IEEE International Conference on Big Data 2023, Dec 2023

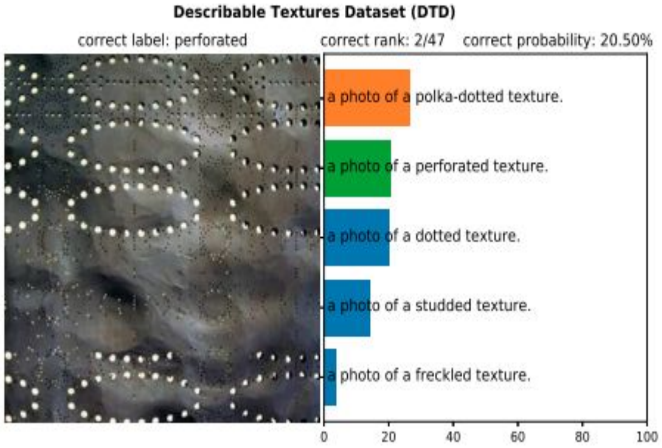
Contrastive Learning: Big Data Foundations and Applications

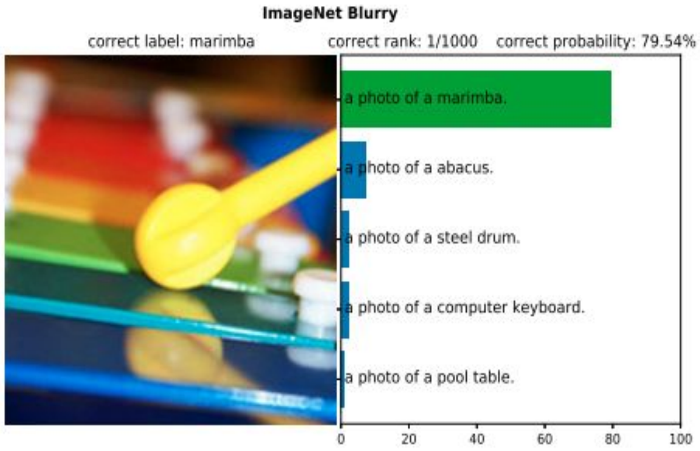
Sandhya Tripathi, PhD **Dr. Christopher R. King, MD, PhD**

Department of Anesthesiology
Washington University School of Medicine
St Louis, Missouri

Website: https://github.com/sandhyat/ContrastiveLearning_Tutorial

Motivating examples : CLIP



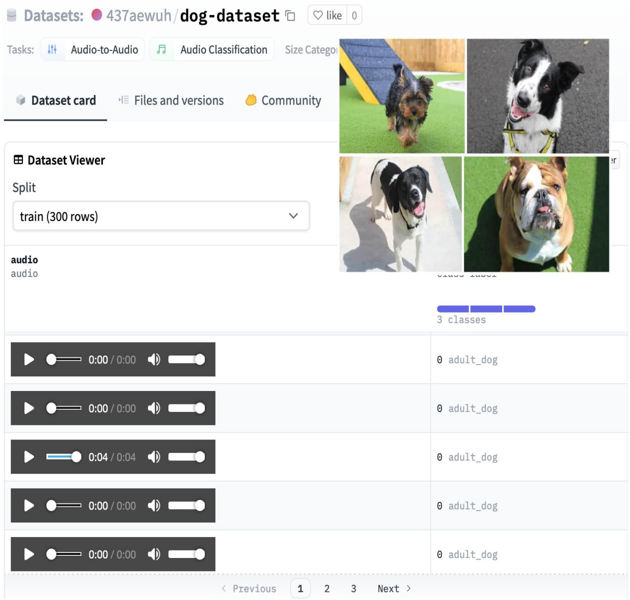



a shiba inu wearing a beret and black turtleneck

https://arxiv.org/pdf/2103.00020.pdf
https://cdn.openai.com/papers/dall-e-2.pdf

Motivating examples: CLAP

Dog bark dataset



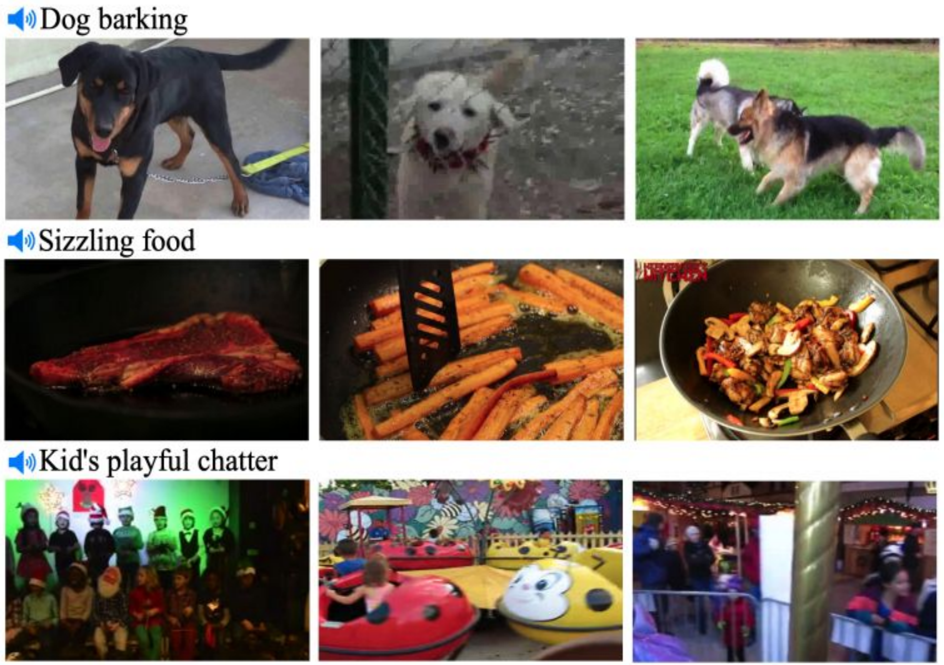

```
In [58]: # details about this particualr dataset available at https://github.com/suzuki256/dog-dataset
dog_barks = load_dataset("437aewuh/dog-dataset", split="train", streaming=True) # dataset of dogbarks. 3 classes: a
...

In [96]: dog_barks = dog_barks.shuffle(seed=42, buffer_size=100)  #shuffles the iterable dataset
sample = next(iter(dog_barks))   # selects the first row from the dog_barks
audio_sample_array = sample["audio"]["array"] # numberical array of the selected sample

In [97]: # checking the id and verifying the sound
print(sample)
Audio(audio_sample_array, rate = 44100)

{'audio': {'path': 'hf://datasets/437aewuh/dog-dataset@73810687d1afab5e8318f4d7e51bfd7175d1848d/adult_dog/adult_do
g_0014.wav', 'array': array([ 0.        ,  0.        ,  0.        , ..., -0.00198364,
       -0.00170898, -0.0017395 ]), 'sampling_rate': 44100}, 'label': 0}

Out[97]: -0:00

In [107]: # Different candidate labels for the zero shot evaluation
candidate_labels = ["Sound of a puppy", "Sound of an adult dog","Sound of a dog"]
candidate_labels1 = ["Sound of an toy dog", "Sound of real dog"]
candidate_labels2 = ["Sound of an adult dog", "Sound of a puppy"]
```

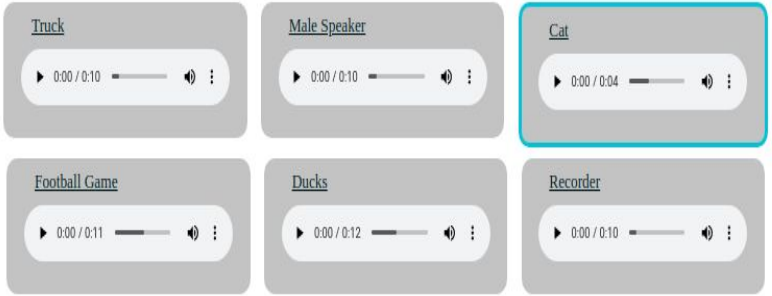
Loading CLAP model

```
In [103]: classifier = pipeline(task="zero-shot-audio-classification", model="laion/clap-htsat-unfused")

In [108]: classifier(audio_sample_array, candidate_labels=candidate_labels)

Out[108]: [{'score': 0.6315874457359314, 'label': 'Sound of a dog'},
 {'score': 0.3127667307853699, 'label': 'Sound of an adult dog'},
 {'score': 0.055645886808633804, 'label': 'Sound of a puppy'}]

In [109]: classifier(audio_sample_array, candidate_labels=candidate_labels1)

Out[109]: [{'score': 0.9551675319671631, 'label': 'Sound of real dog'},
 {'score': 0.044832486659288406, 'label': 'Sound of an toy dog'}]

In [110]: classifier(audio_sample_array, candidate_labels=candidate_labels2)

Out[110]: [{'score': 0.8489577174186707, 'label': 'Sound of an adult dog'},
 {'score': 0.15104229748249054, 'label': 'Sound of a puppy'}]
```

Reproducible here

Audio to image retrieval without seeing the (audio, image) pair data





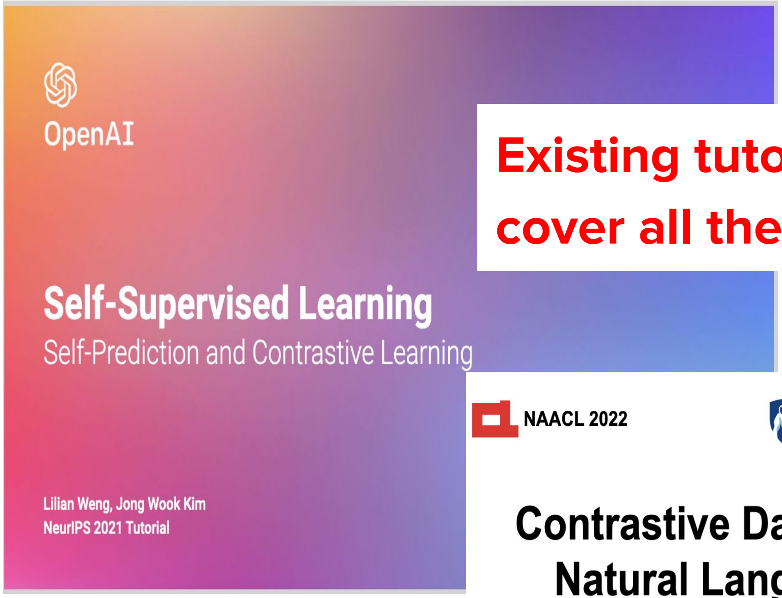
lts

Connecting- Multimodal contrastive learning

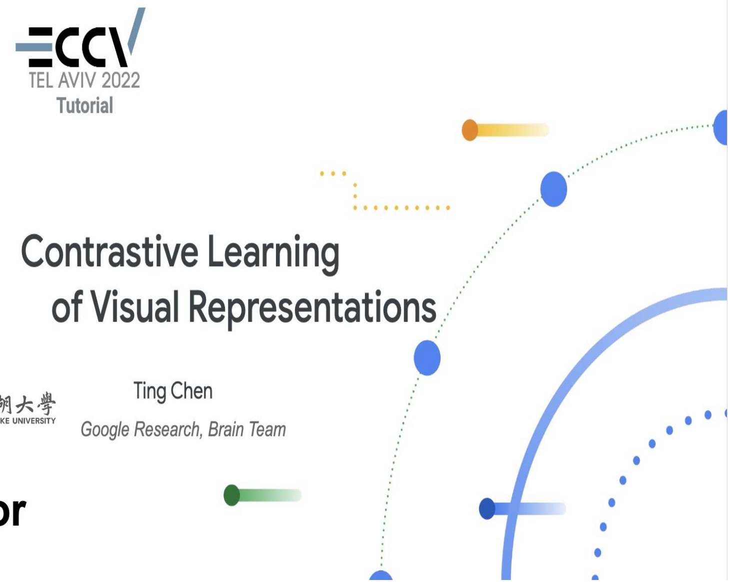
https://arxiv.org/pdf/2305.14381.pdf

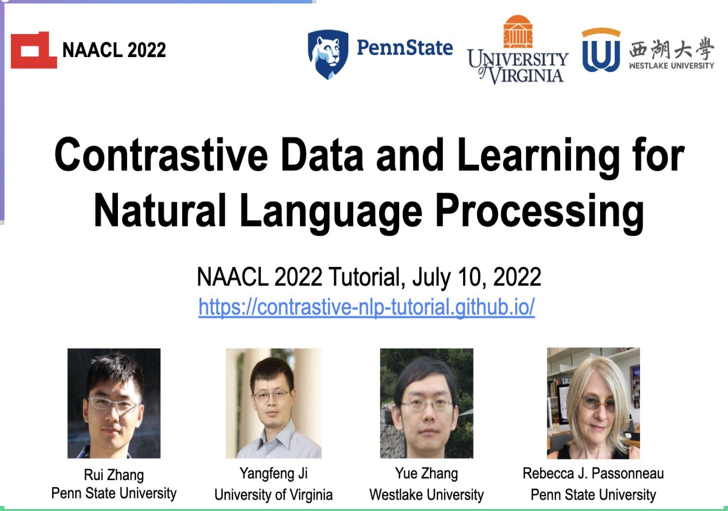
https://c-mcr.github.io/C-MCR/index.html

Why this tutorial?



Existing tutorials do not cover all the modalities





Outline

Part 1: Contrastive learning foundations

Part 2: Contrastive learning in different modalities and applications

Part 3: Hands on demo for time series and tabular data representations examples

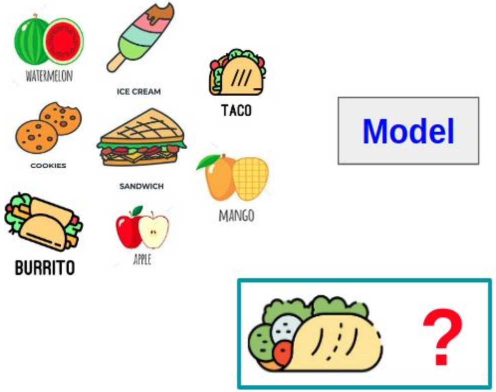
Part 4: Beyond unimodal contrastive learning

Part 1: Contrastive learning foundations

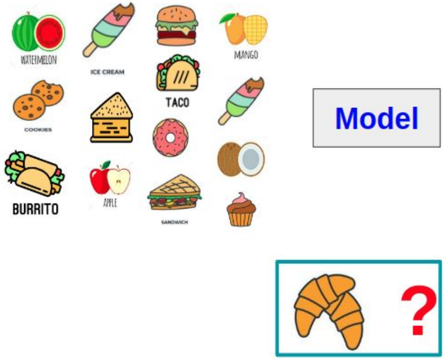
- What is contrastive learning?
- Main components in learning: augmentations/views, loss functions
- Augmentation and loss function examples
- Evaluation and applications of CL
- Why does contrastive learning work?

Different learning paradigms

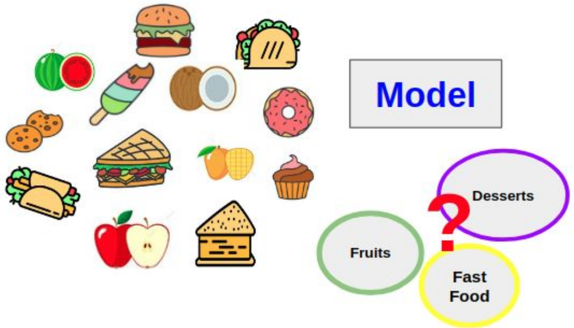
Supervised Learning



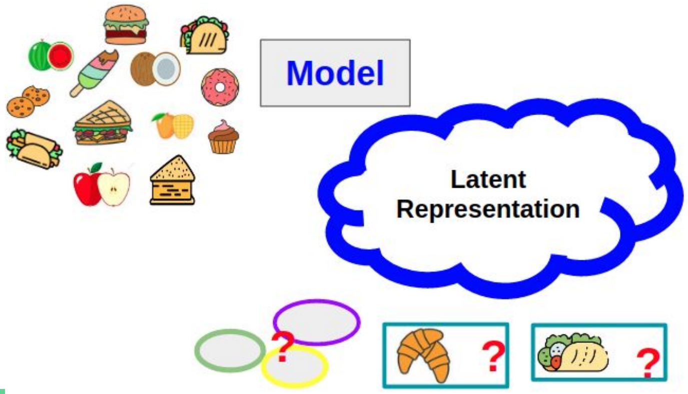
Semi-supervised Learning



Unsupervised Learning



Self-supervised learning



Self-reconstruction

Learn representation by reconstructing example after noise / dimension reduction

Multi-view Self-supervised learning

Learn representation by comparing different views / encoders of the same object

Contrastive

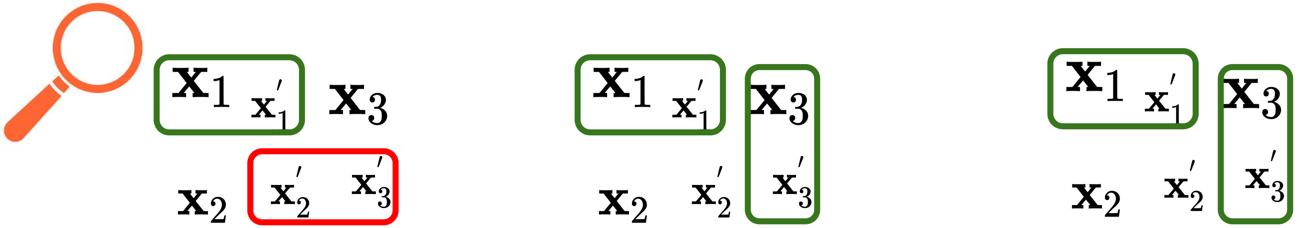
Compare projections from the same example (positive pairs) versus projections from different examples (negative pairs)

Distillation based

Projections of one encoder used as target for other

Clustering based

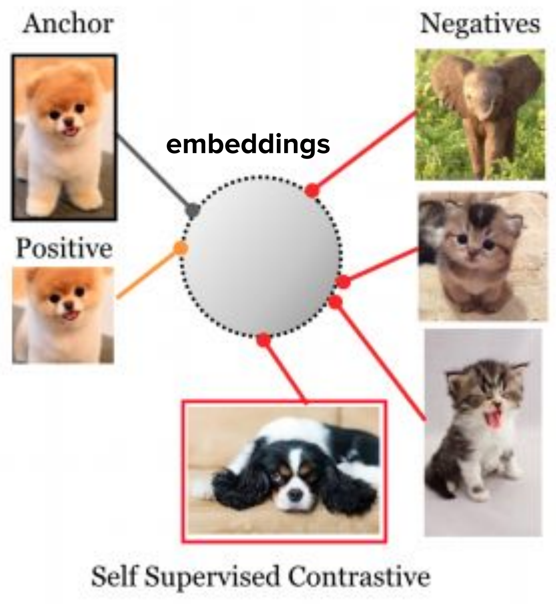
Predicting cross-cluster codes using clustered projections



What is contrastive learning?

