CLIPPO: Image-and-Language Understanding from Pixels Only

arxiv.org/abs/2212.08045

Michael Tschannen, Basil Mustafa and Neil Houlsby
Google DeepMind Zurich

WED-AM, ABC 264, CVPR 2023

CLIPPO: Image-and-Language Understanding from Pixels Only

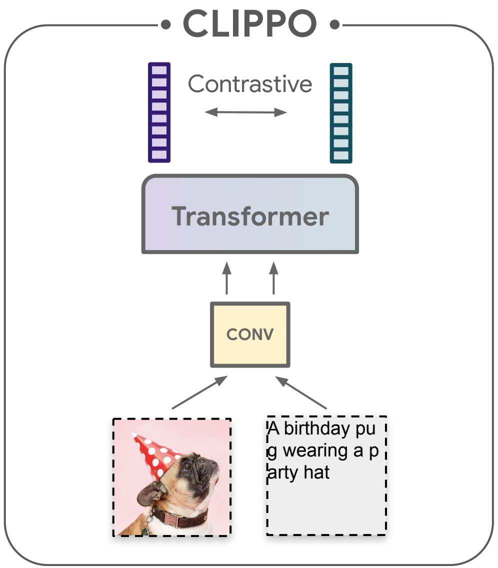


Image credit:
https://burst.shopify.com/photos/party-hat-frenchie?c=fun

- We propose a novel vision & language model which uses
  - **a single ViT to process visual input, or text, or both together, all rendered as RGB images**
  - a CLIP-style contrastive loss
- This simplifies input pipeline and transfer procedures, and side-steps tokenizer design
- **CLIPPO matches performance of an equivalent contrastive model with tokenizer** in
  - zero-shot image classification
  - EN and multilingual image/text retrieval
  - visual question answering
- CLIPPO outperforms prior pixel-based language modeling work on GLUE

Training details and data

- **Character lookup-based Unifont renderer**, font size 16px
- Baselines:
  - CLIP*: our CLIP implementation
  - 1T-CLIP: one-tower model with separate embeddings for images and tokenized text
- **Training recipe tuned for CLIP*** and used for CLIPPO and 1T-CLIP without modifications
- **Training data: WebLI** (Chen et al. 2023): 10B images with alt-texts in 109 languages

(Chen et al. 2023, arxiv:2209.06794)

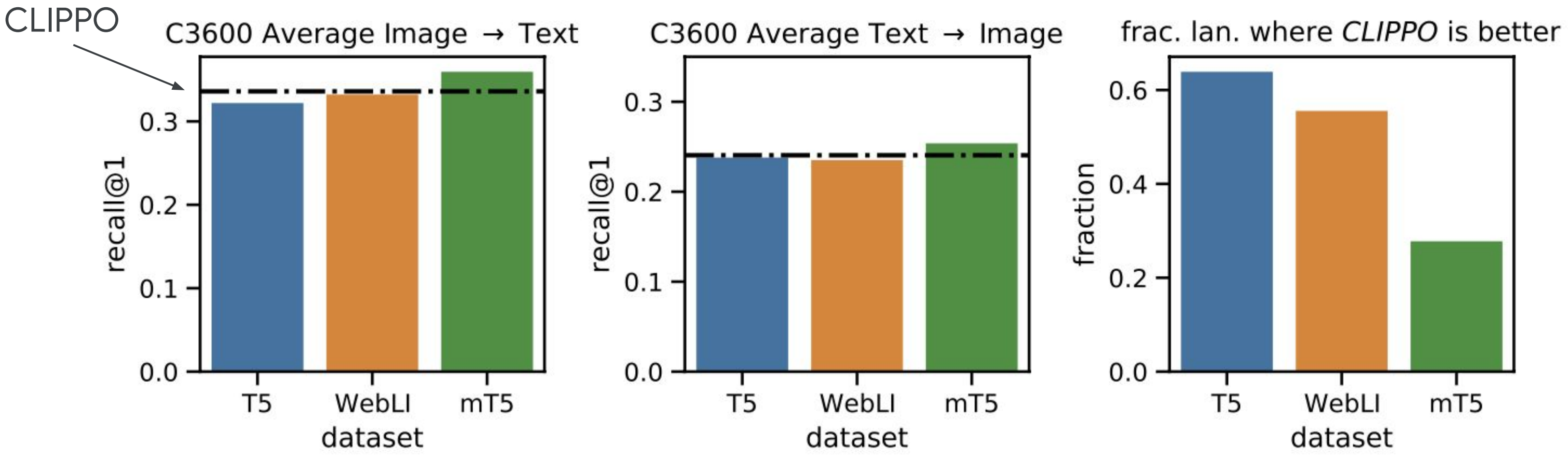
- Optional: **Co-training with text-pairs** (consecutive sentences) from C4

Vision-language results: Zero/few-shot classification and retr.

