Self-Supervised Learning Methods for Chemical Property Prediction

2022. 04. 15

발표자: 허종국

발표자 소개

❖ 이름 : 허종국 (Jong Kook, Heo)

- Data Mining & Quality Analytics Lab
- M.S. Student (2021.03~)
- 지도 교수 : 김성범 교수님

❖ 관심 연구 분야

- Deep Reinforcement Learning
- Graph Neural Networks
- Self-Supervised Learning

❖ 연락망

- E-mail : hjkso1406@korea.ac.kr

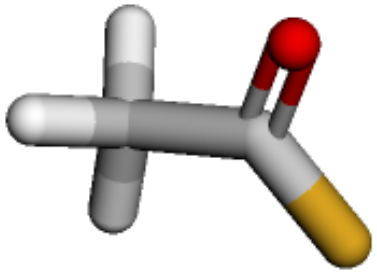
목 차

Introduction

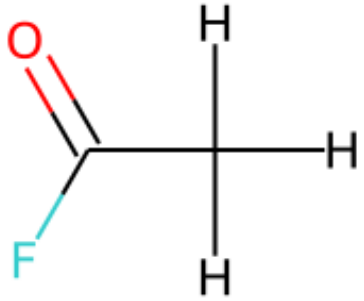
How to represent molecular structure?

❖ 3 Ways of Representing Molecular Structure

- 3D Coordinates : 결합의 길이나 각도 등의 3차원 형태가 중요한 경우에 사용
- 2D Connectivity Graph : 원자를 node, 결합을 edge 로 표현하는 그래프 생성
- Character Sequence : 분자 표기 규칙(SMILES)에 따라 구조를 문자열로 표현



3D Coordinates



2D Connectivity Graph

[H]C([H])([H])C(=O)F

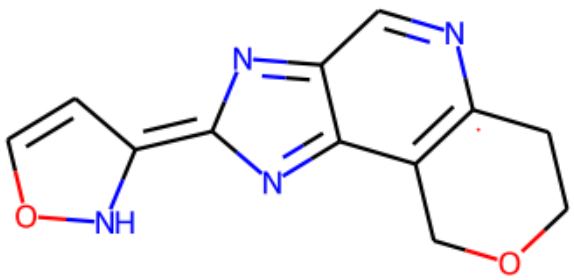
Sequence(SMILES String)

Introduction

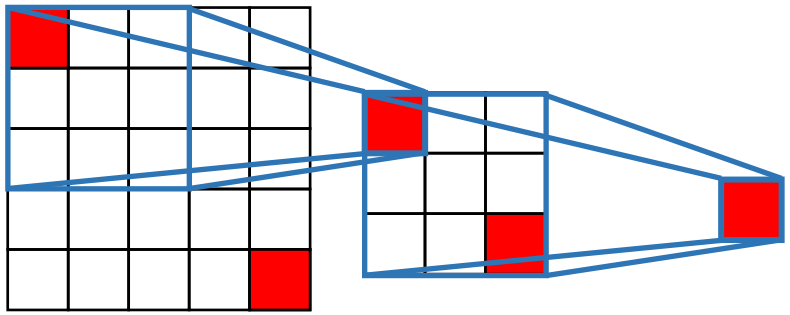
How to represent molecular structure?

❖ 2D Connectivity Graph

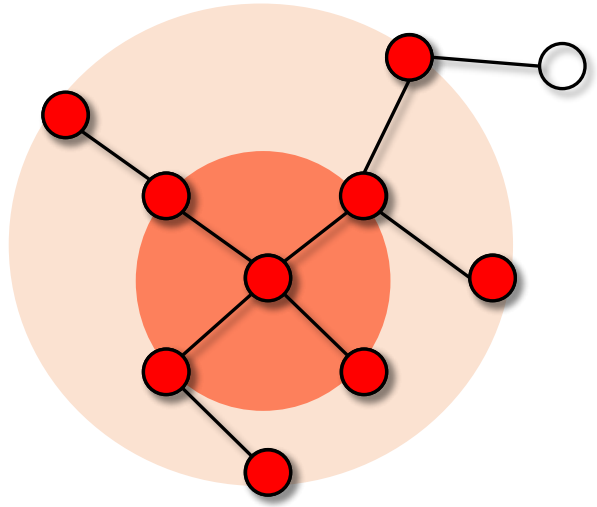
- GNN 계열의 모델을 통해 분자의 요약 정보를 추출
- 원자간 결합 정보 등의 필수적인 구조적 특징을 포착하기 쉬움
  - ✓ 복잡한 고리가 뒤얽힌 분자의 구조를 쉽게 파악 가능
- 원자간 최대 거리가 매우 큰 분자의 경우 특징을 포착하기 어려움
  - ✓ 원자 사이의 관계를 포착하기 위한 레이어가 매우 많이 필요



Complex Molecule



2D Grid Convolution



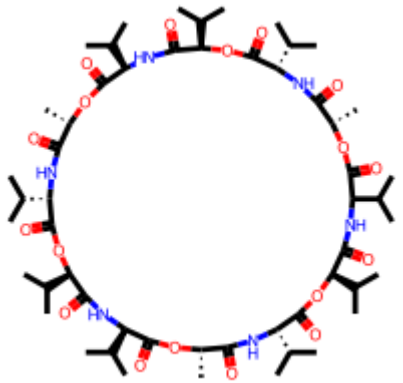
Graph Convolution

Introduction

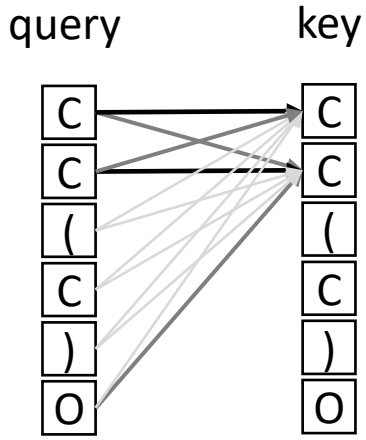
How to represent molecular structure?

❖ Character Sequence

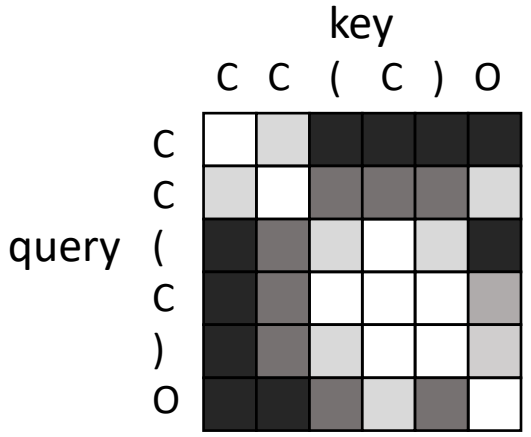
- Transformer 계열의 언어 모델을 통해 분자의 요약 정보를 추출
- 원자간 길이가 긴 분자에 대해 관계를 포착하기 쉬움
  - ✓ Layer 개수에 상관 없이 Self-Attention 연산으로 한번에 모든 원자 간 관계를 포착 가능
- 문자열이 나타내는 분자의 구조적 특징을 포착하기 힘듦
  - ✓ 문자열이 내포하고 있는 문법 규칙을 모델이 학습하는데 오랜 시간이 걸림



Large Molecule



Self-Attention



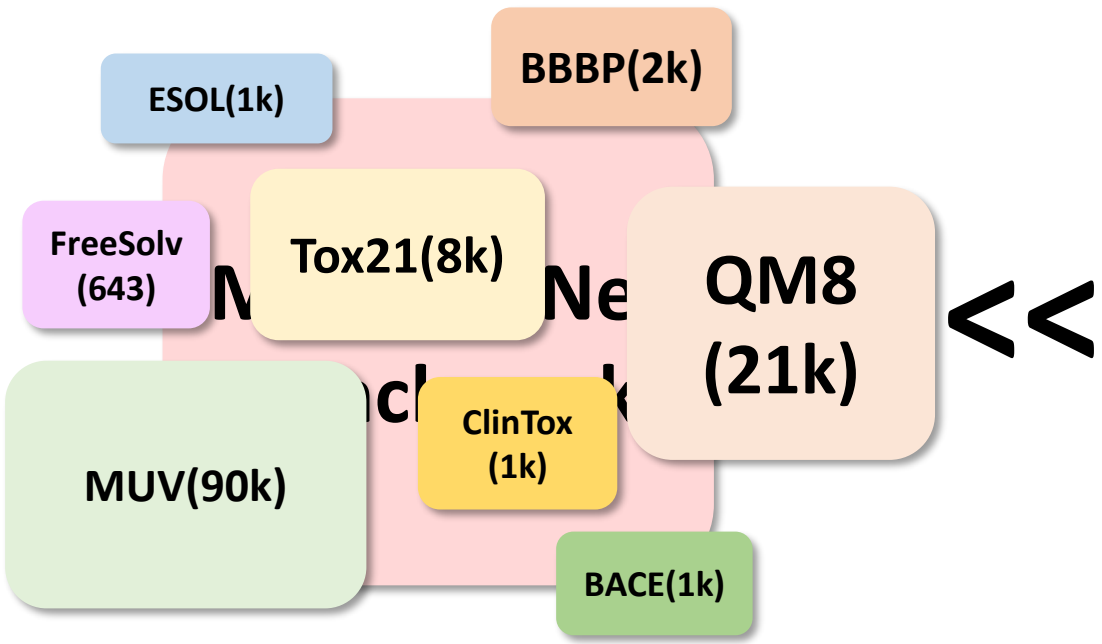
Self-Attention Map

Introduction

Deep Learning Applications in Chemical Domain

❖ What is Property Prediction??

- 특정 화학 분자의 성질(끓는점, 전도성, 방향성, 독성)을 예측하는 것
- MoleculeNet : 물성 예측을 위한 Benchmark Dataset 집합(0.6k~439k)
  - ✓ BBBP : 특정 분자의 뇌혈관장벽 투과성 추정(Binary Classification)



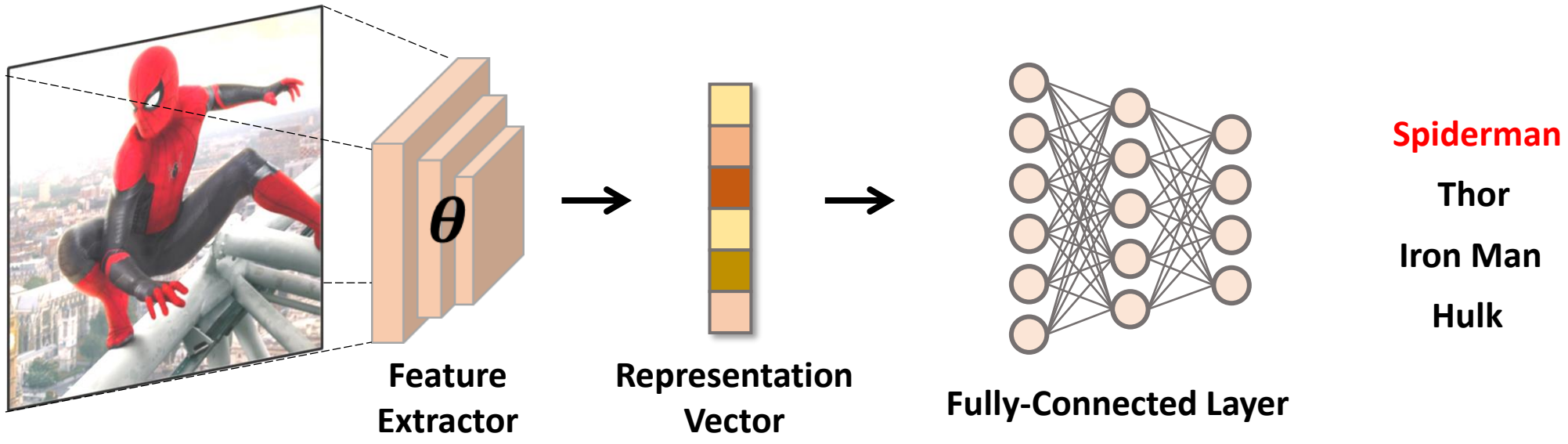
Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical science, 9(2), 513-530.