Learn higher level feature representation where the data itself provides supervision via comparison

Similar data points close, dissimilar ones are far apart



$\mathbf{x}$ anchor point

$\mathbf{x}^+$ positive

$\mathbf{x}^-$ negative

f encoder

$$sim(f(\mathbf{x}, \mathbf{x}^+)) >> sim(f(\mathbf{x}, \mathbf{x}^-))$$

Source: Khosla et. al. 2021

Contrastive learning components

Input data + Models + Optimization over an objective

What ??

How??

Different views

Augmentations

Different modalities

Different sources of information

Encoders

Projection heads

What ??

How ??

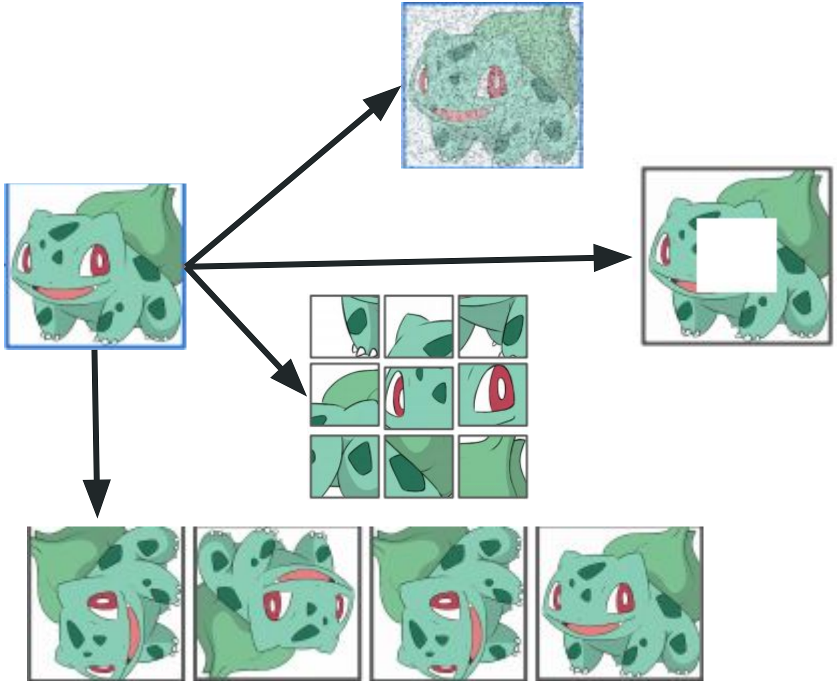
Loss maximizing similarity between representations

Training and batch construction (choice of negative pairs) strategies

Examples of data augmentations

Label preserving: A good set of views are those that share the minimal information necessary to perform well at the downstream task.

Vision: Random crop, Rotation, Adding noise, masking, color jitter



Text: lexical editing, back translation

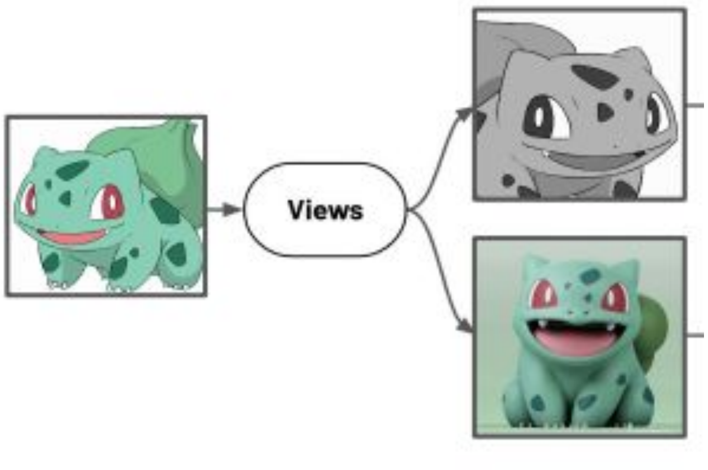
Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

[EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks.]

[What makes for good views for contrastive learning?]

Multiple views of the same data

Images

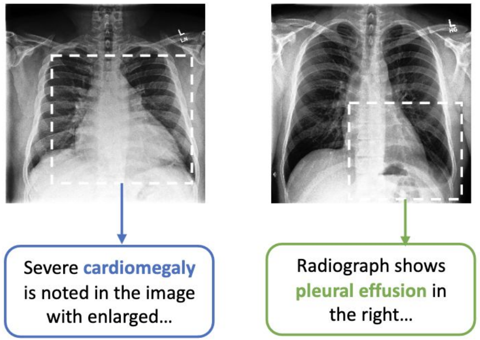


Text

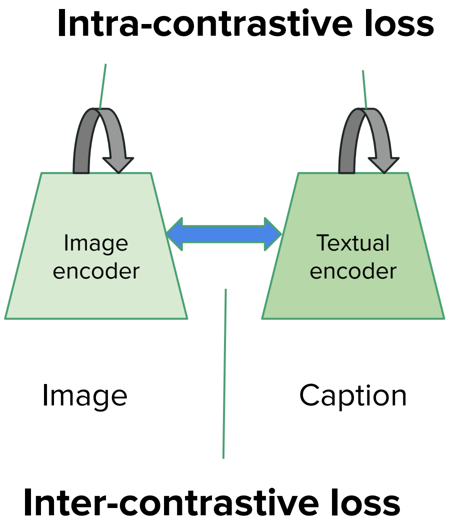
Two views

More than one modalities or data types?

- Availability of additional source of information
  - Captions of images
  - Subtitles in videos
- Issues with unimodal self supervised learning
  - High inter-class similarity in some domains
- Can be used in
  - Zero shot learning by using one set of embeddings and matching
  - Combining representation from different modality sources
    - Demographics + Vitals + medications for readmission prediction



Solution class: [ConVIRT]



Contrastive loss objectives

Contrastive loss (2005)

Face verification

Min: embedding distance if x1,x2 genuine pair
Max: if x1, x2 imposter pair

Noise contrastive estimation 2010

Learning by comparison of target (positive) and noise (negative) distribution

Relates to the log-likelihood in a logistic regression model for discrimination

Triplet loss (2015)

Optimize s.t.

Dist(+ , anchor+margin) **<** **Dist**(-, anchor)

N pair loss 2016

Lifted structured loss 2016

Soft nearest neighbour loss 2019

Noise contrastive estimation based losses

InfoNCE loss 2018

Context c

N samples (positive + negatives)

X = [x1, x2,...,xN]

f is the density ratio used to predict the future x observations using the context

$$\mathcal{L} = -\log \frac{\boxed{f(\mathbf{x},\mathbf{c})}}{\sum_{\mathbf{x}' \in \mathbf{X}} f(\mathbf{x}',\mathbf{c})}$$

NT-Xent 2020

InfoNCE with

f as cosine similarity

Positive and negatives are normalized embeddings

Addition of temperature parameter

$$\mathcal{L} = -\log \frac{\boxed{exp(sim(\mathbf{x}_i,\mathbf{x}^+)/\tau)}}{exp(sim(\mathbf{x}_i,\mathbf{x}^+)/\tau) + \sum_{j=1}^{N-1} exp(sim(\mathbf{x}_i,\mathbf{x}_j^-)/\tau)}$$

Contrastive Learning evaluation and application

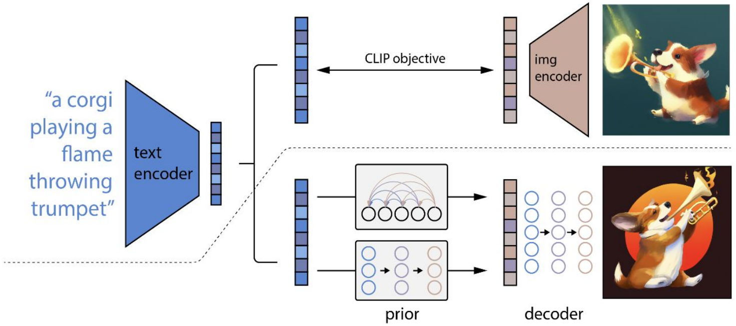
CL trained encoder

Linear evaluation
Fine tune and attach a classifier head

Transfer learning
Use CL learnt representation for another task

Semi-supervised learning

Generative tasks

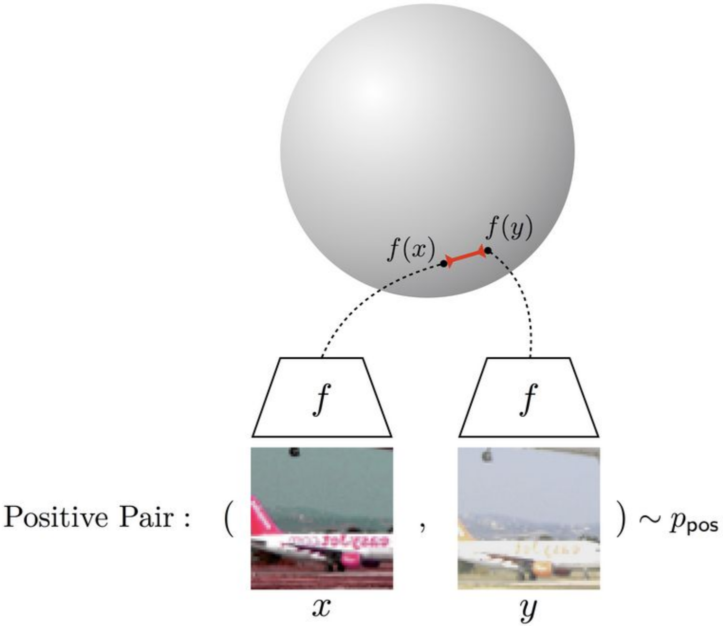


Zero shot learning by similarity scores of embeddings of modalities

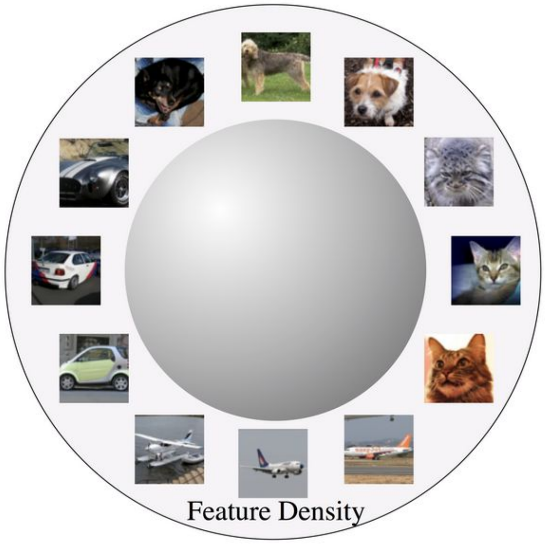
Explanation for the working of contrastive learning

- Geometric interpretation in embedding space
- Relation to mutual information

Understanding contrastive learning through alignment and uniformity on the hypersphere



Alignment: Similar samples have similar features



Uniformity: Preserve maximal information

Mutual information (MI), Entropy and reconstruction (ER) based explanation for CL

InfoNCE loss lower bounds the MI between learnt representations of different views $\Longleftrightarrow$ MI maximization

$$I(Z_1; Z_2) \geq \underbrace{H(Z_2)}_{\text{Entropy}} + \underbrace{\mathbb{E}[\log q_{Z_2|Z_1}(Z_2)]}_{\text{Reconstruction term}} := I_{\text{ER}}(Z_1; Z_2),$$

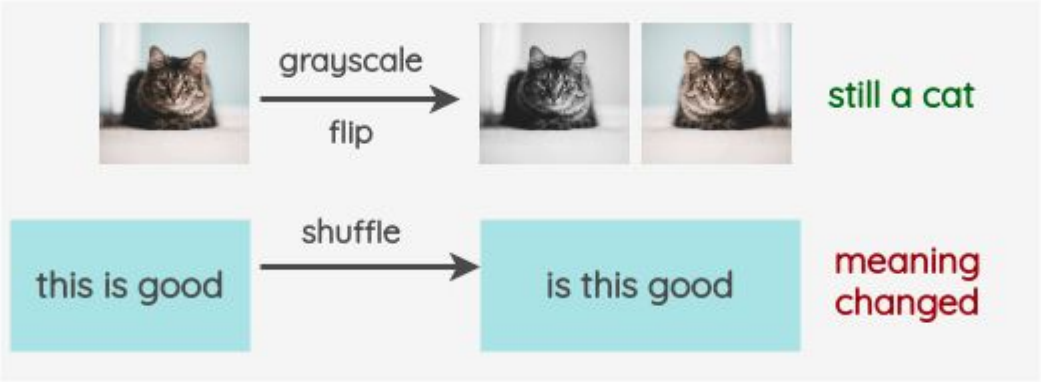
Another approach for contrastive optimization can be to maximize the ER bound.

