SuperStableRep: Data-Efficient Curated Multi-Sample Contrastive Learning for ViTs

Alex Wilentz
SEAS, Applied Mathematics
Harvard University
Cambridge, MA
awilentz@college.harvard.edu

Andrew Lu
SEAS, Computer Science
Harvard University
Cambridge, MA
alu@college.harvard.edu

Carl Ho
SEAS, Computer Science
Harvard University
Cambridge, MA
carlho@college.harvard.edu

***Abstract*—While StableRep constitutes a significant increase in performance over previous visual representation learning, it does not implement negative contrasting learning. In this paper, we present SuperStableRep: an implementation of StableRep that both implements negative contrastive loss with the use of Stable Diffusion to generate images with prompts consisting of one-word deviations. The combination of these two allows the model to better differentiate between close, but distinct categories. Ultimately, results in this work were hampered by limitations on computation and storage, immediate future work should involve training SuperStableRep on an appropriate amount of images, as well as comparing its performance on tasks in which there is large variance in decisions based on small variations in the input data.**

I. Introduction

A model improves when its inputs are higher in quality. While the search for gold-standard data continues in every domain, significant interesting work has developed in the area of synthetic image data for image-text representations. Foundation models consume vast amounts of and heavily rely on internet scraping to make associations and predictions on content. In particular, the quality, quantity, and breadth of data inputs is a significant determinant to the performance of these models. However, real-world data is substantially limited on all these fronts – and it poses a significant challenge for the collection, cleaning, and processing of large amounts of disparate data. We seek to investigate alternative approaches, extending upon recent promising research into the area of synthetic image generation for image-text pairs.

Most notably, recent work investigating synthetic image generation for contrastive learning has proved the efficiency and efficacy of alternative approaches to data inputs to foundation models including vision transformers. Preliminary work by Tian et al. (2023) [1] on the model StableRep, yielded promising results: relative to CLIP's 50M image-text pair dataset, StableRep yielded comparable, even better, performance with 30M fewer images.

We present SuperStableRep, an extension upon the StableRep model to capture the utility of negative constrative learning through novel data generation techniques. Using single-word deviations, we effectively teach a vision transformer model concepts, resulting in faster convergence toward conceptual object-based learning, achieving better results compared to other synthetic-image based models.

[@AW add more on results when ready]

II. Problem Statement

Synthetic image-based foundation vision model training is still in nascent stages. While proposed methods have already demonstrated significant advancements in this area, there is still very substantial room for improvement. Even with rapidly advancing computational resources, the costs of training a large-scale image modality model is very substantial, and we seek to develop methods that significantly reduce costs and development time through better training processes. We draw upon existing research including the StableRep model, to propose a image-text representation model that surpasses state-of-the-art research. Models that span across modalities, such as vision and text, are especially important and emerging fields of research. Our goal is to demonstrate the applicability of intuitive heuristics of learning to vision transformers. As such, we seek to carefully craft text generation prompts that allow models to learn more incrementally and progress through conceptual learning. Because we have neither sufficient time nor resources thus far to pursue a full-scale training of ViTs, we selectively choose objects for training and test accuracy on an image benchmark to compare between SuperStableRep and its predecessor, StableRep.

III. Proposed Approach and Intellectual Points

We propose a novel data generation method and model training process for improving upon existing synthetically-trained ViTs. Synthetically-trained ViTs are valuable in that they can generate data for novel or underrepresented tasks. The key differentiator in the data generation method is the careful curation of data particularly relevant to concept-based learning. Instead of allowing for a wide range of prompts, as used in CLIP and CC12M (used by StableRep), we control for the quality of the synthetically generated data by regulating the quality of the prompts provided. We create prompt templates, which often include some relational characteristics, and generate images across all object-concepts. In doing so, our goal is to enable the model to easily learn concepts and the distinctions between them, without needing to generalize

(a) Golden *retriever* in a field of sunflowers

(b) Golden *labrador* in a field of sunflowers

Fig. 1: Single-word deviations, Stable Diffusion-v1.5

excessively. On the other hand, the image input dataset still retains a significant degree of variety, as the model will view the same object in multiple concepts, which assists in generalization.

The second implementational crux is the development of a multi-positive, multi-negative contrastive learning process. In doing so, we seek to utilize recent advances in contrastive learning where other models only utilized partial data training procedures through multi-positive learning, we seek to innovate and improve upon this by incorporating positive and negative contrastive learning for even faster convergence. With both of these strategies, we seek to develop robust, creative methods for improving accuracy and reducing training costs.

IV. Methods and Implementation

Our implementation of Super Stable Rep occurred in three major parts: image generation, network training, and model testing and comparison.

A. Image Generation

We construct a dataset of around 90,000 images based on Stable Diffusion v1.5 using template prompts that we manually create. For each template prompt, we iterate across all classes within our given object set, and create a set of 16-24 quality synthetic images from SD-1.5. Due to the limited generation capacity, we focus on selecting objects from a narrower domain, choosing to use ImageNet's 400 synsets as the objects to use for prompt generation. Within each synset, there is also several objects which are treated as synonyms. We then parallelize the generation process.

B. Network Training and Development

Network training was based off of the StableRep repository, which is publicly available. The code was adapted to load the previously generated images and their captions, grouping together images generated from the same prompt as the positive examples.

