Explainable Representations in Self-Supervised Learning for Image Understanding

Darpan Gaur, Abhinav Vsk Kompella , Aditya Bacharwar, Naik Avaneesh Radhakrishnan
Indian Institute of Technology Hyderabad
{co21btech11004, es21btech11002, es21btech11003, es21btech11022}@iith.ac.in

Abstract

Self-supervised learning (SSL) has emerged as a powerful alternative to supervised learning, enabling models to learn meaningful representations from unlabeled data. However, these learned representations often lack interpretability, making it difficult to understand their decision-making process. In this project, we aim to investigate the explainability of SSL-based feature representations by integrating explainable artificial intelligence (XAI) techniques. Specifically, we will leverage contrastive learning methods such as SimCLR and employ Grad-CAM and similar concept-based techniques to interpret learned representations. Our objective is to evaluate the trade-off between explainability and performance, providing insights into how SSL embeddings capture meaningful features. By improving the interpretability of self-supervised representations, we aim to enhance their applicability in critical domains such as healthcare, autonomous systems, and scientific discovery.

1. Introduction

Supervised learning has traditionally been the dominant paradigm in computer vision, requiring large-scale labeled datasets for training deep neural networks. However, acquiring labeled data is costly and time-consuming. Self-supervised learning (SSL) addresses this issue by enabling models to learn feature representations from unlabeled data using pretext tasks. Contrastive learning, particularly SimCLR, has shown remarkable success in this domain by maximizing the similarity between augmented views of the same image while pushing apart different images.

Despite the success of SSL in learning powerful feature representations, a critical limitation remains: lack of interpretability. Unlike supervised learning, where class labels provide a semantic understanding of features, SSL representations are often opaque, making it challenging to understand what the model has learned. Explainability is crucial for:

- Trust and Transparency: Understanding SSL feature representations can improve trust in AI systems, particularly in safety-critical applications.
- Bias Detection: XAI techniques can help reveal biases in SSL models, preventing unintended discriminatory behavior.
- Downstream Task Adaptability: Interpretable SSL representations can improve transferability to various tasks by ensuring that learned features align with meaningful concepts.

2. Problem Statement

Our project aims to bridge the gap between self-supervised learning and explainable AI by investigating how to interpret SSL-learned representations. Specifically, we will:

- Pre-train an SSL model using contrastive learning (e.g., SimCLR[1] or MoCo).
- Apply Grad-CAM[3], RELAX[4] or other explainability techniques to visualize the important regions in SSL feature maps.
- Explore concept-based explainability techniques (e.g., TCAV, ProtoPNet) to align SSL representations with human-interpretable concepts.
- Evaluate the trade-off between interpretability and performance.

3. Literature Review

3.1. SimCLR: Simple Framework for Contrastive Learning of Visual Representations

SimCLR[1] is a contrastive self-supervised learning framework that learns visual representations of images without requiring labeled data. The paper explores various components of the model and showcases the importance of each component via numbers from experiments.

The first key component of the SimCLR model is the augmentations it applies that leads to the positive and negative pairs needed to perform contrastive learning. The mentioned model stochastically chooses two augmentations

from the chosen three augmentations to get the training pairs. The paper also mentions various other augmentations but reinforces the choice of the three augmentations (random crop, random colorization, and random Gaussian noise).

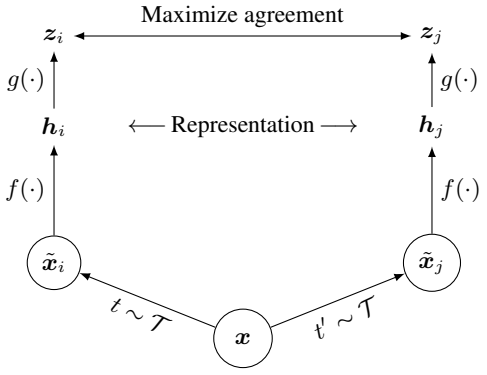

Figure 1. A simple framework for contrastive learning of visual representations. Two separate data augmentation operators are sampled from the same family of augmentations ($t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$) and applied to each data example to obtain two correlated views. A base encoder 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.

Batches of such augmented images are fed to the encoder component where a ResNet architecture was used. The paper then experimented with different projection layers and showed via experimentation that a non-linear projection component worked the best. The contrastive loss is then applied to these projected outputs. SimCLR shows various key components that can be used to extract representations in a self-supervised manner which can be made use of to come up with an explainable self-supervised learning model.

