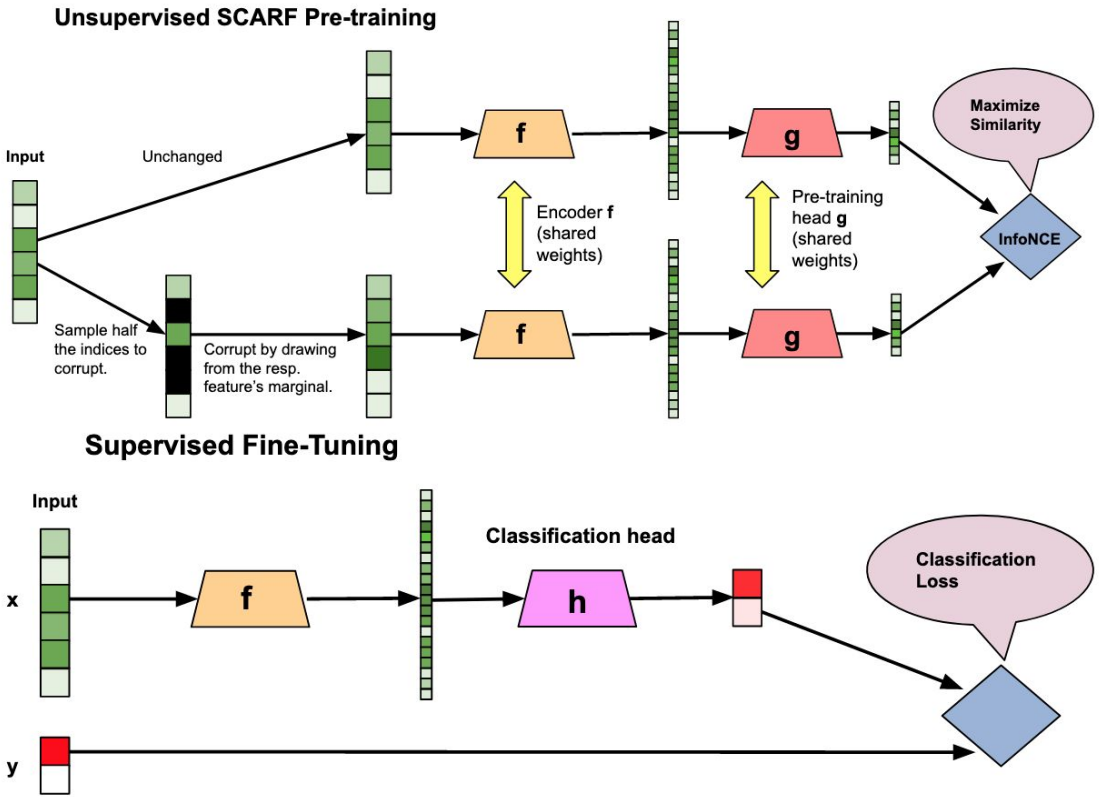
SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption

Котельников Аким
Колесников Георгий
Коган Александра
Михненко Наталья

What is SCARF?



Preprocessing, algorithm, loss

- ***g*** and ***h*** are 2-layer FCN with 256 hidden dim. hidden size, ***f*** is 4-layer
- Categorical features – OHE. Numerical – z-normalization
- Corruption rate 0.6

Algorithm 1 SCARF pre-training algorithm.

