VLP models for vision

IFT 6765 2023
Mini-Lecture presented by Diganta Misra

2023 Mar 14

Content

Mila

01

Learning Transferable Visual Models From Natural Language Supervision

Radford et. al.
OpenAI

Overview

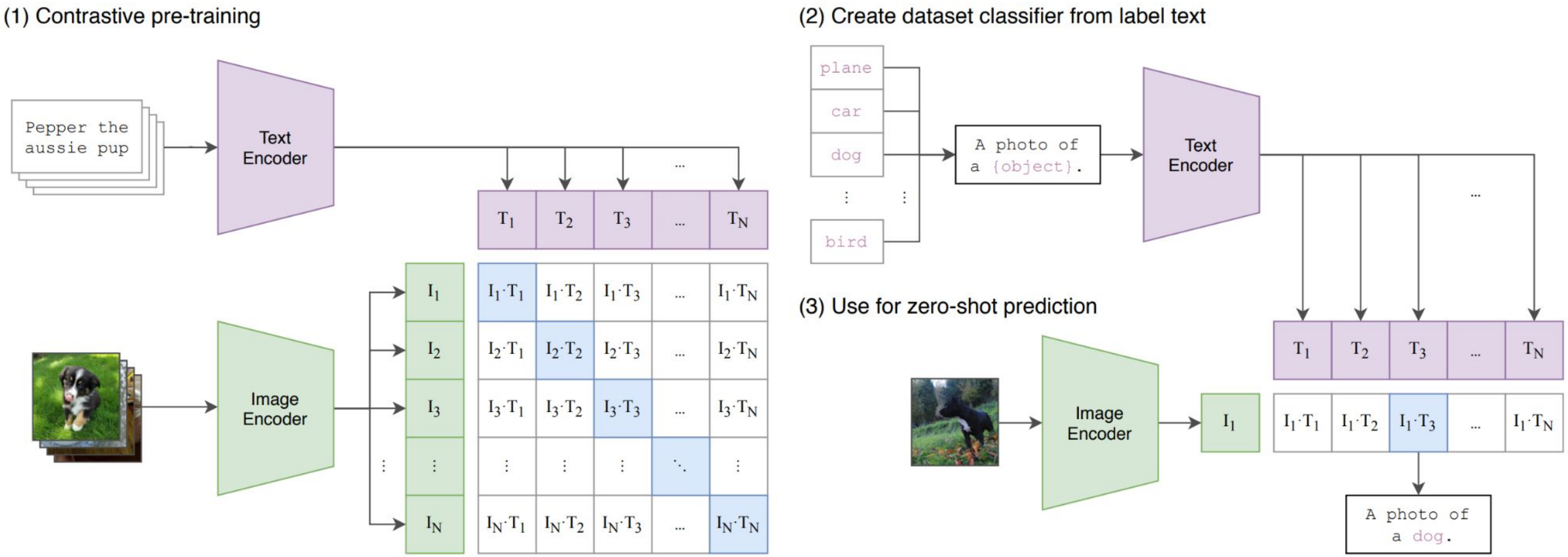


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Findings

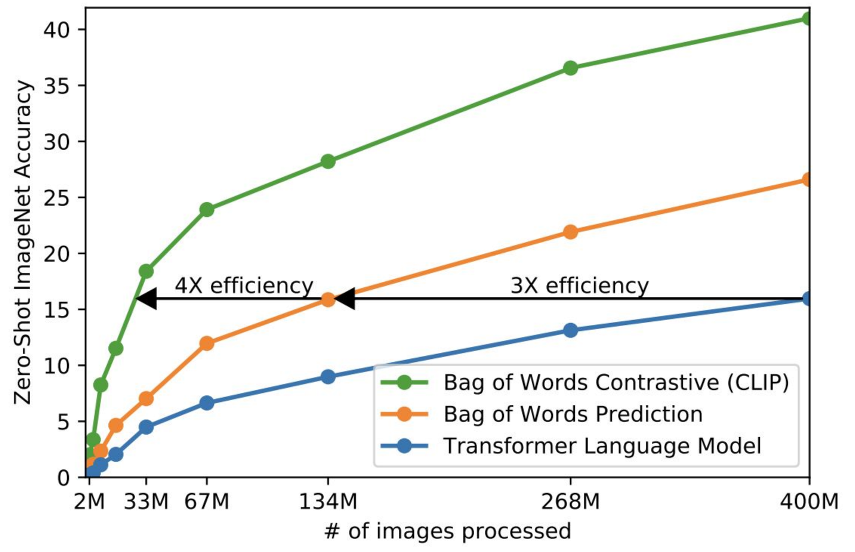


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

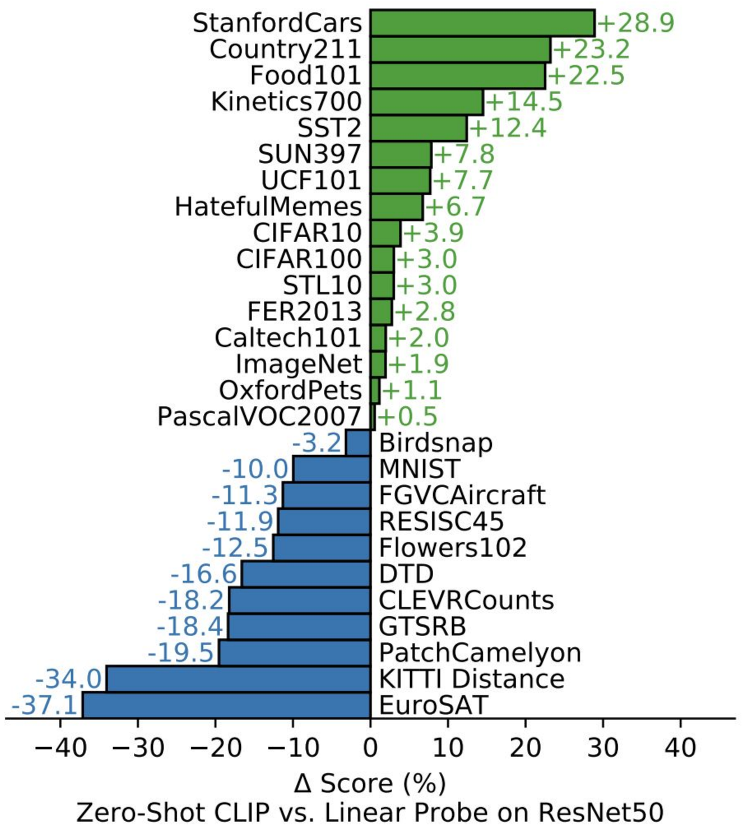


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

Mila

Findings

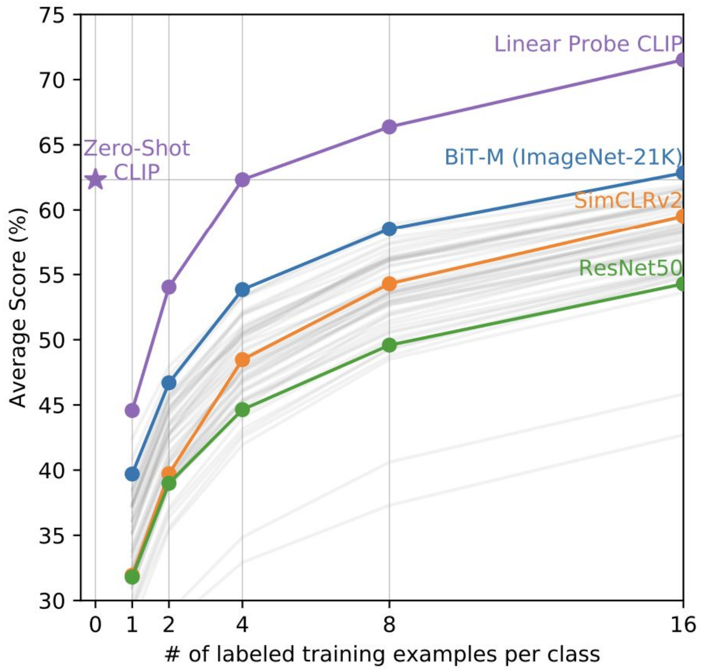


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

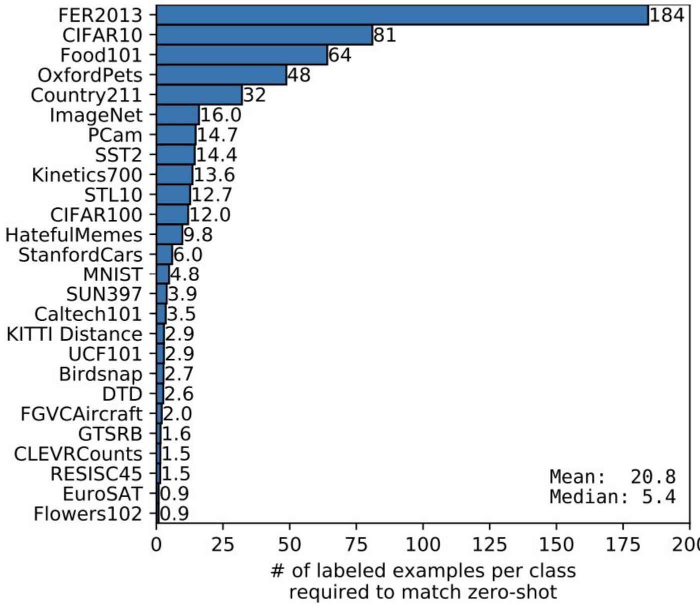


Figure 7. **The data efficiency of zero-shot transfer varies widely.** Calculating the number of labeled examples per class a linear classifier on the same CLIP feature space requires to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer. Values are estimated based on log-linear interpolation of 1, 2, 4, 8, 16-shot and fully supervised results. Performance varies widely from still underperforming a one-shot classifier on two datasets to matching an estimated 184 labeled examples per class.

