DMQA Open Seminar

Self/Semi-Supervised Learning for Tabular Data

2022.10.14

Byeongeun Ko

Data Mining & Quality Analytics Lab.

발표자 소개

❖ 고병은 (Byeongeun Ko)

- 고려대학교 산업경영공학과 대학원 재학 중
- Data Mining & Quality Analytics Lab (김성범 교수님 연구실)
- M.S. Student (2022.03 ~)

❖ Research Interest

- Anomaly Detection
- Self-supervised Learning

❖ E-mail

- byeongeun_ko@korea.ac.kr

Contents

Introduction

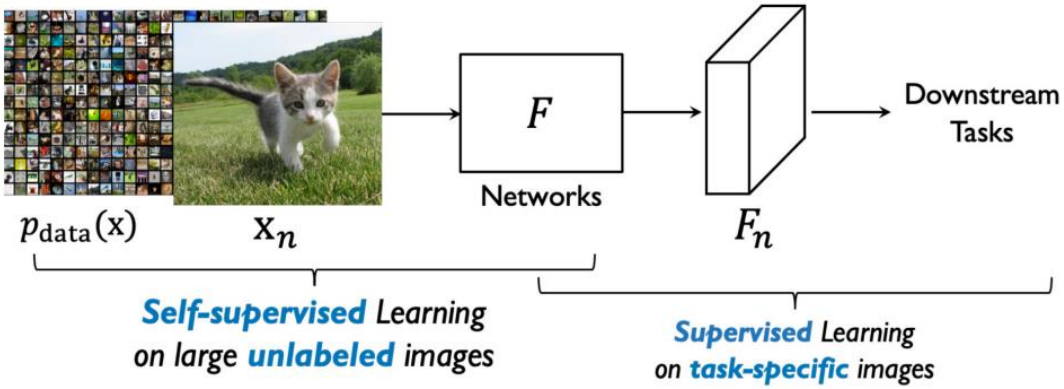
Introduction

❖ Self/Semi-Supervised Learning? - Definition

- Self-Supervised Learning
  - ➢ 외부에서 레이블을 주입 받지 않고 독립변수에 임베딩된 레이블을 사용하여 모델을 학습[1)]
  - ➢ 데이터에서 스스로 레이블을 생성하고 지도 학습 기법으로 레이블된 데이터셋에서 모델을 훈련[2)]

- Semi-Supervised Learning
  - ➢ 레이블이 있는 데이터와 레이블이 없는 데이터가 훈련집합을 구성하는 상황에서 이루어지는 학습[3)]

4)

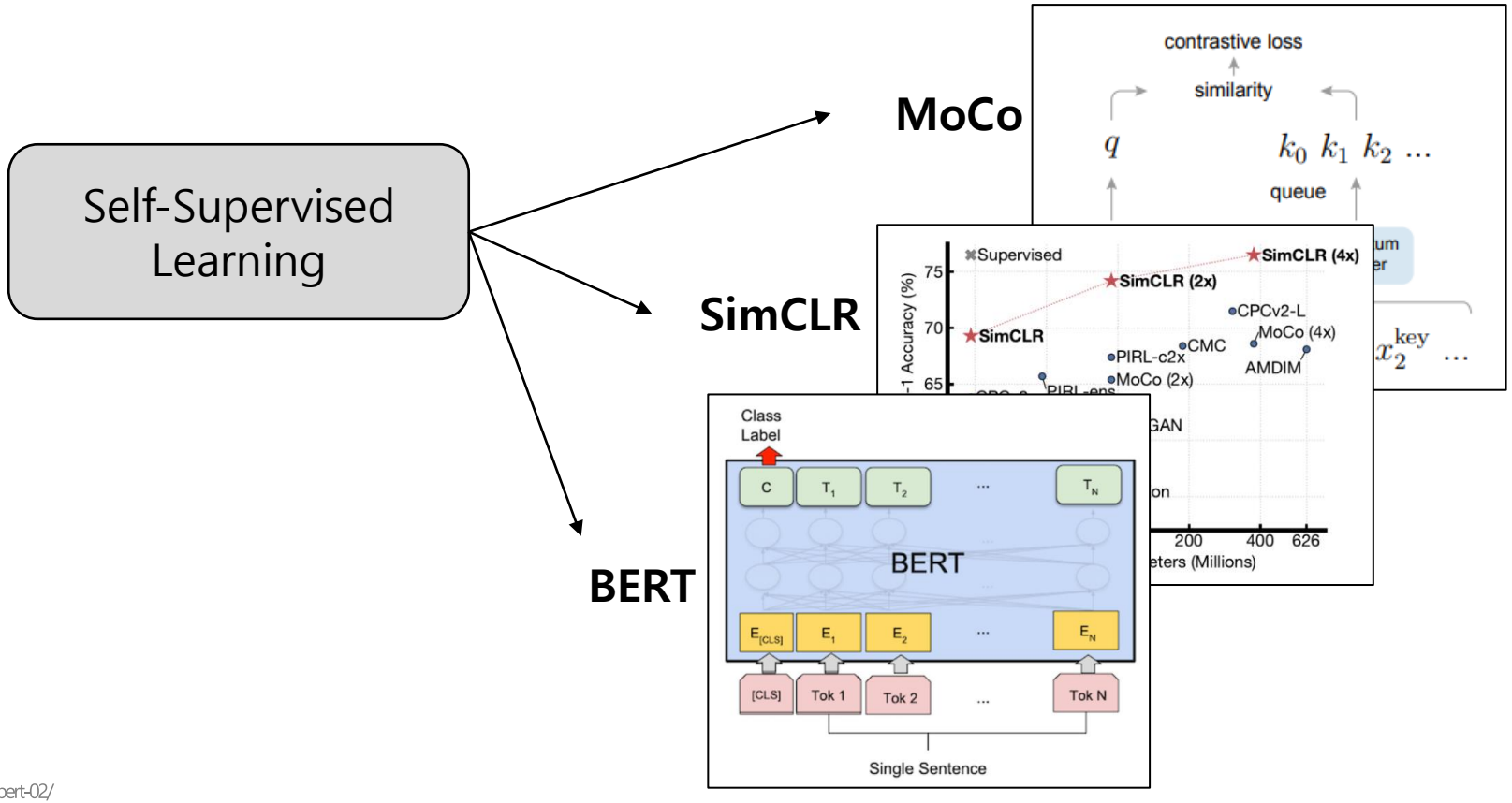


1) Jeremy Howard, Sylvain Gugger, Deep Learning for Coders with fastai & PyTorch, Hanbit Media
2) Aurelien Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFLow, Hanbit Media
3) Il Seok Oh, Machine Learning, Hanbit Academy
4) Kim, S. R. (2022). 최신 자가 학습 기반의 인공지능 기술 동향. Broadcasting and Media Magazine, 27(2), 19-25.

Introduction

❖ **Self/Semi-Supervised Learning?**

- 최근 굉장히 활발하게 연구되고 있으며 높은 성능을 보임



https://paul-hyun.github.io/bert-02/

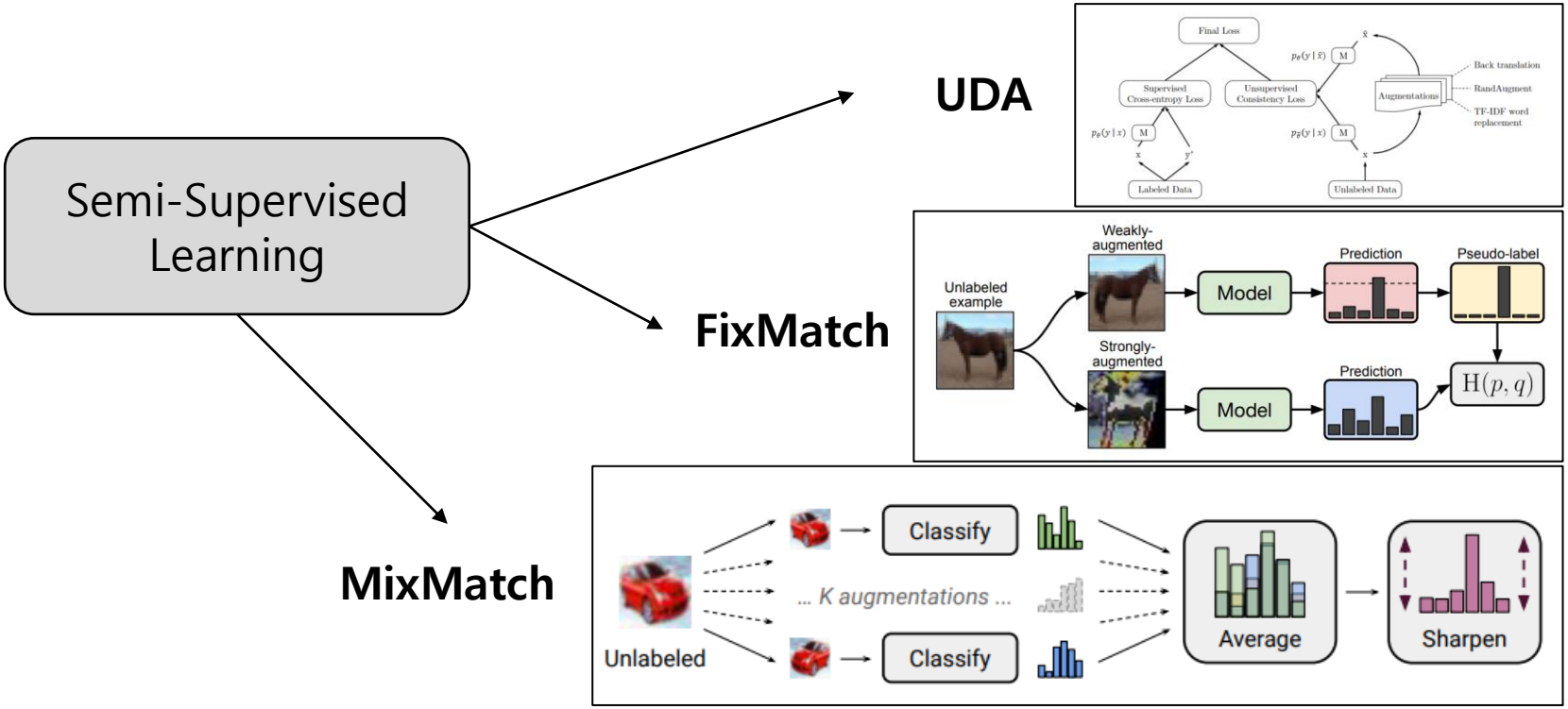
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

Introduction

❖ **Self/Semi-Supervised Learning?**

- 최근 굉장히 활발하게 연구되고 있으며 높은 성능을 보임



Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33, 6256-6268.
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33, 596-608.
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32.

Introduction

❖ Self/Semi-Supervised Learning?

