ECCV TEL AVIV 2022

Open-Vocabulary Visual Perception upon Frozen Vision and Language Models

Yin Cui
Research Scientist, Google

Google Research

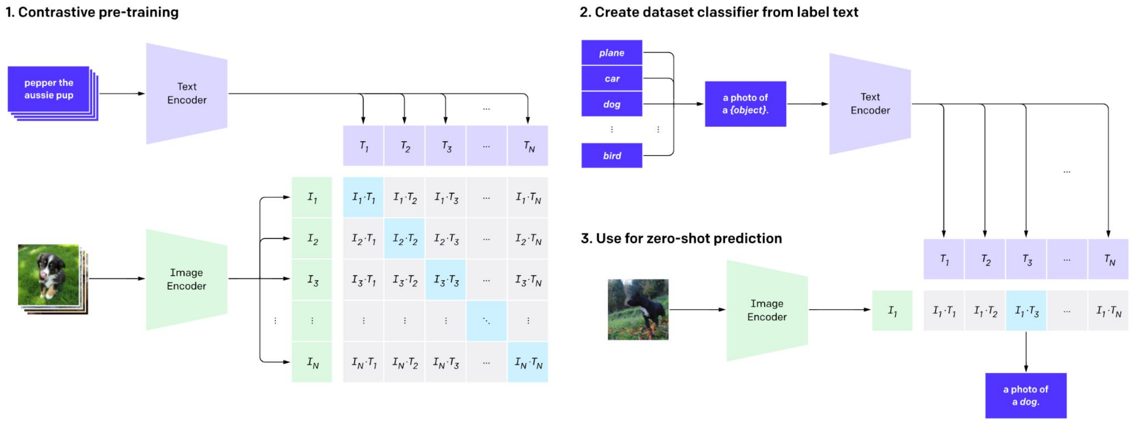
Workshop on Computer Vision in the Wild @ECCV 2022
Oct 23, 2022

01

Background

Vision and Language Models

- Traditional vision models:
  - Pre-training: fixed set of discretized labels (e.g., ImageNet, 1000 classes)
  - Downstream transfer: weight initialization for fine-tuning, **no open-vocabulary capability**
- Vision and language models:
  - Pre-training: large-scale image-text pairs (e.g., CLIP, 400M web image-text pairs)
  - Downstream transfer: direct "zero-shot" or fine-tuning, **open-vocabulary capability through vision-language alignment**



Google Research

https://openai.com/blog/clip/

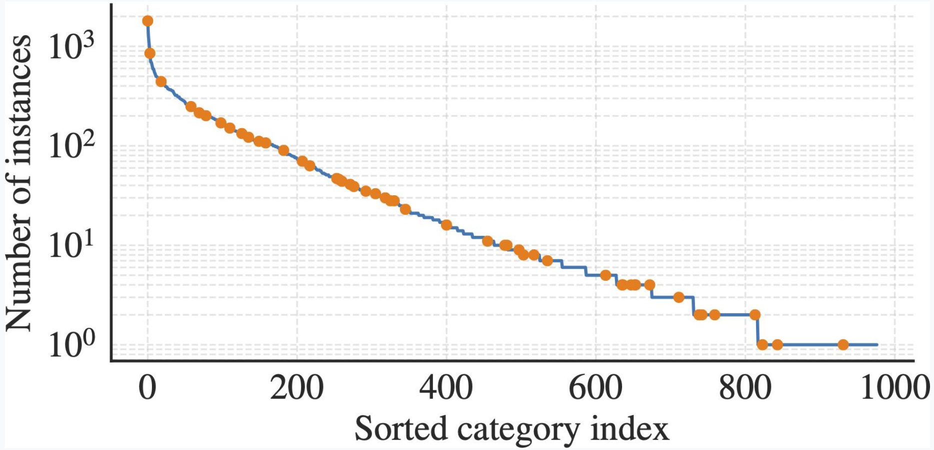
Dataset collection for large vocabulary detection

Dataset	# images	# boxes	# categories
Pascal VOC	11.5k	27k	20
COCO	159k	896k	80
Objects 365	1800k	29,000k	365
LVIS v1.0	159k	1,514k	**1203**

Challenge:
Long-tailed distribution

- A few dominant classes claim most of the data, while most classes have few examples
- Exponentially more data might be needed for rare classes
  - Expensive for tasks requiring extensive annotations: detection, segmentation, video, etc.
- Alternatives?

Number of instances per class in LVIS dataset



Google Research

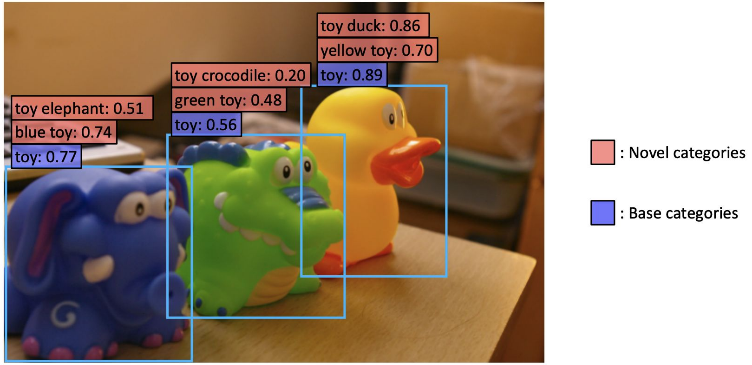
Gupta et al., 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation.