Preliminaries

What is Self-Supervised Learning?

❖ Supervised Learning Framework

- Feature Extractor : 입력 데이터로부터 중요한 정보를 요약, 추출
- Fully-Connected Layer : 요약정보로부터 목적에 알맞는 타겟값을 예측

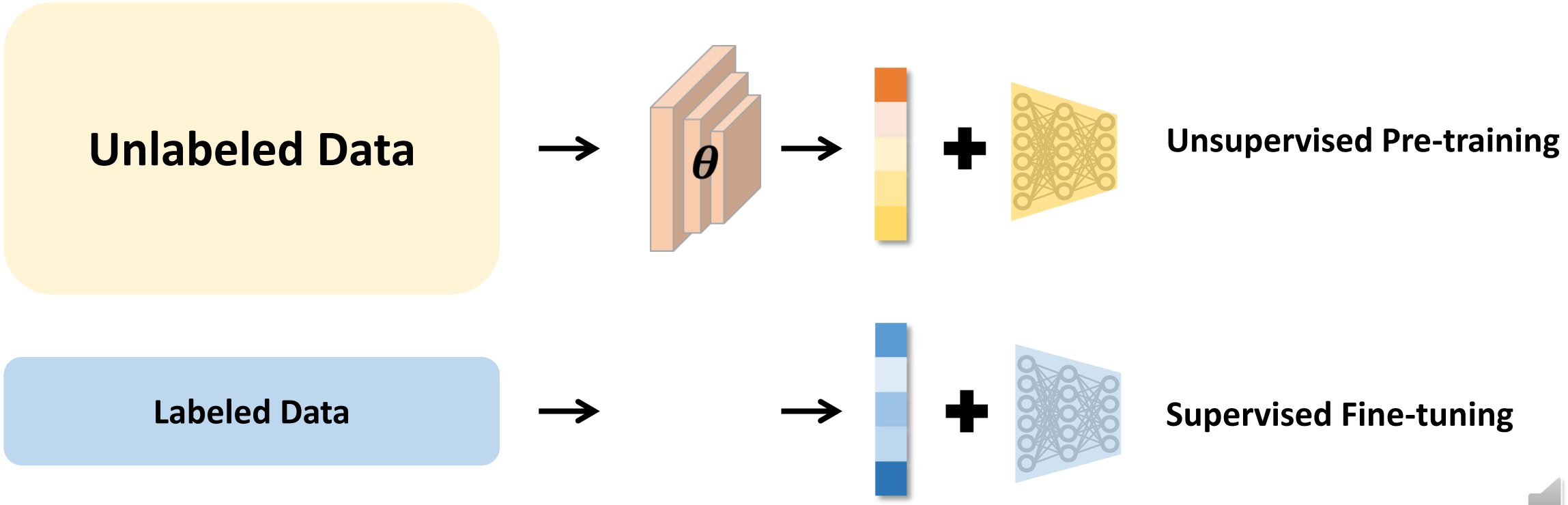


Preliminaries

What is Self-Supervised Learning?

❖ Self-Supervised Learning Framework

- Supervised Learning 을 위한 Labeled Data 를 구축하는 것은 시간/비용적 소모가 큼
- 대량의 Unlabeled Data 를 활용하여 Feature Extractor 를 학습할 수 없을까?

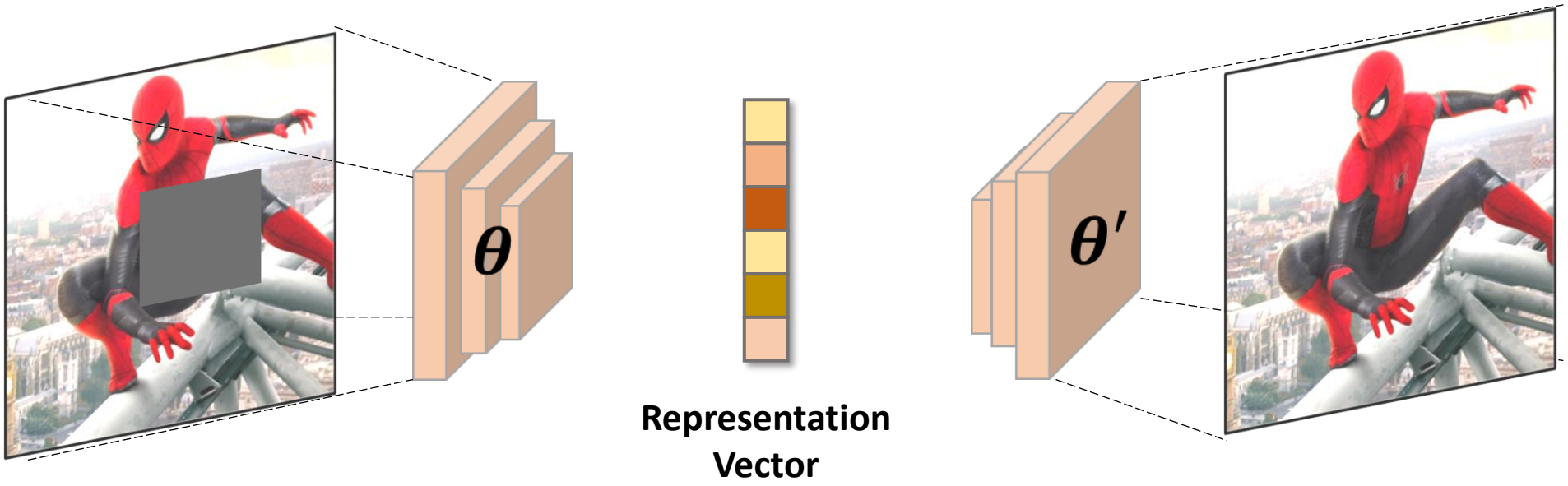


Preliminaries

Self-Supervised Learning in CV

❖ Pretext Task-based Learning

- Inpainting : 주변 이미지 픽셀 정보로부터 마스킹된 패치를 예측



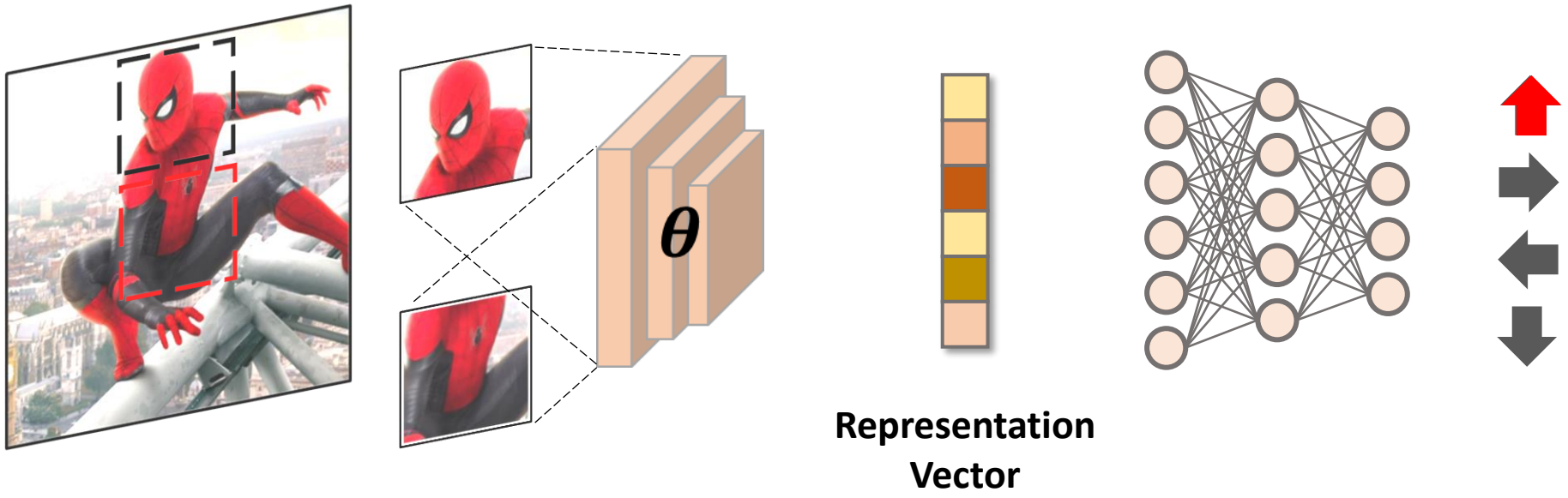
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536-2544).

Preliminaries

Self-Supervised Learning in CV

❖ Pretext Task-based Learning

- Context Prediction : 특정 패치를 중심으로 다른 패치의 상대적인 위치를 예측



Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision (pp. 1422-1430).

