AAAI-22 Main Technical Track Paper

Augmentation-Free Self-Supervised Learning on Graphs

Namkyeong Lee, Junseok Lee, Chanyoung Park

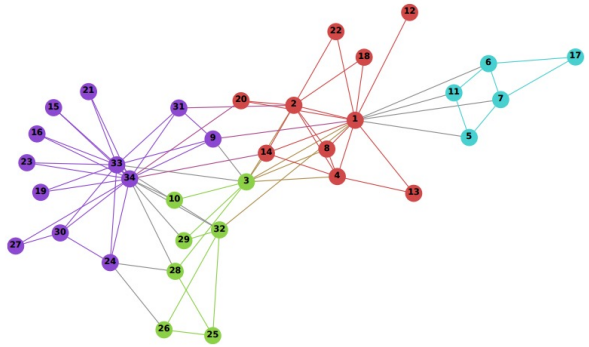
Korea Advanced Institute of Science and Technology (KAIST)

DSAIL Data Science & Artificial Intelligence

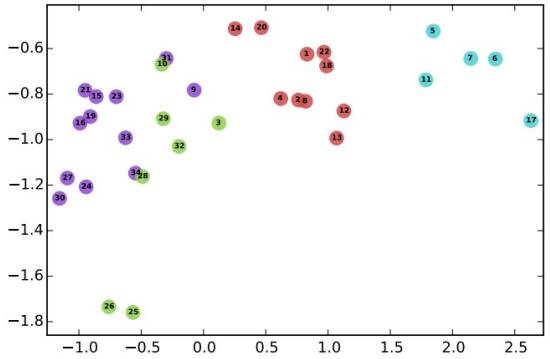
TABLE OF CONTENTS

BACKGROUND GRAPH REPRESENTATION LEARNING

Graph is ubiquitous data structure, employed extensively within computer science and related fields.



Graph Neural Network



Node Classification

Link Prediction

Community Detection

Network Similarity

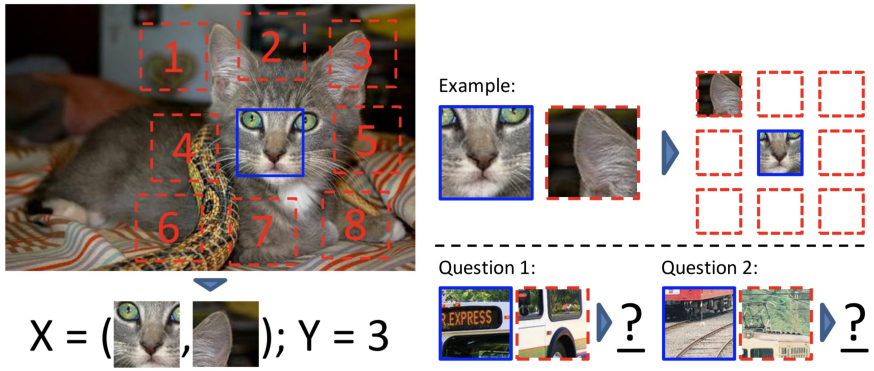
Graph representation learning means mapping the nodes or entire graphs, as points in a low-dimensional vector space.

Graph representation learning has been a powerful strategy for analyzing graph-structured data such as social network, especially by using Graph Neural Networks!

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES

Self-Supervised Learning automatically generate some kind of supervisory signal to solve some task.
(Typically, to learn representations of data or to automatically label a dataset.)

Key Idea
→ Define pretext training task that captures the information of the input data.
→ Use the dependencies among different dimensions of the input data!



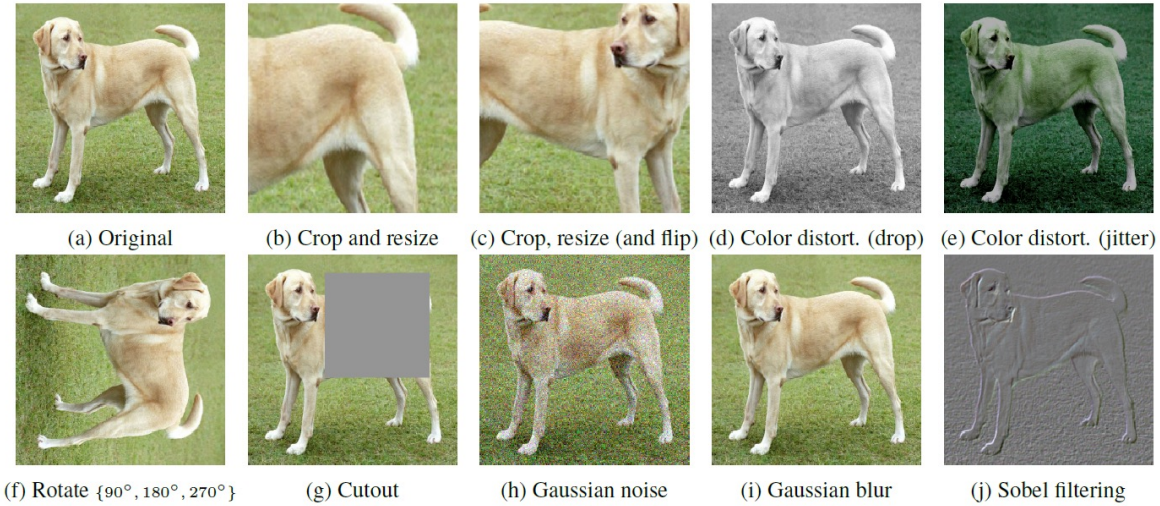
Self-Supervised Learning uses way more supervisory than supervised learning, and enormously more than reinforcement learning.
That's why calling it "unsupervisory" is totally misleading.

