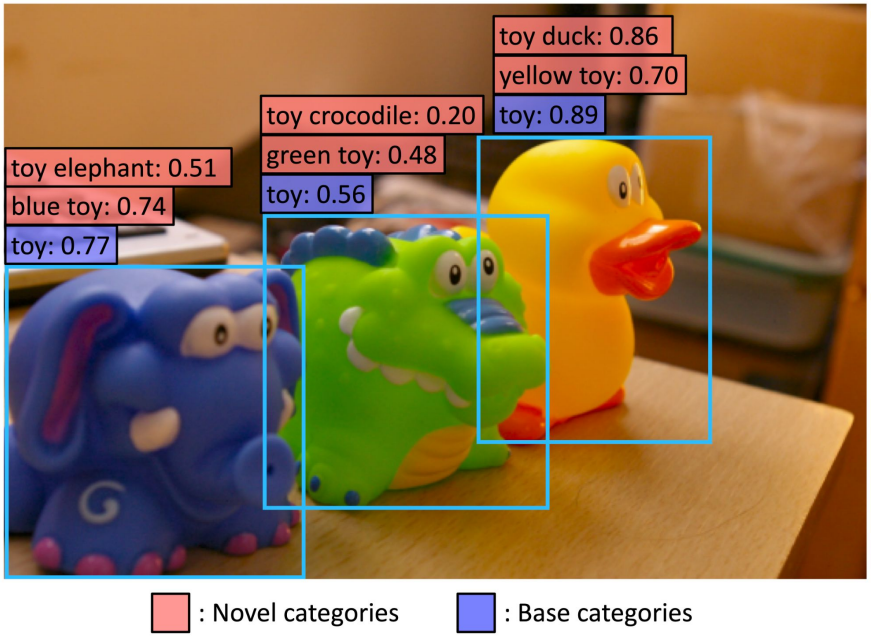
Open-vocabulary Object Detection via Vision and Language Knowledge Distillation

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, Yin Cui

Google Research

Open-vocabulary detection



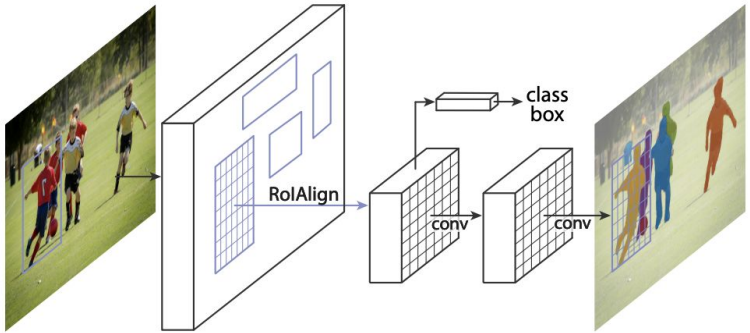
- A new direction for large-vocabulary detection
  - Instead of collecting costly annotations
- Change detection categories during inference
  - Without retraining

Google Research

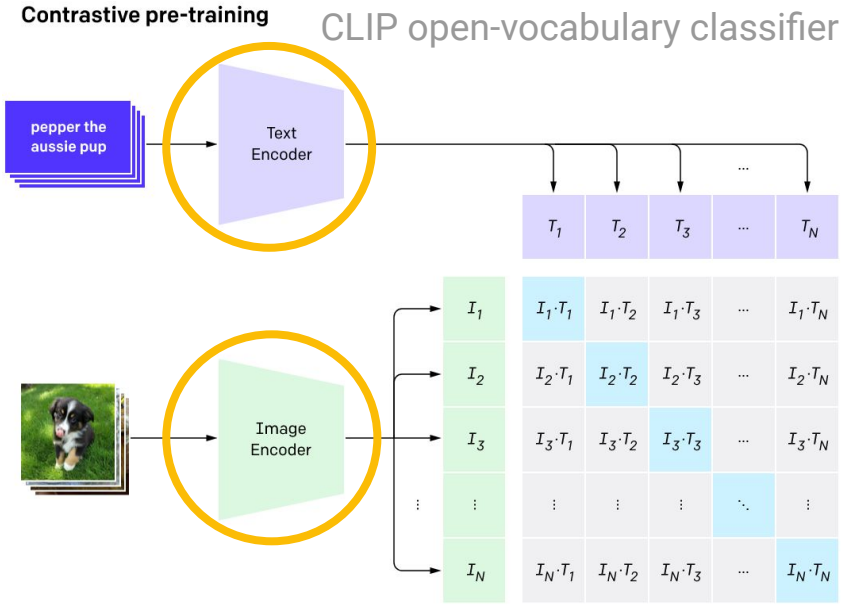
Borrow knowledge from open-vocabulary classification model