Open-Vocabulary Visual Perception upon Vision and Language Models

Google Research

ViLD: Detection via Distillation

- Mask R-CNN with frozen CLIP text encoder: class embeddings as classifiers
- Training:
  - Offline extract CLIP image embeddings on some cropped and resized regions (similar to R-CNN)
  - Online training of Mask R-CNN on **base** classes with distillation between Mask R-CNN region embeddings and offline CLIP image embeddings
- Inference: Online extract **novel** class embeddings, then use pre-trained Mask R-CNN

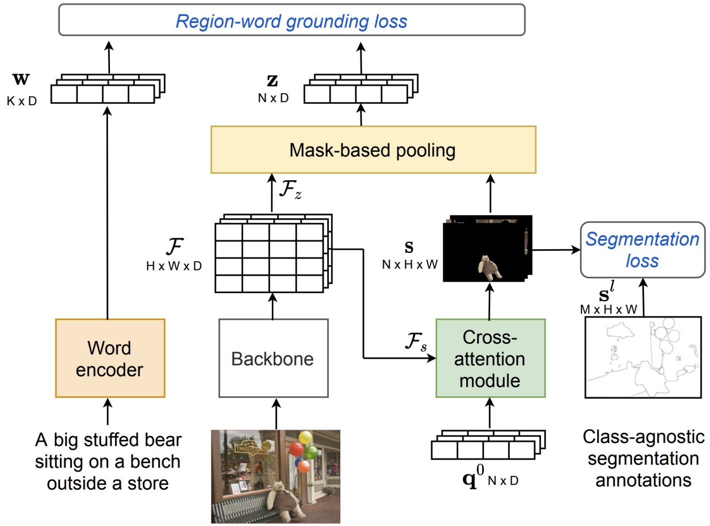


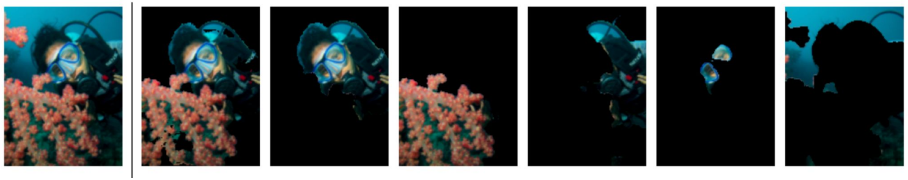
Paper: Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge. ICLR 2022.
Code: https://github.com/tensorflow/tpu/tree/master/models/official/detection/projects/vild
Demo: https://colab.sandbox.google.com/github/tensorflow/tpu/blob/master/models/official/detection/projects/vild/ViLD_demo.ipynb

Google Research

OpenSeg: Segmentation via Aligning Regions with Captions

- Supervision: class-agnostic segmentation + image caption





Paper: Golnaz Ghiasi, Xiuye Gu, Yin Cui, Tsung-Yi Lin. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. ECCV 2022. (Vision + Language Session)
Code: https://github.com/tensorflow/tpu/tree/master/models/official/detection/projects/openseg
Demo: https://colab.sandbox.google.com/github/tensorflow/tpu/blob/master/models/official/detection/projects/openseg/OpenSeg_demo.ipynb

