DGI

Deep Graph Infomax

이수연

2022.02.03

Contents

1. Backgrounds
2. DGI Methodology
3. Classification Performance
4. Conclusion

Backgrounds

the dominant algorithms for unsupervised representation learning with graph-structured data rely on **random walk**-based objectives

nodes that are "close" in the input graph are also "close" in the representation space

Limitation

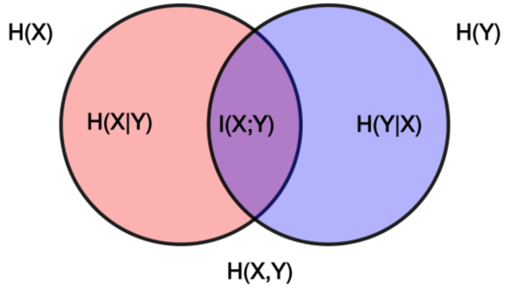
1. over-emphasizing **proximity** information at the expense of **structural** information
2. highly dependent on hyperparameter
3. with stronger encoder, unclear whether it actually provide any useful signal

→ Alternative objective based on **mutual information**

Backgrounds

Mutual Information(MI) measures how much one random variables tells us about another (dependence)

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y}$$



distribution을 알기 어려워서 MI 직접 계산 힘듦

→ lower-bound를 maximize, $I(X,Z) \geq I_\Theta(X,Z)$

DV (Donsker-Varadhan) $\mathcal{I}(X;Y) := \mathcal{D}_{KL}(\mathbb{J}||\mathbb{M}) \geq \widehat{\mathcal{I}}_\omega^{(DV)}(X;Y) := \mathbb{E}_{\mathbb{J}}[T_\omega(x,y)] - \log \mathbb{E}_{\mathbb{M}}[e^{T_\omega(x,y)}]$ → **MINE** (Mutual Information Neural Estimation)

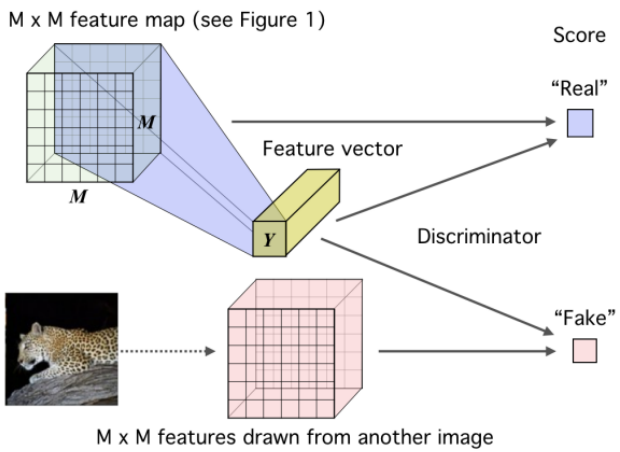
JSD (Jensen-Shannon) $\widehat{\mathcal{I}}_{\omega,\psi}^{(\text{JSD})}(X;E_\psi(X)) := \mathbb{E}_{\mathbb{P}}[-\text{sp}(-T_{\psi,\omega}(x,E_\psi(x)))] - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[\text{sp}(T_{\psi,\omega}(x',E_\psi(x)))]$

infoNCE (Noise-Contrastive Estimation + softmax) $\widehat{\mathcal{I}}_{\omega,\psi}^{(\text{infoNCE})}(X;E_\psi(X)) := \mathbb{E}_{\mathbb{P}}\left[T_{\psi,\omega}(x,E_\psi(x)) - \mathbb{E}_{\tilde{\mathbb{P}}}\left[\log \sum_{x'} e^{T_{\psi,\omega}(x',E_\psi(x))}\right]\right]$