3.2. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (Selvaraju et al., ICCV 2017)

Gradient-weighted Class Activation Mapping (Grad-CAM)[3] is a widely adopted method that provides visual explanations for model predictions by leveraging gradient information to highlight the most important regions in an input image. Unlike earlier Class Activation Mapping (CAM) approaches, which required model modifications, Grad-CAM is architecture-agnostic and can be applied to various types of model architectures.

Grad-CAM generates class-specific saliency maps by computing the importance of feature maps in a CNN for a given target class. The method consists of the following steps:

1. **Forward Pass:** The input image is processed through the CNN, and feature maps A^k from a selected convolutional layer are extracted.

2. **Gradient Computation:** The gradients of the target class score y^c are computed with respect to each feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where Z represents the spatial dimensions of the feature map.

3. **Weighted Feature Map Combination:** A weighted sum of the feature maps is computed using the importance weights α_k^c:

$$L_c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

The ReLU function ensures that only positive influences contribute to the visualization, preventing the inclusion of irrelevant or misleading features.

4. **Heatmap Generation and Overlay:** The resulting saliency map is re-sampled to the original image resolution and overlayed on the input image to highlight the key regions that contribute to the model decision.

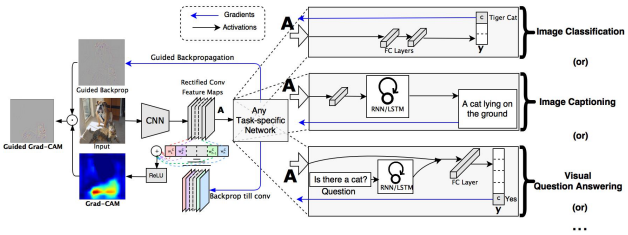

Figure 2. Grad-CAM visualization of a ResNet-50 model for the class "zebra." The heatmap highlights the regions in the input image that are most relevant for the model's prediction.

Although Grad-CAM is highly effective in supervised learning scenarios, its applicability to self-supervised learning (SSL) models presents significant challenges. SSL approaches, such as SimCLR and MoCo, do not rely on explicit class labels, making the traditional Grad-CAM formulation which depends on class scores for gradient computation. To overcome these challenges, we propose several modifications to adapt Grad-CAM for self-supervised learning:

1. **Using Contrastive Loss Gradients:** Instead of computing gradients with respect to class scores, we leverage gradients from the contrastive loss function (e.g., InfoNCE loss). This allows us to identify which regions of an image contribute the most to effective representation learning.

2. **Feature Importance Analysis at Intermediate Layers:** Since SSL models do not use final classification layers,

Grad-CAM can be applied to intermediate feature representations before projection layers to understand which regions are most influential in encoding meaningful representations.

3. **Downstream Evaluation via Linear Probing:** To bridge SSL and interpretability, a linear classifier can be trained on SSL embeddings for a downstream task (e.g., object classification). Grad-CAM can then be used to visualize how well SSL representations capture class-discriminative features.

3.3. RELAX: Representation Learning Explainability

- Representation Learning Explainability (RELAX)[4] is a novel framework for explaining representations that also quantify their uncertainty.
- It measures the change in the representation of an image when compared with the masked version of the image.
- The central idea is that when informative parts are masked out, the representation should change significantly, i.e., the similarity between masked and unmasked representations should be low when informative parts are masked out and high when uninformative parts are masked out.

Let $\mathbf{X} \in \mathbb{R}^{H \times W}$ represents image of dimensions $H \times W$ and f be the feature extraction model that extracts representation $\mathbf{h} = f(\mathbf{X}) \in \mathbb{R}^D$.

To create masked images, we apply a stochastic mask $\mathbf{M} \in [0,1]^{H \times W}$, where M_{ij} is drawn from some distribution. The masked representation is given by $\bar{\mathbf{h}} = f(\mathbf{X} \odot \mathbf{M})$, where $\odot$ denotes element-wise multiplication.

The similarity between masked and unmasked representations is measured using cosine similarity,

$$s(\mathbf{h}, \bar{\mathbf{h}}) = \frac{\langle \mathbf{h}, \bar{\mathbf{h}} \rangle}{\|\mathbf{h}\| \|\bar{\mathbf{h}}\|},$$

, where $\|\cdot\|$ denotes the Euclidean norm of a vector.