- Decouple localization and classification
- **Localization**: class-agnostic modules generalize well
- **Classification**: Pretrained open-vocabulary classifier ⟶ two-stage detector
- ***ViLD***: **Vi**sion and **L**anguage Knowledge **D**istillation

Mask R-CNN detector



CLIP open-vocabulary classifier

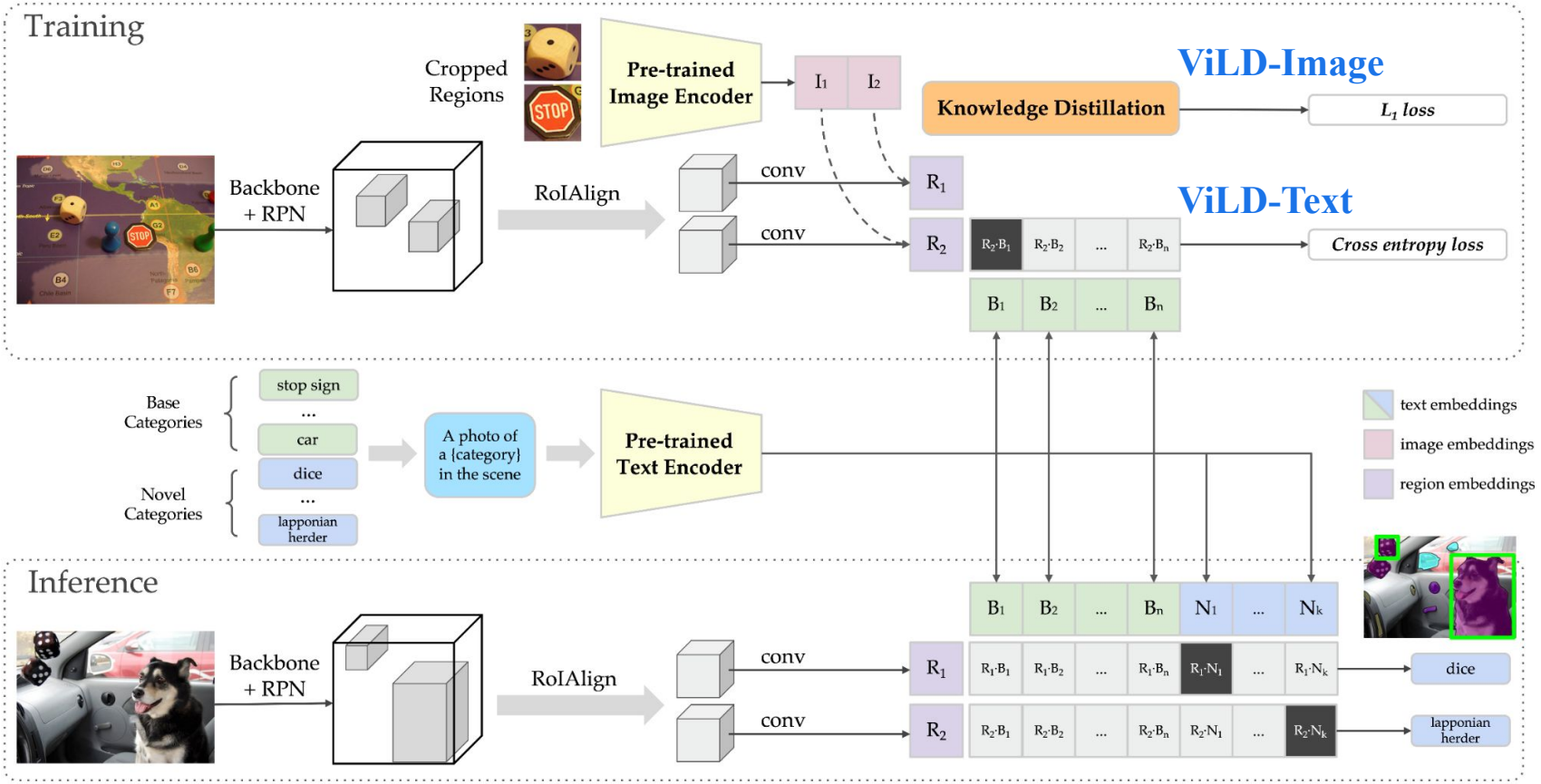


https://openai.com/blog/clip/

Radford et al., 2021. Learning Transferable Visual Models From Natural Language Supervision.

He et al., 2017. Mask R-CNN.

ViLD overview



Quantitative results on LVIS

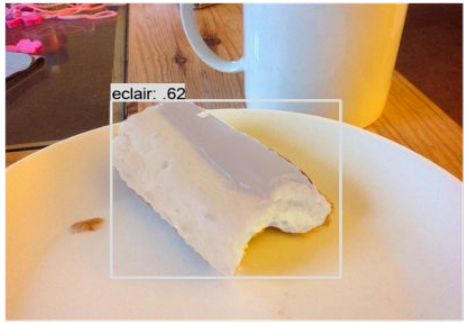
- Hold out LVIS rare categories as novel categories (APr), which is the main metric
- Outperformed supervised baseline
- Strongest model is close to fully-supervised challenge winner (SOTA)

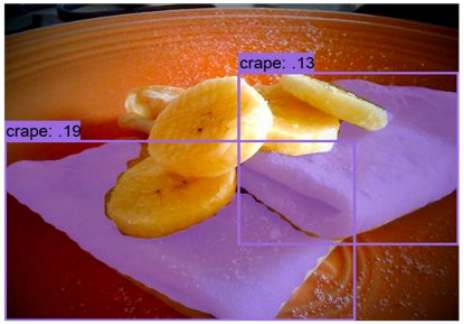
Backbone: ResNet152 (except last row)