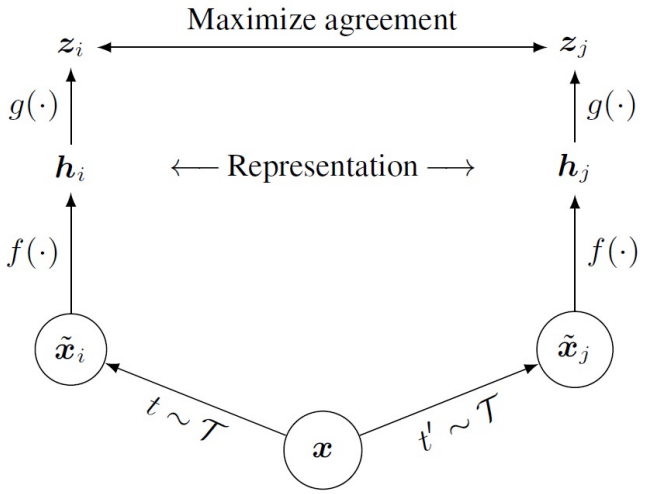
Yann LeCun, 2019

DOERSCH, Carl; GUPTA, Abhinav; EFROS, Alexei A. Unsupervised visual representation learning by context prediction. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 1422-1430.

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES _ SIMCLR

SimCLR is trained by reducing the distance between representations of different augmented views of the same image (Positive), and increasing the distance between representations of augmented views from different images (Negative).





Sample mini batch of N examples.

→ Create 2N data points via Data Augmentation.

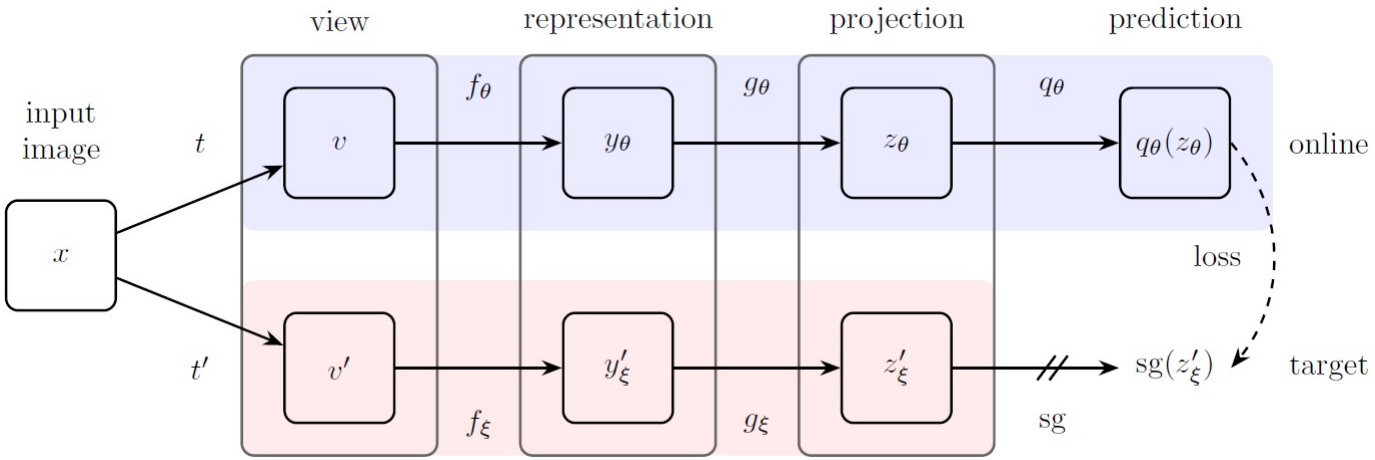
→ Given a positive pair, treat other 2(N-1) points as negative examples.

