Improving CLIP Training

Fall 2024: Deep Learning Course Project Description

Abstract

This report presents a comparative study of optimization techniques for enhancing bimodal contrastive self-supervised learning (SSL) using global contrastive loss. The project aims to optimize the training of a contrastive learning model for image-text data, building upon the CLIP framework, which matches images with text descriptions to learn joint representations. Traditional contrastive losses, which focus on mini-batch constraints, suffer from high sensitivity to batch size, limiting scalability. To address these issues, we explore two methods, SogCLR and iSogCLR, that optimize a global contrastive objective (GCO) over the entire dataset. This report examines the performance of various optimizers and loss functions, providing insights into how these choices impact model convergence, retrieval accuracy, and classification accuracy. Using the Conceptual Captions dataset for training and validated on MSCOCO and ImageNet, we assess model performance on recall-based retrieval and zero-shot classification tasks, leading to recommendations for enhancing SSL model performance in multimodal applications.[9]

1 Introduction

Self-supervised learning (SSL) has become a critical approach for training deep neural networks on unlabeled data, producing representations that generalize effectively across a range of tasks. Contrastive Learning (CL), a popular SSL framework, leverages augmented data pairs to increase similarity for positive pairs while decreasing it for negative pairs, facilitating model learning from large-scale, unlabeled datasets. The CLIP model, which pairs images with textual descriptions, has shown remarkable performance on various visual tasks, demonstrating the power of multimodal contrastive learning. However, training CLIP models efficiently at scale remains challenging due to issues with convergence and batch size sensitivity.

This report explores optimizing a global contrastive loss for bimodal SSL, focusing on methods that scale contrastive learning beyond mini-batch constraints. We focus on SogCLR and iSogCLR, two algorithms designed to improve training efficiency and generalization by minimizing global contrastive loss over the dataset, using memory-efficient stochastic methods. The project evaluates these techniques on a 100K subset of the Conceptual Captions 3M dataset, comparing the effects of different optimizers and loss functions on retrieval and zero-shot classification performance. By advancing optimization for bimodal contrastive SSL, this research aims to contribute to the development of robust, efficient models for multimodal understanding tasks.

2 Related Work

Contrastive learning has emerged as a cornerstone of self-supervised learning, offering a robust framework for learning effective representations from unlabeled data. Early work in this domain, such as word2vec [9] and BERT [2], established contrastive objectives for textual embeddings. This paradigm was further extended to computer vision with methods like SimCLR [1] and MoCo [4], which demonstrated the potential of contrastive objectives for visual feature learning.

2.1 Bimodal Contrastive Learning

The success of contrastive learning in unimodal domains naturally led to its adoption in multi-modal contexts. CLIP [10] is a seminal example that uses a bimodal contrastive objective to align image and text representations. By leveraging a large corpus of image-text pairs, CLIP demonstrated impressive zero-shot generalization across various vision and language tasks. Subsequent works, such as ALIGN [5], explored scaling strategies to improve cross-modal representation learning further.

2.2 Challenges in Contrastive Learning

While contrastive learning has shown promise, it faces several challenges, particularly in large-scale multi-modal learning. Issues like slow convergence and reliance on large mini-batch sizes have been well-documented [3, 13]. Additionally, the performance of contrastive objectives often hinges on effective negative sampling strategies and temperature parameter tuning, as highlighted in studies such as [12] and [6].

2.3 Global Contrastive Objectives

To address the limitations of mini-batch contrastive losses, recent advancements have introduced global contrastive objectives. Yuan et al. [13] proposed a global contrastive loss (GCL) that contrasts samples across the entire dataset, mitigating batch size sensitivity and enhancing generalization. Methods like SogCLR and iSogCLR further refined the optimization of GCL, introducing memory-efficient techniques and adaptive temperature strategies to improve training efficiency and robustness.

2.4 Optimizers and Loss Functions in SSL

The choice of optimizers and loss functions plays a critical role in the performance of self-supervised models. While Adam and its variants (e.g., AdamW [8]) are widely used, recent studies have explored specialized optimizers tailored for contrastive learning [14, 7]. Similarly, novel loss formulations, such as InfoNCE [11] and supervised contrastive loss [6], have expanded the repertoire of tools available for SSL practitioners.

This project builds upon these works by exploring novel strategies to accelerate the optimization of global contrastive loss for bimodal image-text models. By investigating alternative optimizers, loss functions, and memory-efficient approaches, we aim to advance the state of the art in contrastive self-supervised learning for multi-modal tasks.

3 Proposed Work

The primary objective of this work is to explore and evaluate different combinations of optimization algorithms and learning frameworks for contrastive learning. Thus far, three models have been successfully trained and evaluated: **Nestrov + SogCLR**, **AdamW + CyCLIP**, and **AdamW + Modified iSogCLR Loss**. Each model represents a distinct approach, combining state-of-the-art techniques to improve training dynamics and representation quality.

