Detecting Manual Alterations in Biological Image Data Using Contrastive Learning and Pairwise Image Comparison

Georgii Nekhoroshkov
MIPT
Moscow, Russia
nekhoroshkov.gs@phystech.edu

Daniil Dorin
MIPT
Moscow, Russia
dorin.dd.contact@gmail.com

Andrii Hraboviy
MIPT
Moscow, Russia
grabovoy.av@phystech.edu

Abstract

In this paper, we address the problem of detecting manipulations in biological images. Ensuring the integrity of biological image data is essential for reliable scientific research. The study focuses on developing a model for pairwise image comparison using contrastive learning, demonstrating high pairwise comparison metrics to detect manual modifications or more subtle alterations. The proposed method outperforms state-of-the-art models, including CLIP and Barlow Twins, in the task of biological image comparison on fMRI scans and cell datasets. This work contributes to automated fraud detection and data validation in biological research.

Keywords: Machine Learning, Pairwise Image Comparison, Self-Supervised Learning, Fine-Tuning, Automated Fraud Detection, Detecting Data Alterations

1 Introduction

Our work aims to develop a machine learning solution for the problem of reusing existing biological and medical snapshots to demonstrate results in newly published biological articles. Fake images negatively impact on medicine by providing false or fabricated results and undermining the credibility of new scientific work in these fields. Existing state-of-the-art self-supervised learning approaches demonstrate remarkable results in pairwise image comparison tasks (`SimCLR` [1], `CLIP` [2], Barlow Twins [3]). However, their performance significantly worsens when applied to complex biological data. It requires developing model that is more sensitive to subtle changes in the image content while remaining resistant to various manual alterations, such as color jittering, flipping, rotation, noise application, and random affine transformations. At present, the problem of matching biological and medical images remains unsolved due to the complexity of distinguishing snapshots of similar objects, where differences can only be identified by experts in the field.

We propose a solution, based on `Barlow Twins` [3], trained and fine-tuned specifically for complex biological scans. The model belongs to the family of self-supervised learning (SSL) methods, which have been proven to be competitive with supervised representations ([1], [2], [3], [4], [5]). By leveraging a pretrained model, it does not require large snapshot datasets to achieve state-of-the-art accuracy on the `Haxby fMRI`, `CIL Epithelial Cell`, `CIL Lymphocyte Cell` datasets. This

Preprint. Under review.

solution can be widely used by biological articles proofreaders to verify the authenticity of provided images and detect borrowings from known datasets.

2 Problem

Given dataset $\mathcal{D}$, consisting of N biological snapshots:

$$\mathcal{D} = \{d_i \in \mathcal{S}, i \in [0, N)\}$$

where $\mathcal{S}$ is the image space.

For simplicity, we will refer to a pair of images with the same content before alterations as a *similar* pair; otherwise, it will be called *dissimilar*. Our goal is to build a model $\mathcal{M}$ using self-supervised contrastive learning (SSCL), which should be able to distinguish dissimilar pairs of images and identify similar ones.

Let x and y be two images ($x, y \in \mathcal{S}$). The model consists of two main parts:

$$\mathcal{M}(x, y) = h(f_\theta(x), f_\theta(y)))$$

where f_θ is an encoder with a trainable parameter set θ, and h is the linear classifier:

$$f_\theta(x) = v_x \in \mathbb{R}^d$$

$$h(v_x, v_y) = s \in \{0, 1\}$$

Value $s = 1$ corresponds to similar pairs, $s = 0$ represents dissimilar pairs.

The model's accuracy will be evaluated by counting the number of correctly classified similar pairs and incorrectly classified ones, producing a single accuracy value to compare with other state-of-the-art models.

3 Computational Experiment

In this section, we conduct an experiment to train the projector of the `Barlow Twins` model on a blood cells dataset. Our goal is to determine if the model, with its state-of-the-art pretrained encoder, can be successfully adapted to our problem.

The model's encoder is the pretrained `ResNet50` from the original article [3]. The dataset consists of 320 images of lymphocyte cells, split into a training set of 256 images and a validation set of 64 scans. The projector is composed of three groups of layers; each group contains a linear layer, batch normalization, and a ReLU activation function. The input dimensions for the groups are 2048, 1024, and 1024, respectively.

The loss function is identical to that described in the original article. For optimization, we use the `Adam` optimizer with the following learning rate schedule:

$$q_k = \frac{1}{2} \cdot (1 + \cos(\pi \cdot \frac{k}{K}))$$

$$\gamma_k = \gamma_{start} \cdot q_k + \gamma_{end} \cdot (1 - q_k)$$

where k is the epoch number, K is the total number of epochs, $\gamma_{start} = 5 \cdot 10^{-3}$, $\gamma_{end} = 5 \cdot 10^{-4}$.