Preliminaries

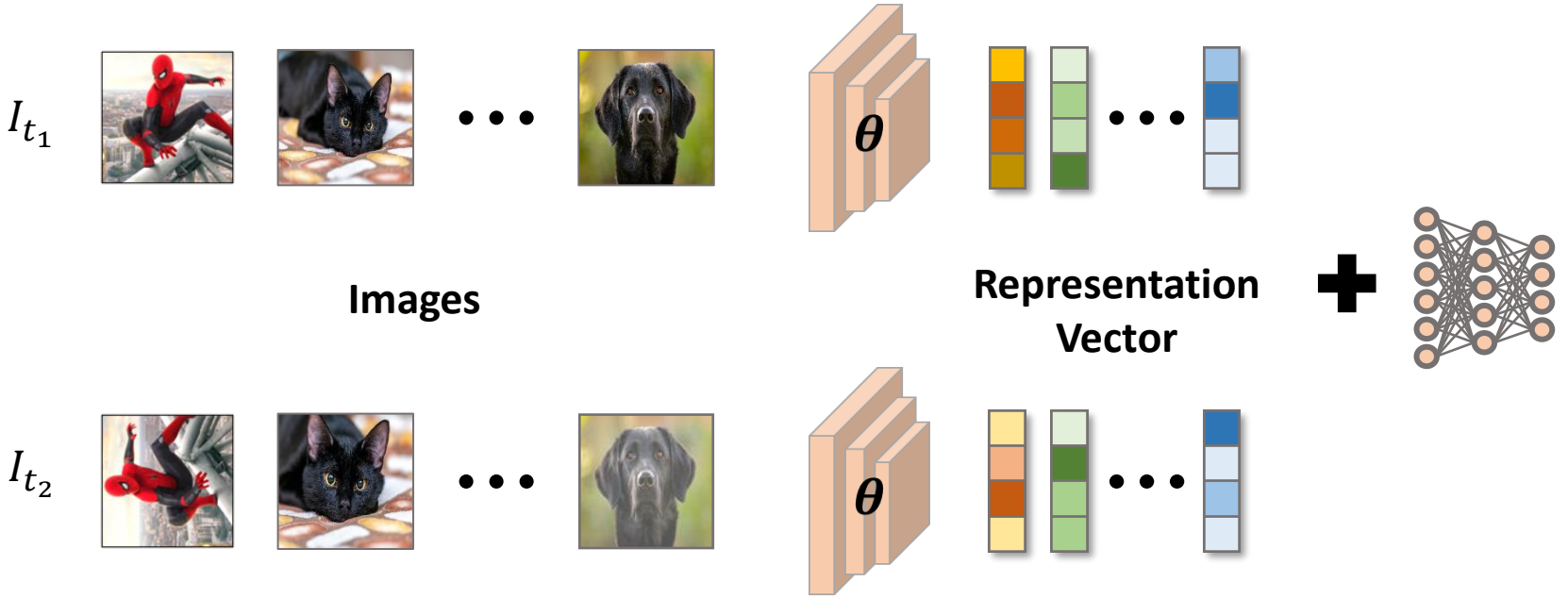
Self-Supervised Learning in CV

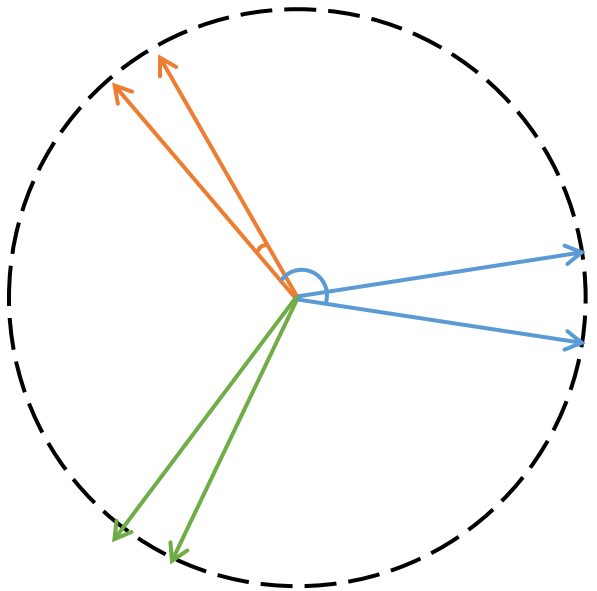
$$s_{i,j} = \frac{z_i^T z_j}{\|z_i\|\|z_j\|}$$

$$l_{i,j} = -\log\left(\frac{\exp\left(\frac{s_{i,j}}{\tau}\right)}{\Sigma_{k=1}^{2N} 1_{k\neq i} \exp\left(\exp\left(\frac{s_{i,k}}{\tau}\right)\right)}\right)$$

$$InfoNCE = \frac{1}{2N}\Sigma_{k=1}^{N}(l_{2k-1,2k} + l_{2k,2k-1})$$

- ❖ Contrastive Learning(SimCLR)
  - Positive Pair : 같은 이미지에서 서로 다른 데이터 증강 기법을 적용한 샘플
  - Negative Pair : 서로 다른 이미지로부터 데이터 증강 기법을 적용한 샘플
  - Positive Pair 는 가깝게, Negative Pair는 멀어지도록 학습(Cosine Similarity)





Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

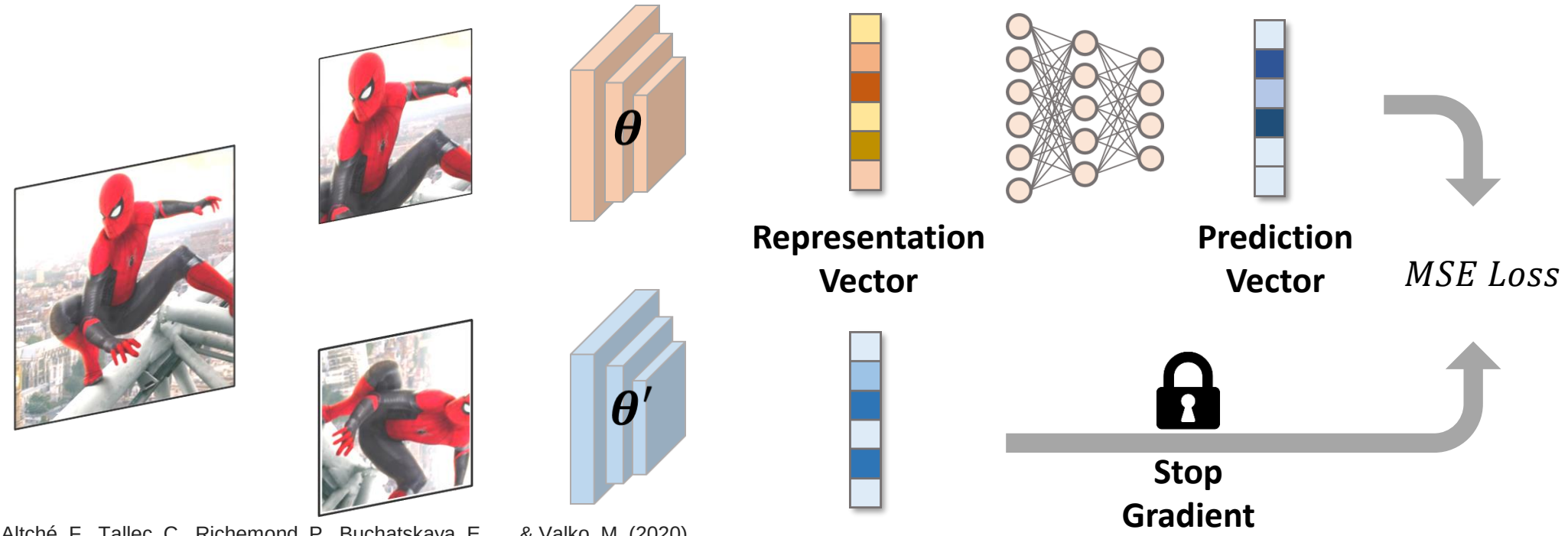
$$l_{I_i,I_j} = MSE\left(\|q(z_\theta(I_i))\|_2, \|SG(z_{\theta'}(I_j))\|_2\right)$$
$$Loss_{BYOL} = l_{I_i,I_j} + l_{I_j,I_i}$$

Preliminaries

Self-Supervised Learning in CV

❖ Non-Contrastive Learning(BYOL)

- 이미지에서 서로 다른 데이터 증강 기법을 적용하여 패치를 생성
- Online Network 와 Target Network에서 Representation Vector 를 추출
- Online Network 의 Representation Vector 로부터 Target Network 의 Representation Vector 를 예측



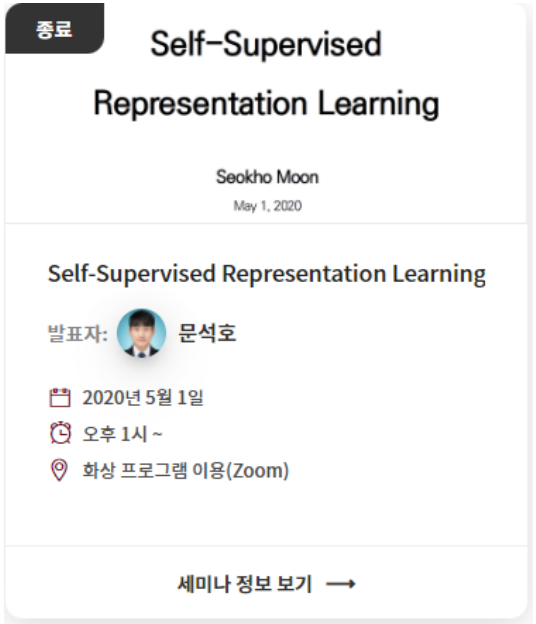
Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271-21284.

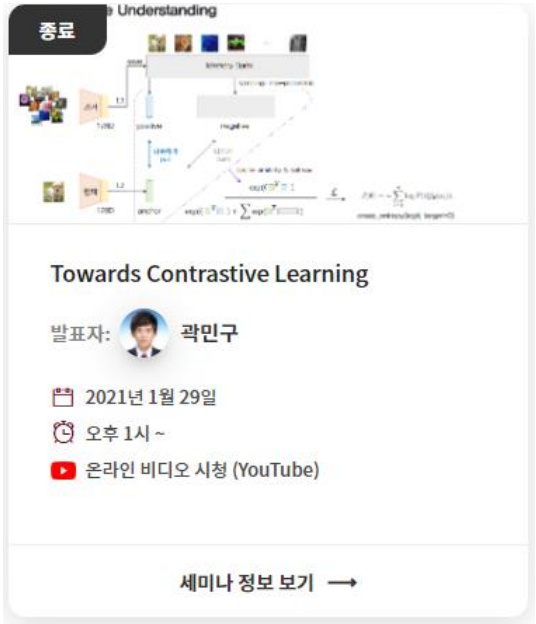
Preliminaries

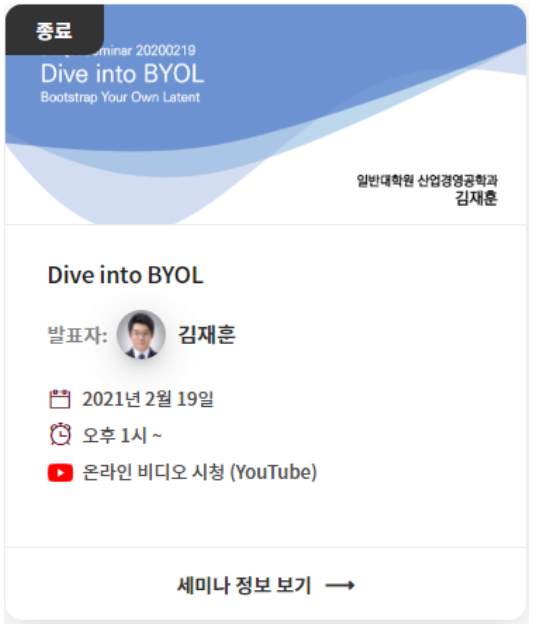
Self-Supervised Learning in CV

❖ Reference Materials of Self-Supervised Learning in CV

- Self-Supervised Representation Learning - Exemplar, Context Prediction, Jigsaw, Colorization, Inpainting, Count, Rotation
- Towards Contrastive Learning - Concept of SSL, NPID, MoCo v1&v2, SimCLR, False Negative Cancellation
- Dive into BYOL - Concept of Representation Learning, Details of BYOL