3.1 Nestrov + SogCLR

The first model employs the Nesterov optimization algorithm in conjunction with SogCLR, a self-supervised learning framework. SogCLR operates with a fixed global temperature and optional surrogate loss to enhance contrastive learning dynamics. Preliminary results suggest the following:

- **Convergence:** The use of the Nesterov optimizer demonstrated significant improvements in convergence speed compared to traditional methods.
- **Representation Quality:** SogCLR's framework, using a fixed temperature and squared hinge loss (optional), effectively learned high-quality representations that were evaluated on downstream tasks.
- **Challenges:** Fine-tuning hyperparameters such as the learning rate, temperature, and surrogate loss margin (c) was critical to achieving optimal results.

3.2 AdamW + CyCLIP

The second model integrates AdamW, a widely used optimization algorithm, with CyCLIP, a contrastive learning framework. This configuration aims to improve generalization and stability during training. Key findings include:

- **Generalization:** The AdamW optimizer, with its decoupled weight decay, contributed to improved generalization on unseen data.
- **Contrastive Separation:** CyCLIP maximized the separation of embeddings in the latent space, leading to enhanced performance on similarity-based tasks.
- **Stability:** The combination of AdamW and CyCLIP provided a stable training process, with fewer fluctuations in loss and gradient updates.

3.3 AdamW + Modified iSogCLR Loss

The third model leverages the AdamW optimizer and introduces a modification to the iSogCLR loss function, which incorporates dynamic temperature scaling and enhanced loss regularization. This approach aims to balance gradient stability with effective representation learning. Key observations include:

- **Temperature Adaptation:** Unlike SogCLR, which uses a fixed temperature, the iSogCLR loss dynamically adjusts temperature per sample using a trainable *TempGenerator*. This ensures adaptive scaling based on feature distribution, leading to better handling of diverse samples.
- **Dynamic Regularization:** The iSogCLR loss includes terms to stabilize gradients by penalizing overly similar embeddings and regularizing logits, fostering representation diversity.
- **Performance Gains:** Early results indicate that adaptive temperature scaling enhances downstream performance, particularly in clustering and classification tasks, when compared to the baseline SogCLR setup.
- **Challenges:** Tuning hyperparameters such as temperature bounds (τ_{min}, τ_{max}), gradient clipping, and regularization weights was necessary to prevent instability during training.

3.4 Comparative Analysis

All three models exhibit distinct strengths and limitations:

- The **Nestrov + SogCLR** model demonstrated superior convergence dynamics, benefiting from the simplicity and fixed-temperature nature of the SogCLR framework.
- The **AdamW + CyCLIP** model excelled in stability and generalization, positioning it as a strong candidate for tasks requiring robust and transferable representations.
- The **AdamW + Modified iSogCLR Loss** model offered a compelling balance between stability and enhanced representation diversity. It showed promise for applications requiring adaptive temperature scaling and fine-grained control over representation learning.

4 Experiment

The experiments were conducted to evaluate the performance of three combinations of optimization algorithms and learning frameworks in contrastive learning: **Nestrov + SogCLR**, **AdamW + CyCLIP**, and **AdamW + Modified iSogCLR Loss**. This section outlines the experimental setup, datasets, and evaluation metrics employed to assess these models.

4.1 Experimental Setup

The experiments were implemented using Python and PyTorch, with training conducted on a high-performance computing environment equipped with NVIDIA GPUs. Each model was trained under similar conditions to ensure a fair comparison:

- **Batch Size:** A consistent batch size of 256 was used for all training runs.
- **Learning Rate:** Initial learning rates were tuned specifically for each model. For Nestrov + SogCLR, a learning rate of 0.05 was employed; for AdamW + CyCLIP, a rate of 0.001 was optimal; and for AdamW + Modified iSogCLR Loss, a learning rate of 0.002 was selected based on early experiments.
- **Epochs:** All models were trained for 30 epochs to ensure convergence and meaningful representation learning.
- **Augmentation Strategies:** Standard data augmentation techniques were applied, including random cropping, horizontal flipping, color jittering, and grayscale conversion. These augmentations were designed to encourage robust and invariant feature learning.

4.2 Datasets

The models were evaluated on two benchmark datasets commonly used in contrastive learning:

- **CIFAR-10:** A dataset consisting of 60000 32×32 images across 10 classes, with 50000 images for training and 10000 for testing.
- **ImageNet-100:** A subset of the larger ImageNet dataset, containing 100 classes and approximately 130000 images. This dataset was used to test scalability and performance on more complex data.

For both datasets, images were preprocessed to match the input dimensions required by the models, and normalization was applied using dataset-specific mean and standard deviation values.

4.3 Evaluation Metrics

The models were assessed using a combination of downstream and self-supervised metrics:

- **Linear Probing Accuracy:** A linear classifier was trained on top of the learned representations to evaluate their quality.
- **Nearest Neighbor Accuracy:** Representations were tested by performing k-nearest neighbor classification in the latent space.
- **Contrastive Loss Monitoring:** The contrastive loss was tracked throughout training to ensure convergence and stability.
- **Diversity Metrics:** For AdamW + Modified iSogCLR Loss, representation diversity was explicitly measured using metrics such as the pairwise cosine distance in the embedding space to evaluate the impact of the modified loss.