Note that even though during image generation, there are images generated from prompts with single-word deviations which may help to create more similar negative examples, for some set of images of a prompt, all other images generated from other prompts were taken to be negative examples.

Two models were trained over 15 epochs, which we deemed sufficient given the scale of our training set: one with only positive contrastive learning as implemented previously, and our implementation of positive and negative contrastive learning. Normalized temperature-scaled cross entropy loss (NT-Xent), a positive and negative contrastive learning method, was used as implemented as used in SimCLR [2], a previous benchmark for learning visual representations, with the following form:

$$\frac{\mathbf{u}^T\mathbf{v}^+}{\tau} - \log\left(\sum_{\mathbf{v}\in\{\mathbf{v}^+,\mathbf{v}^-\}} \exp\left\{\frac{\mathbf{u}^T\mathbf{v}}{\tau}\right\}\right)$$

Where all vectors $\mathbf{v}, \mathbf{u}^+, \mathbf{v}^-$ are ℓ_2 normalized.

C. Testing

Once the two models are trained, they were tested against the ImageNet dataset by training a linear classifier against the inferred representations generated from each of the models. Due to limitations in compute and storage, we used a subset of the ImageNet dataset of 50,000 images from which to train and evaluate the models.

V. Results

After training our models (the first with only multi-positive contrastive learning and the second with both multi-positive and multi-negative contrastive learning), we added a linear classification layer which we trained on evaluated on the ImageNet 2012 dataset. We trained and evaluated the linear classifier with a 50,000-image subset from the ImageNet dataset, randomly splitting the dataset into a equally-sized training and test sets of 25,000 images each.

After 10 epochs (which we determined sufficient to learn on our smaller dataset), we compared the classification performance between the models. Both models performed poorly, though this is not surprising and for several reasons quite predictable, as the base StableRep model had a top-1 accuracy of 0.132% and a top-5 accuracy of 1.466%, while our model with multi-positive and multi-negative contrastive learning had a top-1 accuracy of 0.288% and a top-5 accuracy of 1.638% over the 1000 classes. Note that a random classified would have an expected accuracy of 0.1%. The models' poor performances are attributable to several factors. The first of those is a lack of training data due to the computational cost of generating images and training on a large amount of images. ViTs usually need several million examples to

train, and the StableRep models trained in the original paper were trained on datasets ranging from 2.7 million images to 11.6 million images in size, which is not feasible given our accessible compute. This at worst constituted a difference of more than 100 times the number of images in our dataset, which was less than 90000, a number of examples with which it is difficult to train a ViT model. The second factor is that our generation of the synthetic data was based upon ImageNet's 400 synsets, each of which represents a class. However, the ImageNet dataset has 1,000 classes, and thus many classes were not represented in our training data. Though this allowed our models to learn visual representations for the included classes, this effectively excluded our model from learning representations for the remaining classes. Those that were in the synsets, however, had just above 200 examples per class, which is quite few to learn robust visual representations. Our relatively small training set for the linear classifier likewise probably contributed to error in classification. In general, these limitations are caused by a lack of compute and time. Notably, even with the general poor performance, while the base StableRep model only performed just marginally better than random under our difficult conditions, the SuperStableRep model indeed showed slightly better performance, as its accuracy on ImageNet was significantly higher than both a random classifier and StableRep, despite its poor accuracy.

VI. Related Work

The central related work to this project is StableRep [1], of which our project was based. Our SuperStableRep model is a StableRep model with a modified loss function for both multi-positive and multi-negative contrastive loss. SimCLR [2] was a previous benchmark for learning visual represention in StableRep, and a paper we considered in implementing our loss function.

VII. Conclusion

While the results of this study were limited by the computational resources available, they do suggest a promising improvement in training rate that Super Stable Rep has over the original implementation of Stable Rep. Assuming that the negative contrastive model trains at a rate proportional to the original model which our resulting accuracies might suggest, introducing negative contrastive learning would be able to significantly reduce the number of images needed to train the model.

While results from this current project were limited, future work on this project would entail generating an appropriate number of images (on the same scale as in the original StableRep paper, or slightly less to test the limits of the SuperStableRep model, around 10 to 20 million images) with a more representative dataset. This would ultimately validate the preliminary results of the performance gains of negative contrastive learning for visual representation learners.

Other potential avenues include tailoring the negative contrastive model to visual tasks where subtle changes can have large impacts: some examples include medical imaging tasks and vision systems for self-driving vehicles.

VIII. Group Contribution Statement

Our project consisted of three major parts: image generation, network training and development, and model testing and comparison. Andrew Lu led the team on image generation, both in developing the process to generate positive and negative examples and the generation of the images with Stable Diffusion. Alex Wilentz led the implementation of the multi-negative contrastive learning and deployment of the SuperStableRep model, with consultation and discussion from Andrew and Carl Ho. Carl led the the testing of the models and analysis of their performances, with help from Alex. Regardless of the lead, all stages of the process were very collaborative, and discussion was involved at all phases.

IX. Supplemental

For details and implementation, our code repository can be found at https://github.com/andrewmlu/super-stable-rep

X. References

References

[1] Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. (2023). StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. arXiv preprint arXiv:2306.00984.

[2] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. arXiv preprint arXiv:2002.05709.

[3] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575, 2014.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.