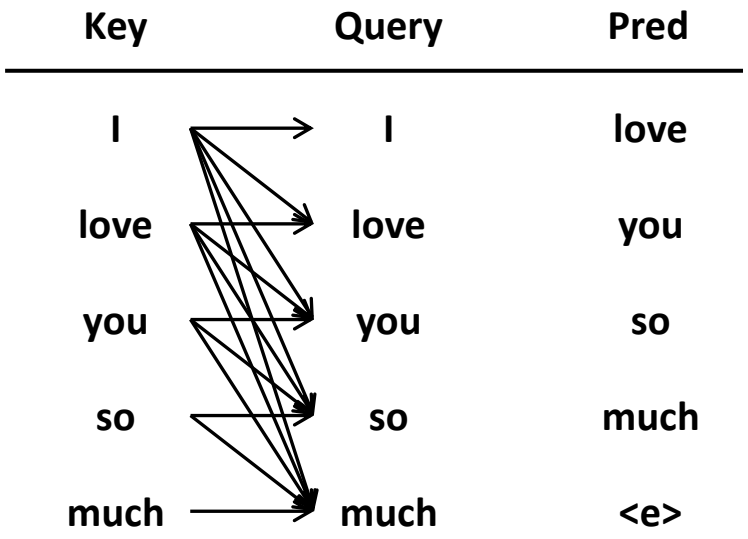
$$\boldsymbol{x} = [x_1, x_2, \ldots, x_T]$$
$$Loss_{LM} = \Sigma_i log P_\theta(x_i | x_{i-1}, x_{i-2}, \ldots.)$$

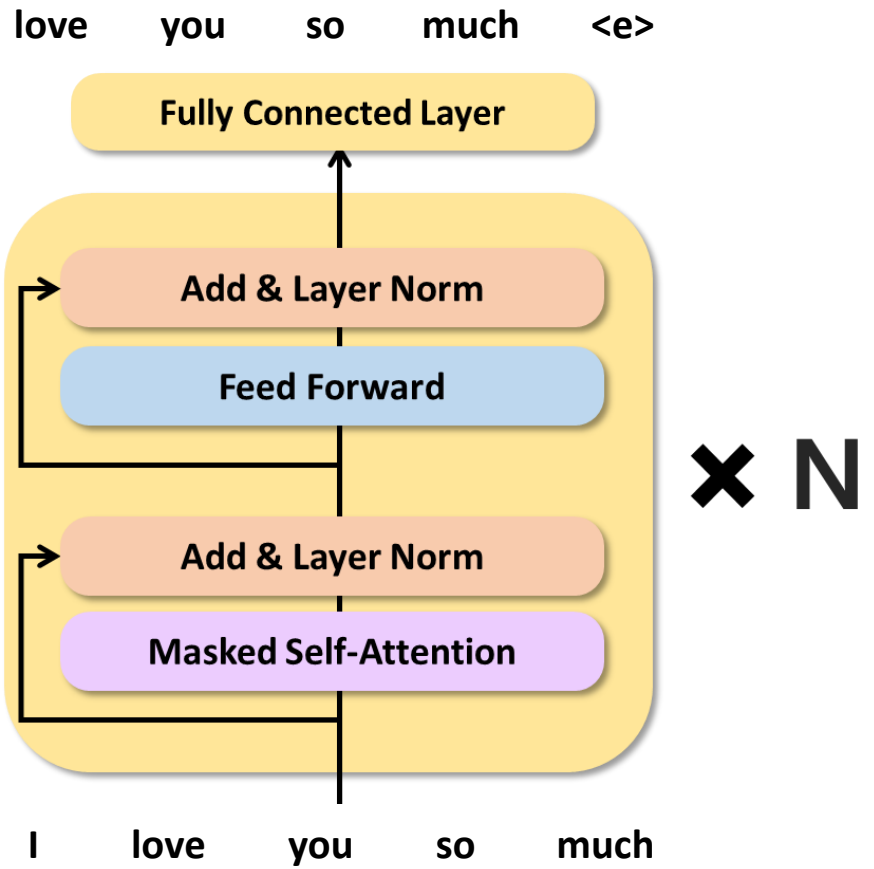
Preliminaries

Self-Supervised Learning in NLP

❖ Pretext Task-based Learning

- Generative Language Model(GPT) : 이전 시점까지의 시퀀스를 통해 다음 시점의 토큰을 예측





Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

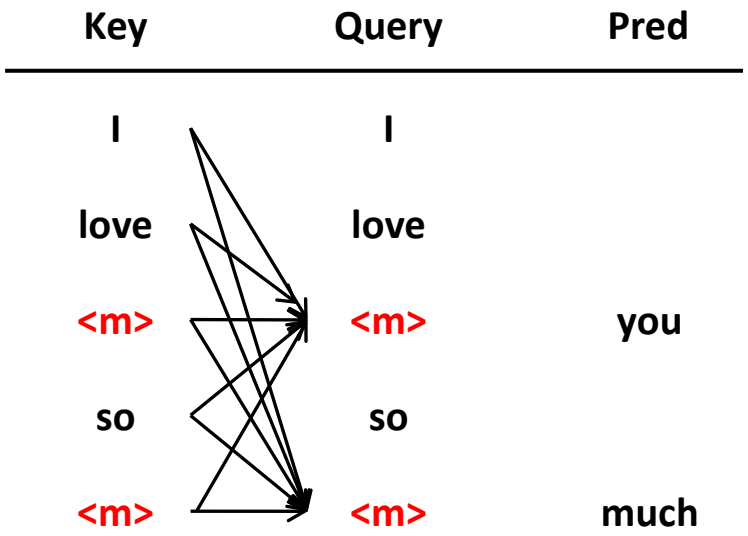
Preliminaries

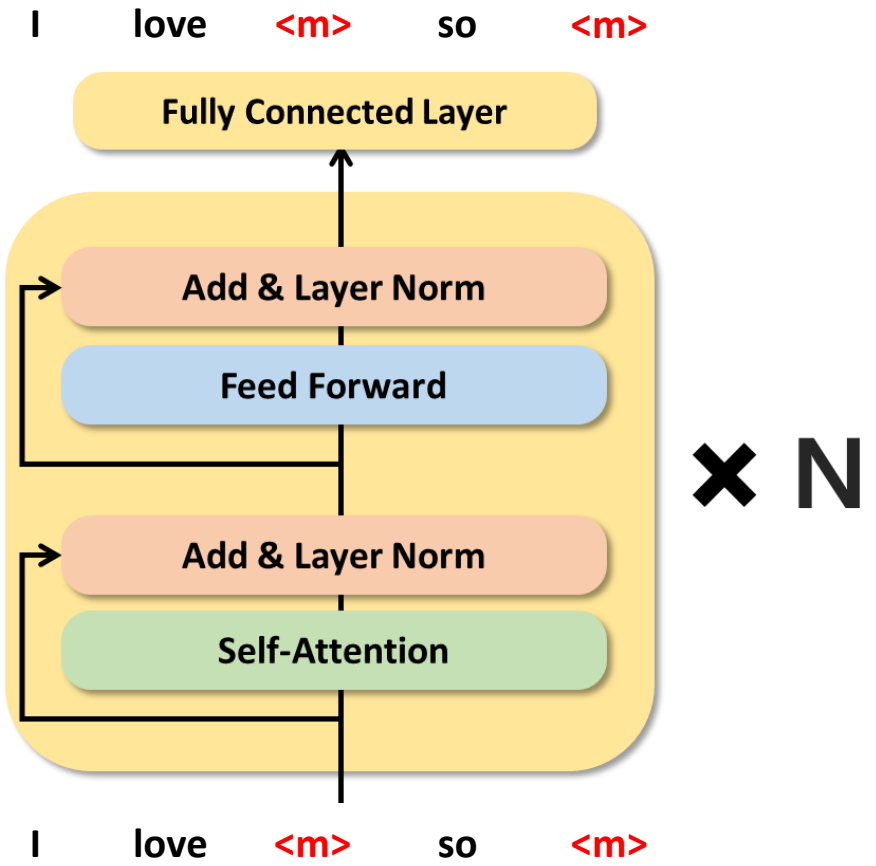
Self-Supervised Learning in NLP

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_{T-1}, x_T]$$
$$m(\boldsymbol{x}) = [x_1, \boldsymbol{m}, \ldots, x_{T-1}, \boldsymbol{m}]$$
$$Loss_{MLM} = \Sigma_i 1_{x_i=\boldsymbol{m}} log P_\theta(x_t | m(\boldsymbol{x}))$$

❖ Pretext Task-based Learning

- Masked Language Model(BERT) : 주변 토큰의 정보를 통해 가려진 토큰 예측





Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Methods

❖ Self-Supervised Learning for Chemical Property Prediction

Pre-training Strategy	Architecture	Algorithm
Masked Language Model	Transformer	**ChemBERTa**
Contrastive Learning	Graph Neural Network	**MolCLR**
BYOL	Both	**DMP**

Methods

ChemBERTa

❖ ChemBERTa : Large-Scale Self-Supervised Pretraining for Molecular Property Prediction(Chithrananda et al, 2020)

- University of Toronto 에서 연구
- Masked Language Model Pre-training 방식으로 RoBERTa 모델을 학습
- 기존 지도학습 모델보다 뛰어난 성능을 보이지 않았음
  - ✓ Pre-train Dataset 의 크기가 커질수록 Fine-tuning 성능이 증가하는 것을 통해 제안 방법론의 타당성 입증

ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction

Seyone Chithrananda
University of Toronto
seyone.chithrananda@utoronto.ca

Gabriel Grand
Reverie Labs
gabe@reverielabs.com

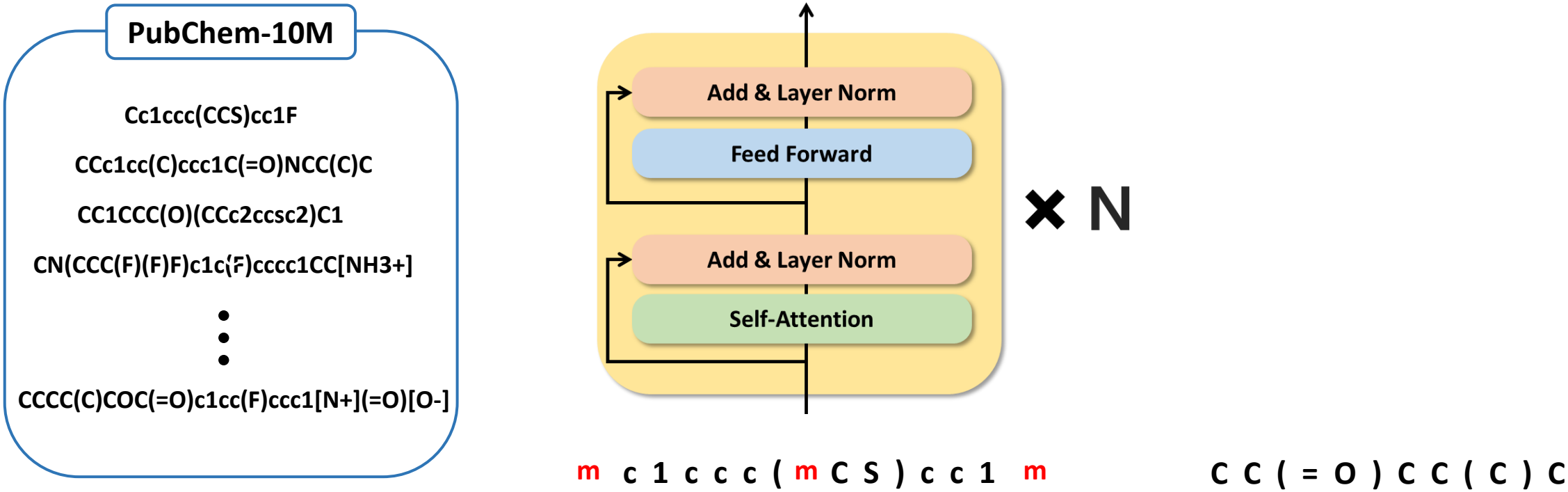
Bharath Ramsundar
DeepChem
bharath.ramsundar@gmail.com