input: unlabeled training data $\mathcal{X} \subseteq \mathbb{R}^M$, batch size N, temperature τ, corruption rate c, encoder network f, pre-train head network g.
let $\widehat{\mathcal{X}_j}$ be the uniform distribution over $\mathcal{X}_j = \{x_j : x \in \mathcal{X}\}$, where x_j denotes the j-th coordinate of x.
let $q = \lfloor c \cdot M \rfloor$ be the number of features to corrupt.
for sampled mini-batch $\left\{x^{(i)}\right\}_{i=1}^{N} \subseteq \mathcal{X}$ **do**
    for $i \in [N]$, uniformly sample subset $\mathcal{I}_i$ from $\{1, ..., M\}$ of size q and define $\tilde{x}^{(i)} \in \mathbb{R}^M$ as follows: $\tilde{x}_j^{(i)} = x_j$ if $j \notin \mathcal{I}_i$, otherwise $\tilde{x}_j^{(i)} = v$, where $v \sim \widehat{\mathcal{X}_j}$. # generate corrupted view.
    let $z^{(i)} = g\left(f\left(x^{(i)}\right)\right), \tilde{z}^{(i)} = g\left(f\left(\tilde{x}^{(i)}\right)\right)$, for $i \in [N]$. # embeddings for views.
    let $s_{i,j} = {z^{(i)}}^{\top} \tilde{z}^{(j)} / \left(\|z^{(i)}\|_2 \cdot \|\tilde{z}^{(j)}\|_2\right)$, for $i, j \in [N]$. # pairwise similarity.
    define $\mathcal{L}_{\text{cont}} := \frac{1}{N} \sum_{i=1}^{N} -\log\left(\frac{\exp(s_{i,i}/\tau)}{\frac{1}{N}\sum_{k=1}^{N} \exp(s_{i,k}/\tau)}\right)$.
    update networks f and g to minimize $\mathcal{L}_{\text{cont}}$ using SGD.
end for
return encoder network f.

Experiments

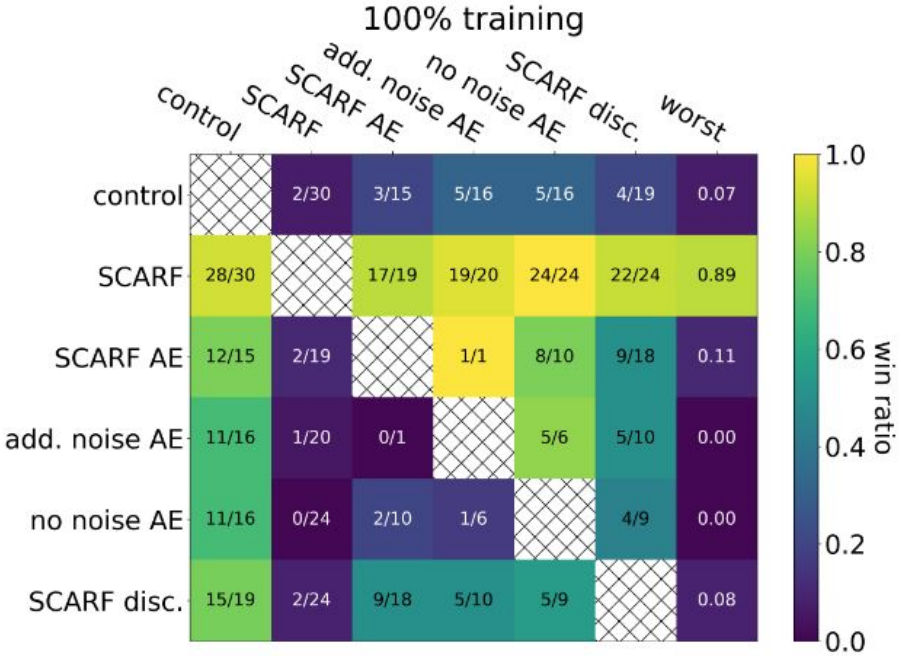
- 69 OpenML datasets
- Full dataset; dataset where 25% have labels; 30% label corruption;
- All runs repeated 30 times with different splits
- Pre-train 1000 epochs, finetune 200

