1 Definitions: Contrastive Language–Image Pre-training (CLIP)

For the original paper, see Radford et al. (2021).

1.1 The Big Picture

1. **Goal**: Learn from (image, caption) pairs, with no fixed list of class labels, so that an image and a text describing the same thing land *close together* in one **shared latent space**, while mismatched images and texts land far apart. When this holds, we say the two modalities are **aligned**. A trained CLIP does not generate anything: it turns an image or a text into a vector and compares vectors.

2. **A training batch**: N image–text pairs $B = \{(I_1, T_1), \ldots, (I_N, T_N)\}$, where image i and text i belong together. These N pairs are **positives**. The other $N(N-1)$ combinations (I_i, T_j) with $j \neq i$ are treated as **negatives**.

1.2 Encoders and Translation Layers

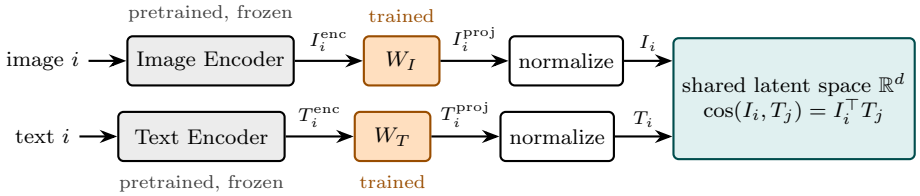
1. **Encoders**: The image encoder (a ResNet or Vision Transformer) extracts visual features, and the text encoder (a Transformer) summarizes the token sequence: $I_i^{\text{enc}} = \text{ImageEncoder}(\text{image } i)$ and $T_i^{\text{enc}} = \text{TextEncoder}(\text{text } i)$. Their outputs live in different spaces and may have different dimensions, so we cannot compare them directly.

2. **Translation layers**: Learned linear maps (*projection layers* in lecture) put both feature vectors into the same dimension d. This is the step that makes the comparison possible:

$$I_i^{\text{proj}} = W_I \, I_i^{\text{enc}} \in \mathbb{R}^d, \qquad T_i^{\text{proj}} = W_T \, T_i^{\text{enc}} \in \mathbb{R}^d. \tag{1.1}$$

3. **Normalization**: Scaling both embeddings to unit length removes the vector's length from the score. Otherwise the model could raise a score by making a vector longer instead of pointing it the right way. The dot product is then the cosine similarity:

$$I_i = \frac{I_i^{\text{proj}}}{\|I_i^{\text{proj}}\|_2}, \qquad T_i = \frac{T_i^{\text{proj}}}{\|T_i^{\text{proj}}\|_2}, \qquad I_i^\top T_j = \cos(I_i, T_j) \in [-1, 1]. \tag{1.2}$$



The encoders can be trained separately

Take a vision model and a language model that were each trained on their own, freeze them, and train only W_I and W_T. The encoders describe their own modality. The translation layers describe how the two relate, by putting them in the same latent space. (The original CLIP trains the encoders and translation layers together from scratch, with the same loss.)

1.3 The Contrastive Loss

1. **Score matrix**: Compute every image–text score at once, with $S_{ij} = I_i^\top T_j / \tau$ (the logit for pair (i, j)). The **diagonal** holds the positives, and everything off the diagonal is a negative. The temperature $\tau > 0$ controls the softmax: a smaller τ magnifies score differences. For example, cosines $[0.30, 0.25, 0.10]$ give probabilities $[0.36, 0.34, 0.30]$ with $\tau = 1$ but $[0.57, 0.35, 0.08]$ with $\tau = 0.1$. In CLIP, τ is learned, not hand-tuned.