→ Instance Discrimination!

CHEN, Ting, et al. A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. PMLR, 2020. p. 1597-1607.

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES _ BYOL

BYOL learns representations of images without using negative samples
→ predicting the target representation with a given online representation



$$\theta \leftarrow \mathrm{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\mathtt{BYOL}}_{\theta,\xi}, \eta\right)$$

Online network

$$\xi \leftarrow \tau\xi + (1-\tau)\theta$$

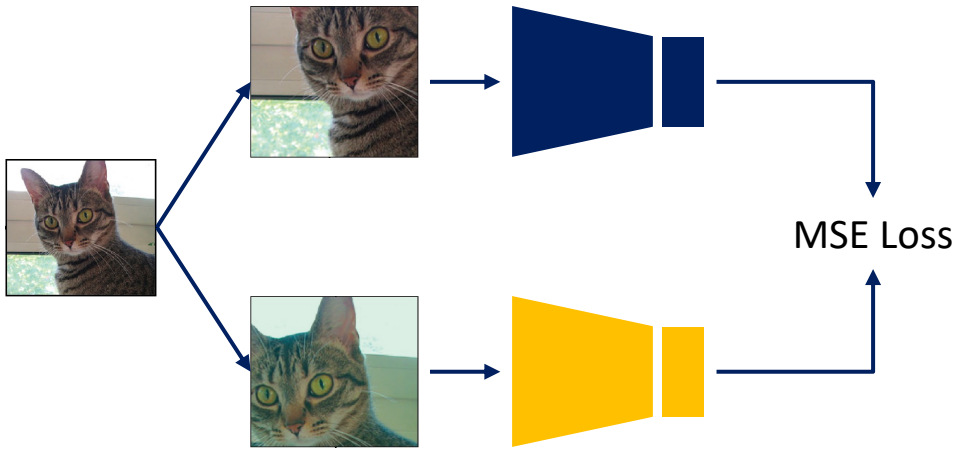
Target network Online network

At each training iteration, online network is trained to minimize the cosine similarity loss,
while target network's parameters are updated using the exponential moving average of online network's parameter.

GRILL, Jean-Bastien, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

BACKGROUND SELF-SUPERVISED LEARNING ON IMAGES

SimCLR and BYOL are cross-view prediction framework.
→ Learn representations by predicting different views of the same image which are created by using augmentation.



Cross-view prediction framework

Learning features that are invariant under the augmentation!

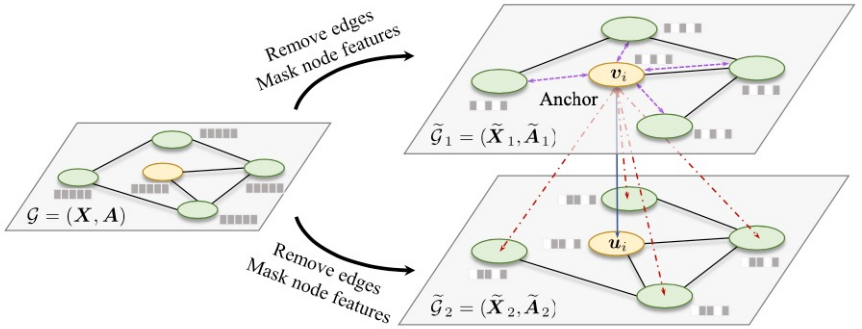
Augmentation was originally used to artificially expand the dataset.
But in here, augmentation is used to create different views of a image!

BACKGROUND SELF-SUPERVISED LEARNING ON GRAPHS

Inspired by the success of contrastive methods in computer vision applied on images, those methods have been recently adopted to graph-strucutured data.

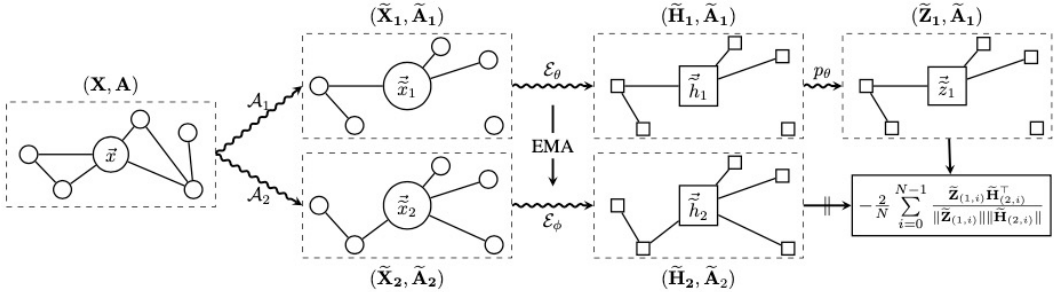
GRACE (Inspired by SimCLR)

Learns representations by pulling the representation of the same node in the two augmented views of graph while pushing apart representations of every other node.