- Self/Semi-Supervised Learning이 주목받는 이유
  - ➢ 수많은 데이터가 생성되어 대용량 데이터를 모으는 일은 어렵지 않으나, 레이블된 데이터를 구하는 비용이 매우 높음

→ 사과

EASY

→ 블루베리

작가 Racool_studio 출처 Freepik

Introduction

❖ **Self/Semi-Supervised Learning?**

- Self/Semi-Supervised Learning이 주목받는 이유
  - ➢ 수많은 데이터가 생성되어 대용량 데이터를 모으는 일은 어렵지 않으나,

    레이블된 데이터를 구하는 비용이 매우 높음

→ 정상

→ 비정상

Difficult & Expensive

작가 kstudio 출처 Freepik

Introduction

❖ Self/Semi-Supervised Learning? – Flow

- ➢ Semi-Supervised Learning
  - Pseudo Labeling - 새로운 라벨을 생성, Hybrid Methods - 여러 아이디어를 혼합하여 사용
    - Self-Training (2007), MixMatch (2019), …

- ➢ Self-Supervised Learning
  - Pretext Task - 사용자가 새로운 문제를 정의하고 학습함으로써 데이터에 대한 이해 ↑
    - Exemplar (2014), …
  - Contrastive Learning - Positive/Negative Pair를 구성하여 유사한 것은 가깝게 상이한 것은 멀게 학습
    - SimCLR (2020), MoCo (2020), …

Zhu, X. (2007, June). Semi-supervised learning tutorial. In International conference on machine learning (ICML) (pp. 1-135).
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems (pp. 5049-5059).
Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems, 27.
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
He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738)..

Introduction

❖ Self/Semi-Supervised Learning?

종료

Self-Supervised Representation Learning

Seokho Moon

May 1, 2020

Self-Supervised Representation Learning

발표자: 문석호

2020년 5월 1일

오후 1시 ~

화상 프로그램 이용(Zoom)

세미나 정보 보기 ⟶

종료

elf-Supervised Learning

(Algorithm & application)

Seokho Moon

Nov 20, 2020

Self-Supervised Learning (algorithm & ap

발표자: 문석호

2020년 11월 20일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

종료

A Holistic Approach to Semi-Supervised Learning

Semi-supervised learning in deep neural

발표자: 이민정

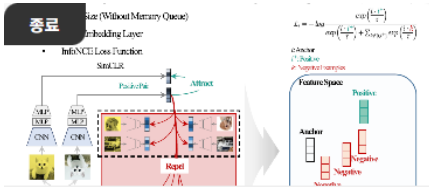
2020년 12월 4일

오후 1시 ~

온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

종료



Applications of Self-Supervised Learning

발표자: 이영재

2022년 3월 4일

오전 3시 ~

온라인 비디오 시청(YouTube)

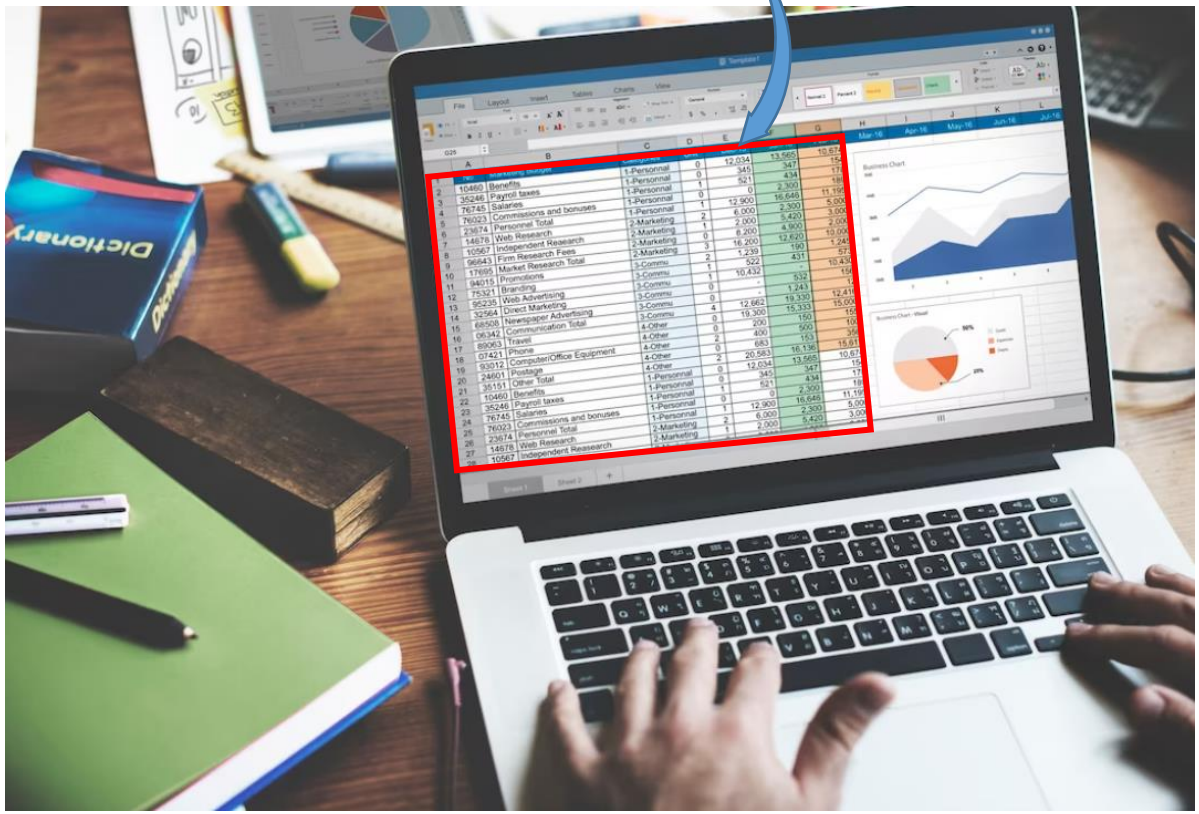
온라인 비디오 시청 (YouTube)

세미나 정보 보기 ⟶

Introduction

❖ Tabular Data

- 2차원 형태로 구성된 (행, 열) 정형 데이터

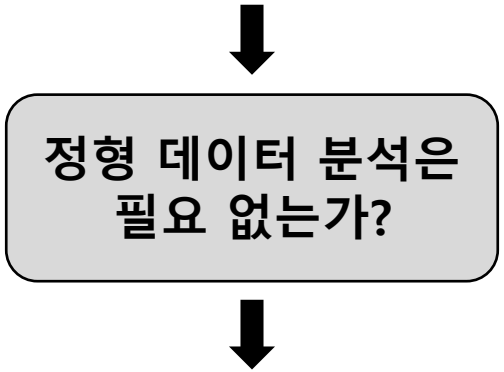


작가 rawpixel.com 출처 Freepik

Introduction

❖ Tabular Data

- TIBCO Software Inc 曰 (데이터 분석 플랫폼 제공 등)
  - ➢ 정형 데이터는 현재 조직에서 사용하는 데이터 유형의 20% 정도이며 그 비율은 떨어지고 있다
  - ➢ 빠른 속도로 증가하는 비정형 및 반정형 데이터의 엄청난 증가로 정형 데이터의 점유율이 감소



  - ➢ 현재 정형 데이터는 비즈니스에 대한 예측이 점점 더 강조되면서 여전히 가치 있다.
  - ➢ 정형 데이터는 비정형 데이터보다 훨씬 더 쉽게 액세스할 수 있기 때문에 가치가 있다.

산업현장(ex. 제조) 에서는 여전히 분석의 수요가 매우 높다

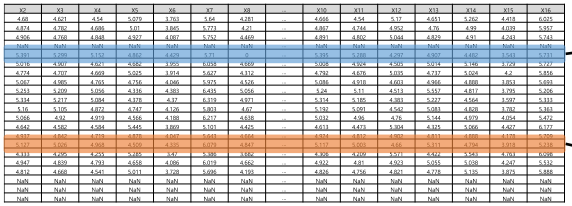
https://www.tibco.com/ko/reference-center/what-is-structured-data

Introduction

❖ Tabular Data

- 정형 데이터도 여전히 레이블링에 소요되는 비용은 높다.

사과 **비전문가 레이블링 가능**



정상

불량

??? **전문가도 힘듦**

추가 시각화, Rule, … 필요

정형 데이터에 Self/Semi-Supervised Learning의 적용 필요

작가 Racool_studio 출처 Freepik

Introduction

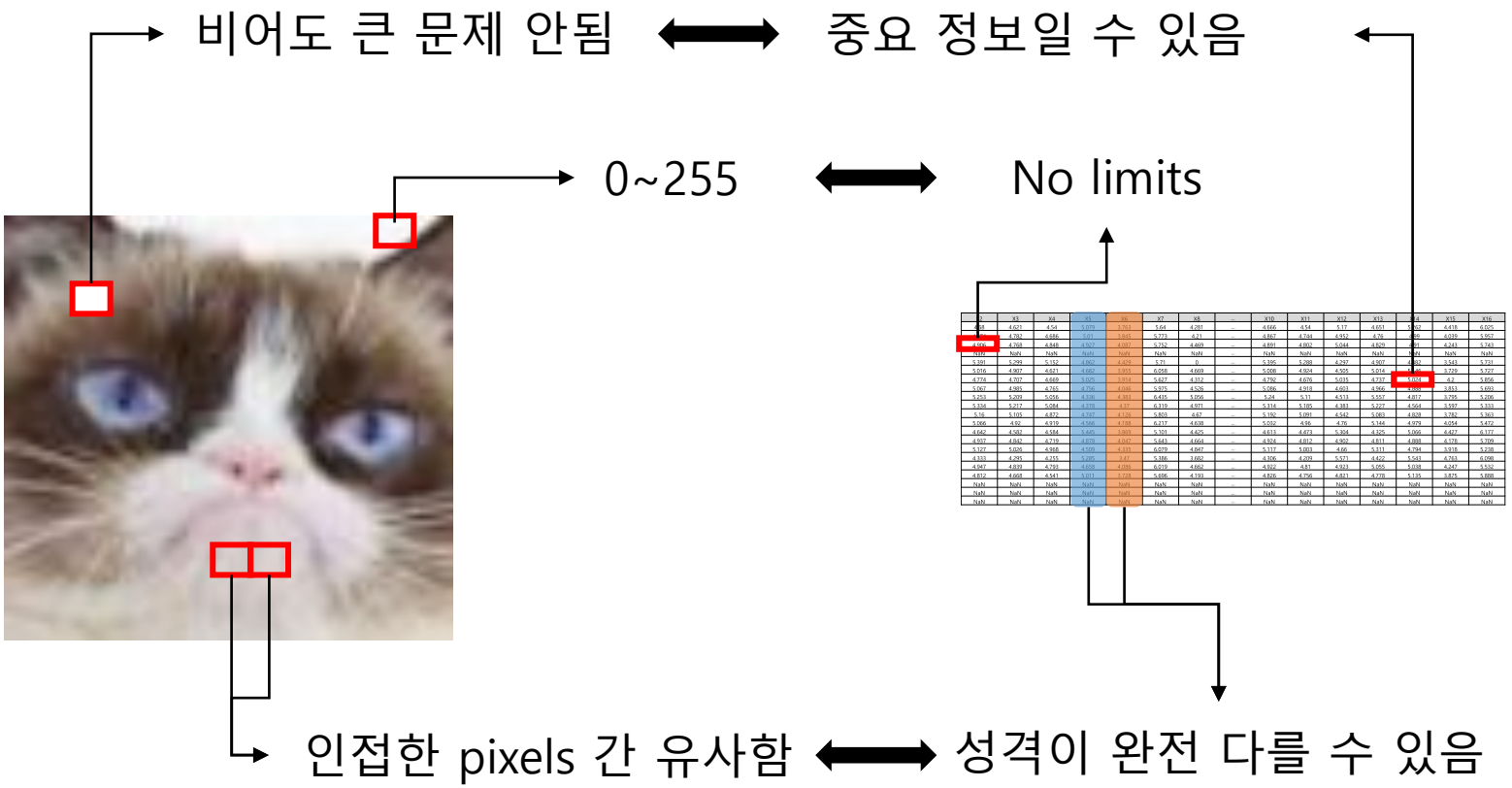
❖ **Why does it perform poorly on tabular data?**

- 적절하지 않은 데이터
  - ➢ Missing Value – 이미지는 딱히 결측치가 없지만 정형 데이터에선 Critical 함
  - ➢ Outlier가 매우 극단적
  - ➢ 주로 클래스 불균형이 심한 문제들
- 특징(feature, 열)간의 관계
  - ➢ 특징 간의 관계가 매우 큰 경우도 있고 전혀 없는 경우도 있고… 불규칙
- 전처리가 어려움
  - ➢ 정형 데이터는 범주형 데이터도 포함 (굉장히 Sparse한 데이터를 만듦 → 차원의 저주)
  - ➢ 기존 데이터 증강 기법이 정형 데이터에 적용하기 어려움 (Rotation, Crop ?!)
- 많은 Self/Semi 방법론들은 비정형 데이터(Image, …) 대상이며 지나치게 해당 구조에 의존적

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.