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

Methods

ChemBERTa

❖ Pre-training Process(Masked Language Model) & Fine-tuning



Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

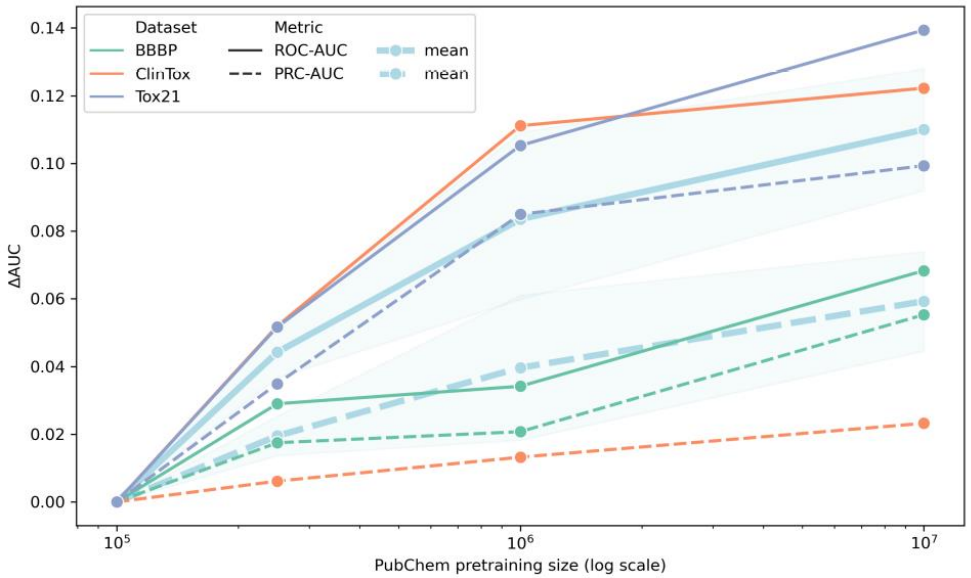
Methods

ChemBERTa

❖ Results

- 기존 지도학습 모델보다 뛰어난 성능을 보이진 않음
- 사용한 Fine-tune Dataset 또한 제한적
- Pre-train Dataset 크기에 따라 성능이 증가한다는 Scalability 를 보여줌

	BBBP 2,039		ClinTox (CT_TOX) 1,478		HIV 41,127		Tox21 (SR-p53) 7,831	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
ChemBERTa 10M	0.643	0.620	0.733	0.975	0.622	0.119	**0.728**	0.207
D-MPNN	**0.708**	0.697	**0.906**	**0.993**	0.752	0.152	0.688	**0.429**
RF	0.681	0.692	0.693	0.968	**0.780**	**0.383**	0.724	0.335
SVM	0.702	**0.724**	0.833	0.986	0.763	0.364	0.708	0.345

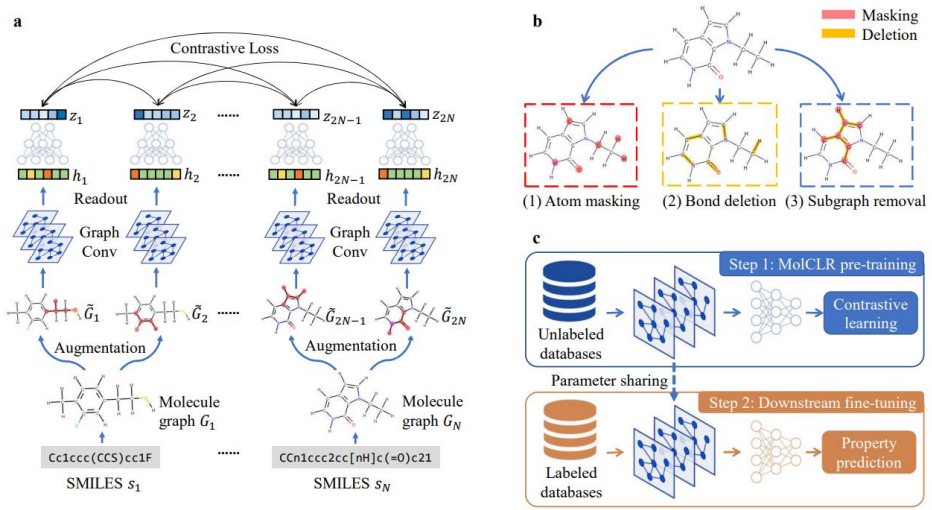


Chithrananda, S., Grand, G., & Ramsundar, B. (2020). Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

Methods

MolCLR

- Molecular Contrastive Learning of Representations via Graph Neural Networks(Wang et al, 2022)
  - Carnegie Mellon 대학에서 연구하였으며 2022년 3월 Nature Machine Intelligence 에 게재
  - edge feature 로써 **결합 정보를 활용 할 수 있도록 GNN 구조를 변형**
  - 분자 도메인에 Contrastive Learning 을 적용하기 위해 **다양한 분자 데이터 증강 기법**을 제안
  - MolCLR를 통해 추출한 **분자의 Representation Vector 에 대한 해석 제공**



MolCLR Architecture

Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

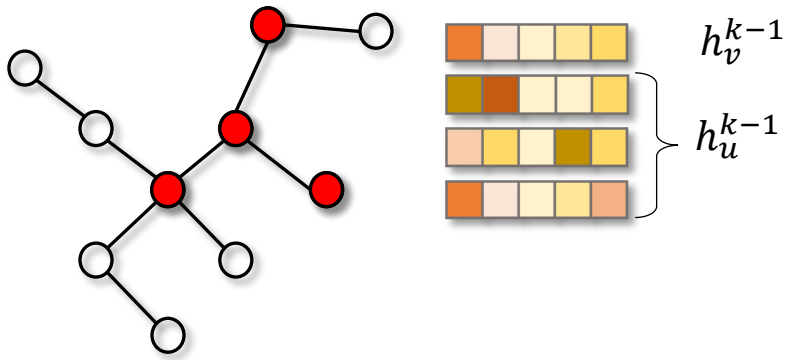
Methods

MolCLR

❖ Architecture Change

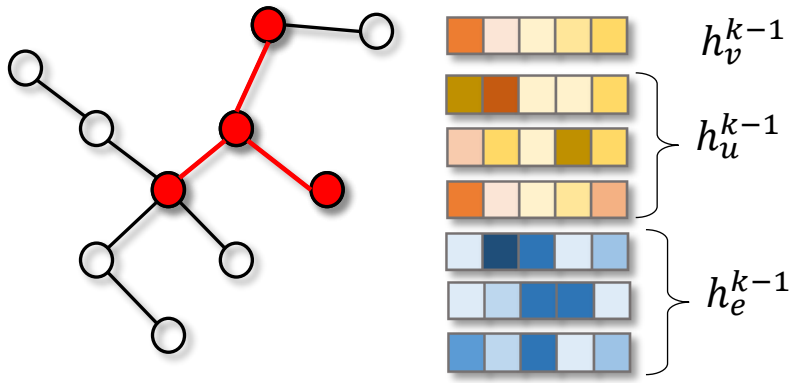
- 기존의 GCN/GIN Layer 는 edge attribute 를 반영하지 않음
- 결합의 정보로써 edge feature 를 활용할 수 있도록 GNN Layer 를 변형

GIN Convolution



$$h_v^k = \mathrm{ReLU}\left(\mathrm{MLP}^{(k)}\left(\sum_{u \in N(v) \cup \{v\}} h_u^{k-1}\right)\right)$$

GINE Convolution



$$h_v^k = \mathrm{ReLU}\left(\mathrm{MLP}^{(k)}\left(\sum_{u \in N(v) \cup \{v\}} h_u^{k-1} + \sum_{e=(v,u):u \in N(v)} h_e^{k-1}\right)\right)$$

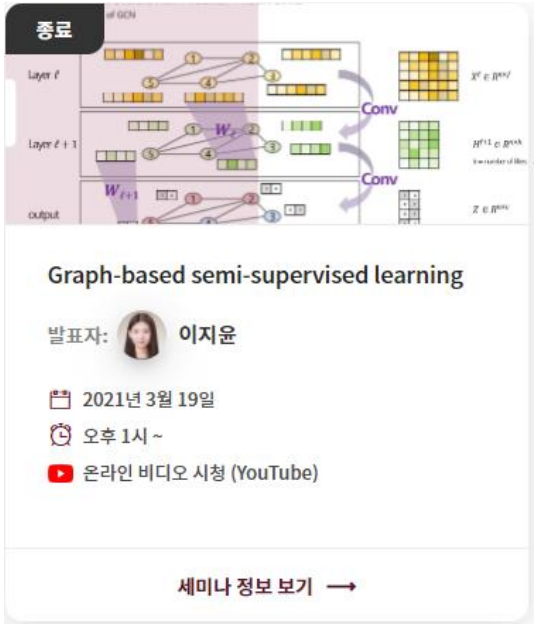
Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

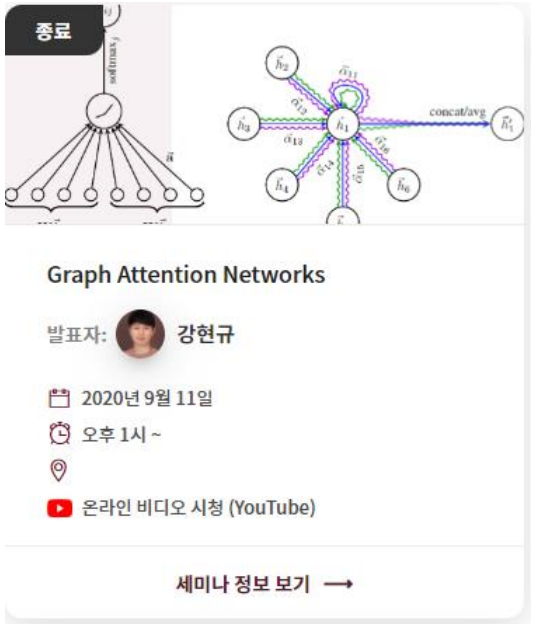
Methods

MolCLR

❖ Reference Materials of Graph Neural Networks

- Graph-based semi-supervised learning - Basic concept of GNN, Label Propagation with GCN
- Graph Attention Networks - Basic concept of GNN, GCN, GGNN, GraphSAGE, GAT



