What is Next for Tabular Data?

Exploring Advances in Self-Supervised Learning

2024. 04. 05

Data Mining & Quality Analytics Lab.

채고은

발표자 소개

❖ 채고은 (Goeun Chae)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석사 과정 (2022. 03 ~ Present)

❖ Research Interest

- Self-Supervised Learning
- Hard Negative Sampling in Contrastive Learning

❖ Contact

- goeunchae@korea.ac.kr

Contents

1. Introduction

Introduction

Background

❖ Tabular Data

- 다양한 분야에서 사용되는 가장 일반적인 데이터 유형
- 데이터 포인트를 나타내는 Row와 속성이나 변수를 나타내는 Column 으로 구성
- 정보를 체계적이고 직관적인 방식으로 구성하므로 데이터 관리에 용이

Columns
Attributes for those observations

Rows
Observations

Superhero	Real Name	Power Level	Home Planet	Weapon	Member Since
Spider-Man	Peter Parker	85	Earth	Web-shooters	1962
Iron Man	Tony Stark	88	Earth	Powered Armor	1963
Hulk	Bruce Banner	90	Earth	Super Strength	1962
Thor	Thor Odinson	95	Asgard	Mjolnir	1962

Tabular Data

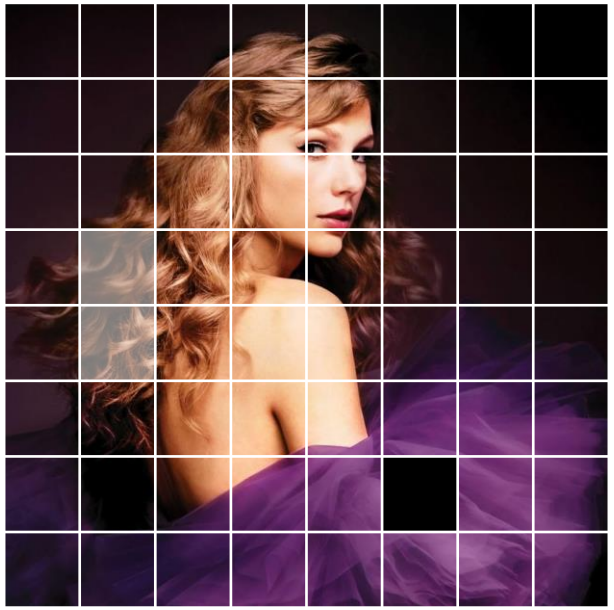
Introduction

Background

❖ Why is it hard to model deep learning methods to tabular data?

- 정형 데이터의 분석은 주로 트리 기반 모델들의 앙상블 모델들이 사용됨
- 정형 데이터에서는 비정형 데이터에서 높은 성능을 보이는 deep learning 에 비해 여전히 전통적 통계 기반의 방법론들이 우세
- Deep learning은 Input 데이터의 locality & spatial dependency 에 집중하여 발전

  → Tabular data 에서는 위의 특성들을 고려하기 어려움



Superhero	Real Name	Power Level	Home Planet	Weapon	Member Since
Spider-Man	Peter Parker	85	Earth	Web-shooters	1962
Iron Man	Tony Stark	88	Earth	Powered Armor	1963
Hulk	Bruce Banner	90	Earth	Super Strength	1962
Thor	Thor Odinson	95	Asgard		1962

Introduction

Background

❖ Why is it hard to model deep learning methods to tabular data?

- Heterogeneous features : 개별 변수간 형태적 다양성
- Small sample size : 학습에 부족한 작은 사이즈의 데이터
- Extreme value : 결측치, 범주형 변수들, outlier 등 데이터의 값들이 stable 하지 않음

80 Rows

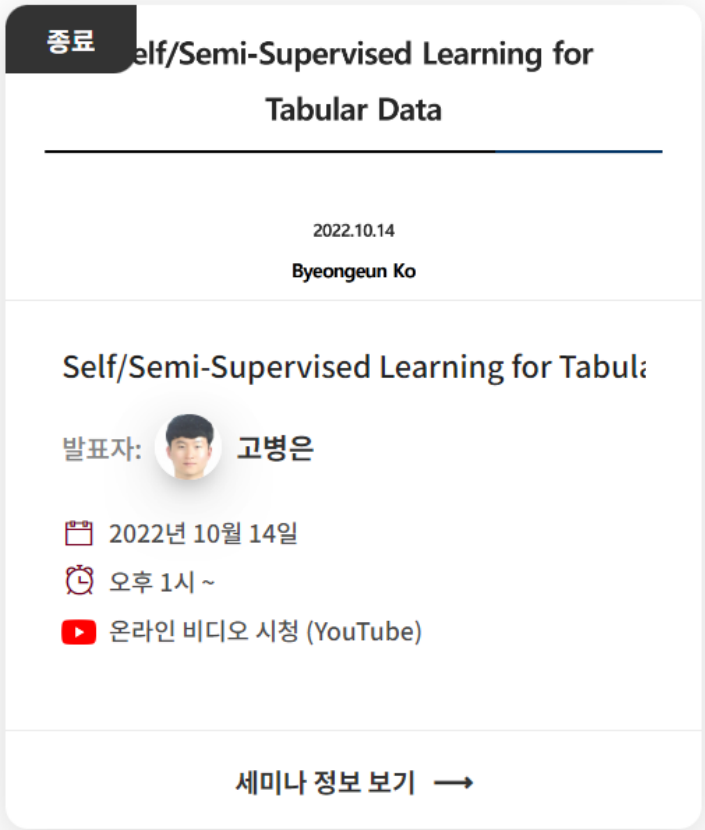
Superhero	Real Name	Power Level	Home Planet	Weapon	Member Since
Spider-Man	Peter Parker	85		Web-shooters	1962
Iron Man		88	Earth	Powered Armor	1963
Hulk	Bruce Banner	90	Earth	Super Strength	1962
Thor	Thor Odinson	95	Asgard	Mjolnir	
…	…	…	…	…	…
Black Widow	Natasha Romanoff		Earth	Combat Skills	1964

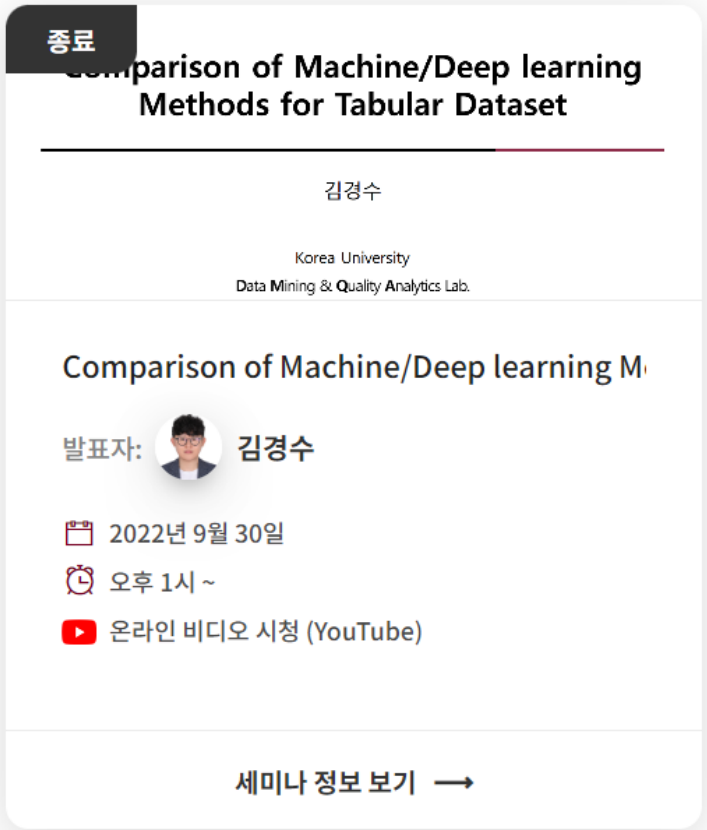
Introduction

Background

❖ Why is it hard to model deep learning methods to tabular data?

- Heterogeneous features : 개별 변수간 형태적 다양성
- Small samp
- Extreme va





Superl			Member Since
Spider-			1962
Iron M			1963
Hull			1962
Tho			
...			...
Black W			1964

80 Rows

Introduction

Self-Supervised Learning for Tabular Data

❖ Self-Supervised Learning for Tabular Data

- Unlabeled 데이터에 대한 접근성이 확대됨에 따라 Self-supervised Learning(SSL)의 필요성과 역할이 중요
- 기존 SSL은 text, image, audio 등 다양한 도메인에서의 확장성 증명
- Tabular 데이터에서 표현 학습을 위해 크게 3가지 접근으로 분류 가능

Self-Supervised Learning for Tabular Data	Predictive Learning	다양한 예측과제를 통해 모델이 데이터로부터 배경 지식을 학습하게 하여 downstream 성능 향상	Vime, TabNet, STUNT, SwitchTab
	Contrastive Learning	대조학습을 통해 instance 간의 유사성과 차이를 학습시켜 Task-agnostic 한 학습 목표	SCARF, STab, TransTab, PTaRL
	Hybrid Learning	위의 두 학습을 결합하여 통합된 학습 지향 두 방식의 장점을 모두 가짐	SubTab, SAINT, DoRA, CT-BERT

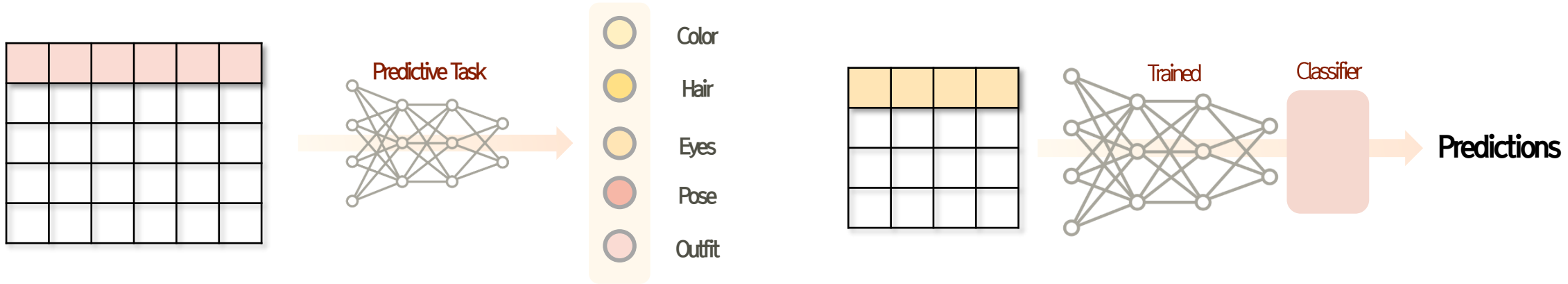
[1] Wang, W. Y., Du, W. W., Xu, D., Wang, W., & Peng, W. C. (2024). A Survey on Self-Supervised Learning for Non-Sequential Tabular Data. arXiv preprint arXiv:2402.01204.