BGRL (Inspired by BYOL)

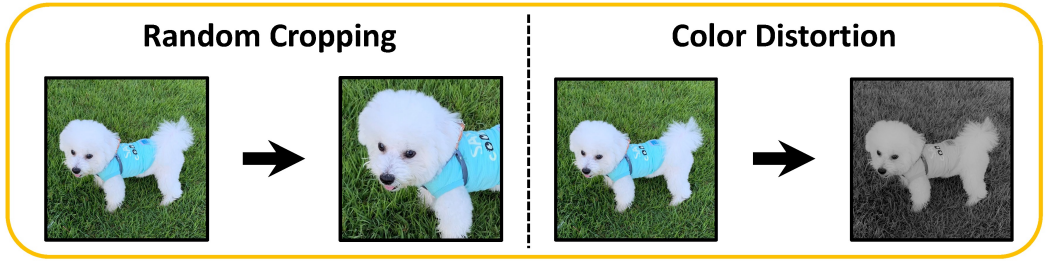
Learns representations by predicting the augmented view of node itself without using negative samples.



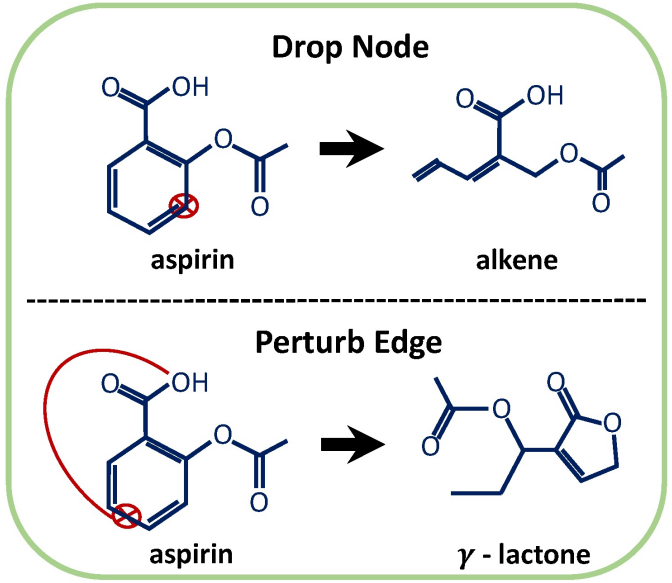
Simple augmentation operation is performed to generate views of the graph during training.

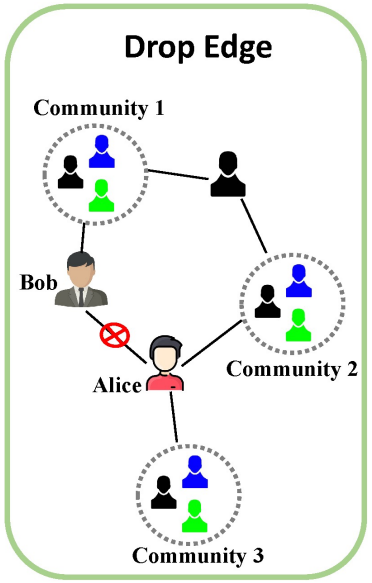
MOTIVATION IS AUGMENTATION APPROPRIATE FOR GRAPHS?

Image's underlying semantic is hardly changed after augmentation.



However in the case of graphs,
we cannot ascertain whether the augmented graph would be positively related to original graph.