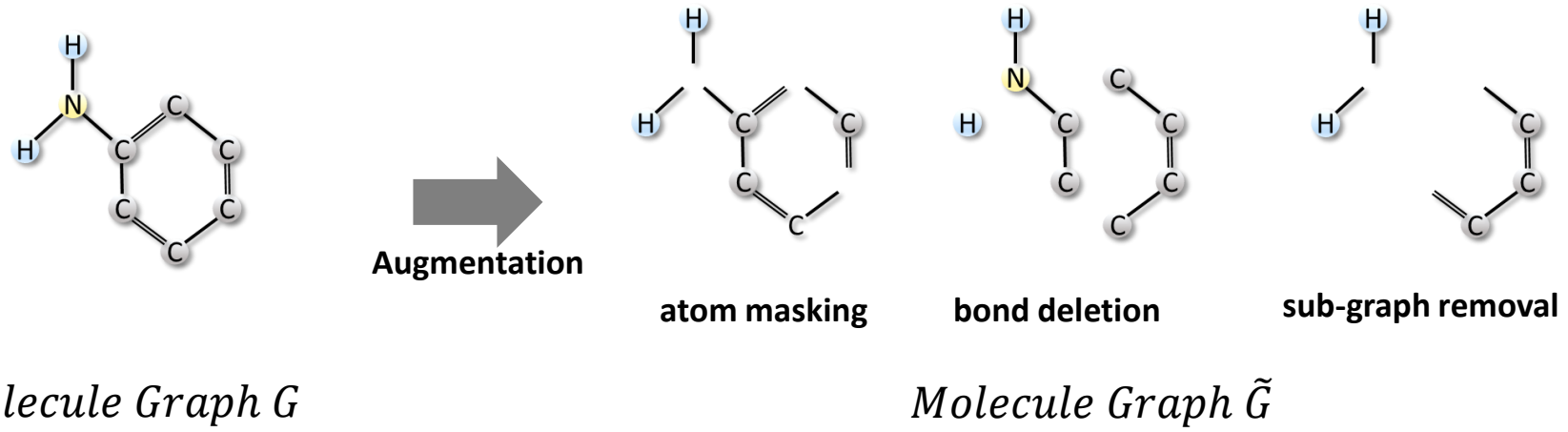


Methods

MolCLR

❖ Molecule Graph Augmentation

- Contrastive Learning 을 학습하기 위해 분자 도메인에서 Positive/Negative Pair 를 정의
- Molecule Graph Augmentation : 특정 분자의 원자/결합/sub-graph 정보를 제거

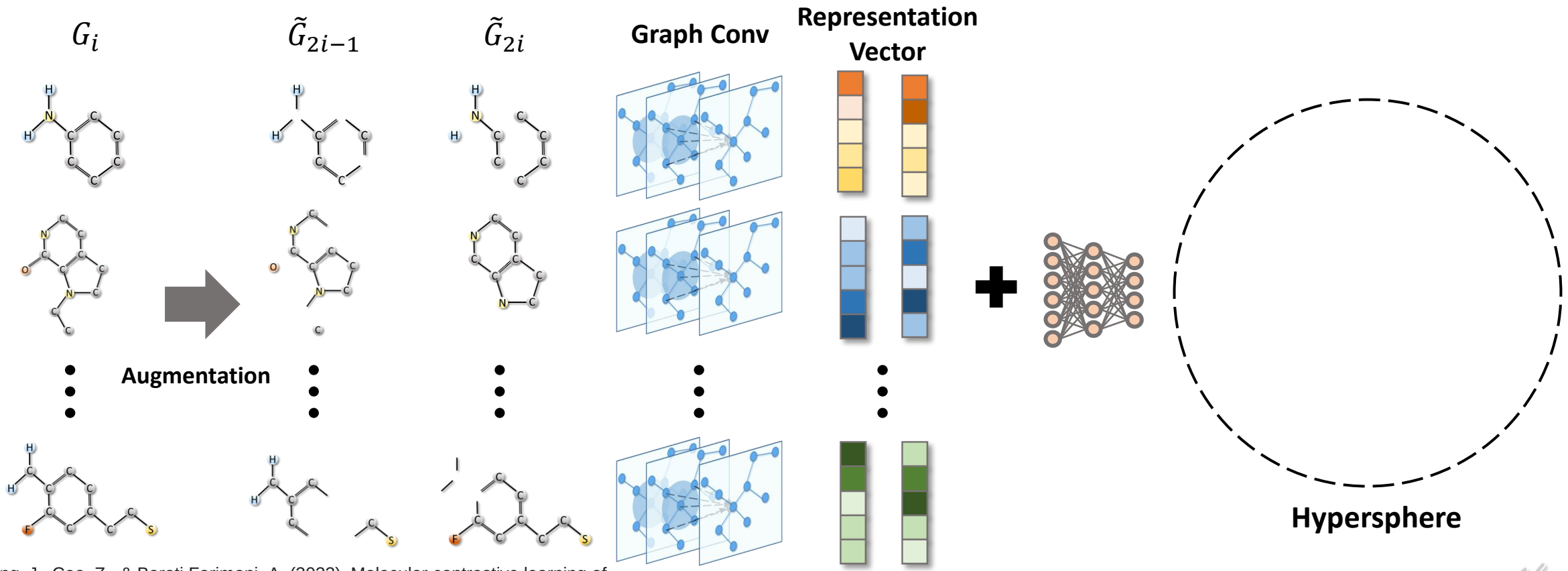


Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

Methods

MolCLR

❖ Pre-training Process(SimCLR)



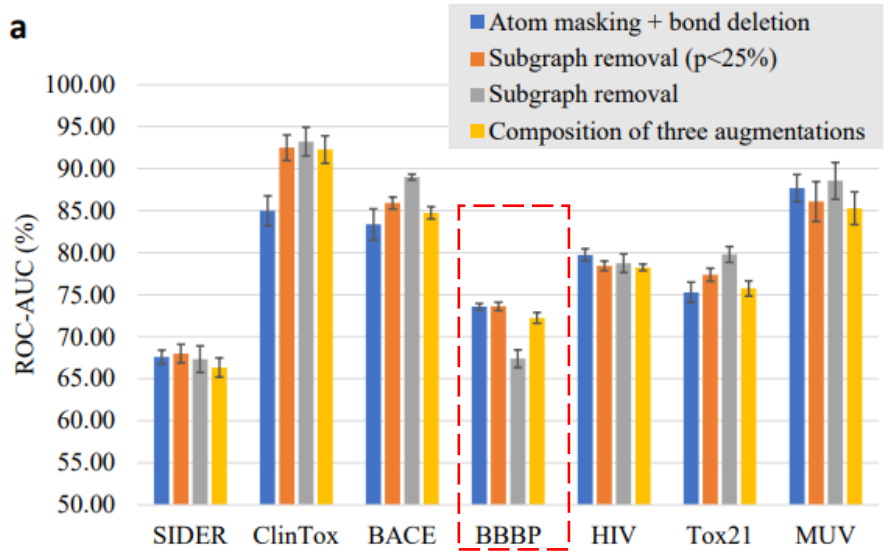
Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

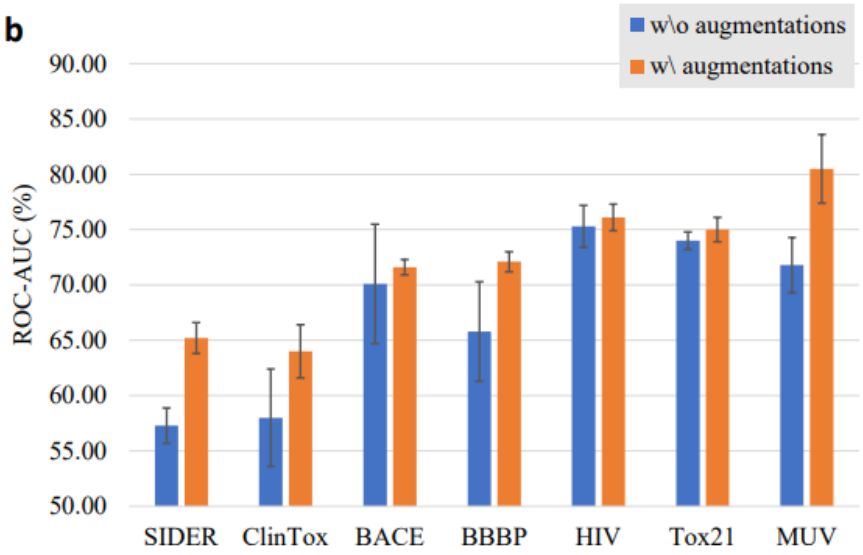
Methods

MolCLR

❖ Results

- PubChem 10M Dataset 으로 사전 학습 후, MoleculeNet 에서 7개의 Classification Task에 대해 finetuning
- Sub-graph Removal 증강 기법을 사용하였을 때, Fine-tuning 성능이 가장 좋음
  - ✓ BBBP 데이터 셋의 경우, 분자 구조의 위상 변화에 따라 성질이 크게 바뀌기 때문에 성능 하락
- 지도 학습 실험에서 제안한 데이터 증강 기법이 성능 향상에 영향을 미치는 것을 증명





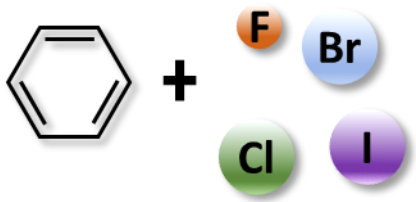
Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