	T_1	T_2	T_3	T_4
I_1	$I_1^\top T_1/\tau$	$I_1^\top T_2/\tau$	$I_1^\top T_3/\tau$	$I_1^\top T_4/\tau$
I_2	$I_2^\top T_1/\tau$	$I_2^\top T_2/\tau$	$I_2^\top T_3/\tau$	$I_2^\top T_4/\tau$
I_3	$I_3^\top T_1/\tau$	$I_3^\top T_2/\tau$	$I_3^\top T_3/\tau$	$I_3^\top T_4/\tau$
I_4	$I_4^\top T_1/\tau$	$I_4^\top T_2/\tau$	$I_4^\top T_3/\tau$	$I_4^\top T_4/\tau$

image-to-text, (1.3): softmax across row 2

text-to-image, (1.4): softmax down column 3

2. **Image-to-text (compare within a row)**: Fix image i and ask: *which of the N texts is my caption?* Each image solves an N-way classification problem whose correct class is $j = i$, so the true caption must outscore every other caption *for this image*:

$$p_{IT}(i \text{ matches } j \mid I_i, T_1, \ldots, T_N) = \frac{\exp(I_i^\top T_j/\tau)}{\sum_{k=1}^{N} \exp(I_i^\top T_k/\tau)}. \tag{1.3}$$

3. **Text-to-image (compare within a column)**: Fix text i and ask: *which of the N images do I describe?* The scores are the same but the competitors differ: the true image must outscore every other image *for this caption*. Below we drop the conditioning and write $p_{IT}(i \text{ matches } j)$ and $p_{TI}(i \text{ matches } j)$.

$$p_{TI}(i \text{ matches } j \mid T_i, I_1, \ldots, I_N) = \frac{\exp(I_j^\top T_i/\tau)}{\sum_{k=1}^{N} \exp(I_k^\top T_i/\tau)}. \tag{1.4}$$

4. **Why take the log?** We want the correct match to be likely for *every* pair at once, i.e., to maximize $\prod_i p_{IT}(i \text{ matches } i)$. The log keeps the same maximizer, but turns the product into a sum that does not underflow. It also undoes the exponential, and $-\log p$ punishes a confident mistake heavily while an already-correct pair costs almost nothing:

$$\ell_i^{IT} = -\log p_{IT}(i \text{ matches } i) = \underbrace{-\frac{I_i^\top T_i}{\tau}}_{\text{pull positive up}} + \underbrace{\log \sum_{j=1}^{N} \exp\Big(\frac{I_i^\top T_j}{\tau}\Big)}_{\text{push competitors down}}. \tag{1.5}$$

5. **The symmetric CLIP objective**: Average the cross-entropy over the batch in both directions, using the same score matrix:

$$\mathcal{L}_{IT} = -\frac{1}{N}\sum_i \log p_{IT}(i \text{ matches } i), \quad \mathcal{L}_{TI} = -\frac{1}{N}\sum_i \log p_{TI}(i \text{ matches } i), \quad \mathcal{L}_{\text{CLIP}} = \frac{\mathcal{L}_{IT} + \mathcal{L}_{TI}}{2}. \tag{1.6}$$

1.4 Using a Trained CLIP

1. **Off-the-shelf classifier (zero-shot)**: To classify a new image I_1 among C classes, write one sentence per class, encode and normalize the sentences to get $T_1, \ldots, T_C$, encode and normalize the image, and choose the sentence with the highest similarity. No new classifier is trained, and no labeled images are used. A softmax turns the scores into a distribution over that list:

$$p_{IT}(1 \text{ matches } c \mid I_1, T_1, \ldots, T_C) = \frac{\exp(I_1^\top T_c/\tau)}{\sum_{c'=1}^{C} \exp(I_1^\top T_{c'}/\tau)}, \qquad \hat{y} = \arg\max_c \; I_1^\top T_c. \tag{1.7}$$

For example, for a photo of a fox with $\tau = 0.1$:

Sentence	"this is a cat"	"this is a dog"	**"this is a fox"**
Cosine similarity	0.55	0.60	**0.80**
Probability	6.7%	11.1%	**82.1%**

2. **Prompt ensembling**: Different templates ("a photo of a {}", "a drawing of a {}", ...) give different class embeddings. Average the normalized template embeddings, then normalize again.

3. **Limitations**: CLIP aligns one global vector per image with one per sentence, so it can miss counting, negation, and object relationships.

2 What Does CLIP Learn?

1. A trained CLIP model is given one new image. What does it output for this image, and for an (image, text) pair? Can CLIP, on its own, write a caption for the image?