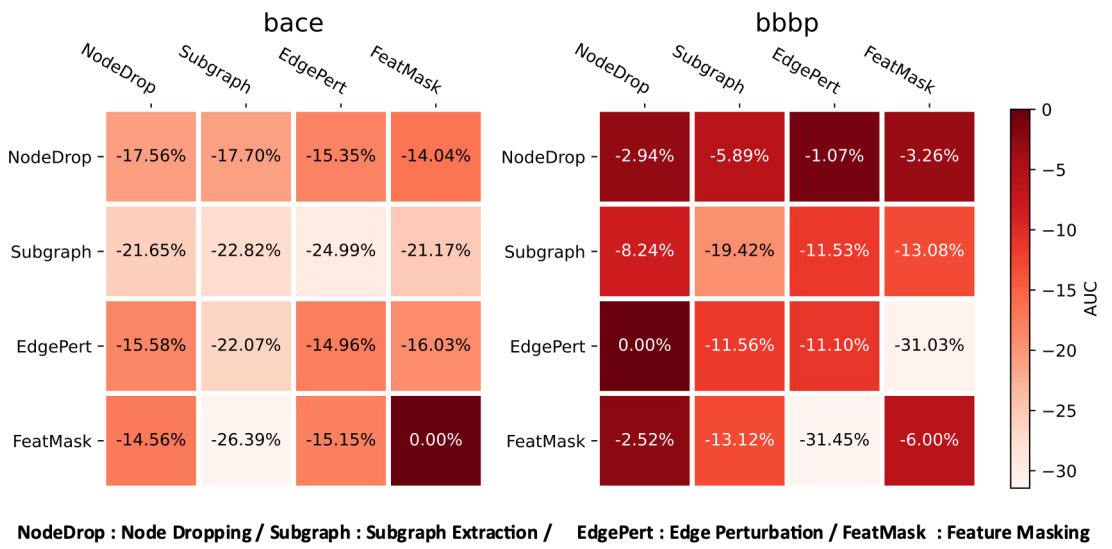




This is mainly because graphs contain not only the semantic but also the structural information.

MOTIVATION IS AUGMENTATION APPROPRIATE FOR GRAPHS?

		Comp.	Photo	CS	Physics
Node Classi.	BGRL	-4.00%	-1.06%	-0.20%	-0.69%
	GCA	-19.18%	-5.48%	-0.27%	OOM
Node Clust.	BGRL	-11.57%	-13.30%	-0.78%	-6.46%
	GCA	-26.28%	-23.27%	-1.64%	OOM



Performance sensitivity according to hyperparameters for augmentations on node level task (left) and graph level task (right).

We observed that the quality of the learned representations relies on the choice of augmentation scheme.
→ Performance on various downstream tasks varies greatly according to the choice of augmentation hyperparameters.

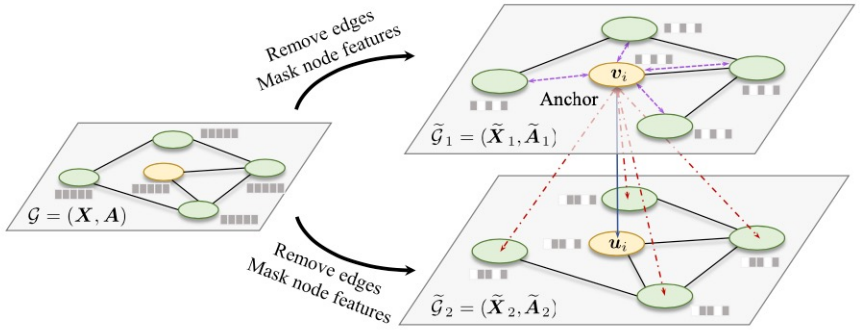
We need more stable and general framework for generating alternative view of the original graph without relying on augmentation.

MOTIVATION IS INSTANCE DISCRIMINATION APPROPRIATE FOR GRAPHS?

Another limitation arises due to the inherent philosophy of contrastive learning.

Limitations

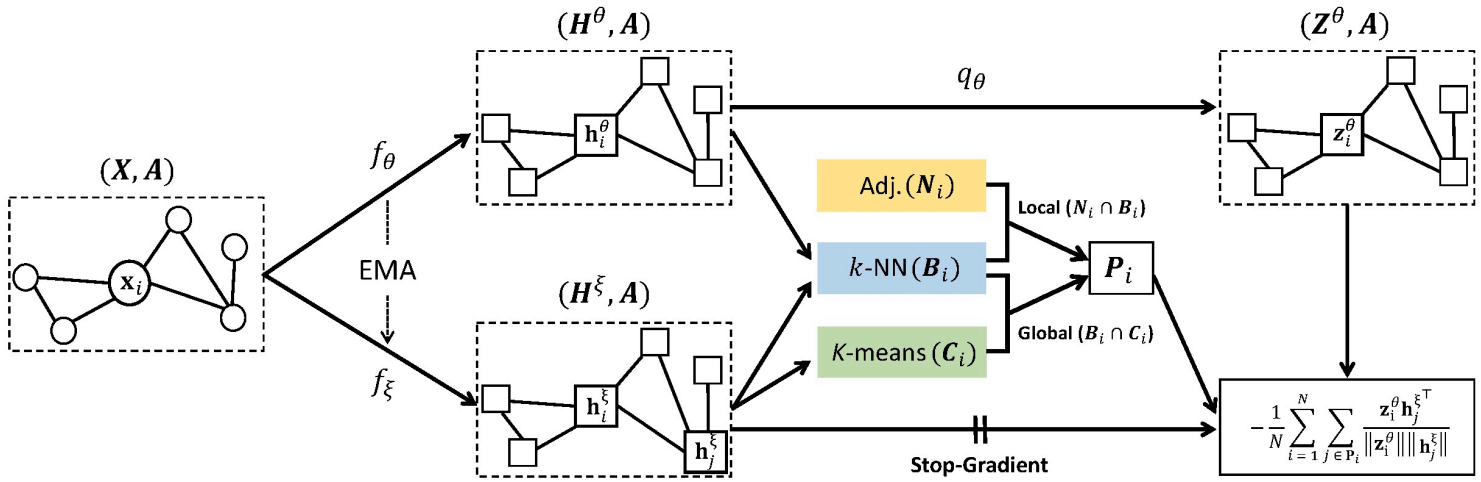
1. Treating all other nodes apart from the node itself as negatives overlooks the structural information of graphs.
2. Large amount of negative samples requires high computational and memory costs.



