COMP 648: Computer Vision Seminar

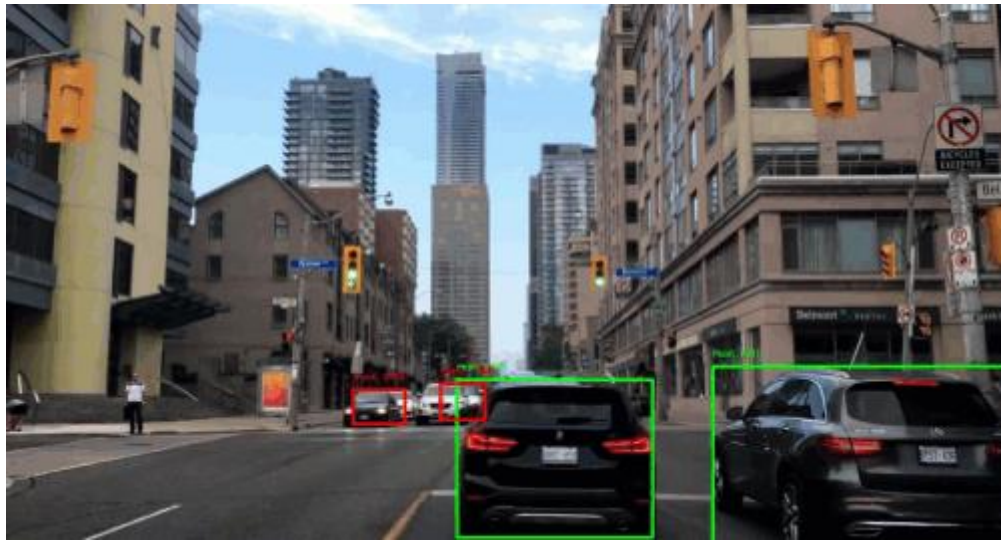
Visual grounding: Learning to localize objects

Atanu Dahari

RICE UNIVERSITY

Object Detection and Localization

- One of the most important and challenging tasks in Computer Vision.
- Recent surge of interest in object detection.
  - Self driving cars
  - Robot vision
  - Video surveillance



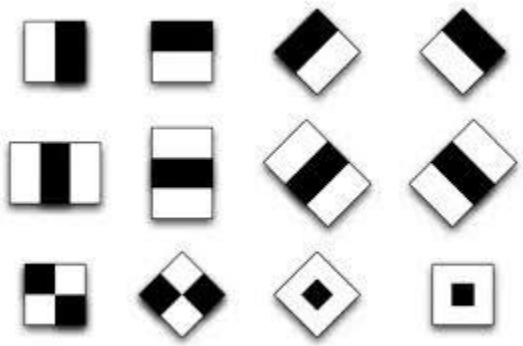
Background: A Road Map of Object Detection

Traditional era – back in 2001

1. **Viola Jones Detectors:**

Paper: https://arxiv.org/pdf/1905.05055.pdf

Goal – Detection of human faces in real time using sliding windows.

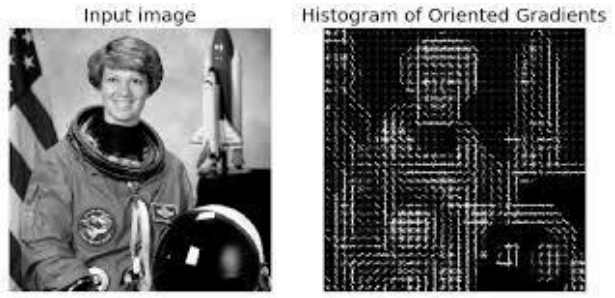


Haar like features

- Haar wavelet is used as the feature representation of an image

2. HOG (Histogram of Oriented Gradients)

- Mainly designed for human detection
- Based on the idea that the object's shape can be defined by the length and density of gradient vectors.
- One of the first object detection algorithms to use normalization to avoid feature invariance.



3. DPM

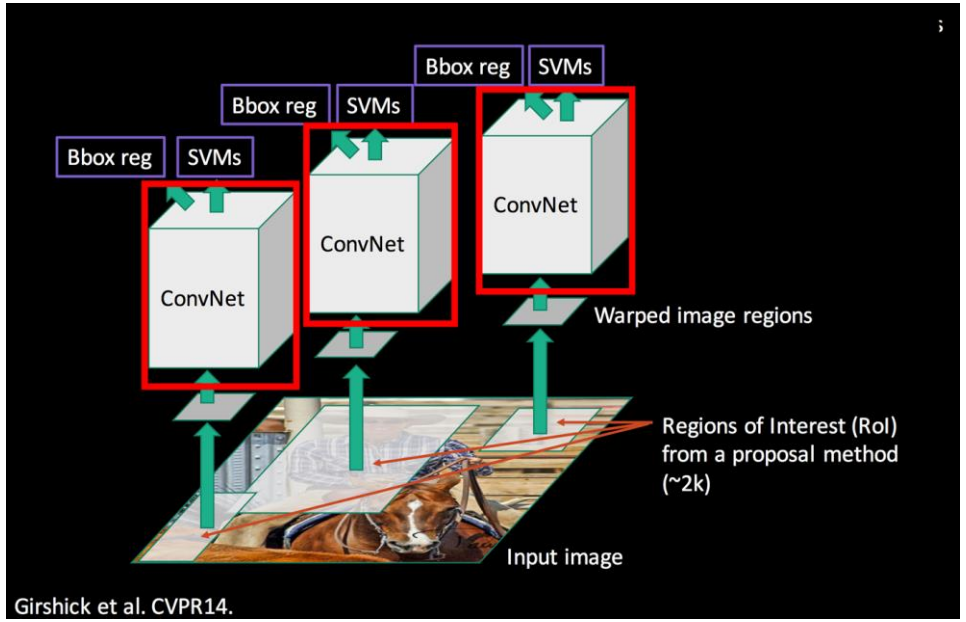
- Goal – To detect small parts of an object ensemble them to detect the whole object