Methods

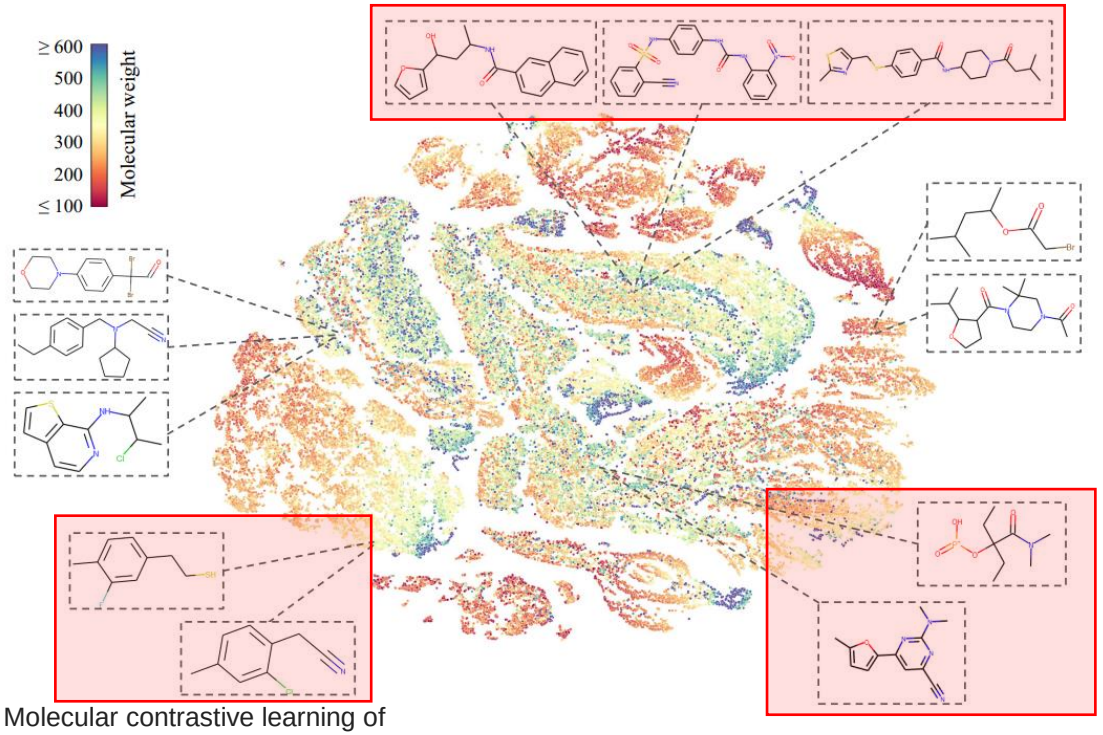
MolCLR

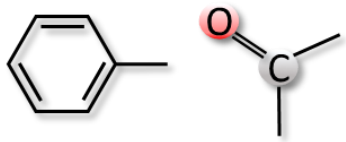
❖ Results

- MolCLR 을 통해 학습된 Representation Vector 에 대한 t-SNE Embedding
- 유사한 구조나 작용기를 가진 분자끼리 Clustering 이 되는 것을 확인

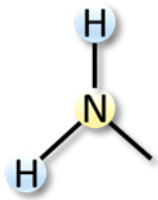


Benzene + Halogen





Arlys + Cabonyls



Amines

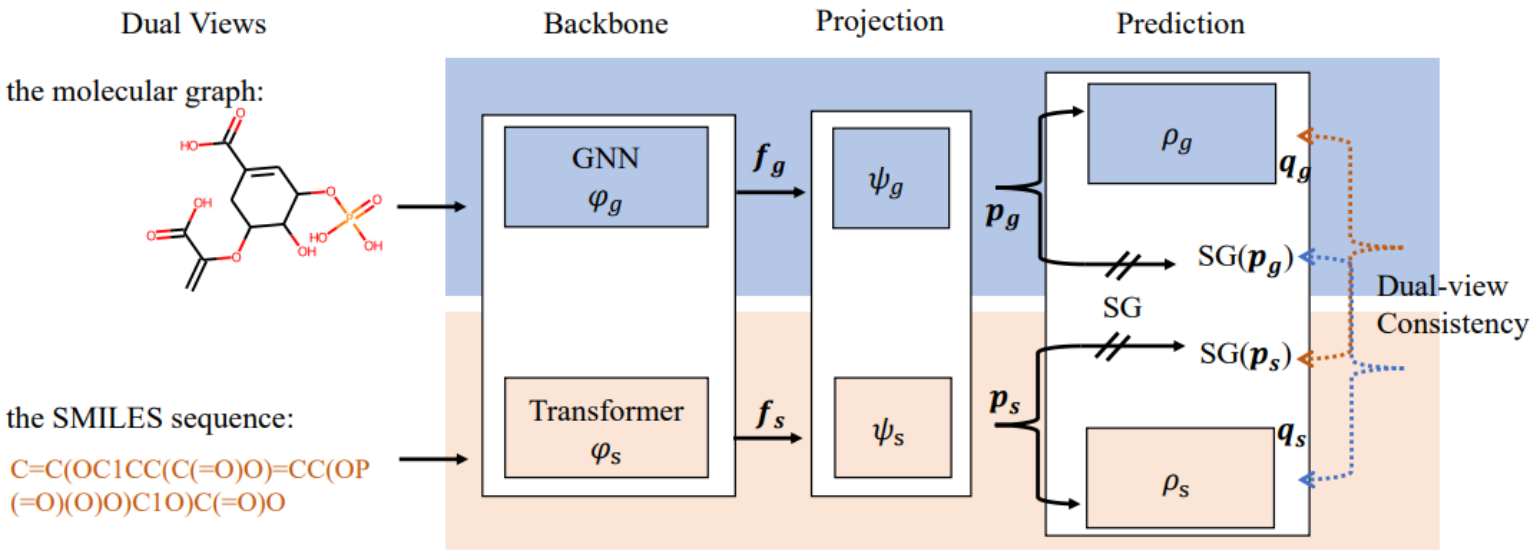
Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.

Methods

DMP

❖ Dual-view Molecule Pre-training(Zhu et al, 2022)

- University of Sci&Tech of China 와 Microsoft Research Asia 에서 연구
- 문자열과 그래프 구조가 가지는 장단점이 서로 상호보완적임을 보여줌
- BYOL 로부터 영감을 받아 서로 다른 포맷의 표현 벡터를 예측하는 사전 학습 수행



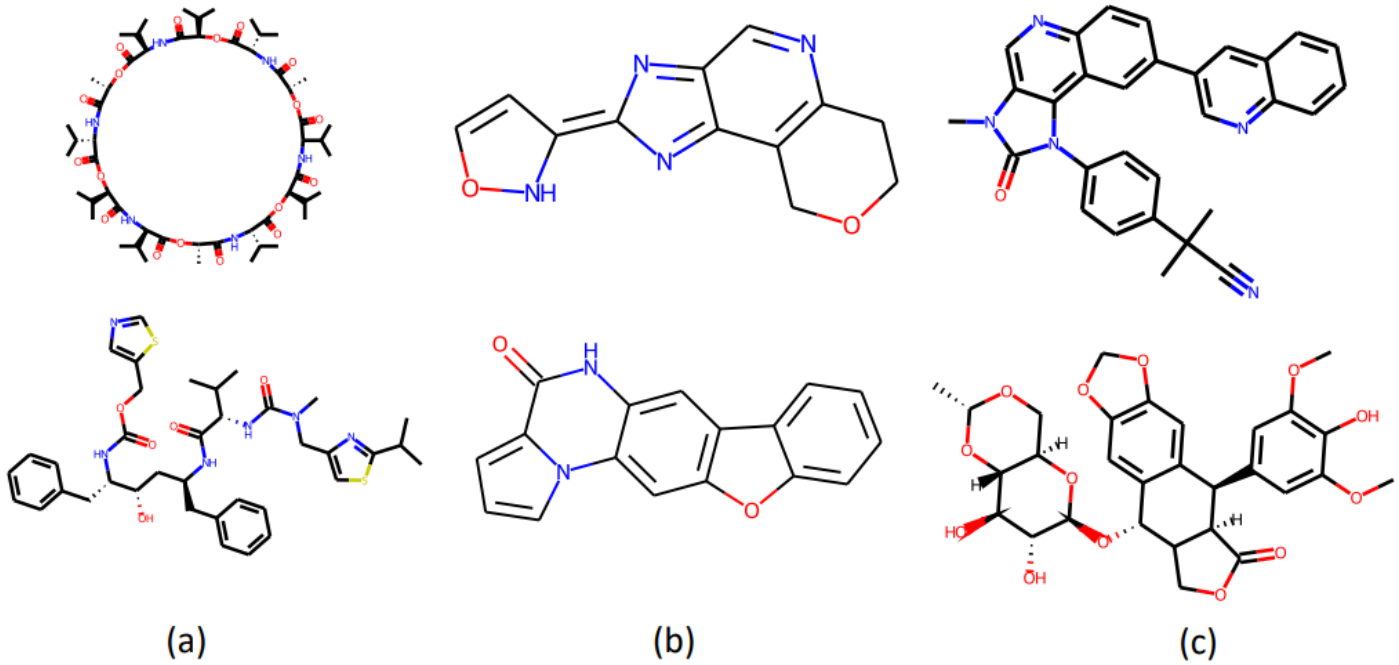
Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

Methods

DMP

❖ Pros & Cons of each representations

- Transformer 계열의 모델은 원자 사이 거리가 먼 큰 분자의 표현을 잘 학습
- GNN 계열의 모델은 여러 개의 고리가 합쳐 복잡하게 뒤얽힌 분자의 표현을 잘 학습

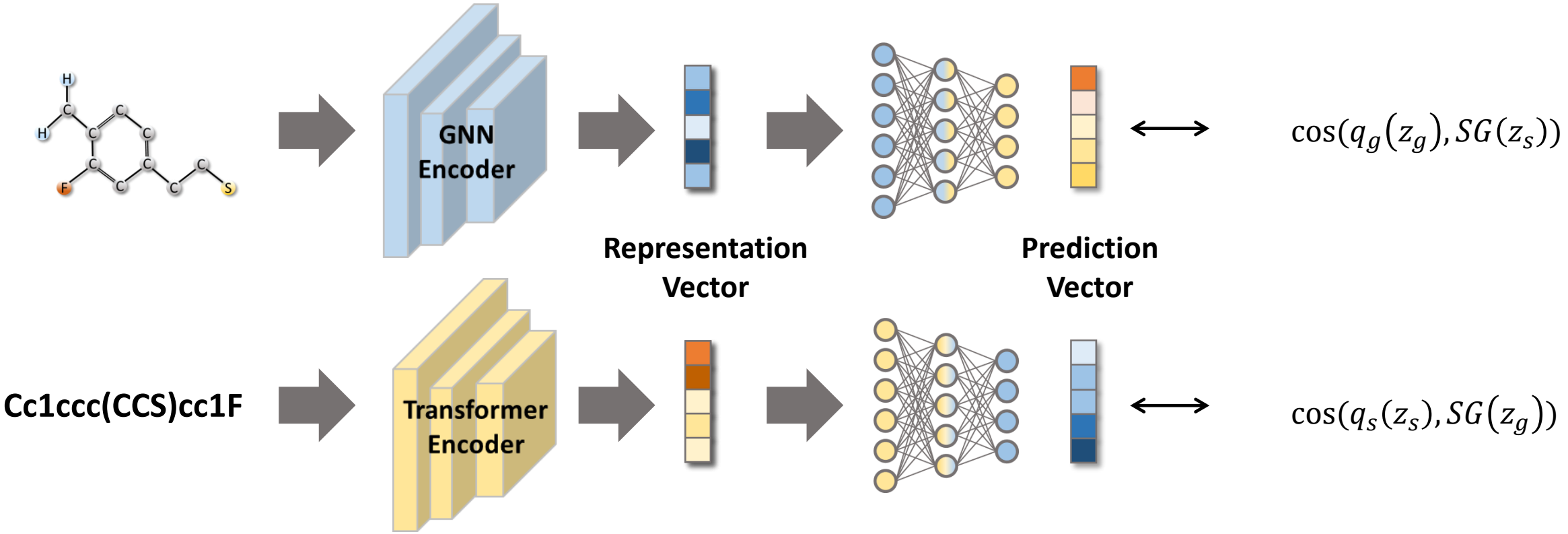


Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

Methods

DMP

❖ Pre-training Process1 (Multi-Format BYOL)



