TEXAS A&M UNIVERSITY
Engineering

TEXAS A&M UNIVERSITY
Department of Computer Science & Engineering

CSCE 636 DEEP LEARNING: Improving CLIP Training

By
Sonjoy Kumar Paul
Phanender Chalasani

Introduction

Contrastive Learning Image Pretraining (CLIP) is a neural network that uses natural language supervision to learn visual concepts. CLIP can perform zero-shot visual classification when provided with the names of visual categories. CLIP models are trained to understand the relationship between images and their textual descriptions by learning joint representations in a shared embedding space.

Advantages of CLIP

- **Highly Efficient**: CLIP learns from diverse, unstructured, and noisy datasets, designed for flexible zero-shot inference across different visual tasks.

- **Resource-Intensive Datasets**: CLIP learns efficiently from text-image pairs freely available on the internet compared to deep learning models, which traditionally relies on expensive, manually labeled datasets with limited visual concepts.

CLIP MODAL ARCHITECTURE

CLIP uses a dual-encoder architecture to map images and text into a shared latent space. It works by jointly training two encoders. One encoder for images (Vision Transformer) and one for text (Transformer-based Language Model).

Image Encoder: The image encoder extracts salient features from the visual input. This encoder takes an **image as input** and produces a high-dimensional vector representation. It typically uses a convolutional neural network (CNN) architecture, like **ResNet**, for extracting image features.

Text Encoder: The text encoder encodes the semantic meaning of the corresponding textual description. It takes a **text caption/label as input** and produces another high-dimensional vector representation. It often uses a transformer-based architecture, like a **Transformer** or **BERT**, to process text sequences.

Shared Embedding Space: The two encoders produce embeddings in a shared vector space. These shared embedding spaces allow CLIP to compare text and image representations and learn their underlying relationships.

ARCHITECTURE

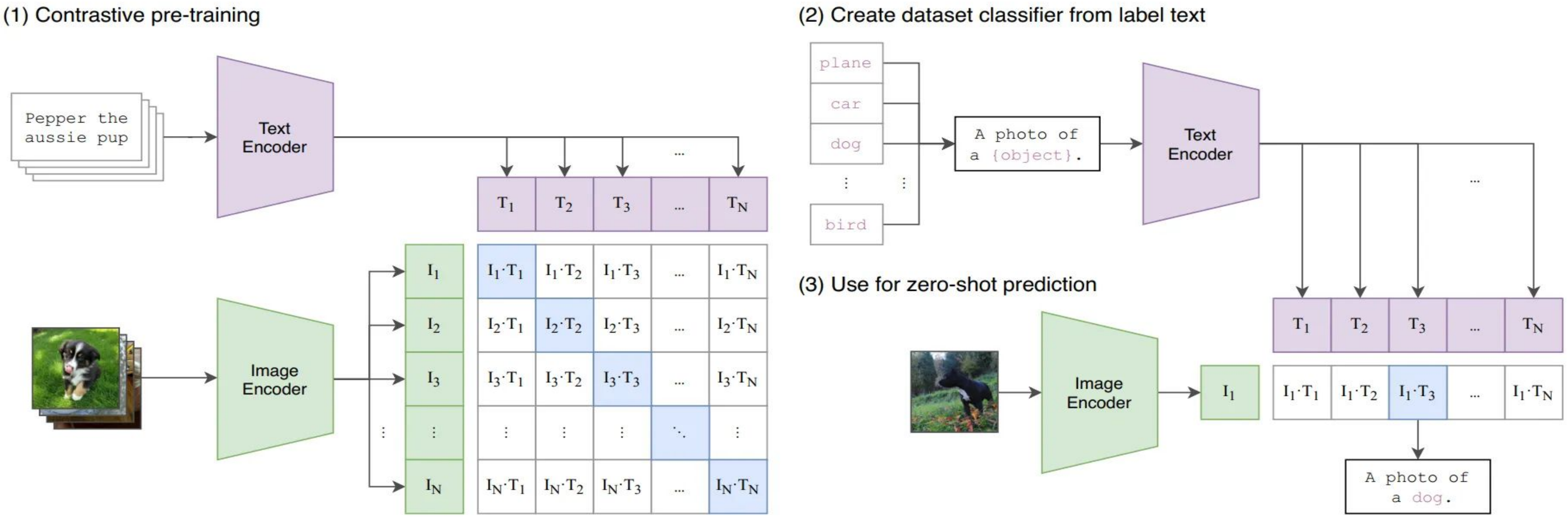

Fig 1: **Architecture overview of CLIP** [Source]

Project Requirements

1) To compare with at least two different optimizers:

We tried three optimizers.

- AdamW
- Stochastic Gradient Descent (SGD)
- Adam

2) To compare at least 3 different loss functions:

We tried five loss functions

- Contrastive Language-Image Pre-training loss (CLIP Loss)
- Stochastic Optimization for Global Contrastive Learning loss (SogCLR Loss)
- Cyclic Contrastive Language-Image Pretraining Loss (CyCLIP Loss)
- Variance-Invariance-Covariance Regularization (VICReg Loss)
- Online Contrastive Learning Representation (OnlineCLR Loss)

Model (Training and Validation)

Dataset

- To train the model, we used a 100k subset of the **Conceptual Captions 3M (CC3M)** dataset, and for validation, we used the **MSCOCO** validation dataset (for retrieval) and the ImageNet validation dataset (for zero-shot classification).