	#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

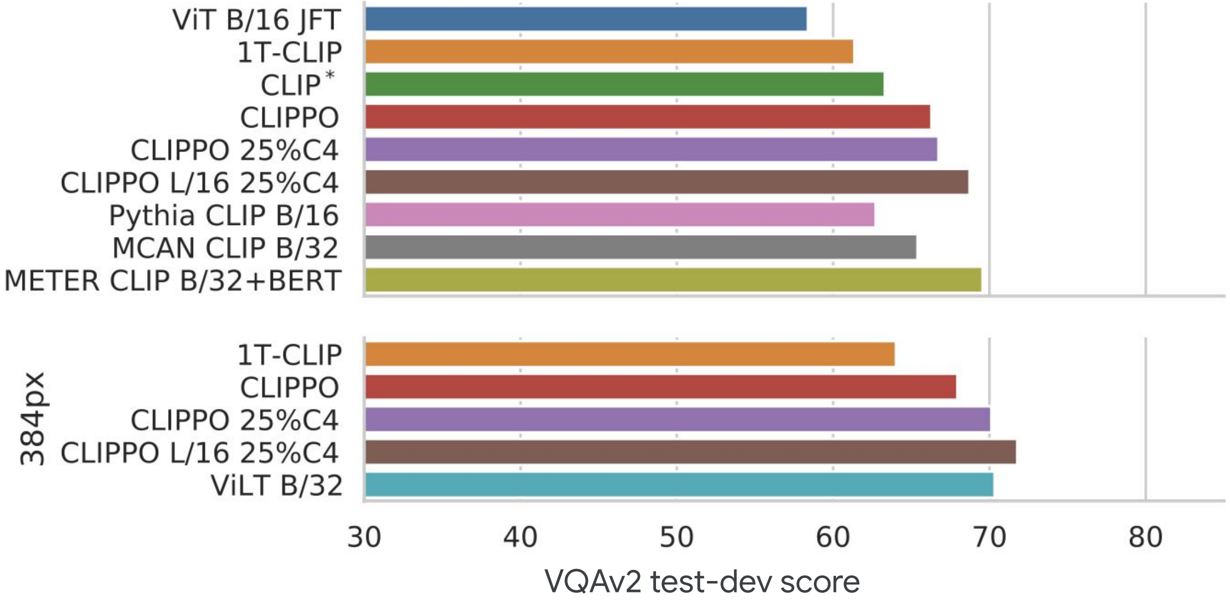
- CLIP* has the best results but has > 2x #params
- CLIPPO and 1T-CLIP achieve comparable performance
- Co-training with text/text pairs reduces performance (but also reduces the number of image/text pairs in the mini-batch)

Vision-language results: Multilingual zero-shot retrieval



- CLIPPO and 1T-CLIP trained on WebLI with **multilingual alt-texts**
- **CrossModal3600:** Diverse set of images each with captions in 36 languages
- CLIPPO: Matches/outperforms 1T-CLIP unless a large multilingual corpus (mC4) is used for tokenizer construction

Vision-language results: Visual question answering



Answer: surfing

https://farm6.staticflickr.com/5118/5802043331_8a76835c1d_z.jpg

example input image

- Fine-tuning with fused image + question input, answer by classification
- CLIPPO clearly outperforms other CLIP-style models
- CLIPPO performs comparably with VQA models using complex data and loss mix

Language results: GLUE benchmark

	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

- 1T-CLIP text enc. and CLIPPO perform similarly to BiLSTM-style models
- Very low CoLA score: alt-texts are rarely grammatical sentences
- CLIPPO w/ 50%C4 matches PIXEL; CLIPPO w/ C4 only outperforms PIXEL

Conclusion

CLIPPO shows that pixels alone are sufficient for multimodal Vision & Language learning, using only image-level contrastive losses (no word level loss!)

Check out the paper for an analysis of

- different contrastive text-only co-training tasks
- the efficiency of traditional and pixel-based tokenizers
- the modality gap of CLIPPO and 1T-CLIP
- the robustness of CLIPPO to typographic attacks

Paper: arxiv.org/abs/2212.08045
Code, models, colab: github.com/google-research/big_vision

Code & models

Colab