Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data

Paul Hager, Martin J. Menten, Daniel Rueckert

Technical University of Munich, Imperial College London
2023 IEEE/CVPR

January 31, 2025

Presented by Angel Huang

Introduction

Purpose

- Develop a multimodal contrastive learning framework for medical applications, particularly in scenarios with limited labeled data

Intuition

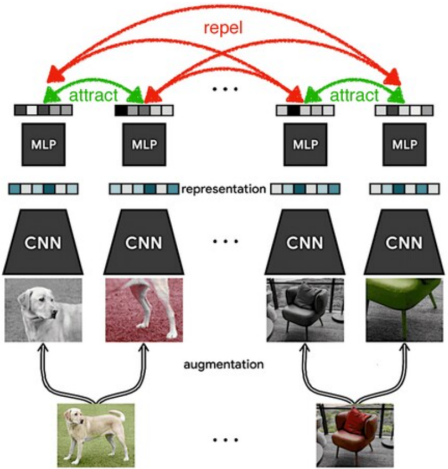
- Medical datasets often include multimodal data, such as images (e.g., MRI scans) and tabular data (e.g., clinical and demographic information). Combining these can potentially enhance the performance of deep learning models.
- **Problem**: Medical datasets, especially for rare diseases, often have limited labeled data, making it challenging to train supervised deep learning models effectively . Multiple modalities of data are not always available or easy to use at inference.
- **Solution**: Pretrain models multimodally and predict unimodally (with images). Use self-supervised pretraining that combines SimCLR (for images) and SCARF (for tabular data) to boost performance in predicting rare conditions.

Applications

- Predicting (rare) medical conditions
- Predicting car models

Background: SimCLR

Simple Framework for Contrastive Learning of Visual Representations



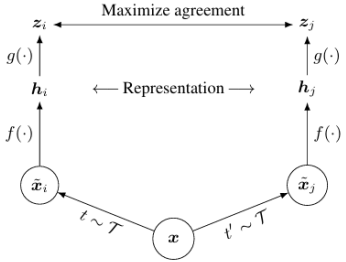


Figure 2. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder network $f(\cdot)$ and a projection head $g(\cdot)$ are trained to maximize agreement using a contrastive loss. After training is completed, we throw away the projection head $g(\cdot)$ and use encoder $f(\cdot)$ and representation $\boldsymbol{h}$ for downstream tasks.

Figure: SimCLR (Chen, et. al., 2020)

Background: SimCLR NT-Xent Loss (normalized temperature-scaled cross entropy loss)

$$\ell_{i,j} = -\log \frac{\exp(\cos(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\cos(z_i, z_k)/\tau)} \tag{1}$$

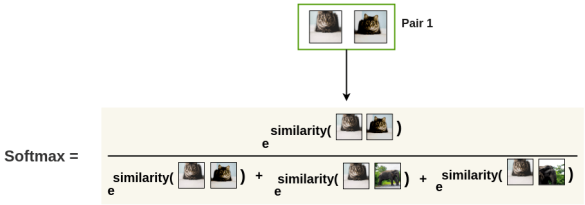


Figure: Amit Choudhary's Blog

- **Balance** between pulling positives and pushing negatives apart
- **Preventing Collapse:** embeddings spread too far apart rather than forming meaningful clusters