2. **Alignment.**

   (a) Take an image encoder trained only for ImageNet classification and a text encoder trained only on text, and add randomly initialized translation layers W_I, W_T into the same $\mathbb{R}^d$. Before any further training, is the cosine $I^\top T$ between an image and a caption meaningful? What training makes it meaningful?

   (b) After training, we multiply every image embedding by the same rotation (orthogonal) matrix R, i.e., $I_i \mapsto RI_i$, and leave the text embeddings unchanged. Does image–text retrieval still work? What if we rotate *both* by the same R?

3 Building the CLIP Loss by Hand

Consider a toy batch of $N = 3$ pairs with $d = 3$. After the encoders and translation layers (1.1), the (not yet normalized) embeddings are:

	$i = 1$	$i = 2$	$i = 3$
Image I_i^{proj}	(2, 0, 0)	(0, 3, 0)	(0, 0, 4)
Text T_i^{proj}	(4, 3, 0)	(6, 8, 0)	(0, 3, 4)

1. **Normalize and score.** Compute the unit-length embeddings using (1.2), then the 3×3 matrix of cosines $I_i^\top T_j$ (rows: images, columns: texts). Which entries are the positives? In each row, which caption is the hardest negative?

2. **Why normalize?** Compute the *raw* dot products $I_i^{\text{proj}\top} T_j^{\text{proj}}$ for all nine pairs. Which caption does image 1 prefer now? What goes wrong without normalization?

3. **Temperature.** To keep the arithmetic simple, use $\tau = 1/(k \ln 2)$, so that every exponential is a power of 2: $\exp(I_i^\top T_j/\tau) = 2^{k\, I_i^\top T_j}$. Look only at image 1 (row 1) and fill in the table.

τ	**Logits** $I_1^\top T_j/\tau$	exp(logits)	$p_{IT}(1 \text{ matches } j)$ for $j = 1, 2, 3$
$\tau \to \infty$			
$\frac{1}{5 \ln 2} \approx 0.29$			
$\frac{1}{10 \ln 2} \approx 0.14$			

How does the distribution change as τ decreases? Does the predicted caption change?

4. **The loss in both directions.** Use $\tau = 1/(5 \ln 2)$, so that $\left[\exp(S_{ij})\right] = \left[\begin{smallmatrix} 16 & 8 & 1 \\ 8 & 16 & 8 \\ 1 & 1 & 16 \end{smallmatrix}\right]$. Compute $p_{IT}(i \text{ matches } i)$ for each row (1.3) and $p_{TI}(i \text{ matches } i)$ for each column (1.4) as fractions, and then $\mathcal{L}_{IT}$, $\mathcal{L}_{TI}$ and $\mathcal{L}_{\text{CLIP}}$ (1.6). You may use $\ln 2 \approx 0.693$, $\ln 3 \approx 1.099$ and $\ln 5 \approx 1.609$.

5. **Rows vs. columns.** Which of the six matching probabilities in part 4 is the lowest? Why is this pair worse in one direction than in the other, and what does this say about using both directions?

4 Zero-Shot Inference

1. A frozen CLIP model classifies an image using the prompts "a photo of a dog", "a photo of a cat" and "a photo of a car", with $\tau = 0.1$. The cosines between the image and the three prompts are $[0.32,\ 0.27,\ 0.05]$. Compute the class probabilities (1.7) and the predicted class. Now add a fourth prompt, "a photo of a wolf", with cosine 0.33. What are the new prediction and the new probability of "dog"? What does this show?

2. Why is this called *zero-shot* classification? Why do we use the prompt "a photo of a {class}" instead of just the class name?

5 Who's on the Porch?

A camera on your porch runs CLIP on every frame to detect whether there is a **non-human animal** on the porch. Each sentence reads "this is a photo of ...". For one frame, with $\tau = 0.1$, the similarities are:

	T_1: an empty porch	T_2: a person on a porch	T_3: a dog	T_4: a cat	T_5: a raccoon
$I^\top T_j$	0.60	0.20	0.55	0.55	0.50

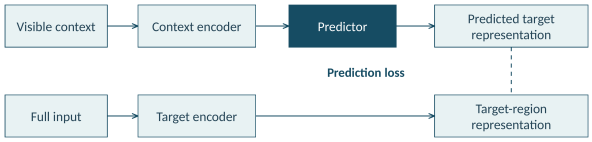
1. What does "choose the highest-similarity sentence" predict? Now sum the probabilities of T_3, T_4, T_5 to get $P(\text{animal})$. Why do the two approaches disagree? Would $P(\text{animal})$ change if we added more animal sentences?