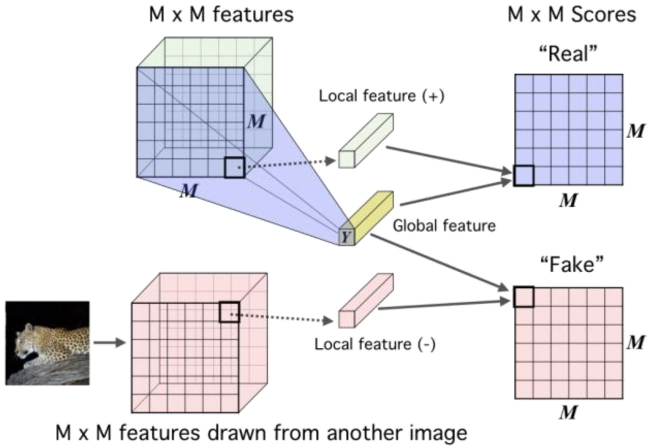
Backgrounds

Deep InfoMax (DIM) unsupervised learning of representations by maximizing mutual information between an input and the output of a deep neural network encoder

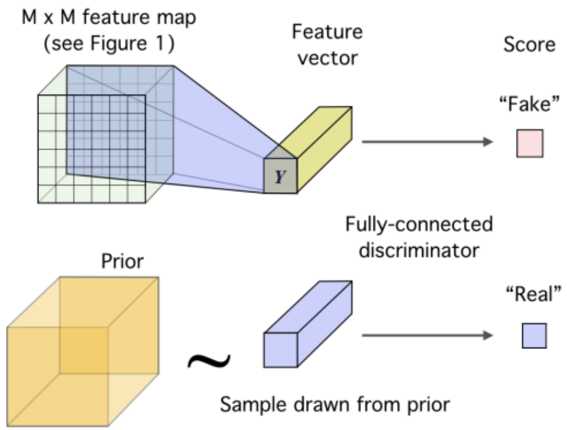
Global DIM



Local DIM



matching to **prior**

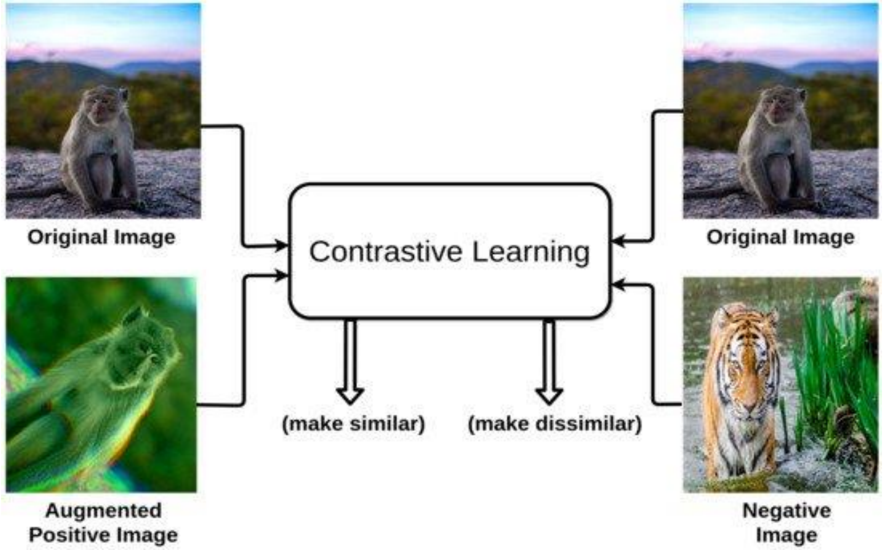


objective

$$\arg\max_{\omega_1,\omega_2,\psi} \left(\alpha\widehat{\mathcal{I}}_{\omega_1,\psi}(X;E_\psi(X)) + \frac{\beta}{M^2}\sum_{i=1}^{M^2}\widehat{\mathcal{I}}_{\omega_2,\psi}(X^{(i)};E_\psi(X))\right) + \arg\min_{\psi}\arg\max_{\phi}\gamma\widehat{\mathcal{D}}_\phi(\mathbb{V}||\mathbb{U}_{\psi,\mathbb{P}})$$

Backgrounds

Contrastive Learning an important approach for **self-supervised** learning of representations



Basic intuition behind contrastive learning paradigm
from: A Survey on Contrastive Self-Supervised Learning

"push original and augmented images closer and push original and negative images away"

real(i.e. positive)과 fake(i.e. negative) example을 **대조(contrast)**하여 representation 학습

Backgrounds

DIM을 그래프에 적용!