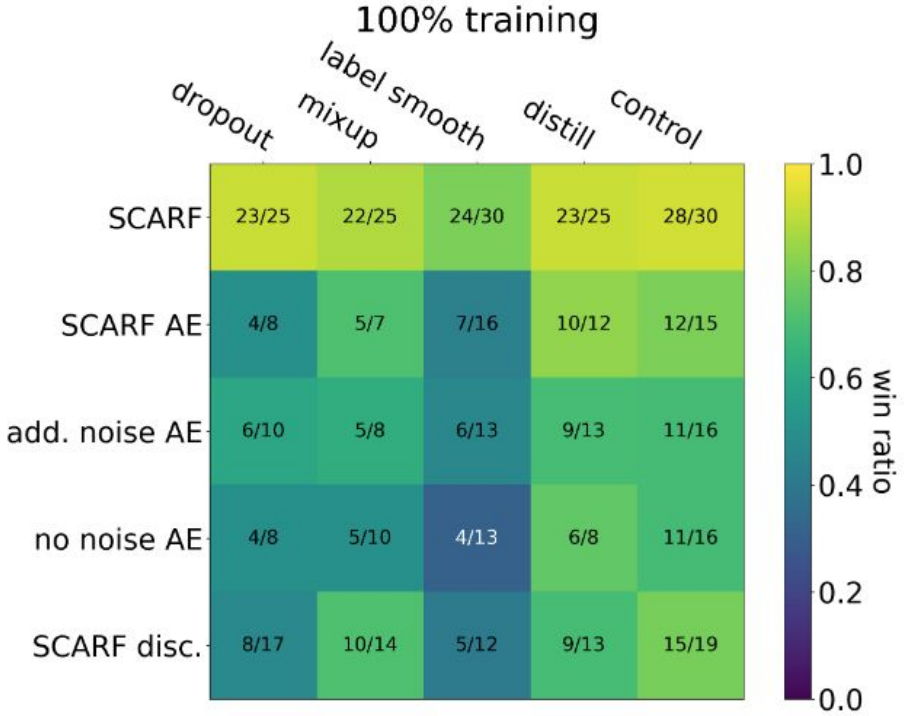
Evaluation method: **“Win matrix”**

$$W_{i,j} = \frac{\sum_{d=1}^{69} \mathbb{1}[\text{method } i \text{ beats } j \text{ on dataset } d]}{\sum_{d=1}^{69} \mathbb{1}[\text{method } i \text{ beats } j \text{ on dataset } d] + \mathbb{1}[\text{method } i \text{ loses to } j \text{ on dataset } d]}.$$

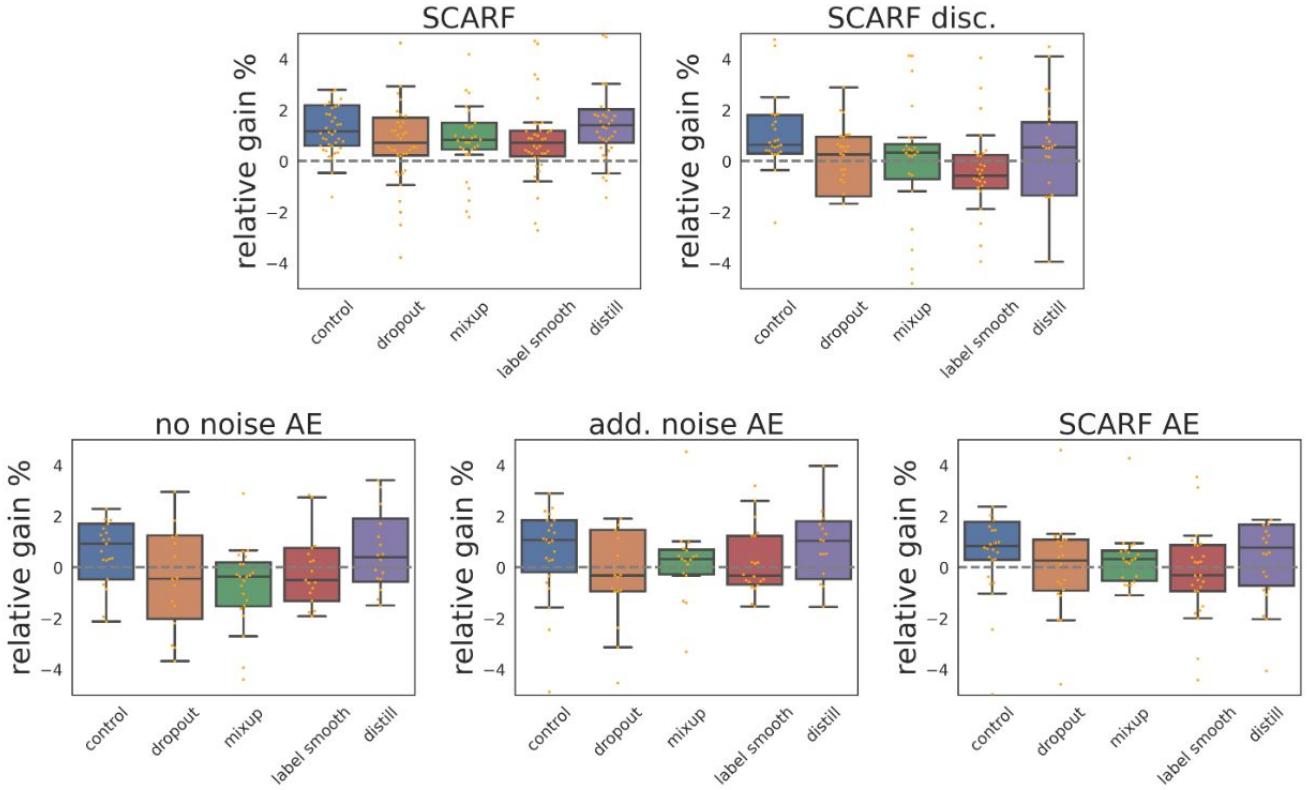
“Beats” and “loses” are only defined when the means are not a statistical tie (using Welch’s t-test with unequal variance and a p-value of 0.05)

