Contrastive Language Image Pretraining (CLIP)

OUTLINE

Introduction to CLIP – What it is and why it is important?

Contrastive Learning in CLIP – How CLIP learns from paired image-text data?

CLIP’s Architecture and Training – VLM components and training

Limitations and Challenges – Key challenges and limitations in CLIP’s approach

CONTRASTIVE LANGUAGE–IMAGE PRETRAINING

1. **Multimodal**: vision + language (text)
2. **Internet Scale**
   - Trained on image-caption pairs
   - Broad visual understanding
3. **Zero-Shot Capabilities**
   - No training data needed for many classification tasks


Man wearing blue mask


Colorful buttons


Hot air balloon in foggy mountains


Peach fruit on a tree

CLIP DATASET & SCALE

Web ImageText (WIT) Dataset

- 400M image–text pairs gathered from the internet
- Derived from alt text or captions
- Similar word-count to used in GPT-2
- Not publicly available

Training Time

- Largest ResNet-based (RN50x64) CLIP trained in ~18 days on 592 V100 GPUs
- Largest ViT-based CLIP in ~12 days on 256 V100 GPUs.

OUTLINE

Introduction to CLIP – What it is and why it is important?

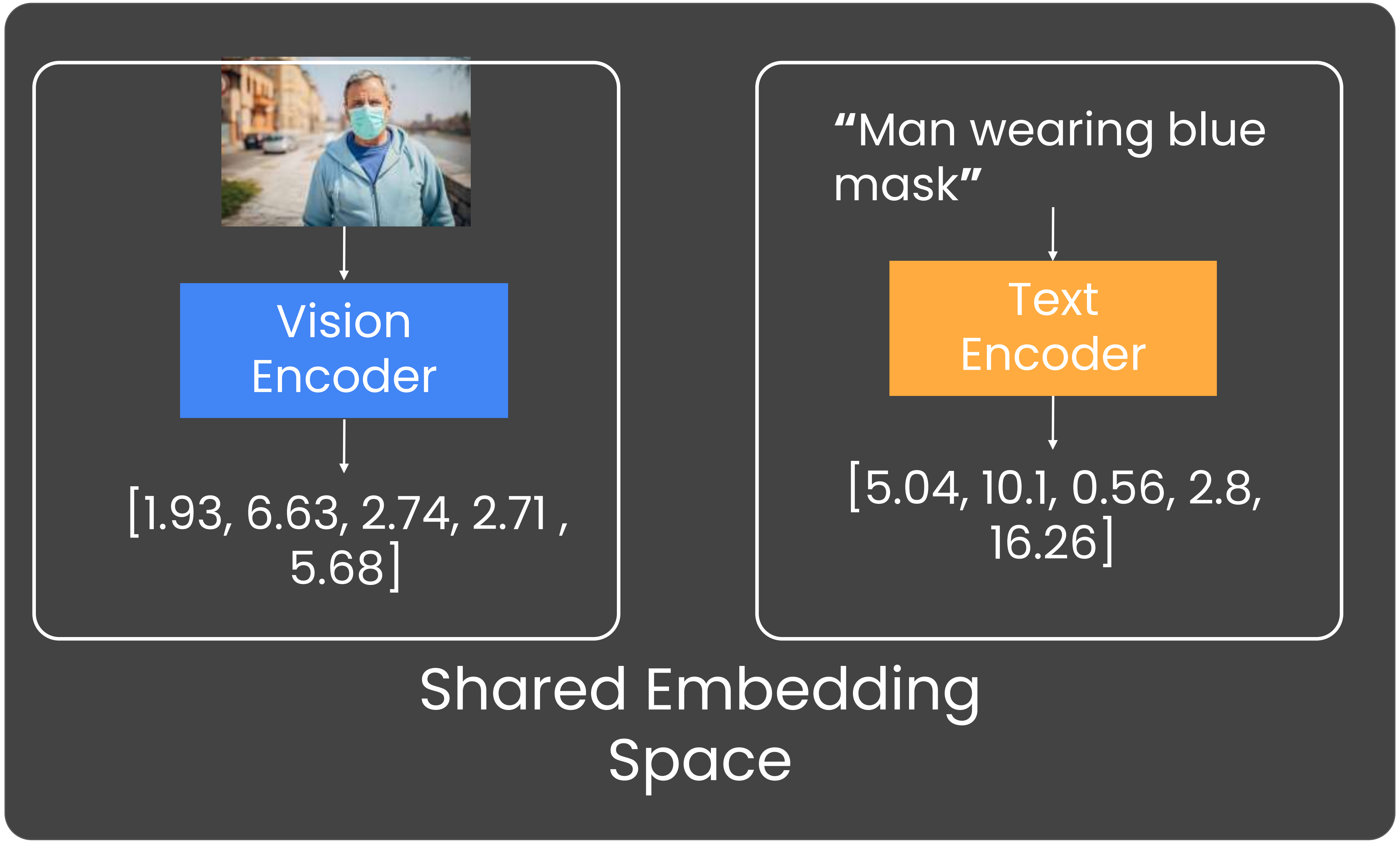
Contrastive Learning in CLIP – How CLIP learns from paired image-text data?