The Role of Entropy and Reconstruction in Multi-View Self-Supervised Learning

Part 2: Contrastive learning in different modalities and applications

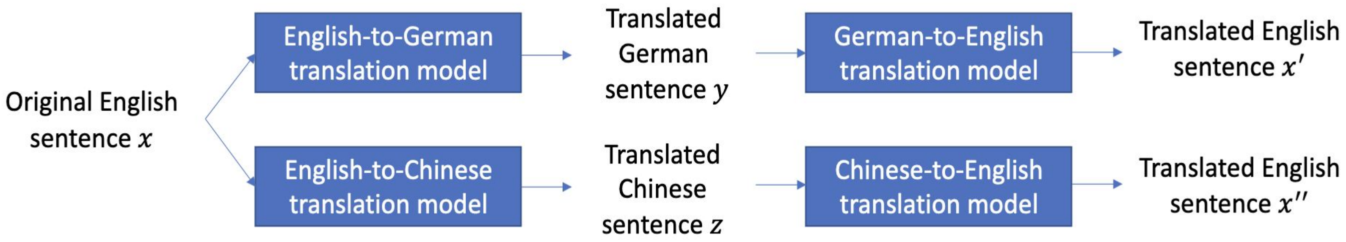
- Augmentations are not universal
- Augmentations in different modalities
  - In input space
  - In embedding space
- Choice of good augmentation/views and beyond
- Batch construction strategies in different modalities
- Optimization objective and training strategies as adapted by different modalities
- CL application in other modalities

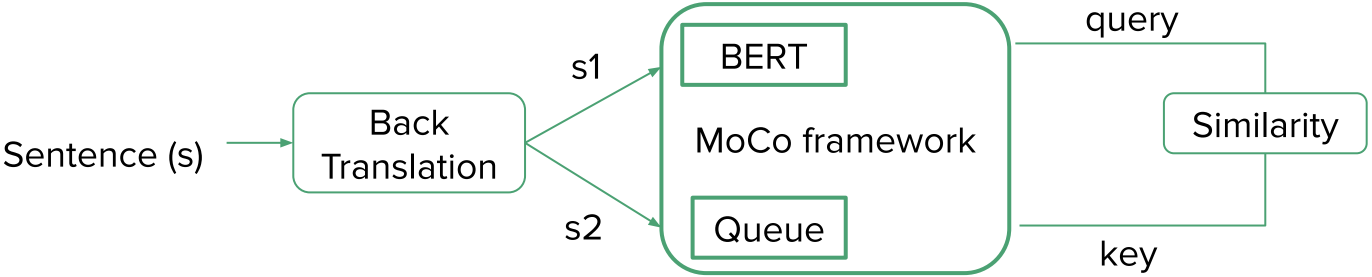
Need for specific augmentations in different modalities



Rotations don't make sense in time-series

Sentence based Back translation augmentation





- Token level target might not capture global semantics, hence sentence based augmentations

[CERT: Contrastive Self-supervised Learning for Language Understanding]

Input data augmentations

Use of different set of augmentation functions

Cross view task that uses the context of one augmentation to predict the future embedding of another augmentation

1) Weak augmentations: scaling and time shifting
2) Strong : permutation, strong jittering

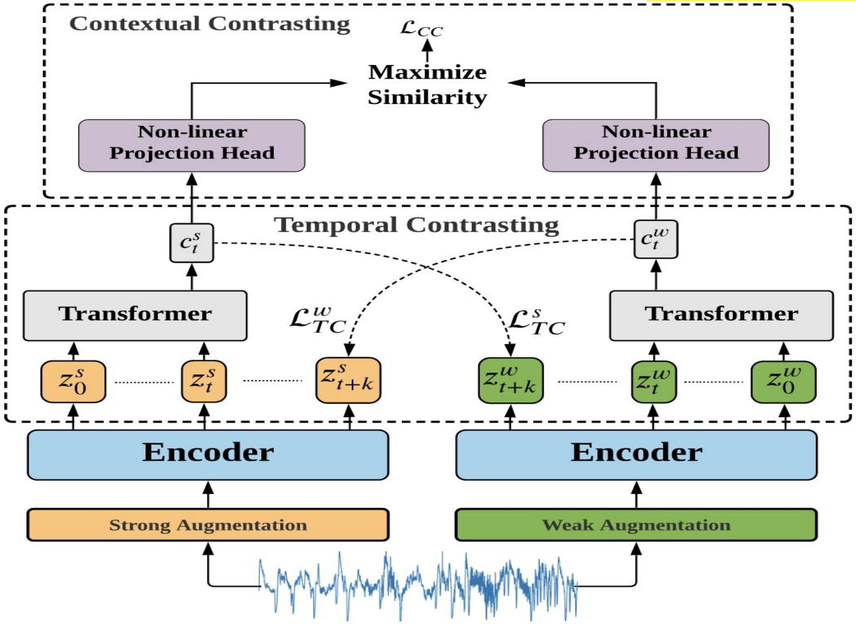


Fig. 1: The overall architecture of the proposed TS-TCC. The Temporal Contrasting module learns robust temporal features through a tough cross-view prediction task. The Contextual Contrasting module learns discriminative features by maximizing the similarity between the contexts of the same sample while minimizing its similarity with the other samples within the mini-batch.

Self-supervised contrastive representation learning for semi-supervised time-series classification

Input data augmentations

Domain dependent combination of augmentation and views

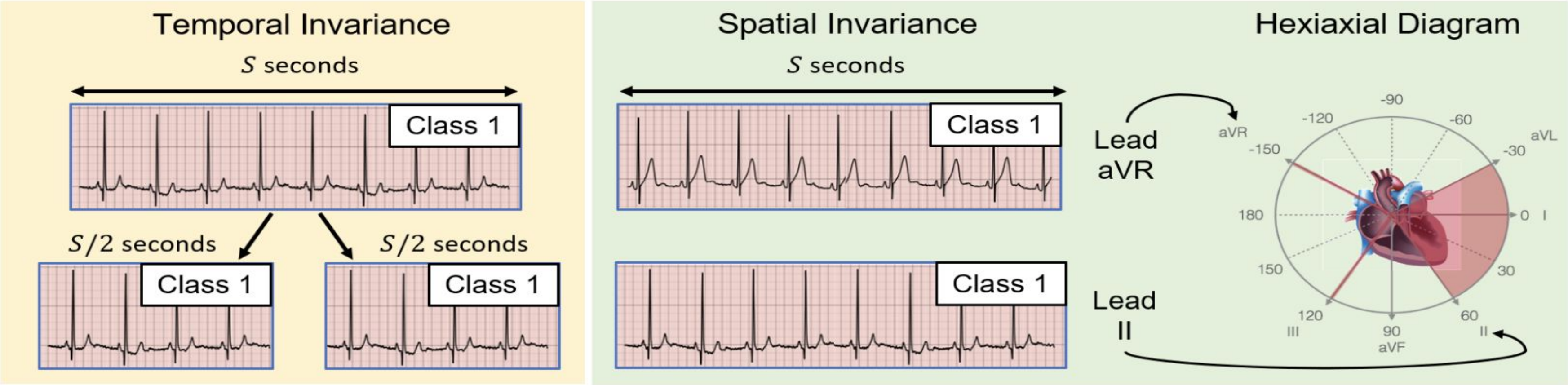


Figure 1. ECG recordings reflect both temporal and spatial information. This is because they measure the electrical activity of the heart using different leads (views) over time. **Temporal Invariance.** Abrupt changes to the ECG recording are unlikely to occur on the order of seconds, and therefore adjacent segments of shorter duration will continue to share context. **Spatial Invariance.** Recordings from different leads (at the same time) will reflect the same cardiac function, and thus share context.

Exploit different kind of invariances for a single modality for a specific patient

[CLOCS: Contrastive Learning of cardiac signals across space, time and patients]

Frequency based augmentations

Input data augmentations

Time-based and freq-based representations are closer in time-freq space where the consistency loss is defined.

Instead of both augmented views, original and augmented view are used for maximizing representation similarity.

$$\mathcal{L}_{C,i} = \sum_{S^{\text{pair}}} \left(S_i^{\text{TF}} - S_i^{\text{pair}} + \delta \right), \quad S^{\text{pair}} \in \{ S_i^{\text{T}\widetilde{\text{F}}}, S_i^{\widetilde{\text{T}}\text{F}}, S_i^{\widetilde{\text{T}}\widetilde{\text{F}}} \},$$

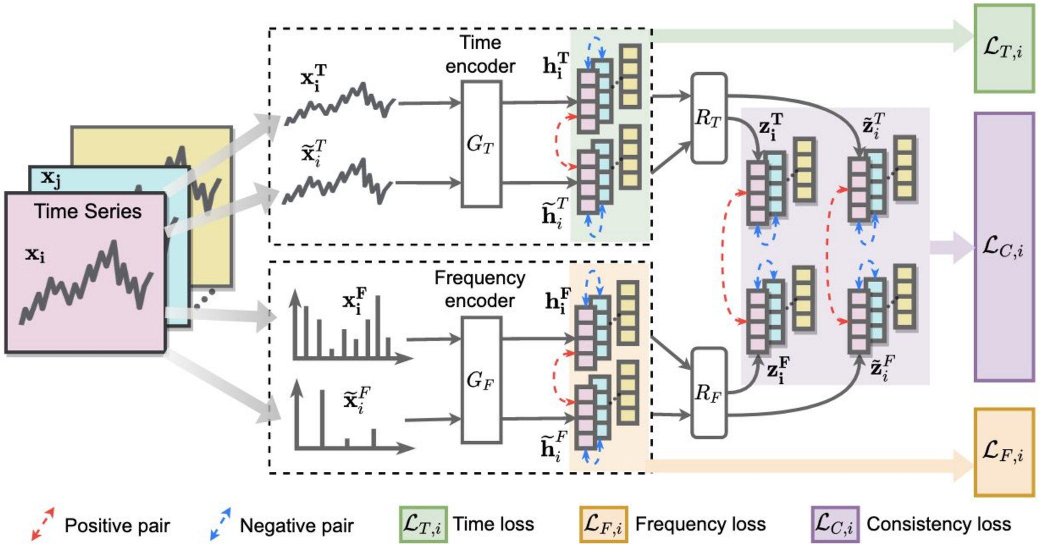
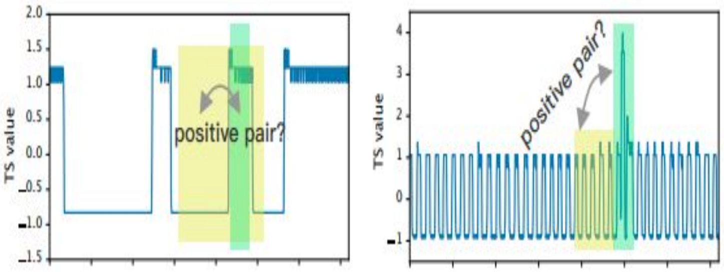


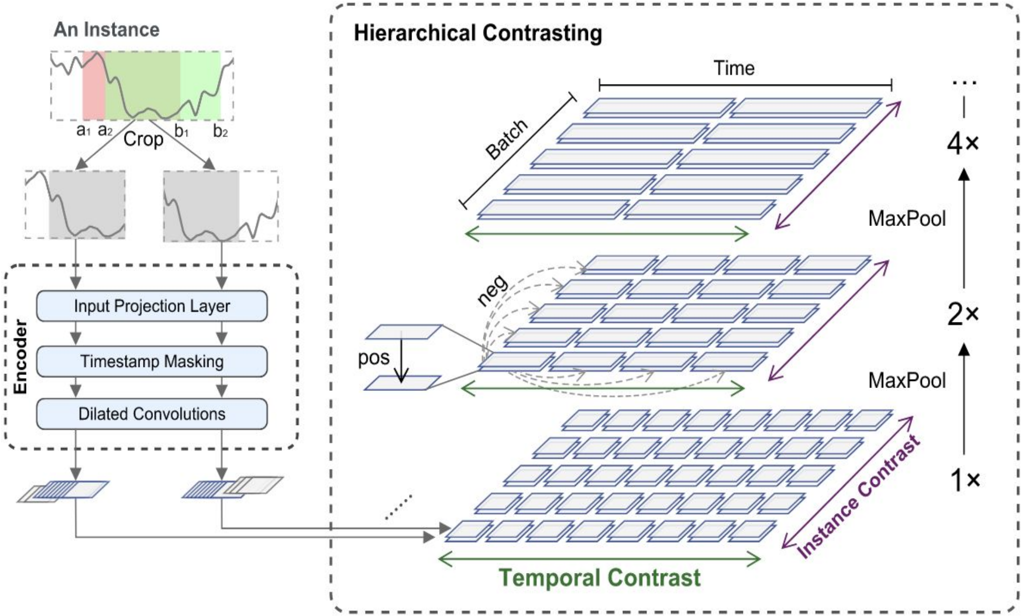
Figure 2: Overview of TF-C approach. Our TF-C pre-training model $\mathcal{F}$ has four components: a time encoder G_{T}, a frequency encoder G_{F}, and two cross-space projectors R_{T} and R_{F}. For an input time series $\boldsymbol{x}_i$, the model produces time-based representations (*i.e.*, $\boldsymbol{z}_i^{\text{T}}$ and $\widetilde{\boldsymbol{z}}_i^{\text{T}}$ of input $\boldsymbol{x}_i$ and its augmented version, respectively) and frequency-based representations (*i.e.*, $\boldsymbol{z}_i^{\text{F}}$ and $\widetilde{\boldsymbol{z}}_i^{\text{F}}$ of input $\boldsymbol{x}_i$ and its augmented version, respectively). The TF-C property is realized by promoting the alignment of time- and frequency-based representations in the latent time-frequency space, providing a vehicle for transferring $\mathcal{F}$ to a target dataset not seen before.

[Self Supervised Contrastive Pre-Training for time-series via time frequency consistency]

Augmentations that avoid level shift and anomalies

Input data augmentations

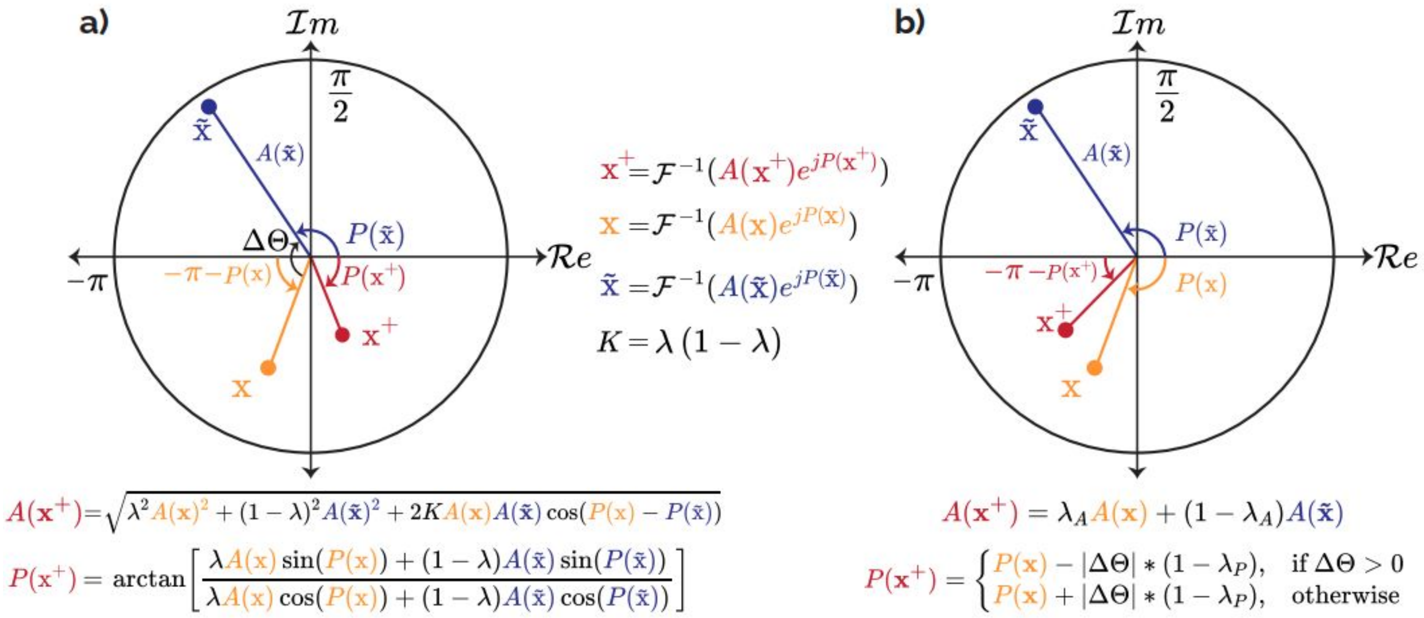




- Importance of contextual consistency: representations at the same time stamps in two augmented contexts (cropping and masking) as positive pair.
- Hierarchical contrastive loss

[TS2Vec: Towards universal representation of time series]

Separate augmentation for phase and amplitude of a time series

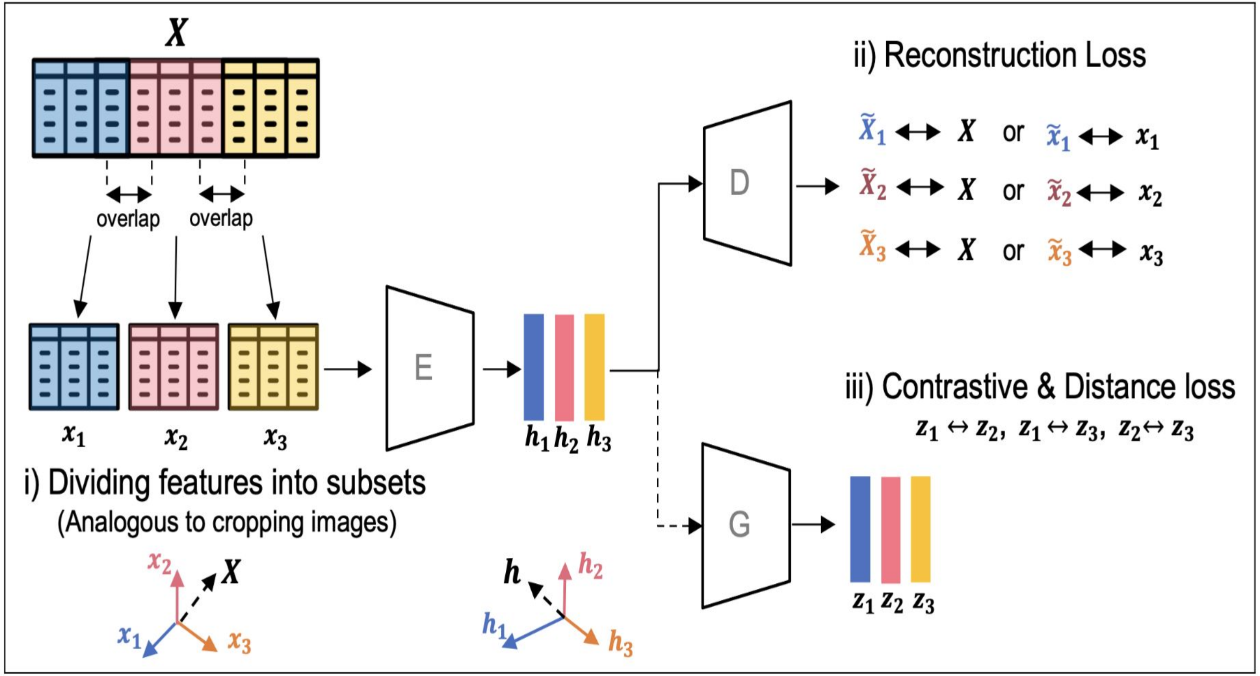


[Finding Order in Chaos: A Novel Data Augmentation Method for Time Series in Contrastive Learning]

