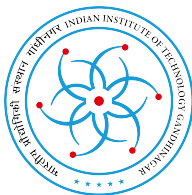


# Mini-CLIP: Contrastive Language–Image Pretraining

Aditya Borate, Aryan Solanki & Nishchay Bhutoria

ES 667 - Indian Institute of Technology Gandhinagar

November 27, 2025





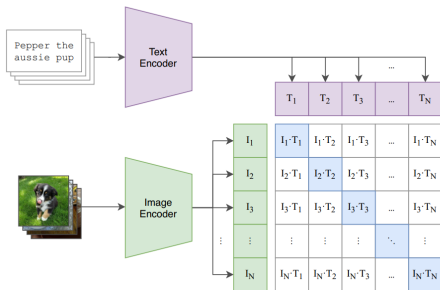
# Why is CLIP So Useful?

- **Natural Language Supervision** Learns visual concepts directly from text descriptions instead of fixed labels (unlike in a neural network).
- **Zero-Shot Generalization** Performs tasks on new datasets without any fine-tuning by comparing image and text embeddings.
- **Unified Multimodal Space** Maps images and text into a shared embedding space, enabling flexible retrieval and matching.
- **Scales Extremely Well** Larger CLIP models trained on billions of image-text pairs show strong cross-domain robustness.
- **Enables Many Applications**
  - Image  $\rightarrow$  Text and Text  $\rightarrow$  Image retrieval
  - Zero-shot classification
  - Caption ranking and filtering
  - Vision-language pretraining for downstream tasks

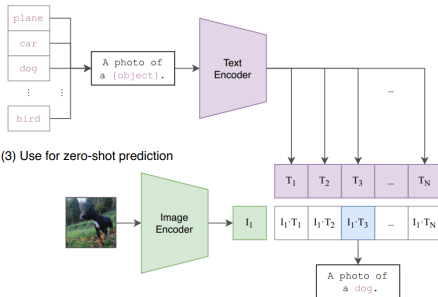
- Goal: Implement a miniature version of CLIP (Contrastive Language–Image Pretraining).
- Learn joint embeddings of images and text using contrastive learning.
- Enables:
  - Zero-shot image classification
  - Cross-modal image–text retrieval
  - Multimodal representation learning

# Mini-CLIP Architecture: Contrastive Pre-training

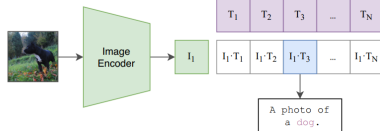
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



- **ViT-S/16**

- Vision Transformer with small capacity and  $16 \times 16$  patch size.
- Pretrained on large-scale image data.
- Used as the primary image encoder in Mini-CLIP.

- **ResNet-101**

- Deep convolutional residual network (101 layers).
- Serves as a stronger CNN baseline encoder.
- Allows comparison of transformer vs. CNN for CLIP-style training.

- **RoBERTa-base**
  - Transformer-based masked language model.
  - Pretrained on large corpora with robust optimization.
  - Encodes captions into sentence embeddings.
- Output text embeddings are projected into the shared image–text embedding space for contrastive learning.

# Dataset: Flickr30k

- 31k images, each with 5 human-written captions.
- 150k total image–caption pairs.
- Suitable for contrastive multimodal training within compute limits.

- **Image Encoder:** ViT-S/16 (pretrained)
- **Text Encoder:** RoBERTa-base (pretrained)
- **Projection Heads:**
  - Linear layers mapping to shared embedding space.
- Joint embedding dimension:  $d$

# Training Objective: InfoNCE Loss

## Setup

$$(I_i, T_i)_{i=1}^N$$

## Embeddings

$$\tilde{z}_i^I = \frac{f_{\text{img}}(I_i)}{\|f_{\text{img}}(I_i)\|}, \quad \tilde{z}_i^T = \frac{f_{\text{text}}(T_i)}{\|f_{\text{text}}(T_i)\|}$$

## Similarity

$$S_{ij} = (\tilde{z}_i^I)^\top (\tilde{z}_j^T), \quad \text{logits}_{ij} = \frac{S_{ij}}{\tau}$$

## Temperature

Lower  $\tau$  makes softmax sharper.  
Higher  $\tau$  smooths it.

## Image to text

$$\mathcal{L}_{\text{I2T}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\text{logits}_{ii}}}{\sum_{j=1}^N e^{\text{logits}_{ij}}}$$

## Text to image

$$\mathcal{L}_{\text{T2I}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\text{logits}_{ii}}}{\sum_{j=1}^N e^{\text{logits}_{ji}}}$$

## Final loss

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2} (\mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}})$$

# Training Strategy

- Pretrained encoders + projection heads.
- Optimizer: AdamW with learning rate scheduling.
- Two regimes:
  - Last-layer fine-tuning
  - Full fine-tuning
- Data augmentation on images (resize-crop, flipping).
- Tokenization and batching for text captions.

# Evaluation Setup

- **Image  $\rightarrow$  Text Retrieval (I2T)**

Given an image, retrieve the correct caption.

- **Text  $\rightarrow$  Image Retrieval (T2I)**

Given a caption, retrieve the correct image.

- **Metrics: Recall@1, Recall@5, Recall@10.**

The Recall@K metric is defined as:

$$\text{Recall@K} = \frac{\text{Number of images where correct caption was in Top K}}{\text{Total Number of Images}} \quad (1)$$

**Table:** Performance Comparison on Flickr30k Test Set

*Inference evaluated on a subset of 1000 Flickr30k test images.*

Encoder Models	Fine-tune Method	Image-to-Text (I2T)			Text-to-Image (T2I)		
		R@1	R@5	R@10	R@1	R@5	R@10
ViT(Small) + RoBERTa	Setup 1	21.6	53.6	65.2	22.4	50.4	64.2
ViT + RoBERTa	Setup 2	28.2	68.2	75.4	28.4	62.7	74.6
ViT(Small) + RoBERTa	Setup 3	36.6	72.7	83.8	33.9	71.9	84.0
ResNet101 + RoBERTa	Setup 1	5.7	6.9	15.2	5.5	6.3	14.9
ResNet101 + RoBERTa	Setup 2	22.1	49.4	61.7	20.7	48.3	61.0
CLIP (OpenAI)	Zero-shot	32.3	53.7	62.9	27.6	49.7	59.0

*Setup 1: Fine-tuning only the projection heads.*

*Setup 2: Fine-tuning the entire model.*

*Setup 3: Full fine-tuning followed by stabilization using only the last two layers.*

- Full fine-tuning offers better Recall@K but risks overfitting.
- Last-layer tuning is more stable on small datasets.
- Partial fine-tuning (last 2 transformer blocks) is a strong middle ground.

- Larger batch sizes for stronger negatives.
- Partial fine-tuning of upper transformer layers.
- Experiment with stronger augmentations.
- Evaluate zero-shot classification on new datasets.

# Conclusion

- Mini-CLIP successfully learns a joint vision-language embedding space.
- Enables retrieval and zero-shot tasks.
- Demonstrates the strength of contrastive multimodal learning.

# Group Members

<b>Name</b>	<b>Roll Number</b>	<b>Programme</b>
Aditya Borate	23110065	BTech 2023–27
Aryan Solanki	23110049	BTech 2023–27
Nishchay Bhutoria	23110222	BTech 2023–27

- **GitHub Repository:**

[https://github.com/Aryan-IIT/miniCLIP\\_DL](https://github.com/Aryan-IIT/miniCLIP_DL)

- **Paper Link:**

<https://arxiv.org/pdf/2103.00020>