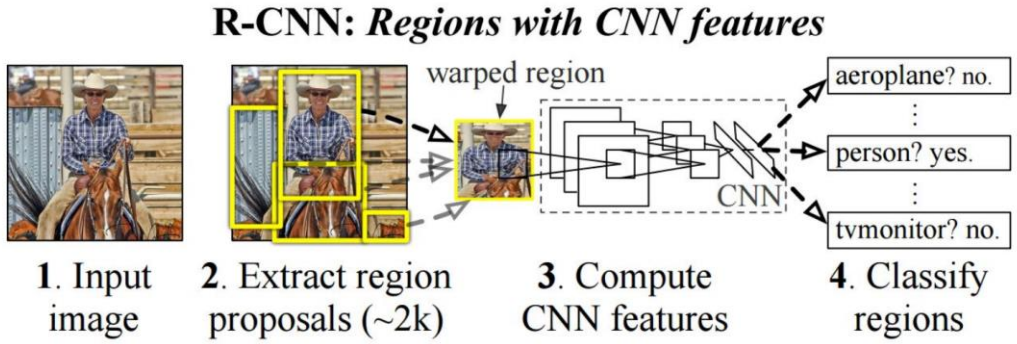
Deep learning era - 2014

- **Problem with traditional algorithms** – Huge number of features to calculate making them computationally difficult.

1. **RCNN – Region based Convolutional Neural Networks**

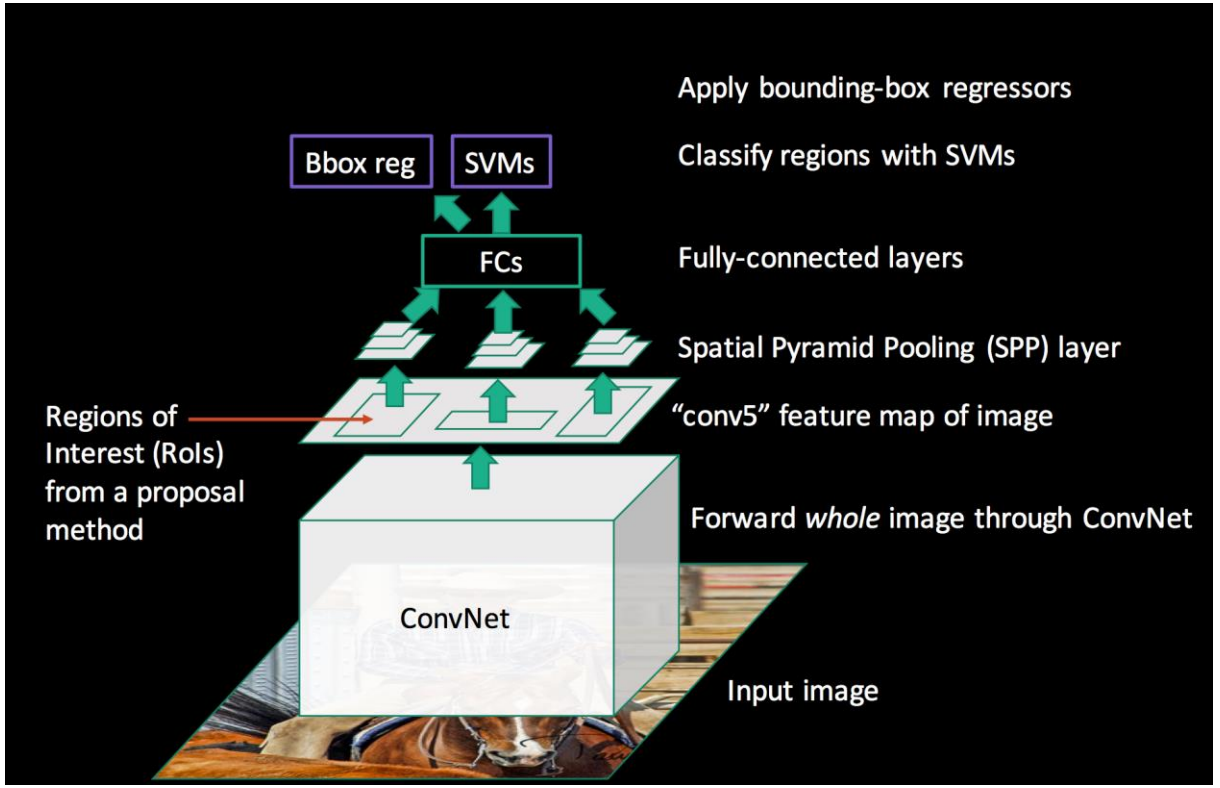
- Uses greedy algorithm to recursively combine similar regions to extract 2000 regions known as region proposals.