Input data augmentations

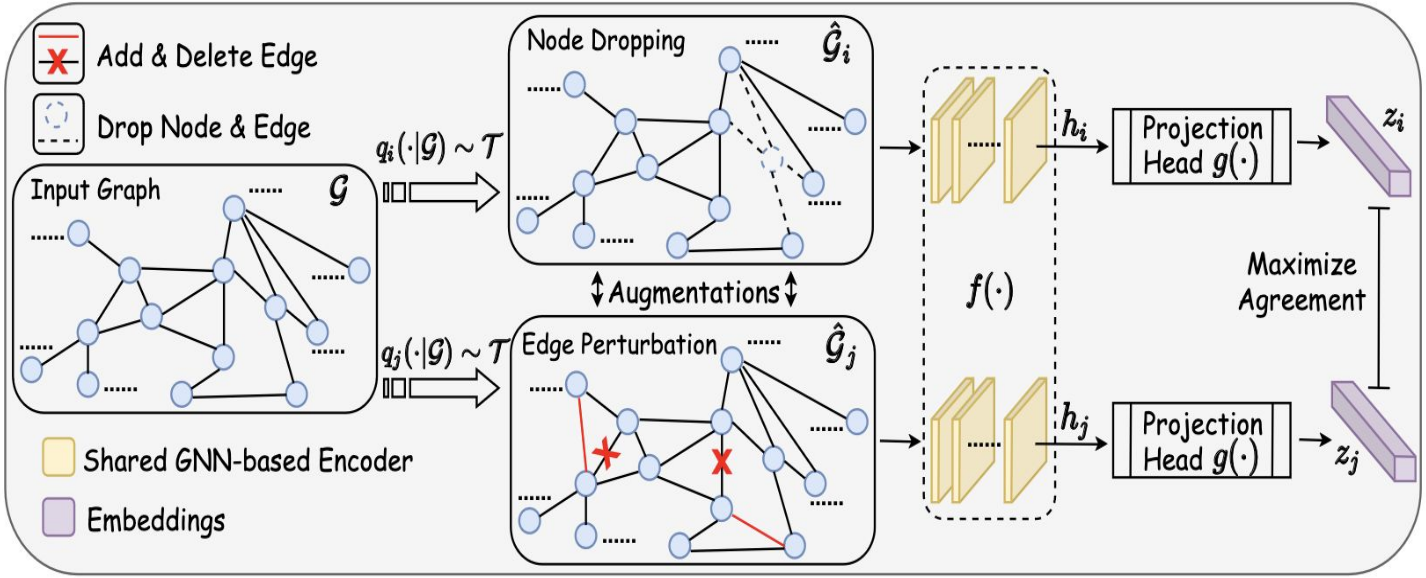
Column subsets of a tabular modality are the augmentations

- Contrasting between the views (overlapping column subsets)
- Full table reconstruction based on column subset (better generalization)



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

Data augmentations for graphs

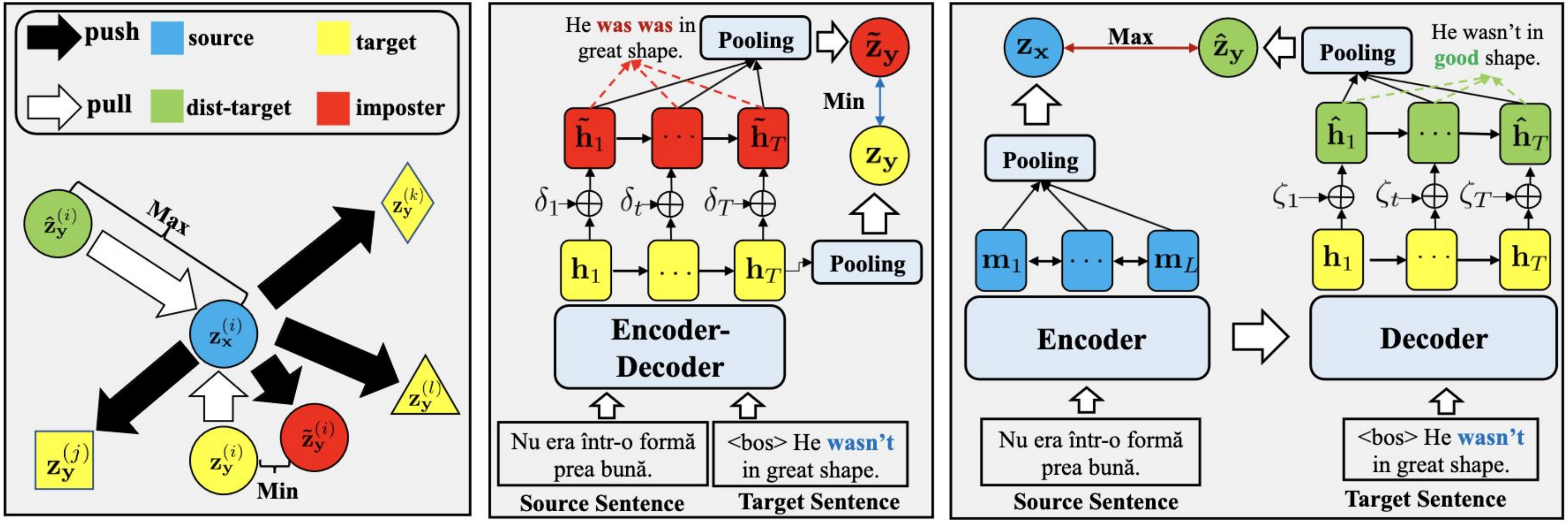


Identification and evaluation of different augmentations: node dropping, edge perturbation, attribute masking and subgraph

[Graph contrastive learning with augmentations](#)

Adversarial perturbations as contrastive views

Embedding space augmentations



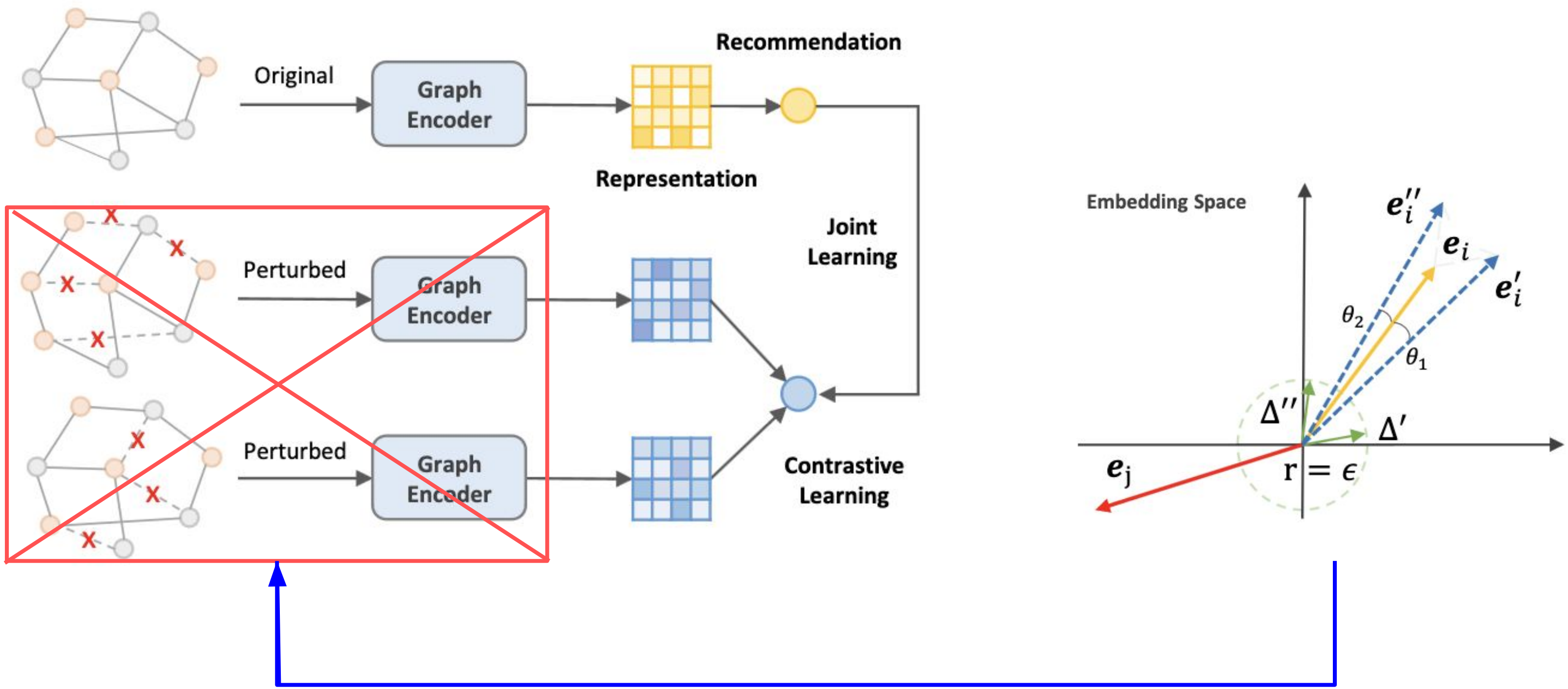
(a) Contrastive Learning with perturbation (b) Generation of Imposters (c) Generation of Distant-Targets

➕ Improves generalization by tackling exposure bias: Never exposed to incorrect tokens during training

[Contrastive learning for adversarial perturbations for conditional text generation](#)

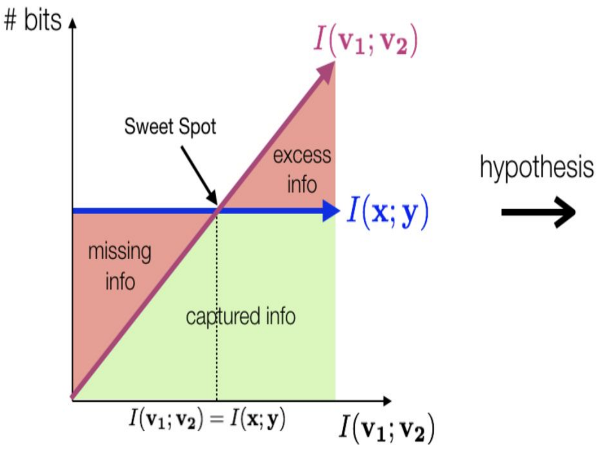
Contrasting on noisy embeddings as views

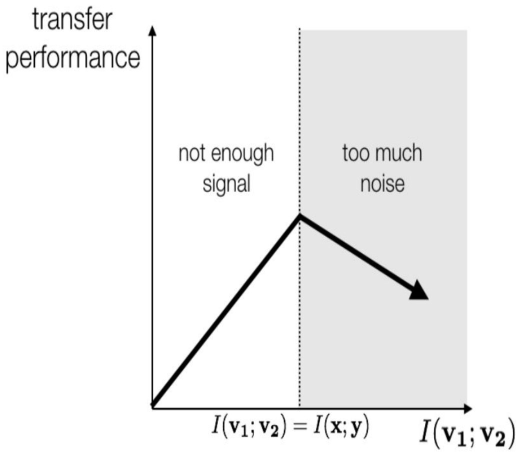
Embedding space augmentations



[Are Graph Augmentations Necessary? Simple Graph Contrastive Learning for Recommendation]

InfoMin principle: Only share label information w.r.t the downstream task





Source: [Blog]

Identify InfoMin aug set with reverse U downstream performance
RandomResizedCrop, Color jittering, Gaussian Blur, RandAugment, Color Dropping, JigSaw

[What makes for good views for contrastive learning?]

Label preserving to Label destroying augmentations

Specifying invariance → augmentations of input from different classes can collide

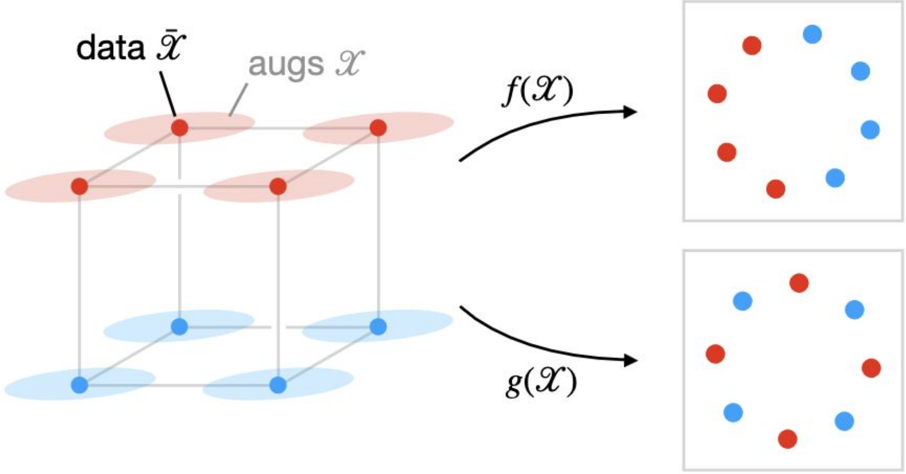
Viewmaker networks: Learn views/augmentations jointly with the representation.

Stochastically alter different parts of input → Not label preserving

Can also serve as feature dropout: preventing any one feature from becoming a shortcut feature and suppressing the learning of other features

Feature dropout: Revisiting the role of augmentations in contrastive learning

Does choice of augmentation and contrastive loss always explain the success of contrastive learning?



Pretraining: $L_{\text{cont}}(g) \approx L_{\text{cont}}(f)$

Downstream: $L_{\text{clf}}(g) \gg L_{\text{clf}}(f)$

[Understanding contrastive learning requires incorporating inductive biases.]

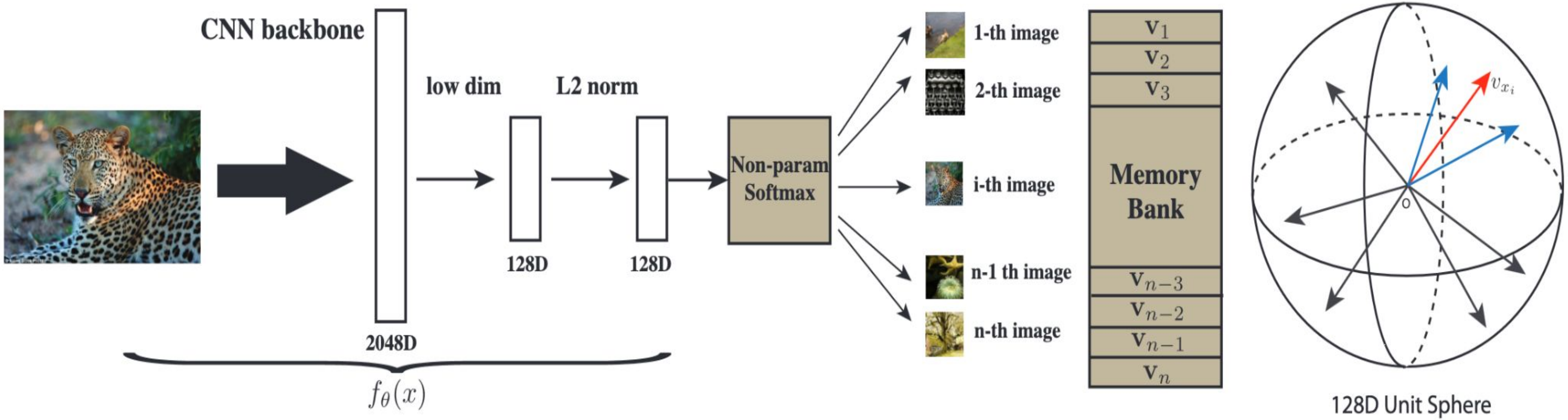
Batch construction strategies in Contrastive Learning

