TEXAS A&M UNIVERSITY®

Deep Learning (CSCE 636)

Project Report

Improving CLIP Training

22 November 2024

Adithi Srinath - 535005885

Venkata Sumanth Reddy Kota - 236002825

Abstract

This project focuses on experimenting on multiple loss functions and optimizers and also enhancing the existing Contrastive Language-Image Pre-training (CLIP) framework by optimizing its global contrastive loss for bimodal self-supervised learning (SSL). CLIP has demonstrated exceptional performance across diverse vision-language tasks by leveraging paired image-text datasets. However, training CLIP with large datasets comes with challenges such as slow convergence and significant resource requirements.

To overcome these issues, we propose two novel loss functions - **Modified_CLIP and Debiased Contrastive Loss** - to improve alignment between image-text representations while mitigating overfitting. This project also explores combinations of advanced optimizers, like AdamW, Nadam and Novograd, to enhance gradient stability. Experiments performed on subsets of Conceptual Captions and evaluated against MSCOCO and ImageNet benchmarks demonstrate the accuracy of our methods in retrieval (Image-to-Text and Text-to-Image Recall) and zero-shot classification tasks. Our contributions emphasize the critical role of global contrastive loss in achieving efficient SSL for multimodal models, providing a foundation for future enhancements in this field.

Introduction

As an effective model for learning from huge volumes of unlabeled data, self-supervised learning (SSL) has attracted a lot of attention. SSL has transformed domains such as natural language processing (NLP) and computer vision by employing models to extract valuable features without the need for labeled data. Contrastive learning (CL) is one of the best methods among the several SSL frameworks, greatly improving model performance in a variety of domains. Increasing the similarity between "positive" pairs - such as enhanced views of the same image - and decreasing the similarity between "negative" pairs - such as augmented views from different images are the objectives of CL. This approach makes CL especially useful for large-scale unsupervised learning since it produces strong representations that can generalize across a variety of tasks.

Contrastive learning has proven especially effective in multimodal settings, aligning different data modalities such as images and text. Strong cross-modal understanding is made possible in this situation by seeing text and visual representations as distinct viewpoints of the same underlying idea. One prominent example is CLIP (Contrastive Language-Image Pretraining), which achieves remarkable performance on tasks like text-based image retrieval and zero-shot image classification without requiring task-specific fine-tuning by learning a unified embedding space from extensive image-text datasets. But there are drawbacks to CLIP and other contrastive learning techniques that use mini-batch-based local contrastive loss, including sensitivity to batch size and optimization inefficiencies, particularly when dealing with big datasets. This approach's reliance on mini-batches for negative sampling hampers scalability, slowing convergence and limiting the quality of learned representations, thus reducing the broader applicability of CL-based models.

An alternate strategy called global contrastive loss uses the complete dataset for negative sampling as opposed to mini-batches in an effort to overcome these issues. This method enhances representation quality by lowering optimization errors caused by small batch sizes. Developments such as SogCLR and iSogCLR, which employ dynamic temperature scaling and moving averages to boost training efficiency, have further demonstrated the feasibility of global contrastive objectives.

Building on these developments, our project suggests new optimizers and loss functions specifically designed for CLIP training. We investigate how customized optimization strategies can speed up convergence and enhance downstream task performance by combining global contrastive loss with methods like Modified_CLIP and Debiased Contrastive Loss. We implemented techniques like dynamic temperature scaling and positive pair weighting to overcome issues like slow convergence and ineffective use of large-scale datasets. These improvements stabilize gradients during training and improve the alignment of image-text representations. The model's ability to generalize is greatly enhanced by Debiased Contrastive Loss, which specifically lessens the impact of easy negatives while emphasizing hard negatives. Better management of the inherent complexity in multimodal SSL is ensured by these customized changes.

