Dissecting Representation Misalignment in Contrastive Learning via Influence Function

Huanyi Xie, Chenyang Ren, Khouloud Saadi, Shu Yang, Zhen Tan, Jingfeng Zhang, Lijie Hu, Di Wang
KAUST PRADA Lab MBZUAI Arizona State University University of Auckland

Motivation

Contrastive learning, commonly applied in large-scale multimodal models, often relies on data from diverse and often unreliable sources, which can include misaligned or mislabeled text-image pairs.

- This frequently leads to robustness issues and hallucinations, ultimately causing performance degradation.
- Data valuation is an efficient way to detect and trace these misalignments.

This paper solves a central challenge:

How can we efficiently quantify the exact influence of misaligned positive and negative samples in contrastive learning?

Why Classical Influence Functions Fail

Classical influence functions estimate data contributions using gradient information to avoid expensive retraining. However, they fail for contrastive learning:

- **Pointwise vs. Pairwise**: Classical IF was initially designed for M-estimators, which operate with pointwise loss. Contrastive learning encourages models to draw positive pairs closer while pushing negative pairs apart.
- **Negative Samples**: The original definition of the influence function does not consider the roles of positive and negative samples.
- **Dimensionality**: Computing gradients and Hessians in high-dimensional multimodal spaces is computationally demanding.

Core Idea: ECIF Formulation

We introduce the **Extended Influence Function for Contrastive Loss (ECIF)**. ECIF considers both positive and negative samples and provides a closed-form approximation of contrastive learning models, eliminating the need for retraining.
The standard self-supervised contrastive loss is defined as:

$$L_{Batch}(U, V; \theta) \triangleq \sum_{i=1}^{N} -log(e_i \cdot \sigma(S_{i,*})) - log(e_i \cdot \sigma(S^T_{*,i}))$$

where S is the similarity matrix, e_i is the standard basis vector, and σ is the softmax function.

Dual-Perspective Analysis

1. Influence as Positive Samples
By up-weighting the targeted positive pair by ϵ, we isolate its contribution. The positive influence function is deduced as:

$$positive\text{-}IF((x^T, x^I); \hat{\theta}) = -H_{\hat{\theta}}^{-1} \cdot \nabla_\theta Pos((x^T, x^I); \hat{\theta})$$

where $H_{\hat{\theta}}$ is the Hessian matrix.

2. Influence as Negative Samples
When data serves as a negative sample, its information is coupled with other data in the denominator. We add a matrix $\log \zeta \times E_n$ to the similarity matrix and perform a Taylor expansion at $\zeta = 1$ to separate the influence. The negative influence is:

$$negative\text{-}IF(\mathcal{D}^*, Seg; \hat{\theta}) = -H_{\hat{\theta}}^{-1} \cdot \nabla_\theta Neg(\mathcal{D}^*, Seg; \hat{\theta})$$

with the gradient corresponding to the impact of the negative terms.

Definition: ECIF
The Extended Influence Function for Contrastive Loss (ECIF) is defined jointly:

$$ECIF(\mathcal{D}^*, Seg; \hat{\theta}) = (positive\text{-}IF(\mathcal{D}^*, Seg; \hat{\theta}), negative\text{-}IF(\mathcal{D}^*, Seg; \hat{\theta}))$$

This dual-perspective view is crucial because each sample acts as a positive pair once but as a negative pair across the entire batch.

Applications of ECIF

Building upon ECIF, we develop algorithms for three primary tasks:

- **Valuable Data Identification**: Utilizing Task-related Influence Score (IS) to find data that significantly boosts model performance.
- **Misprediction Trace-Back**: Imposing constraints on parameter perturbations to accurately identify training samples responsible for specific test errors (Relative-IS).
- **Misalignment Detection**: Using validation text-image sets to identify pairs causing high negative impact on task metrics.

Experimental Setup

Datasets: FGVC-Aircraft, Food101, Flowers102, Cifar-100, etc.
Baselines: Full Retraining, IF-EKFAC, TRAK, and TracIN.
Metrics: Accuracy after removing samples and total runtime (s).

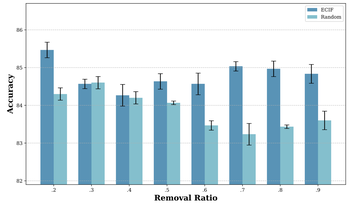
Main Quantitative Result

Data Influence Removal Evaluation

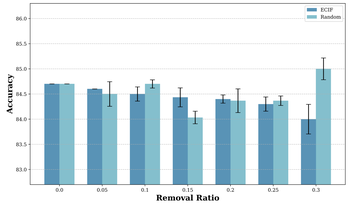
Dataset	Method	Accuracy (%)	RT (s)
Food101	Retrain	84.93 ± 0.17	875.4
	ECIF	84.87 ± 0.24	**436.8**
Cifar100	Retrain	73.50 ± 0.35	753.6
	ECIF	73.00 ± 0.20	**444.0**

✓ **High Fidelity Efficiency:** ECIF achieves comparable generalization performance to full retraining (accuracy difference is only 0.06% on Food101) while saving 50-80% computational time.

Visual Evidence



Harmful data removal on Food101



Valuable data removal on Food101