Method	AP_r	AP_c	AP_f	AP
ViLD-text	11.7	25.8	34.4	26.7
ViLD-image	10.8	10.0	8.7	9.6
ViLD ($w = 1.0$)	**18.7**	21.1	28.4	23.6
ViLD-ensemble ($w = 2.0$)	**18.7**	24.9	30.6	26.0
Supervised-RFS (base + novel)	14.4	26.8	34.2	27.6
ViLD-ensemble-ALIGN-b7	**26.3**	27.2	32.9	29.3
2020 Challenge winner	30.0	41.9	46.0	41.5

Google Research

Qualitative examples of detecting novel objects (LVIS)





Google Research

Finetuning-free transfer

- Finetuning-free transfer to PASCAL / COCO / Objects365 by simply replacing text embeddings

Backbone: ResNet50

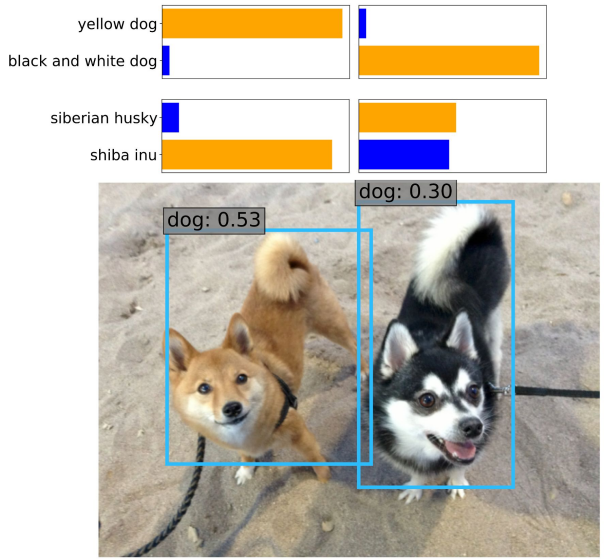
Method	PASCAL VOC†		COCO			Objects365		
	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
ViLD-text	40.5	31.6	28.8	43.4	31.4	10.4	15.8	11.1
ViLD	72.2	56.7	36.6	55.6	39.8	11.8	18.2	12.6
Finetuning	78.9	60.3	39.1	59.8	42.4	15.2	23.9	16.2
Supervised	78.5	49.0	46.5	67.6	50.9	25.6	38.6	28.0