Additionally, the research project looks into how well lightweight optimizers like AdamW, Nadam, and Novograd perform in comparison. AdamW's decoupled weight decay makes it extremely effective at avoiding overfitting, especially on a variety of datasets such as ImageNet and MSCOCO. This investigation guarantees that the suggested techniques can be adjusted to settings with computational limitations while preserving an equilibrium between efficacy and efficiency. Our tests' outcomes show notable gains in important performance indicators, such as Top-1 Accuracy for classification tasks, Image-to-Text Recall, and Text-to-Image Recall. These results highlight the potential of global contrastive loss in promoting multimodal SSL and pave the way for further research that makes use of these techniques in domain-specific or resource-constrained applications.

Related Work

With the development of SSL frameworks like SimCLR and MoCo, which opened up opportunities for models like CLIP, the study of contrastive learning has advanced quickly. In order to increase similarity within positive pairs, these frameworks contrast enriched perspectives of data in order to focus on representation learning. CLIP built upon this basis by adding a multimodal dimension and achieving state-of-the-art performance in zero-shot tasks through training on extensive paired image-text datasets.

Recent advancements in optimizing CLIP training focus on improving efficiency and scalability. By employing memory-efficient stochastic algorithms for dataset-wide representations, techniques such as SogCLR and iSogCLR improve robustness and optimize global contrastive loss, thereby overcoming the drawbacks of mini-batch sampling. By using techniques like flexible learning schedules and effective gradient reduction, FastCLIP improves CLIP in distributed environments while avoiding memory constraints and compute inefficiencies. This bridges the gap between large-scale and resource-constrained contexts by enabling the training of large-scale CLIP models with fewer resources. Furthermore, MaskCLIP captures both global and localized visual characteristics by combining masked self-distillation and contrastive learning. This method works well for applications that need subtle feature extraction since it enhances image-text alignment and allows for improved generalization for lower vision tasks.

By overcoming significant constraints in CLIP model training and broadening the application of multimodal SSL, our study places itself within this context. This research improves the effectiveness and resilience of global contrastive loss training by introducing new loss functions like Modified_CLIP and Debiased Contrastive Loss and investigating efficient optimizers like AdamW. Our method guarantees that these improvements are scalable and relevant to a range of contexts by utilizing insights from recent developments in memory-efficient and distributed training techniques.

Specifically, our techniques highlight how crucial it is to strike a balance between performance improvements and computational economy. Dynamic temperature scaling and customized positive pair weighting, for instance, directly address convergence problems without sacrificing model accuracy. In addition to raising the bar for CLIP training, these achievements pave the way for the use of multimodal SSL models in fields with severe resource limitations, including as autonomous systems, natural language processing, and medical imaging. Our approach lays the groundwork for future investigations into adaptive and domain-specific improvements by integrating with frameworks like FastCLIP and MaskCLIP.

Proposed Methodology

Our proposed methodology builds upon state-of-the-art techniques in contrastive learning to address limitations in training CLIP models. The methodology revolves around introducing novel loss functions and optimizing existing frameworks to enhance the alignment of image-text representations while minimizing training inefficiencies.

Modified_CLIP Loss

The Modified_CLIP loss refines the original CLIP loss by incorporating two points: dynamic temperature scaling and a weighting factor for positive pairs. These adjustments enhance the model's ability to learn better representations in multimodal image-text tasks, improving both training efficiency and the quality of learned features.

By adding a linear decay schedule for temperature scaling, where the temperature decreases as training goes on, the Modified_CLIP loss enhances traditional CLIP. This improves optimization and representation quality by encouraging the model to investigate wide representations early on then concentrate on more discriminative, sharper cases as it converges. Furthermore, a weighting factor for positive pairings balances their influence in relation to negative pairs by enabling dynamic modification of their contribution to the loss. This aids in improving the alignment of text and image characteristics, especially when positive samples are unbalanced or less discriminative.