- local features to capture information shared across the entire graph
- MI maximization
- Contrastive learning

DGI Methodology

Negative Sampling

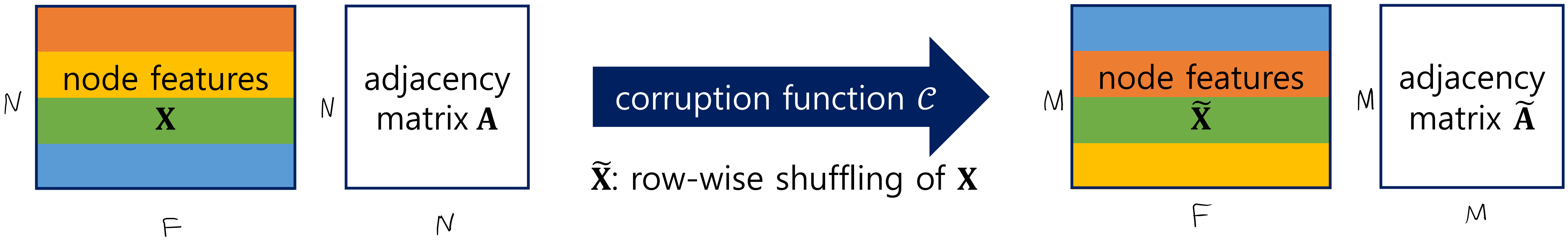
multi-graph setting

corruption function simply samples a different graph from the training set

single-graph setting

original graph $(\mathbf{X}, \mathbf{A})$

alternative graph $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}})$



DGI Methodology

Local & Global features

Local features

Global feature

input graph $(\mathbf{X}, \mathbf{A})$

$\mathbf{H}$

readout function $\mathcal{R}$

$\vec{s}$

summary vector $\vec{s}$

encoder $\mathcal{E}$

$$\mathcal{R}(\mathbf{H}) = \sigma\left(\frac{1}{N}\sum_{i=1}^{N}\vec{h}_i\right)$$

negative example $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}})$

$\widetilde{\mathbf{H}}$

$\vec{h}_i$ summarize a patch of the graph centered around node i
→ **patch representation**

DGI Methodology

Theoretical Motivation

Theorem 1. minimizing the classification error in the discriminator ≡ maximizing the mutual information between the input and output

$$\vec{s}^{\star} = \mathrm{argmin}_{\vec{s}}\mathrm{Err}^{*} \qquad \vec{s}^{\star} = \mathrm{argmax}_{\vec{s}}\mathrm{MI}(\mathbf{X}; \vec{s})$$

Theorem 2. the $\vec{h}_i$ that minimizes the classification error between $p(\vec{h}_i, \vec{s})$ and $p(\vec{h}_i)p(\vec{s})$ also maximizes $\mathrm{MI}(\mathbf{X}_i^{(k)}; \vec{h}_i)$

→ a classifier between samples from the joint(positive examples) and the product of marginals (negative examples), and using the binary cross-entropy (BCE) loss to optimize this classifier

DGI Methodology

Mutual Information Maximization

discriminator $\mathcal{D}(\vec{h}_i, \vec{s}) = \sigma(\vec{h}_i^T \mathbf{W} \vec{s})$

positive example과 negative example을 구별하도록 training

Glorot initialization
Adam SGD optimizer

$$\mathcal{L} = \frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}_{(\mathbf{X},\mathbf{A})}\left[\log \mathcal{D}\left(\vec{h}_i, \vec{s}\right)\right] + \sum_{j=1}^{M}\mathbb{E}_{(\widetilde{\mathbf{X}},\widetilde{\mathbf{A}})}\left[\log\left(1-\mathcal{D}\left(\vec{\widetilde{h}}_j, \vec{s}\right)\right)\right]\right)$$