CLIP's Architecture and Training – VLM components and training

Limitations and Challenges – Key challenges and limitations in CLIP's approach

CONTRASTIVE LEARNING

1. Encode image as vector
2. Encode text as same sized vector

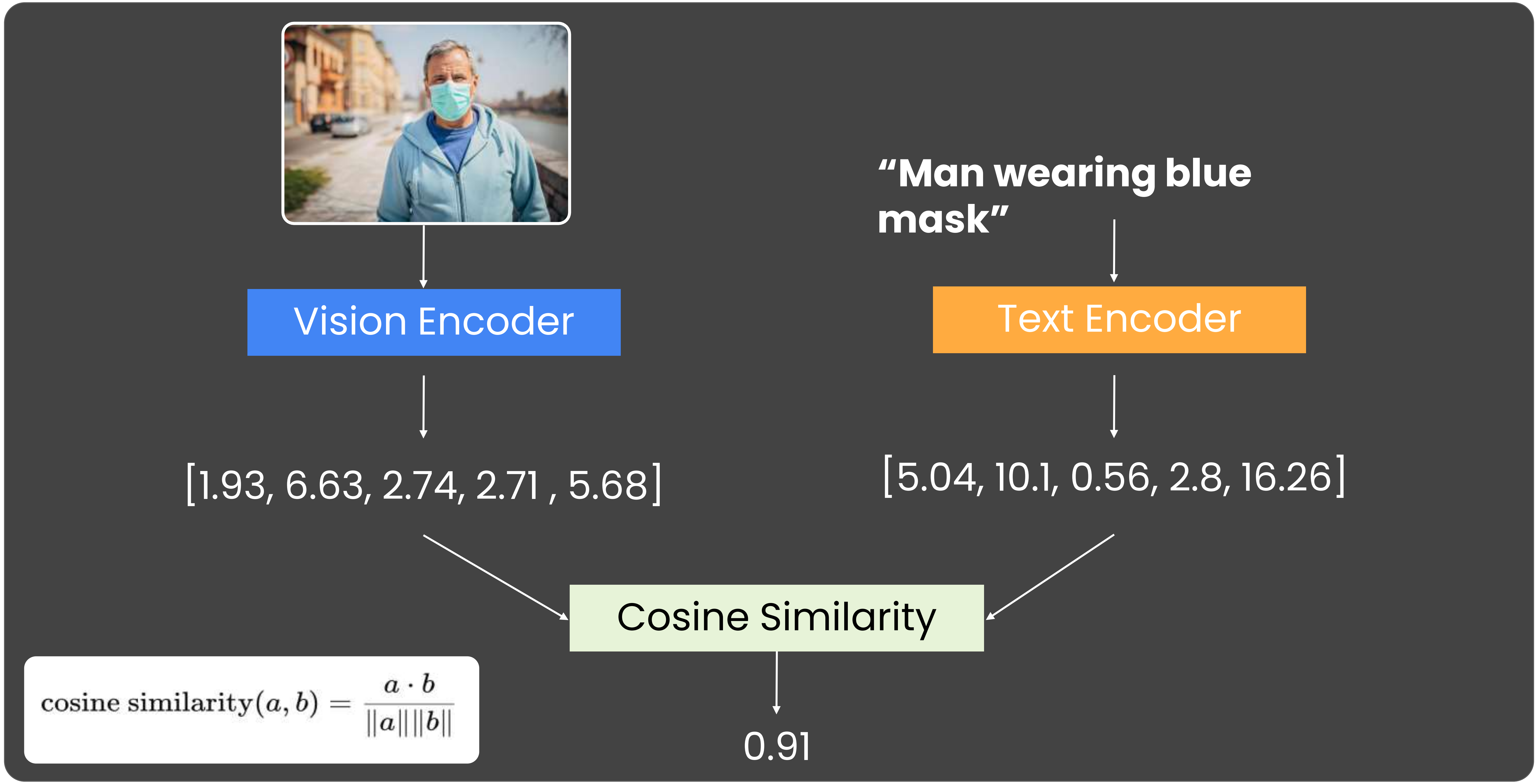


Embedding:

A vector representation of some data (text, image, audio) usually reflecting semantic meaning

