Meta AI

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Scaling Language-Image Pre-training via Masking

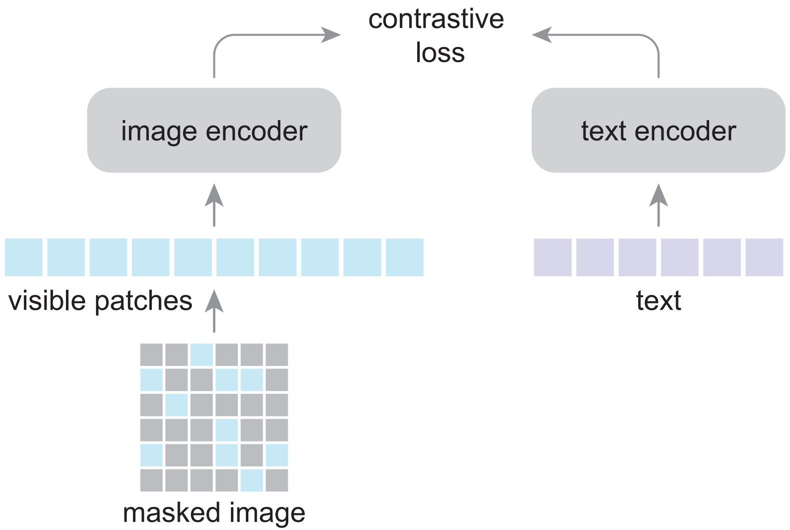
Yanghao Li*, Haoqi Fan*, Ronghang Hu*, Christoph Feichtenhofer†, Kaiming He†
Meta AI, FAIR

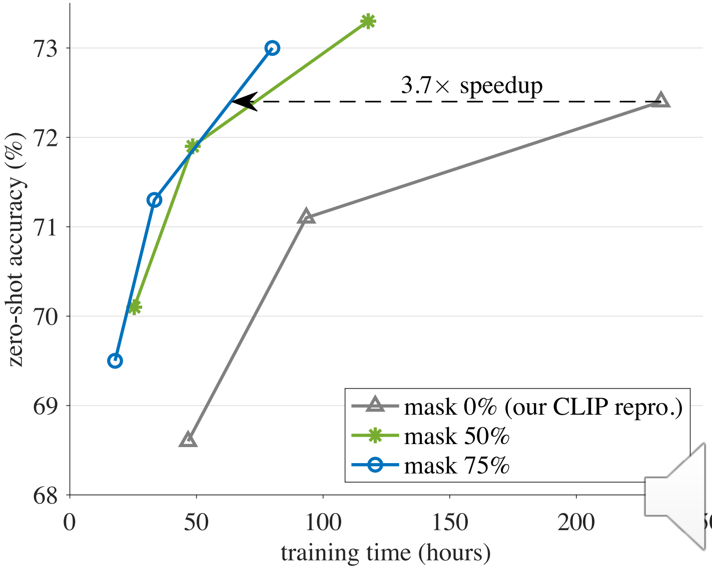
Paper Tag: **THU-PM-266**
Poster date: June 22, 2023

https://github.com/facebookresearch/flip

Fast Language-Image Pre-training (FLIP)