2. Spatial Pyramid Pooling (SPP –Net)

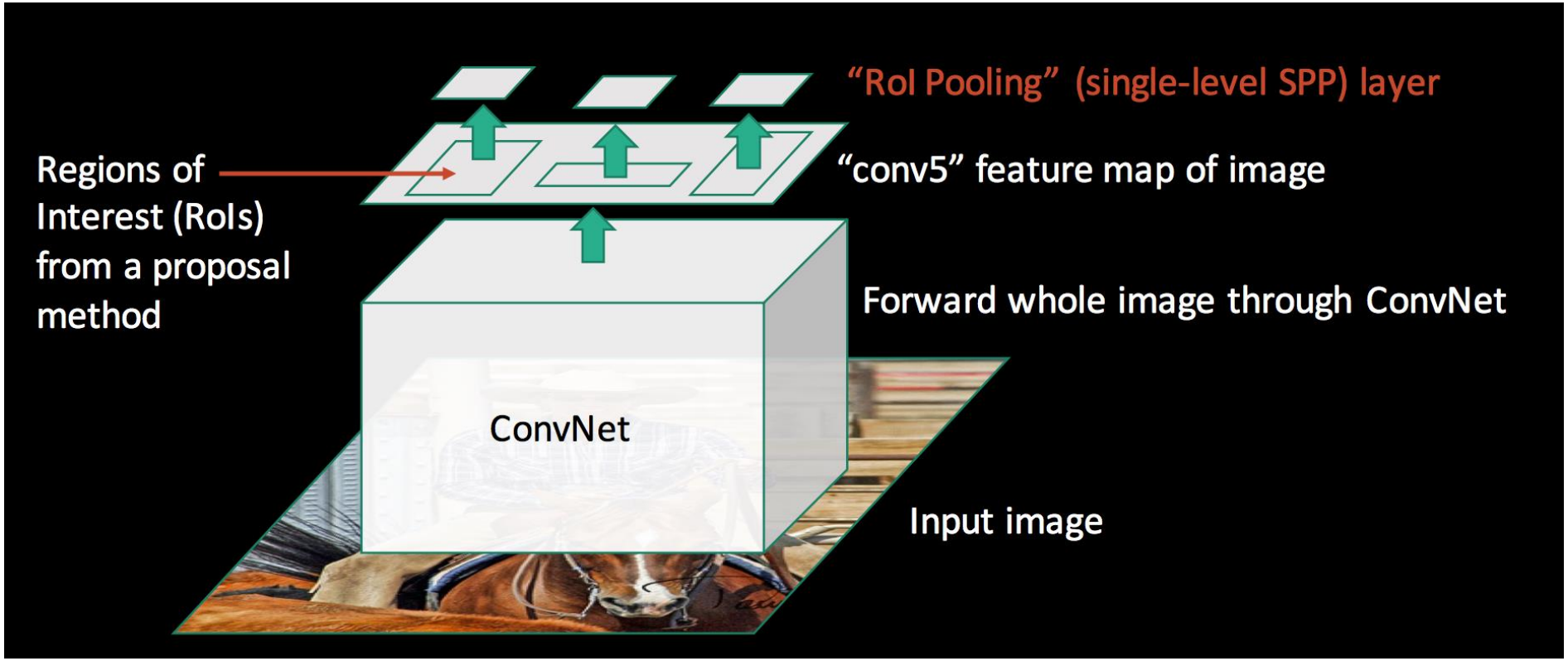
- We feed the input image to the CNN and use the selective search algorithm to generate Region of Interest (RoIs).
- Wrap the RoIs into spatial pyramid pooling (SPP) layers.
- Enables a CNN to generate a fixed-length representation regardless of the size of image/region of interest.



- No need to do convolution each time. Instead convolution is done only once.

3. Fast RCNN

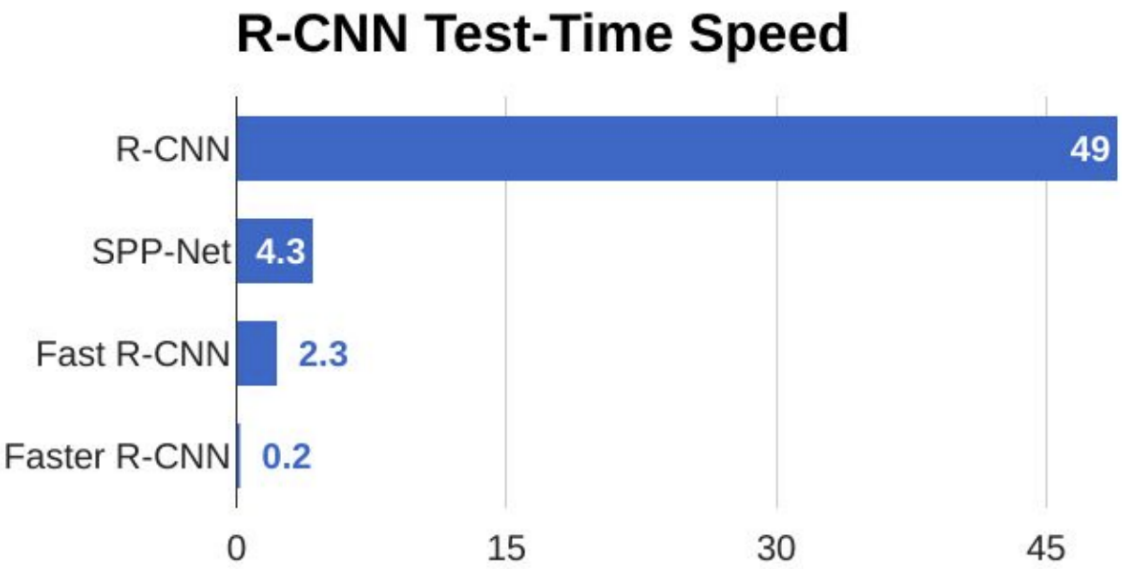
- Instead of generating a pyramid of layers, Fast R-CNN warps ROIs into one single layer using the RoI pooling.