Experiments. Full dataset





Experiments. Full dataset



Experiments. Batch size and corruption rate

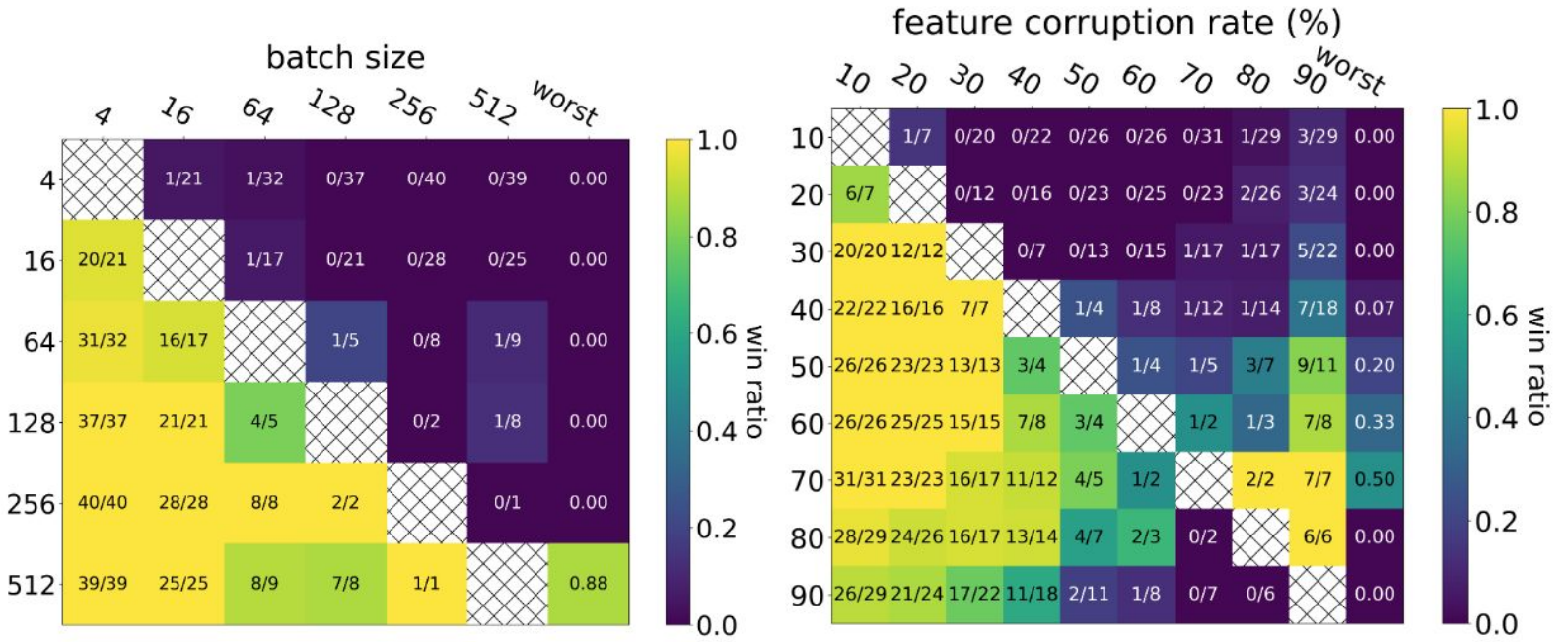
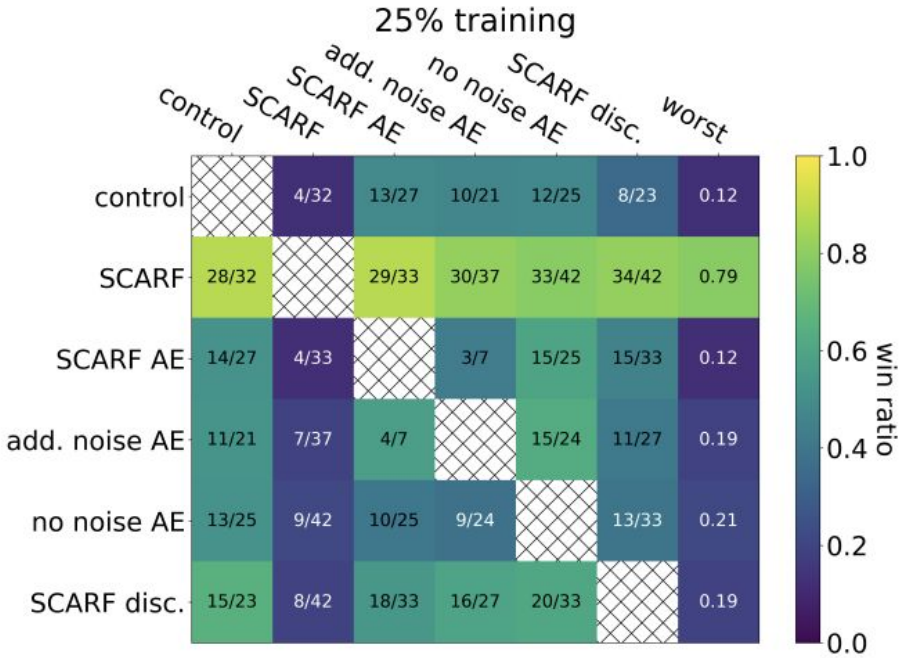