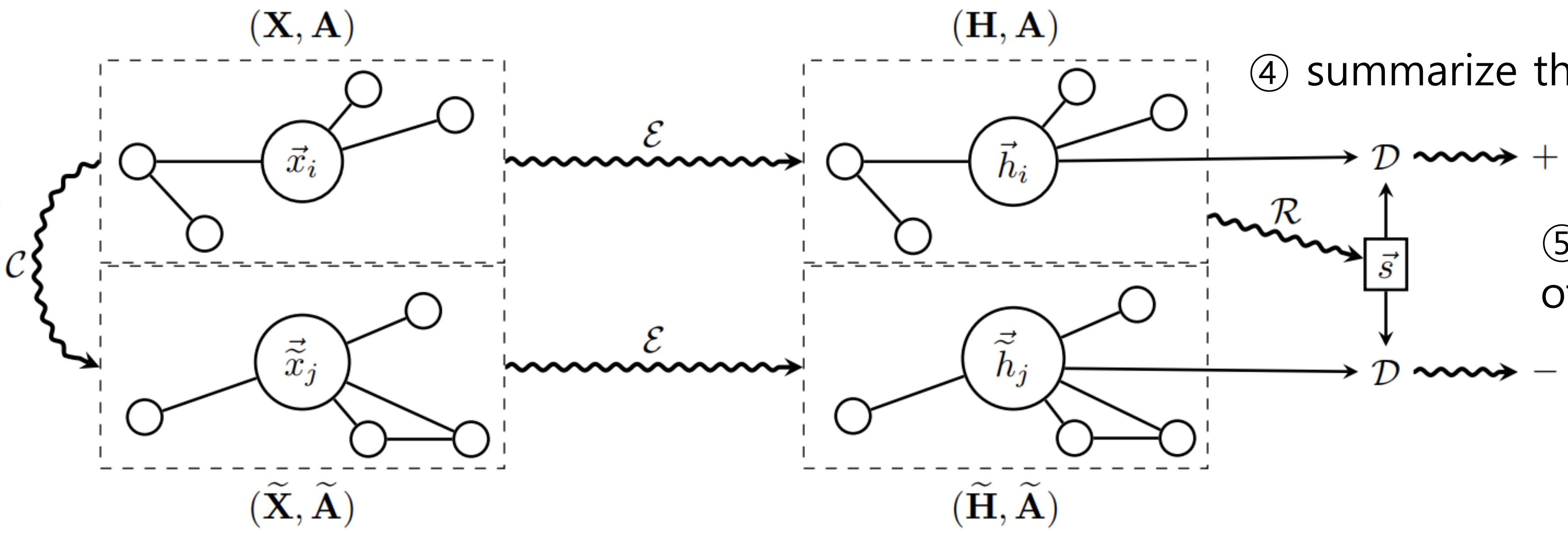
noise-contrastive type objective with a standard binary cross-entropy (BCE) loss

→ mutual information with the global graph summary, patch-level의 similarity

DGI Methodology

Overview of DGI

① negative sampling



④ summarize the input graph

⑤ update parameters of $\mathcal{E}, \mathcal{R}$ and $\mathcal{D}$

②, ③ patch representations for input graph/negative example

Classification Performance

patch representations learned in a fully unsupervised manner
→ evaluating the node-level classification utility

Datasets

(1) Cora, Citeseer, Pubmed citation networks: classifying research papers into topics (transductive)

(2) Reddit social network: predicting the community structure (inductive)

(3) PPI network: classifying protein roles (inductive, multiple graphs)

Classification Performance

Encoder

Transductive learning

one-layer GCN

$$\mathcal{E}(\mathbf{X},\mathbf{A}) = \sigma\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}\right)$$

the learned filters rely on a fixed and known adjacency matrix: not appropriate to Inductive learning

Inductive learning on large graphs

three-layer mean-pooling model with skip connections

$$\mathrm{MP}(\mathbf{X},\mathbf{A}) = \hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}\mathbf{X}\mathbf{\Theta}$$

$$\mathcal{E}(\mathbf{X},\mathbf{A}) = \widetilde{\mathrm{MP}}_3(\widetilde{\mathrm{MP}}_2(\widetilde{\mathrm{MP}}_1(\mathbf{X},\mathbf{A}),\mathbf{A}),\mathbf{A})$$