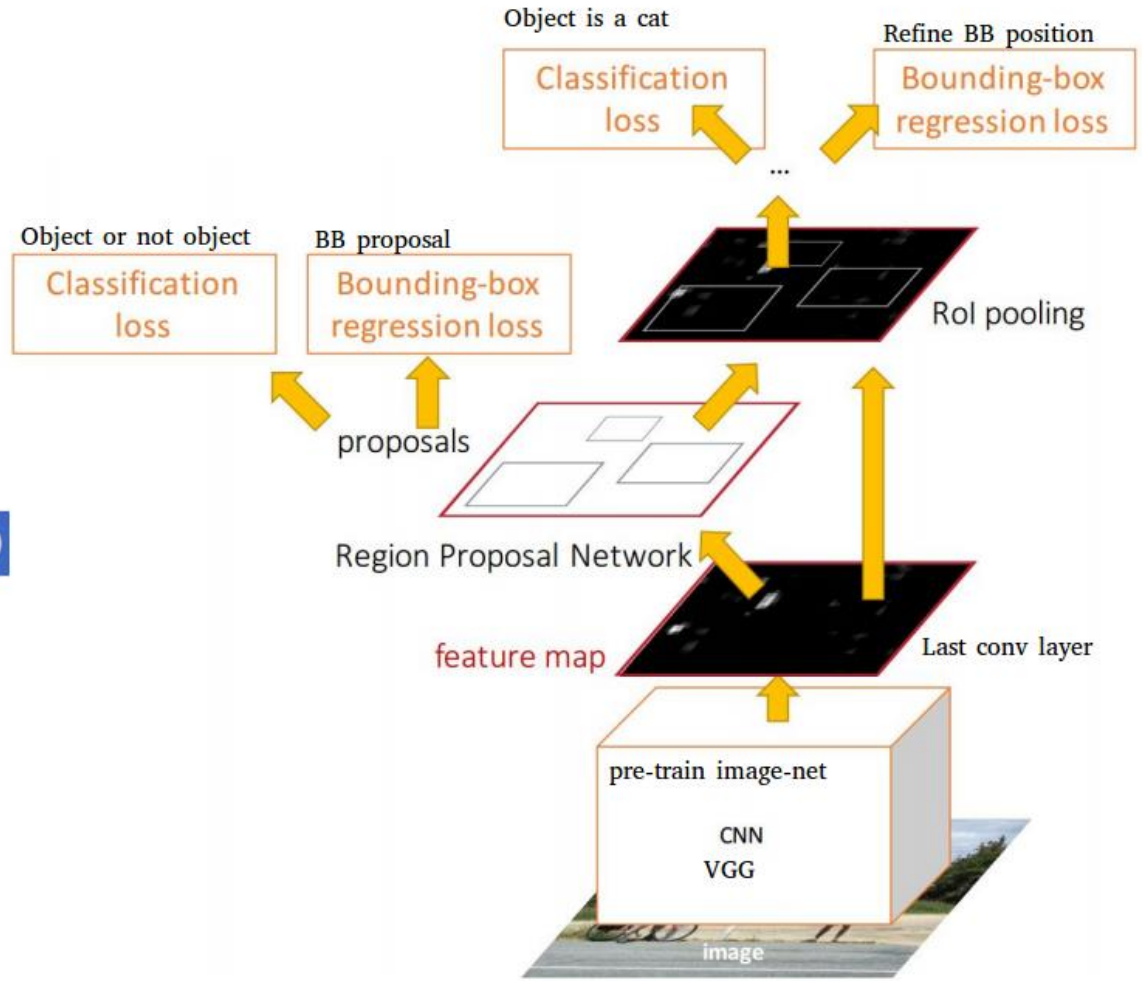


- SPP-net cannot update parameters below SPP layer during training while all parameters of Fast CNN can be trained together.

4. Faster RCNN

Idea: Eliminate selective search and Integrate the Bounding Box Proposals as part of the CNN predictions

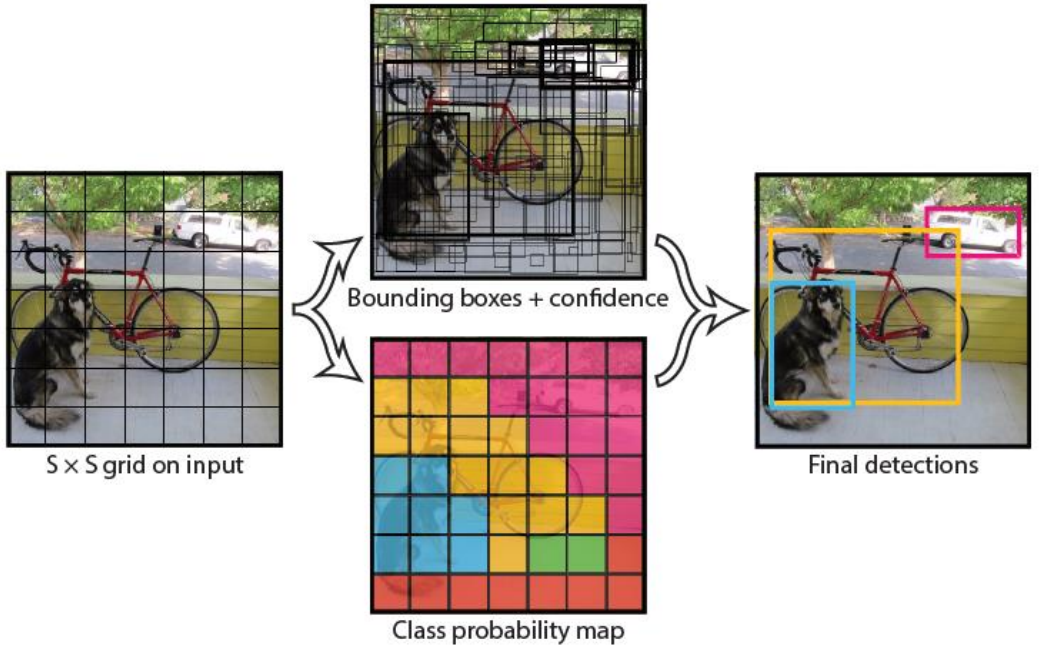




Single Shot Detectors

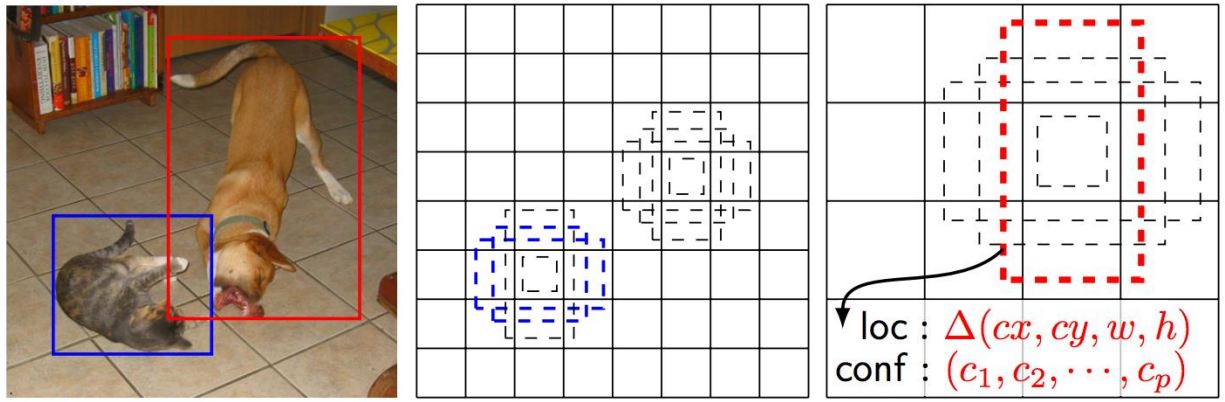
1. You Only Look Once (YOLO):