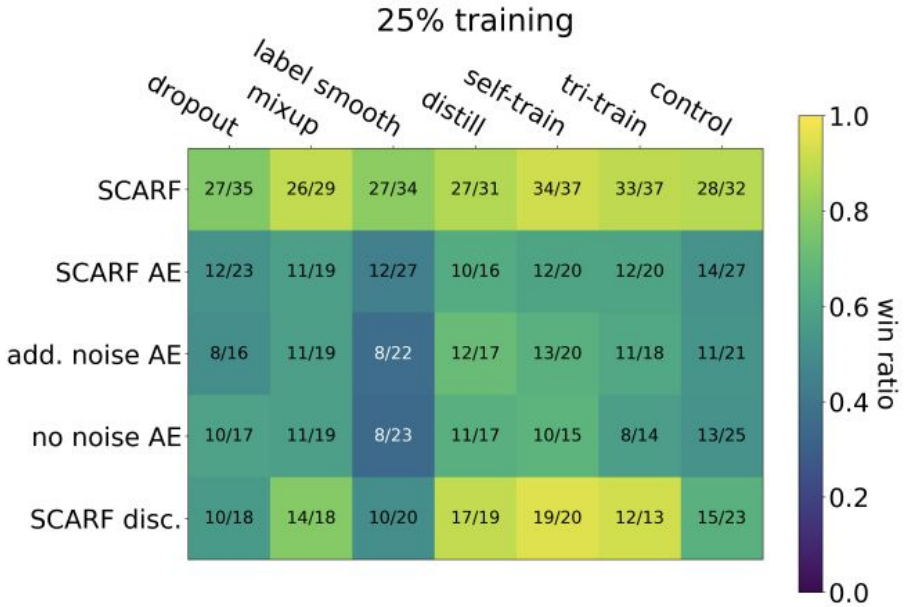


Figure 6: Win matrix for various batch sizes (**Left**) and corruption rates (**Right**) for the fully labeled, noiseless setting.