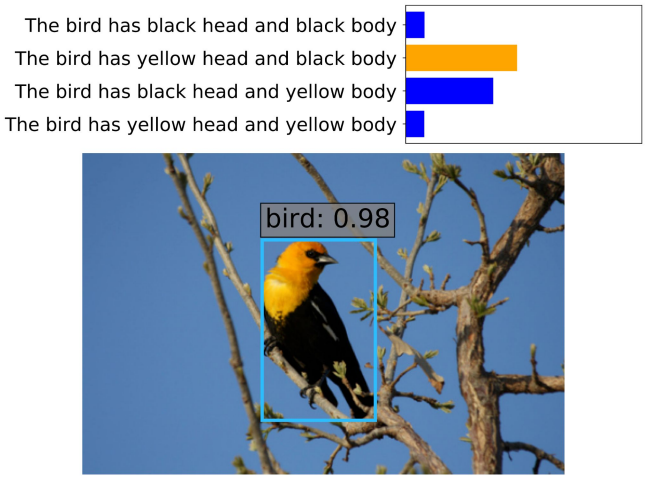
PASCAL VOC

COCO

On-the-fly interactive detection

- After detecting pre-defined categories, use on-the-fly free-form text embeddings to recognize more details.

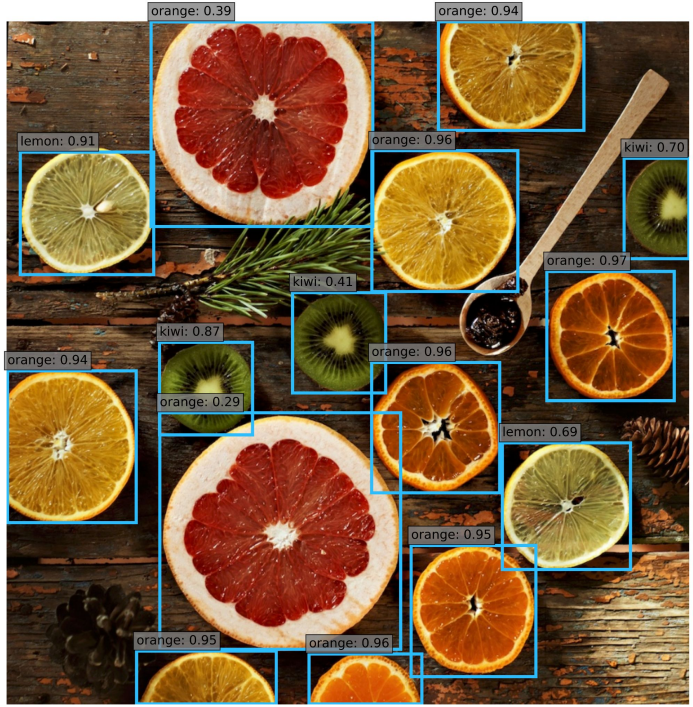




Google Research

Systematic expansion of dataset vocabulary

- Detect fruit with color attributes (expand LVIS vocabulary with 11 colors).



Original dataset vocabulary

Expanded with color attributes

Google Research

Thank You

Code: https://github.com/tensorflow/tpu/tree/master/models/official/detection/projects/vild

Colab demo:
https://colab.sandbox.google.com/github/tensorflow/tpu/blob/master/models/official/detection/projects/vild/ViLD_demo.ipynb

Google Research