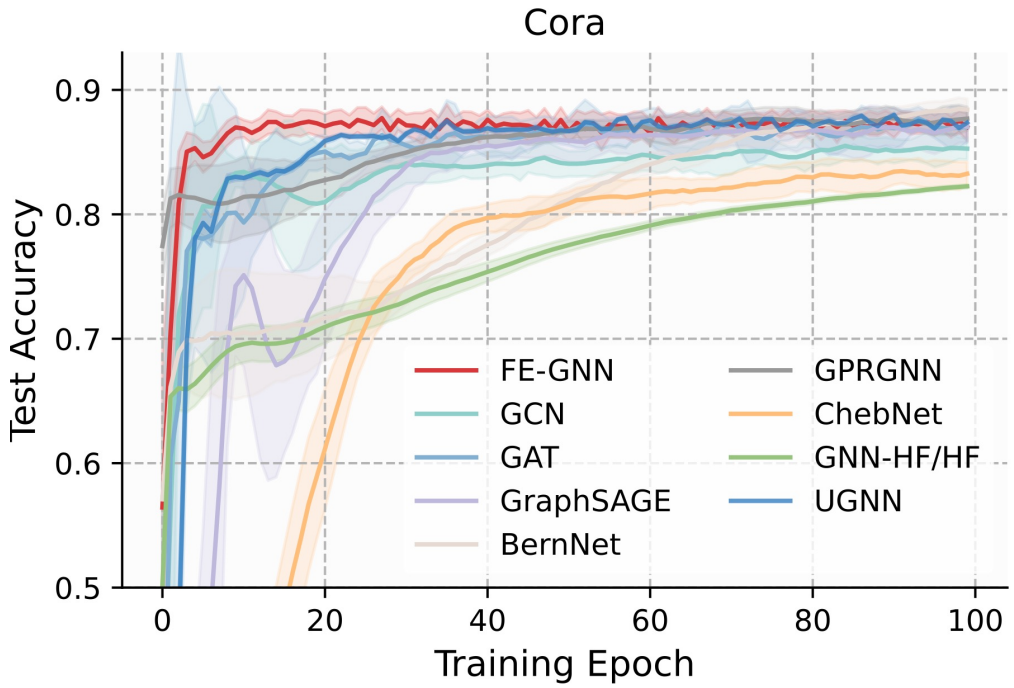

Figure 5. Convergence curve

CONCLUSIONS & OPEN QUESTIONS

- In this paper, we provide the feature space view to analyze GNNs, which separates the feature space and the parameters. Together, we provide a theoretical analysis of the existing feature space of GNNs and summarize two issues. We propose 1) feature subspace flattening and 2) structural principal components for these issues. Extensive experimental results verify their superiority.
- Limitations:
  - Nonlinear cases are not included in our work and will be considered in future work.
  - Also, the correlation between the subspaces should be studied more carefully beyond the linear correlation property; in a sense, the parameters can be further reduced by introducing reasonable constraints.
  - Finally, more feature space construction methods should be discovered for future work.

清華大學
Tsinghua University

THANK YOU!

ANY FEEDBACK OR DISCUSSIONS ARE WELCOMED!!!

CONTACT: SUNJQ20@MAILS.TSINGHUA.EDU.CN