- No two step region proposals + Classification
- a single CNN simultaneously predicts multiple bounding boxes and class probabilities for those boxes



2. Single Shot MultiBox Detector (SSD):

- YOLO – Increase in detection speed but suffers in localization accuracy.
- Multiresolution detection techniques – Denser grid map + multiscale grid map



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

3. Retina Net: Uses a ResNet + FPN (Feature Pyramid Network)

- Introduced focal loss to address class imbalance problem.
- Penalizes hard negative examples more than easy examples.
- Achieves state-of-the-art performance

All the object detection models discussed so far can predict a fixed set of pre-determined object categories.

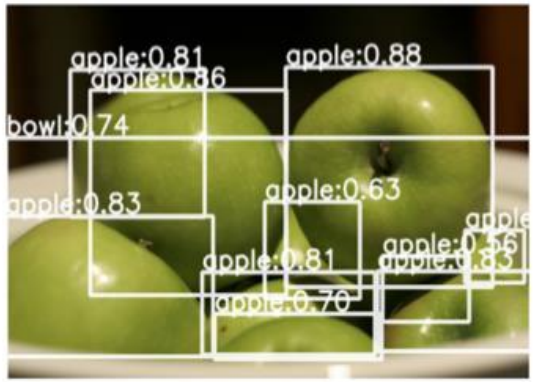
Can object detection models can have zero shot capabilities?

By language supervision

Grounded Language-Image Pre-training (GLIP)

Paper address: https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.pdf

- **Phrase grounding**: Identifying the fine-grained correspondence between phrases in a sentence and objects (or regions) in an image

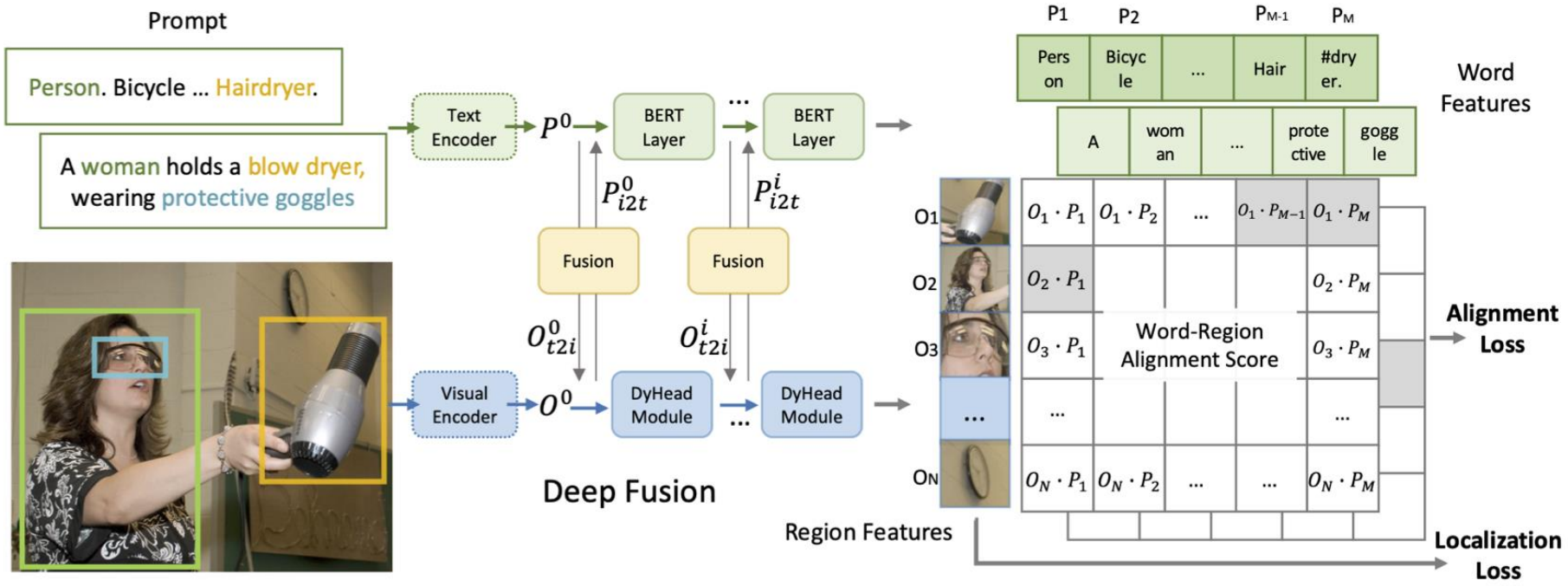


Prompt : person. bicycle. car. motorcycle...

Prompt : pistol

Prompt : there are some holes on the road

GLIP – Model Architecture



- A grounding model replaces the object classification logits with the word alignment scores.
- Dot product of the region (or box) visual features and the token (or phrase) language features.
- Aligns each region/box to phrases in text prompts.