Background: SCARF

Self-Supervised Contrastive Learning using Random Feature Corruption

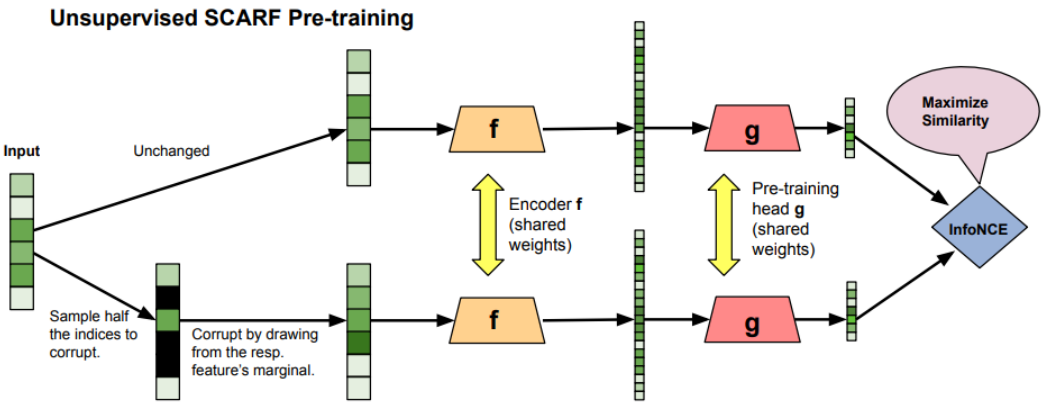
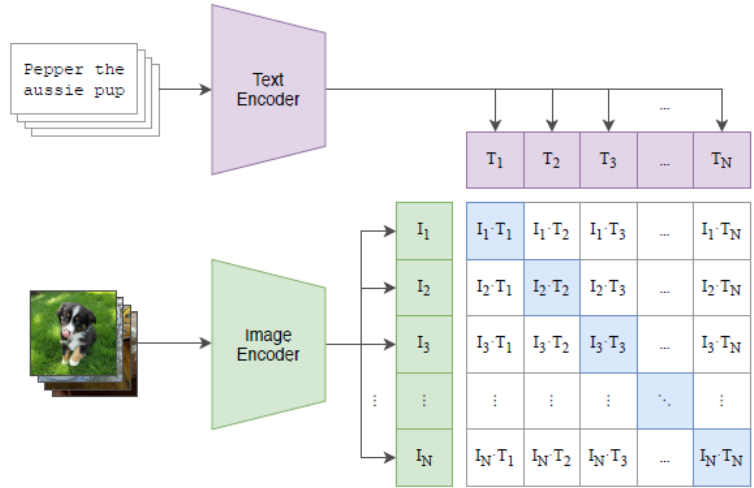


Figure: SCARF (Bahri, et. al., 2022)

Background: CLIP

Contrastive Language-Image Pre-training



Methods: CLIP Loss

The projection of an image z_{ji} and a tabluar sample z_{jt} is defined as:

$$\begin{aligned} z_{ji} &= f_{\phi_I}(f_{\theta_I}(x_{ji})) \\ z_{jt} &= f_{\phi_T}(f_{\theta_T}(x_{jt})) \end{aligned} \tag{2}$$

For a batch of size N, the loss for the imaging modality is ($\ell_{t,i}$ is calculated analagously):

$$\ell_{i,t} = -\sum_{j \in N} \log \frac{\exp(\cos(z_{ji}, z_{jt})/\tau)}{\sum_{k \in N, k \neq j} \exp(\cos(z_{ji}, z_{kt})/\tau)} \tag{3}$$

where τ is the temperature parameter. A lower τ makes the model focus more on hard negatives (samples that are close but should be dissimilar). The total loss is:

$$L = \lambda \ell_{i,t} + (1 - \lambda)\ell_{t,i} \tag{4}$$

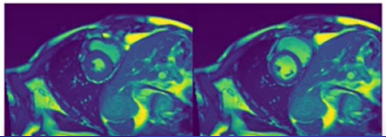
Where λ balances the contribution of the image and tabular losses.
Note: Exclude the positive pair from the denominator to avoid self-competition.

Methods: Dataset1

UK BioBank

- 50k imaging subjects
  - selected cardiac MRI
  - end-systolic, end-diastolic time points
- 1k+ tabular features
  - lifestyle, questionnaire, interview, physical measures, etc.
  - selected 117 with published cardiac effect
  - diverse features: age, diet, smoker, BP, BMI, fitness, alcohol, anxiety, etc.
- Prediction targets (based on ICD codes):
  - Myocardial Infarction (3% rate)
  - Coronary Artery Disease (CAD) (6% rate)
- Subjects: 29k training, 7k validation, 4k test-pairs
- Finetune train splits were balanced using all positive subjects and a static set of randomly chosen negative subjects. The test and validation sets were left untouched.

Smoker	Age	BP	BMI	Sex	Fitness	Alcohol
FALSE	62	150/90	29.2	Male	High	Moderate

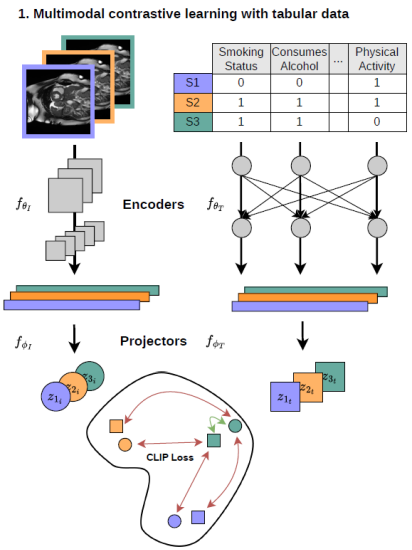


Methods: Dataset2

Data Visual Marketing (DVM) Cars

- Created from used car advertisements
- 1M images of cars from various angles
- Tabular: sales and technical data, excluding brand and model year
- Paired tabular data with a single random image from each advertisement
- Cars: 70k training, 17k validation, 88k test-pairs
- Prediction target: Car Model (286 classes)
- Missing features: after normalization, use iterative multivariate imputer, modeling missing features as a function of existing features over multiple imputation rounds.