Importance R_{ij} of pixel (i, j) is defined as:

$$R_{ij} = \mathrm{E}_{\mathbf{M}}\big[s(\mathbf{h}, \bar{\mathbf{h}}) M_{ij}\big].$$

$$\bar{R}_{ij} = \frac{1}{N} \sum_{n=1}^{N} s(\mathbf{h}, \bar{\mathbf{h}}_n) M_{ij}(n).$$

The uncertainty U_{ij} of pixel (i, j) is defined as:

$$U_{ij} = \mathrm{Var}_{\mathbf{M}}[s(\mathbf{h}, \bar{\mathbf{h}}) M_{ij}].$$

$$\bar{U}_{ij} = \frac{1}{N} \sum_{n=1}^{N} (s(\mathbf{h}, \bar{\mathbf{h}}_n) - \bar{R}_{ij})^2 M_{ij}(n).$$

The figure 3, shows an example where RELAX is used to investigate the explanations and uncertainities for a selection of widely used feature extraction models. Red indicates high values, and blue indicates low values. In the figure, two birds are present, one prominently displayed in the foreground and another in the background. The plot shows all models emphasize the bird in the foreground with low uncertainty. However, the emphasis on the bird in the background varies with different degrees of uncertainty.

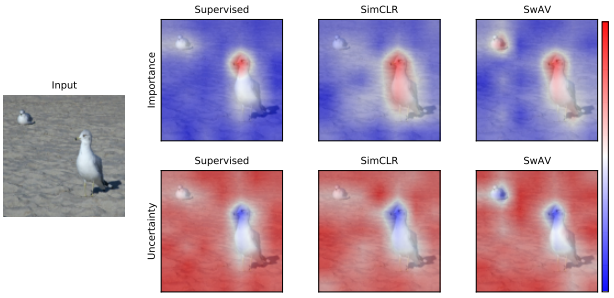

Figure 3. RELAX explations and uncertainty estimates for a VOC image.

3.4. Prototypical Networks for Unsupervised Disentanglement

ProtoVAE is a novel Variational Autoencoder (VAE)-based framework[2] that incorporates Prototypical Networks to enforce structured disentanglement of latent representations in a completely self-supervised manner. The model achieves this by applying interventions in latent space, where a single latent dimension is modified while keeping others unchanged. The VAE acts as the generative component, while the Prototypical Network clusters latent representations, ensuring that each dimension encodes a distinct and interpretable factor of variation. A local isometry constraint is introduced to maintain smooth and meaningful changes in data space when traversing latent dimensions. Additionally, a discriminator network regularizes the latent space, ensuring that modified representations still map to valid data distributions. Empirical results demonstrate state-of-the-art disentanglement on benchmark datasets such as dSprites, 3DShapes, MPI3D, and CelebA.

In our project, we aim to improve interpretability in Self-Supervised Learning (SSL) representations, particularly for contrastive learning methods like SimCLR. We can leverage ProtoVAE's self-supervised latent disentanglement framework to introduce structured clustering in SSL embeddings, making them more explainable. By integrating a Prototypical Network, we can group semantically similar representations, providing insights into how SSL learns features. Furthermore, the interventional approach used in ProtoVAE can be adapted to analyze individual feature importance in SSL embeddings. By combining this with Grad-CAM and concept-based methods, we can provide more transparent and interpretable explanations of SSL representations, bridging the gap between performance and trust in

self-supervised models.

4. Dataset

Since our objective is to obtain explainable representations using self-supervised learning, we are not restricted to a specific dataset. Instead, we aim to experiment with datasets that provide diverse visual concepts, enabling a better evaluation of both representation learning and explainability. For this purpose, we consider using either the CIFAR-10 or ImageNet datasets. The choice of dataset will depend on the computational constraints and the level of detail required for explainability analysis.

5. Future Works

We have currently explored key components from multiple research papers to understand both self-supervised learning and explainability techniques. Building on these insights, we will try to develop a hybrid framework that integrates elements from these approaches to obtain explainable representations using self-supervised learning. Our goal is to design a model that not only learns robust representations without labels but also provides meaningful and interpretable visual features.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 1

[2] Vaishnavi Patil, Matthew Evanusa, and Joseph JaJa. Protovae: Prototypical networks for unsupervised disentanglement, 2023. 3

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[4] Kristoffer K. Wickstrøm, Daniel J. Trosten, Sigurd Løkse, Ahcène Boubekki, Karl Øyvind Mikalsen, Michael C. Kampffmeyer, and Robert Jenssen. Relax: Representation learning explainability, 2022. 1, 3