Experiments. 25% training





Рецензия

Новизна: Авторы предлагают новый метод аугментации для self-supervised обучения путем случайного искажения признаков.

Значимость: Получен универсальный метод улучшающий результаты машинного обучения на таблицах, обгоняющий по своим показателям существующие бейзлайны.

Обоснованность: Метод проверен на 69 наборах реальных табличных данных, запущенных по 30 раз.

Предложения к улучшению: Сравнение с методом BYOL и, возможно, больше абсолютных (а не относительных) показателей.

План практика-исследователя (гуглера)

- Краткая информация о статье
- Рассказ об авторах
- Рассказ про несколько статей, которые цитируются в данной
- Рассказ про несколько статей, которые цитируют данную
- ?Предложение дальнейших путей развития
- ?Возможное применение статьи в индустрии
- Шутка, связанная с названием статьи (“что-то любопытное” из формы)

Исследование контекста статьи SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption

Гуглер Коган Александра

Основная информация

Что? SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption

и Кто? Dara Bahri, Heinrich Jiang, Yi Tay, Donald Metzler

Где? ICLR 2022 (spotlight presentation)

Когда? 29 Sept 2021 (modified: 23 Nov 2021)

Google Research

Авторы

Dara Bahri
Research Scientist
в Google Research

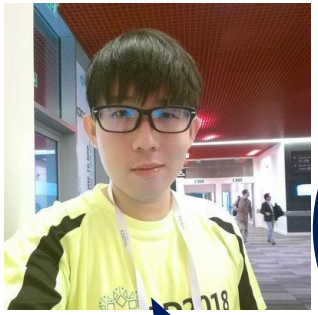
37 работ

Про: трансформеры, attention

Heinrich Jiang
Research Scientist
в Google Research

39 работ

Про: кластеризация и классификация



Yi Tay
Research Scientist
в Google Research

86 работ

Про: рекомендации, трансформеры

Donald Metzler
Senior Staff Software Engineer в Google

138 работ

Про: markov random fields

Общие статьи

Bahri	Jiang	Tay	Metzler	#papers
■	■	■	■	2
	■	■	■	2
■		■	■	21
■	■		■	2
■	■	■		2
		■	■	24
	■		■	2
	■	■		2
■			■	21
■		■		25
■	■			14

Label Smoothed Embedding Hypothesis for Out-of-Distribution Detection – вторая общая статья этого коллектива

Кого цитируют

94 цитирования

- Одно «самоцитирование»: Deep k-nn for noisy labels (D Bahri, H Jiang, M Gupta)

- Из схожего: Self-supervised Learning for Large-scale Item Recommendations (отличие: masking random features и dropout)

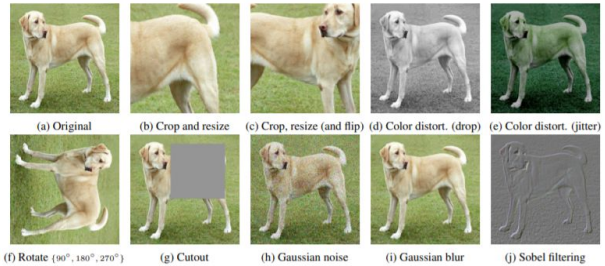
Идейно сравниваются с

NLP

token masking
(MPNet: Masked and Permuted Pre-training for Language Understanding)

CV

color distortion
(Colorful Image Colorization)



cropping
(A Simple Framework for Contrastive Learning of Visual Representations)