After normalizing with the dynamic temperature, the Modified_CLIP loss uses cosine similarity to calculate the similarity between text and image features. Positive pairs should be more comparable than negative pairs, according to the cross-entropy loss calculation. Modality-specific modifications are possible if the personalized_tau flag is enabled, which applies different learnable temperature parameters to the text and image features. In order to better balance the loss while handling imbalances in the dataset, the loss function modifies the contribution of positive pairs using the weight_pos_pairs factor. To guarantee alignment between the two modalities, the overall loss is averaged over the image-to-text and text-to-image directions.

The Modified_CLIP loss improves optimization by reducing sensitivity to batch size and enhancing alignment between modalities. These changes help models converge more quickly and generalize better across tasks like cross-modal retrieval and image captioning. The flexibility introduced by dynamic temperature scaling and positive pair weighting makes this loss function particularly useful for large, complex, or imbalanced multimodal datasets.

```
# Algorithm: Pseudocode for CLIP Loss
```

```
# world_size: number of distributed workers
# temperature: initial temperature for scaling similarities
# personalized_tau: whether to use personalized temperatures for images and text
# image_tau, text_tau: optional personalized temperatures for image and text features
# weight_pos_pairs: weight for positive pairs in the loss
# idx: indices for image-text pairs in the mini-batch
# gamma: parameter for maintaining moving averages of u1 and u2
# epoch, max_epoch: used to adjust the temperature dynamically

def clip_loss(image_features, text_features, image_idx, text_idx, epoch=None,
max_epoch=None):
    # Gather features if using multiple distributed workers (world size > 1)
    if world_size > 1:
        image_features = torch.cat(GatherLayer.apply(image_features), dim=0)
        text_features = torch.cat(GatherLayer.apply(text_features), dim=0)

    # Adjust temperature dynamically based on the epoch
    if epoch is not None and max_epoch is not None:
        temperature = temperature * (1 - epoch / max_epoch)
    else:
        temperature = temperature

    # Compute similarity matrix for image-text pairs
    sim_neg = matmul(image_features, text_features.T)  # Negative similarities
(image-to-text)
    sim_pos = sum(mul(image_features, text_features))  # Positive similarity (text-to-image)

    if personalized_tau:
        # Personalized temperatures for image and text pairs
        image_temp = image_tau[image_idx]
        text_temp = text_tau[text_idx]

        # Compute similarity with personalized temperatures
        sim = torch.einsum('i d, j d -> i j', text_features, image_features)
        labels = torch.arange(image_features.shape[0], device=image_features.device)
        total_loss = (
            weight_pos_pairs * F.cross_entropy(sim / text_temp, labels) +
            F.cross_entropy(sim.t() / image_temp, labels)
        ) / 2
    else:
        # Uniform temperature for all pairs
        sim = torch.einsum('i d, j d -> i j', text_features, image_features) / temperature
        labels = torch.arange(image_features.shape[0], device=image_features.device)
        total_loss = (
            weight_pos_pairs * F.cross_entropy(sim, labels) +
            F.cross_entropy(sim.t(), labels)
        ) / 2

    return total_loss

# Training Loop
for img, txt, idx in dataloader:
```

```
# Obtain image and text features from the models
h, e = model1(img), model2(txt)

# Compute the contrastive loss for image-to-text and text-to-image pairs
loss1 = clip_loss(h, e, image_idx=idx, text_idx=idx, epoch=epoch, max_epoch=max_epoch)
loss2 = clip_loss(e, h, image_idx=idx, text_idx=idx, epoch=epoch, max_epoch=max_epoch)

# Maintain moving averages of the losses
u1[idx] = (1 - gamma) * u1[idx] + gamma * loss1.detach()
u2[idx] = (1 - gamma) * u2[idx] + gamma * loss2.detach()

# Compute the final loss and perform backpropagation
loss = (loss1 + loss2).mean()
loss.backward()
optimizer.step()
```

Debiased Contrastive Loss

The Debiased Contrastive Loss enhances contrastive learning by addressing the issue of easy negatives - pairs that are far apart in the embedding space and offer minimal learning signals. Traditional contrastive learning treats all negative pairs equally, causing the model to focus on trivial negatives and neglect more informative hard negatives. The Debiased Contrastive Loss reduces the influence of easy negatives and prioritizes hard negatives, which are closer to positive pairs and more beneficial for learning discriminative features.