To this end, we propose a self-supervised learning framework for graphs called Augmentation-Free Graph Representation Learning (AFGRL).

Requires neither augmentation techniques nor negative samples for learning representations of graphs!

AUGMENTATION-FREE GRAPH REPRESENTATION LEARNING



Model architercture

Instead of creating two arbitrarily augmented views of graph, we use the original graph per se as one view.
And generate another view by discovering nodes that can be serve as positive samples via k-nearest neighbor search in embedding space.

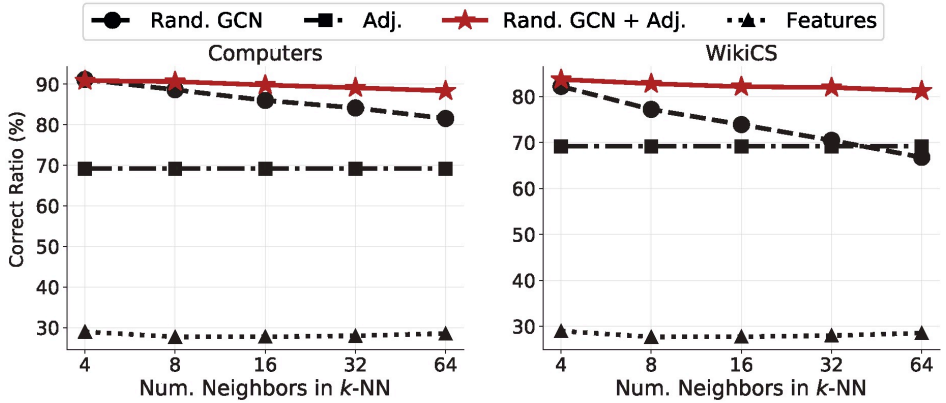
However, naively selected positive samples with kNN includes false positives.
→ Let's filter out false positives regarding local and global perspective!

AUGMENTATION-FREE GRAPH REPRESENTATION LEARNING

More than 10 % of positive candidates are false when using only k-NN.
→ We should filter out positive false positives from the samples discovered by k-NN search.

- Capturing Local Structural Information
  - AFGRL filters out nodes that are non-adjacent with query node
    → Following smoothness assumption of graph-structured data.

- Capturing Global Semantics
  - AFGRL filters out nodes that are not in the same cluster with query node
    → Discovering non-adjacent nodes that share the global semantics.



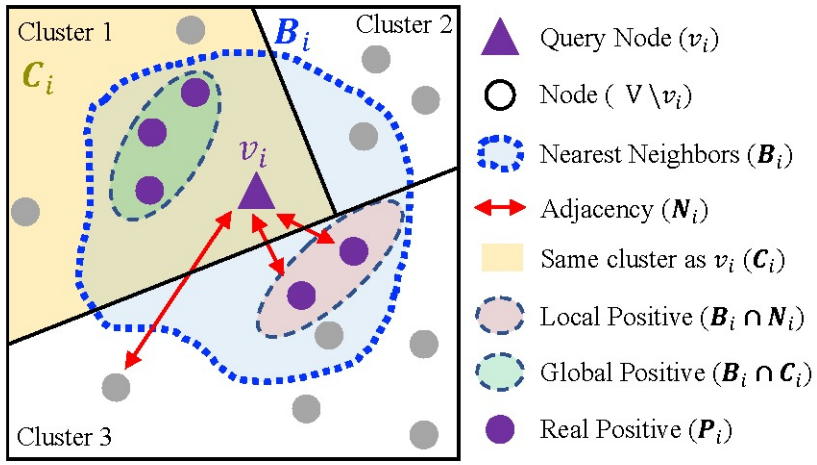
AUGMENTATION-FREE GRAPH REPRESENTATION LEARNING

B_i : Set of k-nearest neighbors of query node v_i
N_i : Set of adjacent nodes of query node v_i
C_i : Set of nodes that are in the same cluster with query node v_i