Object detection as Phrase grounding

- Instead of classifying each region/box into c classes.
- Align each region to c phrases in a text prompt

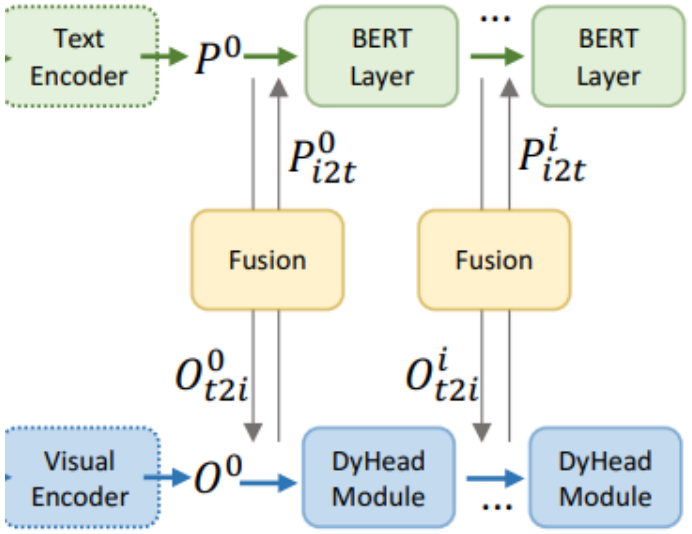
- **How to design a text prompt for a detection task?**

- Given object classes [person, bicycle, car, ..., toothbrush], one simple way is

  Prompt = “Detect: person, bicycle, car, ... , toothbrush”,

Prompt : aerosol can…
lollipop… pendulum…

Language-Aware Deep Fusion

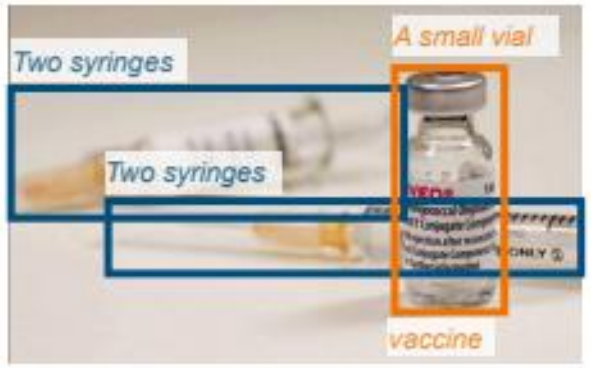


Deep Fusion

To boost the performance of phrase grounding a deep fusion is done to fuse the image and text information before computing the alignment scores at the end

Scalability of training with grounding models

- Grounding data can learn a much larger vocabulary of visual concepts than existing detection data
- Scaling up detection vocabulary – Still no more than 2000 categories
- Grounding models can expand the vocabulary to cover any concepts that appear in the grounded captions
- Due to language supervision, GLIP can learn very rare categories.



Two syringes and a small vial of vaccine.

playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

Training

- Trained in somewhat a self supervised training way.
- Pre-train a teacher GLIP on 3M human-annotated detection and grounding data.
- Use this teacher model to predict boxes and phrases for 24M web-collected image-text data.
- Train a student model on the total of 27M data.

Zero Shot Evaluations

Model	Backbone	Pre-Train Data	Zero-Shot 2017val	Fine-Tune 2017val / test-dev
Faster RCNN	RN50-FPN	-	-	40.2 / -
Faster RCNN	RN101-FPN	-	-	42.0 / -
DyHead-T [9]	Swin-T	-	-	49.7 / -
DyHead-L [9]	Swin-L	-	-	58.4 / 58.7
DyHead-L [9]	Swin-L	O365,ImageNet21K	-	60.3 / 60.6
SoftTeacher [58]	Swin-L	O365,SS-COCO	-	60.7 / 61.3
DyHead-T	Swin-T	O365	43.6	53.3 / -
GLIP-T (A)	Swin-T	O365	42.9	52.9 / -
GLIP-T (B)	Swin-T	O365	44.9	53.8 / -
GLIP-T (C)	Swin-T	O365,GoldG	**46.7**	55.1 / -
GLIP-T	Swin-T	O365,GoldG,Cap4M	46.3	54.9 / -
GLIP-T	Swin-T	O365,GoldG,CC3M,SBU	46.6	**55.2** / -
GLIP-L	Swin-L	FourODs,GoldG,Cap24M	**49.8**	**60.8** / 61.0
GLIP-L	Swin-L	FourODs,GoldG+,COCO	-	- / **61.5**

Zero-shot domain transfer and fine-tuning on **COCO**.
Note: GLIP even outperforms prior supervised models (e.g. GLIP-T under Zero-Shot v.s. Faster RCNN under Fine-Tune)

Model	Backbone	MiniVal [19]				Val v1.0			
		APr	APc	APf	AP	APr	APc	APf	AP
MDETR [19]	RN101	20.9	24.9	24.3	24.2	-	-	-	-
MaskRCNN [19]	RN101	26.3	34.0	33.9	33.3	-	-	-	-
Supervised-RFS [13]	RN50	-	-	-	-	12.3	24.3	32.4	25.4
GLIP-T (A)	Swin-T	14.2	13.9	23.4	18.5	6.0	8.0	19.4	12.3
GLIP-T (B)	Swin-T	13.5	12.8	22.2	17.8	4.2	7.6	18.6	11.3
GLIP-T (C)	Swin-T	17.7	19.5	**31.0**	24.9	7.5	11.6	**26.1**	16.5
GLIP-T	Swin-T	**20.8**	**21.4**	**31.0**	**26.0**	**10.1**	**12.5**	25.5	**17.2**
GLIP-L	Swin-L	**28.2**	**34.3**	**41.5**	**37.3**	**17.1**	**23.3**	**35.4**	**26.9**

Zero-shot domain transfer on **LVIS**
Note: GLIP outperforms strong supervised baselines (shown in gray).