Scaling

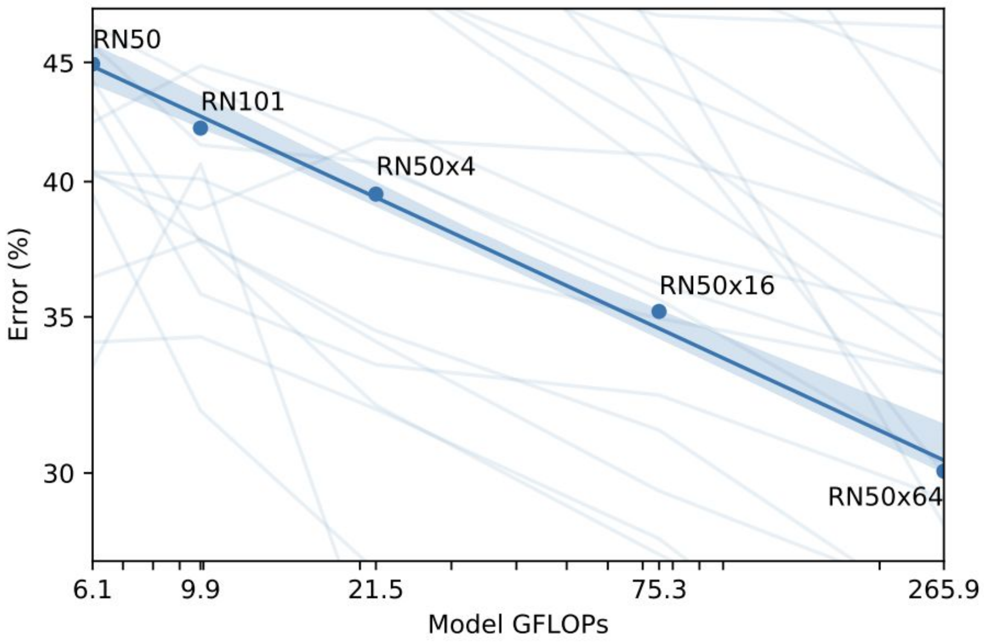


Figure 9. **Zero-shot CLIP performance scales smoothly as a function of model compute.** Across 39 evals on 36 different datasets, average zero-shot error is well modeled by a log-log linear trend across a 44x range of compute spanning 5 different CLIP models. Lightly shaded lines are performance on individual evals, showing that performance is much more varied despite the smooth overall trend.

Mila

Robustness

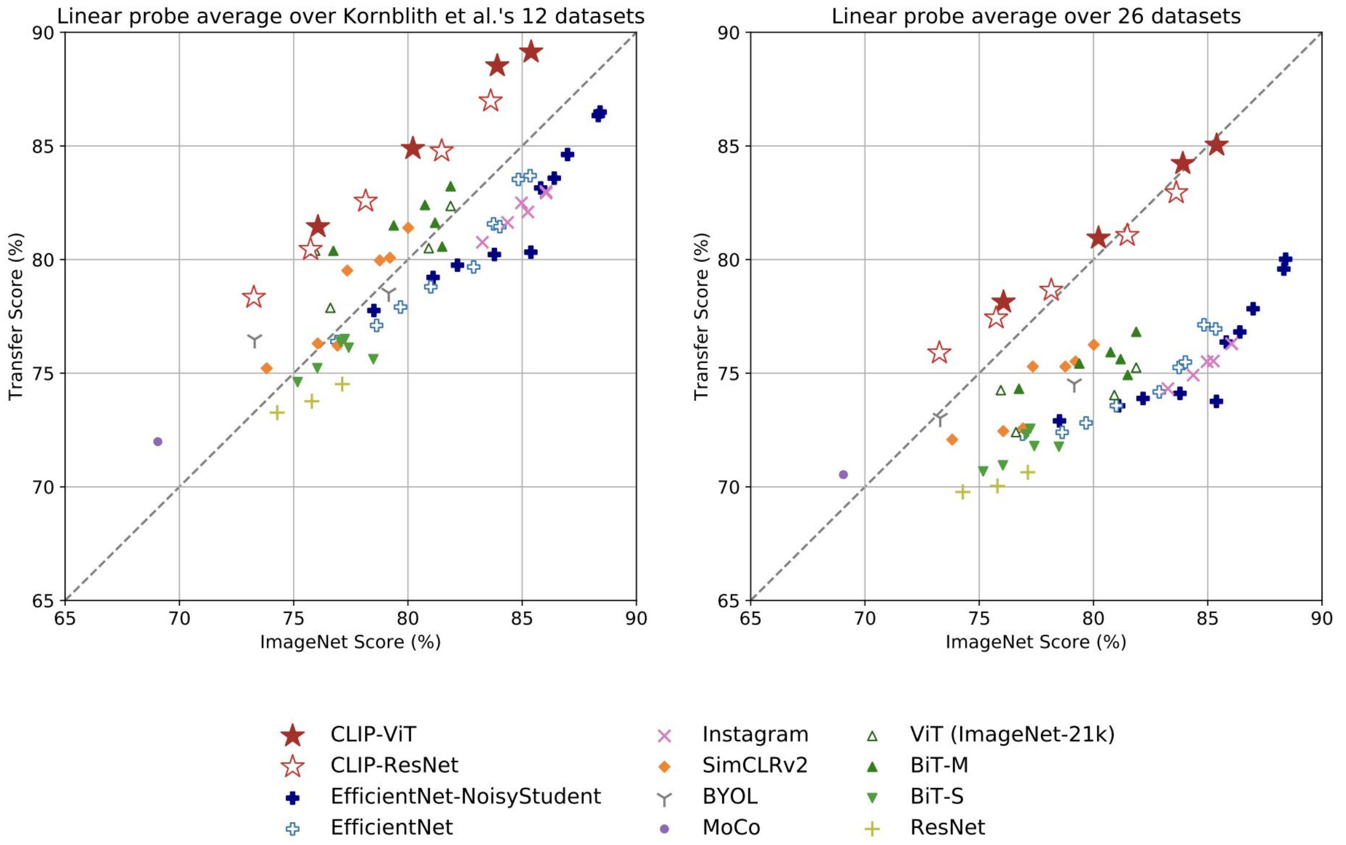
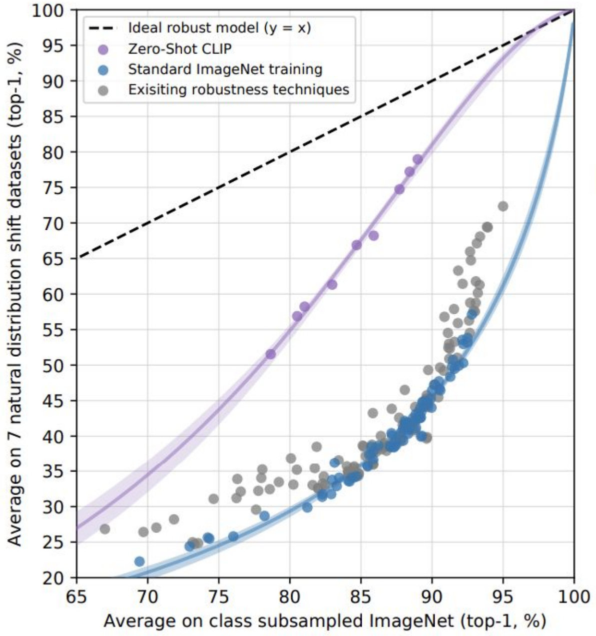
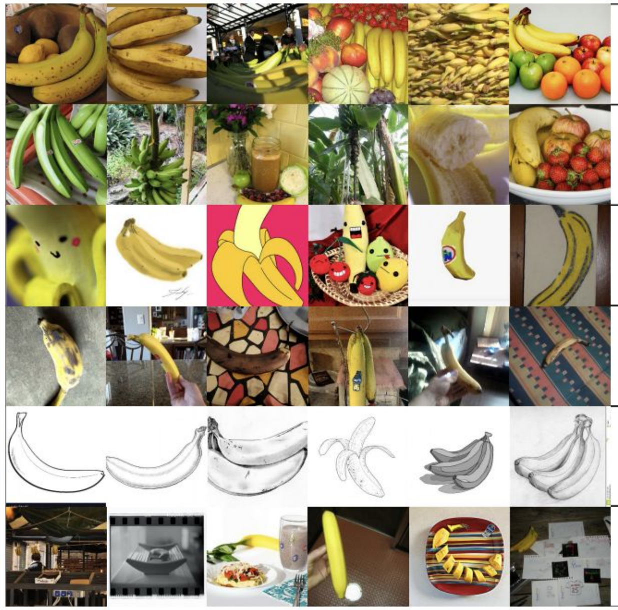


Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

Robustness





	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

Mila

Worst Subgroup Generalisation

Model	Race	Gender	Age
FairFace Model	**93.7**	94.2	59.7
Linear Probe CLIP	93.4	**96.5**	**63.8**
Zero-Shot CLIP	58.3	95.9	57.1
Linear Probe Instagram	90.8	93.2	54.2

Table 3. Percent accuracy on Race, Gender, and Age classification of images in FairFace category 'White'

Model	Race	Gender	Age
FairFace Model	75.4	94.4	60.7
Linear Probe CLIP	**92.8**	**97.7**	**63.1**
Zero-Shot CLIP	91.3	97.2	54.3
Linear Probe Instagram	87.2	93.9	54.1

Table 4. Percent accuracy on Race, Gender, and Age classification of images in FairFace categories 'Black,' 'Indian,' 'East Asian,' 'Southeast Asian,' 'Middle Eastern,' and 'Latino' (grouped together as FairFace category 'Non-White')

Model	Gender	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian	Average
Linear Probe CLIP	Male	96.9	96.4	98.7	96.5	98.9	96.2	96.9	97.2
	Female	97.9	96.7	97.9	99.2	97.2	98.5	97.3	97.8
		97.4	96.5	98.3	97.8	98.4	97.3	97.1	97.5
Zero-Shot CLIP	Male	96.3	96.4	97.7	97.2	98.3	95.5	96.8	96.9
	Female	97.1	95.3	98.3	97.8	97.5	97.2	96.4	97.0
		96.7	95.9	98.0	97.5	98.0	96.3	96.6	
Linear Probe Instagram	Male	92.5	94.8	96.2	93.1	96.0	92.7	93.4	94.1
	Female	90.1	91.4	95.0	94.8	95.0	94.1	94.3	93.4
		91.3	93.2	95.6	94.0	95.6	93.4	93.9	

Table 5. Percent accuracy on gender classification of images by FairFace race category

Mila

02

Image-and-Language Understanding from Pixels Only

Tschannen et al.
Google AI

Overview

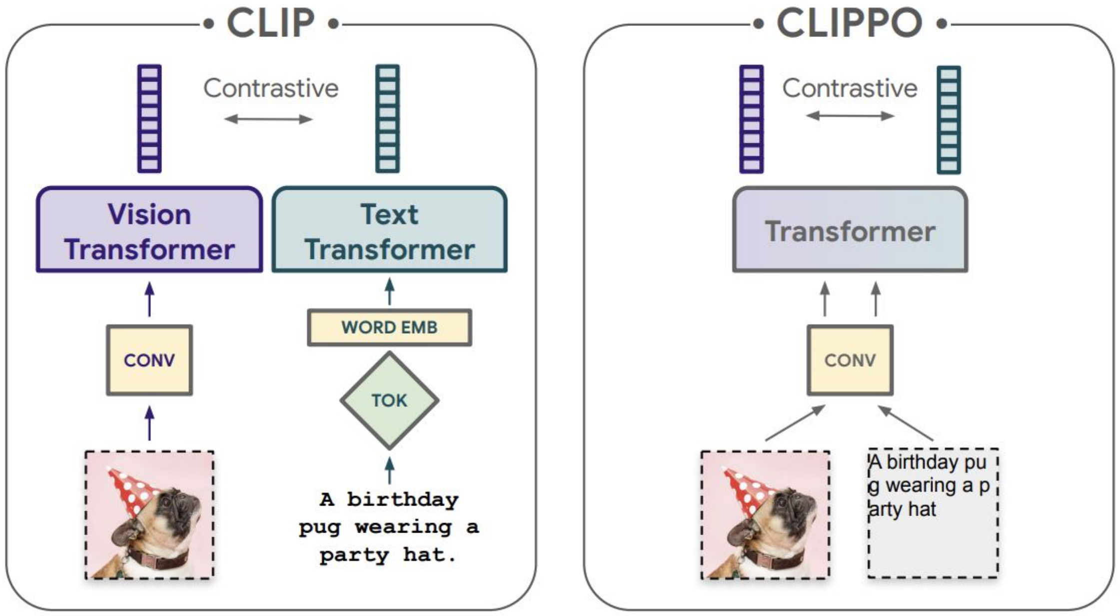


Figure 1. CLIP [50] trains separate image and text encoders, each with a modality-specific preprocessing and embedding, on image/alt-text pairs with a contrastive objective. CLIPPO trains a pure pixel-based model with equivalent capabilities by rendering the alt-text as an image, encoding the resulting image pair using a shared vision encoder (in two separate forward passes), and applying same training objective as CLIP.

Findings

	#param.	training dataset	I1k 10s.	I1k 0s.	C I→T	C T→I	F I→T	F T→I
CLIP*	203M	WebLI	55.8	65.1	48.5	31.3	79.2	59.4
1T-CLIP	118M	WebLI	53.9	62.3	48.0	30.3	77.5	58.2
CLIPPO	93M	WebLI	53.0	61.4	47.3	30.1	76.4	57.3
CLIPPO	93M	WebLI + 25%C4	52.1	57.4	40.7	26.7	68.9	51.8
CLIPPO	93M	WebLI + 50%C4	48.0	53.1	35.2	23.4	64.8	47.2
1T-CLIP L/16	349M	WebLI	60.8	67.8	50.7	32.5	81.0	61.0
CLIPPO L/16	316M	WebLI	60.3	67.4	50.6	33.4	79.2	62.6
CLIPPO L/16	316M	WebLI + 25%C4	60.5	66.0	44.5	29.8	72.9	57.3
CLIPPO L/16	316M	WebLI + 50%C4	56.8	61.7	39.7	27.3	70.1	54.7

Table 1. Vision and vision-language cross-modal results. We report ImageNet-1k 10-shot linear transfer validation accuracy (I1k 10s.), ImageNet-1k zero-shot transfer validation accuracy (I1k 0s.), image-to-text and text-to-image retrieval recall@1 on MS-COCO (C I→T and C T→I) and on Flickr30k (F T→I and F I→T). CLIPPO and 1T-CLIP incur a minor drop in these evaluations compared to CLIP*, while only using about half of the model parameters. Co-training with text pairs from C4 (models with + xx%C4) degrades performance on some cross-modal tasks (but leads to improved language understanding capabilities, see Table 2).

Findings

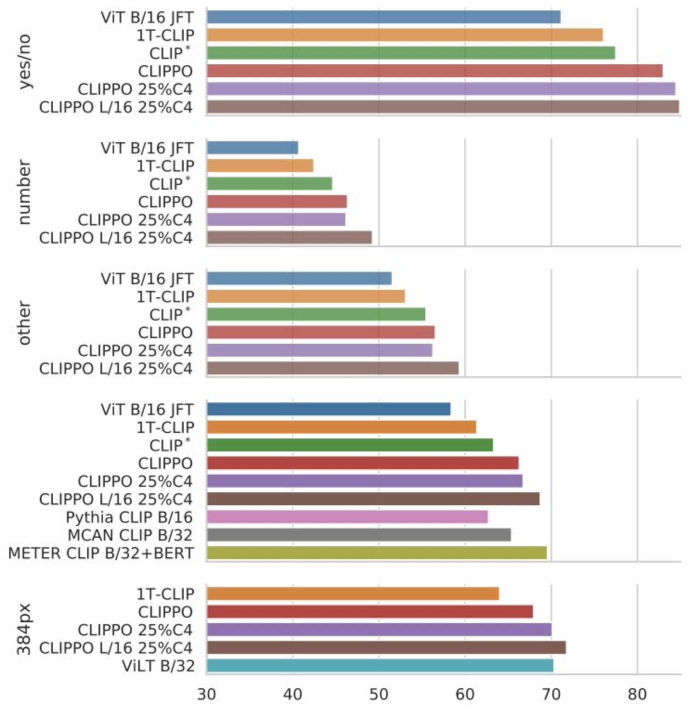


Figure 2. Results on the VQAv2 benchmark (test-dev set). In addition to CLIPPO and baselines produced in this work, we also compare to Pythia and MCAN models with ViT encoders from [61], and with comparably sized METER [16] and ViLT [34] models. CLIPPO outperforms CLIP* and 1T-CLIP clearly on "yes/no" questions and gets similar performance as task-specific models.

