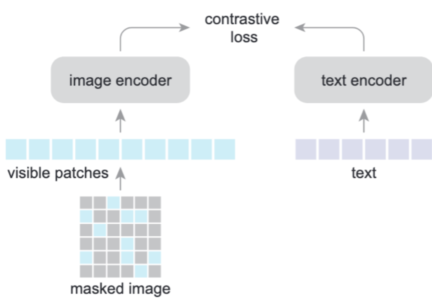
FLIP

A Preprint

Authors introduce FLIP, which is a simple method for efficient CLIP training. Idea behind it is to randomly remove patches/mask. FLIP trains >3× faster in wall-clock time for reaching similar accuracy as its CLIP counterpart; with the same number of epochs, FLIP reaches higher accuracy than its CLIP counterpart. The whole idea was inspired by MAE(Masked Autoencoder) which sparselly applies ViT encoder to visible content.



The intuition behind method is reducing computation so throughout training more image-text pairs can be used under the same time and also have a contrastive objective over a larger batch under the same memory constraint. As for Image masking authors adapt ViT; they mask 50-75 % of image and feed it into ViT encoder(by it meant the visible part). It is also possible to encode text the same way as was presented with image masking. however, as the text encoder is smaller, speeding it up does not lead to a better overall trade-off.

Unmasking: The simplest strategy is using the the same encoder pre-trained on masked images with just one simple setting: masking ratio 0 %

The training is performed on LAION-400M and evaluate zeroshot accuracy on ImageNet-1K validation. As results show, batchsize has a major impact on accuracy:

mask	batch	FLOPs	time	acc.
0%	16k	1.00×	1.00×	68.6
50%	32k	0.52×	0.50×	**69.6**
75%	64k	0.28×	0.33×	68.2

(a) **Image masking** yields higher or comparable accuracy and speeds up training. Entries are subject to the same memory limit.

batch	mask 50%	mask 75%
16k	68.5	65.8
32k	69.6	67.3
64k	**70.4**	**68.2**

(b) **Batch size**. A large batch has big gains over smaller batches.

text mask	text len	time	acc.
baseline, 0%	32	1.00×	**68.2**
random, 50%	16	0.92×	66.0
prioritized, 50%	16	0.92×	67.8

(c) **Text masking** performs decently, but the speed gain is marginal as its encoder is smaller. Here the image masking ratio is 75%.

	mask 50%	mask 75%
w/ mask	66.4	60.9
w/ mask, ensemble	68.1	65.1
w/o mask	**69.6**	**68.2**

(d) **Inference unmasking**. Inference on intact images performs strongly even without tuning.

	mask 50%	mask 75%
baseline	69.6	68.2
+ tuning	**70.1**	**69.5**

(e) **Unmasked tuning**. The distribution shift by masking is reduced by a short tuning.

	mask 50%	mask 75%
baseline	**69.6**	**68.2**
+ MAE	69.4	67.9

(f) **Reconstruction**. Adding the MAE reconstruction loss has no gain.

Table 1. **Zero-shot ablation experiments**. Pre-training is on LAION-400M for 6.4 epochs, evaluated by zero-shot classification accuracy

FLIP and CLIP both use the same contrastive loss:

CLIP uses a symmetric contrastive loss based on InfoNCE, defined as:

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[- \log \frac{\exp(\mathrm{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(I_i, T_j)/\tau)} - \log \frac{\exp(\mathrm{sim}(T_i, I_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(T_i, I_j)/\tau)} \right]$$

where:

- N — batch size;
- I_i — image embedding;
- T_i — text embedding;
- $\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$ — cosine similarity between vectors;
- τ — learnable temperature parameter.