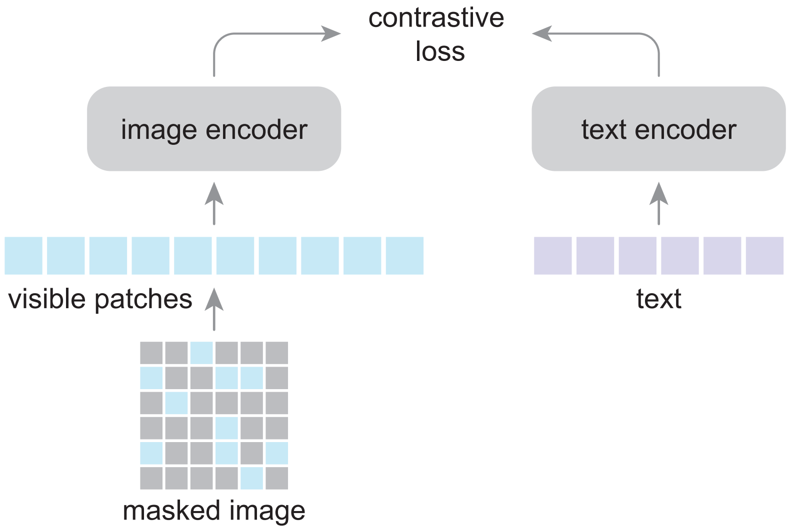
- A simple method for ***efficient*** CLIP training via ***Masking***
  - Randomly masking out image patches with a <u>*high masking ratio*</u>





FLIP Overview

- Benefits from masking
  - ***See more sample pairs*** under the same wall-clock training time
  - ***Contrast more sample pairs*** by larger batches under similar memory constraint



Properties of FLIP – Image Masking

- **Image masking** yields higher or comparable accuracy and speeds up training

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

Properties of FLIP – Batch Size

- **A large batch** has big gains over smaller batches

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

Properties of FLIP – Unmasked tuning

- A short tuning (0.32 epoch) greatly reduce distribution gap

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

FLIP Results

- Zero-shot ImageNet accuracy

case	data	epochs	B/16	L/16	L/14	H/14
CLIP [52]	WIT-400M	32	68.6	-	75.3	-
OpenCLIP [36]	LAION-400M	32	67.1	-	72.8	-
CLIP, our repro.	LAION-400M	32	68.2	72.4	73.1	-
FLIP	LAION-400M	32	68.0	74.3	74.6	75.5

For ViT-L/14, FLIP is *better* than both OpenCLIP and our reproduced CLIP pre-trained on the *same* data

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *ICML, 2021*
Ilharco, Gabriel, et al. OpenCLIP." *2021*

FLIP Results

- Linear-probing and fine-tuning on ImageNet

case	data	epochs	model	zero-shot	linear probe	fine-tune
CLIP [52]	WIT-400M	32	L/14	75.3	83.9†	-
CLIP [52], our transfer	WIT-400M	32	L/14	75.3	83.0	87.4
OpenCLIP [36]	LAION-400M	32	L/14	72.8	82.1	86.2
CLIP, our repro.	LAION-400M	32	L/16	72.4	82.6	86.3
FLIP	LAION-400M	32	L/16	74.3	83.6	86.9

FLIP outperforms OpenCLIP and CLIP counterparts
pre-trained on the same data

FLIP Results

- FLIP performs better on **zero-shot image/text retrieval**

			text retrieval						image retrieval					
			Flickr30k			COCO			Flickr30k			COCO		
case	model	data	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [52]	L/14@336	WIT-400M	88.0	98.7	99.4	58.4	81.5	88.1	68.7	90.6	95.2	37.8	62.4	72.2
CLIP [52], our eval.	L/14@336	WIT-400M	88.9	98.7	99.9	58.7	80.4	87.9	72.5	91.7	95.2	38.5	62.8	72.5
CLIP [52], our eval.	L/14	WIT-400M	87.8	99.1	99.8	56.2	79.8	86.4	69.3	90.2	94.0	35.8	60.7	70.7
OpenCLIP [36], our eval.	L/14	LAION-400M	87.3	97.9	99.1	58.0	80.6	88.1	72.0	90.8	95.0	41.3	66.6	76.1
CLIP, our impl.	L/14	LAION-400M	87.4	98.4	99.5	59.1	82.5	89.4	74.4	92.2	95.5	43.2	68.5	77.5
FLIP	L/14	LAION-400M	89.1	98.5	99.6	60.2	82.6	89.9	75.4	92.5	95.9	44.2	69.2	78.4