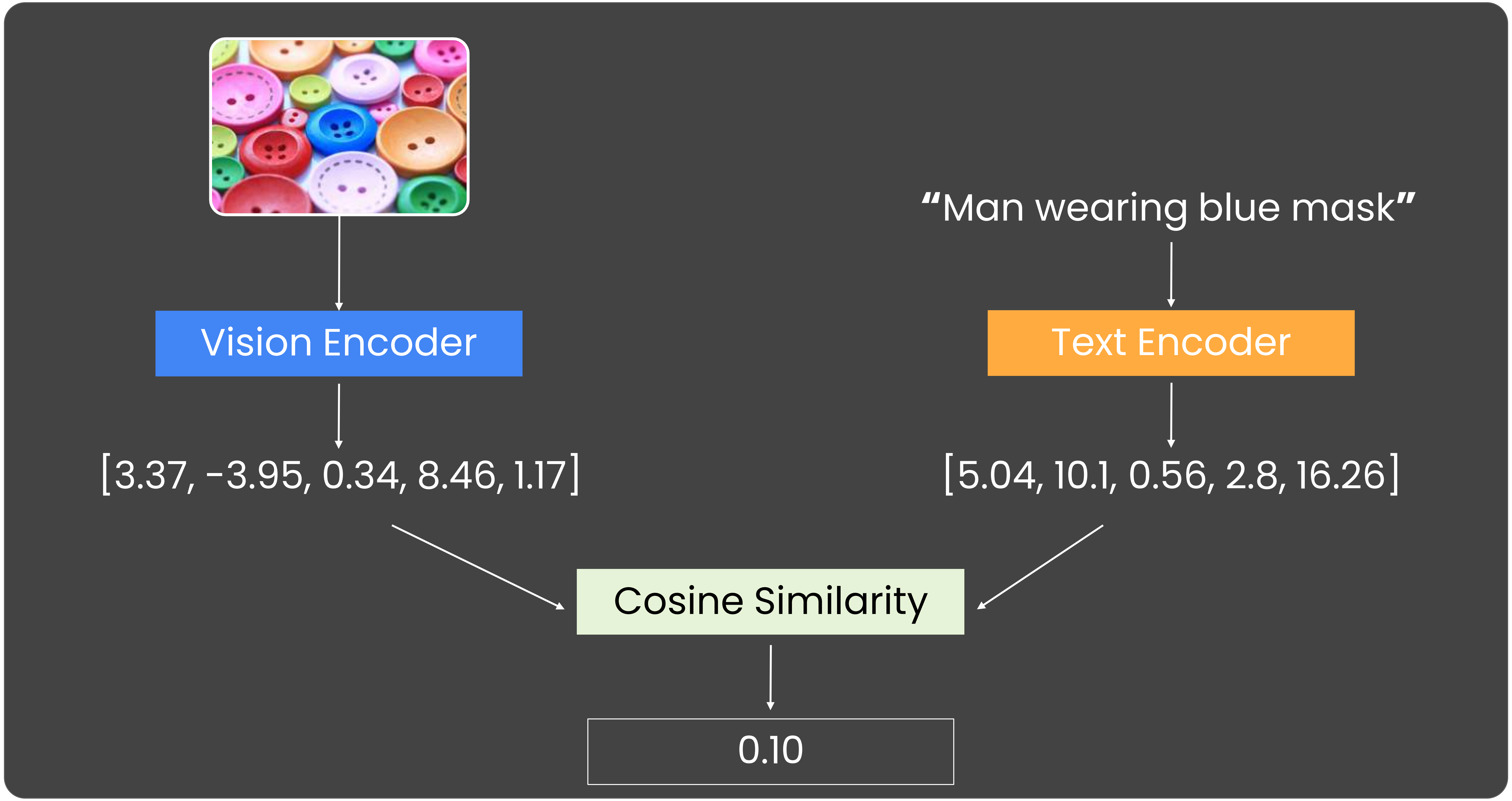
CONTRASTIVE LEARNING

HIGH SIMILARITY FOR CORRECT PAIR



CONTRASTIVE LEARNING

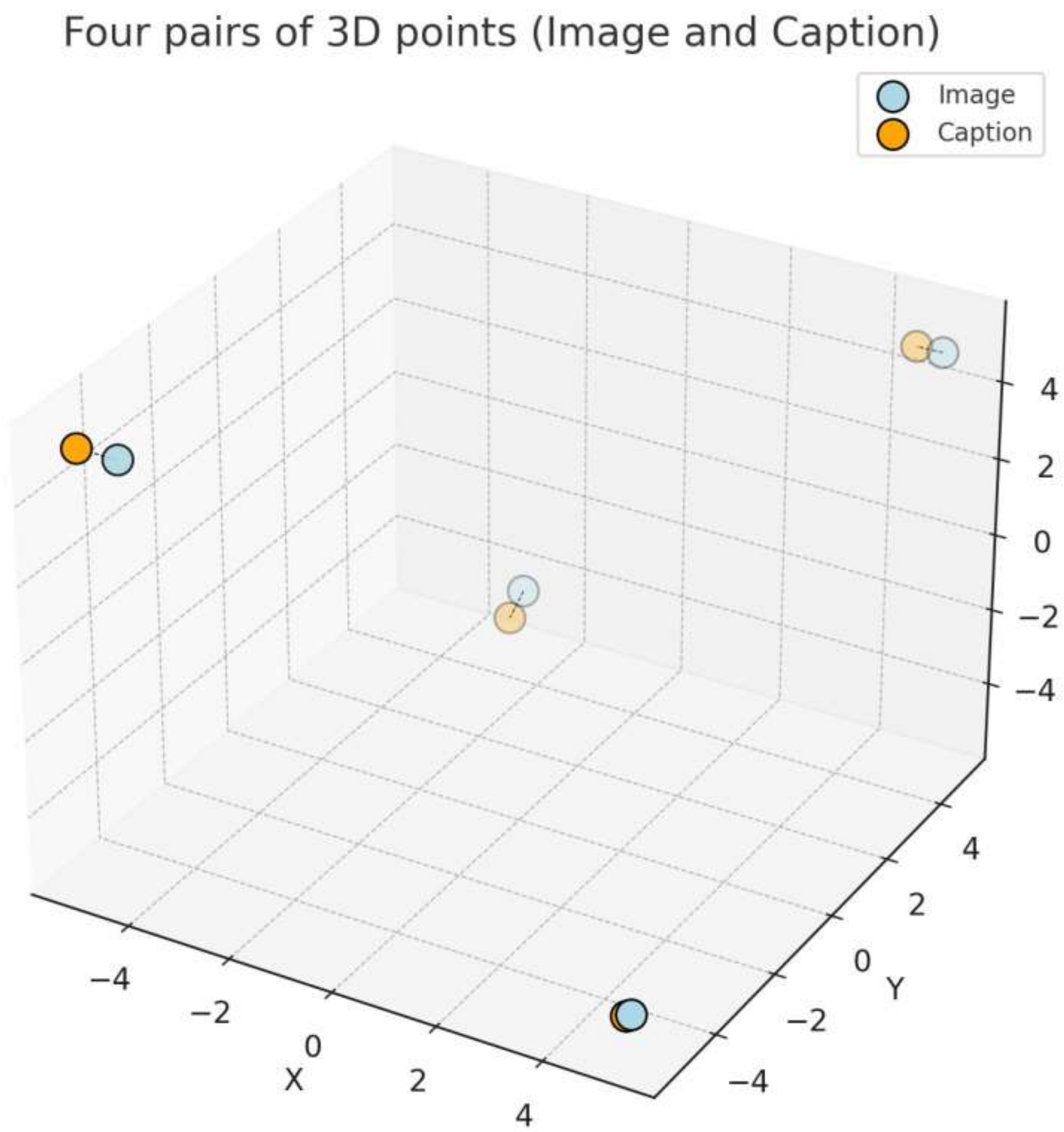
LOW SIMILARITY FOR PAIR



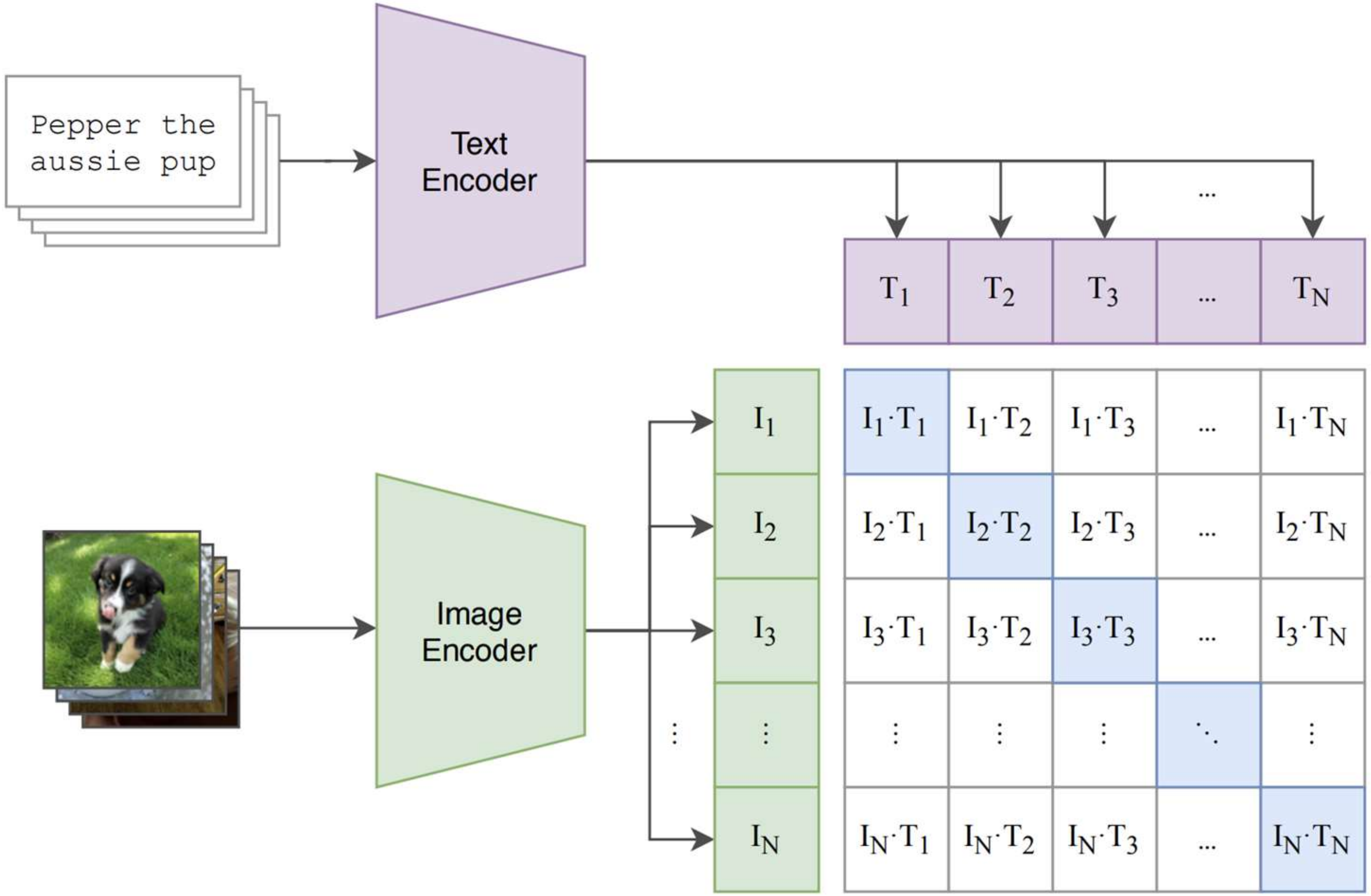
CONTRASTIVE LEARNING: EMBEDDING

Batch size of N image-text pairs

- **Image-text similarity**: Maximize similarity between an image and it's caption while minimize it similarity between the image and other captions.
- **Text-image similarity**: Maximize similarity between a caption and it's image while minimize the similarity between the caption and other images.



CONTRASTIVE LEARNING: TRAINING LOSS



CONTRASTIVE LEARNING: TRAINING LOSS

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\left[\underbrace{-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(I_i, T_j)/\tau)}}_{\text{image-to-text loss}} + \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(T_i, I_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(T_i, I_j)/\tau)}}_{\text{text-to-image loss}}\right]$$