Google Research

Open-Vocabulary Visual Perception upon Frozen Vision and Language Models

Google Research

02

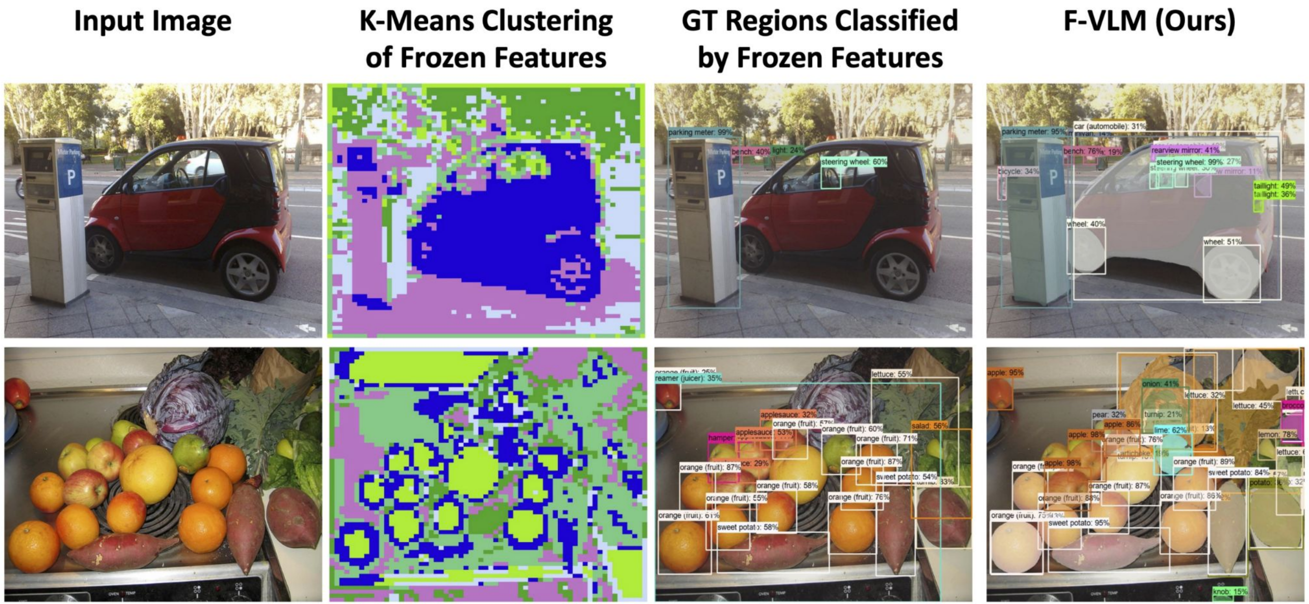
Frozen VLMs → Object Detection

F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models
https://arxiv.org/abs/2209.15639

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, Anelia Angelova

Frozen CLIP is a good localizer and a good region classifier

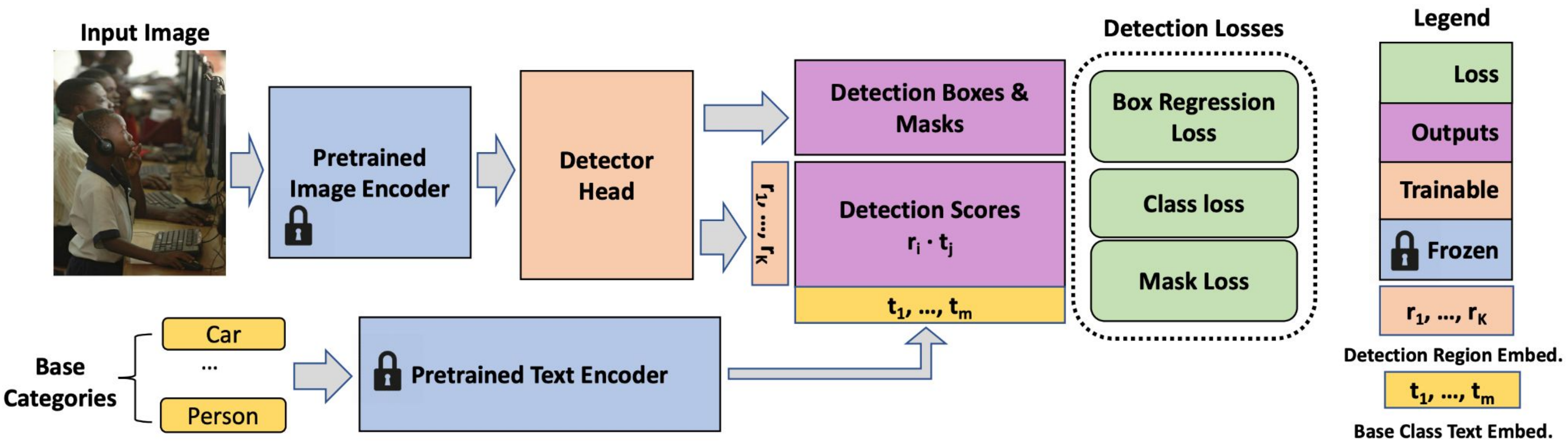
- CLIP features are
  - locality sensitive for describing object shapes via simple k-means clustering (k=6)
  - discriminative for region classification on ground truth regions from LVIS



Google Research

F-VLM Training

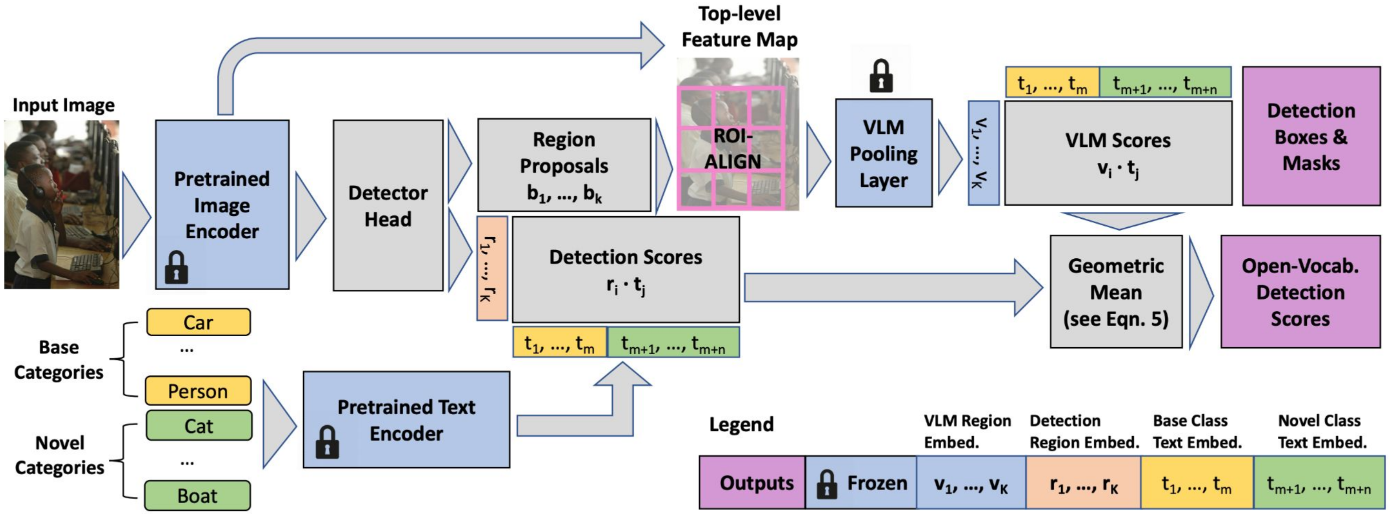
- During training, F-VLM is simply a detector with the last classification layer replaced by base-category text embeddings.
- We only train the detector head (RPN, FPN and Mask R-CNN heads) and keep the pretrained VLM image and text encoder frozen.