Methods: Multimodal CL



Implementation

- Images (based on SimCLR)
  - Imaging encoder: ResNet50 that generate 2048-d embeddings
  - Imaging projector: 1-layer MLP with hidden layer of size 2048
- Tabular (based on SCARF)
  - Tabular encoder: 2-layer MLP with hidden layer of size 2048, generate 2048-d embeddings
  - Tabular projector: projects directly from the embeddings with a fully connected layer, then size 128
- CLIP Loss pull projections from a subject together and against all other subjects in batch
- After pretraining, the projection head is removed and a fully connected layer to the output class nodes is added

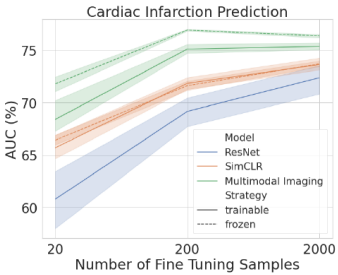
Results

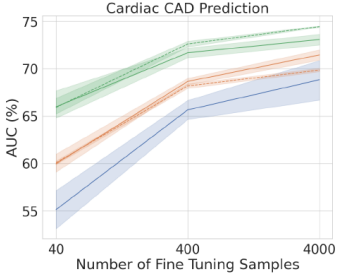
Multimodal pertaining improves unimodal prediction

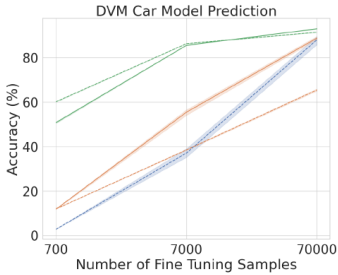
Model	AUC (%) Frozen / Infarction	AUC (%) Trainable / Infarction	AUC (%) Frozen / CAD	AUC (%) Trainable / CAD	Top-1 Accuracy (%) Frozen / DVM	Top-1 Accuracy (%) Trainable / DVM
Supervised ResNet50	72.37±1.80	72.37±1.80	68.84±2.54	68.84±2.54	87.97±2.20	87.97±2.20
SimCLR	73.69±0.36	73.62±0.70	69.86±0.21	71.46±0.71	65.48±0.48	88.76±0.81
BYOL	69.18±0.43	70.69±2.09	66.91±0.19	70.66±0.22	59.73±0.28	89.18±0.90
SimSiam	71.72±0.18	72.31±0.26	67.79±0.12	70.13±0.35	22.11±2.83	87.43±0.88
BarlowTwins	66.06±1.11	71.35±1.23	62.90±0.23	69.63±0.58	52.57±0.08	85.47±0.82
Multimodal Imaging	**76.35±0.19**	**75.37±0.43**	**74.45±0.09**	**73.08±0.75**	**91.43±0.13**	**93.00±0.18**

Results

Multimodal pertaining is beneficial in low-data regimes







Methods: Integrated Gradients

Integrated gradients (IG): measure the importance of individual features in generating the embeddings.

$$IG_i(x) = (x_i - x_i^{'}) \int_0^1 \frac{\partial F(x^{'} + \alpha(x - x^{'}))}{\partial x_i} d\alpha \tag{5}$$

Where:

- x is the input feature vector.
- x' is the baseline (usually a zero vector).
- F is the model (in this case, the tabular encoder).
- α is a scaling parameter that interpolates between the baseline x' and the input x.
- The integral computes the average gradient of the model output with respect to the feature x_i along the path from x' to x.