Кто цитирует

1 цитирование: Deep Neural Networks and Tabular Data: A Survey

	Method	Interpr.	Key Characteristics
Encoding	SuperTML Sun et al. (2019)		Transform tabular data into images for CNNs
	VIME Yoon et al. (2020)	✓	Self-supervised learning and contextual embedding
	IGTD		Transform tabular data into images for CNNs
	SCARF Bahri et al. (2021)		Self-supervised contrastive learning
	Wide&Deep		Embedding layer for categorical features

Схожие работы

Данная работа: SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption

randomizing random features based on the features' respective marginal training distribution

Схожая, которую цитировали: Self-supervised Learning for Large-scale Item Recommendations

masking random features in a correlated manner and applying a **dropout** for categorical features

Конкурент: SubTab: Subsetting Features of Tabular Data forSelf-Supervised Representation Learning

learning from tabular data into a multi-view representation learning problem by **dividing** the input **features** to **multiple subsets**

Дальнейшая судьба

Направления дальнейшей работы:

- Сравнение с другими моделями, например с SubTab (подсчет accuracy)
- Возможно, можно как-нибудь использовать лейблы и в схеме предобучения. (Понятно, что метод self-supervised, но есть этап supervised fine-tuning и, возможно, на нем можно воспользоваться старой схемой с «навотором» для учета лейблов)

Применение:

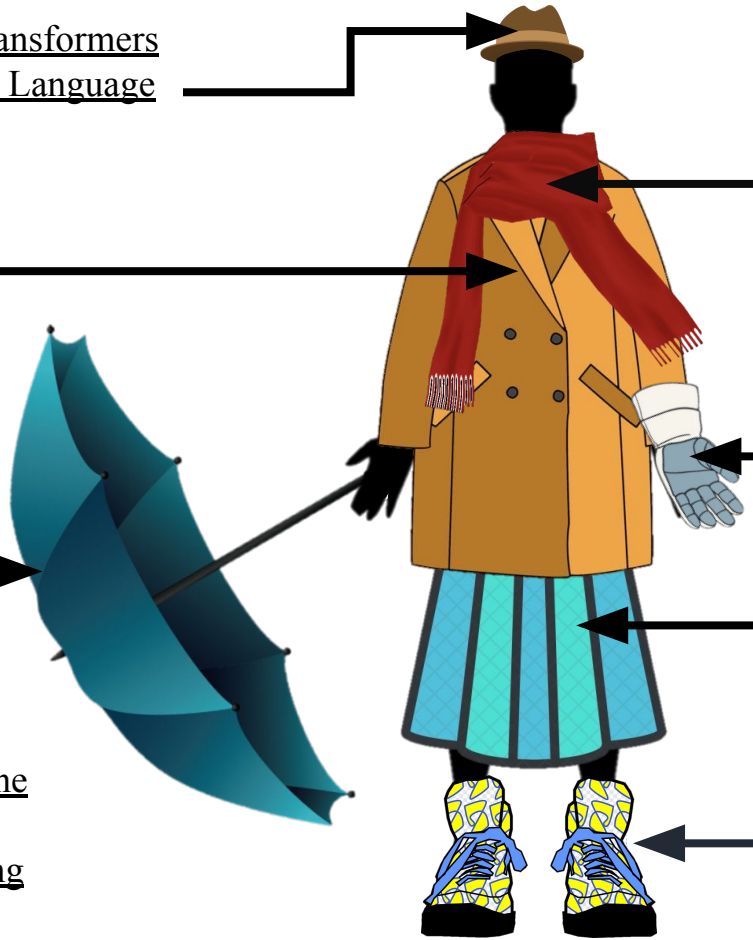
Предложенную модель можно использовать тогда, когда ведется работа с tabular data в задаче классификации, когда мало размеченных данных или есть шум в лейблах.

HAT
Hardware-Aware Transformers for Efficient Natural Language Processing

SCARF
Self-Supervised Contrastive Learning using Random Feature Corruption

CoaT
Co-Scale Conv-Attentional Image Transformers

GloVe
Global Vectors for Node Representations



SKIRT
an Advanced Dust Radiative Transfer Code with a User-Friendly Architecture

UMBRELLA
Uncertainty-Aware Model-Based Offline Reinforcement Learning Leveraging Planning

SHOE
Supervised Hashing with Output Embeddings