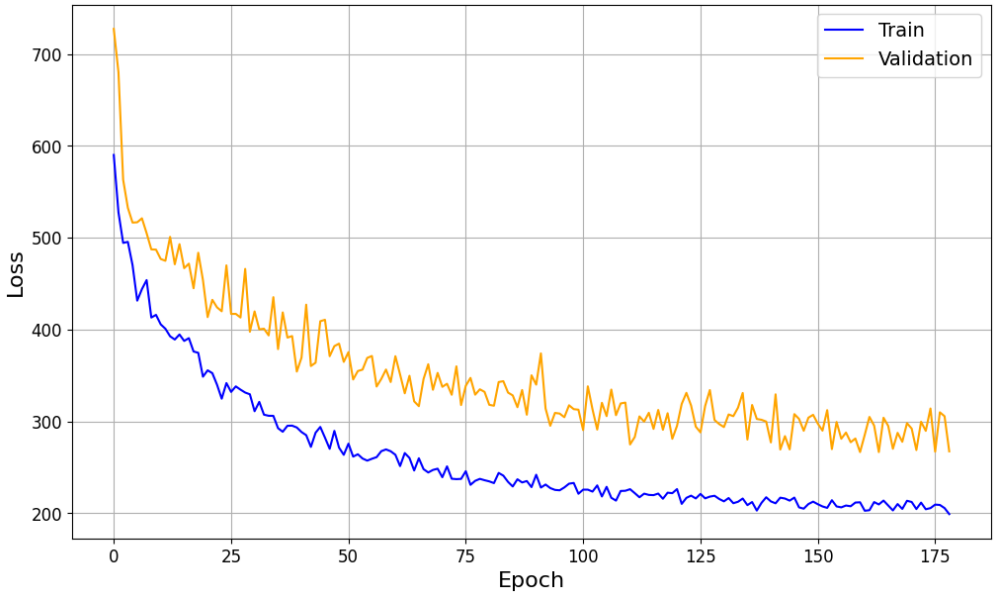

Figure 1: Graph of the loss function for training and validation samples. The experiment was conducted over 180 epochs. The validation loss decreases alongside the training loss, indicating that the model can be trained further. Training takes approximately 25 minutes on a Google Colab GPU.

4 Method

The challenge of detecting reused biological and medical images lies in the difficulty of distinguishing between visually similar images while maintaining invariance to various transformations. Our dataset consists of biological snapshots obtained from publicly available sources and is stored as an array $\mathcal{D}$ with elements $d_i \in [0, 256)^{l \times l \times 3}$, where l denotes the length of the image sides. We define a *solution* as any method intended to address the stated problem, and we refer to the *model* as the machine learning solution proposed in our work.

The structure of our model is inspired by the `Barlow Twins` [3]. It consists of three deep networks: an embedding function f, a projector function p, and a similarity function s that returns a single number in the range $[0, 1]$. The embedding network is implemented using a fine-tuned, pretrained `ResNet50`.

To train the projector, we begin with a batch of images X. Each image x_i is augmented in two different ways to produce two modified versions, y_i^A and y_i^B. These two batches of augmented images, Y^A and Y^B, are then processed by the embedding function f to yield embeddings E^A and E^B. The embeddings are subsequently passed through the projector function p, resulting in the projected embeddings Z^A and Z^B. These projected embeddings are used to compute the loss function $\mathcal{L}$ as proposed in `Barlow Twins` [3]:

$$\mathcal{L}(Z^A, Z^B) = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2,$$

where λ is a positive constant, and $\mathcal{C}_{ij}$ is the cross-correlation matrix between the outputs of the two networks along the batch dimension, defined as:

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}}.$$

For simplicity, we refer to a pair of images with the same content before augmentation as a *similar* pair, and a pair with different content as a *dissimilar* pair. The objective of the loss function is to minimize the distance between embeddings of similar pairs while maximizing the distance for dissimilar pairs.

For accuracy evaluation, we train the similarity network function s while keeping the model's weights frozen. Given a batch of images X, we generate two batches, Y^A and Y^B, where Y^A comprises images from X with random modifications, and Y^B is a shuffled version of X with different modifications applied. Consequently, for each image x_i, there is a corresponding modified image y_i^A and a randomly transformed image y_i^B, with a probability of 0.3 that they form a *similar* pair. These batches are processed sequentially through the embedding function f and the projector p, and the resulting embeddings Z^A and Z^B are then fed into the similarity function s. The function s produces a vector P with values in the range $[0, 1]$, where each element P_i represents the estimated likelihood that the pair (y_i^A, y_i^B) is similar. The similarity network is trained using Binary Cross-Entropy (BCE) loss.

The model is trained on 70% of the dataset $\mathcal{D}$, validated on an additional 10%, and tested on the remaining 20%.

Topic #1. TODO

Topic #2. TODO

5 Preliminaries

5.1 General notation

In this section, we introduce the general notation used in the rest of the paper and the basic assumptions.

5.2 Assumptions

TODO

6 Experiments

To verify the theoretical estimates obtained, we conducted a detailed empirical study...

7 Discussion

TODO

8 Conclusion

TODO

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 (PMLR)*, 2020.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[3] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

[4] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese network features for image matching. In *23rd International Conference on Pattern Recognition (ICPR)*, 2016.

[5] Grill Jean-Bastien, Strub Florian, Altché Florent, Tallec Corentin, H. Richemond Pierre, Buchatskaya Elena, Doersch Carl, Avila Pires Bernardo, Daniel Guo Zhaohan, Gheshlaghi Azar Mohammad, Piot Bilal, Kavukcuoglu Koray, Munos Rémi, and Valko Michal. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

A Appendix / supplemental material

A.1 Additional experiments

TODO