Real positives for node v_i → $\mathrm{P}_i = (\mathrm{B}_i \cap \mathrm{N}_i) \cup (\mathrm{B}_i \cap \mathrm{C}_i)$

Minimize the cosine distance between query node v_i and real positives P_i!

$$\mathcal{L}_{\theta,\xi} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{v_j \in \mathbf{P}_i} \frac{\mathbf{z}_i^{\theta}\mathbf{h}_j^{\xi\top}}{\left\|\mathbf{z}_i^{\theta}\right\|\left\|\mathbf{h}_j^{\xi}\right\|}$$



EXPERIMENTS AFGRL OUTPERFORMS PREVIOUS SOTA METHODS!

We performed several experiments to measure the quality of learned embeddings.

	WikiCS	Computers	Photo	Co.CS	Co.Physics
Sup. GCN	77.19 ± 0.12	86.51 ± 0.54	92.42 ± 0.22	93.03 ± 0.31	95.65 ± 0.16
Raw feats.	71.98 ± 0.00	73.81 ± 0.00	78.53 ± 0.00	90.37 ± 0.00	93.58 ± 0.00
node2vec	71.79 ± 0.05	84.39 ± 0.08	89.67 ± 0.12	85.08 ± 0.03	91.19 ± 0.04
DeepWalk	74.35 ± 0.06	85.68 ± 0.06	89.44 ± 0.11	84.61 ± 0.22	91.77 ± 0.15
DW + feats.	77.21 ± 0.03	86.28 ± 0.07	90.05 ± 0.08	87.70 ± 0.04	94.90 ± 0.09
DGI	75.35 ± 0.14	83.95 ± 0.47	91.61 ± 0.22	92.15 ± 0.63	94.51 ± 0.52
GMI	74.85 ± 0.08	82.21 ± 0.31	90.68 ± 0.17	OOM	OOM
MVGRL	77.52 ± 0.08	87.52 ± 0.11	91.74 ± 0.07	92.11 ± 0.12	95.33 ± 0.03
GRACE	**77.97 ± 0.63**	86.50 ± 0.33	92.46 ± 0.18	92.17 ± 0.04	OOM
GCA	77.94 ± 0.67	87.32 ± 0.50	92.39 ± 0.33	92.84 ± 0.15	OOM
BGRL	76.86 ± 0.74	89.69 ± 0.37	93.07 ± 0.38	92.59 ± 0.14	95.48 ± 0.08
AFGRL	77.62 ± 0.49	**89.88 ± 0.33**	**93.22 ± 0.28**	**93.27 ± 0.17**	**95.69 ± 0.10**

Performance on node classification

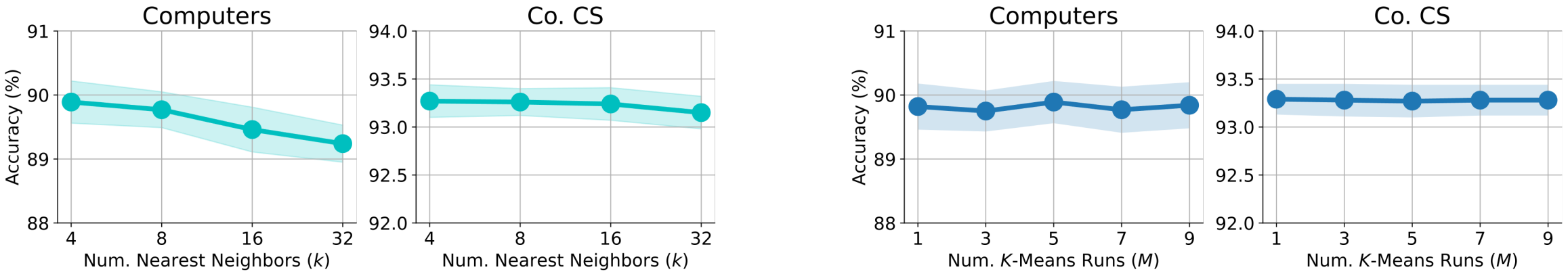
		GRACE	GCA	BGRL	AFGRL
WikiCS	NMI	**0.4282**	0.3373	0.3969	0.4132
	Hom.	**0.4423**	0.3525	0.4156	0.4307
Computers	NMI	0.4793	0.5278	0.5364	**0.5520**
	Hom.	0.5222	0.5816	0.5869	**0.6040**
Photo	NMI	0.6513	0.6443	**0.6841**	0.6563
	Hom.	0.6657	0.6575	**0.7004**	0.6743
Co.CS	NMI	0.7562	0.7620	0.7732	**0.7859**
	Hom.	0.7909	0.7965	0.8041	**0.8161**
Co.Physics	NMI	OOM	OOM	0.5568	**0.7289**
	Hom.	OOM	OOM	0.6018	**0.7354**

Performance on node clustering

AFGRL outperforms previous augmentation-based methods without using augmentation techniques nor negative samples.