2. Predictive Learning

Predictive Learning

❖ Predictive Learning Approach

- 가장 넓게 사용되는 접근 방식으로, downstream 이전에 predictive task 설계
- 모델이 주어진 predictive task를 수행하며 데이터로부터 표현벡터 학습
- Upstream 데이터와 downstream 데이터의 관계를 고려하여 효과적인 predictive task 고안이 중요
  - ✓ Learning from masked features & perturbation in latent space

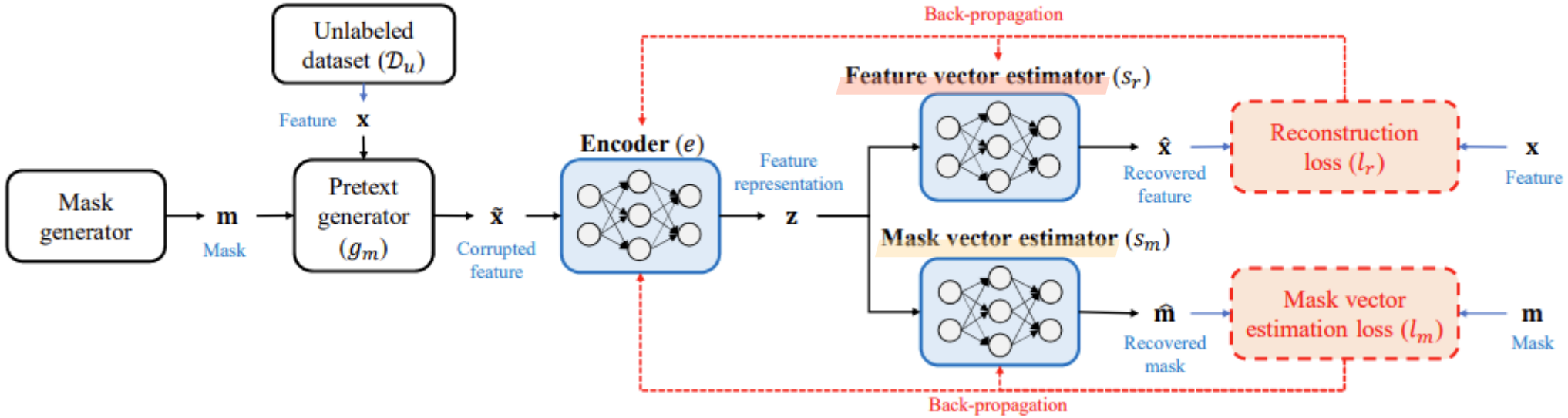


Predictive Learning

Learning from Masked Features

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

- Masked Autoencoder를 기반으로 데이터의 무작위 마스킹을 복구하며 모델 최적화
- Mask vector estimator : masked 특성 식별
- Feature vector estimator : 관련된 non-masked 특성으로 부터 masked 특성 추정
- Reconstruction loss & masked vector estimation loss 를 통해 표현벡터 학습



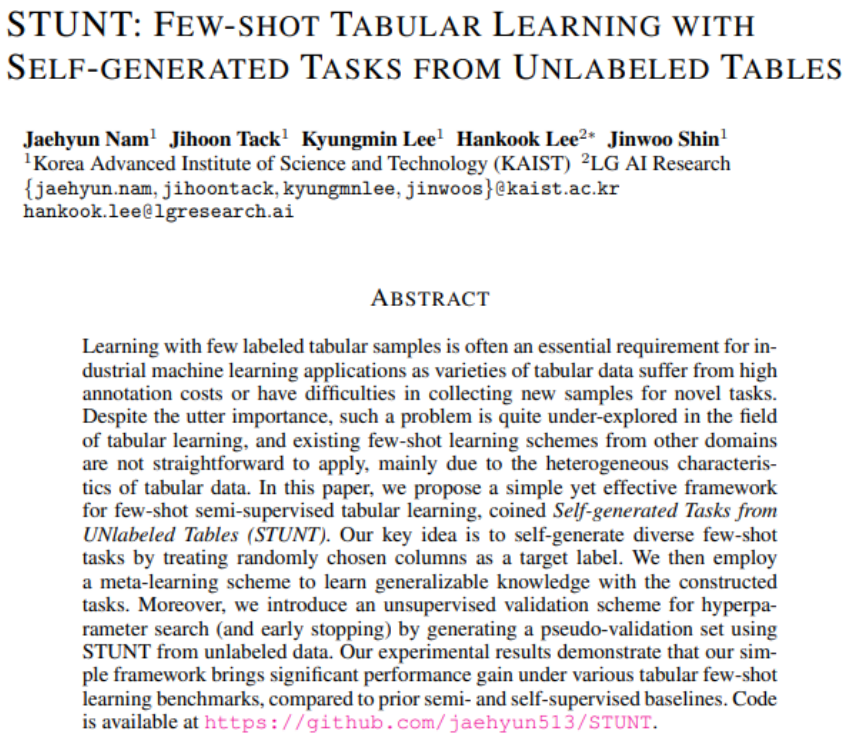
[2] Yoon, J., Zhang, Y., Jordon, J., & Van der Schaar, M. (2020). Vime: Extending the success of self-and semi-supervised learning to tabular domain. Advances in Neural Information Processing Systems, 33, 11033-11043.

Predictive Learning

Perturbation in latent space

❖ STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables (ICLR, 2023)

- Tabular 데이터의 heterogeneous features 에서 일반화 가능한 특성을 학습하기 위해 self-generated tasks 메타학습 진행
- 표의 column 특성을 유용한 target으로 취급하여 unlabeled tabular 데이터로부터 다양한 set of tasks 생성



STUNT: FEW-SHOT TABULAR LEARNING WITH SELF-GENERATED TASKS FROM UNLABELED TABLES

Jaehyun Nam[1] **Jihoon Tack**[1] **Kyungmin Lee**[1] **Hankook Lee**[2*] **Jinwoo Shin**[1]
[1]Korea Advanced Institute of Science and Technology (KAIST) [2]LG AI Research
{jaehyun.nam, jihoontack, kyungmnlee, jinwoos}@kaist.ac.kr
hankook.lee@lgresearch.ai

ABSTRACT

Learning with few labeled tabular samples is often an essential requirement for industrial machine learning applications as varieties of tabular data suffer from high annotation costs or have difficulties in collecting new samples for novel tasks. Despite the utter importance, such a problem is quite under-explored in the field of tabular learning, and existing few-shot learning schemes from other domains are not straightforward to apply, mainly due to the heterogeneous characteristics of tabular data. In this paper, we propose a simple yet effective framework for few-shot semi-supervised tabular learning, coined *Self-generated Tasks from UNlabeled Tables (STUNT)*. Our key idea is to self-generate diverse few-shot tasks by treating randomly chosen columns as a target label. We then employ a meta-learning scheme to learn generalizable knowledge with the constructed tasks. Moreover, we introduce an unsupervised validation scheme for hyperparameter search (and early stopping) by generating a pseudo-validation set using STUNT from unlabeled data. Our experimental results demonstrate that our simple framework brings significant performance gain under various tabular few-shot learning benchmarks, compared to prior semi- and self-supervised baselines. Code is available at https://github.com/jaehyun513/STUNT.

[3] Nam, J., Tack, J., Lee, K., Lee, H., & Shin, J. (2023). Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. arXiv preprint arXiv:2303.00918.