OUTLINE

Introduction to CLIP – What it is and why it is important?

Contrastive Learning in CLIP – How CLIP learns from paired image-text data?

CLIP's Architecture and Training – VLM components and training

Limitations and Challenges – Key challenges and limitations in CLIP's approach

CLIP ARCHITECTURE

Vision Encoder

- ResNet-50 convolutional network or a vision transformer
- Output vector of size “512”

Text Encoder

- 12-layer transformer (with ~63 million parameters, 8 attention heads) similar in architecture to GPT-2
- Output vector size “512”

Parameters of both encoders are adjusted so that matching image/text pairs have higher similarity scores than mismatches

CLIP Architecture and Training

Training

- No pre-trained weights from ImageNet or NLP models.
- Large mini batch sizes were used – on the order of 32,768 examples per batch – enabled by distributed training across many GPUs.

Contrastive Loss Choice

- Predicting captions word-for-word (a generative approach) was much slower to learn visual features.
- A contrastive loss was much faster.

OUTLINE

Introduction to CLIP – What it is and why it is important?

Contrastive Learning in CLIP – How CLIP learns from paired image-text data?

CLIP’s Architecture and Training – VLM components and training

Limitations and Challenges – Key challenges and limitations in CLIP’s approach

CLIP LIMITATIONS

Abstract and Fine-Grained Tasks

- Struggles with counting objects or determining distances, which require more reasoning.
- Has difficulty distinguishing very similar subcategories (e.g., bird species or car models) in zero-shot settings.

Out-of-Distribution Generalization Gaps

- Performs poorly on specialized data like handwritten digits (e.g., MNIST), indicating limited coverage.
- Sensitive to prompt phrasing.

CLIP LIMITATIONS

Biases from Uncurated Web Data

- Absorbs societal biases present in image–caption pairs, leading to potentially harmful outcomes.
- Misclassification can be aggravated by incomplete label sets.

Ethical and Practical Concerns

- Could be misused for surveillance or facial recognition, raising privacy and ethical issues.
- Its open-ended design makes failures less predictable, highlighting the need for cautious deployment.

MODEL CARD

Name	“openai/clip-vit-base-patch32”
Summary	Dual-encoder vision-language model (ViT-B/32 image encoder + Transformer text encoder) trained contrastively on ~400 M image–text pairs to enable zero-shot recognition, retrieval and embedding tasks.
Parameters	151 M
Release Date	January 2021
Developer	OpenAI
License	Code — MIT License; Weights released for research only under OpenAI terms (“no un-evaluated commercial deployment”)

THANK YOU