- Batch independent negative pairs
- Batch dependent negative pairs

Negative samples from a memory bank

- Non parametric classification problem at the instance level

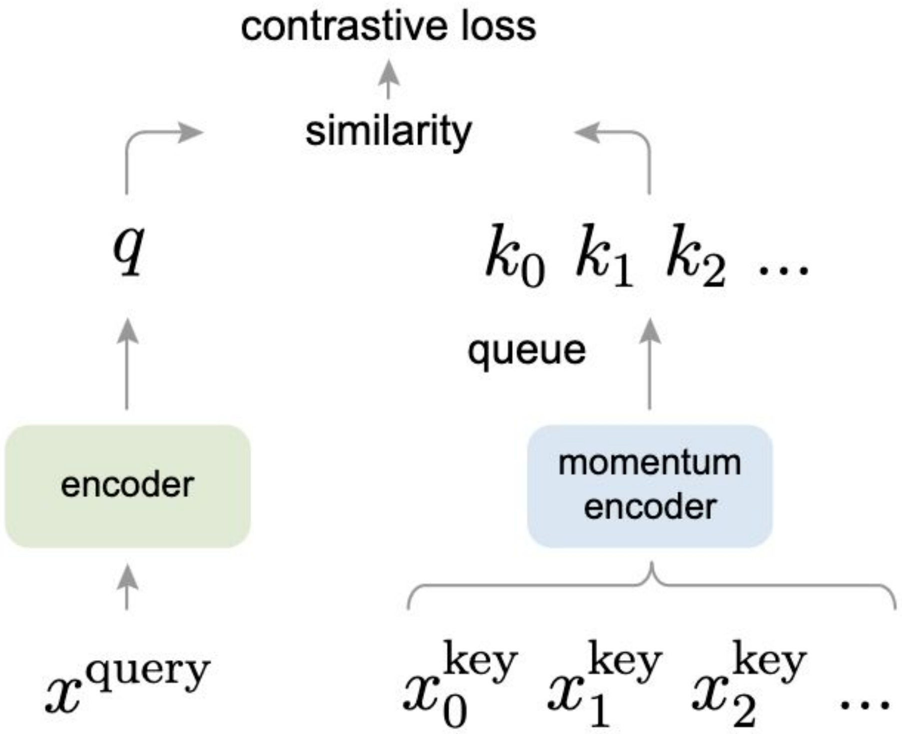
$$P(i|\mathbf{v}) = \frac{exp(\mathbf{v}_i^T \mathbf{v}/\tau)}{\sum_{j=1}^{n} exp(\mathbf{v}_j^T \mathbf{v}/\tau)} \quad \text{where} \quad \mathbf{v} = f_\theta(x)$$



[Unsupervised Feature Learning via Non-Parametric Instance Discrimination]

Avoiding the less consistent representations from the memory bank

- Dictionary as a queue
  - Independent of minibatch size
- Momentum update
  - $\theta_k = m\theta_k + (1-m)\theta_q$



[Momentum Contrast for Unsupervised Visual Representation Learning](#)

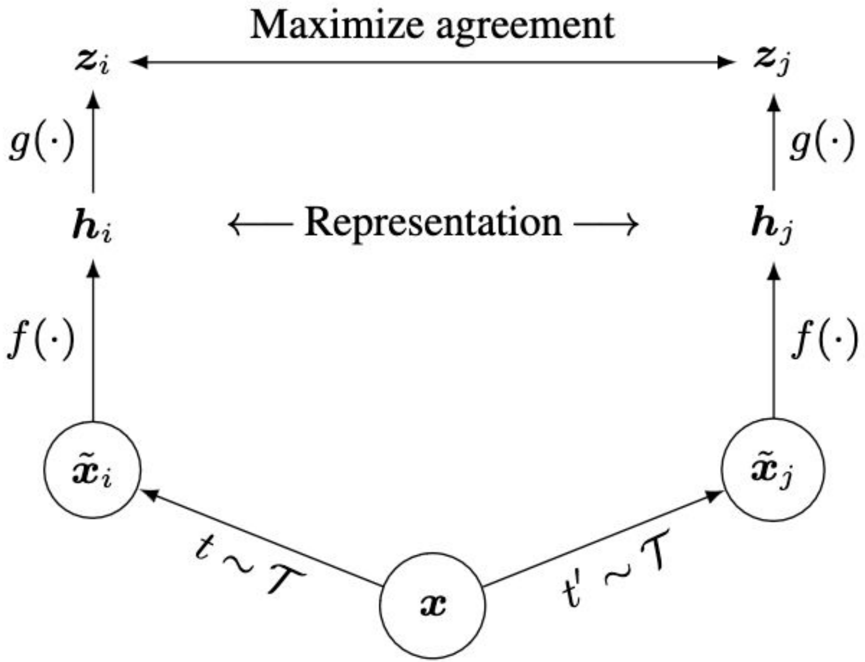
Introduction of non-linear projection h and large batch sizes

- Focus on
  - Composition of augmentation
  - Longer training
- Introduce NT-Xent

Positive pair

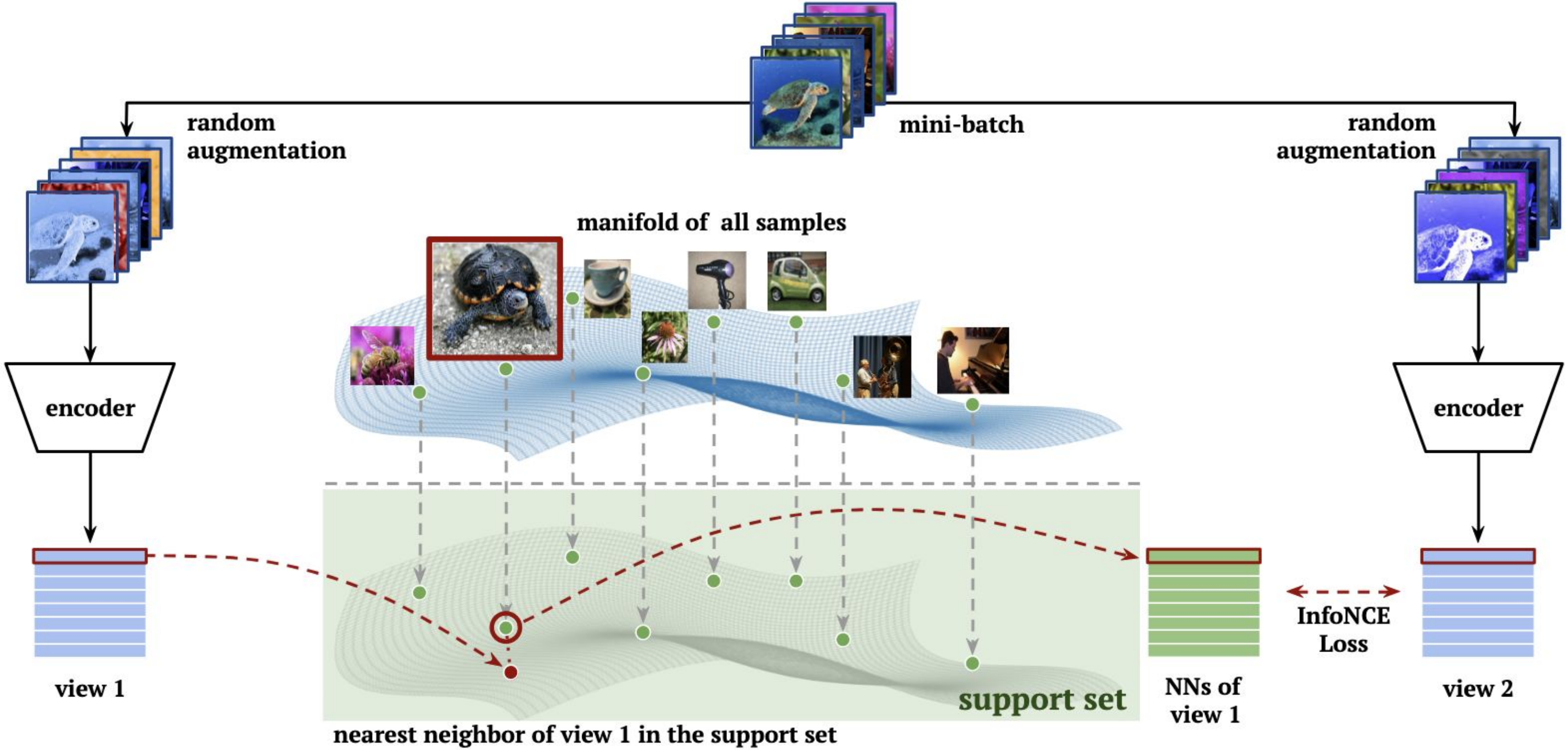
$$l_{i,j} = -log\frac{exp(sim(\mathbf{z}_i,\mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} 1[k\neq i]exp(sim(\mathbf{z}_i,\mathbf{z}_k)/\tau)}$$

Negative pair



[A Simple Framework for Contrastive Learning of Visual Representations]

Positive samples from the nearest neighbour set of the anchor's representation

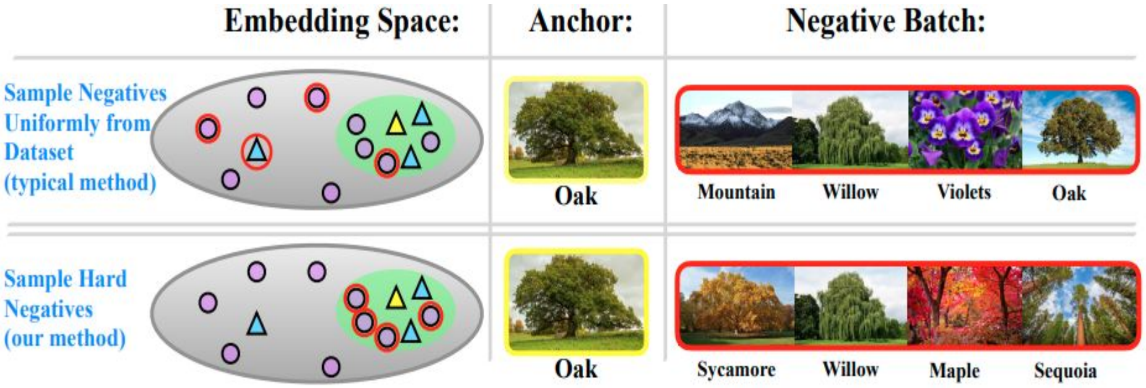


[With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations]

Hard negative mining

Identifying negative samples that have different label from the anchor but the embedding features may be very close.

Makes the discriminative task difficult→ learning of better representations

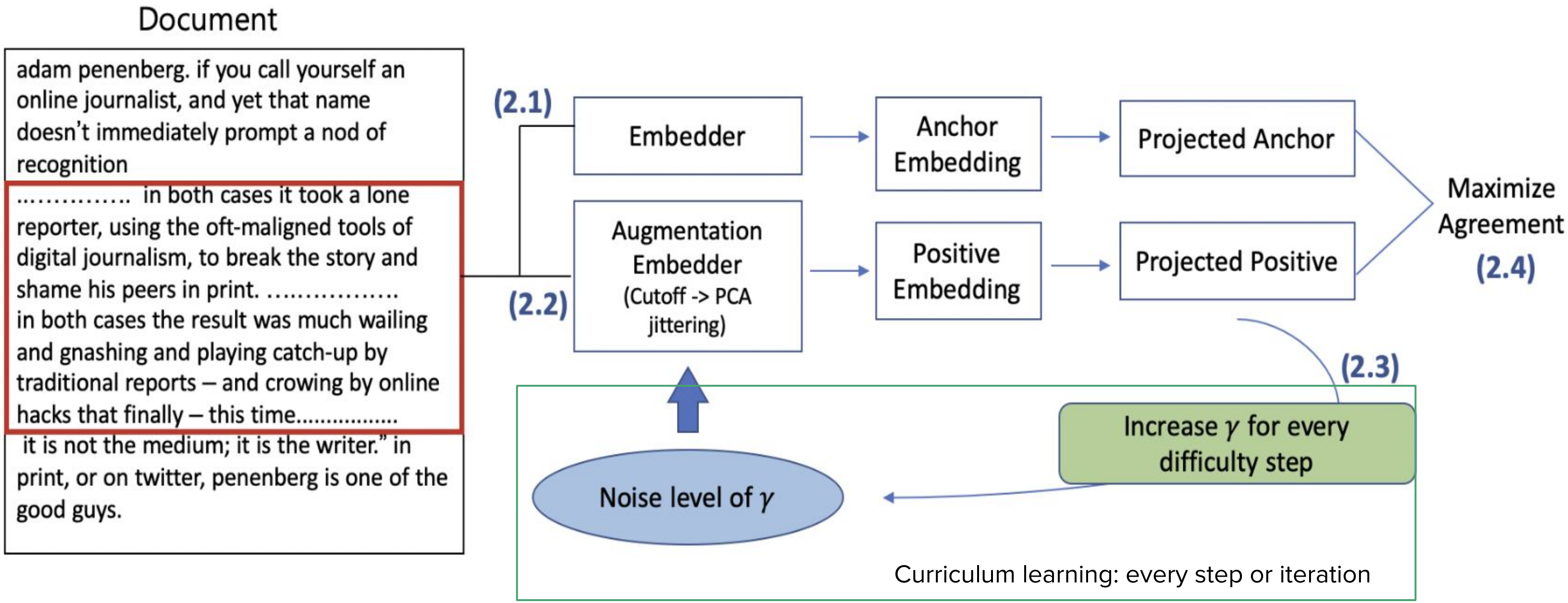


[Contrastive learning with hard negative samples]

Optimization objective and training strategies

- Combining with curriculum learning
- Using teacher student model approach
- Neighbourhood based contrastive loss
- Mixing in input space based contrastive loss
- Object-level contrastive loss
- Tricks and tweaks for tabular contrastive learning

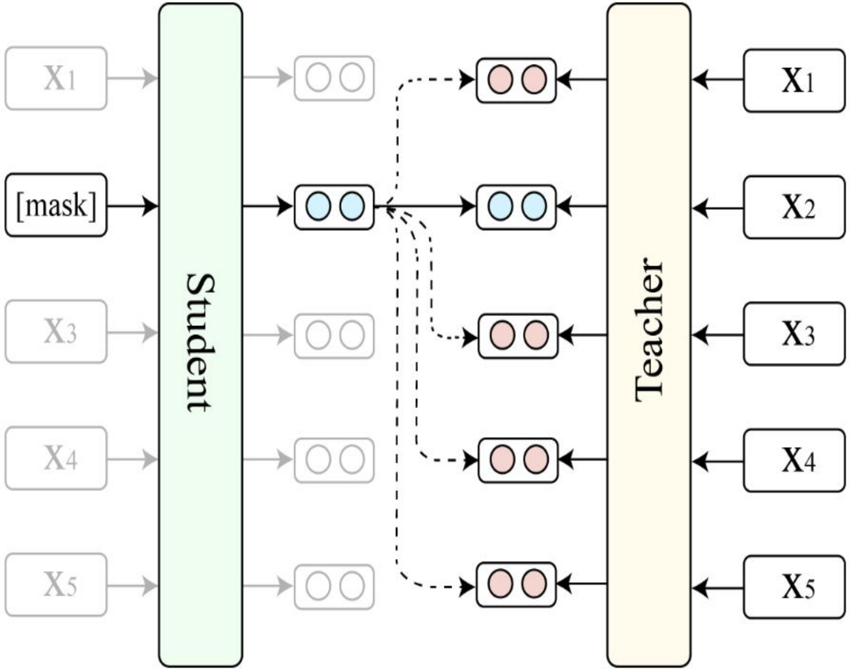
Stacked augmentations with increasing strength over the iterations



[Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning]()

Learning diverse token representation from frozen teacher model contrastively

- Builds on top of the BERT model
- Improved performance demonstrated on English and Chinese language tasks
- Focuses on learning token level representations



[TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning]

Neighbourhood embedded space as a supervisor

- Neighbourhood for sample i
  N (i) = {k != i | n(x_i, x_k) = 1}
- Neighbourhood alignment loss: positive pairs are projection and the corresponding neighbours in the momentum projection
- Neighbourhood discriminative loss: positive pairs are projection and the momentum based projection from other view

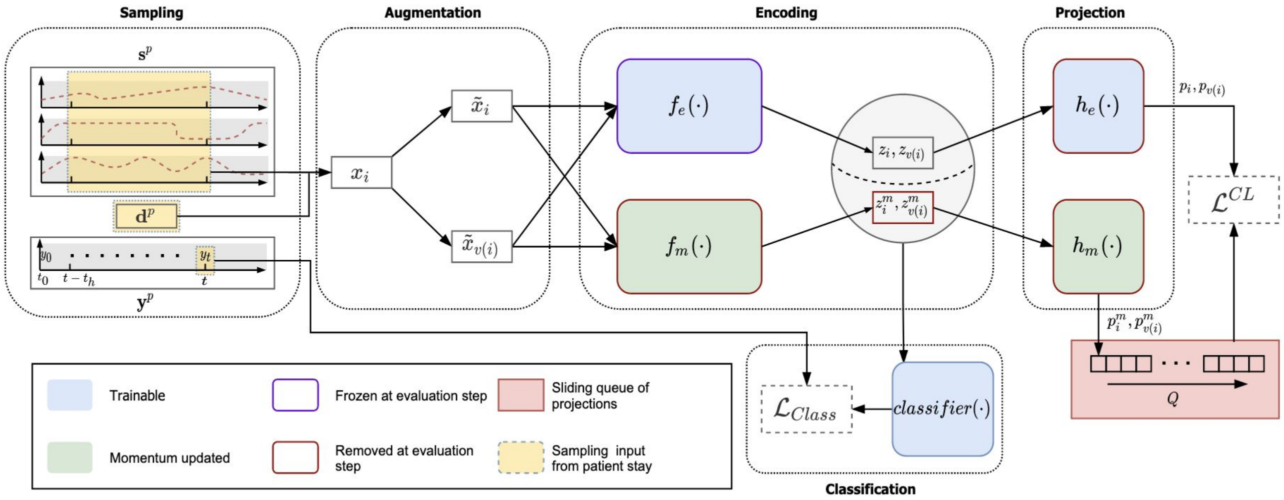


Figure 2. Schema of the contrastive pipeline initially proposed by (Chen et al., 2020c). From a patient stay p, we sample $x_i = (\mathbf{s}_t^p, \mathbf{d}^p)$ corresponding to the patient state at time t. We augment it twice and pass both views, $\tilde{x_i}$ and $x_{\tilde{v(i)}}$, through an encoder f_e and a momentum encoder f_m. At training time, the representations are further projected with h_e and h_m. From these projections and the sliding momentum queue Q, we compute the contrastive objective $\mathcal{L}^{\mathrm{CL}}$. At evaluation time, we freeze f_e and train a classifier on top of the learned representation.