Introduction

❖ **Why does it perform poorly on tabular data?**



Introduction

❖ **Why does it perform poorly on tabular data?**

- 원-핫 인코딩(one-hot encoding)을 수행하면 굉장히 sparse한 데이터 생성
  - 차원의 저주 (curse of dimensionality)

City
Seoul
Busan
Masan
…
Suwon

→

Seoul	Busan	Masan	…	Suwon
1	0	0	…	0
0	1	0	…	0
0	0	1	…	0
…	…	…	…	…
0	0	0	…	1

Introduction

❖ Different Approaches

- 그렇다면 정형 데이터의 Self/Semi-Supervised Learning은 어떻게 할 것인가?

➢ 정형 데이터를 이미지화

  ▪ SuperTML(2019), IGTD(2021)

➢ 어텐션 메커니즘 사용

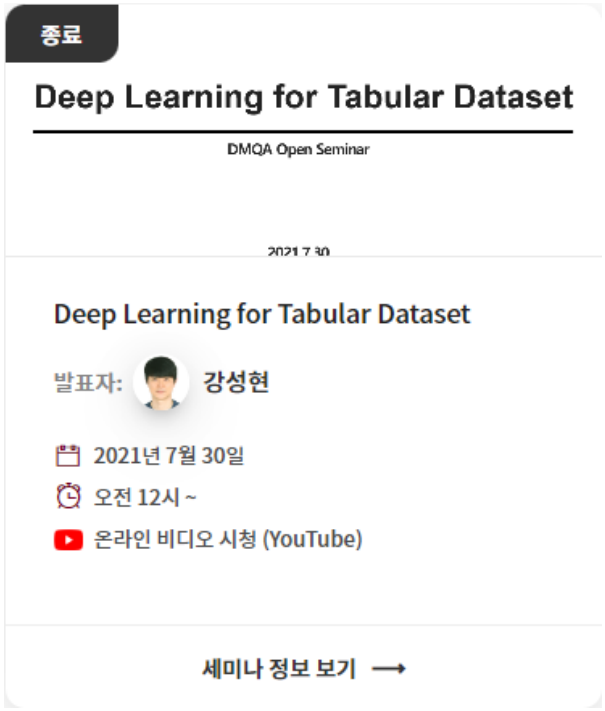
  ▪ TabNet(2019)

➢ 오토인코더 사용 Contextual Embedding

  ▪ VIME(2020), SubTab(2021)

➢ 대조 학습

  ▪ SCARF(2021), Contrastive Mixup(2021)



Methods

Methods

❖ **VIME:** Extending the Success of Self- and Semi-supervised Learning to Tabular Domain

- NeurIPS (2020), 2022년 10월 5일 기준 48회 인용
- 자기지도학습과 준지도학습 프레임워크 제안

VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain

Jinsung Yoon
Google Cloud AI, UCLA
jinsungyoon@google.com

Yao Zhang
University of Cambridge
yz555@cam.ac.uk

James Jordon
University of Oxford
james.jordon@wolfson.ox.ac.uk

Mihaela van der Schaar
University of Cambridge
UCLA, Alan Turing Institute
mv472@cam.ac.uk

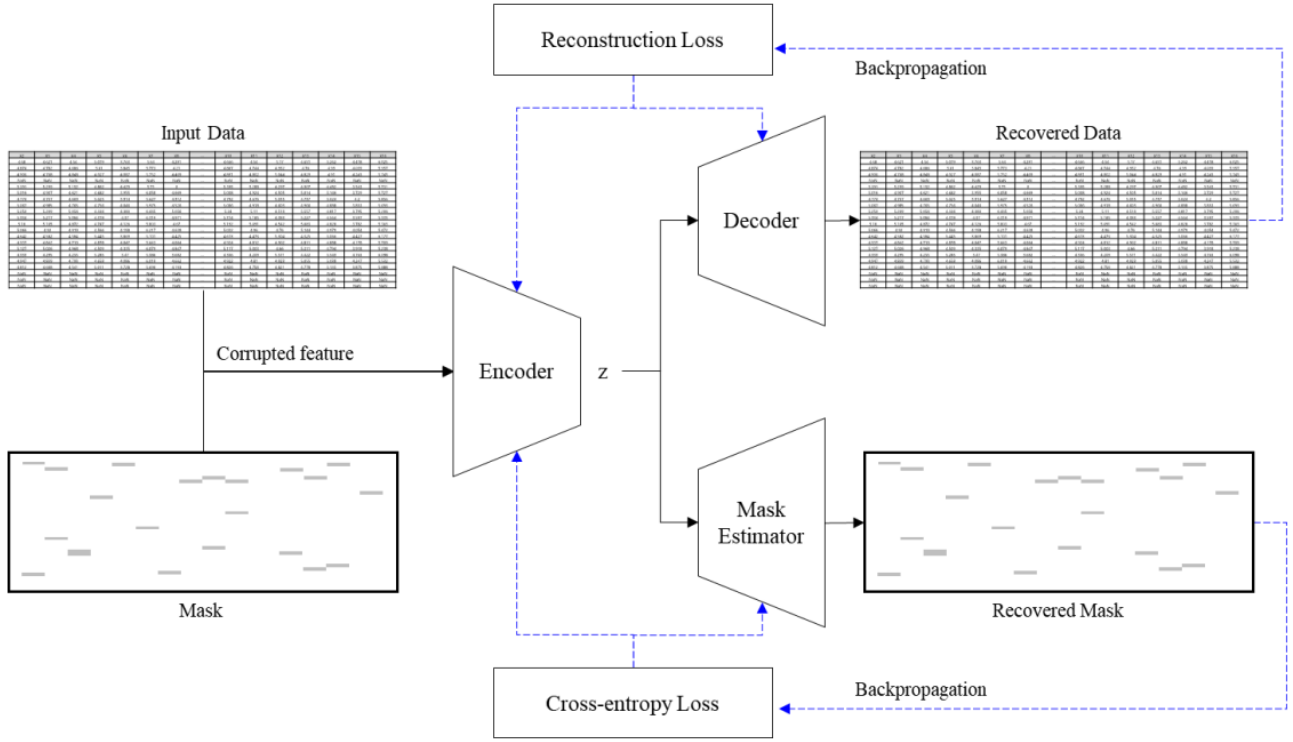
Abstract

Self- and semi-supervised learning frameworks have made significant progress in training machine learning models with limited labeled data in image and language domains. These methods heavily rely on the unique structure in the domain datasets (such as spatial relationships in images or semantic relationships in language). They are not adaptable to general tabular data which does not have the same explicit structure as image and language data. In this paper, we fill this gap by proposing novel self- and semi-supervised learning frameworks for tabular data, which we refer to collectively as VIME (Value Imputation and Mask Estimation). We create a novel pretext task of estimating mask vectors from corrupted tabular data in addition to the reconstruction pretext task for self-supervised learning. We also introduce a novel tabular data augmentation method for self- and semi-supervised learning frameworks. In experiments, we evaluate the proposed framework in multiple tabular datasets from various application domains, such as genomics and clinical data. VIME exceeds state-of-the-art performance in comparison to the existing baseline methods.

Methods

❖ VIME - Self

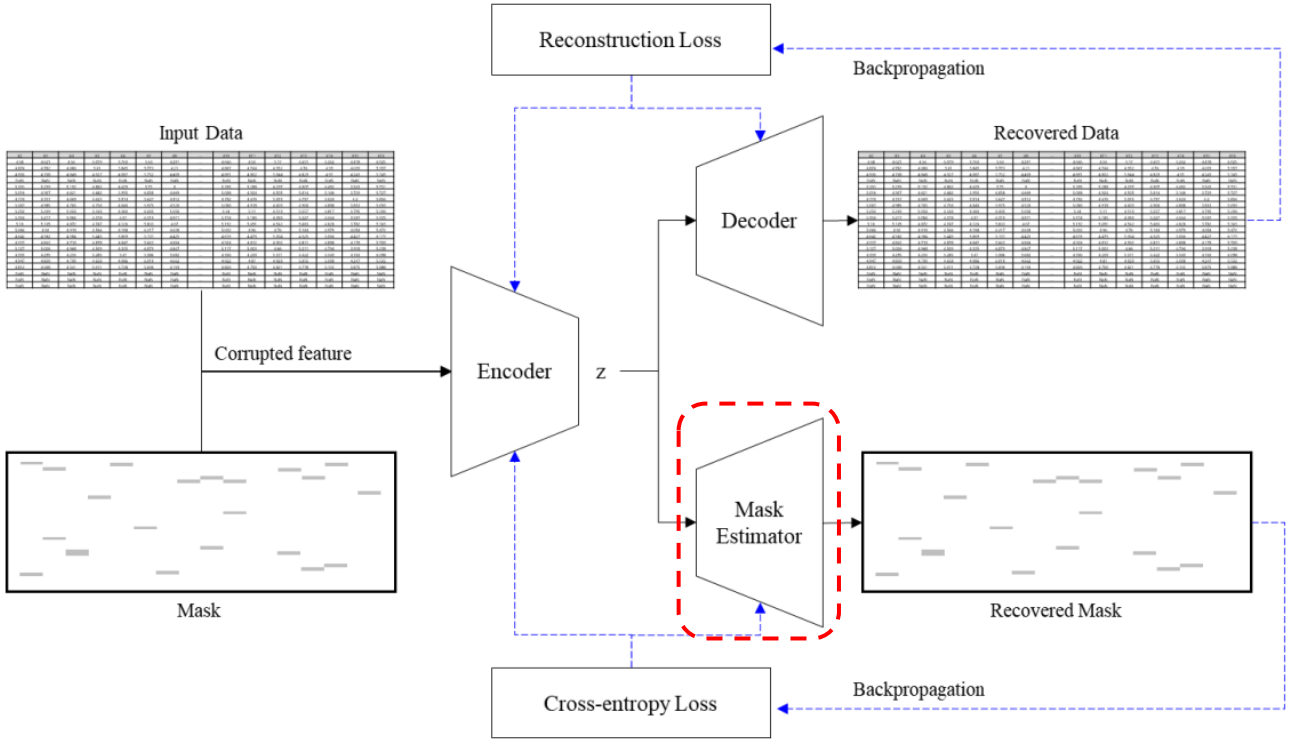
- Network : Encoder, Decoder, Mask Estimator
- Losses : Reconstruction Loss, Cross-entropy Loss