2. A classmate proposes just two sentences: "a porch with an animal" and "a porch with no animal". Why is this risky?

3. At night the camera switches to infrared and accuracy drops. You freeze the encoders and train only the translation layers on captioned night frames. With $I_i^{\text{enc}} \in \mathbb{R}^{768}$, $T_i^{\text{enc}} \in \mathbb{R}^{512}$, and $d = 256$, how many entries do W_I and W_T have? The encoders have 120 million parameters. Why does this suit a small dataset?

6 JEPA on the Porch

JEPA (Assran et al. 2023) starts from the same idea as a masked autoencoder: hide part of an image and try to fill it in. An autoencoder decodes back to pixels. JEPA has no decoder: it encodes the visible and hidden parts into a shared latent space and fills in the *representation* of the hidden part.



With s_x the context representation and s_y the target representation, $\hat{s}_y = \text{Pred}(s_x,\ \text{target location})$ and the loss $E = D(s_y, \hat{s}_y)$ compares features, not pixels. The target encoder follows a moving average of the context encoder.

1. Your porch video has blowing leaves, moving shadows, and rain. Why might filling in the latent space be better here than filling in the pixels?

2. A JEPA predicts the features of the next frame, $\hat{s}_y = (0.60,\ 0.80)$, and $E = \|s_y - \hat{s}_y\|_2^2$ measures surprise. Compute E for each frame. Can this tell you that frame 3 contains a raccoon? How could you combine JEPA and CLIP in the detector?

Frame	1: leaves blowing	2: rain	3: something walks in
s_y	$(0.62,\ 0.78)$	$(0.57,\ 0.82)$	$(0.80,\ 0.60)$

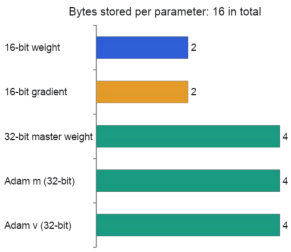
3. You want to use your JEPA encoder with sentences instead of CLIP's image encoder. How would you connect them, and what would you train?

4. Suppose every encoder maps all inputs to one constant vector. What is E? What is $\mathcal{L}_{\text{CLIP}}$ for a batch of N pairs? What does this say about the role of negatives?

7 Definitions: Distributed Training

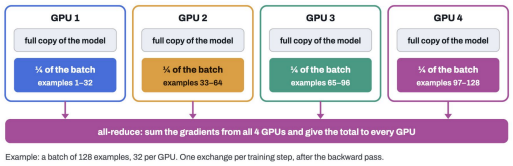
GPU size vs. Adam size: Adam keeps two running averages, m and v, for *every* parameter, plus a 32-bit master copy of each weight. Training with N parameters takes 16 bytes each, before activations:

$$16N = \underbrace{2N}_{\text{weights}} + \underbrace{2N}_{\text{gradients}} + \underbrace{12N}_{\text{Adam}}$$



Llama 3 8B needs 16 GB to run, so it fits on an 80 GB H100. Training it needs 128 GB, so it does not.

Data parallelism: Copy the whole model onto every GPU and split each batch. Each GPU computes its own gradients, then an **all-reduce** ($2N$ bytes, once per step) gives every GPU the same sum, so the copies stay identical. Each GPU still holds $16N$ bytes: more GPUs add speed but no memory.

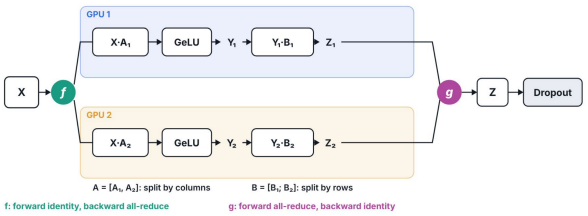


Tensor parallelism: Split each matrix multiplication *inside* a layer across p GPUs. Each GPU stores only its slice of the weights, with the gradients and Adam state for that slice, so memory per GPU is $16N/p$. For $Y = XW$, we can cut W (called A in the figures) by columns, which needs a sum in the backward pass, or by rows, which needs a sum in the forward pass.