Predictive Learning

STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

❖ STUNT: Self-generated Tasks from Unlabeled Tables

- Step 1 : Unlabeled tabular 데이터에서 다양한 과제 자체 생성
- Step 2 : 생성된 과제에 대해 메타학습 수행을 통해 일반화된 분류기 f 학습
- Step 3: Labeled 데이터를 사용하여 분류기 f 적응

https://arxiv.org/pdf/1703.05175.pdf

Predictive Learning

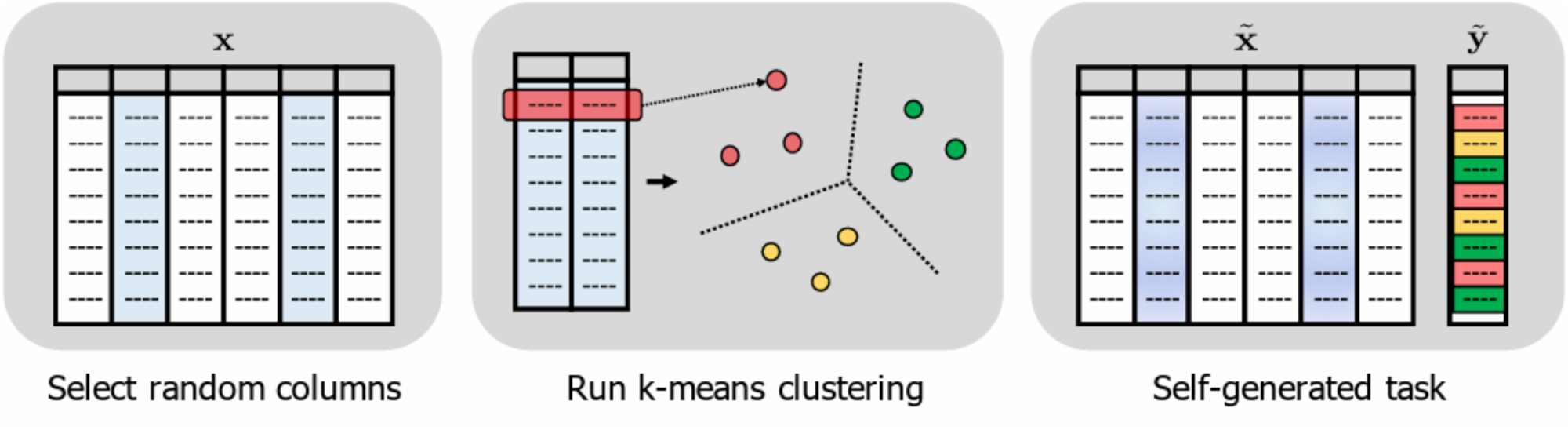
STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

❖ Step 1 : Task generation from unlabeled tables

- 무작위로 선택된 columns에 대해 k-mean clustering을 통해 pseudo-label 생성
- 다양성을 향상시키고 원래 label과 높은 상관관계를 가지는 column 을 sampling 할 가능성 증가
- Uniform distribution $U(r_1, r_2)$에서 masking 비율 p 샘플링

  → Random binary mask $m = [m_1, m_2, \ldots, m_d]^T \in \{0,1\}^d$ 생성

  → Generated task from STUNT $\tau_{STUNT} := \{\tilde{x}_{u,i}, \tilde{y}_{u,i}\}_{i=1}^{N_u}$



[3] Nam, J., Tack, J., Lee, K., Lee, H., & Shin, J. (2023). Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. arXiv preprint arXiv:2303.00918.

Predictive Learning

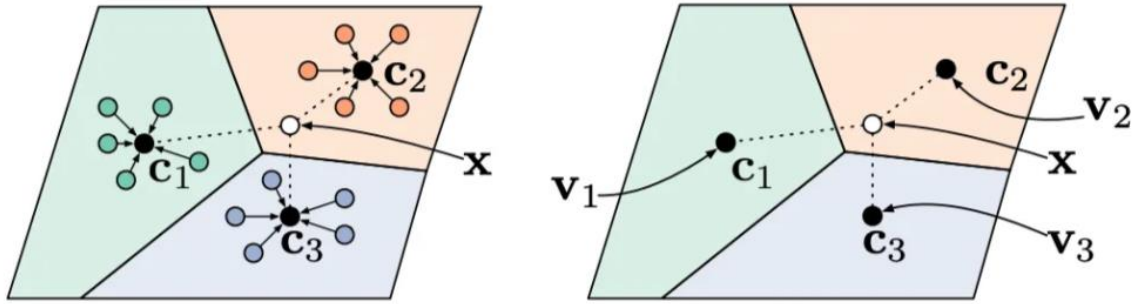
STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

❖ Step 2 : Meta-learning with STUNT

- 생성된 task를 바탕으로 ProtoNet을 활용하여 네트워크의 메타학습 제안
- ProtoNet은 각 class 샘플의 평균 embedding vector 까지의 거리 계산을 통해 분류를 수행할 수 있는 embedding space 학습
  - ✓ Class number flexibility, model and data agnosticism, simplicity with superior performance

ProtoNet

Embedding space + k-means for meta-learning



$$p_{\tilde{c}} = \frac{1}{|S_{\tilde{c}}|} \sum_{(\tilde{x}_u, \tilde{y}_u) \in S_{\tilde{c}}} z_\theta(\tilde{x}_u)$$

$$f_\theta(y = \tilde{c}|x; S) = \frac{\exp(-\| z_\theta(x) - p_{\tilde{c}} \|_2)}{\sum_{\bar{c}'} \exp(-\| z_\theta - p_{\bar{c}'} \|_2)}$$

Predictive Learning

STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

❖ Experiments

- Few-shot classification 을 위해 각 class 별로 1개, 5개의 labeled samples 가 가능하도록 설정
- kNN과 unsupervised meta-learning 방법을 제외하고 100개의 추가 labeled samples 사용하여 hyperparameter 탐색
- Labeled validation set 없이 few-shot tabular 분류 성능을 크게 향상

1 Shot Performance Evaluation

Type	Method	Val.	income	cmc	karhunen	optdigit	diabetes	semeion	pixel	dna	Avg.
# shot = 1											
Sup.	CatBoost	✓	57.00	34.60	55.67	61.32	60.02	43.21	59.16	41.35	52.06
	MLP	✓	60.52	35.06	48.67	61.02	57.25	40.88	55.62	44.39	50.43
	LR	✓	59.64	35.08	55.05	65.19	57.61	42.90	59.71	44.28	52.43
	kNN	-	61.22	34.99	54.42	65.58	58.56	44.35	61.48	42.67	52.82
Semi-sup.	Mean Teacher	✓	60.63	35.58	54.57	66.10	58.05	43.56	61.02	46.58	53.26
	ICT	✓	61.83	36.53	58.37	69.12	58.08	43.48	60.88	46.55	54.36
	Pseudo-Label	✓	60.52	34.97	49.44	61.50	57.03	41.42	56.12	44.26	50.66
	MPL	✓	60.85	35.13	47.66	61.52	57.39	41.82	56.01	44.22	50.58
	VIME-Semi	✓	56.40	32.97	57.40	66.85	58.16	40.43	52.86	39.18	50.53
Self-sup.	SubTab + Fine-tune	✓	59.74	35.65	41.11	49.88	59.35	30.49	42.23	40.86	44.91
	SubTab + LR	✓	61.88	35.68	50.32	67.05	58.06	40.27	60.40	45.68	52.42
	SubTab + kNN	-	61.58	35.87	48.74	66.05	59.22	39.99	61.30	44.16	52.36
	VIME + Fine-tune	✓	60.50	34.98	47.50	61.31	57.23	41.09	53.79	44.30	50.09
	VIME + LR	✓	61.99	35.30	59.62	70.52	56.95	47.20	64.17	51.36	55.89
	VIME + kNN	-	62.16	35.55	58.56	69.31	58.35	46.99	64.62	50.29	55.78
Unsup.-Meta.	UMTRA	-	57.23	35.46	49.05	49.87	57.64	26.33	34.26	25.13	41.87
	SES	-	56.39	34.59	49.19	56.30	59.97	33.73	49.19	39.56	47.37
	CACTUs	-	**64.02**	36.10	65.59	71.98	58.92	48.96	67.61	65.93	59.89
	STUNT (Ours)	-	63.52	**37.10**	**71.20**	**76.94**	**61.08**	**55.91**	**79.05**	**66.20**	**63.88**

5 Shot Performance Evaluation

# shot = 5											
Sup.	CatBoost	✓	64.51	39.75	82.38	84.05	65.75	68.69	84.49	63.46	69.14
	MLP	✓	66.25	37.40	77.56	83.30	64.32	66.25	81.97	59.73	67.10
	LR	✓	66.53	37.15	81.02	86.22	64.19	67.87	85.02	58.88	68.36
	kNN	-	70.49	38.56	79.98	84.89	67.32	68.33	84.02	61.45	69.38
Semi-sup.	Mean Teacher	✓	67.05	37.73	81.08	86.66	65.45	69.67	85.24	61.47	69.29
	ICT	✓	70.13	38.09	84.58	87.01	65.47	70.26	86.12	63.37	70.63
	Pseudo-Label	✓	66.26	37.49	78.60	83.71	64.46	67.49	82.94	60.06	67.63
	MPL	✓	67.61	37.47	77.85	83.70	64.51	67.08	82.39	59.65	67.53
	VIME-Semi	✓	65.13	37.32	80.53	87.13	65.39	64.80	82.83	52.08	66.90
Self-sup.	SubTab + Fine-tune	✓	66.01	37.60	67.80	75.40	66.69	56.46	75.34	55.62	62.62
	SubTab + LR	✓	70.12	37.67	73.25	86.07	64.92	61.34	82.14	58.90	66.80
	SubTab + kNN	-	71.91	39.51	69.56	83.60	68.79	59.87	80.13	61.57	66.87
	VIME + Fine-tune	✓	65.97	37.25	77.82	83.13	64.40	63.63	81.01	59.58	66.60
	VIME + LR	✓	67.80	37.51	82.87	87.42	64.29	71.53	86.79	69.62	70.98
	VIME + kNN	-	72.16	39.28	79.15	83.86	66.94	68.45	84.07	71.09	70.63
Unsup.-Meta.	UMTRA	-	65.78	38.05	67.28	73.29	64.41	35.90	51.32	25.08	52.64
	SES	-	68.27	39.04	74.80	78.46	66.61	52.74	74.80	52.25	63.37
	CACTUs	-	72.03	38.81	82.20	85.92	66.79	65.00	85.25	**81.52**	72.19
	STUNT (Ours)	-	**72.69**	**40.40**	**85.45**	**88.42**	**69.88**	**73.02**	**89.08**	79.18	**74.77**