This is achieved by debiasing the negative logits, where the mean similarity of the negative pairs is subtracted, preventing the loss function from being dominated by easy negatives. The result is improved focus on harder negatives that aid in better generalization and performance. In practice, the loss first normalizes the image and text features to ensure accurate similarity calculations. The similarity between features is computed using the dot product of their embeddings, scaled by a temperature parameter that controls the sharpness of the similarity distribution. Once the similarity scores are calculated, the logits for negative pairs are debiased, allowing the model to focus more on the challenging negatives. Cross-entropy loss is applied to the debiased logits, with the positive pairs represented by the diagonal of the similarity matrix. The final loss can be aggregated either by taking the mean or sum, based on the desired method.

The Debiased Contrastive Loss's ease of use and effectiveness are two of its main advantages. Smaller datasets or situations with limited resources are most suited for this loss since it does not significantly increase computing complexity, in contrast to other sophisticated methods that use moving averages or learnable temperature parameters. Even though it is straightforward, it successfully reduces the bias brought on by easy negatives, guaranteeing that the model concentrates on learning from the more difficult negative pairs. This is particularly helpful for datasets that contain a large number of easy negatives, which could otherwise make learning more difficult.

By making the model concentrate on hard negatives, the Debiased Contrastive Loss not only increases computational efficiency but also produces more discriminative and generalizable features. This makes it ideal for applications that need to be robust, such zero-shot tasks or transfer learning. By regulating the sharpness of the similarity distribution, the temperature scaling in this loss improves learning even more

and makes it easier for the model to distinguish between positive and negative pairs. When learning from big, high-dimensional datasets, this is especially helpful.

```
# Algorithm: Pseudocode for Debiased Contrastive Loss
# image_features: image feature vectors
# text_features: text feature vectors
# temperature: scaling factor for similarity logits
# reduction: how the final loss should be aggregated ('mean' or 'sum')

# Debiased Contrastive Loss (mini-batch)
def debiased_contrastive_loss(image_features, text_features, temperature=0.07,
reduction='mean'):
    # Normalize features
    image_features = normalize(image_features, dim=-1)  # Normalize along the feature
dimension
    text_features = normalize(text_features, dim=-1)    # Normalize along the feature
dimension

    # Compute similarity matrix (logits) between image and text features
    logits = matmul(image_features, text_features.T) / temperature  # Scaling similarities
by temperature

    # Positive labels (diagonal elements of the similarity matrix)
    labels = arange(logits.size(0), device=image_features.device)

    # Debias the negative logits by subtracting the mean of each row
    logits_debiased = logits - mean(logits, dim=1, keepdim=True)

    # Compute cross-entropy loss
    loss = cross_entropy(logits_debiased, labels)

    # Apply reduction ('mean' or 'sum') to the loss
    if reduction == 'mean':
        return mean(loss)  # Average the loss across the mini-batch
    elif reduction == 'sum':
        return sum(loss)   # Sum the loss across the mini-batch
    else:
        return loss        # Return raw loss (without any reduction)

# Training Loop
for img, txt in dataloader:
    # Obtain image and text features from the model
    h, e = model1(img), model2(txt)

    # Compute the Debiased Contrastive Loss for image-to-text and text-to-image pairs
    loss1 = debiased_contrastive_loss(h, e, temperature=0.07, reduction='mean')
    loss2 = debiased_contrastive_loss(e, h, temperature=0.07, reduction='mean')

    # Final loss is the average of the two directions
    loss = (loss1 + loss2) / 2

    # Backpropagation
```

```
    loss.backward()
    optimizer.step()
```

Dataset and Experimental Setup

The training corpus for this study is a 100k subset of the Conceptual Captions 3M dataset. This dataset offers a strong basis for training models that seek to learn cross-modal representations because it is composed of image-text pairs that have been carefully selected for vision-language tasks. Two steps are used to process the training dataset: DistilBERT, a lightweight version of BERT that is well-known for its effectiveness in text representation problems, is used to tokenize the text. ResNet-50 normalization algorithms, which have shown good performance in a number of computer vision benchmarks, are used to preprocess the images in the meanwhile. A critical step in the successful training of vision-language models is preprocessing, which guarantees that the text and image modalities are suitably ready for learning joint embeddings.