FLIP Results

- FLIP performs better on **image captioning** and **visual question answering**

			COCO caption					nocaps		VQAv2
case	model	data	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	CIDEr	SPICE	acc.
CLIP [52], our transfer	L/14	WIT-400M	37.5	29.6	58.7	126.9	22.8	82.5	12.1	76.6
OpenCLIP [36], our transfer	L/14	LAION-400M	36.7	29.3	58.4	125.0	22.7	83.4	12.3	74.5
CLIP, our repro.	L/16	LAION-400M	36.4	29.3	58.4	125.6	22.8	82.8	12.2	74.5
FLIP	L/16	LAION-400M	37.4	29.5	58.8	127.7	23.0	85.9	12.4	74.7

Scaling Behavior of FLIP

- The speed-up of FLIP facilitates scaling explorations

				zero-shot transfer					transfer learning				
				zero-shot	text retrieval		image retrieval		lin-probe	fine-tune	captioning		vqa
case	model	data	sampled	IN-1K	Flickr30k	COCO	Flickr30k	COCO	IN-1K	IN-1K	COCO	nocaps	VQAv2
baseline	Large	400M	12.8B	74.3	88.4	59.8	75.0	44.1	83.6	86.9	127.7	85.9	74.7
model scaling	**Huge**	400M	12.8B	75.5	89.2	62.8	76.4	46.0	84.3	87.3	130.3	91.5	76.3
data scaling	Large	**2B**	12.8B	75.8	91.7	63.8	78.2	47.3	84.2	87.1	128.9	87.0	75.5
schedule scaling	Large	400M	**25.6B**	73.9	89.7	60.1	75.5	44.4	83.7	86.9	127.9	86.8	75.0

Model and **data scaling** consistently **outperform** baselines

- *Data scaling* is favored for ***zero-shot transfer***
- *Model scaling* is favored for ***transfer learning***

Model scaling: ViT-L to ViT-H (~2x params)

Data scaling: LAION-400M to LAION-2B (image-text pairs)

Schedule scaling: 12.8B sampled data to 25.6B

Scaling Behavior of FLIP

- The speed-up of FLIP facilitates scaling explorations

				zero-shot transfer					transfer learning				
				zero-shot	text retrieval		image retrieval		lin-probe	fine-tune	captioning		vqa
case	model	data	sampled	IN-1K	Flickr30k	COCO	Flickr30k	COCO	IN-1K	IN-1K	COCO	nocaps	VQAv2
baseline	Large	400M	12.8B	74.3	88.4	59.8	75.0	44.1	83.6	86.9	127.7	85.9	74.7
model scaling	**Huge**	400M	12.8B	75.5	89.2	62.8	76.4	46.0	84.3	87.3	130.3	91.5	76.3
data scaling	Large	**2B**	12.8B	75.8	91.7	63.8	78.2	47.3	84.2	87.1	128.9	87.0	75.5
schedule scaling	Large	400M	**25.6B**	73.9	89.7	60.1	75.5	44.4	83.7	86.9	127.9	86.8	75.0
model+data scaling	**Huge**	**2B**	12.8B	77.6	92.8	67.0	79.9	49.5	85.1	87.7	**130.4**	**92.6**	77.1
joint scaling	**Huge**	**2B**	**25.6B**	**78.8**	**93.1**	**67.8**	**80.9**	**50.5**	**85.6**	**87.9**	130.2	91.2	**77.3**

- Model and data scaling are **highly complementary**
  - Scaling both (+3.3%) > model + data scaling alone (+1.2% + 1.5%)
- **Joint scaling** with schedule scaling leads to the best in most cases

Scaling Language-Image Pre-training via Masking

Yanghao Li[*], Haoqi Fan[*], Ronghang Hu[*], Christoph Feichtenhofer[†], Kaiming He[†]
Meta AI, FAIR

Paper Tag: **THU-PM-266**
Poster date: June 22, 2023

https://github.com/facebookresearch/flip