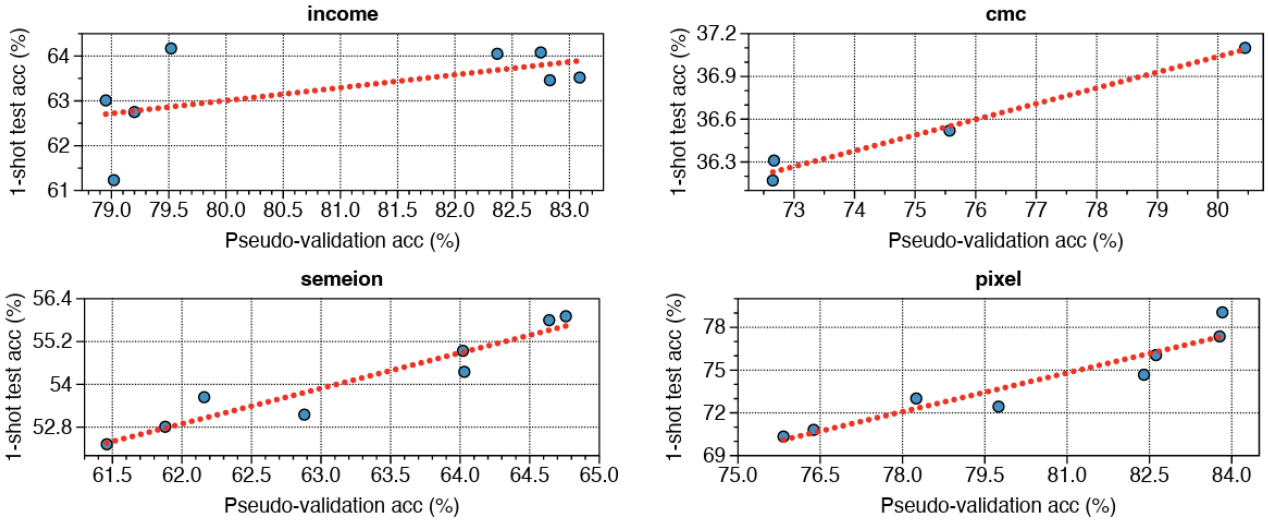
[3] Nam, J., Tack, J., Lee, K., Lee, H., & Shin, J. (2023). Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. arXiv preprint arXiv:2303.00918.

Predictive Learning

STUNT: Few-shot Tabular Learning with Self-generated Tasks from Unlabeled Tables

❖ Experiments

- 10 shot 상황에서도 성능 검증
- Pseudo-validation acc 와 1-shot test acc 가 양의 상관관계를 가짐을 확인
- 과적합 문제 완화 & 다양한 데이터셋에서 최적의 훈련 단계를 알 수 없는 상황에서도 모델의 조기 중단 시점을 합리적으로 결정 가능



10 Shot Performance Evaluation

Method \ Dataset	income	cmc	karhunen	optdigit	diabetes	semeion	pixel	dna	Avg.
kNN	**74.27**	41.07	85.63	87.44	71.32	74.64	87.52	71.15	74.13
ICT	71.56	38.00	**88.25**	**90.84**	67.63	74.67	**89.13**	69.55	73.70
VIME + LR	69.17	37.92	86.63	89.63	66.56	**77.66**	88.71	74.73	73.88
CACTUs	**73.63**	**42.14**	85.48	87.92	70.75	68.22	87.21	**84.40**	74.97
STUNT (Ours)	**74.08**	**42.01**	86.95	**89.91**	**72.82**	74.74	**89.90**	80.96	**76.42**

Early stopping performance with the pseudo-validation set

	income		cmc		semeion		pixel	
Problem	Last	Early	Last	Early	Last	Early	Last	Early
1-shot	61.58	**63.52**	36.94	**37.10**	51.94	**55.91**	74.92	**79.05**
5-shot	70.84	**72.69**	**40.43**	40.40	71.55	**73.02**	87.60	**89.08**

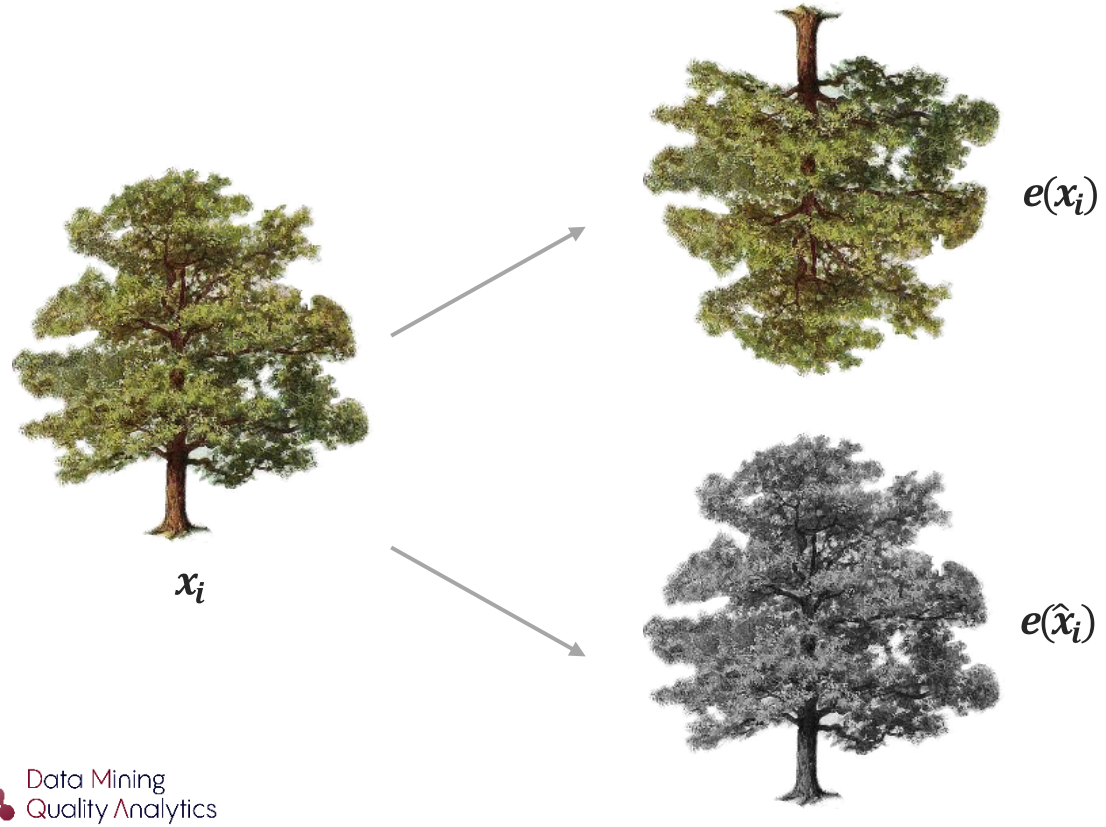
[3] Nam, J., Tack, J., Lee, K., Lee, H., & Shin, J. (2023). Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. arXiv preprint arXiv:2303.00918.

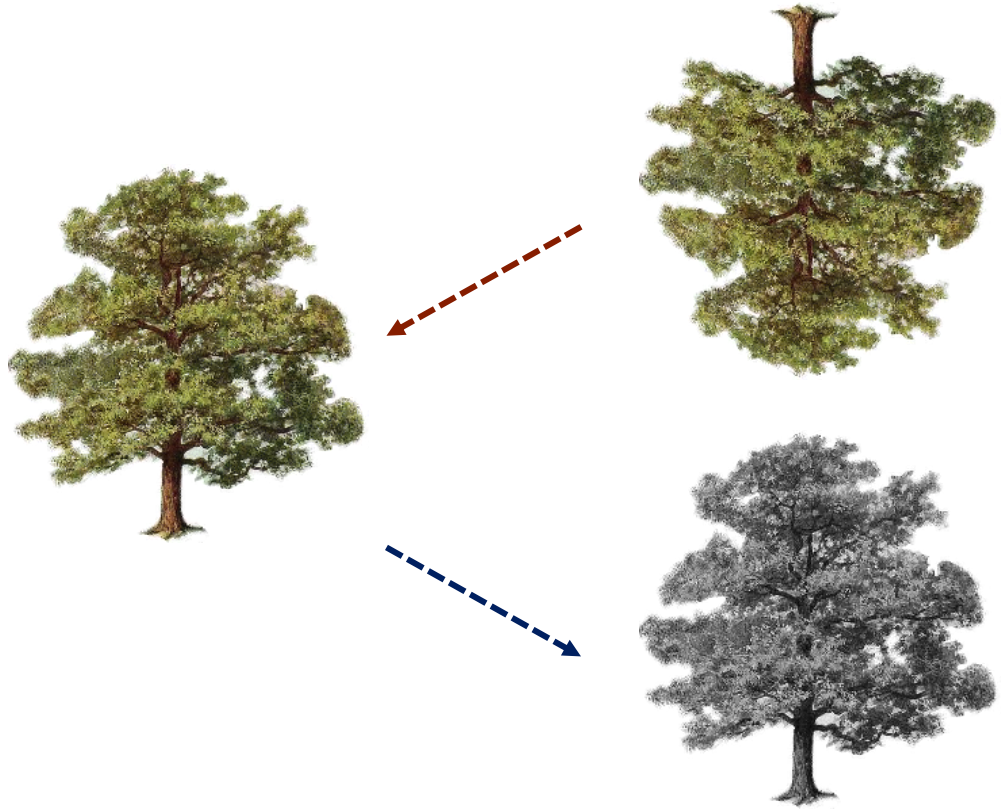
3. Contrastive Learning

Contrastive Learning

❖ Contrastive Learning Approach

- 같은 input 에 대해서 다른 views, corruptions 를 통해 robust 표현벡터 학습
- 유사한 instance 간의 유사성을 극대화하고 유사하지 않은 instance 들은 멀리 위치하도록 학습
- $\mathcal{L}_{contrastive} = \phi(e(x_i), e(\hat{x}_i))$