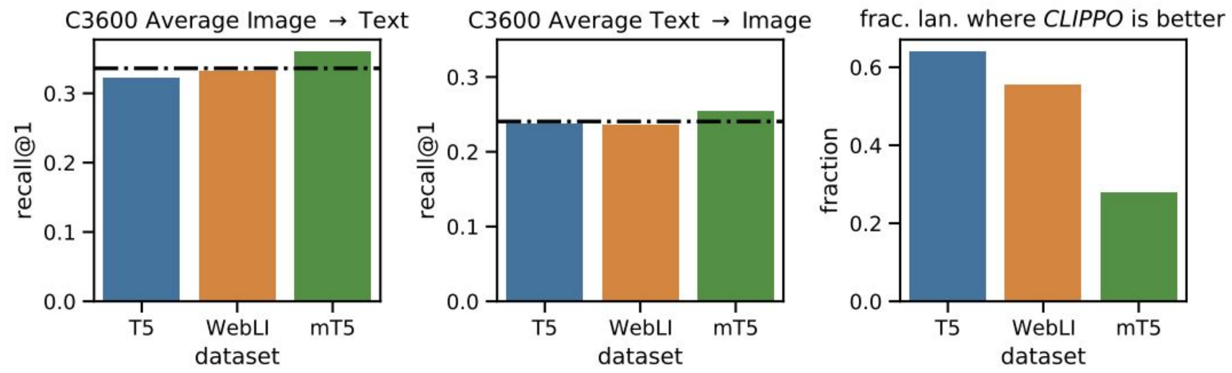


Figure 3. Zero-shot image/text retrieval performance on Cross-Modal3600 [64]. Although specialized (mc4) tokenizers can be leveraged to improve multilingual performance CLIPPO (dashed black line) broadly matches or exceeds comparable 1T-CLIP models trained with vocabulary size 32,000 (the word embeddings result in a 27% increase in parameter count compared to CLIPPO).

Results

	training dataset	MNLI-M/MM	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	avg
BERT-Base	Wiki + BC	84.0 / 84.0	87.6	91.0	92.6	60.3	88.8	90.2	69.5	83.1
PIXEL	Wiki + BC	78.1 / 78.1	84.5	87.8	89.6	38.4	81.1	88.2	60.5	76.3
BiLSTM		66.7 / 66.7	82.0	77.0	87.5	17.6	72.0	85.1	58.5	68.1
BiLSTM+Attn, ELMo		72.4 / 72.4	83.6	75.2	91.5	44.1	56.1	82.1	52.7	70.0
CLIP* img enc.	WebLI	66.4 / 66.4	78.6	69.4	78.6	0.0	5.2	81.2	52.7	55.5
CLIP* text enc.	WebLI	71.8 / 71.8	82.7	73.0	86.2	6.6	65.0	81.4	53.8	65.9
1T-CLIP text enc.	WebLI	72.6 / 72.6	83.8	80.7	84.9	0.0	79.6	83.3	57.0	68.3
CLIPPO	WebLI	73.0 / 73.0	84.3	81.2	86.8	1.8	80.5	84.1	53.4	68.6
CLIPPO	WebLI + 25%C4	77.7 / 77.7	85.3	83.1	90.9	28.2	83.4	84.5	59.2	74.4
CLIPPO	WebLI + 50%C4	79.2 / 79.2	86.4	84.2	92.9	38.9	83.4	84.8	59.9	76.6
CLIPPO	C4	79.9 / 79.9	86.7	85.2	93.3	50.9	84.7	86.3	58.5	78.4
CLIPPO L/16	WebLI + 25%C4	76.6 / 76.6	87.1	79.9	93.2	48.2	84.1	84.6	56.0	76.1
CLIPPO L/16	WebLI + 50%C4	82.3 / 82.3	87.9	86.7	94.2	55.3	85.8	85.9	59.2	80.0

Table 2. Results for the GLUE benchmark (dev set). The metric is accuracy except for the performance on QQP and MRPC, which is measured using the F_1 score, CoLA which uses Matthew's correlation, and STS-B which evaluated based on Spearman's correlation coefficient. "avg" corresponds to the average across all metrics. The results for BERT-Base and PIXEL are from [54, Table 3], and BiLSTM and BiLSTM+Attn, ELMo from [66, Table 6]. All encoders considered here have a transformer architecture comparable to BERT-Base (up to the text embedding layer), except for CLIPPO L/16 which uses a ViT L/16, and the two BiLSTM model variants. Wiki and BC stand for (English) Wikipedia and Bookcorpus [78] data, respectively.

Results

Figure 7. Example training images with rendered questions (black letters on gray background) from the VQAv2 dataset (image size 224×224). After fine-tuning CLIPPO on VQAv2 it can process images and question jointly in this form. Note that the answers (on white background) are not part of the image.

Mila

Results

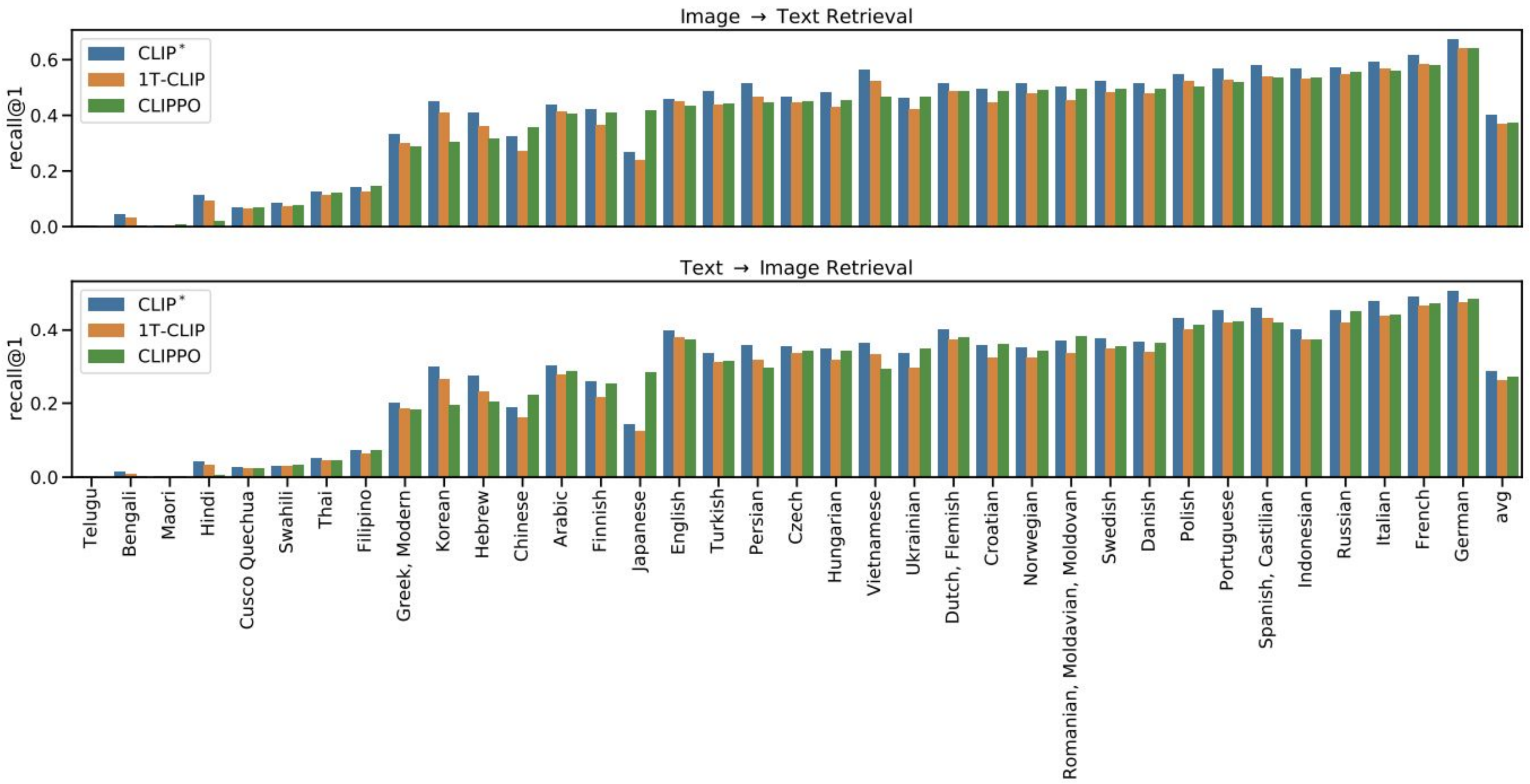


Figure 8. Per-language and average image-to-text and text-to-image recall@1 on the Crossmodal3600 data set. All the models are trained for 250k iterations on WebLI with multilingual alt-texts. CLIP* and 1T-CLIP use a SentecePiece tokenizer with vocabulary size 32,000 built from 300M randomly sampled WebLI alt-texts, whereas CLIPPO is tokenizer-free by design.