Methods: Integrated Gradients and Explainability

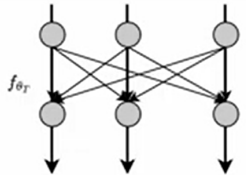
Results: Integrated Gradients and Explainability

Baseline

Smoking Status	Consumes Alcohol	...	Physical Activity
0	0		0

Original Tabular Entry

	Smoking Status	Consumes Alcohol	...	Physical Activity
S1	0	0		1
S2	1	1		1
S3	1	1		0

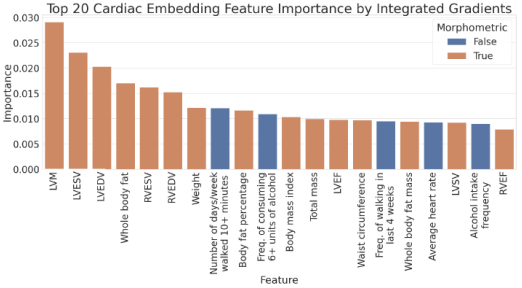


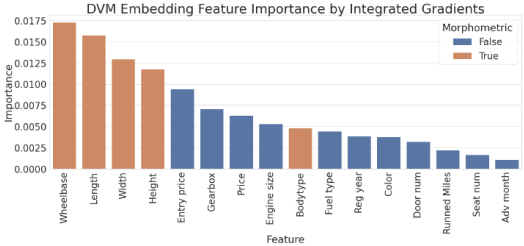
Integrated Gradients

Smoking Status	Consumes Alcohol	...	Physical Activity
0.145	0.678		-0.365

Results: Integrated gradients and explainability

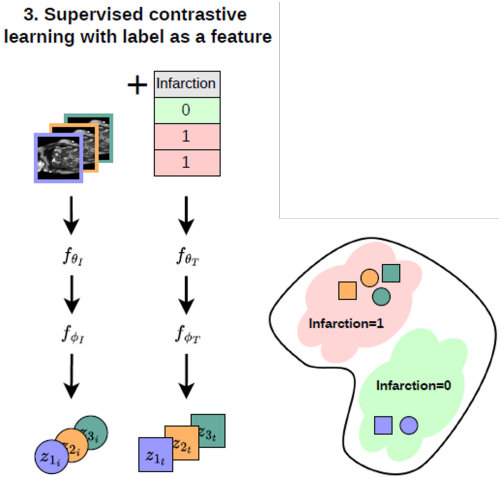
Morphometrics (size and shape) features are important for tabular embeddings





- This is also supported by the Guided Gra-CAM results which shows that left ventricle is important in image processing.
- If only uses morphometric features, they achieve similar loss as using all data.
- This morphometric features are so important that they found if only use morphometric features, they even improve DVM accuracy compared to using all features.

Results: Label as a Feature (LaaF)



- Appending the ground truth label as a tabular feature during pretraining.
- During fine-tuning, the label feature is removed, and the model is trained to predict the label based on the learned representations.

Contrastive	Label Used	Model	AUC (%) Infarction	AUC (%) CAD	Top-1 Accuracy (%) DVM
✓		Multimodal Imaging Baseline	76.35±0.19	**74.45±0.09**	91.43±0.13
	✓	Supervised ResNet50	72.37±1.80	68.84±2.54	87.97±2.20
✓	✓	Label as a Feature (LaaF)	**76.60±0.42**	73.76±0.31	93.56±0.08
✓	✓	FN Elimination	75.38±0.06	72.45±0.09	92.39±0.18
✓	✓	FN Elimination + LaaF	75.30±0.05	72.39±0.08	94.07±0.05
✓	✓	SupCon	—	—	93.82±0.11
✓	✓	SupCon + LaaF	—	—	**94.40±0.04**

Figure: LaaF Boosts Supervised Contrastive Strategies

Conclusions

- Presents a novel multimodal contrastive learning framework that combines tabular data and imaging data for self-supervised pretraining
- Combines SimCLR and SCARF, two leading contrastive learning strategies
- Morphometric features are found to be important during the pre-training process via the integrated gradients analysis.
- A new way for supervised contrastive learning with label as a feature.

Conclusions

Strength

- Highly relevant for medical applications, where multimodal data are often available but underutilized in combination.
- Self-supervised which is especially useful for rare diseases where data is scarce. The model can be pretrained on large, unlabeled datasets (e.g., BioBanks) and fine-tuned on smaller, labeled datasets for rare diseases
- Performs well in low-data regimes

Weakness

- Only included White subjects from the BioBank dataset due to other race is underrepresented.
- Only examined classification tasks, future on segmentation and regression.
- Relies on the availability of high-quality tabular data (BioBank), which may not always be feasible in clinical settings (missingness)
- Multimodal pretraining can be computationally expensive, especially when dealing with large datasets

Recommendations

- Results compelling? – Yes.
- Recommend reading? – Yes, clearly written.
- Implement in my research? – Worth trying for Tabular + Notes data
- Github: https://github.com/paulhager/MMCL-Tabular-Imaging