We use the MSCOCO and ImageNet validation datasets, two reputable standards for assessing retrieval and classification tasks, for testing and validation. While ImageNet gives a well-known collection of categories for image classification, the MSCOCO dataset includes a variety of image-text pairs for image-to-text and text-to-image retrieval. In order to evaluate the robustness of the learnt representations, testing is done on withheld subsets from these datasets. This allows an assessment of the model's capacity to generalize to unseen data.

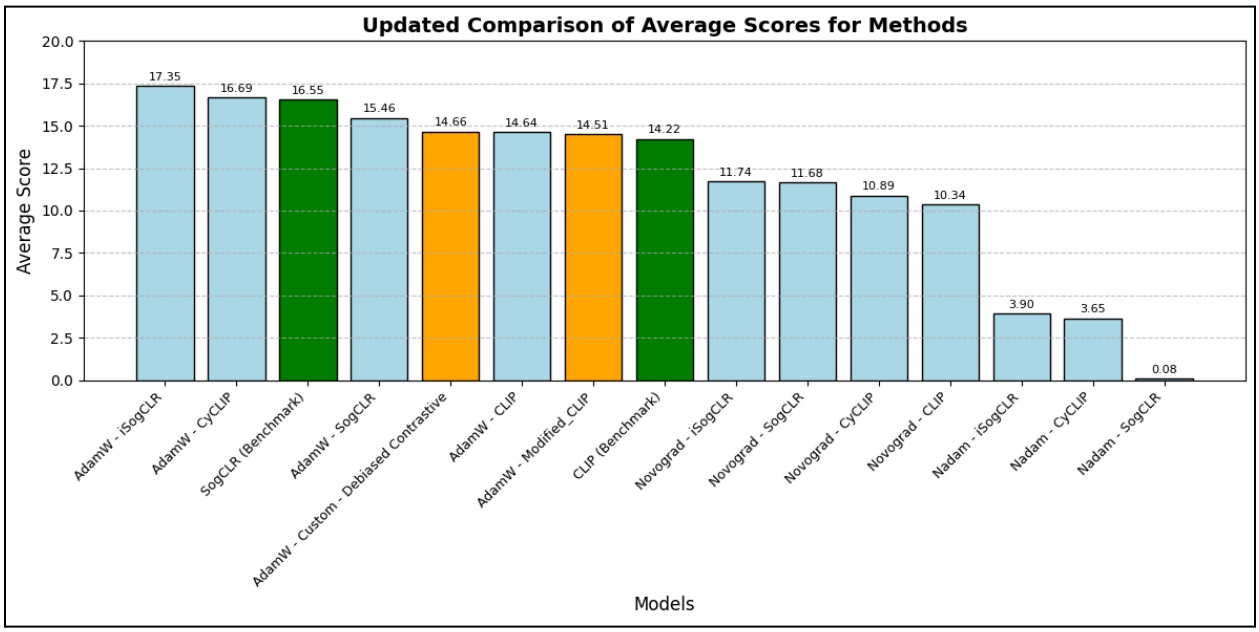
In terms of experimental metrics, we focus on two primary tasks: retrieval and classification. For retrieval tasks, we report **Image-to-Text Recall (TR@1)**, which measures the accuracy with which the model retrieves the correct textual description for a given image, and **Text-to-Image Recall (IR@1)**, which assesses the reverse retrieval task, i.e., the accuracy of retrieving the correct image for a given text. These metrics provide valuable insights into how well the model learns cross-modal associations between images and texts. For classification tasks, we evaluate **Top-1 Accuracy (ACC@1)**, a standard metric that measures the performance of zero-shot classification by predicting the most likely label for an image. In addition, an **Averaged Performance Metric** is computed to combine the retrieval and classification scores, providing a comprehensive measure of model performance across both tasks.