Modality Gap Analysis

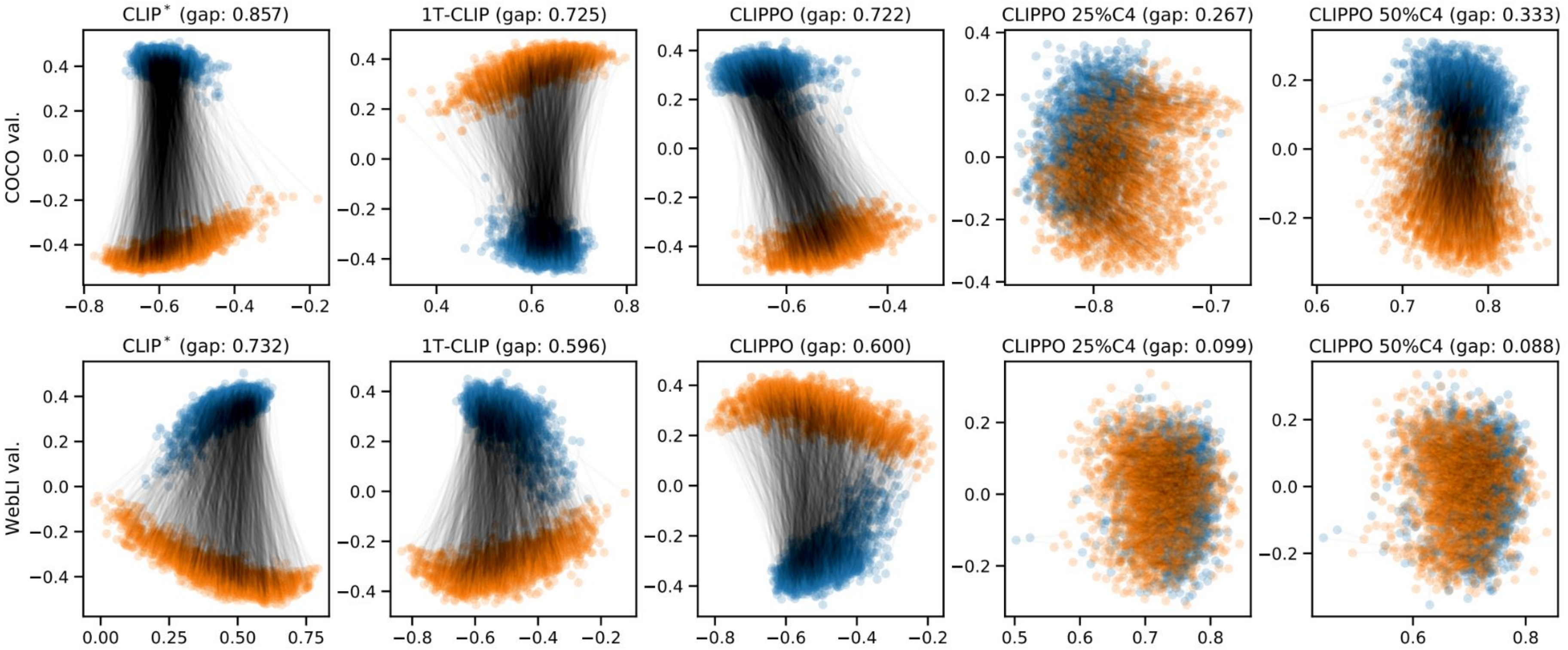


Figure 10. Visualization of the modality gap for examples from the WebLI and MS-COCO validation sets. The visualization follows the analysis from [41] and shows embedded images (blue dots) and corresponding alt-text (orange dots), projected to the first two principal components of the validation data matrix.

03

Scaling Language-Image Pre-training via Masking

Li et al.
Meta AI

Overview

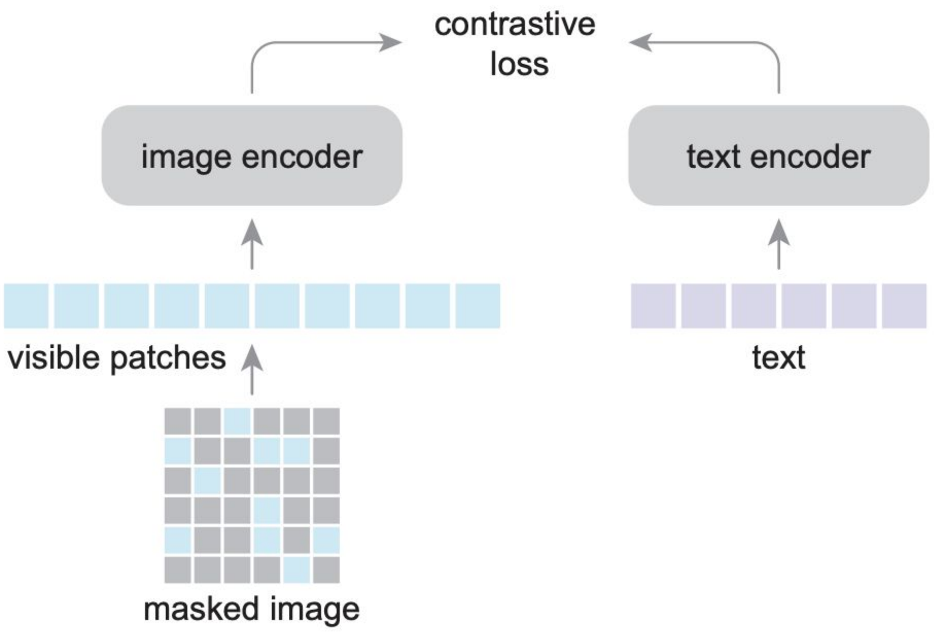


Figure 2. **Our FLIP architecture**. Following CLIP [52], we perform contrastive learning on pairs of image and text samples. We randomly mask out image patches with a high masking ratio and encode only the visible patches. We do not perform reconstruction of masked image content.

MAE

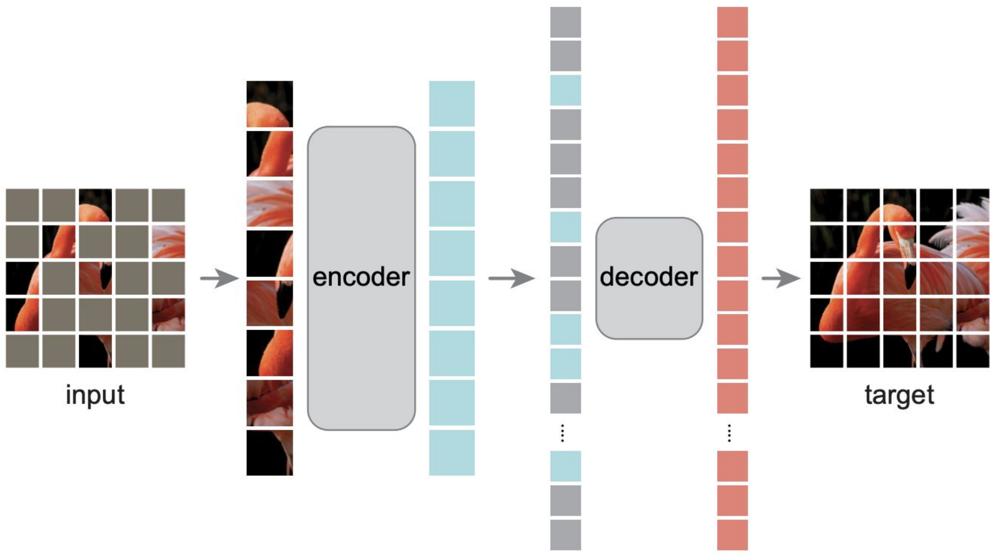


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

Table 2. **Wall-clock time** of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

Efficiency

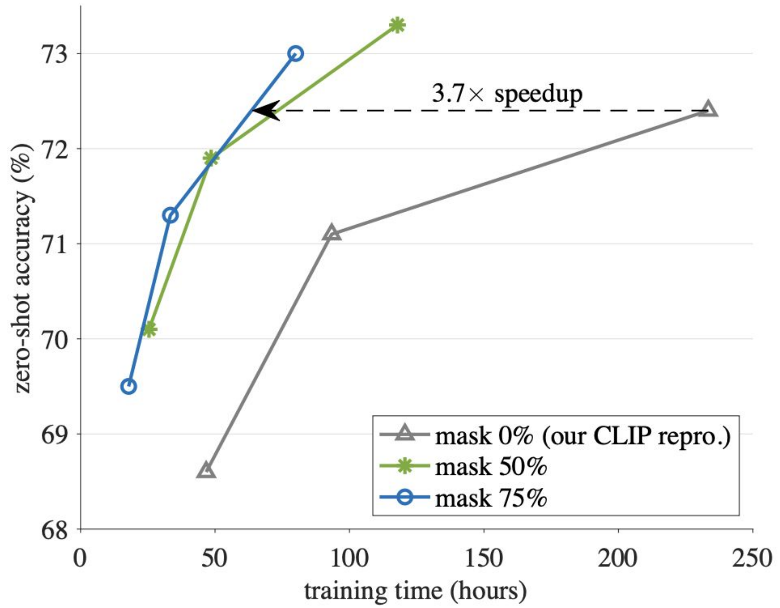


Figure 1. **Accuracy *vs.* training time trade-off.** With a high masking ratio of 50% or 75%, our FLIP method trains faster and is more accurate than its CLIP counterpart. All entries are benchmarked in 256 TPU-v3 cores. Training is done on LAION-400M for 6.4, 12.8, or 32 epochs, for each masking ratio. Accuracy is evaluated by zero-shot transfer on the ImageNet-1K validation set. The model is ViT-L/16 [20]. More details are in Fig. 3. As the CLIP baseline takes ~2,500 TPU-days training, a speedup of 3.7× can save ~1,800 TPU-days.