Google Research

F-VLM Inference

- At test time, we use the region proposals to crop out the top-level features of the VLM vision encoder and compute the VLM score per region. VLM pooling is an attention layers used in CLIP.
- We combine the detection score and the VLM score for open-vocabulary detection of unseen classes.



Google Research

Results on LVIS Open-Vocabulary Benchmark

- Base: 866 LVIS frequent + common classes (10-1977 images per class)
- Novel: 337 LVIS rare classes (<10 images per class)
- State-of-the-art on LVIS novel classes (+6.5 AP)

Backbone	Pretrained CLIP	Method	Distill	Trainable Backbone	AP_r	AP
R50 Comparison:						
R50	ViT-B/32	ViLD (Gu et al., 2022)	✓	✓	16.1	22.5
R50	ViT-B/32	ViLD-Ens. (Gu et al., 2022)	✓	✓	16.6	25.5
R50	ViT-B/32	DetPro (Du et al., 2022)‡	✓	✓	19.8	25.9
R50	ViT-B/32	Detic-ViLD (Zhou et al., 2022c)*	✗	✓	17.8	26.8
R50	R50	RegionCLIP (Zhong et al., 2022)†	✓	✓	17.1	28.2
R50	R50	F-VLM (Ours)	✗	✗	18.6	24.2
System-level Comparison:						
R152	ViT-B/32	ViLD (Gu et al., 2022)	✓	✓	18.7	23.6
R152	ViT-B/32	ViLD-Ens. (Gu et al., 2022)	✓	✓	18.7	26.0
EN-B7	ViT-L/14	ViLD-Ens. (Gu et al., 2022)	✓	✓	21.7	29.6
EN-B7	EN-B7*	ViLD-Ens. (Gu et al., 2022)	✓	✓	26.3	29.3
R50	ViT-B/32	DetPro-Cascade (Du et al., 2022)‡	✓	✓	20.0	27.0
R50	ViT-B/32	Detic-CN2 (Zhou et al., 2022c)*	✗	✓	24.6	32.4
R50x4	R50x4	RegionCLIP (Zhong et al., 2022)†	✓	✓	22.0	32.3
ViT-L/14	ViT-L/14	OWL-ViT (Minderer et al., 2022)	✗	✓	25.6	34.7
R50x4	R50x4	F-VLM (Ours)	✗	✗	26.3	28.5
R50x16	R50x16	F-VLM (Ours)	✗	✗	30.4	32.1
R50x64	R50x64	F-VLM (Ours)	✗	✗	**32.8**	34.9