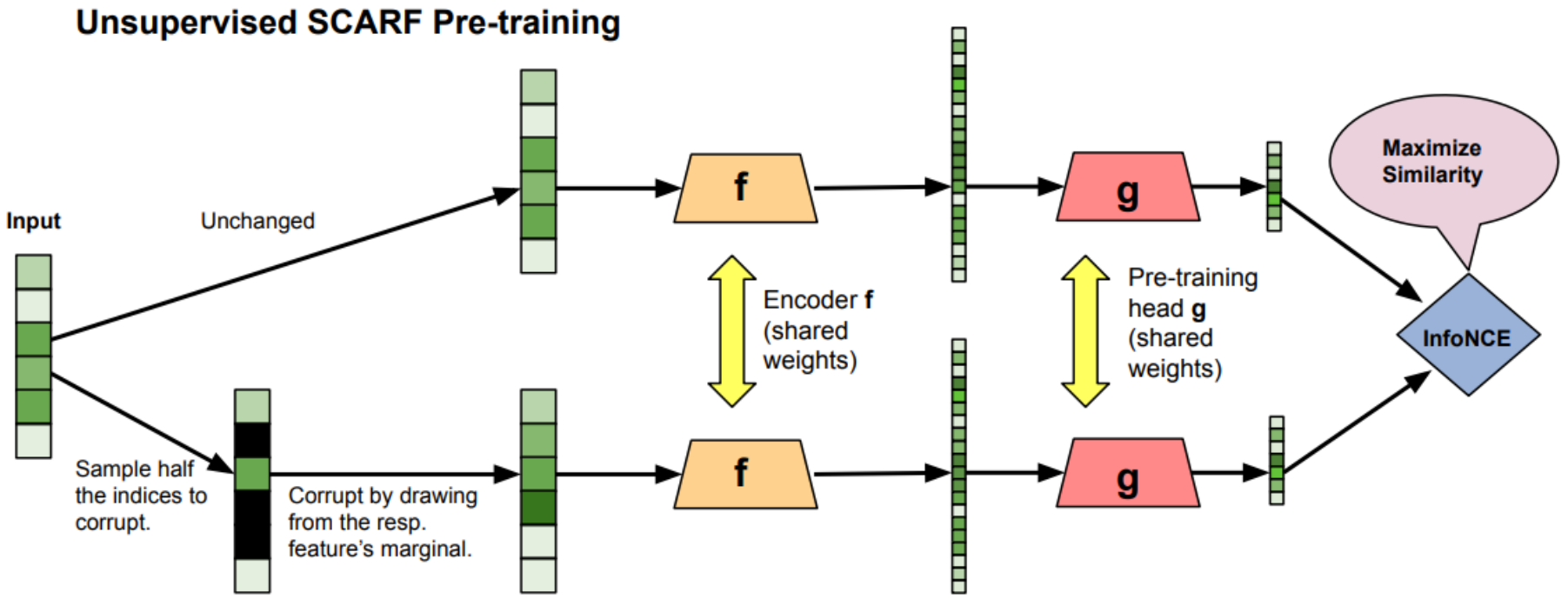




Contrastive Learning

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

- MLP 기반 framework로, contrastive learning pre-training & supervised fine-tuning 두 단계로 구성
- 주어진 input에 대해 무작위 부분집합만큼 손상되고 해당 특성들의 주변 분포에서 무작위 view 로 대체
- InfoNCE loss function을 활용하여 sample 과 해당 sample 의 변형은 가까워지도록, 다른 sample 의 변형은 멀어지도록 대조학습 진행

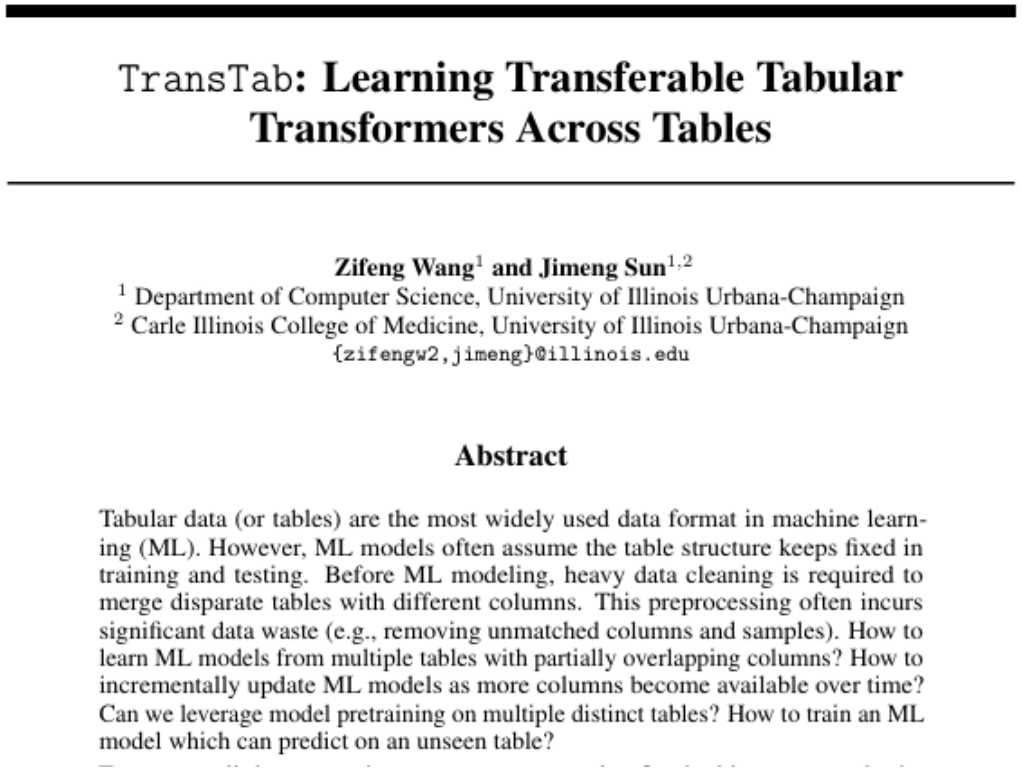


[4] Bahri, D., Jiang, H., Tay, Y., & Metzler, D. (2021). Scarf: Self-supervised contrastive learning using random feature corruption. arXiv preprint arXiv:2106.15147.

Contrastive Learning

❖ TransTab: Learning Transferable Tabular Transformers Across Tables (NeurIPS, 2022)

- Transfer learning, feature incremental learning, zero-shot inference 를 위한 다양한 열의 맥락을 학습하는데 목표
- Transferable Transformers for Tabular analysis (TransTab) 을 도입하여 고정된 테이블 구조를 완화하는 방법론 제안



TransTab: Learning Transferable Tabular Transformers Across Tables

Zifeng Wang[1] and Jimeng Sun[1,2]
[1] Department of Computer Science, University of Illinois Urbana-Champaign
[2] Carle Illinois College of Medicine, University of Illinois Urbana-Champaign
{zifengw2,jimeng}@illinois.edu

Abstract

Tabular data (or tables) are the most widely used data format in machine learning (ML). However, ML models often assume the table structure keeps fixed in training and testing. Before ML modeling, heavy data cleaning is required to merge disparate tables with different columns. This preprocessing often incurs significant data waste (e.g., removing unmatched columns and samples). How to learn ML models from multiple tables with partially overlapping columns? How to incrementally update ML models as more columns become available over time? Can we leverage model pretraining on multiple distinct tables? How to train an ML model which can predict on an unseen table?

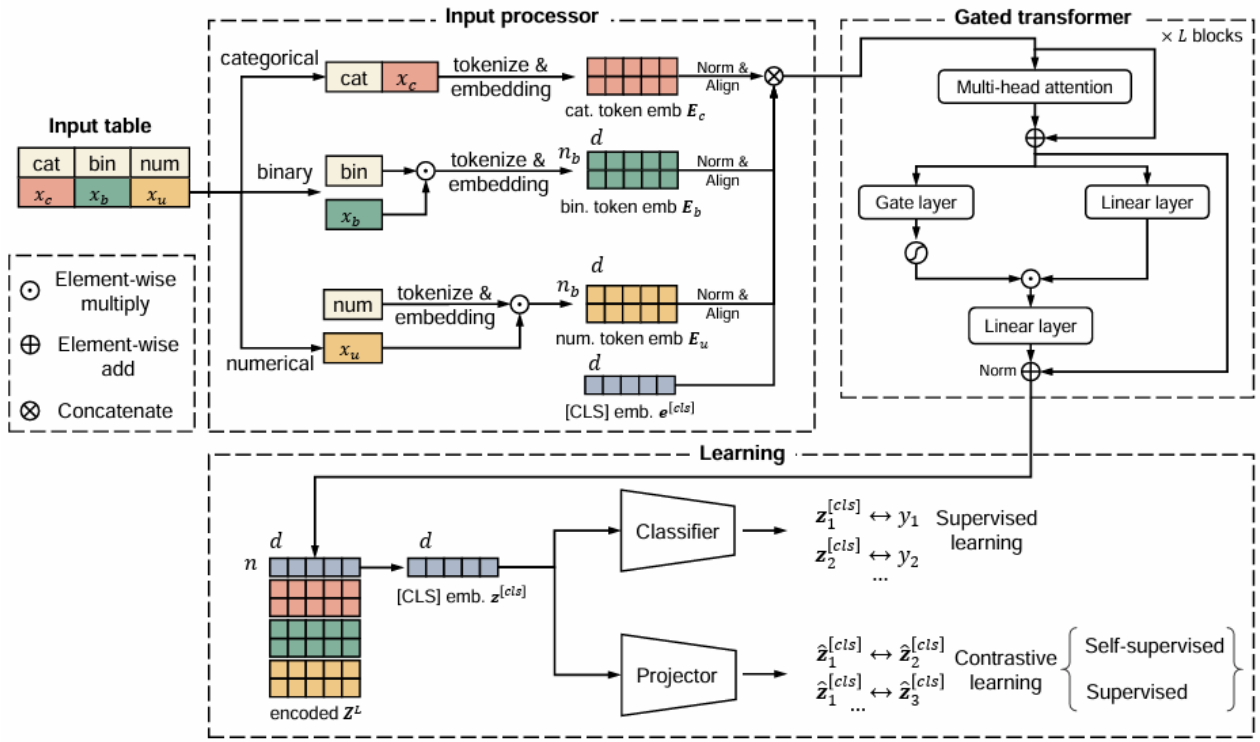
[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