Methods

❖ VIME - Self

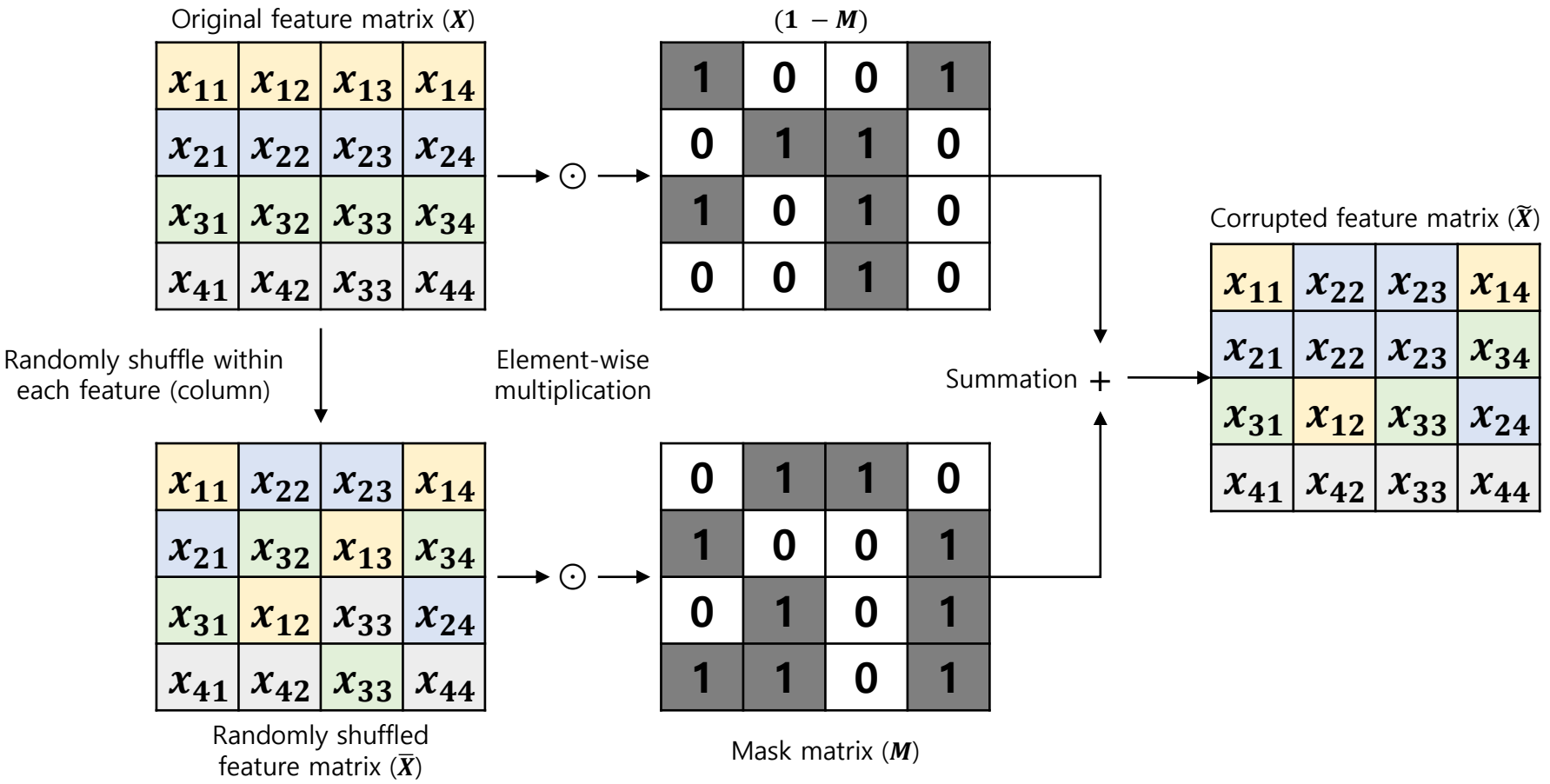
- Mask estimator는 주어진 데이터에 변형을 가한 후 어떤 데이터가 변형이 되었는지를 찾음
  - ➢ 이를 통해 이미지나 자연어에 비하여 약한 feature들 간의 상관관계를 잘 찾음



Methods

❖ VIME - Self

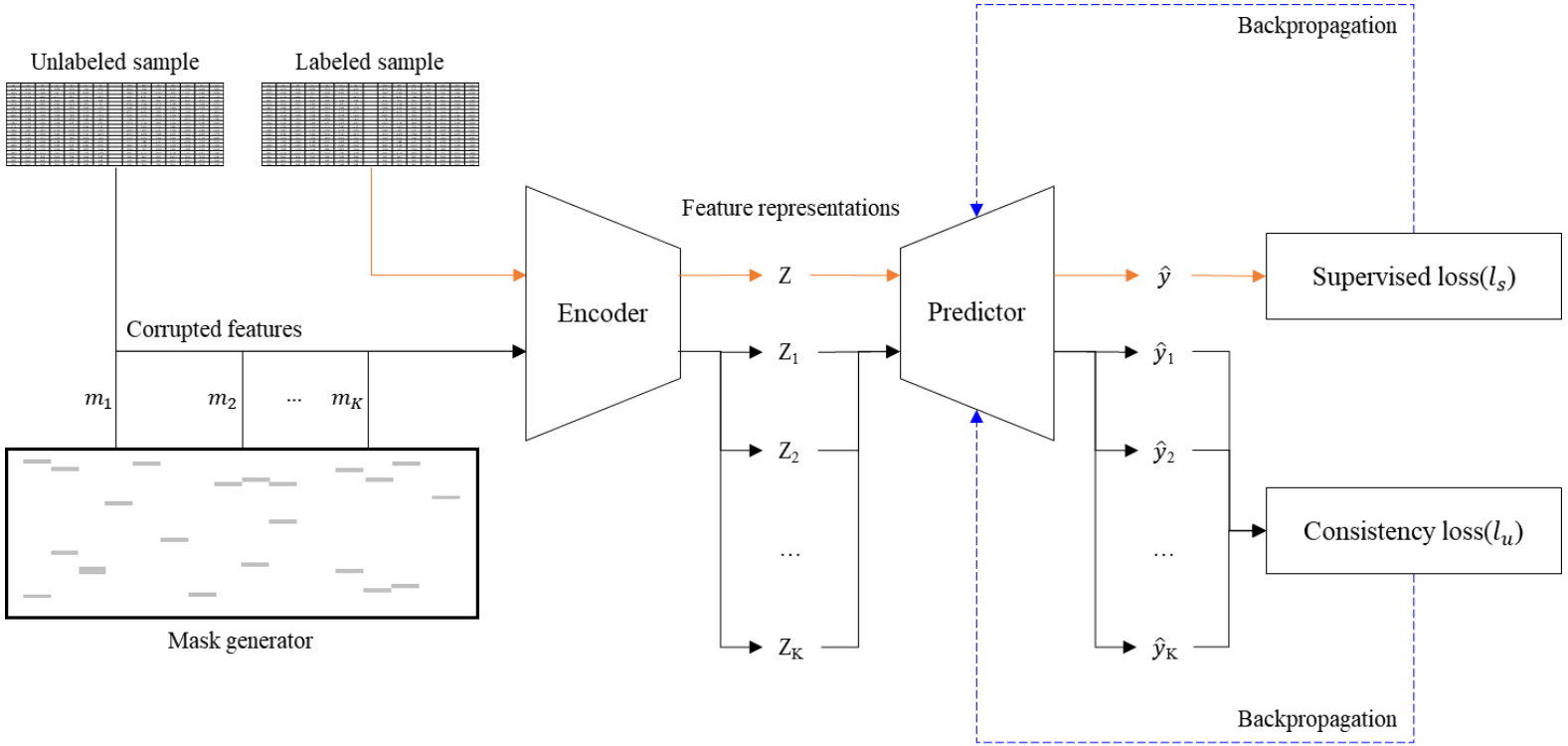
- Noise strategy : Column-wise swap noise → 정형 데이터에 적합한 노이즈 방식



Methods

❖ VIME - Semi

- VIME-Self에서 학습했던 Encoder를 재사용 (고정)
- 레이블이 없는 데이터는 Mask generator를 활용하여 K개의 증강된 데이터를 만듦
  - ➢ 증강된 데이터는 같음 sample에서 유래되었으므로 같은 레이블을 가져야함 → Consistency loss



Methods

❖ VIME

- 실험 결과

Table 1: AUROC Performances of patient treatment predictions on Hormone and Radical therapy (higher the better). (Mean ± Standard deviations are computed over 10 runs)

Type	Models	Hormone	Radical
SL	Logistic Regression	.8371±.0013	.8036±.0015
	2-layer Perceptron	.8351±.0023	.8146±.0022
	XGBoost	.8423±.0018	.8166±.0011
Self-SL	DAE	.8335±.0049	.8144±.0061
	Context Encoder	.8308±.0051	.8134±.0066
Semi-SL	MixUp	.8448±.0021	.8214±.0029
VIME		**.8602±.0029**	**.8391±.0021**

Table 2: Prediction accuracy of the methods on UCI Income, MNIST and UCI Blog datasets (Mean ± Std are computed over 10 runs).

Type	Models	Income	MNIST	Blog
Supervised models	Logistic Regression	.8425±.0013	.8989±.0023	.6915±.0029
	2-layer Perceptron	.8520±.0023	.9387±.0014	.7972±.0058
	XGBoost	.8623±.0021	.9413±.0026	.7975±.0030
Self-supervised models	DAE	.8578±.0028	.9431±.0032	.8001±.0039
	Context Encoder	.8611±.0027	.9455±.0048	.8033±.0051
Semi-supervised models	MixUp	.8701±.0021	.9461±.0023	.8088±.0038
Variants of VIME	Supervised only	.8520±.0023	.9387±.0014	.7972±.0058
	Self-SL only	.8599±.0026	.9406±.0019	.8147±.0037
	Semi-SL only	.8771±.0031	.9548±.0023	.8361±.0041
VIME		**.8804±.0030**	**.9577±.0022**	**.8389±.0044**

Methods

❖ **SubTab: Sub**setting Features of **Tab**ular Data for Self-Supervised Representation Learning

- NeurIPS (2021), 2022년 10월 5일 기준 12회 인용
- 비교적 간단한 구조를 활용하여 높은 성능을 달성

SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation Learning

Talip Uçar, Ehsan Hajiramezanali, Lindsay Edwards

Respiratory and Immunology, R&D, AstraZeneca
{talip.ucar, ehsan.hajiramezanali, lindsay.edwards}@astrazeneca.com