[Neighbourhood contrastive learning applied to online patient monitoring]()

Convex combination of samples as an augmentation

- lambda in [0,1] following a Beta distribution with parameter alpha. Higher value of alpha→ more mixing

- Contrastive comparison between original sample and the convex combination

- Convex combination multiplier reflected in loss too

- Performance demonstration on ECG datasets

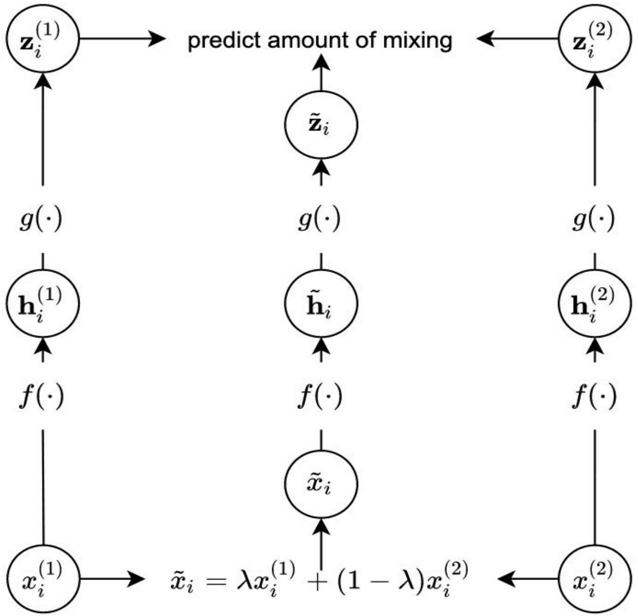


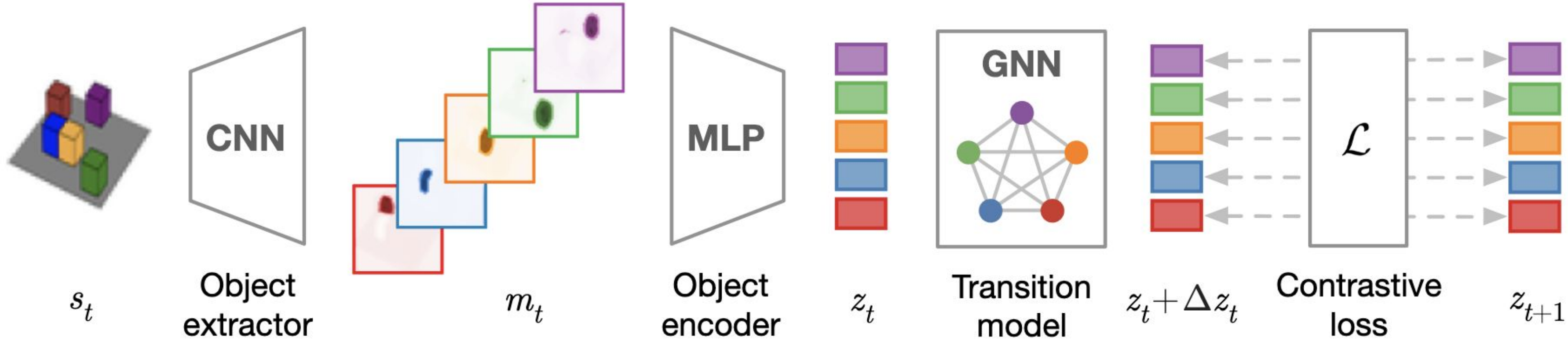
Fig. 2. The proposed framework. Two minibatches are sampled randomly from the data and combined using Equation 1. All samples are passed though an encoder $f(\cdot)$ resulting in a representation that can be used for down-stream tasks. Next, this representation is transformed using a projection head $g(\cdot)$ into a representation where the proposed contrastive loss is applied.

Mixing Up Contrastive Learning: Self Supervised Learning for Time Series

Object level contrastive loss

- Learning in environments with compositional structures where scenes can be disentangled into objects, their properties, and relations between them

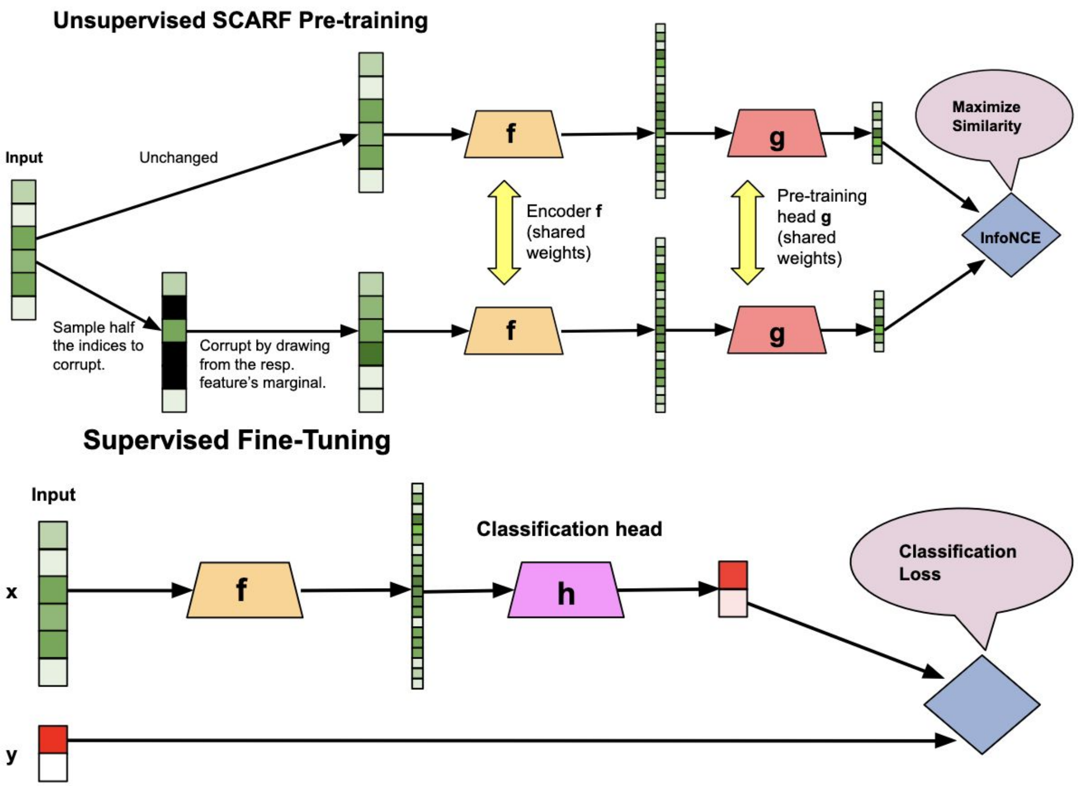
$$\mathcal{B} = \{(s_t, a_t, s_{t+1})\}_{t=1}^{T} \longleftrightarrow \mathcal{K} = \{(e_t, r_t, o_t)\}_{t=1}^{T}$$



Contrastive learning of structured world models

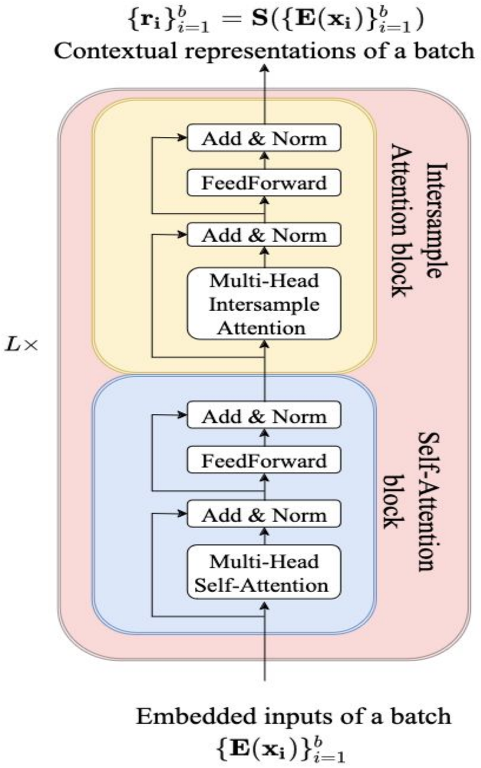
Feature noise as encoders and shared encoders for tabular datasets

- Relies on augmentations generated by random corruption of features using values from feature's empirical distribution.
- Performs well in limited label and label noise settings
- Insensitive to batch size, feature corruption type and rate, metric for maximizing similarity.

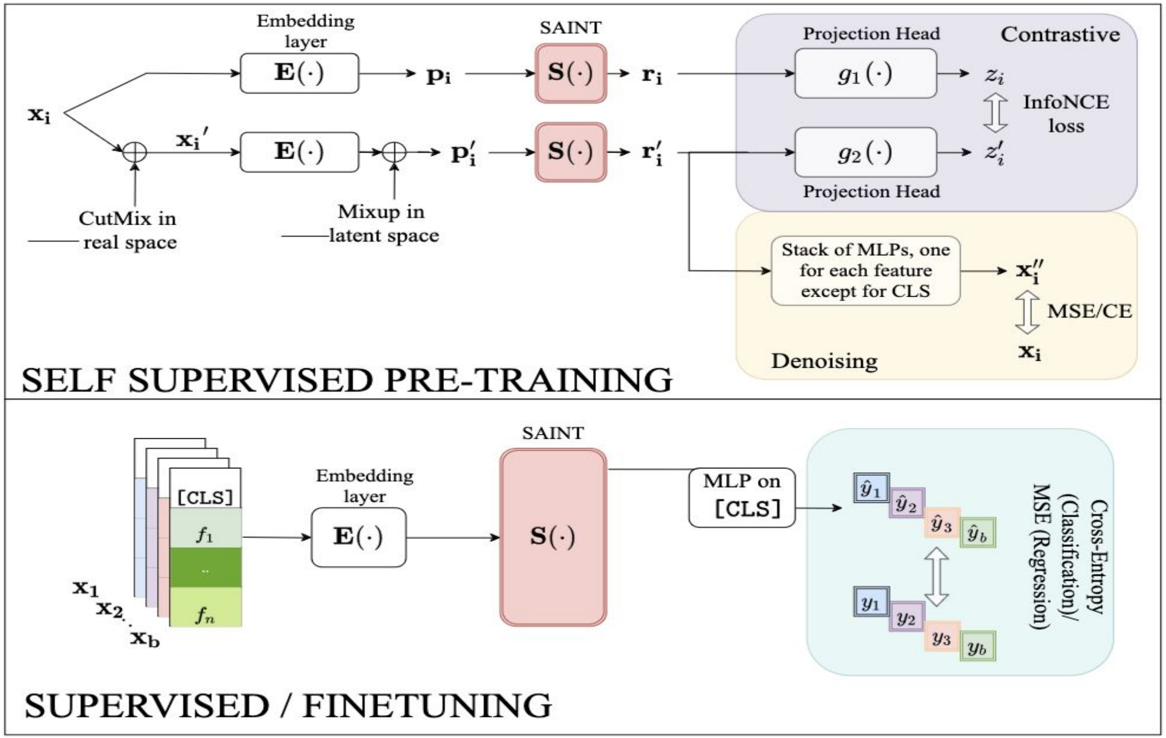


[SCARF: Self supervised contrastive learning using random feature corruption]

Dual attention on compositely augmented tabular datasets



(a)

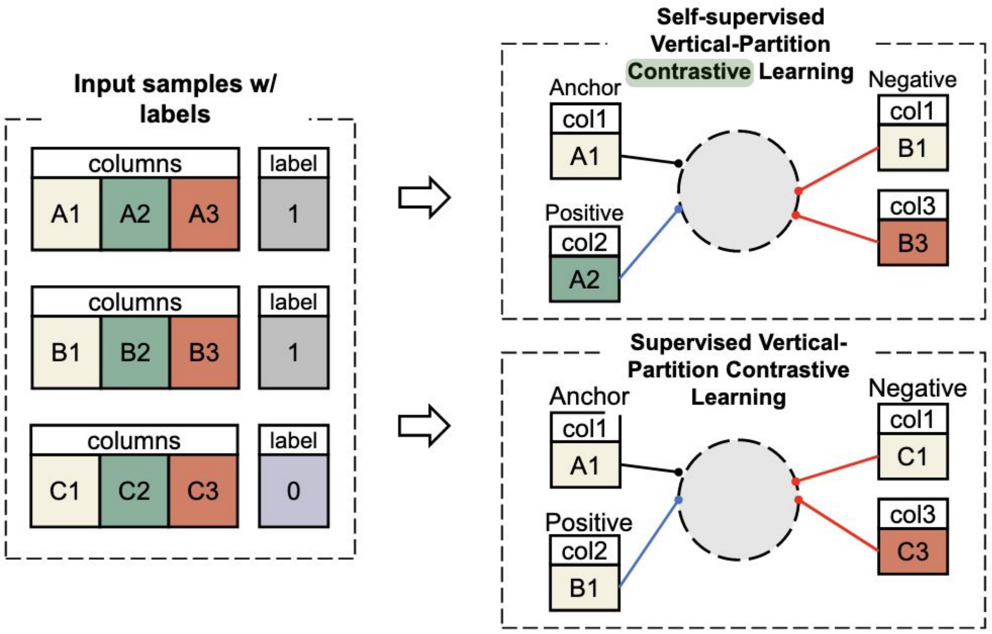


(b)

[SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training]

Column subset with metadata as views across multiple tables

- Multiple tables are tokenized using a common standard and fed into a gated transformer
- The classifier token is used while optimizing the contrastive loss with column subsets as views.
- Significantly better performance for zero-shot and tabular learning



[TransTab: Learning Transferable Tabular Transformers Across Tables]

10 mins break

- Questions from previous parts
- Water or restroom break
- Time to setup the laptop
- Get started with the git repo.
- Any issues in starting those jupyter notebook

Part 3: Getting started with the demo

1) Clone the tutorial repo from
   https://github.com/sandhyat/ContrastiveLearning_Tutorial/tree/main
2) Datasets are included in the repository
3) Make sure you have a python compiler tested in a jupyter notebook

TS2Vec implementation CL_for_TimeseriesDataset.ipynb

Robustness to missingness (parameter 'irregular')

No missingness

```
In [121]: # Linear evaluation of the model
          # modelname can 'knn', 'svm', 'xgbt', 'linear'
          modelname = 'svm'
          out, eval_res = eval_classification(model, train_data, train_labels, test_data, test_labels, eval_protocol=modelname)
```

```
In [122]: # Saving the model and printing the results
          pkl_save(f'{run_dir}/out.pkl', out)
          pkl_save(f'{run_dir}/eval_res.pkl', eval_res)
          print("Dataset : ", dataset, " trained on a ", modelname, " classifier " )
          print('Evaluation result:', eval_res)

          Dataset :  ECG200  trained on a  svm  classifier
          Evaluation result: {'acc': 0.94, 'auprc': 0.9821420927458668, 'auroc': 0.9639756944444444}
```

20%

```
In [100]: # Saving the model and printing the results
          pkl_save(f'{run_dir}/out.pkl', out)
          pkl_save(f'{run_dir}/eval_res.pkl', eval_res)
          print("Dataset : ", dataset, " trained on a ", modelname, " classifier " )
          print('Evaluation result:', eval_res)

          Dataset :  ECG200  trained on a  svm  classifier
          Evaluation result: {'acc': 0.84, 'auprc': 0.9526658932094338, 'auroc': 0.9201388888888888}
```

40%

```
In [111]: # Saving the model and printing the results
          pkl_save(f'{run_dir}/out.pkl', out)
          pkl_save(f'{run_dir}/eval_res.pkl', eval_res)
          print("Dataset : ", dataset, " trained on a ", modelname, " classifier " )
          print('Evaluation result:', eval_res)

          Dataset :  ECG200  trained on a  svm  classifier
          Evaluation result: {'acc': 0.8, 'auprc': 0.9269376459829015, 'auroc': 0.8802083333333334}
```

TS2Vec implementation [CL_for_TimeseriesDataset.ipynb]

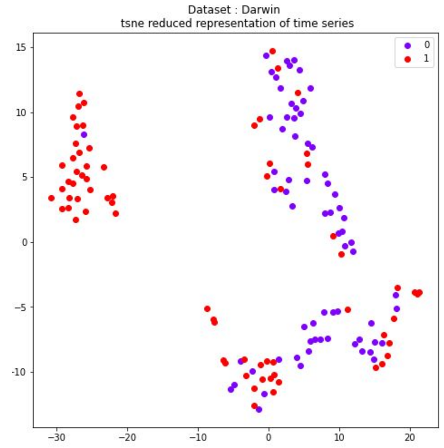
Questions:

1) Impact of batchsize
2) Impact of representation dimension (rep-dims)

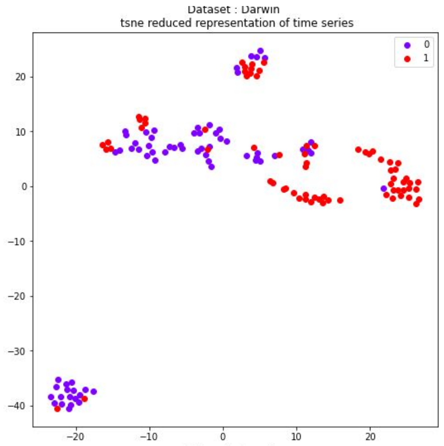
SCARF implementation CL_for_TabularDataset.ipynb

Task: Predicting Alzheimer's present or not
Samples: 139+35
Features: 450

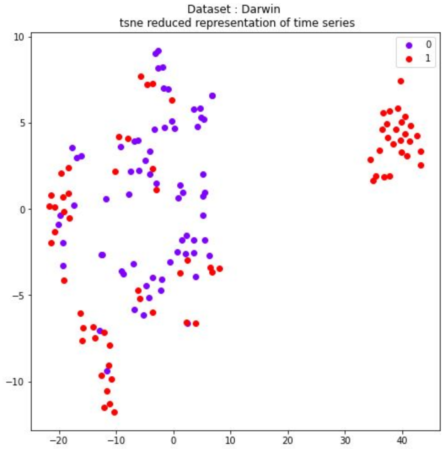
Batch size: 128
Repr_dim: 16



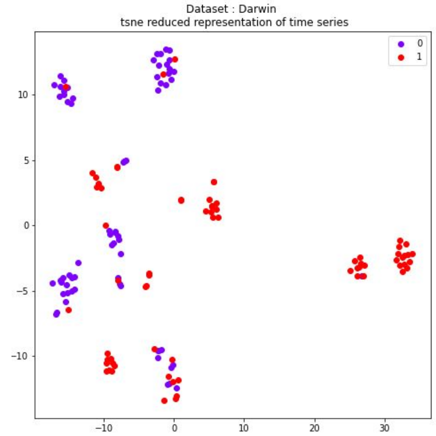
Batch size: 128
Repr_dim: 100



Batch size: 256
Repr_dim: 16



Batch size: 256
Repr_dim: 100



SCARF implementation CL_for_TabularDataset.ipynb

Questions:

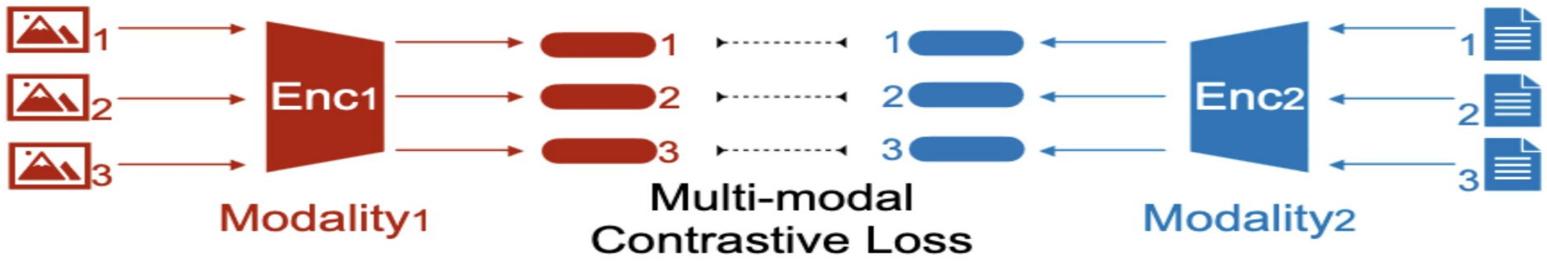
1) Impact of corruption rate on quality of embeddings
2) Label noise robustness threshold

Part 4: Beyond unimodal contrastive learning

- Multimodal contrastive learning
  - Combination of different modalities
  - Shared vs unique information between modalities
  - Teacher student approach in multi-modal CL
  - Modality Gap
  - Issues
- Twist to Conventional CL losses
- Competitive non-contrastive approaches

Multimodal contrastive learning

If two modalities can be treated as two views/augmentations as in unimodal cases, then direct CL application possible.

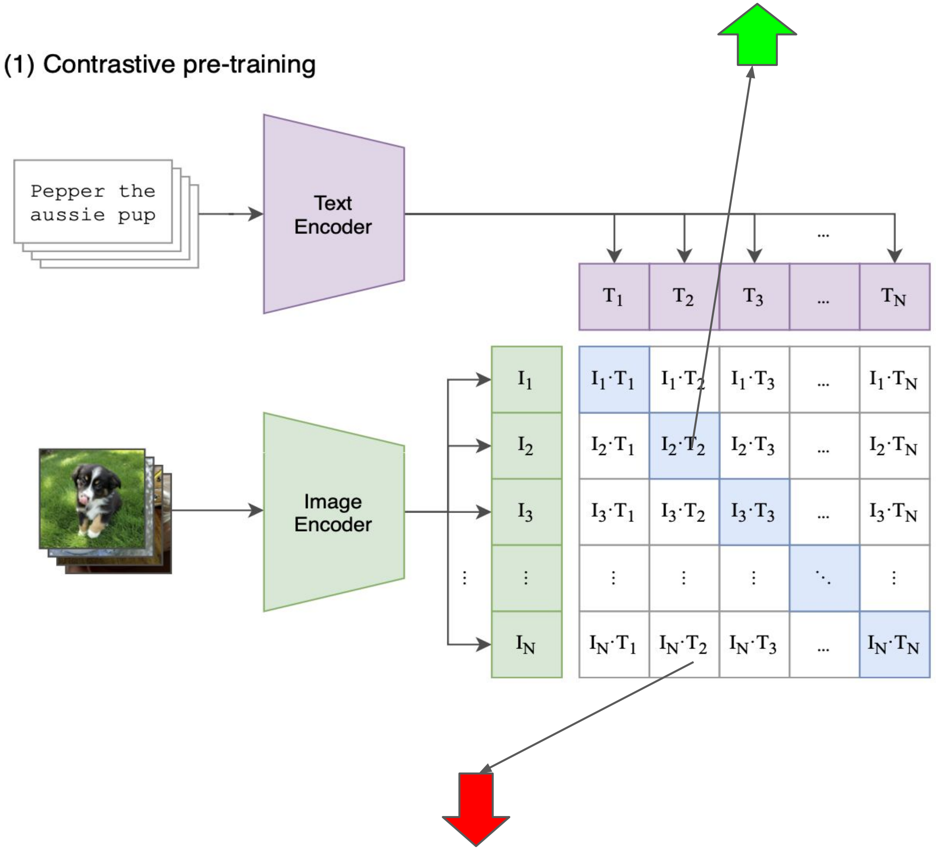


$$
\begin{aligned}
(x_k \in \mathrm{Modal}_1, y_k \in \mathrm{Modal}_2) &\sim D \\
\mathbf{x}_k &= \mathrm{Normalize}(\mathrm{Enc}_1(x_k)) \\
\mathbf{y}_k &= \mathrm{Normalize}(\mathrm{Enc}_2(y_k)) \\
s_{i,j} &= \mathbf{x}_i \cdot \mathbf{y}_j
\end{aligned}
$$

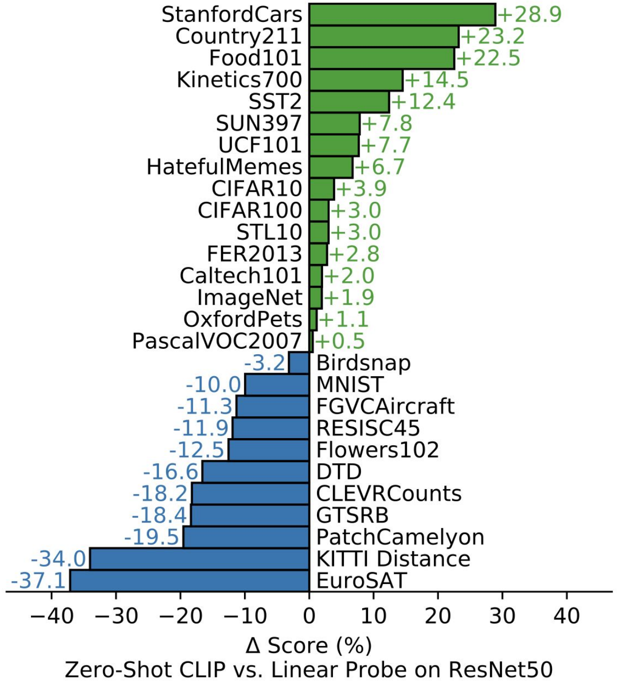
$$
\begin{aligned}
\mathcal{L}_{\mathcal{M}_1 \to \mathcal{M}_2} &= -\frac{1}{N} \sum_i \log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{i,j}/\tau)} \\
\mathcal{L}_{\mathcal{M}_2 \to \mathcal{M}_1} &= -\frac{1}{N} \sum_i \log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{j,i}/\tau)} \\
\mathcal{L} &= \frac{1}{2}(\mathcal{L}_{\mathcal{M}_1 \to \mathcal{M}_2} + \mathcal{L}_{\mathcal{M}_2 \to \mathcal{M}_1})
\end{aligned}
$$

Image sources: Mind the Gap paper, 2022

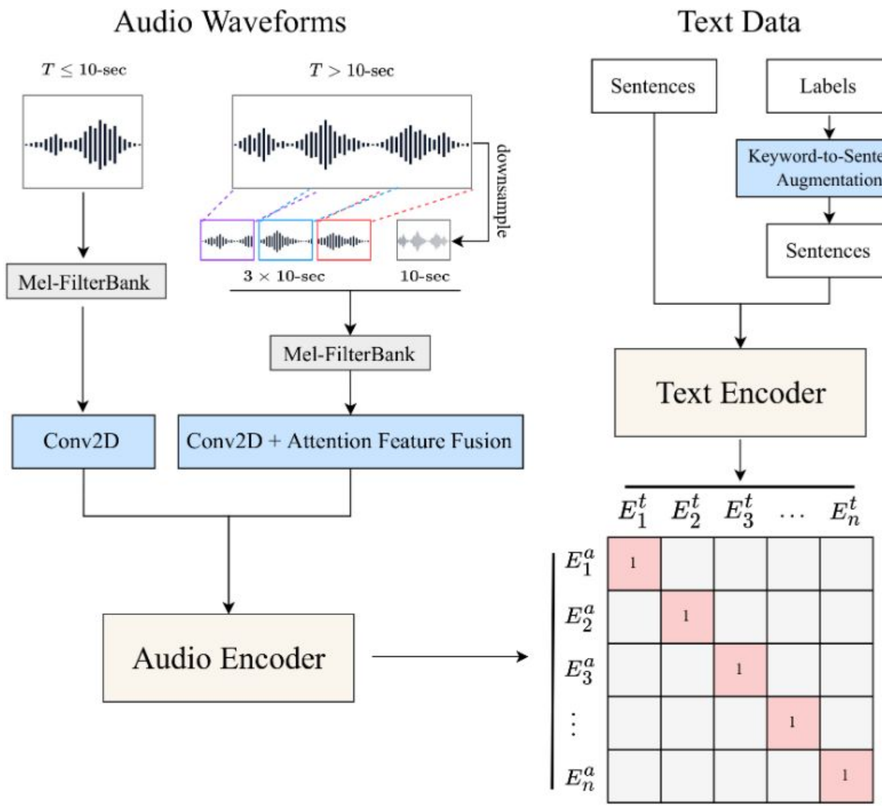
Contrastive Language-Image Pre-training, CLIP



Create a 400 million image-pair dataset
Demonstrated the zero shot performance on various evaluation datasets



Contrastive Language Audio Pre-training (CLAP) (Improved)

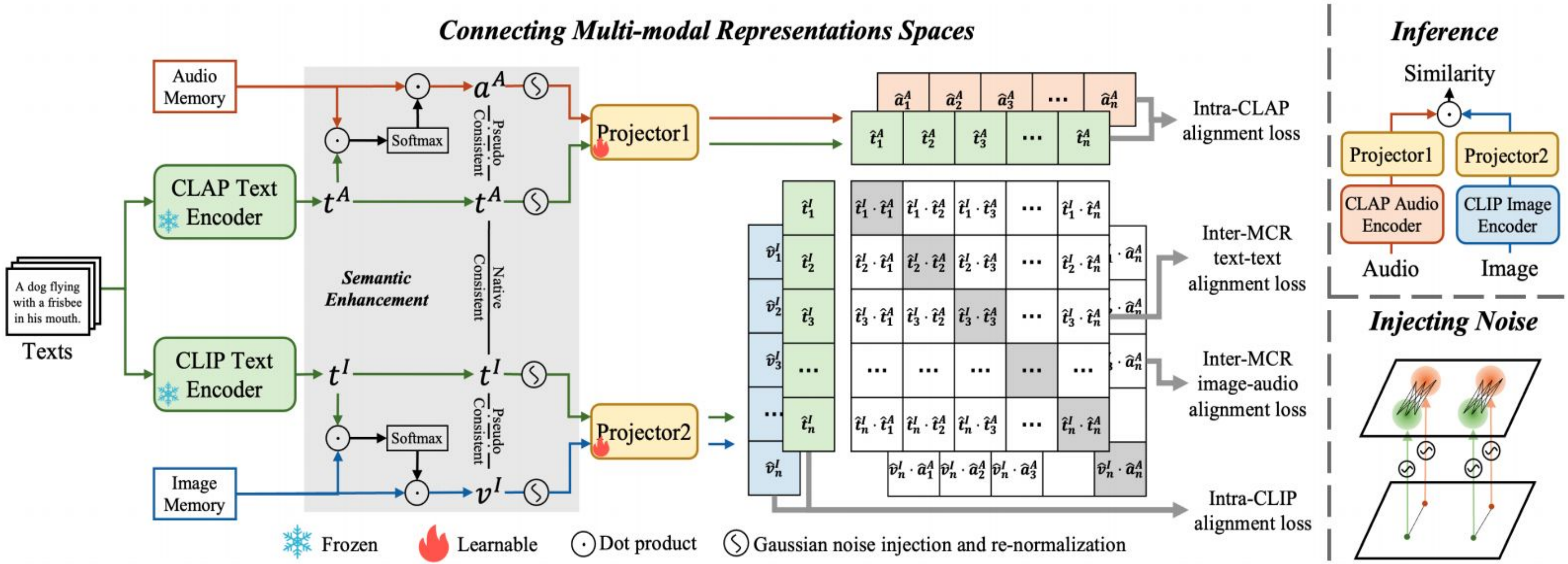


- Feature fusion for variable length signals
- Keyword to sentence augmentation when only tags available

Model	Audio Classification Dataset & Setting				
	ESC-50	US8K	VGGSound		FSD50K
	ZS.	ZS.	ZS.	SV.	SV.
Wav2CLIP [9]	41.4	40.4	10.0	46.6	43.1
AudioClip [3]	69.4	65.3	-	-	-
CLAP [5]	82.6	73.2	-	-	58.6
Ours	89.1	73.2	29.1	**75.4**	64.9
Ours+Fusion	88.0	75.8	26.3	75.3	64.4
Our+K2C Aug.	**91.0**	**77.0**	**46.2**	75.3	59.7
SoTA*	82.6 [5]	73.2 [5]	10.0 [9]	64.1 [25]	**65.6** [26]

Zero shot and fully supervised performance

Connecting multimodal contrastive learning (CLIP + CLAP)



No need of paired data from the modalities to connect

Factorized contrastive learning: Going beyond multi-view redundancy

Assumption: shared information between modalities is relevant for downstream tasks

Low shared info

High unique info

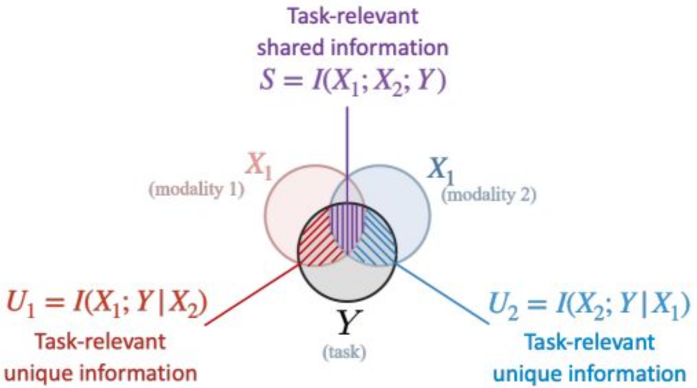
Blanket of snow (IRFL dataset)

Demographics
Comorbidities

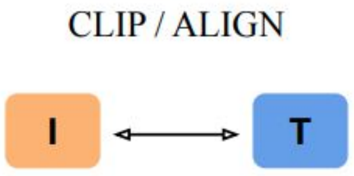
Task	IRFL
Zero-shot CLIP [60]	89.15%
SimCLR [13]	91.57%
Cross+Self [79, 87]	95.18%
FACTORCL-IndAug	92.77%
FACTORCL-SSL	**95.18%**
Fine-tuned CLIP [60]	96.39%
SupCon [39]	89.16%
FACTORCL-SUP	**98.80%**