Contrastive Learning

TransTab: Transferable Transformers for Tabular analysis

❖ Key Components of TransTab

- Input processor : tabular input을 token-level embedding 변환
- Gated transformer : 스택된 gated transformer 층을 통한 token-level embedding의 추가 인코딩
- Learning module : labeled 데이터에 대한 classifier와 contrastive learning을 위한 projector 포함



[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

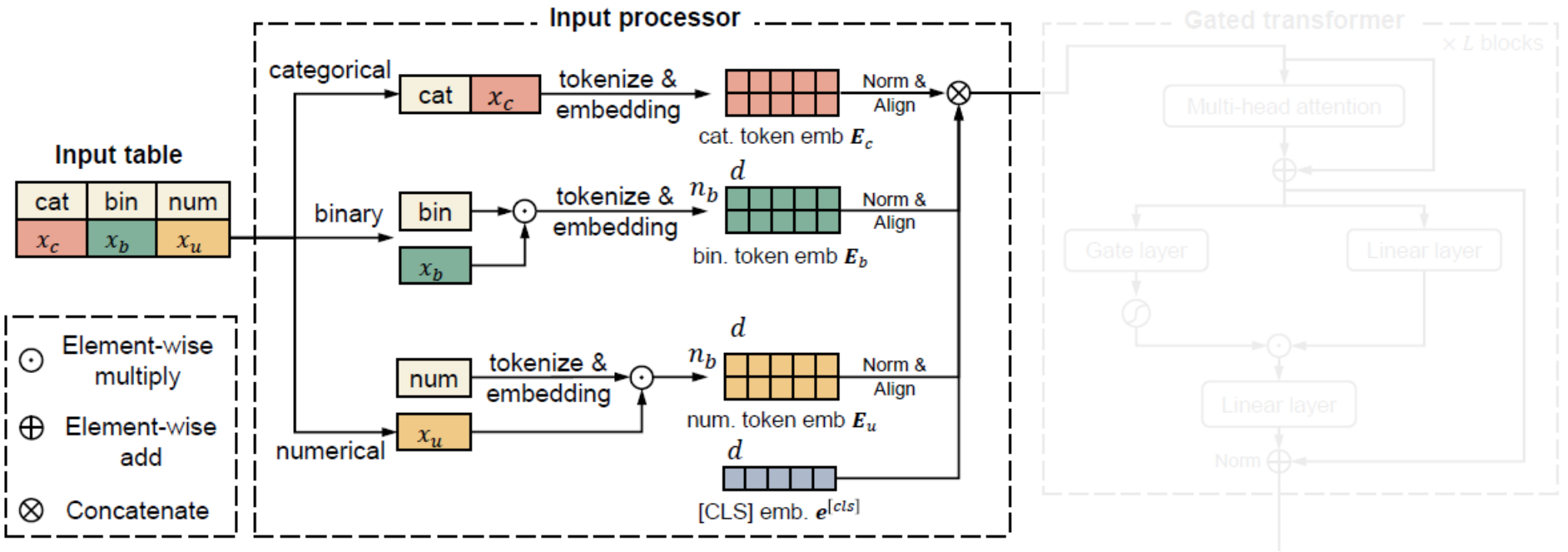
Contrastive Learning

TransTab: Transferable Transformers for Tabular analysis

❖ Input Processor for Columns and Cells

- Variable-column tables 수용 & 데이터 간 지식 유지를 위해 input processor 구축
- 각 column의 cell (데이터)을 의미론적으로 인코딩된 토큰의 시퀀스로 변환
- ex) 'Weight' column의 경우 값이 60일때, 60 years old가 아닌 60kg 의미

  → 열 이름을 모델링에 포함시켜 총 4가지 타입 categorical/ text (cat) & binary (bin) & numerical 처리



[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

Contrastive Learning

TransTab: Transferable Transformers for Tabular analysis

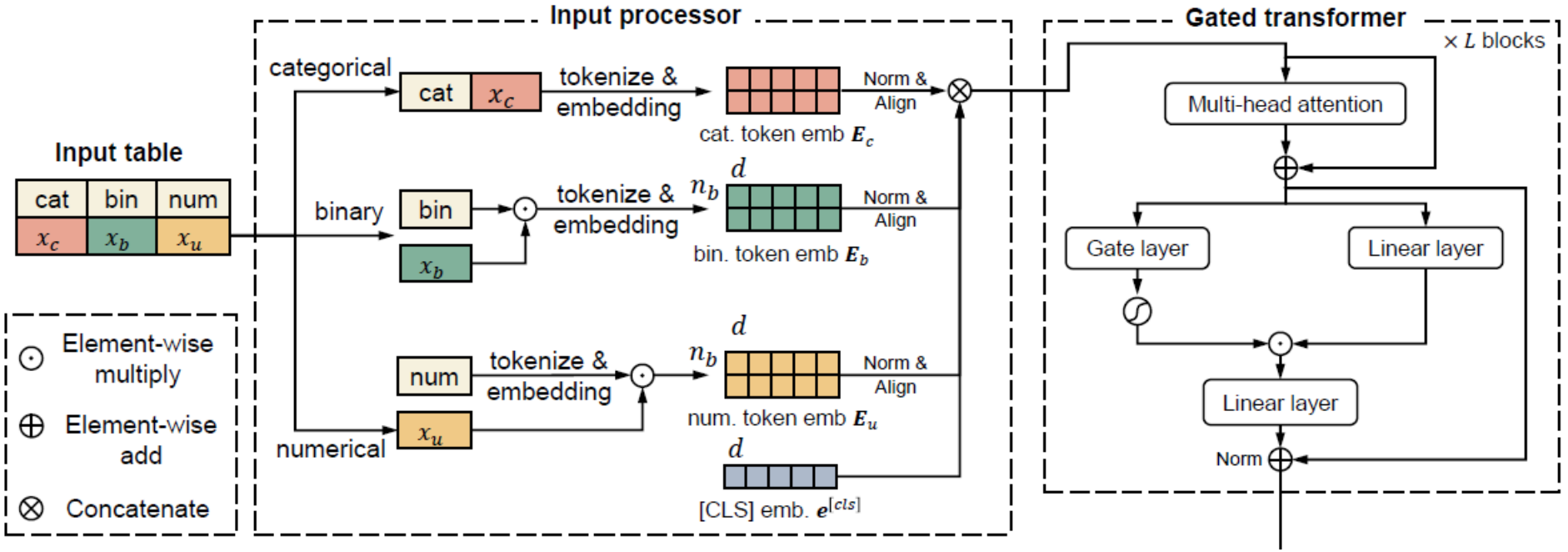
❖ Gated Transformers

- Self-attention layer & feedforward layer로 구성된 NLP의 classical transformer 적용
- Features 간의 상호작용 탐색을 위해 l^{th} input representation Z_l 먼저 선택
- Z_l^{att} 는 token-wise gating layer에 의해 추가로 변환, Z_{l+1} 을 얻기 위해 linear output 변환
- Token에 대한 attention을 재분배함으로써 중요한 features에 집중하도록 학습

$$Z_{att}^{l} = \text{MultiHeadAttn}(Z^{l}) = [head_1, head_2, \ldots, head_h]W^{o},$$

$$head_i = \text{Attention}(Z^{l}W_i^{Q}, Z^{l}W_i^{K}, Z^{l}W_i^{V}),$$

$$Z^{l+1} = \text{Linear}((g^{l} \odot Z_{att}^{l}) \oplus \text{Linear}(Z_{att}^{l}))$$



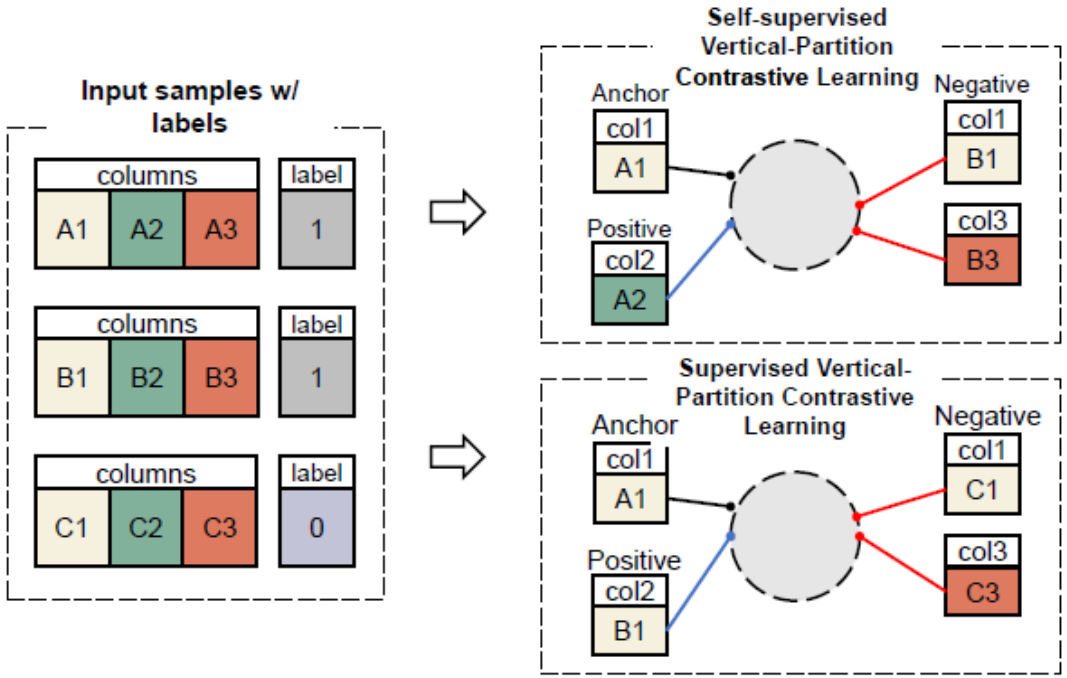
[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