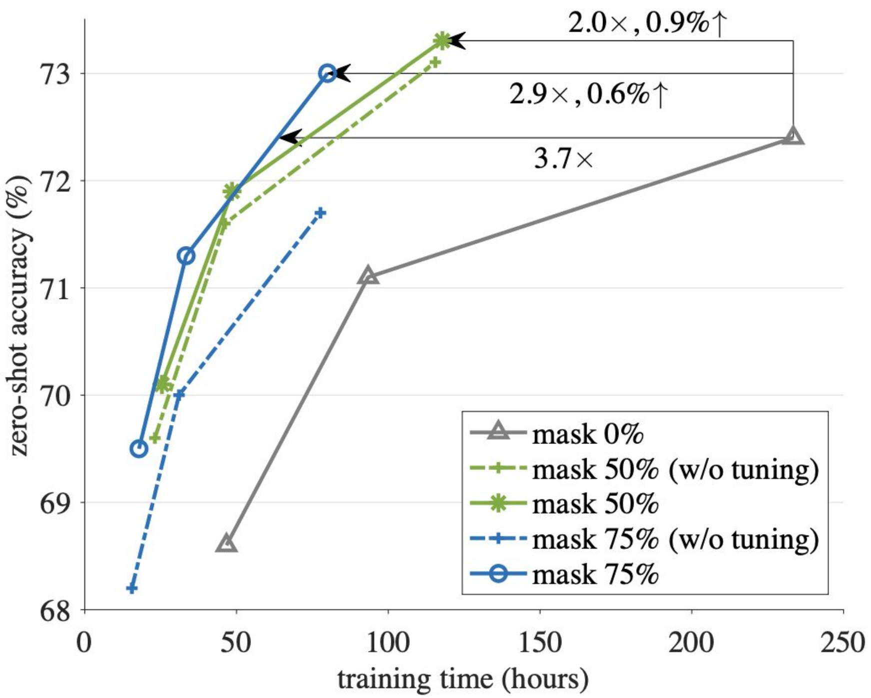


Figure 3. **Accuracy *vs.* training time trade-off in detail.** The setting follows Table 1a. Training is for 6.4, 12.8, or 32 epochs, for each masking ratio. Unmasked tuning, if applied, is for 0.32 epoch. All are benchmarked in 256 TPU-v3 cores. Zero-shot accuracy is on IN-1K validation. The model is ViT-L/16. Our method speeds up training and increases accuracy.

Mila

Results

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

Table 2. **Zero-shot accuracy on ImageNet-1K classification**, compared with various CLIP baselines. The image size is 224. The entries noted by grey are pre-trained on a different dataset. Our models use a 64k batch, 50% masking ratio, and unmasked tuning.

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

Table 3. **Linear probing and fine-tuning accuracy on ImageNet-1K classification**, compared with various CLIP baselines. The entries noted by grey are pre-trained on a different dataset. The image size is 224. †: CLIP in [52] optimizes with L-BFGS; we use SGD instead.

Results

	data	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	MNIST	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCam	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
CLIP [52]	WIT-400M	92.9	96.2	77.9	48.3	67.7	77.3	36.1	84.1	55.3	93.5	92.6	78.7	87.2	99.3	59.9	71.6	50.3	23.1	32.7	58.8	76.2	60.3	24.3	63.3	64.0
CLIP [52], our eval.	WIT-400M	91.0	95.2	75.6	51.2	66.6	75.0	32.3	83.3	55.0	93.6	92.4	77.7	76.0	99.3	62.0	71.6	51.6	26.9	30.9	51.6	76.1	59.5	22.2	55.3	67.3
OpenCLIP [36], our eval.	LAION-400M	87.4	94.1	77.1	61.3	70.7	86.2	21.8	83.5	54.9	90.8	94.0	72.1	71.5	98.2	53.3	67.7	47.3	29.3	21.6	51.1	71.3	50.5	22.0	55.3	57.1
CLIP, our repro.	LAION-400M	88.1	96.0	81.3	60.5	72.3	89.1	25.8	81.1	59.3	93.2	93.2	74.6	69.1	96.5	50.7	69.2	50.2	29.4	21.4	53.1	71.5	53.5	18.5	53.3	57.2
FLIP	LAION-400M	89.3	97.2	84.1	63.0	73.1	90.7	29.1	83.1	60.4	92.6	93.8	75.0	80.3	98.5	53.5	70.8	41.4	34.8	23.1	50.3	74.1	55.8	22.7	54.0	58.5

Table 4. **Zero-shot accuracy on more classification datasets**, compared with various CLIP baselines. This table follows Table 11 in [52]. The model is ViT-L/14 with an image size of 224, for all entries. Entries in green are the best ones using the LAION-400M data.

			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

Table 5. **Zero-shot image/text retrieval**, compared with various CLIP baselines. The image size is 224 if not noted. Entries in green are the best ones using the LAION-400M data.

Results

	model	data	IN-V2 top-1	IN-A top-1	IN-R top-1	ObjectNet top-1	IN-Sketch top-1	IN-Vid PM-0	IN-Vid PM-10	YTBB PM-0	YTBB PM-10
CLIP [52]	L/14@336	WIT-400M	70.1	77.2	88.9	72.3	60.2	95.3	89.2	95.2	88.5
CLIP [52], our eval.	L/14@336	WIT-400M	70.4	78.0	89.0	69.3	59.7	95.9	88.8	95.3	89.4
CLIP [52], our eval.	L/14	WIT-400M	69.5	71.9	86.8	68.6	58.5	94.6	87.0	94.1	86.4
OpenCLIP [36], our eval.	L/14	LAION-400M	64.0	48.3	84.3	58.8	56.9	90.3	81.4	86.5	77.8
CLIP, our repro.	L/14	LAION-400M	65.6	46.3	84.7	58.0	58.7	89.3	80.5	85.7	77.8
FLIP	L/14	LAION-400M	66.8	51.2	86.5	59.1	59.9	91.1	83.5	89.4	83.3

Table 6. **Zero-shot robustness evaluation,** compared with various CLIP baselines. This table follows Table 16 in [52]. The image size is 224 if not noted. Entries in green are the best ones using the LAION-400M data.

case	model	data	COCO caption BLEU-4	COCO caption METEOR	COCO caption ROUGE-L	COCO caption CIDEr	COCO caption SPICE	nocaps CIDEr	nocaps SPICE	VQAv2 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

Table 7. **Image Captioning and Visual Question Answering**, compared with various CLIP baselines. Entries in green are the best ones using the LAION-400M data. Here the results are on the COCO captioning test split of [38], nocaps val split, and VQAv2 test-dev split, respectively.

Scaling

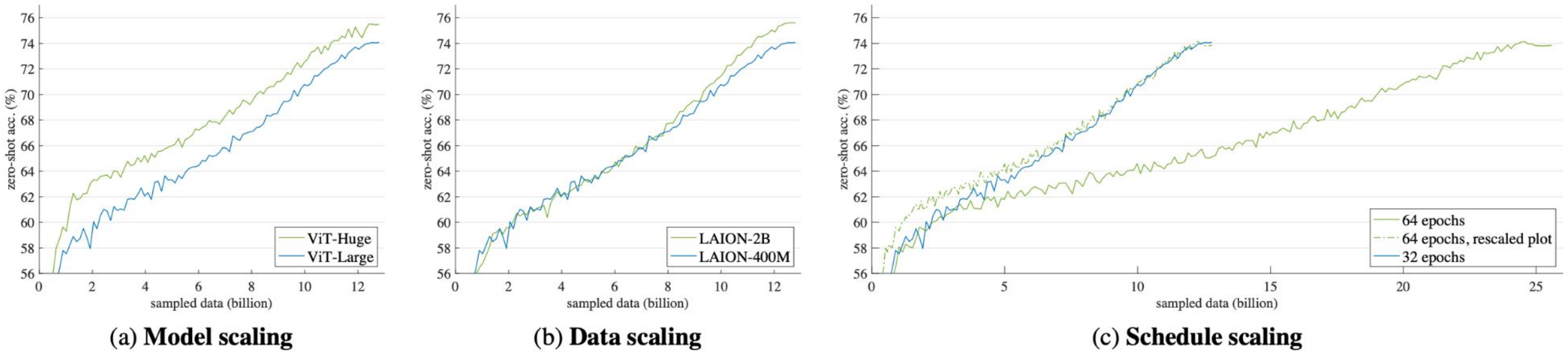


Figure 4. **Training curves of scaling**. The x-axis is the number of sampled data during training, and the y-axis is the zero-shot accuracy on IN-1K. The blue curve is the baseline setting: ViT-Large, LAION-400M data, 32 epochs (12.8B sampled data). In each subplot, we compare with scaling one factor on the baseline. In schedule scaling (Fig. 4c), we plot an extra hypothetical curve for a better visualization.

Mila

Scaling

case	model	data	sampled	zero-shot transfer: zero-shot IN-1K	zero-shot transfer: text retrieval Flickr30k	zero-shot transfer: text retrieval COCO	zero-shot transfer: image retrieval Flickr30k	zero-shot transfer: image retrieval COCO	transfer learning: lin-probe IN-1K	transfer learning: fine-tune IN-1K	transfer learning: captioning COCO	transfer learning: captioning nocaps	transfer learning: vqa 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**	**78.1**	92.1	66.8	79.3	49.2	85.0	87.5	130.1	91.1	76.9

Table 8. **Scaling behavior of FLIP**, evaluated on a diverse set of downstream tasks: classification, retrieval (R@1), captioning (CIDEr), and visual question answering. In the middle three rows, we scale along one of the three axes (model, data, schedule), and the green entries denote the best ones among these three scaling cases. Data scaling is in general favored under the zero-shot transfer scenario, while model scaling is in general favored under the transfer learning scenario (*i.e.*, with trainable weights in downstream).