Phrase grounding Evaluations

Row	Model	Data	Val			Test		
			R@1	R@5	R@10	R@1	R@5	R@10
1	MDETR-RN101	GoldG+	82.5	92.9	94.9	83.4	93.5	95.3
2	MDETR-ENB5	GoldG+	83.6	93.4	95.1	84.3	93.9	95.8
3	GLIP-T	GoldG	84.0	95.1	96.8	84.4	95.3	97.0
4		O365,GoldG	84.8	94.9	96.3	85.5	95.4	96.6
5		O365,GoldG,Cap4M	**85.7**	**95.4**	**96.9**	**85.7**	**95.8**	**97.2**
6	GLIP-L	FourODs,GoldG,Cap24M	**86.7**	**96.4**	**97.9**	**87.1**	**96.9**	**98.1**

Note: Phrase grounding performance on Flickr30K

Analysis on Scalability of grounding data

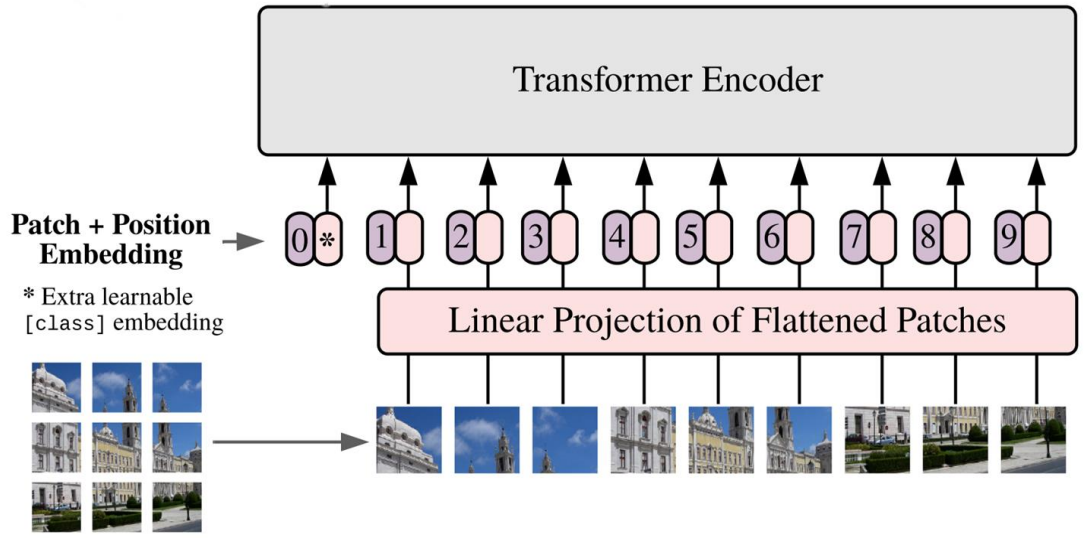
Row	Pre-Training Data	COCO 2017val	LVIS MiniVal			
			AP_r	AP_c	AP_f	AP
1	VG w/o COCO	26.9	4.9	10.4	23.2	16.1
2	+ GoldG	29.2	7.8	14.0	24.5	18.5
3	OpenImages	29.9	12.8	12.1	17.8	14.9
4	+ GoldG	33.6	15.2	16.9	24.5	20.4
5	O365	44.9	13.5	12.8	22.2	17.8
6	+GoldG	**46.7**	17.7	19.5	31.0	24.9
7	O365,GoldG,Cap4M	46.3	**20.8**	21.4	31.0	26.0
8	FourODs	46.3	15.0	**22.5**	**32.8**	**26.8**

- Adding grounding data brings consistent improvement with different detection data (Row 1-6).

Note: The model trained with 2.66M detection data with 1500 categories (Row8) does not match performance with 0.66M detection data and 0.8M grounding data (Row6).

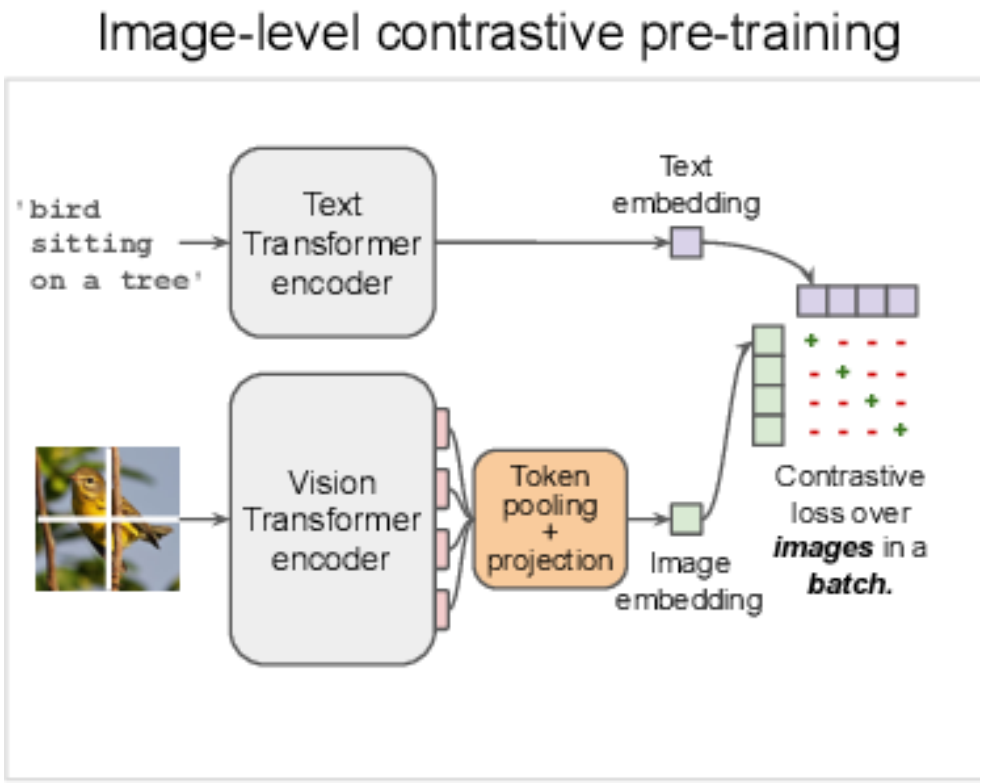
Visual Transformers (ViT) – From Jefferson's Talk

The paper AN IMAGE IS WORTH 16X16 WORDS introduces the main way to tokenize images for transformers, just split then into patches of 16 by 16 pixels and pass then through a linear layer