X, A are not needed

sampling node neighborhoods with replacement

Inductive learning on multiple graphs

three-layer mean-pooling model with dense skip connections

$$\begin{aligned}
\mathbf{H}_1 &= \sigma\left(\mathrm{MP}_1(\mathbf{X},\mathbf{A})\right)\\
\mathbf{H}_2 &= \sigma\left(\mathrm{MP}_2(\mathbf{H}_1 + \mathbf{X}\mathbf{W}_{\mathrm{skip}},\mathbf{A})\right)\\
\mathcal{E}(\mathbf{X},\mathbf{A}) &= \sigma\left(\mathrm{MP}_3(\mathbf{H}_2 + \mathbf{H}_1 + \mathbf{X}\mathbf{W}_{\mathrm{skip}},\mathbf{A})\right)
\end{aligned}$$

Classification Performance

Result the DGI approach is competitive with the results reported for the GCN model with the supervised loss

Transductive mean classification accuracy

Available data	Method	Cora	Citeseer	Pubmed
X	Raw features	47.9 ± 0.4%	49.3 ± 0.2%	69.1 ± 0.3%
A, Y	LP (Zhu et al., 2003)	68.0%	45.3%	63.0%
A	DeepWalk (Perozzi et al., 2014)	67.2%	43.2%	65.3%
X, A	DeepWalk + features	70.7 ± 0.6%	51.4 ± 0.5%	74.3 ± 0.9%
X, A	Random-Init (ours)	69.3 ± 1.4%	61.9 ± 1.6%	69.6 ± 1.9%
X, A	**DGI** (ours)	**82.3** ± 0.6%	**71.8** ± 0.7%	**76.8** ± 0.6%
X, A, Y	GCN (Kipf & Welling, 2016a)	81.5%	70.3%	79.0%
X, A, Y	Planetoid (Yang et al., 2016)	75.7%	64.7%	77.2%

note that… supervised transductive SOTA(GraphSGAN)는 못 넘음

Inductive micro-averaged F_1 score

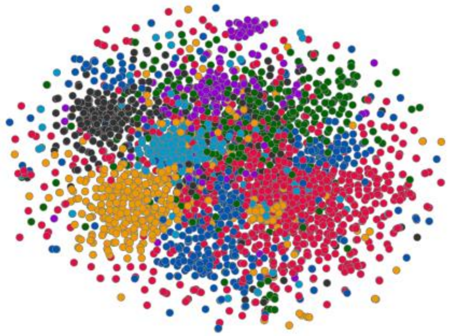
Available data	Method	Reddit	PPI
X	Raw features	0.585	0.422
A	DeepWalk (Perozzi et al., 2014)	0.324	—
X, A	DeepWalk + features	0.691	—
X, A	GraphSAGE-GCN (Hamilton et al., 2017a)	0.908	0.465
X, A	GraphSAGE-mean (Hamilton et al., 2017a)	0.897	0.486
X, A	GraphSAGE-LSTM (Hamilton et al., 2017a)	0.907	0.482
X, A	GraphSAGE-pool (Hamilton et al., 2017a)	0.892	0.502
X, A	Random-Init (ours)	0.933 ± 0.001	0.626 ± 0.002
X, A	**DGI** (ours)	**0.940** ± 0.001	**0.638** ± 0.002
X, A, Y	FastGCN (Chen et al., 2018)	0.937	—
X, A, Y	Avg. pooling (Zhang et al., 2018)	0.958 ± 0.001	0.969 ± 0.002

potential in the inductive node classification

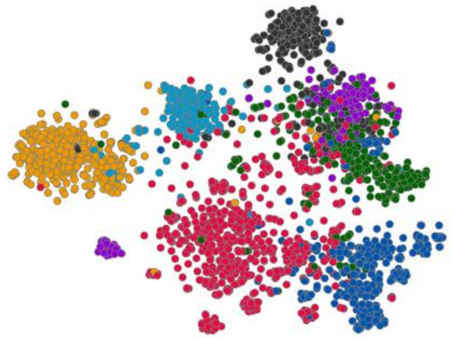
large gap… attributed to the extreme sparsity of available node features

Classification Performance

t-SNE embeddings of the nodes in the Cora dataset



raw features



learned DGI model

Silhouette score of 0.234

Conclusions

Deep InfoMax → Deep Graph Infomax

local mutual information maximization across the graph's patch representations
→ node embeddings that are mindful of the global structural properties of the graph

competitive performance across a variety of both **transductive** and **inductive** classification tasks

감사합니다