EXPERIMENTS AFGRL IS ROBUST TO THE CHOICE OF HYPERPARAMETERS

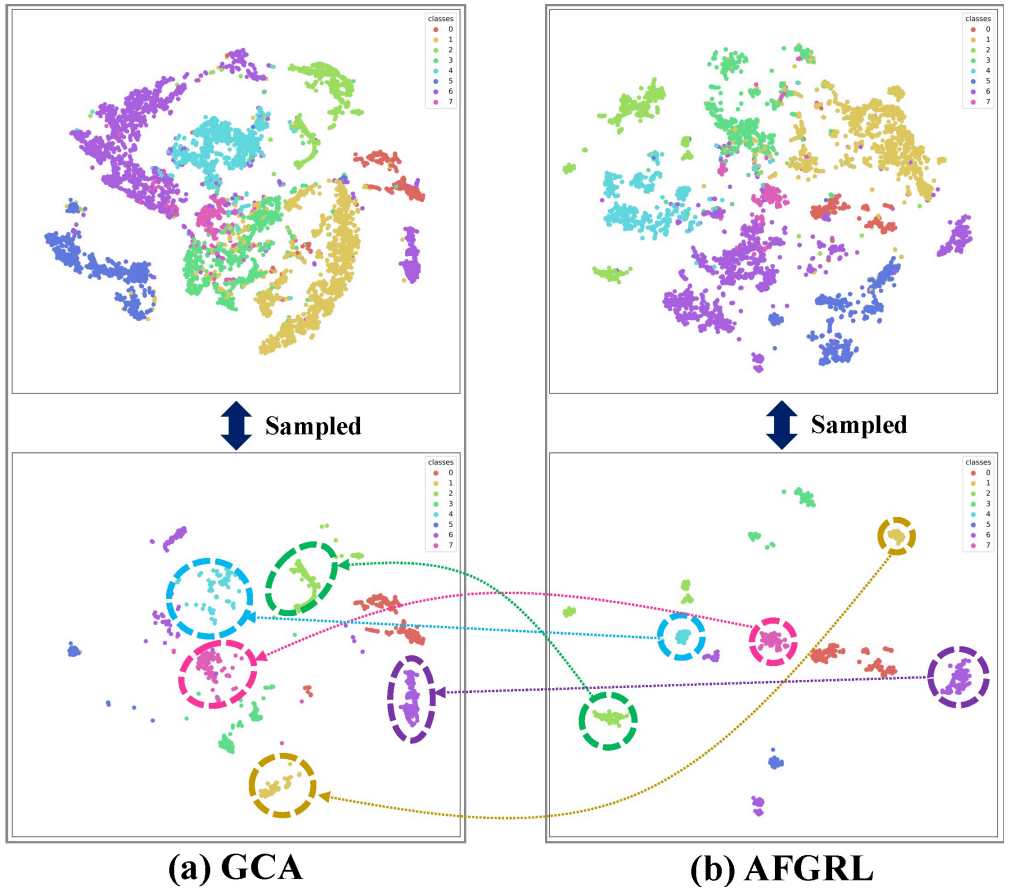
We performed sensitivity analysis on several hyperparameters to show robustness of AFGRL.



Sensitivity analysis on model hyperparameters

AFGRL shows stability over various hyperparameters.
→ Can be easily trained compared with other augmentation-based methods.

EXPERIMENTS AFGRL CAPTURES MORE FINE-GRAINED CLASS INFORMATION



AFGRL captures more fine-grained class information compared with GCA.
→ Nodes are more tightly grouped!

CONCLUSION

Previous graph representation learning methods have following limitations :

1. Using augmentation, which may alter the semantics of graph-structured data.
2. Treating all other nodes apart from node itself as negatives overlooking the structural information of graphs.

To this end we propose AFGRL, which requires neither augmentation techniques nor negative samples.

Instead of creating views of a graph using arbitrary augmentation,
we generate view of node by discovering nodes that can serve as positive samples via k-NN.

Regarding local and global semantics of graph, we filter out false positive candidates generated by k-NN.

- Local semantics are considered by using adjacency
- Global semantics are considered by using K-Means clustering

SUPPLEMENTARY MATERIALS

[Full paper] https://arxiv.org/abs/2112.02472

[Code] https://github.com/Namkyeong/AFGRL

[Author Email] namkyeong96@kaist.ac.kr