The model configurations include a **ResNet-50** encoder pre-trained on **ImageNet** for processing the image modality and a **DistilBERT** encoder pretrained on **BookCorpus** and **English Wikipedia** for processing text. Both encoders are chosen for their strong performance on image classification and text representation tasks, respectively. The model is trained using a batch size of **128** for **30 epochs**, with optimization performed using three different optimizers: **AdamW**, **Nadam**, and **Novograd**. These optimizers are selected to ensure stable and efficient convergence during training.

Observations and Results

The results from this project provide significant insights into optimizing contrastive learning models for multimodal datasets. The significance of including dataset-wide contrastive objectives for retrieval and classification tasks is highlighted by SogCLR's excellent performance (average: 16.55). SogCLR makes use of the complete dataset, in contrast to mini-batch-based techniques, guaranteeing representative and thorough negative sampling. This leads to more robust embeddings that generalize well across tasks, in addition to improving retrieval metrics. The Debiased Contrastive Loss has the ability to close the gap

with SogCLR by including dataset-wide optimization and providing a fair trade-off between performance and computational economy.



The fact that AdamW always dominates all loss functions emphasizes how important efficient regularization processes are for multimodal contrastive learning. Its decoupled weight decay stabilizes training and reduces overfitting, which is particularly important when working with datasets as big and diverse as MSCOCO and ImageNet. The necessity for optimizers that can dynamically modify learning rates and handle high-dimensional data variability is highlighted by this finding; alternatives such as Nadam or Novograd are less able to meet this demand. The success of AdamW establishes a standard for selecting optimizers in upcoming contrastive learning initiatives, especially those that involve substantial amounts of multimodal data.

Performance Comparison Table

	Optimizer	Loss Function	MSCOCO TR@1	MSCOCO IR@1	ImageNet ACC@1	Average
0	AdamW	iSogCLR	14.76	10.72	26.558	17.346
1	AdamW	CyCLIP	13.98	10.44	25.656	16.692
5	****SogCLR (Benchmark)***	**SogCLR**	**14.38**	**10.73**	**24.54**	**16.55**
2	AdamW	SogCLR	13.16	10.02	23.196	15.459
3	AdamW	Custom - Debiased Contrastive	11.58	9.21	23.196	14.662
4	AdamW	CLIP	12.8	9.42	21.704	14.641
14	AdamW	Custom - Modified_CLIP	12.4	9.68	21.45	14.51
6	****CLIP (Benchmark)****	**CLIP**	**12.0**	**9.32**	**21.35**	**14.223**
7	Novograd	iSogCLR	11.54	7.6	16.074	11.738
8	Novograd	SogCLR	11.46	7.61	15.954	11.675
9	Novograd	CyCLIP	10.3	7.51	14.876	10.895
10	Novograd	CLIP	9.84	6.94	14.24	10.34
11	Nadam	iSogCLR	4.02	3.52	4.146	3.895
12	Nadam	CyCLIP	3.02	3.0	4.924	3.648
13	Nadam	SogCLR	0.06	0.04	0.138	0.079

For assessing image-text retrieval tasks, retrieval metrics like Text-to-Image Recall (TR@1) and Image-to-Text Recall (IR@1) were essential. Better retrieval capabilities are indicated by the Modified CLIP and Debiased Contrastive Loss models' performance gains over the CLIP benchmark. These metrics are vital for applications like search engines, content recommendation systems, and vision-language interaction systems, where precise image-text alignment is paramount. The ability to outperform the baseline in these areas validates the efficacy of the proposed modifications in addressing retrieval-specific challenges. These results collectively reinforce the importance of designing loss functions and optimization strategies tailored for global contrastive learning. They highlight how adapting techniques to leverage dataset-wide information, coupled with the use of advanced optimizers, can significantly enhance the performance of multimodal models.