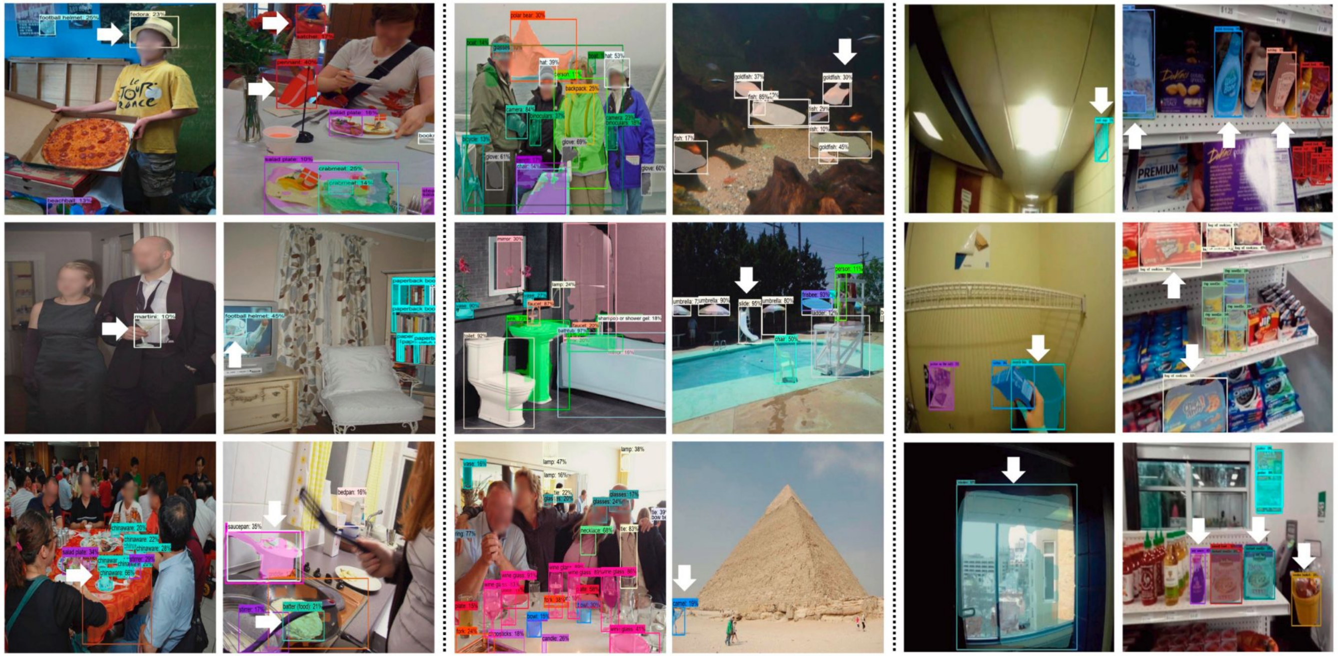
Google Research

Cross-Dataset Transfer

- LVIS base → COCO, Objects 365, Ego4D

Method	COCO			Objects365		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised (Gu et al., 2022)	46.5	67.6	50.9	25.6	38.6	28.0
ViLD-R50 (Gu et al., 2022)	36.6	55.6	39.8	11.8	18.2	12.6
DetPro-R50 (Du et al., 2022)	34.9	53.8	37.4	12.1	18.8	12.9
F-VLM-R50 (Ours)	32.5	53.1	34.6	11.9	19.2	12.6
F-VLM-R50x4 (Ours)	36.0	57.5	38.7	14.2	22.6	15.2
F-VLM-R50x16 (Ours)	37.9	59.6	41.2	16.2	25.3	17.5
F-VLM-R50x64 (Ours)	**39.8**	**61.6**	**43.8**	**17.7**	**27.4**	**19.1**

LVIS Novel Categories **Objects365 Transfer** **Ego4D Transfer**



Google Research

Training Efficiency

- 14x - 226x less training cost (TPUv3 per-core-hour) compared with the best ViLD model
- Great potential to be incorporated with gigantic VLMs for both fine-tuning or co-training

Method	Mask AP_r	#Iters	Epochs	Training Cost (Per-Core-Hour)	Training Cost Savings
ViLD-EN-B7 (Gu et al., 2022)	26.3	180k	460	8000	1×
F-VLM (Ours)	32.8	46.1k	118	565	14×
F-VLM (Ours)	31.0	5.76k	14.7	71	113×
F-VLM (Ours)	27.7	**2.88k**	**7.4**	**35**	**226**×

Google Research

03

Frozen VLMs → Multimodal Video

MOV: Multimodal Open-Vocabulary Video Classification via Pre-Trained Vision and Language Models
https://arxiv.org/abs/2207.07646

Rui Qian, Yeqing Li, Zheng Xu, Ming-Hsuan Yang, Serge Belongie, Yin Cui

Google Research

Frozen VLM is a good video classifier

- Benchmark on Kinetics-400 video action classification
- Frozen CLIP vision encoder with a trainable light-weight head
- Frozen CLIP text encoder to extract class embeddings as classifiers

Model	Head	Prompt	Training	Kinetics Top-1
CLIP ViT-B/16	Avg Pool	80 ImageNet	No (0-shot)	54.9
CLIP ViT-B/16	Avg Pool	28 Video	No (0-shot)	59.3 (+4.4)
CLIP ViT-B/16	Avg Pool	28 Video	Linear	74.5 (+15.2)
CLIP ViT-B/16	TFM + Avg Pool	28 Video	Linear	77.2 (+2.7)
SlowFast	–	–	E2E Training	79.8
TimeSformer	–	–	E2E Training	78.0
ViViT-B	–	–	E2E Training	80.0