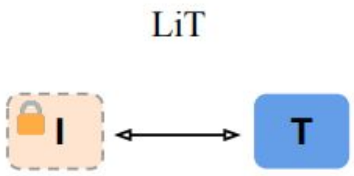
Conditional augmentation

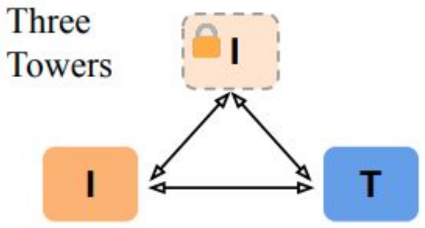
Factorization of the information



Role of supervised pretrained encoders in Multimodal CL







Contrastively trains both the image and text encoders from scratch

Contrastively trains only text encoder while using the embeddings from locked pretrained image encoder

Performs 3-way contrastive comparisons while training the image and text encoder from scratch and also matching the embeddings to a locked pretrained image encoder

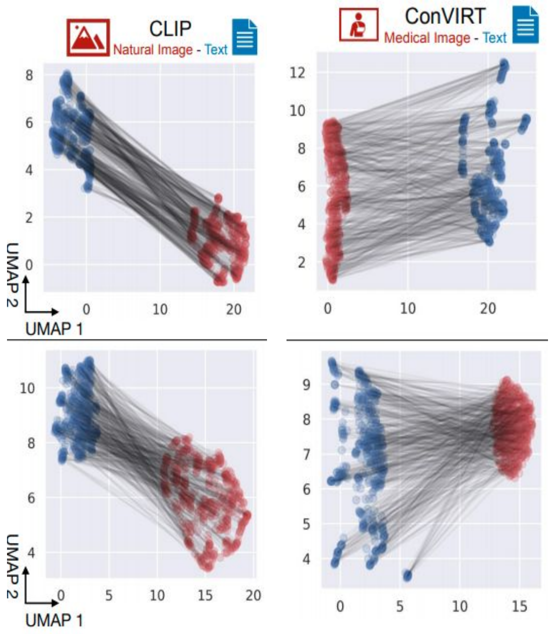
Retrieval and zeroshot classification: 3Towers > LiT
Additional training cost and scale of models: 3Towers < LiT

LiT (https://arxiv.org/pdf/2111.07991.pdf), 3Towers (https://arxiv.org/pdf/2305.16999.pdf)

[Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning]

Pretrained embeddings

Random weights



Reasons:

DNNs create cone effect

Multiple modalities→ different cones

Implications of changing modality gap by shifting the embeddings of CLIP

Zero-shot performance

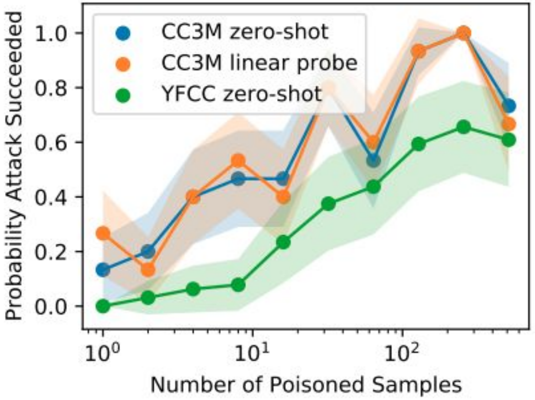
Dataset	Original gap	Modified gap	Direction
Coarse-grained Classification			
CIFAR10	0.9013	**0.9081**	↑
CIFAR100	0.6658	**0.6737**	↓
Fine-grained Classification			
EuroSAT	0.5410	**0.5645**	↓
Optical Character Recognition			
SVHN	0.5389	**0.5396**	↑
HatefulMemes	0.5800	**0.5811**	↑

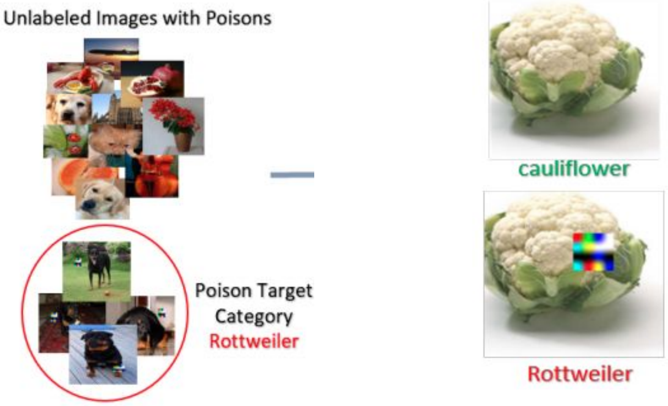
FairFace dataset
MG: 0.82→0.97

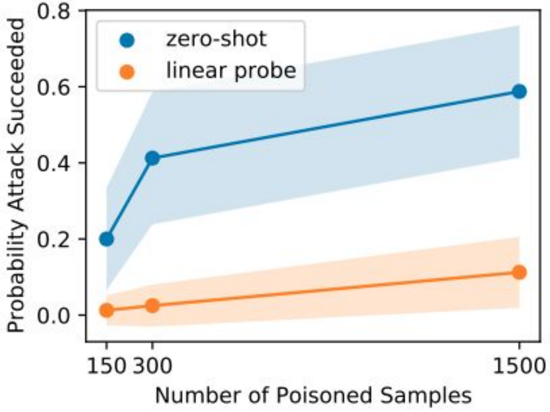
Denigration Biases	Original gap			Modified gap		
	Crime related	Non human	Sum	Crime related	Non human	Sum
Black	1.0%	0.1%	**1.1%**	0.8%	0.1%	**1.0%**
White	15.5%	0.2%	**15.7%**	13.2%	0.4%	**13.7%**
Indian	1.2%	0.0%	**1.2%**	1.1%	0.0%	**1.1%**
Latino	2.8%	0.1%	**2.8%**	1.9%	0.1%	**2.0%**
Middle Eastern	6.3%	0.0%	**6.3%**	5.2%	0.0%	**5.2%**
Southeast Asian	0.5%	0.0%	**0.5%**	0.3%	0.0%	**0.3%**
East Asian	0.7%	0.0%	**0.7%**	0.6%	0.0%	**0.6%**

What can go wrong with multi-modal learnt encoders?

- Uncurated data source
- Higher risk of adversaries



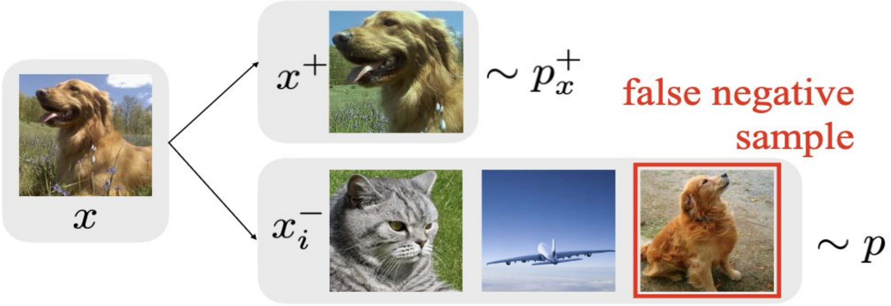


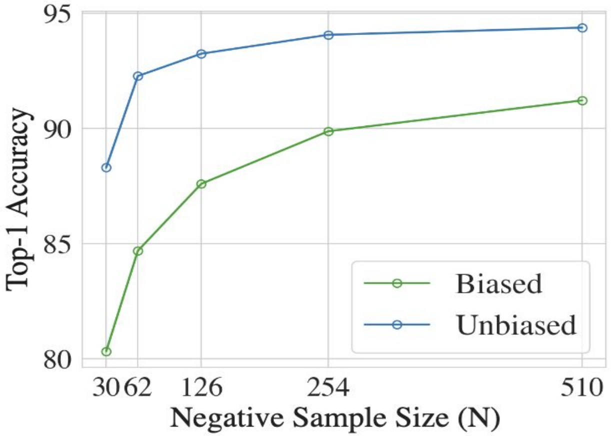


[Poisoning and backdooring contrastive learning]()

Loss modification

What if your negative samples are from the same class?





$$p(x') = \tau^+ p_x^+(x') + \tau^- p_x^-(x')$$

Uniform distribution across c latent classes

Relying on Positive unlabelled learning and estimating the negative sample distribution

Suggests more than one positive example per anchor point.

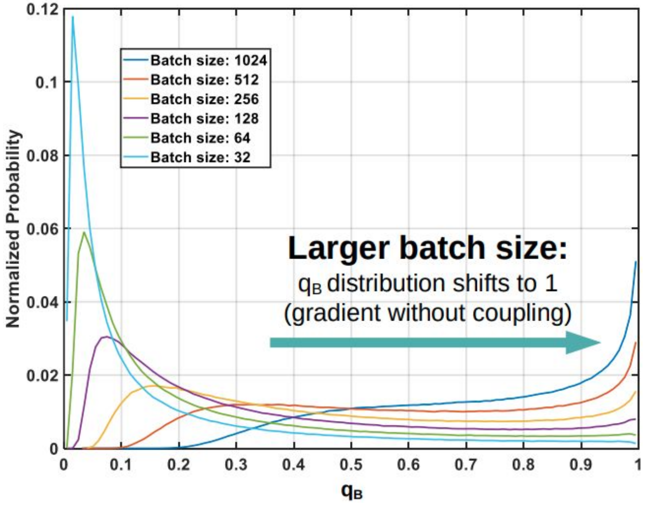
Debiased contrastive learning

Increasing the learning efficiency

InfoNCE gradient

Negative positive coupling : easy classification pretext task

$$\nabla_{z_i^{(1)}} L_i^{(1)} \propto \underbrace{\left(1 - \frac{exp(\langle z_i^{(1)}, z_i^{(2)}\rangle/\tau)}{exp(\langle z_i^{(1)}, z_i^{(2)}\rangle/\tau) + \sum_{l\in\{1,2\}, j\in[1,N], j\neq i} exp(\langle z_i^{(1)}, z_j^{(l)}\rangle/\tau)}\right)}_{q_{B,i}^{(1)}} \left(z_i^2 - \sum_{l\in\{1,2\}, j\neq i} w_j^{(l)} z_j^{(l)}\right)$$



Reverse engineer by removing the multiplier

$$L_{DC,i}^{(k)} = -\langle z_i^{(1)}, z_i^{(2)}\rangle/\tau + log(\sum_{l\in\{1,2\}, j\in[1,N], j\neq i} exp(\langle z_i^{(k)}, z_j^{(l)}\rangle/\tau)$$

Decoupled Contrastive Learning

Expert features guiding similar to class labels for learning representation

- Dataset form : (X, F, Y) where F = {f_1,..,f_N}

$$s_{ij} := 1 - \frac{\|f_i - f_j\|_2}{\max_{k,l} \|f_k - f_l\|_2}$$

Similarity between expert features

- Interested in learning an encoder **E** such that the if the expert features of two points are similar then their corresponding representations should be similar too and vice versa.

$$D_{ij} := \|E(x_i) - E(x_j)\|_2$$

$$\tau \in \mathbb{R}^+$$

Temperature parameter controls hard negative mining

- No augmentations or large batch sizes

$$\mathcal{L}^{\tau}_{ExpCLR}(E(X), F) = \tau \log \left[\sum_{i,j=1}^{N} \frac{\exp\left(\frac{L_{ij}}{\tau}\right)}{N^2} \right]$$

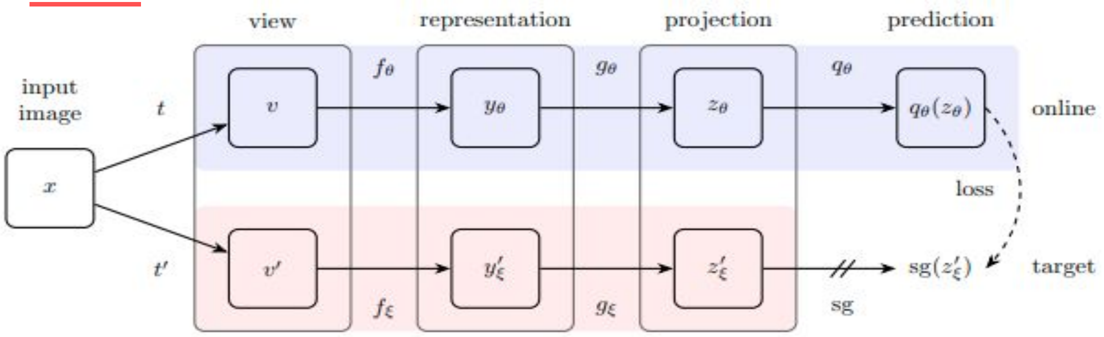
$$L_{ij} := ((1 - s_{ij})\Delta - D_{ij})^2$$

Continuous version of pair-loss

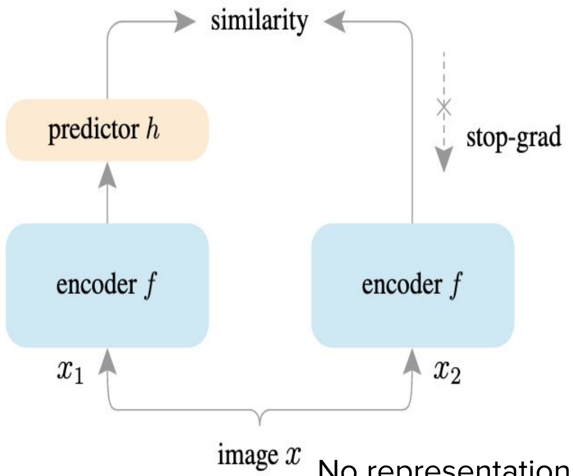
Utilizing Expert Features for contrastive learning of time-series representations

Non-contrastive learning (without negative examples)

BYOL

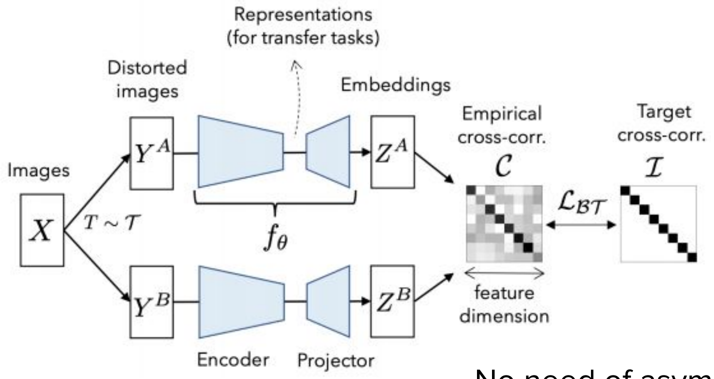


SimSiam



No representation collapse

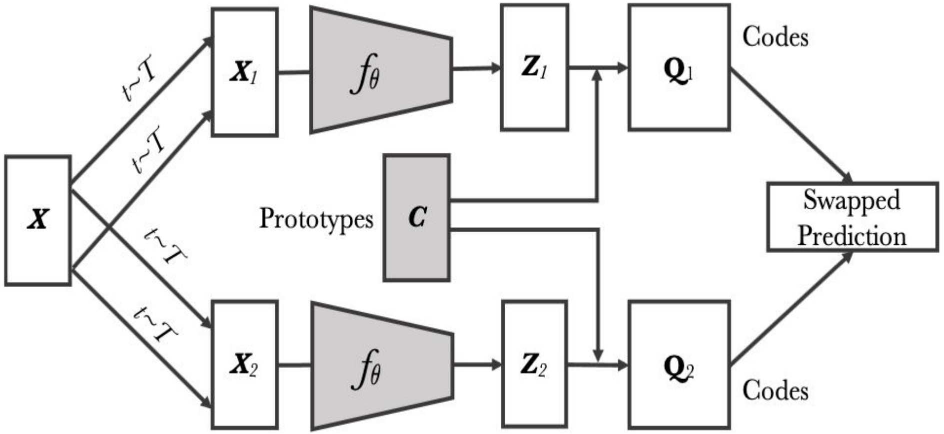
Barlow Twins



No need of asymmetric prediction network or stop gradient

Non contrastive learning (clustering based)

- Avoid need of large batches or memory banks
- Propose a new augmentation strategy called multi-crop
- Improvement on ImageNet



Predicting Q1 from z2 with cross entropy minimization and vice versa

Source: https://github.com/facebookresearch/swav

[Unsupervised Learning of Visual Features by Contrasting Cluster Assignments]()

[Divide and Contrast: Self supervised learning from uncurated data]()

What after this?

1) From theory perspective
    a) Geometry based
    b) Mutual information based (FLO estimators, other estimates)
2) From application perspective
    a) More adaptive augmentations (Rethinking rotation, Triplet teaching)
    b) Explanations in contrastive learning (in NLP, Consistent, CoRTX)
3) Multi-modality to other domains

.....

Thank you!