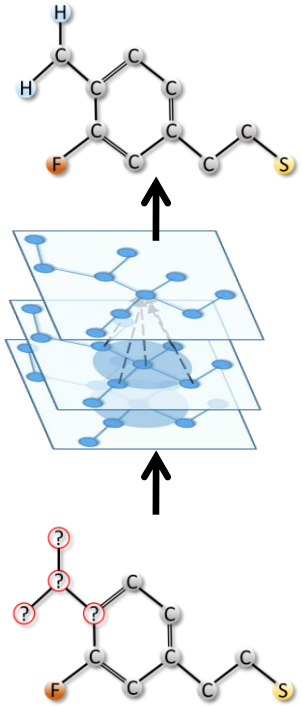
Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

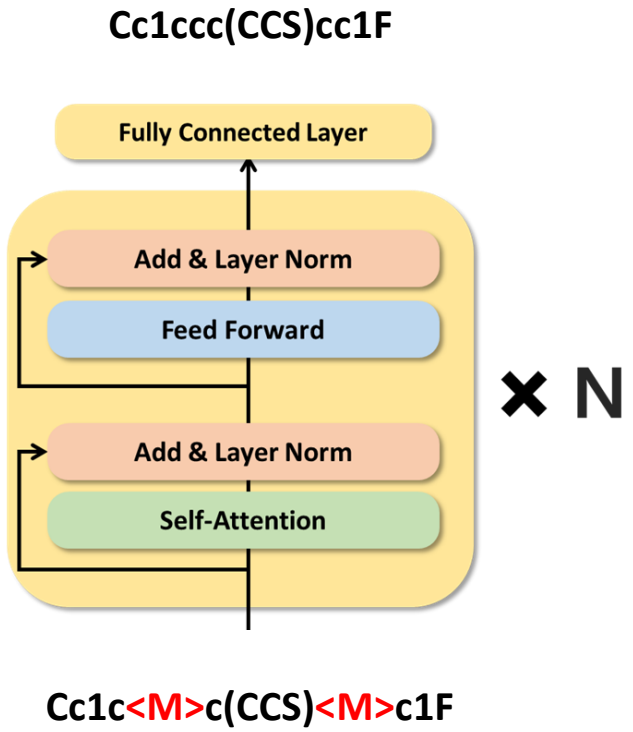
Methods

DMP

❖ Pre-training Process2 (Masked Language Model)

- GNN : Masked Atom Prediction
- Transformer : Masked Token Prediction





Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

Methods

DMP

❖ Results

- PubChem 10M Dataset 으로 사전 학습 수행 후, MoleculeNet Benchmark 에 대해 Finetuning
- DMP 로 학습한 두 architecture 중 Transformer가 GNN 계열보다 우수한 성능을 보임
- Masked Language Model Pre-training 을 추가하였을 때 더욱 우수한 성능을 보임

Dataset # Molecules	BBBP 2039	Tox21 7831	ClinTox 1478	HIV 41127	BACE 1513	SIDER 1478
RF	71.4 ± 0.0	76.9 ± 1.5	71.3 ± 5.6	78.1 ± 0.6	86.7 ± 0.8	68.4 ± 0.9
SVM	72.9 ± 0.0	81.8 ± 1.0	66.9 ± 9.2	79.2 ± 0.0	86.2 ± 0.0	68.2 ± 1.3
MGCN [36]	$\mathbf{85.0 \pm 6.4}$	70.7 ± 1.6	63.4 ± 4.2	73.8 ± 1.6	73.4 ± 3.0	55.2 ± 1.8
D-MPNN [57]	71.2 ± 3.8	68.9 ± 1.3	90.5 ± 5.3	75.0 ± 2.1	85.3 ± 5.3	63.2 ± 2.3
Hu et al. [23]	70.8 ± 1.5	78.7 ± 0.4	78.9 ± 2.4	80.2 ± 0.9	85.9 ± 0.8	65.2 ± 0.9
MolCLR [53]	73.6 ± 0.5	$\mathbf{79.8 \pm 0.7}$	93.2 ± 1.7	80.6 ± 1.1	89.0 ± 0.3	68.0 ± 1.1
TF (MLM)	74.9 ± 0.6	77.6 ± 0.4	92.9 ± 0.5	80.2 ± 0.4	88.0 ± 0.5	68.4 ± 0.4
TF ($\times 2$)	75.6 ± 0.7	77.1 ± 0.5	92.0 ± 0.8	80.4 ± 0.4	88.1 ± 0.5	68.2 ± 1.2
DMP_{TF} w/o MLM	71.1 ± 0.4	75.7 ± 0.4	93.8 ± 0.7	79.1 ± 1.7	88.3 ± 0.7	68.1 ± 0.7
DMP_{TF}	78.1 ± 0.5	78.8 ± 0.5	$\mathbf{95.0 \pm 0.5}$	$\mathbf{81.0 \pm 0.7}$	$\mathbf{89.3 \pm 0.9}$	$\mathbf{69.2 \pm 0.7}$
GNN (MLM)	74.5 ± 0.3	74.8 ± 0.5	92.3 ± 0.7	78.5 ± 0.5	84.1 ± 0.4	67.0 ± 0.5
GNN ($\times 2$)	74.1 ± 0.6	75.1 ± 0.3	92.8 ± 0.7	79.2 ± 0.9	85.1 ± 1.0	69.0 ± 0.4
DMP_{GNN}	74.7 ± 0.2	76.7 ± 0.3	94.2 ± 0.4	79.5 ± 1.0	85.7 ± 0.8	68.4 ± 0.5
TF (MLM) + GNN (MLM)	76.1 ± 0.3	77.8 ± 0.8	94.0 ± 0.4	80.1 ± 0.4	87.5 ± 0.9	69.3 ± 0.9
$\text{DMP}_{\text{TF+GNN}}$	77.8 ± 0.3	79.1 ± 0.4	95.6 ± 0.7	81.4 ± 0.4	89.4 ± 0.8	69.8 ± 0.6
DMP_{TF} (100M)	78.4 ± 0.3	79.0 ± 0.3	95.5 ± 0.2	81.1 ± 0.3	89.6 ± 0.3	70.0 ± 0.6
DMP_{GNN} (100M)	75.2 ± 0.6	77.5 ± 0.6	94.7 ± 0.4	80.3 ± 0.5	86.3 ± 1.0	69.2 ± 0.5

Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

Methods

DMP

❖ Results

- GNN Only/Transformer Only Pre-training 과 제안 방법론의 t-SNE embedding
- 같은 골격 구조(Scaffold)를 가진 분자끼리 Clustering 이 더 잘되는 것을 확인

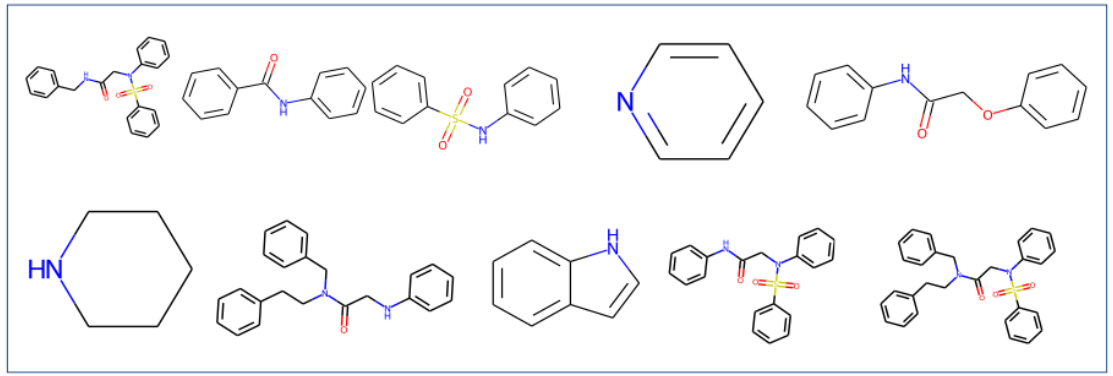
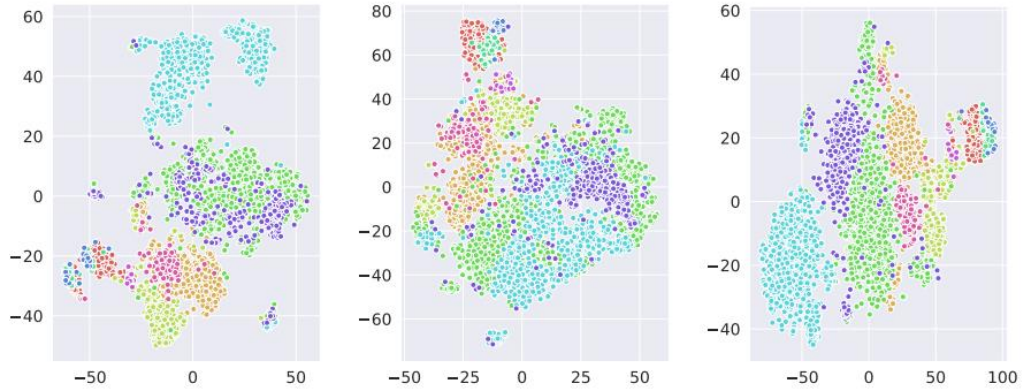


Figure 4: Ten scaffolds used in our visualization.



(a) GNN: DB Index 3.56 (b) Transformer: DB Index 3.59 (c) DMP: DB Index 2.19

Figure 3: Visualization of the representations learned by different models. Different colors indicate different scaffolds.

Zhu, J., Xia, Y., Qin, T., Zhou, W., Li, H., & Liu, T. Y. (2021). Dual-view molecule pre-training. arXiv preprint arXiv:2106.10234.

Conclusion

Summary

❖ Comments

- 분자의 물성 예측을 위한 벤치마크(MoleculeNet)이 등장함에 따라 Deep Learning 을 적용한 다양한 연구가 진행 중

- 지도 학습을 위한 양질의 labeled data 는 구하기 어렵기 때문에, PubChem 등의 Unlabeled Data 를 통해 분자의 Representation 을 Pre-training 하는 방법론이 등장

- 고전적 분자 표현 방식(i.e. ECFP) 보다 Unsupervised Learning 해서 추출한 learned representation 이 더 좋은 성능을 보임

- 화학 분자의 입력 형태에 따른 성능은 우위가 없으며, 장단점이 존재하며 서로 상호보완적

Appendix

Additional Materials

[1] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2019). Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

[2] Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2019, September). Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (pp. 429-436).

[3] Honda, S., Shi, S., & Ueda, H. R. (2019). Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*.

[4] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, *33*, 12559-12571.

[5] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, *59*(8), 3370-3388.