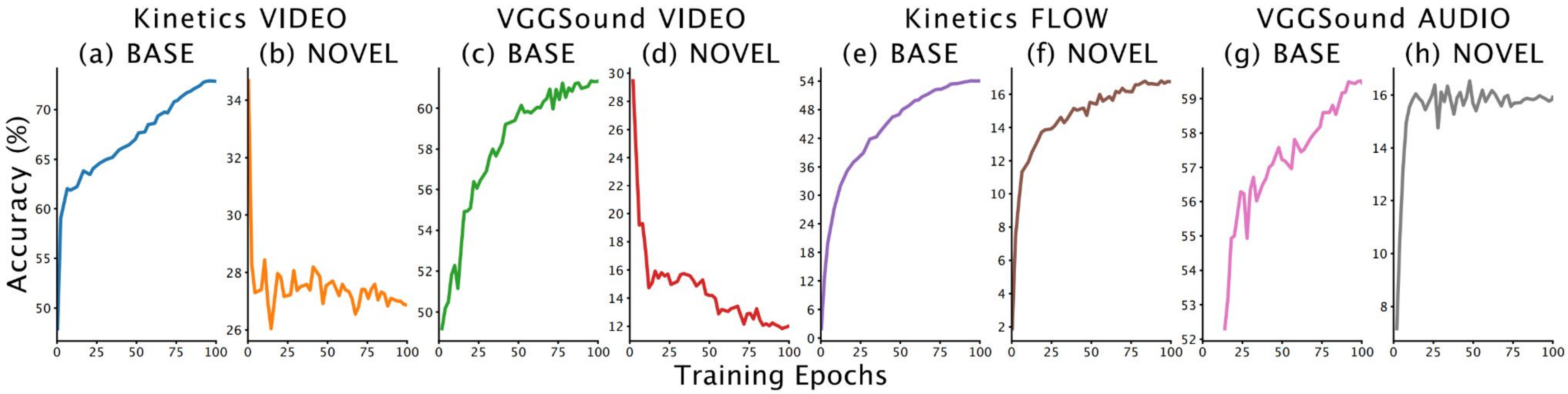
Google Research

Pre-trained VLM for multimodal video

- Frozen CLIP for video
- Audio and motion naturally co-exist with video
- Pre-trained CLIP for
  - Audio Spectrogram
    - duplicate to a 3-channel image; bilinearly interpolating the positional encoding
  - Optical Flow
    - 2-channel x-y motion + zero-padded 3rd channel ⟶ 3-channel image

Fine-tuning VLM for multimodal open-vocabulary video

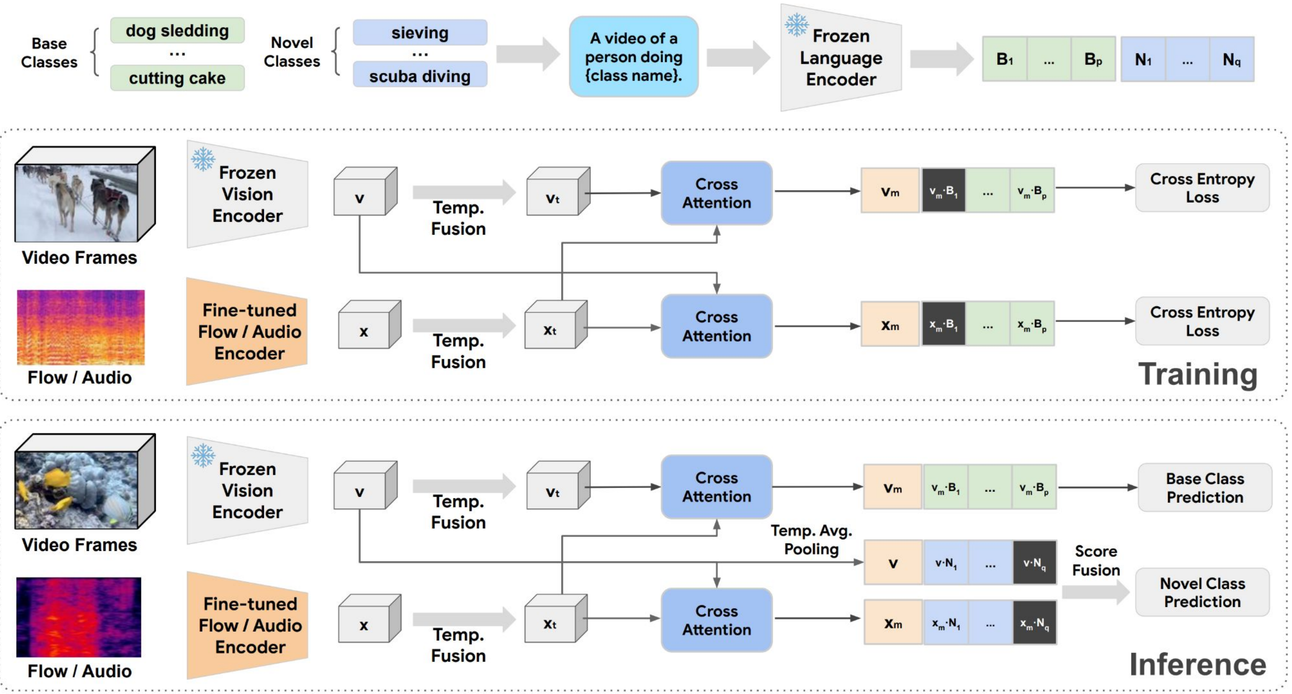
- Kinetics, VGGSound datasets: train on base classes, eval on both base and novel classes
- Observations:
  - All modalities are able to improve on base classes
  - With base class training, flow and audio can generalize to novel classes while video cannot
- How can we build a model to leverage audio and flow to help video?