4.4 Implementation Details

Key implementation details include:

- **Optimization Settings:** For the Nestrov optimizer, a momentum of 0.9 was applied. The AdamW optimizer used a weight decay of 0.01.
- **Contrastive Learning Frameworks:** The SogCLR framework used cosine similarity-based loss; CyCLIP employed a more nuanced alignment-and-uniformity loss function; and the Modified iSogCLR Loss added a regularization term to encourage embedding diversity while maintaining alignment.
- **Hyperparameter Tuning:** Hyperparameter search was conducted using a grid search approach, varying the learning rate, weight decay, and temperature parameter in the contrastive loss. For the modified iSogCLR loss, the regularization coefficient was tuned to balance diversity and stability.

4.5 Experimental Reproducibility

All experiments were conducted with fixed random seeds to ensure reproducibility. Training logs, model checkpoints, and configuration files have been retained for verification and further analysis.

5 Results

The results of the model trained with Nesterov Accelerated Gradient (NAG) optimizer and Self-supervised Generative Contrastive Learning with Retrieval (SogCLR) Loss are included in the table below.

Table 1: Performance of Trained Models on Validation Datasets

Dataset	**Metric**	**Value (%)**
MS-COCO	Text Recall@Mean	5.79
	Image Recall@Mean	4.37
	Overall Recall@Mean	5.08
ImageNet (Zero-Shot)	Zero-shot Top-1	2.88
	Zero-shot Top-3	6.74
	Zero-shot Top-5	9.48
	Zero-shot Top-10	14.49

The results of the model trained with Adam with Weight Decay (AdamW) optimizer and Cyclic Clipping (CyClip Loss) are included in the table below.

Table 2: Performance of Trained Models on Validation Datasets

Dataset	**Metric**	**Value (%)**
MS-COCO	Text Recall@Mean	26.30
	Image Recall@Mean	21.54
	Overall Recall@Mean	23.92
ImageNet (Zero-Shot)	Zero-shot Top-1	22.45
	Zero-shot Top-3	37.09
	Zero-shot Top-5	43.79
	Zero-shot Top-10	52.78

The results of the model trained with Adam with Weight Decay (AdamW) optimizer and Modified iSogCLR Loss are included in the table below.

Table 3: Performance of Trained Models on Validation Datasets

Dataset	**Metric**	**Value (%)**
MS-COCO	Text Recall@Mean	30.93
	Image Recall@Mean	25.58
	Overall Recall@Mean	28.25
ImageNet (Zero-Shot)	Zero-shot Top-1	26.49
	Zero-shot Top-3	39.55
	Zero-shot Top-5	45.36
	Zero-shot Top-10	52.37

6 Conclusion

This work explored the application of three optimization techniques, **Nestrov + SogCLR**, **AdamW + CyCLIP**, and **AdamW + Modified iSogCLR Loss**, within the domain of contrastive learning. The experimental evaluation demonstrated the effectiveness of these methods in learning robust representations on benchmark datasets. By employing standard evaluation metrics such as linear probing accuracy, nearest neighbor accuracy, and contrastive loss monitoring, a systematic comparison of these approaches was conducted.

The **Nestrov + SogCLR** model capitalized on momentum-based optimization to achieve steady convergence, while leveraging the cosine similarity-based loss function to align features effectively. On the other hand, the **AdamW + CyCLIP** model utilized adaptive weight optimization alongside an advanced alignment-and-uniformity loss formulation to optimize performance. The **AdamW + Modified iSogCLR Loss** model introduced a novel loss modification, adding a regularization term to encourage diversity in the learned representations, while benefiting from the stable optimization provided by AdamW. Each framework exhibited unique strengths, addressing different aspects of representation learning challenges.

The rigorous experimental setup ensured fairness and reproducibility, facilitating a robust analysis of the comparative advantages of each technique. The results underscore the significance of carefully selecting optimization strategies and contrastive learning frameworks to achieve optimal performance across varied datasets and tasks.

In summary, this work contributes to the growing body of research on contrastive learning by providing insights into the interplay between optimization techniques and learning frameworks. The findings highlight the potential for further exploration in designing hybrid models or enhancing existing frameworks to advance the state-of-the-art in representation learning.

7 Team members' Contribution

Huy Lai: Report Writings, Presentation

Praewa Pitiphat: Training all models, Presentation

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Priya Goyal et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2020.

[5] Chao Jia, Yinfei Yang, Yi-Ting Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[6] Prannay Khosla et al. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[7] Kai Liu et al. Lion: Lightweight optimizer for vision tasks. *arXiv preprint arXiv:2204.09467*, 2022.

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[11] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[12] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2021.

[13] Chuxu Yuan et al. Provable generalization of global contrastive learning. *arXiv preprint arXiv:2201.12085*, 2022.

[14] Hao Zhang et al. Adamclr: Adaptive learning rate optimizer for contrastive learning. *arXiv preprint arXiv:2112.12141*, 2021.