04

MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning

Eichenberg et al.
Aleph Alpha & Heidelberg University

Overview

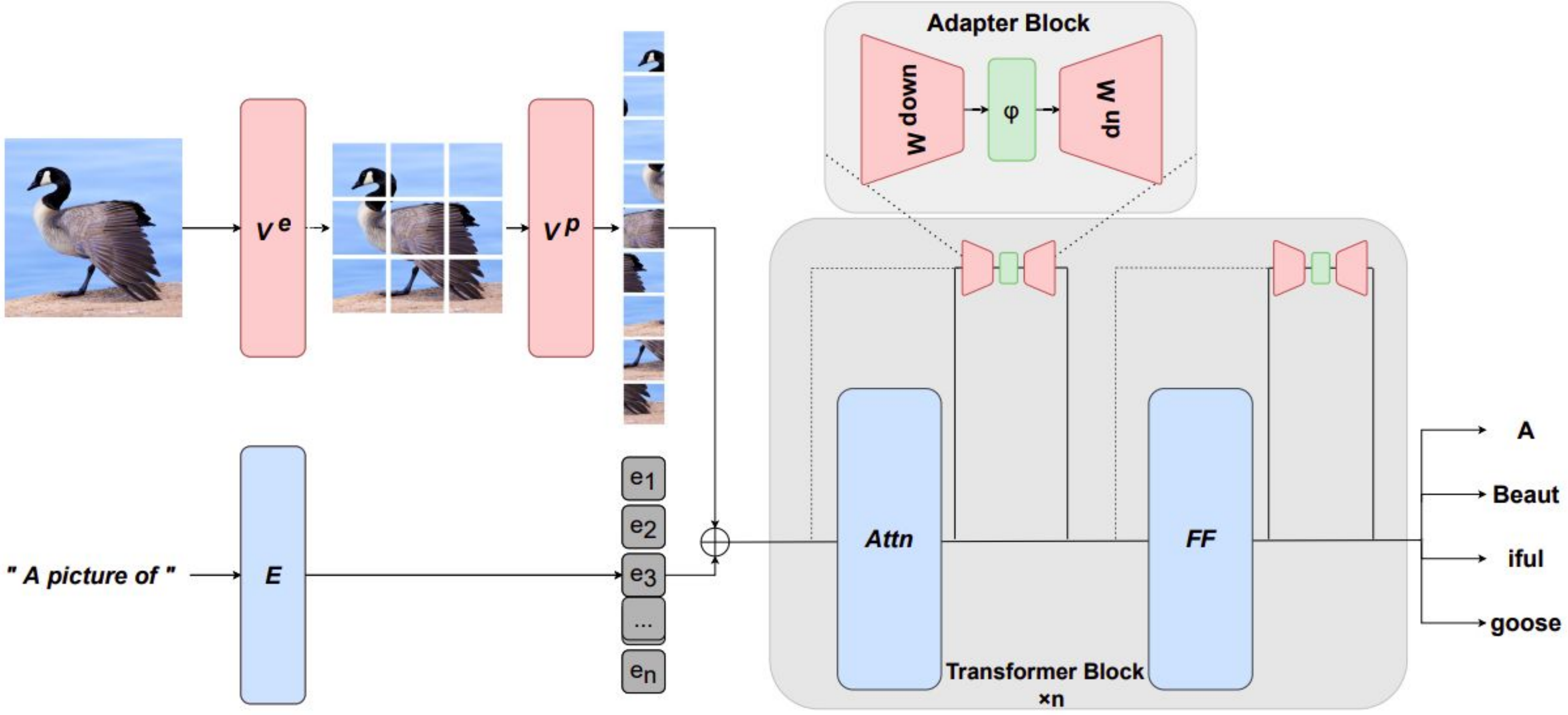
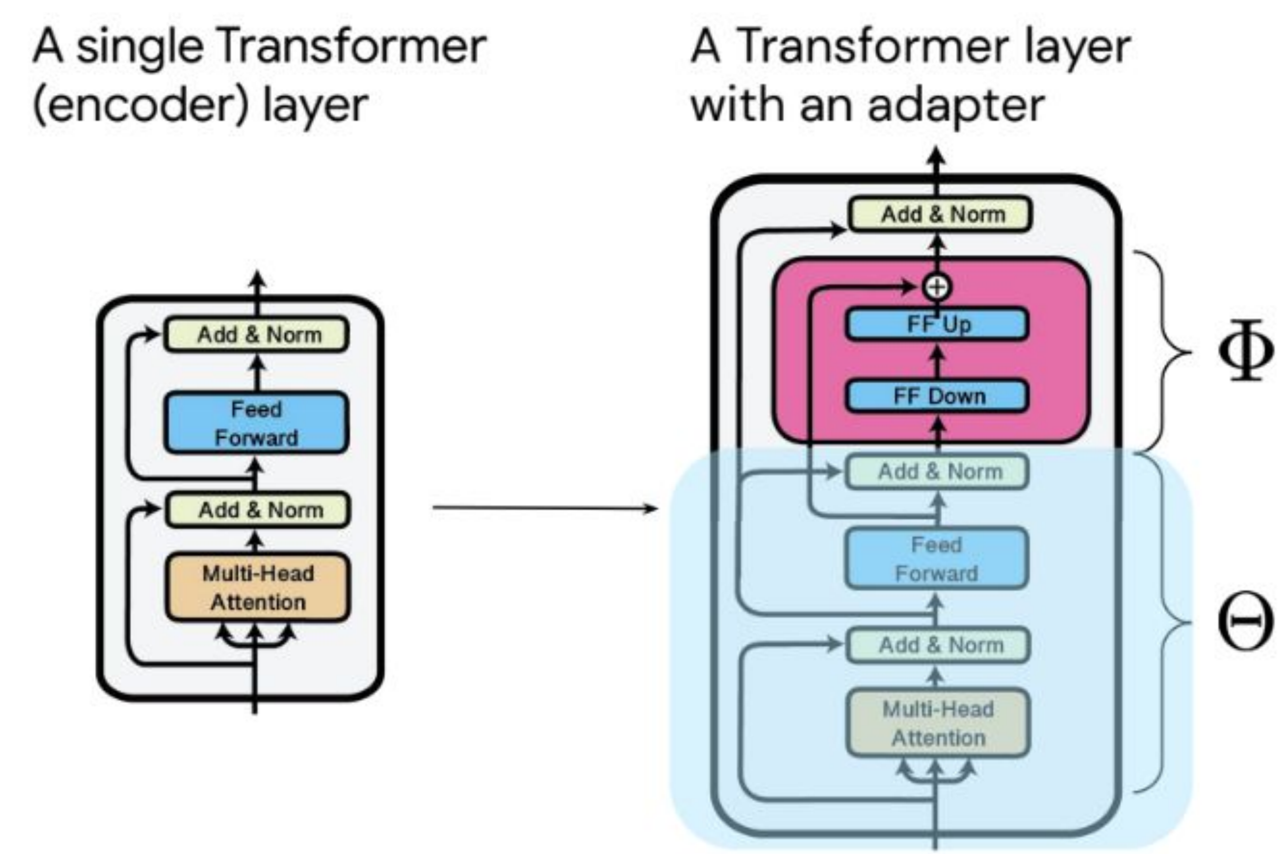


Figure 2: MAGMA's architecture. The layers in red are trained, and the layers in blue remain frozen.

Background: Adapters

- Allocate additional capacity for to enable transfer learning on incoming downstream tasks without training a new model for every task using adapters
- Small bottleneck layers inserted between a pre-trained model's weights
- Adapter parameters are **encapsulated** between transformer layers with parameters which are frozen



Findings

	VQA	OKVQA	GQA	VizWiz	SNLI-VE	NoCaps		Coco	
						CIDEr	B@4	CIDEr	B@4
MAGMA	68.0	**49.2**	54.5	35.4	79.0	93.6	27.8	91.2	31.4
SOTA	**75.5**	48.0	**72.1**	**54.7**	**86.3**	**112.2**	**33.1**	**143.3**	**41.7**
SOTA model	*SimVLM*	*PICa*	*CFR*	*Pythia*	*SimVLM*	*SimVLM*	*VIVO*	*SimVLM*	*OSCAR*

Table 2: MAGMA finetuned performance. **B@4**: NoCaps-all score. SOTA scores are to the best of our knowledge at the time of writing. If available/applicable, we compare to the SOTA score of models solving the task in an open-ended generative fashion like MAGMA (notably *SimVLM* on VQA), otherwise we compare to the general SOTA (classification setting). Models: *SimVLM* (Wang et al., 2021), *PICa* (Yang et al., 2021), *CFR* (Nguyen et al., 2021), *Pythia* (Singh et al., 2019), *VIVO* (Hu et al., 2020), *OSCAR* (Li et al., 2020).