	Column-parallel ($W_i \in \mathbb{R}^{D\times H/p}$)	**Row-parallel** ($W_i \in \mathbb{R}^{D/p\times H}$)
Forward	$Y = [XW_1, \ldots, XW_p]$	$Y = \sum_i X_i W_i$ **(all-reduce)**
Backward	$\partial\ell/\partial X = \sum_i G_i W_i^\top$ **(all-reduce)**	$\partial\ell/\partial X_i = G\,W_i^\top$

In a block: Split A by columns, because GeLU is not linear and cannot be applied to partial sums. Then split B by rows, so $Z = Y_1B_1 + Y_2B_2$. Operator f sums backward and g sums forward, so each block needs one sum each way.



Cost: Attention and MLP are two blocks, so each layer needs **4 all-reduces**, each moving (tokens × hidden size × 2) bytes no matter how large p is. So tensor parallelism stays on fast NVLink, inside one server.

Others: **Pipeline**: each GPU owns consecutive layers; idle share is $(p-1)/(m+p-1)$ for m micro-batches. **Optimizer** (ZeRO): shard the duplicated Adam state, then gradients, then weights. **Token** (TeraPipe): the next stage starts on early tokens while later ones finish. **Expert** (MoE): a router sends each token to a few experts through an all-to-all exchange.

8 Training PorchLM

You now want a language model that writes a nightly log of what the camera saw. Use these numbers:

PorchLM parameters	$N = 16 \times 10^9$	H100 memory	80 GB
Transformer layers	$L = 48$	Achieved compute	400 TFLOP/s per GPU
Hidden size	5,000	NVLink	450 GB/s
Tokens per micro-batch	9,000	Network	50 GB/s per GPU

Assume 16 bytes per parameter, 8 GPUs per server, and 1 GB $= 10^9$ bytes, and ignore activation memory. Training costs $6N$ FLOPs per token. Moving M bytes takes M/B, and an all-reduce takes about $2M/B$ (one pass to sum, one to send the total back), matching lecture's 2.5 GB in about 10 ms over NVLink.

1. **GPU size vs. Adam size**: What is the total training memory, and how much of it is optimizer state? Can one H100 run PorchLM for inference? Can it train PorchLM?

2. **Data parallelism**: What is the memory per GPU on 8 H100s? On 64? What is the largest model data parallelism can train on an H100?

3. You shrink to $N = 4 \times 10^9$. Eight H100s each process 9,000 tokens per step. Find the memory per GPU, the compute time, and the gradient all-reduce time over NVLink. What is the speedup over one GPU processing all 72,000 tokens?

4. **Tensor parallelism by hand**: Let $p = 2$, $X = \left[\begin{smallmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 3 & 1 \end{smallmatrix}\right]$, and $W = \left[\begin{smallmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 2 & 1 \end{smallmatrix}\right]^\top$. Compute what each GPU produces for $Y = XW$ under column-parallel, and under row-parallel with the first two rows of W on GPU 1. If the next step squares each entry, compare entry $(1, 1)$ with squaring before and after the row-parallel sum. Which split must come first in an MLP?

5. Back to PorchLM. What is the smallest p that fits on H100s? With $p = 8$, what is the memory per GPU? Why does the Adam state shrink here when it did not under data parallelism?

6. Train PorchLM with $p = 8$ on one server. How many bytes does one all-reduce move, and how long does it take? What is the total communication time per micro-batch? What are the compute time, the step time, and the share of the step spent communicating?

7. A teammate wants $p = 16$ across two servers: "Twice the GPUs means half the compute." Assume every sum now runs at network speed. Find the new step time. Was your teammate right?

8. **Putting it together**: Use tensor parallelism with $p = 8$ inside each of 4 servers, and data parallelism across them. Each server holds one full copy of the model and processes its own 9,000 tokens. At the end of the step, each GPU all-reduces its gradient slice over the network with its match in the other servers. Find the memory per GPU, the size and time of that all-reduce, and the new step time. What is the speedup in tokens per second over one server?