Google Research

Overview of MOV

- Transformer head for temporal fusion
- Asymmetrical cross-attention to fuse multimodal information



Google Research

Open-Vocabulary video classification

- Kinetics (Video, Flow, Text) and VGGSound (Video, Audio, Text)
  - Outperforms multimodal method VATT, frozen CLIP, and CLIP adaptation methods (CoOp, CLIP-Adapter)

method	modalities	base acc.	novel acc.	harmonic mean
VATT (Akbari et al., 2021)	V, T	19.8	21.6	20.7
CLIP (Radford et al., 2021)	V, T	51.2	56.7	53.8
CoOp (Zhou et al., 2021)	V, T	58.9	45.7	51.5
CLIP-Adapter (Gao et al., 2021)	V, T	66.5	36.2	46.9
MOV (Ours)	V, F, T	(+8.8) **75.3**	(+1.4) **58.1**	(+11.8) **65.6**

method	modalities	base acc.	novel acc.	harmonic mean
VATT (Akbari et al., 2021)	V, A, T	21.6	23.7	22.6
CLIP (Radford et al., 2021)	V, T	48.5	48.8	48.6
CoOp (Zhou et al., 2021)	V, T	56.9	42.0	48.3
CLIP-Adapter (Gao et al., 2021)	V, T	60.0	27.5	37.7
MOV (Ours)	V, A, T	(+8.4) **68.4**	(+2.7) **51.5**	(+10.2) **58.8**

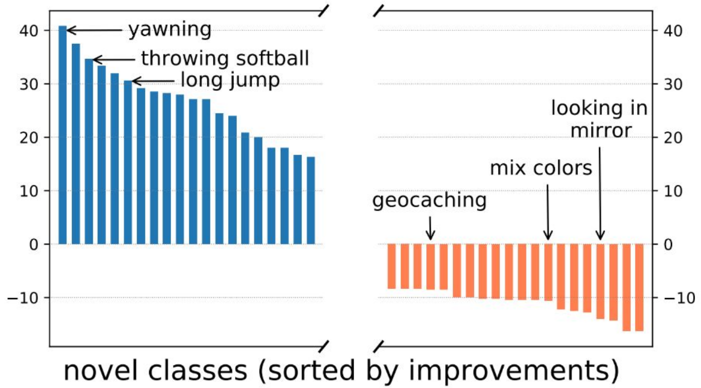
Scalability of MOV

- MOV scales well with a stronger ViT-L/14 backbone.
  - CLIP's improvements translates well to MOV for novel classes

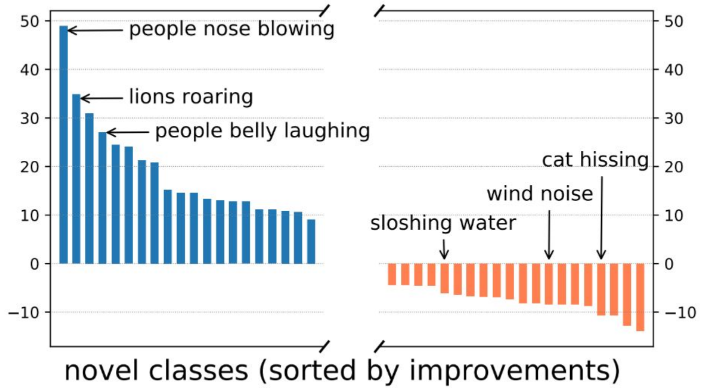
method	backbone	Kinetics-700 base acc.	Kinetics-700 novel acc.	VGGSound base acc.	VGGSound novel acc.
CLIP (Radford et al., 2021)	ViT-B/16	51.2	56.7	48.5	48.8
CLIP (Radford et al., 2021)	ViT-L/14	59.6	65.3	52.6	54.1
MOV (Ours)	ViT-B/16	75.3	58.1	68.4	51.5
MOV (Ours)	ViT-L/14	(+4.8) **80.1**	(+8.8) **66.9**	(+3.5) **71.9**	(+4.6) **56.1**

Multimodal improvement analysis

- Per-class improvement on Kinetics and VGGSound due to additional modality



(a) **Kinetics with additional flow modality**



(b) **VGGSound with additional audio modality**

Google Research

Cross-Dataset Transfer on UCF and HMDB

- UCF* and HMDB* are used extensively by existing zero-shot methods where we randomly choose 50% of the classes for training and evaluate on the rest 50% (10 runs)
- MOV achieves state-of-the-art, outperforming both traditional zero-shot methods and most recent CLIP-adaptation methods by large margins

method	vision† + text‡	pre-train§	UCF*/UCF	HMDB*/HMDB
GA (Mishra et al., 2018)	C3D + W2V	S1M	17.3±1.1 / -	19.3±2.1 / -
TARN (Bishay et al., 2019)	C3D + W2V	S1M	19.0±2.3 / -	19.5±4.2 / -
CWEGAN (Mandal et al., 2019)	I3D + W2V	IN, K400	26.9±2.8 / -	30.2±2.7 / -
TS-GCN (Gao et al., 2019)	GLNet + W2V	IN-shuffle	34.2±3.1 / -	23.2±3.0 / -
PS-GNN (Gao et al., 2020)	GLNet + W2V	IN-shuffle	36.1±4.8 / -	25.9±4.1 / -
E2E (Brattoli et al., 2020)	R(2+1)D + W2V	K700	48.0 / 37.6	32.7 / 26.9
DASZL (Kim et al., 2021)	TSM + Attributes	IN, K400	48.9±5.8 / -	- / -
ER (Chen & Huang, 2021)	TSM + BERT	IN, K400	51.8±2.9 / -	35.3±4.6 / -
ResT (Lin et al., 2022)	RN101 + W2V	K700	58.7±3.3 / 40.6	41.1±3.7 / 34.4
MIL-NCE (Miech et al., 2020)	S3D + W2V	HT100M	- / 29.3	- / 10.4
VideoCLIP (Xu et al., 2021)	S3D + TSF	HT100M	- / 22.5	- / 11.3
VATT (Akbari et al., 2021)	ViT + TSF	HT100M	- / 18.4	- / 13.2
CLIP (Radford et al., 2021)	ViT-B/16 + TSF	WIT	79.9±3.8 / 73.0	54.0±4.1 / 46.1
ActionCLIP (Wang et al., 2021)	ViT-B/16 + TSF	WIT$^+$	- / 69.5	- / 50.5
X-CLIP (Ni et al., 2022)	ViT-B/16 + TSF	WIT$^+$	- / 72.0	- / 44.6
MOV (Ours)	ViT-B/16 + TSF	WIT$^+$	**82.6±4.1 / 76.2**	**60.8±2.8 / 52.1**
MOV (Ours)	ViT-L/14 + TSF	WIT$^+$	**87.1±3.2 / 80.9**	**64.7±3.2 / 57.8**