Results

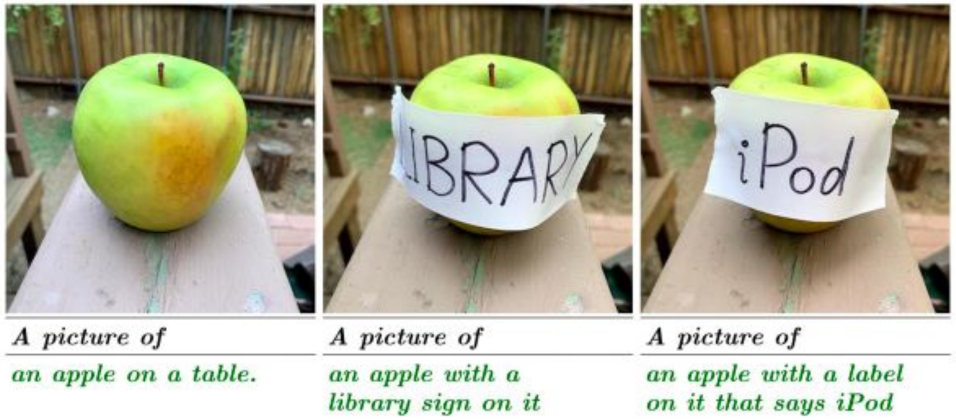


Figure 5: An example of an adversarial *typographic attack* which MAGMA appears robust to, unlike CLIP.

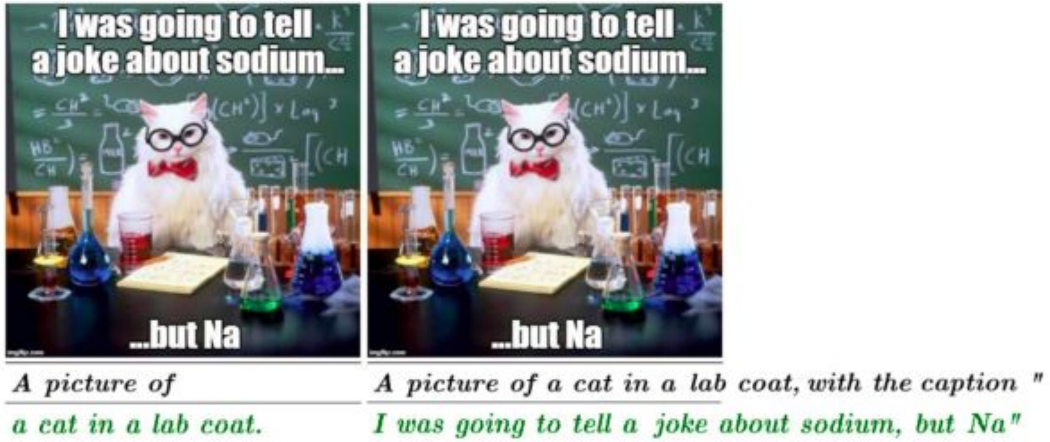


Figure 6: Example of *multi-step prompting*. Using the output of the model (left) again as the input (right), the generation procedure is broken down into atomic steps.

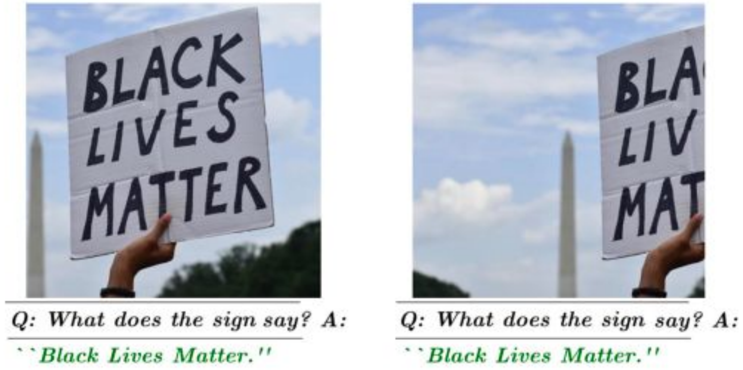


Figure 4: MAGMA's OCR capabilities. Even when

Mila

Ablations:

Adapter ablations					n-shot-VQA				n-shot-OKVQA				n-shot-GQA				n-shot-VizWiz				Avg.
Type	λ	Attn	FF	Params	0	1	2	4	0	1	2	4	0	1	2	4	0	1	2	4	
–	–	–	–	0.1	36.4	41.5	41.7	41.8	12.5	16.2	16.0	16.5	12.6	**20.8**	23.6	26.9	2.9	5.3	5.5	6.7	20.4
s	1	–	2	2	34.7	40.1	42.2	43.2	12.4	16.9	18.6	21.5	8.2	14.1	19.2	24.6	5.3	7.4	7.8	9.7	20.4
s	1	–	4	1	32.7	40.2	42.5	43.8	11.7	16.3	19.1	21.2	6.8	15.6	22.1	27.7	4.2	6.7	6.9	8.6	20.0
s	1	8	8	1	36.6	41.7	**43.8**	45.2	**13.9**	17.1	20.0	22.5	**14.3**	20.7	**24.9**	**28.4**	**5.6**	8.5	8.6	9.8	**22.6**
s	1	12	6	1	**36.9**	41.2	43.6	44.7	13.9	**19.4**	**21.6**	23.2	12.8	18.8	22.5	25.8	5.3	**9.6**	**9.8**	**10.6**	22.5
p	1	–	4	1	36.5	41.7	43.1	43.8	14.5	18.4	20.3	21.8	11.2	16.3	19.9	23.2	4.6	8.4	8.4	9.2	21.3
p	t	8	8	1	34.9	**42.2**	44.1	**45.4**	12.9	17.7	21.4	**23.4**	8.8	15.6	20.2	24.5	4.3	7.9	8.5	9.9	21.4
Encoder ablations																					
NFResnet					32.0	37.0	39.0	39.7	9.8	15.8	18.9	20.8	9.1	**20.2**	**27.1**	**28.7**	2.8	5.6	6.5	8.2	20.1
CLIP-ViT					32.8	33.9	36.7	37.7	10.5	9.2	12.4	14.2	8.4	14.9	22.2	25.7	2.7	5.1	5.2	7.7	17.5
CLIP-RN50x4					**35.2**	40.0	**42.6**	**44.2**	**12.6**	**17.7**	19.0	**21.8**	**10.5**	13.0	16.1	20.5	**5.0**	6.2	6.6	8.3	20.0
CLIP-RN50x16					32.7	**40.2**	42.5	43.8	11.7	16.3	**19.1**	21.2	6.8	15.6	22.1	27.7	4.2	**6.7**	**6.9**	**8.6**	**20.4**
Frozen (NFResnet + no adapters)					28.6	36.7	37.9	38.1	6.2	15.1	16.2	15.8	8.7	23.5	27.0	27.5	1.7	5.4	6.2	8.0	18.9

Type: (s)caled or (p)arallel. λ: 1 or (t)rained. Attn, FF: Downsample factor of the bottleneck in the resp. position. – means not applied. Params: Number of trainable parameters relative to the ablation with sequential FF adapters with downsample factor 4

Ablations:

Adapter ablations					NoCaps - CIDEr				NoCaps - B@4				CoCo - CIDEr	CoCo - B@4
Type	λ	Attn	FF	params	In	Out	Near	All	In	Out	Near	All		
–	–	–	–	0.1	45.1	53.7	43.3	45.7	9.9	5.8	7.9	7.8	36.7	10.3
s	1	–	2	2	37.7	55.5	40.6	43.2	6.2	6.1	6.5	6.4	33.4	9.4
s	1	–	4	1	39.3	56.2	44.0	45.8	6.3	6.7	7.7	7.3	39.6	11.2
s	1	8	8	1	38.2	49.5	40.9	42.2	6.4	4.9	6.7	6.3	37.1	10.6
s	1	12	6	1	**51.9**	**64.8**	**54.6**	**56.2**	**11.4**	**8.4**	**11.3**	**10.8**	**46.3**	**13.9**
p	1	–	4	1	37.5	38.1	35.9	36.0	7.2	5.1	6.7	6.4	36.3	10.8
p	t	8	8	1	40.6	58.3	45.0	47.1	8.0	6.6	7.9	7.7	39.5	11.2
Encoder ablations														
NFResnet					22.5	16.2	22.0	20.9	5.0	1.6	5.3	4.5	22.4	8.2
CLIP-ViT					33.2	44.2	35.3	36.8	5.9	5.2	5.8	5.7	27.2	7.7
CLIP-RN50x4					**47.7**	43.6	**48.1**	**50.2**	**9.3**	6.7	**9.2**	**8.7**	**41.9**	**13.1**
CLIP-RN50x16					39.3	**56.2**	44.0	45.8	6.3	**6.7**	7.7	7.3	39.6	11.2

Type: (s)caled or (p)arallel. λ: 1 or (t)rained. Attn, FF: Downsample factor of the bottleneck in the resp. position. – means not applied. Params: Number of trainable parameters relative to the ablation with sequential FF adapters with downsample factor 4

Mila

Ablations: Insights

1. Applying adapters to the attention layer is key.
2. More adapter parameters to the feed forward layer increases performance on knowledge-based tasks.
3. Balancing attention and feed-forward parameter allocation aids scene understanding.
4. CLIP-RN50x16, on average, performs best at VQA tasks.
5. CLIP-ViT has the worst average score across question answering tasks.

Questions?