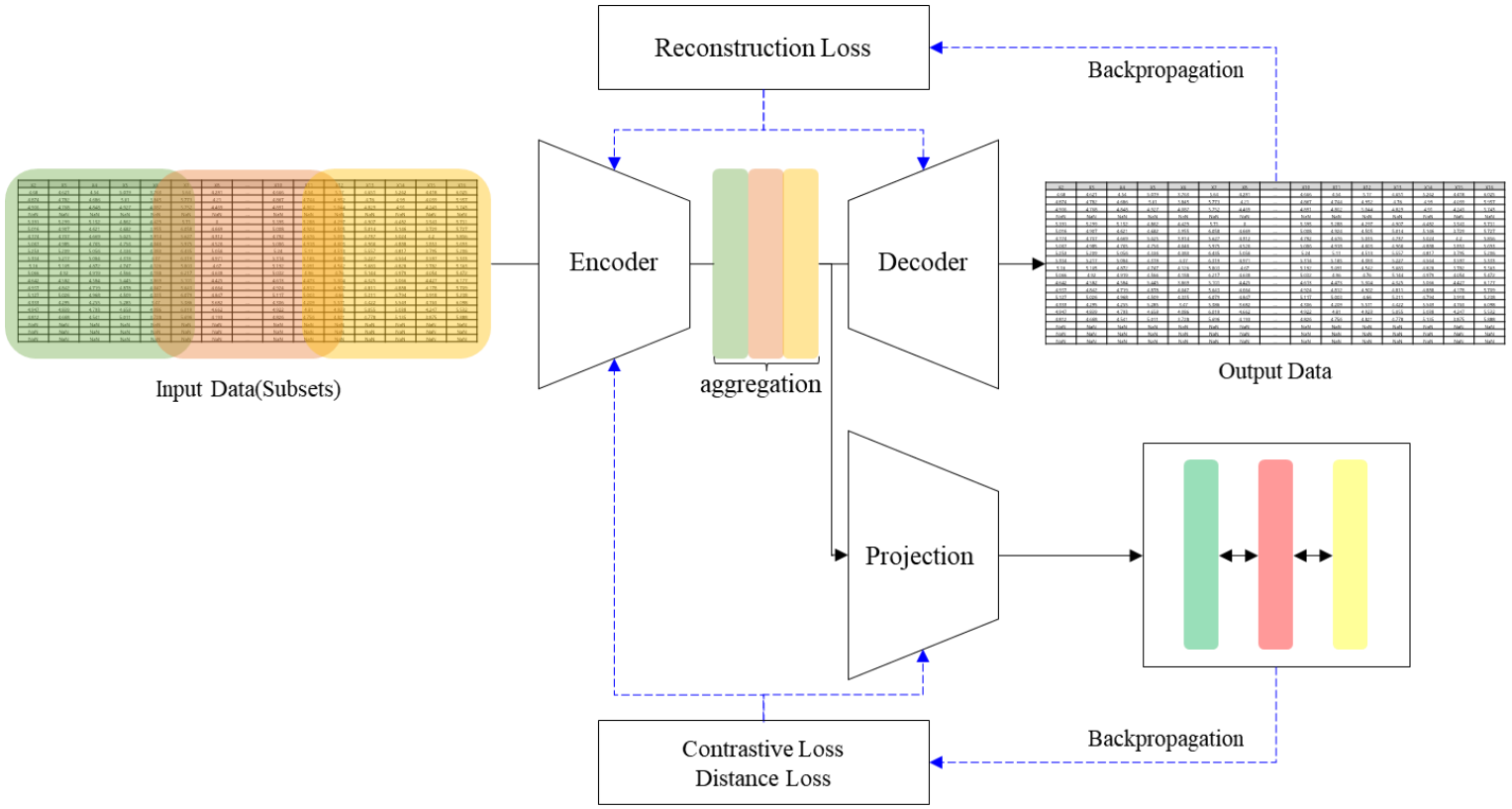
Abstract

Self-supervised learning has been shown to be very effective in learning useful representations, and yet much of the success is achieved in data types such as images, audio, and text. The success is mainly enabled by taking advantage of spatial, temporal, or semantic structure in the data through augmentation. However, such structure may not exist in tabular datasets commonly used in fields such as healthcare, making it difficult to design an effective augmentation method, and hindering a similar progress in tabular data setting. In this paper, we introduce a new framework, **Sub**setting features of **Tab**ular data (SubTab), that turns the task of learning from tabular data into a multi-view representation learning problem by dividing the input features to multiple subsets. We argue that reconstructing the data from the subset of its features rather than its corrupted version in an autoencoder setting can better capture its underlying latent representation. In this framework, the joint representation can be expressed as the aggregate of latent variables of the subsets at test time, which we refer to as *collaborative inference*. Our experiments show that the SubTab achieves the state of the art (SOTA) performance of 98.31% on MNIST in tabular setting, on par with CNN-based SOTA models, and surpasses existing baselines on three other real-world datasets by a significant margin.

Methods

❖ SubTab

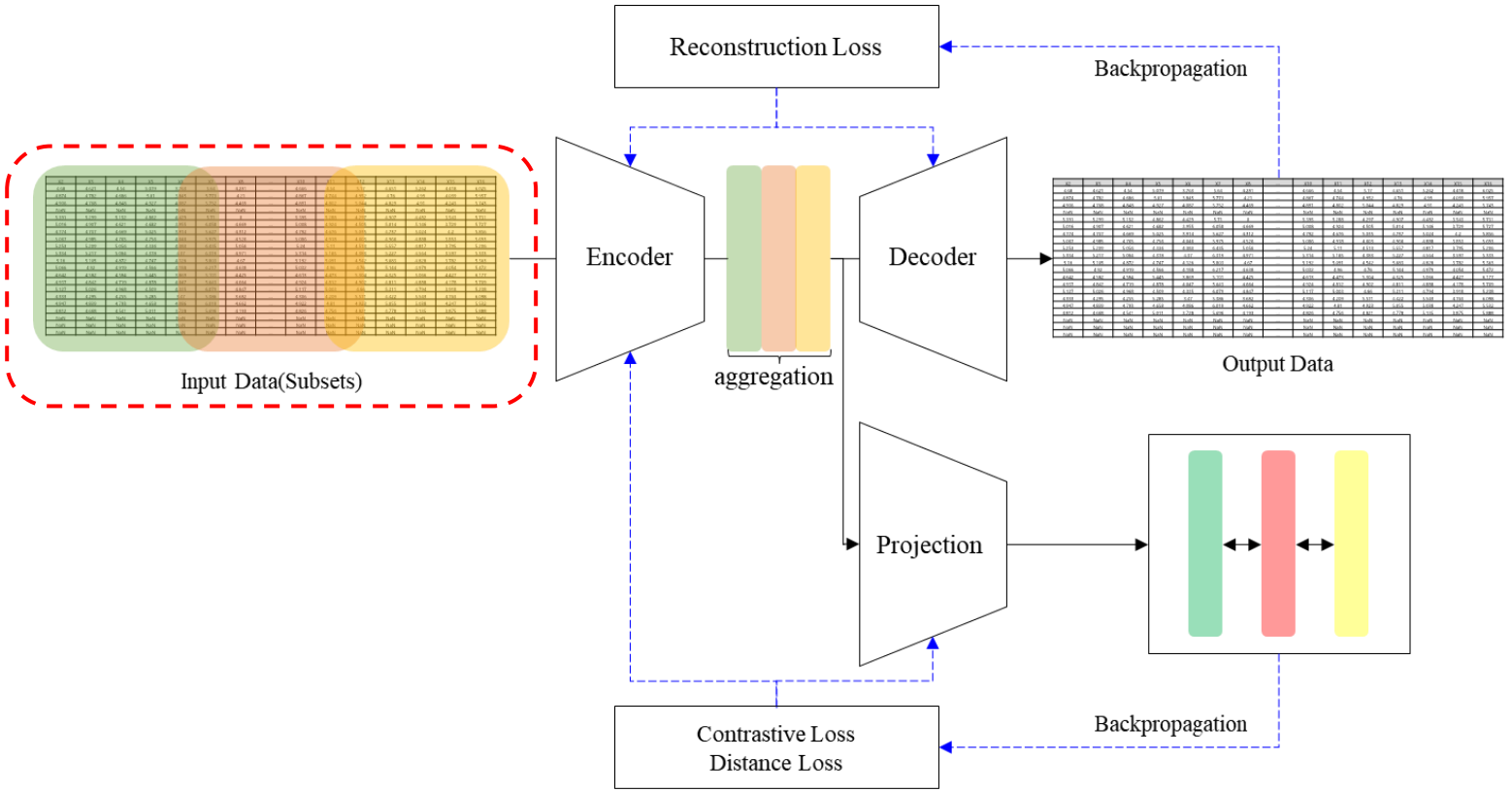
- Network : Encoder, Decoder, Projection
- Losses : Reconstruction Loss, Contrastive Loss, Distance Loss



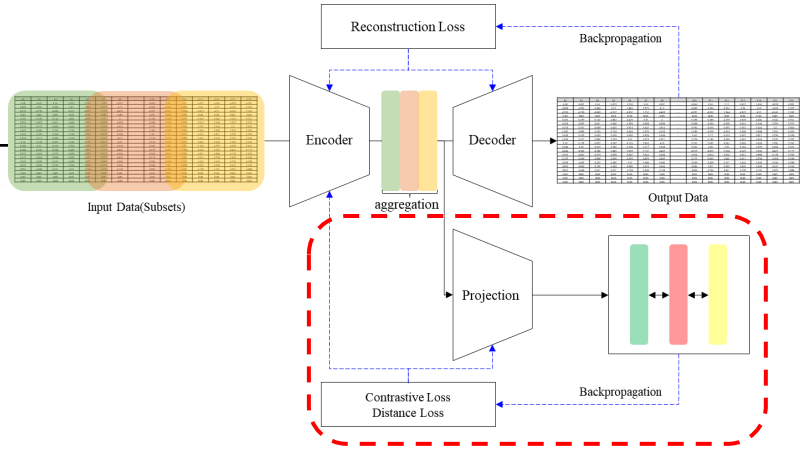
Methods

❖ SubTab

- 입력 데이터를 여러 개의 부분집합(subset)으로 나누어 학습을 진행하는 것이 핵심
- 각 부분집합은 hyperparameter에 의해 일정 수준 교차



Methods



❖ SubTab

- Contrastive/Distance Loss는 optional 한 loss
  - ➢ Negative pair → Contrastive Loss
  - ➢ Positive pair → Contrastive Loss, Distance Loss(MSE)
- 클래스의 수가 많을 수록 Negative pair가 제대로 매칭될 확률이 높으므로 효과가↑

$Subset_i$

X2	X3	X4	X5	X6	X7	X8	...	X10	X11	X12	X13	X14	X15	X16
4.68	4.621	4.54	5.079	3.763	5.64	4.281	...	4.666	4.54	5.17	4.651	5.262	4.418	6.025
4.874	4.782	4.686	5.01	3.845	5.773	4.21	...	4.867	4.744	4.952	4.76	4.99	4.039	5.957
4.906	4.768	4.848	4.927	4.087	5.752	4.469	...	4.891	4.802	5.044	4.829	4.91	4.243	5.743
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5.391	5.299	5.152	4.862	4.429	5.71	0	...	5.395	5.288	4.297	4.907	4.482	3.543	5.731
5.016	4.907	4.621	4.682	3.955	6.058	4.669	...	5.008	4.924	4.505	5.014	5.146	3.729	5.727
4.774	4.707	4.669	5.025	3.914	5.627	4.312	...	4.792	4.676	5.035	4.737	5.024	4.2	5.856
5.067	4.985	4.765	4.756	4.046	5.975	4.526	...	5.086	4.918	4.693	4.966	4.888	3.853	5.693
5.253	5.209	5.056	4.336	4.383	6.435	5.056	...	5.24	5.11	4.513	5.557	4.817	3.795	5.206
5.334	5.217	5.084	4.378	4.37	6.319	4.971	...	5.314	5.185	4.383	5.227	4.564	3.597	5.333
5.16	5.105	4.872	4.747	4.126	5.803	4.67	...	5.192	5.091	4.542	5.083	4.828	3.782	5.363
5.066	4.92	4.919	4.566	4.188	6.217	4.638	...	5.032	4.96	4.76	5.144	4.979	4.054	5.472
4.642	4.582	4.584	5.445	3.869	5.101	4.425	...	4.613	4.473	5.304	4.325	5.066	4.427	6.177
4.937	4.842	4.719	4.878	4.047	5.643	4.664	...	4.924	4.812	4.902	4.811	4.888	4.178	5.709
5.127	5.026	4.968	4.509	4.335	6.079	4.647	...	5.117	5.003	4.66	5.311	4.794	3.918	5.238
4.333	4.295	4.255	5.285	3.47	5.386	3.682	...	4.306	4.209	5.571	4.422	5.543	4.763	6.098
4.947	4.839	4.793	4.658	4.086	6.019	4.662	...	4.922	4.81	4.923	5.055	5.038	4.247	5.532
4.812	4.668	4.541	5.011	3.728	5.696	4.193	...	4.826	4.756	4.821	4.778	5.135	3.875	5.888
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