Contrastive Learning

TransTab: Transferable Transformers for Tabular analysis

❖ Self-supervised and supervised pretraining of TransTab

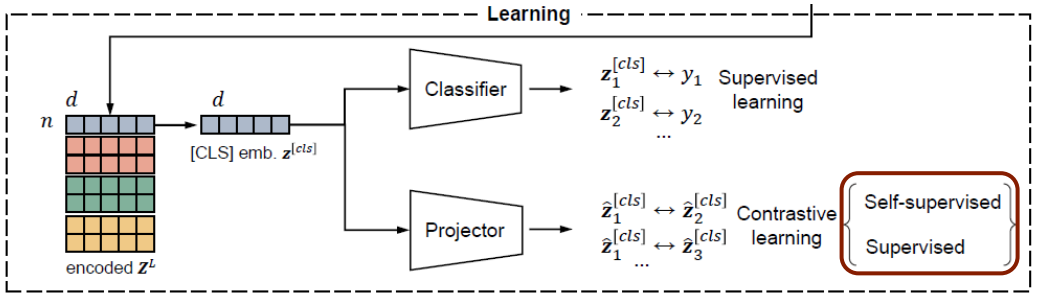
- 대부분의 tabular SSL 방법론은 고정된 column 전체에서 작동하여 높은 계산 비용 & 과적합 위험
- TransTab에서는 view-invariant factors를 학습하기 위해 tabular 수직 분할 사용하여 positive, negative samples 구성
- Columns의 부분집합을 선택하여 동일한 sample의 partition을 positive, 다른 sample의 partition을 negative 로 설정



Loss function of Self-VPCL

$$\ell(K) = -\sum_{i=1}^{B}\sum_{k=1}^{K}\sum_{k' \neq k}^{K} log \frac{\exp \psi(v_i^k, v_i^{k'})}{\sum_{j=1}^{B}\sum_{k^\dagger=1}^{K} \exp \psi(v_i^k, v_j^{k^\dagger})},$$

사전학습 데이터에 labeled 가 있는 경우에 사용하는 Supervised Contrastive Learning 같이 제안



[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

4. MambaTab

MambaTab

❖ MambaTab: A Simple Yet Effective Approach for Handling Tabular Data (arXiv, 2024)

- 구조화된 state-space model (SSM)을 기반으로 tabular 데이터에 적합한 방법론 제안
- Table의 end-to-end 지도학습 방식인 SSM 모델의 변형인 Mamba 활용

MambaTab: A Simple Yet Effective Approach for Handling Tabular Data

Md Atik Ahamed[1], **Qiang Cheng**[1,2*]

[1]Department of Computer Science, University of Kentucky, Lexington, KY, USA

[2]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY, USA

{atikahamed, qiang.cheng}@uky.edu

Abstract

Tabular data remains ubiquitous across domains despite growing use of images and texts for machine learning. While deep learning models like convolutional neural networks and transformers achieve strong performance on tabular data, they require extensive data preprocessing, tuning, and resources, limiting accessibility and scalability. This work develops an innovative approach based on a structured state-space model (SSM), MambaTab, for tabular data. SSMs have strong capabilities for efficiently extracting effective representations from data with long-range dependencies. MambaTab leverages Mamba, an emerging SSM variant, for end-to-end supervised learning on tables. Compared to state-of-the-art baselines, MambaTab delivers superior performance while requiring significantly fewer parameters and minimal preprocessing, as empirically validated on diverse benchmark datasets. MambaTab's efficiency, scalability, generalizability, and predictive gains signify it as a lightweight, "out-of-the-box" solution for diverse tabular data with promise for enabling wider practical applications.

models for tabular data, thereby impeding their wider application. Moreover, almost all existing tabular learning methods, except TransTab [Wang and Sun, 2022], operate under vanilla supervised learning, requiring identical train and test table structures. They are not well-suited for feature incremental learning where features are sequentially added; under such a setting, they have to either drop new features or old data, leading to insufficient use of training data. It is desirable to have the ability to continuously learn from new features.

To address these challenges, we introduce a new approach for tabular data based on structured state-space models (SSMs) [Gu *et al.*, 2021b] [Gu *et al.*, 2021a] [Fu *et al.*, 2022]. These models can be interpreted as a combination of CNNs and recursive neural networks, having advantages of both types of models. They offer parameter efficiency, scalability, and strong capabilities for learning representations from varied data, particularly for sequential data with long-range dependencies. To tap into these potential advantages, we leverage SSMs as an alternative to CNNs or Transformers for modeling tabular data.

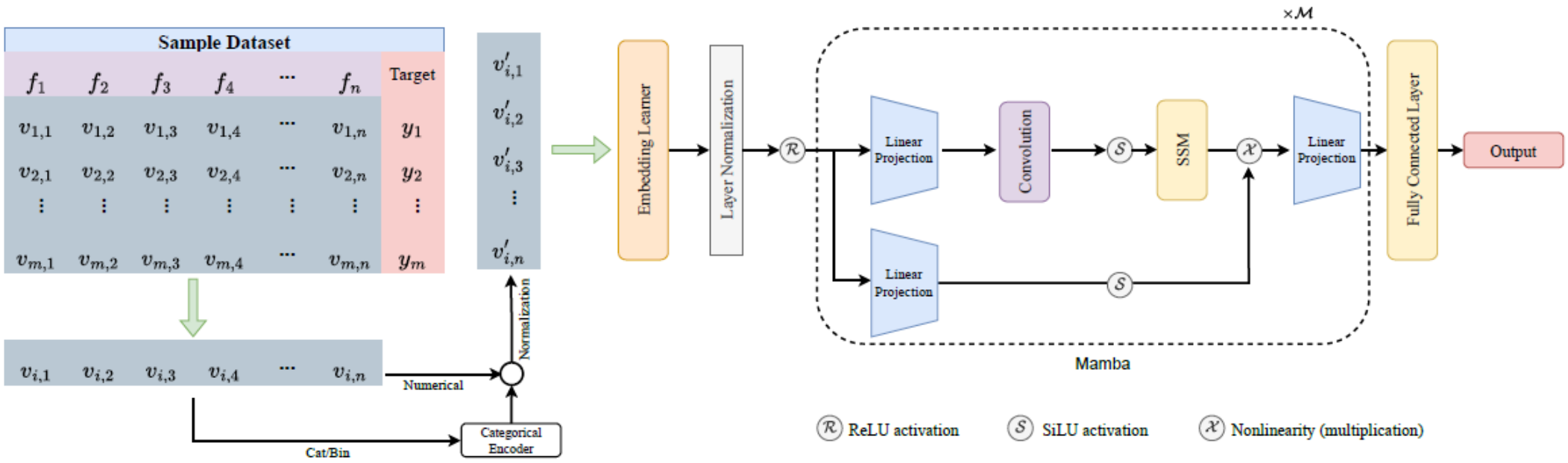
Specifically, we leverage Mamba [Gu and Dao, 2023], an emerging SSM variant, as a critical building block to build a new supervised model for tabular data called *MambaTab*. This proposed model has several key advantages over exist-

[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

MambaTab

❖ MambaTab: A Simple Yet Effective Approach for Handling Tabular Data

- Data preprocessing & Embedding representation learning
- Cascading Mamba blocks
- Output prediction

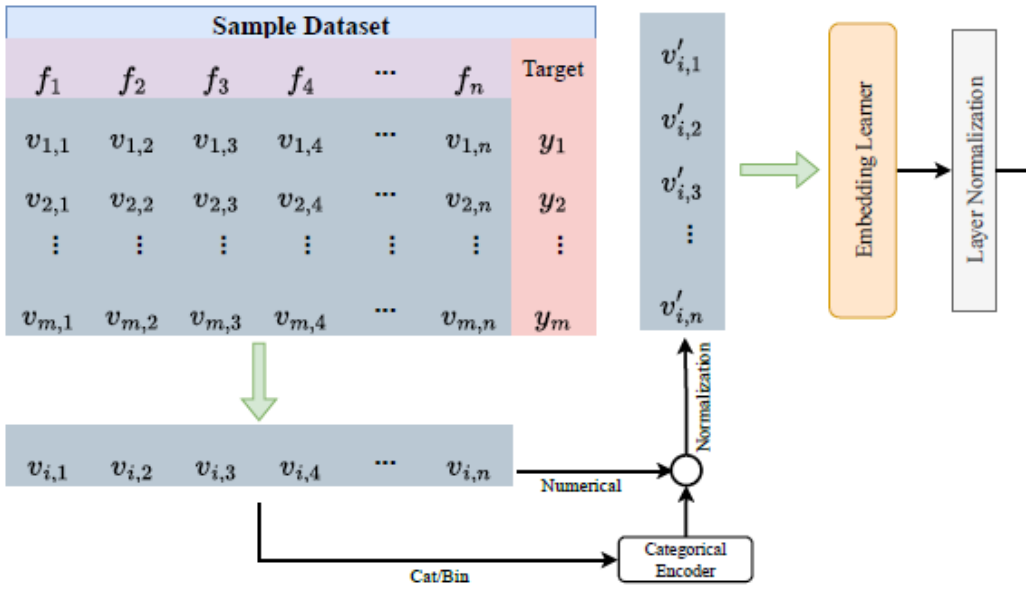


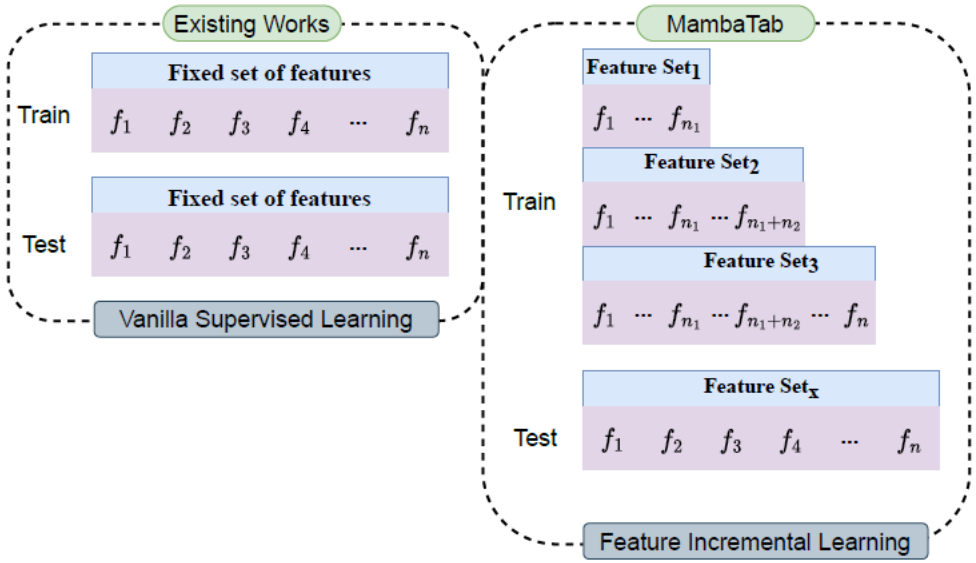
[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