Metric

The evaluation metric is the average of

- Image-to-Text Recall at position 1
- Text-to-Image Recall at position 1 on the retrieval dataset
- Top-1 Accuracy on the classification dataset

Model Architecture

- Image Encoder: ResNet-50 (ImageNet pretrained)
- Text Encoder: DistilBERT (pretrained on BookCorpus & Wikipedia)

Optimizers

Adam (Adaptive Moment Estimation) - Introduced in 2014

- Adam is an optimization algorithm that combines the best properties of AdaGrad and RMSProp to handle sparse gradients in noisy problems
- Adam computes individual adaptive learning rates for different parameters based on gradient history, making it particularly effective for deep neural networks

AdamW - Introduced in 2017

- AdamW is a modified version of Adam. It decouples weight decay from the gradient update.
- AdamW performs weight decay only after controlling the parameter-wise step size, preventing the regularization term from affecting the moving averages. This modification allows models trained with AdamW to generalize much better than those trained with standard Adam

Optimizers

Stochastic Gradient Descent (SGD) - Introduced in 1951

- SGD is a variant of gradient descent that processes one random training example (or a small batch) at a time instead of using the entire dataset for each iteration
- SGD performs frequent updates with high variance, which enables it to jump to potentially better local minima (can lead to complication of convergence to the exact minimum)

Loss Functions

Contrastive Language-Image Pre-training loss (CLIP Loss)

- CLIP loss is a contrastive learning objective that optimizes the similarity between paired image and text embeddings while minimizing similarity between unpaired ones.
- The loss function computes the cosine similarity between all possible image-text pairs in a batch and applies a symmetric cross-entropy loss to maximize the similarity scores of genuine pairs while minimizing scores for incorrect pairings.

Stochastic Optimization for Global Contrastive Learing loss (SogCLR Loss)

- SogCLR (Second-Order Gradient CLIP Learning Rate) loss is an advanced variant of CLIP loss that introduces adaptive weighting of negative pairs using second-order gradient information.
- SogCLR implements a stability mechanism to prevent numerical overflow and optionally includes a square hinge loss surrogate function for better gradient behavior.

Loss Functions

Cyclic Contrastive Language-Image Pretraining Loss (CyCLIP Loss)

- CyCLIP loss enhances the standard CLIP contrastive loss by adding two additional consistency constraints, the in-modal cyclic consistency and cross-modal cyclic consistency.
- The in-modal cyclic consistency ensures that similarity relationships between pairs of images match those between their corresponding text pairs
- The cross-modal cyclic consistency enforces symmetry in image-to-text and text-to-image similarity computations.

Variance-Invariance-Covariance Regularization (VICReg Loss)

- The invariance term ensures similar embeddings for related inputs through MSE loss.
- The variance term prevents representation collapse by maintaining a minimum standard deviation.
- The covariance term decorrelates different dimensions of the embeddings by minimizing the off-diagonal elements of the covariance matrix.
- Three components are weighted by coefficients (sim_coeff, std_coeff, and cov_coeff) to balance their contributions to the final loss

Loss Functions

Online Contrastive Learning Representation (OnlineCLR Loss)

- Improves upon traditional contrastive learning by maintaining running estimates of positive and negative pair distributions to adaptively reweight samples during training.

- The loss function uses different temperature parameters for positive and negative pairs , and maintains moving averages to track the distribution of similarity scores over time.

- This adaptive reweighting strategy helps the model focus on more informative examples and achieve better performance without requiring large batch sizes or memory banks.

Results

Optimizer	Method	MSCOCO TR@1	MSCOCO IR@1	ImageNet ACC@1	Average
AdamW	SogCLR	13.18	10.3	24.55	16.68
AdamW	CLIP	11.62	9.16	21.66	14.81
AdamW	CyCLIP	14.1	10.68	25.91	16.9
AdamW	VicReg	2.86	2.16	5.79	3.6
AdamW	OnlineCLR	10.96	8.64	20.52	13.37
SGD	SogCLR	1.56	1	2.87	1.81
SGD	CLIP	10.3	7	17.01	11.44
SGD	CyCLIP	10.38	7.31	16.86	11.52
SGD	VicReg	2	1.6	2.43	2.01
SGD	OnlineCLR	0.74	0.56	1.5	0.93
Adam	SogCLR	0.1	0.1	0.23	0.14
Adam	CLIP	3.34	3.03	4.71	3.69
Adam	CyCLIP	3.92	3.25	4.04	3.74
Adam	VicReg	0.66	0.63	1.28	0.86
Adam	OnlineCLR	0.02	0.02	0.1	0.05

Benchmarks

Optimizer	Method(Loss Function)	MSCOCO TR@1	MSCOCO IR@1	ImageNet ACC@1	Average
AdamW	CLIP	12.0	9.32	21.35	14.22
AdamW	SOGCLR	14.38	10.73	24.54	16.55

Conclusion

Best Performing Configurations

- AdamW emerges as the clearly superior optimizer across all loss functions
- CyCLIP with AdamW achieves the best overall performance (16.9% average), followed closely by SogCLR (16.68% average)
- The results match or exceed the benchmarks, with CyCLIP and SogCLR showing particularly strong performance

Optimizer Impact

- AdamW consistently outperforms both SGD and Adam by a large margin
- SGD shows moderate performance only with CLIP and CyCLIP
- Adam performs poorly across all loss functions, suggesting it's not suitable for these contrastive learning tasks

THANK YOU