$Subset_j$

X2	X3	X4	X5	X6	X7	X8	...	X10	X11	X12	X13	X14	X15	X16
4.68	4.621	4.54	5.079	3.763	5.64	4.281	...	4.666	4.54	5.17	4.651	5.262	4.418	6.025
4.874	4.782	4.686	5.01	3.845	5.773	4.21	...	4.867	4.744	4.952	4.76	4.99	4.039	5.957
4.906	4.768	4.848	4.927	4.087	5.752	4.469	...	4.891	4.802	5.044	4.829	4.91	4.243	5.743
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5.391	5.299	5.152	4.862	4.429	5.71	0	...	5.395	5.288	4.297	4.907	4.482	3.543	5.731
5.016	4.907	4.621	4.682	3.955	6.058	4.669	...	5.008	4.924	4.505	5.014	5.146	3.729	5.727
4.774	4.707	4.669	5.025	3.914	5.627	4.312	...	4.792	4.676	5.035	4.737	5.024	4.2	5.856
5.067	4.985	4.765	4.756	4.046	5.975	4.526	...	5.086	4.918	4.693	4.966	4.888	3.853	5.693
5.253	5.209	5.056	4.336	4.383	6.435	5.056	...	5.24	5.11	4.513	5.557	4.817	3.795	5.206
5.334	5.217	5.084	4.378	4.37	6.319	4.971	...	5.314	5.185	4.383	5.227	4.564	3.597	5.333
5.16	5.105	4.872	4.747	4.126	5.803	4.67	...	5.192	5.091	4.542	5.083	4.828	3.782	5.363
5.066	4.92	4.919	4.566	4.188	6.217	4.638	...	5.032	4.96	4.76	5.144	4.979	4.054	5.472
4.642	4.582	4.584	5.445	3.869	5.101	4.425	...	4.613	4.473	5.304	4.325	5.066	4.427	6.177
4.937	4.842	4.719	4.878	4.047	5.643	4.664	...	4.924	4.812	4.902	4.811	4.888	4.178	5.709
5.127	5.026	4.968	4.509	4.335	6.079	4.647	...	5.117	5.003	4.66	5.311	4.794	3.918	5.238
4.333	4.295	4.255	5.285	3.47	5.386	3.682	...	4.306	4.209	5.571	4.422	5.543	4.763	6.098
4.947	4.839	4.793	4.658	4.086	6.019	4.662	...	4.922	4.81	4.923	5.055	5.038	4.247	5.532
4.812	4.668	4.541	5.011	3.728	5.696	4.193	...	4.826	4.756	4.821	4.778	5.135	3.875	5.888
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Methods

❖ SubTab

- 실험 결과

➢ 얕은 구조의 MLP만 활용하여 CNN 기반 방법론 보다 비교적 높은 성능을 달성

Rank	Model	Accuracy↑	F1 (%)	Paper	Code	Result	Year	Tags
1	IIC	99.3		Invariant Information Clustering for Unsupervised Image Classification and Segmentation			2018	
2	SCAE (LIN-MATCH)	98.7		Stacked Capsule Autoencoders			2019	
3	SubTab	98.31		SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation Learning			2021	
4	Self-Organizing Map							
5	Bidirectional InfoGAN							
6	Adversarial AE							
7	CatGAN							
8	InfoGAN							
9	PixelGAN Autoencoders							
10	density based							

Table 1: Accuracy scores for all models for various datasets. The abbreviations in the table; NC: Neighbour columns used, RF: Random features used, G: Gaussian noise used, S: Swap noise used.

Type	Models	MNIST	Income	Blog	Obesity	TCGA
Supervised baseline	Logistic Regression	92.60±0.03	84.68±0.05	84.15±0.12	62.35±4.02	36.98± 1.25
	Random Forest	96.96±0.06	84.62±0.07	83.61±0.15	67.45±2.23	61.62± 1.02
	XGBoost	98.02±0.086	86.11±0.20	84.29±0.23	64.05±4.52	72.61±1.31
Autoencoder baseline	AE	92.77±0.32	84.67±0.07	84.06±0.24	61.96±3.28	55.16±0.75
	AE w/ Dropout (p=0.2)	94.31±0.28	85.00±0.10	84.18±0.20	62.74±4.38	56.87±2.26
Self-supervised	DAE (RF)	96.30±0.14 (S)	84.37±0.36 (G)	84.12±0.29 (G)	56.43±5.79 (G)	54.31±1.39 (G)
	CAE (NC)	96.39±0.20 (S)	84.24±0.18 (G)	84.3±0.31 (G)	62.26±5.01 (G)	54.20±1.17 (G)
	VIME-self	95.23±0.17 (S)	84.43±0.08 (G)	84.11±0.27 (G)	66.45±4.54 (G)	55.11±1.37 (G)
	SubTab with:					
	Base model (No noise)	97.26±0.2	85.31±0.08	84.29±0.26	68.01±3.07	57.02±1.50
	+Noise	97.47±0.18 (S)	85.34±0.07 (G)	84.47±0.15 (G)	**71.13±4.08 (G)**	**58.25±1.36 (G)**
	+Distance loss	97.52±0.14 (S)	**85.35±0.06** (G)	**84.64±0.19** (G)	69.25±4.19 (G)	58.15±1.56 (G)
	+LatentDim=512	**97.86±0.07 (S)**	-	-	-	-

Methods

❖ **SCARF:** Self-Supervised Contrastive Learning using Random Feature Corruption

- ICLR (2022), 2022년 10월 5일 기준 9회 인용

SCARF: SELF-SUPERVISED CONTRASTIVE LEARNING USING RANDOM FEATURE CORRUPTION

Dara Bahri, Heinrich Jiang, Yi Tay, Donald Metzler
Google Research
{dbahri,heinrichj,yitay,metzler}@google.com

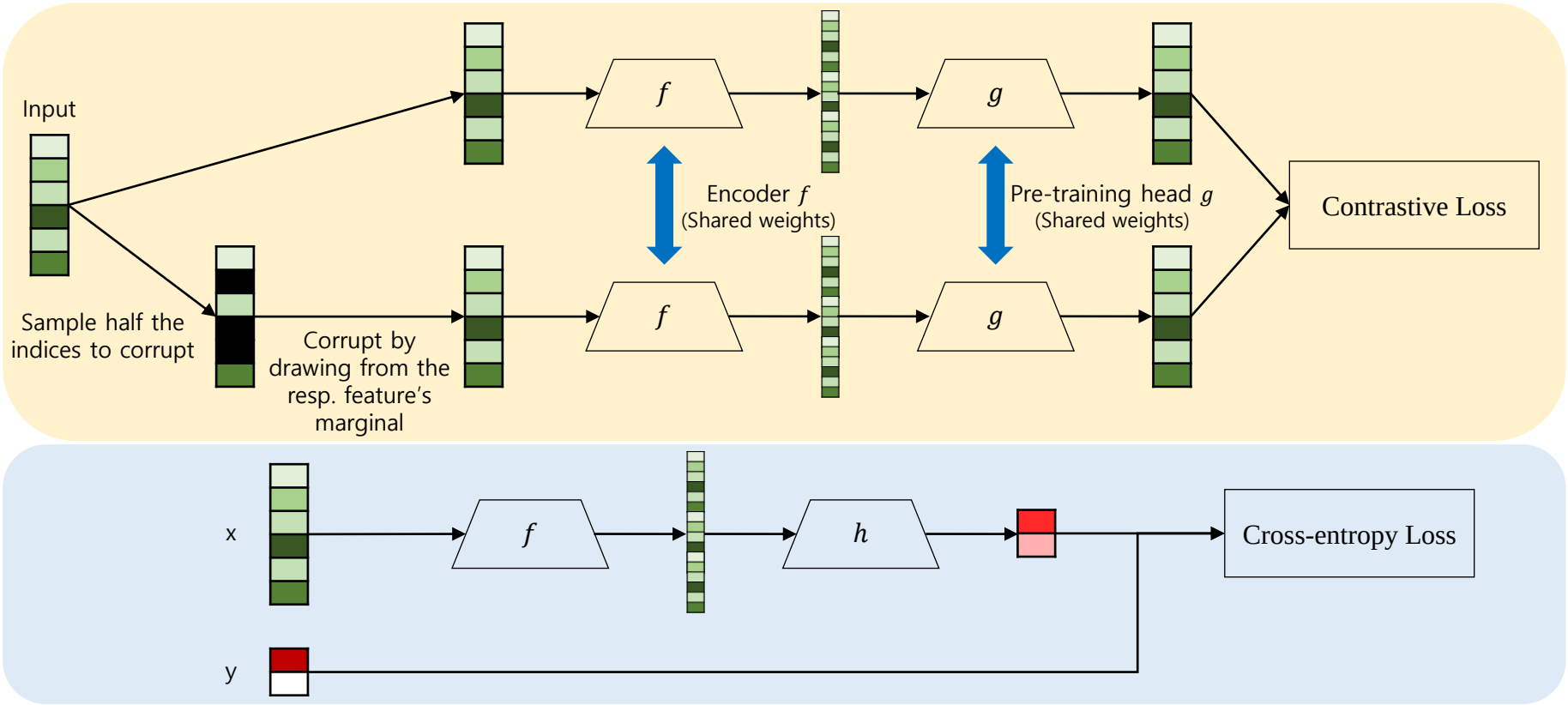
ABSTRACT

Self-supervised contrastive representation learning has proved incredibly successful in the vision and natural language domains, enabling state-of-the-art performance with orders of magnitude less labeled data. However, such methods are domain-specific and little has been done to leverage this technique on real-world *tabular* datasets. We propose SCARF, a simple, widely-applicable technique for contrastive learning, where views are formed by corrupting a random subset of features. When applied to pre-train deep neural networks on the 69 real-world, tabular classification datasets from the OpenML-CC18 benchmark, SCARF not only improves classification accuracy in the fully-supervised setting but does so also in the presence of label noise and in the semi-supervised setting where only a fraction of the available training data is labeled. We show that SCARF complements existing strategies and outperforms alternatives like autoencoders. We conduct comprehensive ablations, detailing the importance of a range of factors.