† vision encoder: C3D (Tran et al., 2015), I3D (Carreira & Zisserman, 2017), GLNet (Szegedy et al., 2015), R(2+1)D (Tran et al., 2018), TSM (Lin et al., 2019), RN101 (He et al., 2016), S3D (Xie et al., 2018), ViT (Dosovitskiy et al., 2021).
‡ text encoder: W2V (Mikolov et al., 2013), BERT (Devlin et al., 2019), TSF (Vaswani et al., 2017).
§ pre-train data: S1M (Karpathy et al., 2014), IN (Deng et al., 2009), K400 (Kay et al., 2017), IN-shuffle (Mettes et al., 2016), K700 (Carreira et al., 2019), HT100M (Radford et al., 2021), WIT (Radford et al., 2021), WIT$^+$ has additional training on Kinetics.

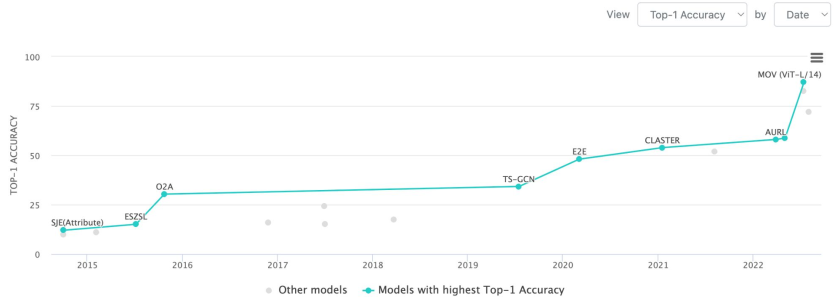
Cross-Dataset Transfer on UCF and HMDB

- 72.0 ⟶ 87.1 on UCF (+15.1)
- 44.6 ⟶ 64.7 on HMDB (+20.1)

Zero-Shot Action Recognition

Zero-Shot Action Recognition on UCF101

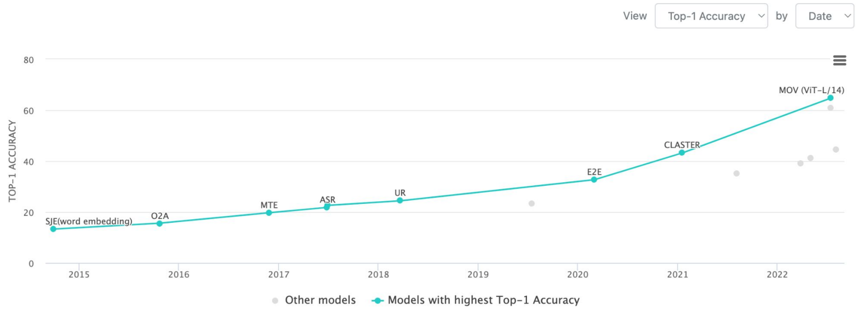
Leaderboard Dataset



Zero-Shot Action Recognition

Zero-Shot Action Recognition on HMDB51

Leaderboard Dataset



Google Research

Conclusion

- Utilizing pre-trained VLMs has become a promising paradigm for open-vocabulary visual perception (Detection, Segmentation, Video, etc.)

- It's possible to greatly simplify the paradigm by directly building upon frozen VLMs with minimal modifications
    - In F-VLM: Frozen VLMs are strong object localizer and region classifier
    - In MOV: Frozen VLMs are strong video classifier and can be further improved by fine-tuning on additional modalities like audio and flow

- Other than being performant models, other advantages of building upon frozen VLMs are
    - Reduced training computation and memory
    - Potential of scaling to gigantic foundation models and co-training with foundation models

Acknowledgement of Co-Authors

Xiuye Gu
Google

Weicheng Kuo
Google

Golnaz Ghiasi
Google

Rui Qian
Cornell

Tsung-Yi Lin
Nvidia

AJ Piergiovanni
Google

Yeqing Li
Google

Zheng Xu
Google

Ming-Hsuan Yang
Google / UC Merced

Anelia Angelova
Google

Serge Belongie
University of Copenhagen