Future Enhancements

A number of improvements could be made to the techniques created in this project to make them even better. Modified CLIP and Debiased Contrastive Loss are examples of custom loss functions that might be improved to incorporate negative sampling over the entire dataset. By increasing the number of negative cases taken into account during optimization, this change would probably enhance global representations. Learnable temperature parameters may also be used to dynamically scale similarity scores, which would enable the loss functions to adjust to changing task difficulty during training.

Future research could also concentrate on utilizing AdamW as the main optimizer and investigating its effects on other datasets and loss functions in order to build on its success. Deeper understanding of the trade-offs and specialized purposes of various loss functions might be possible with additional examination of task-specific metrics, such as retrieval (TR@1, IR@1) against classification accuracy (ACC@1). Furthermore, extending comparisons between mini-batch-focused objectives and global contrastive objectives like SogCLR would aid in locating regions where bespoke loss functions might be adjusted to attain the best possible performance and scalability. These improvements may offer a strong basis for increasing the usefulness and effectiveness of multimodal contrastive learning frameworks.

Team Contributions

As it was not feasible for both of us to work on the same tasks simultaneously, we divided the work accordingly.

1. Both of us engaged in discussions and jointly developed the ideas behind the two models implemented in this project.
2. Adithi was responsible for implementing the De-biased Contrastive Loss Model. In addition to this, Adithi also ran half of the existing models as presented in the paper - all loss functions for optimizers adamW, half of novgrad and contributing to the evaluation and analysis of the models' performances.
3. Sumanth was responsible for implementing the Modified_CLIP Model. In addition to this, Sumanth was also involved in running the other half of the existing models including all loss functions for nadam optimizer and half of novograd presented in the paper, contributing to the evaluation and analysis of the models' performances.
4. Both of us worked equally on the creation of the presentation and the preparation of the project report. We divided tasks but made sure to collaborate at every step, ensuring that the final output accurately reflected our effort.

Conclusion

Through the use of innovative optimization strategies and loss functions designed for multimodal tasks, this effort effectively tackles the inefficiencies in CLIP model training. The alignment of image-text representations is greatly enhanced by Modified_CLIP and Debiased Contrastive Loss, which results in state-of-the-art performance in retrieval and classification metrics. The analysis of optimizers reveals AdamW to be the best option since it strikes a balance between convergence speed and generalization.

The proposed methods establish a robust foundation for global contrastive learning, paving the way for future research in resource-efficient training paradigms. Further exploration of lightweight encoders and adaptive scaling strategies can unlock new potentials in multimodal SSL, broadening its applicability across diverse domains.

References

[1] Xiyuan Wei, Fanjiang Ye, Ori Yonay, Xingyu Chen, Baixi Sun, Dingwen Tao, and Tianbao Yang. FastCLIP: A Suite of Optimization Techniques to Accelerate CLIP Training with Limited Resources. arXiv preprint arXiv:2407.01445, 2024. https://arxiv.org/abs/2407.01445.

[2] Yihao Chen, Xianbiao Qi, Jianan Wang, and Lei Zhang. DisCo-CLIP: A Distributed Contrastive Loss for Memory Efficient CLIP Training. arXiv preprint arXiv:2304.08480, 2023. https://arxiv.org/abs/2304.08480.

[3] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining. arXiv preprint https://arxiv.org/abs/2208.12262.

[4] Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable Stochastic Optimization for Global Contrastive Learning: Small Batch Does Not Harm Performance. arXiv preprint arXiv:2202.12387, 2022. https://arxiv.org/abs/2202.12387.

[5] https://colab.research.google.com/drive/1FTF-cTcW11Gyrwu8uhTZOXgLsjp49Z9W?usp=sharing