Methods

❖ SCARF

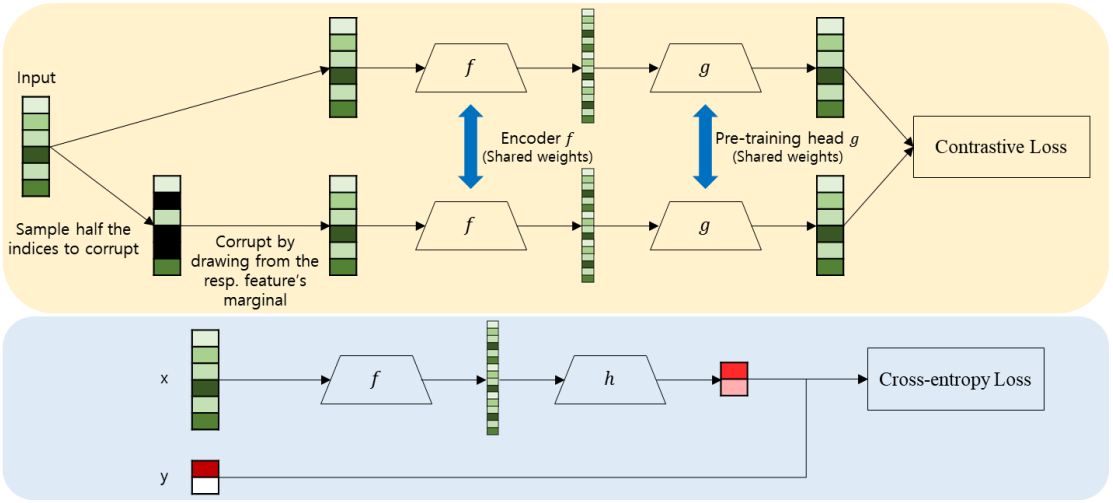
- Network : Encoder(f), Pre-training head(g), Classification head(h)
- Losses : Contrastive loss(maximize similarity), Cross-entropy loss



Methods

❖ SCARF

- 특징
  - ➢ Pre-training에서 학습했던 network중 g는 폐기하고 f만 사용
  - ➢ Fine-tuning 단계에서 f와 h를 모두 학습
  - ➢ 이전 VIME, SubTab과 달리 오토인코더 구조를 통한 기존 데이터 복원을 하지 않음



Methods

❖ **Contrastive Mixup:** Self- and Semi-Supervised Learning for Tabular Domain

- 2022년 10월 5일 현재 4회 인용
- 준지도학습 프레임워크 제안

Contrastive Mixup: Self- and Semi-Supervised learning for Tabular Domain

Sajad Darabi
UCLA
sajad.darabi@cs.ucla.edu

Shayan Fazeli
UCLA
shayan.fazeli@cs.ucla.edu

Ali Pazokitoroudi
UCLA
alipazoki@cs.ucla.edu

Sriram Sankararaman
UCLA
sriram@cs.ucla.edu

Majid Sarrafzadeh
UCLA
majid@cs.ucla.edu

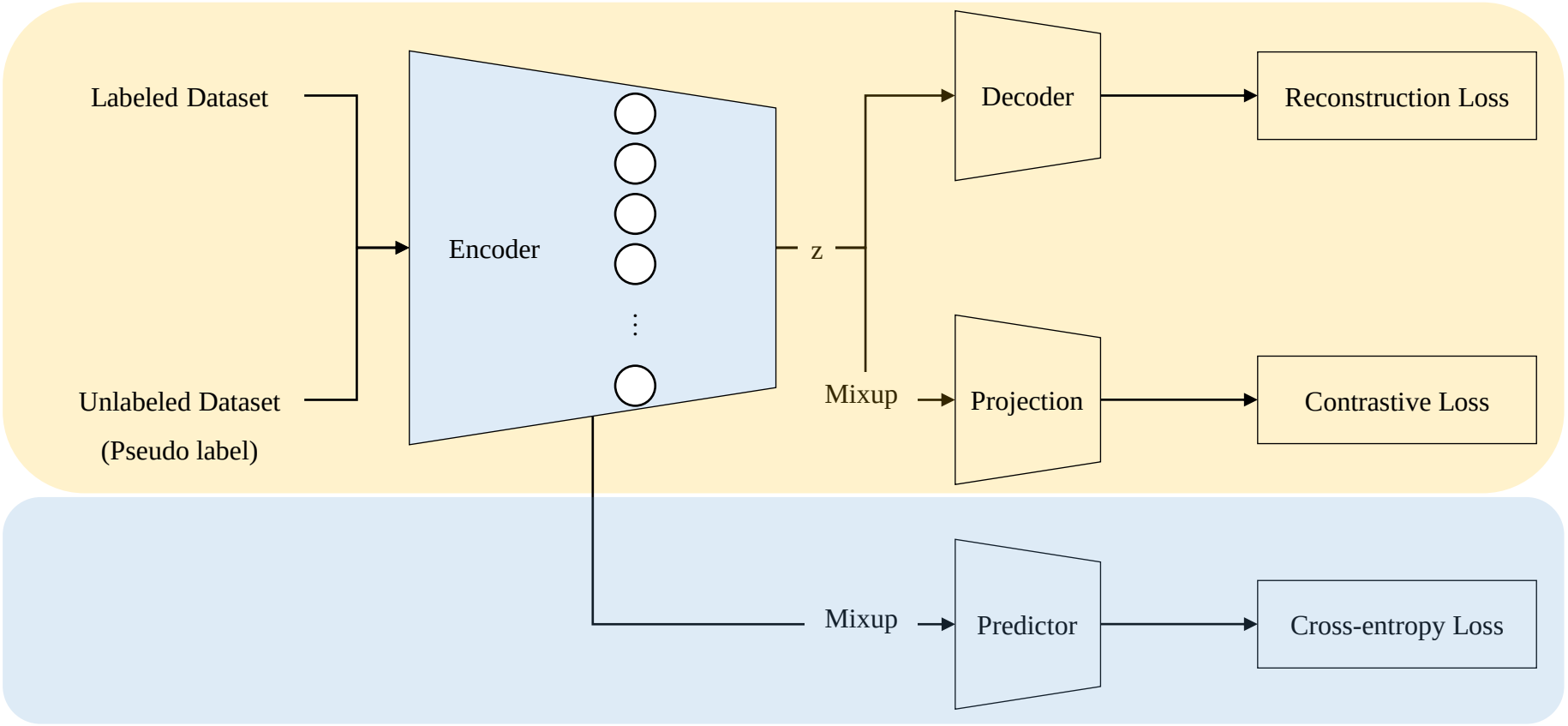
Abstract

Recent literature in self-supervised has demonstrated significant progress in closing the gap between supervised and unsupervised methods in the image and text domains. These methods rely on domain-specific augmentations that are not directly amenable to the tabular domain. Instead, we introduce Contrastive Mixup, a semi-supervised learning framework for tabular data and demonstrate its effectiveness in limited annotated data settings. Our proposed method leverages Mixup-based augmentation under the manifold assumption by mapping samples to a low dimensional latent space and encourage interpolated samples to have high a similarity within the same labeled class. Unlabeled samples are additionally employed via a transductive label propagation method to further enrich the set of similar and dissimilar pairs that can be used in the contrastive loss term. We demonstrate the effectiveness of the proposed framework on public tabular datasets and real-world clinical datasets.

Methods

❖ Contrastive Mixup

- Network : Encoder, Decoder, Projection, Predictor
- Losses : Reconstruction loss, Contrastive loss, Cross-entropy loss