MambaTab

❖ Data preprocessing & Embedding representation learning

- TransTab과 달리 column의 유형을 수동으로 식별하지 않고 데이터 전처리 과정 간소화 & 자동화
- 범주형 feature의 순서를 강제하는 대신, feature의 직접적인 multi-dimensional features 학습함으로써 더 유연한 학습 가능
- Incremental learning을 통해 하류 Mamba block이 동일한 입력 특성 차원을 갖도록 보장
  - ✓ 고정된 feature set에 대해 구애 받지 않고 점진적으로 추가되는 특성을 학습하여 가중치를 전달



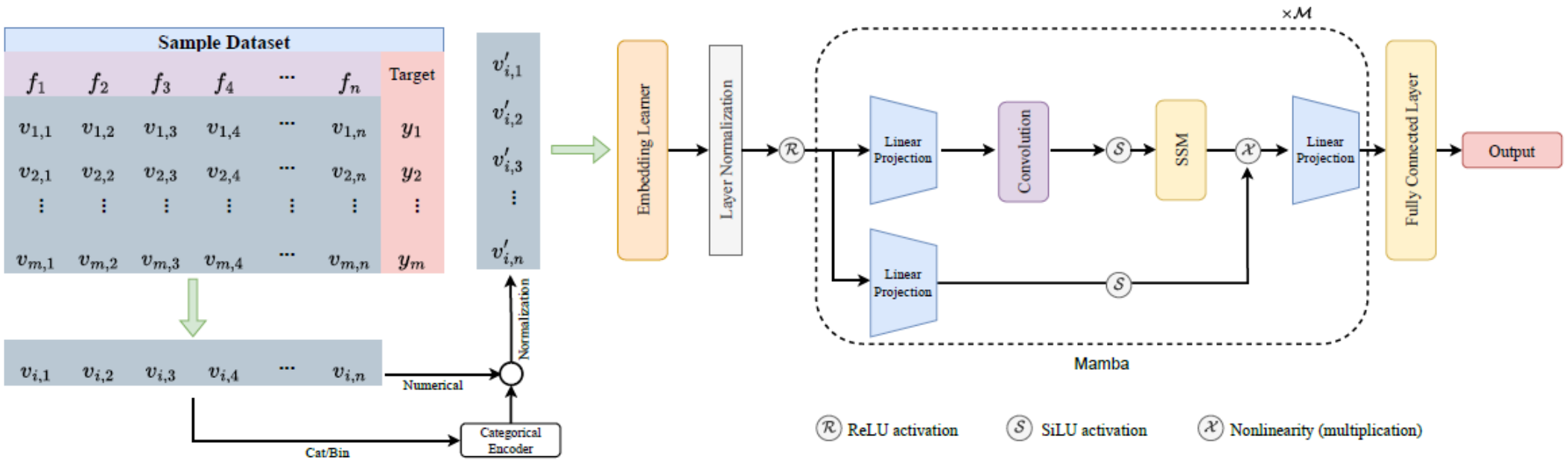


[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

MambaTab

❖ Cascading Mamba blocks

- 계층 정규화와 ReLU 활성화함수를 거친 embedded representation의 처리 담당
- Feature의 배치, 길이, 차원 매핑을 통해 정보 전달 목표
- SSM (State Space Model)의 적용으로 인해 입력 token sequence에 따른 의존적 추론 가능성 제공



[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

MambaTab

❖ Experiments

- 지도학습에서 hyperparameter tuning 후에는 모든 dataset에 대해 최고 성능 달성
- Dataset의 feature 집합을 3개의 non-overlapping subset으로 나눈 후 각 set에 대해 incremental learning 진행
  - ✓ TransTab, MambaTab만 incremental learning / 다른 모델들은 한 set에 대해서만 학습 가능
  - ✓ Hyperparameter tuning을 통해 성능 향상 가능

Vanilla Supervised Learning

Methods	Datasets							
	CG	CA	DS	AD	CB	BL	IO	IC
LR	0.720	0.836	0.557	0.851	0.748	0.801	0.769	0.860
XGBoost	0.726	0.895	0.587	0.912	0.892	0.821	0.758	**0.925**
MLP	0.643	0.832	0.568	0.904	0.613	0.832	0.779	0.893
SNN	0.641	0.880	0.540	0.902	0.621	0.834	0.794	0.892
TabNet	0.585	0.800	0.478	0.904	0.680	0.819	0.742	0.896
DCN	0.739	0.870	0.674	0.913	0.848	0.840	0.768	0.915
AutoInt	0.744	0.866	0.672	0.913	0.808	0.844	0.762	0.916
TabTrans	0.718	0.860	0.648	**0.914**	0.855	0.820	0.794	0.882
FT-Trans	0.739	0.859	0.657	0.913	0.862	0.841	0.793	0.915
VIME	0.735	0.852	0.485	0.912	0.769	0.837	0.786	0.908
SCARF	0.733	0.861	0.663	0.911	0.719	0.833	0.758	0.905
TransTab	0.768	0.881	0.643	0.907	0.851	0.845	**0.822**	0.919
MambaTab-D	0.771	0.954	0.643	0.906	0.862	0.852	0.785	0.906
MambaTab-T	**0.801**	**0.963**	**0.681**	**0.914**	**0.896**	**0.854**	0.812	0.920

Feature Incremental Learning

Methods	Datasets							
	CG	CA	DS	AD	CB	BL	IO	IC
LR	0.670	0.773	0.475	0.832	0.727	0.806	0.655	0.825
XGBoost	0.608	0.817	0.527	0.891	0.778	0.816	0.692	0.898
MLP	0.586	0.676	0.516	0.890	0.631	0.825	0.626	0.885
SNN	0.583	0.738	0.442	0.888	0.644	0.818	0.643	0.881
TabNet	0.573	0.689	0.419	0.886	0.571	0.837	0.680	0.882
DCN	0.674	0.835	0.578	0.893	0.778	0.840	0.660	0.891
AutoInt	0.671	0.825	0.563	0.893	0.769	0.836	0.676	0.887
TabTrans	0.653	0.732	0.584	0.856	0.784	0.792	0.674	0.828
FT-Trans	0.662	0.824	0.626	0.892	0.768	0.840	0.645	0.889
VIME	0.621	0.697	0.571	0.892	0.769	0.803	0.683	0.881
SCARF	0.651	0.753	0.556	0.891	0.703	0.829	0.680	0.887
TransTab	0.741	0.879	0.665	0.894	0.791	0.841	0.739	0.897
MambaTab-D	**0.787**	**0.961**	**0.669**	**0.904**	**0.860**	**0.853**	**0.783**	**0.908**

[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

5. Conclusion

Conclusions

What is Next for Tabular Data?

❖ Predictive Learning

- 모델이 주어진 predictive task를 수행하며 upstream 데이터에서 표현벡터 학습
- VIME : mask vector estimator – feature vector estimator / STUNT : self-generated task

❖ Contrastive Learning

- 동일한 instance에 대해 다양한 관점과 변형을 통해 강건한 표현벡터를 학습
- SCARF: corruption & InfoNCE / TransTab: self-supervised vertical-partition

❖ MambaTab

- Manual data preprocessing 필요 없음
- Transformer-based 방법론들에 비해 적은 memory 사용
- Efficiency & scalability & generalizability

6. Reference

References

[1] Wang, W. Y., Du, W. W., Xu, D., Wang, W., & Peng, W. C. (2024). A Survey on Self-Supervised Learning for Non-Sequential Tabular Data. arXiv preprint arXiv:2402.01204.

[2] Yoon, J., Zhang, Y., Jordon, J., & Van der Schaar, M. (2020). Vime: Extending the success of self-and semi-supervised learning to tabular domain. Advances in Neural Information Processing Systems, 33, 11033-11043.

[3] Nam, J., Tack, J., Lee, K., Lee, H., & Shin, J. (2023). Stunt: Few-shot tabular learning with self-generated tasks from unlabeled tables. arXiv preprint arXiv:2303.00918.

[4] Bahri, D., Jiang, H., Tay, Y., & Metzler, D. (2021). Scarf: Self-supervised contrastive learning using random feature corruption. arXiv preprint arXiv:2106.15147.

[5] Wang, Z., & Sun, J. (2022). Transtab: Learning transferable tabular transformers across tables. Advances in Neural Information Processing Systems, 35, 2902-2915.

[6] Ahamed, M. A., & Cheng, Q. (2024). MambaTab: A Simple Yet Effective Approach for Handling Tabular Data. arXiv preprint arXiv:2401.08867.

Thank You