Methods

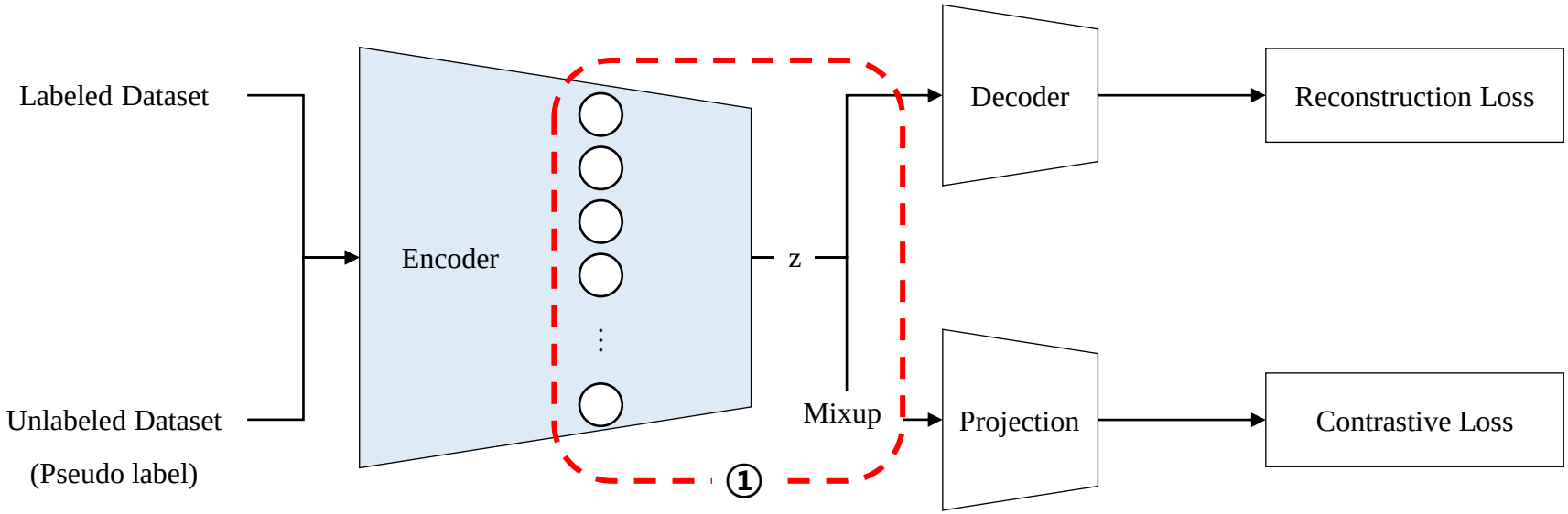
❖ Contrastive Mixup

- 특징 ①

  ➢ Mixup을 통한 데이터 증강을 Encoder 내부의 특정 Layer에서 진행

  동일한 label을 가진 sample을 사용하여 데이터 증강 수행

→ 정형 데이터에 더 적합



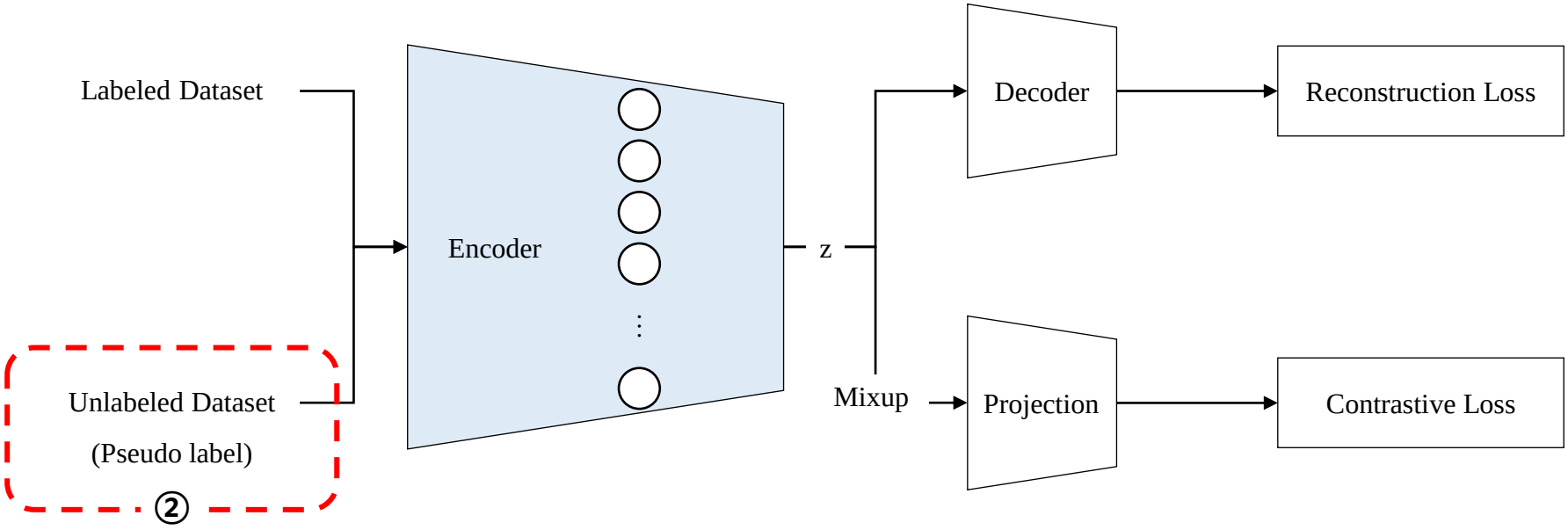
Methods

❖ Contrastive Mixup

- 특징 ②

  ➢ Representation vector z를 활용하여 인접 데이터와의 유사도를 사용하여 pseudo label 생성

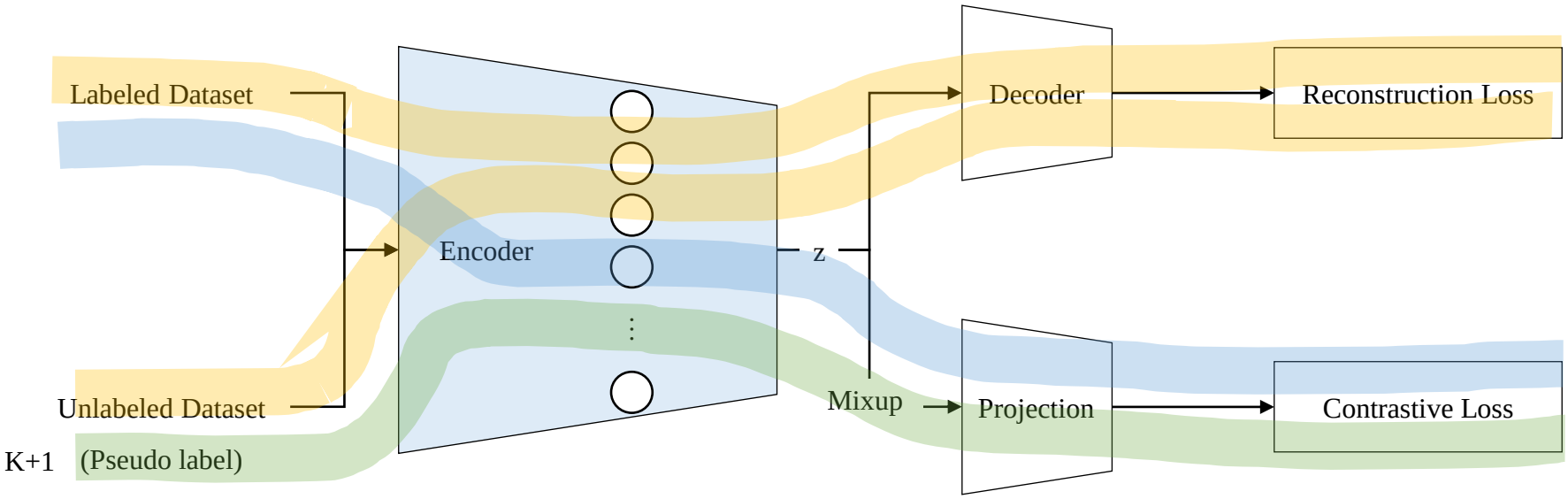
  Pseudo label은 매 epoch 마다 업데이트



Methods

❖ Contrastive Mixup

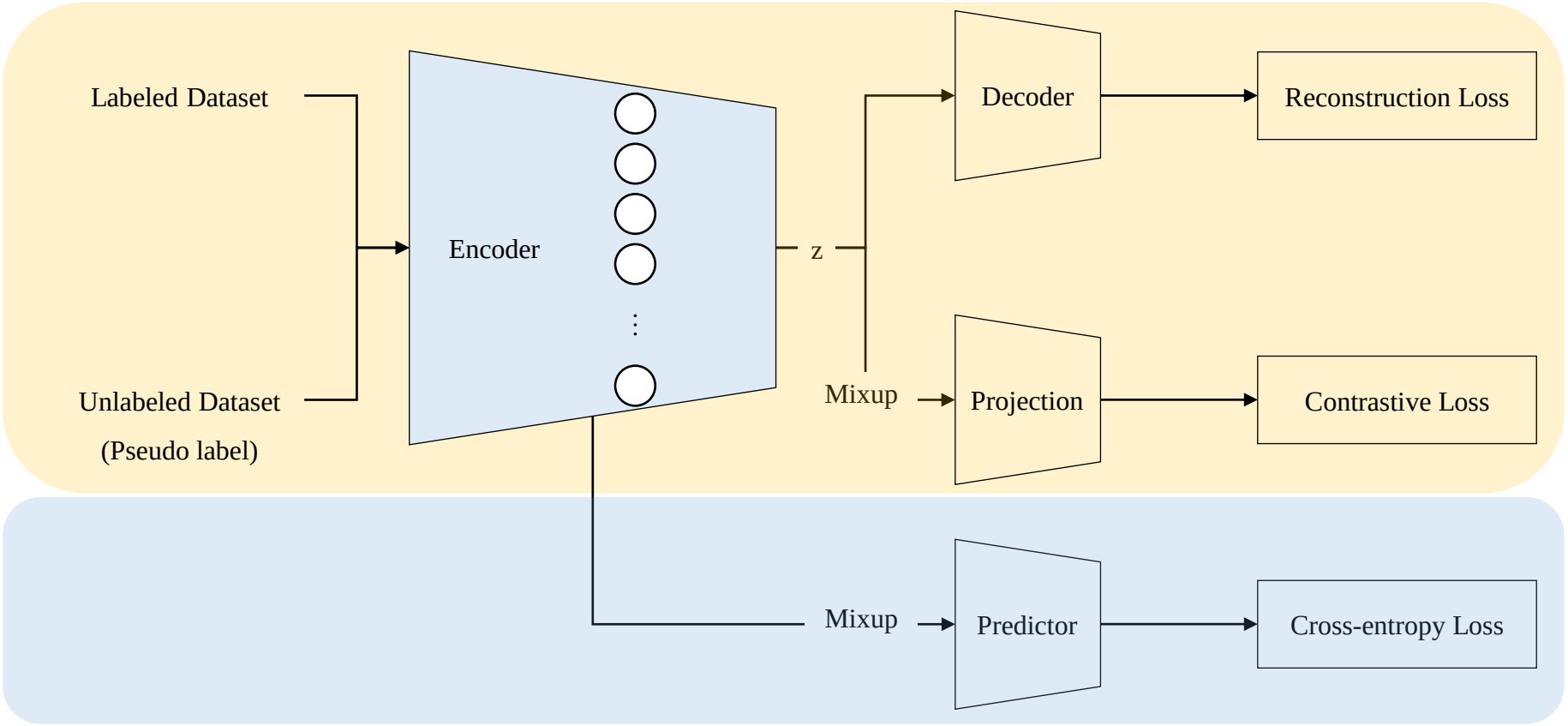
- Pseudo label은 K+1 epoch 부터 활용
  - ➢ 입력 데이터와 출력 데이터의 차이를 이용한 학습
  - ➢ 동일한 label을 가진 데이터로 증강된 데이터는 유사해야 함을 이용한 학습
  - ➢ K+1 epoch 부터 pseudo label을 활용



Methods

❖ Contrastive Mixup

- Pre-training을 통하여 Encoder를 학습 및 고정 후 Downstream task를 통해 Classification 문제 해결
  - ➢ Predictor 학습 시 Pseudo label 사용



Methods

❖ Contrastive Mixup

- 실험 결과

Table 1: Comparison on public tabular datasets.

Type	Method	Dataset			
		MNIST	**Adult**	**Blog Feedback**	**Covertype**
Supervised	Logistic	90.12 (±0.098)	82.41 (±0.413)	78.91 (±0.22)	70.54 (±0.087)
	MLP	93.69 (±0.234)	83.19 (±0.663)	79.63 (±0.519)	75.95 (±0.202)
	CatBoost (100%)	97.41 (±0.098)	87.54 (±0.075)	85.08 (±0.088)	88.64 (±0.077)
Semi-supervised	AE	94.72 (±0.127)	84.18 (±0.078)	80.09 (±0.199)	79.67 (±0.296)
	Manifold Mixup	94.92 (±0.012)	84.68 (±0.279)	80.24 (±0.652)	78.79 (±0.135)
	VIME	95.71 (±0.013)	84.54 (±0.408)	81.36 (±0.301)	79.02 (±0.329)
Ours (Ablation)	Supervised	93.69 (±0.234)	83.19 (±0.663)	79.63 (±0.519)	75.95 (±0.202)
	Self-SL	95.82 (±0.131)	85.16 (±0.249)	81.38 (±0.373)	79.46 (±0.463)
	Self-SL+PL	97.01 (±0.066)	85.26 (±0.207)	81.65 (±0.370)	79.92 (±0.682)
	Ours	**97.58** (±0.078)	**85.42** (±0.210)	**81.88** (±0.123)	**80.41** (±0.205)

Conclusions

Conclusions

❖ Conclusions

- Introduction
  - ➢ 최근 Self/semi-supervised learning의 배경
  - ➢ 정형 데이터(tabular data)와 기존 방법론의 한계
- Methods
  - VIME, SubTab, SCARF, Contrastive Mixup

- 정형 데이터에 자기지도학습 및 준지도학습 적용 연구가 활발히 진행되고 있고 다양한 아이디어 적용을 통한 연구 가능성이 크다고 생각됨

References

❖ References

- Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679-6687.
- Bahri, D., Jiang, H., Tay, Y., & Metzler, D. (2021). Scarf: Self-supervised contrastive learning using random feature corruption. arXiv preprint arXiv:2106.15147.
- Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.
- Darabi, S., Fazeli, S., Pazoki, A., Sankararaman, S., & Sarrafzadeh, M. (2021). Contrastive mixup: Self-and semi supervised learning for tabular domain. arXiv preprint arXiv:2108.12296.
- Sun, B., Yang, L., Zhang, W., Lin, M., Dong, P., Young, C., & Dong, J. (2019). Supertml: Two-dimensional word embedding for the precognition on structured tabular data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
- Ucar, T., Hajiramezanali, E., & Edwards, L. (2021). Subtab: Subsetting features of tabular data for self-supervised representation learning. Advances in Neural Information Processing Systems, 34, 18853-18865.
- Yoon, J., Zhang, Y., Jordon, J., & van der Schaar, M. (2020). Vime: Extending the success of self-and semisupervised learning to tabular domain. Advances in Neural Information Processing Systems, 33, 11033-11043.

Thank you