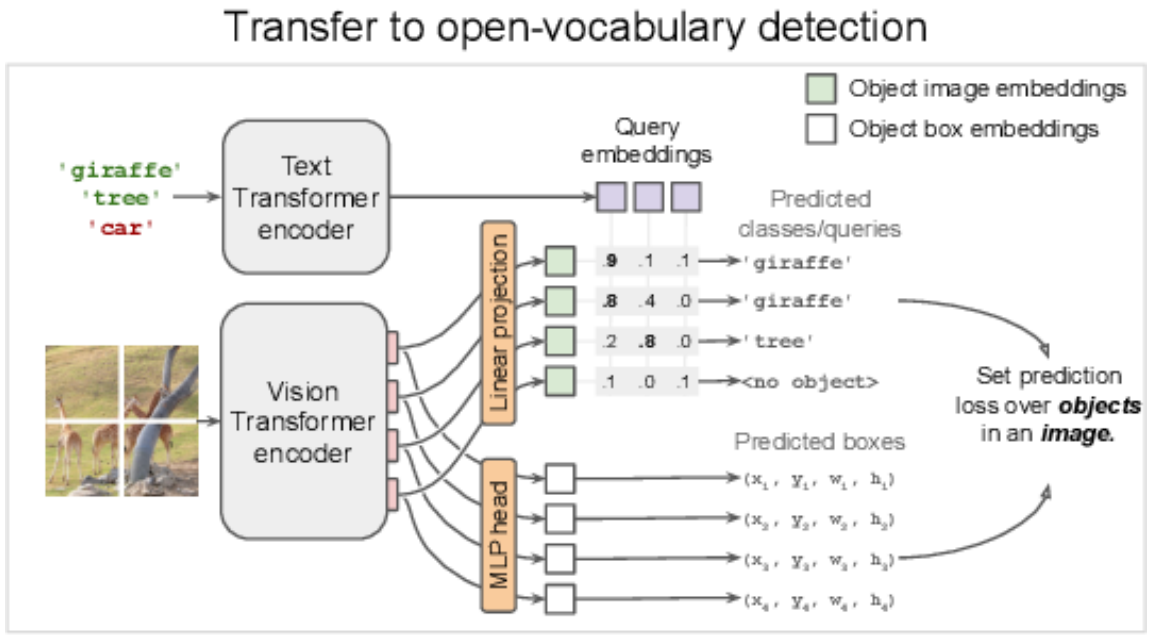
Simple Open-Vocabulary Object Detection with Vision Transformers (OWL- ViT)

Paper address: https://arxiv.org/pdf/2205.06230.pdf



The image and text encoder are pretrained contrastively using image-text pairs, similar to CLIP

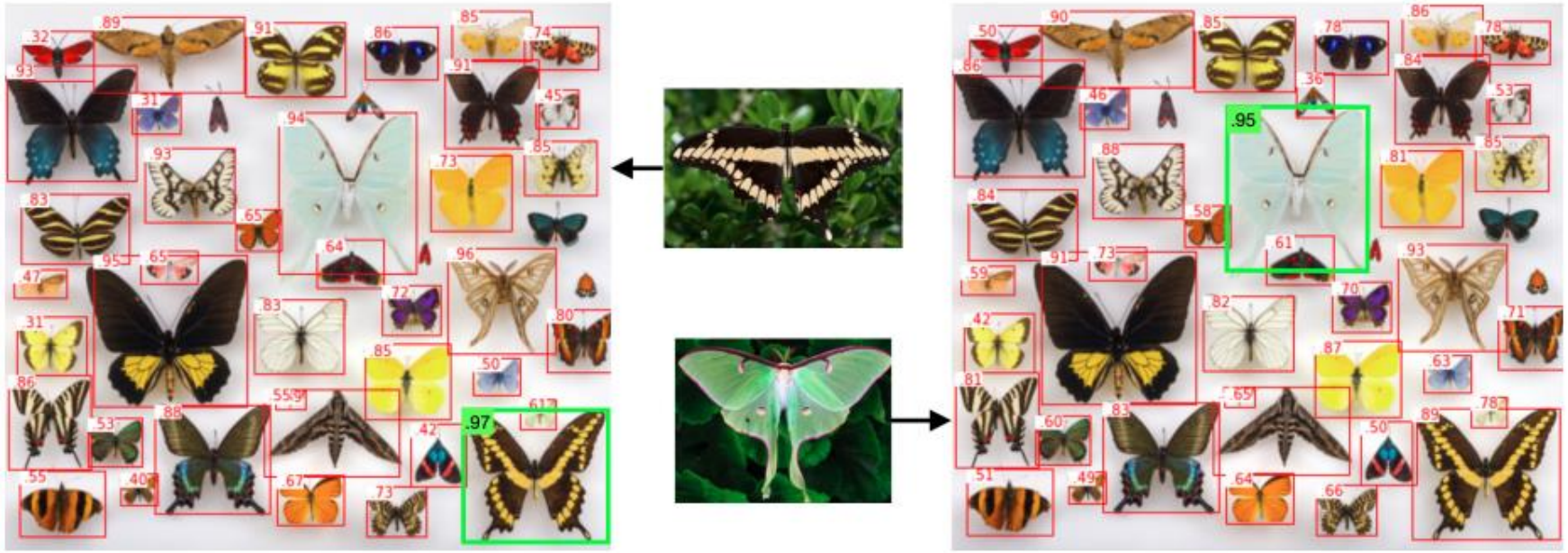
Transfer to object detection



- Image Encoder - Replace token pooling and projection layer and instead linearly project each output token and box coordinates.
- Text Encoder – We pass queries which are class names or other textual object descriptions
- Output – For each object the model predicts a bounding box and a probability with which each query applies to the object.

One-Shot Image conditioned transfer.

- The model does not require query embeddings to be of textual origin.
- We can supply image- instead of text-derived embeddings as queries to the classification head without modifying the model.
- This is called image-conditioned one-shot object detection one-shot object detection because the query image is essentially a single training example.



Example of one-shot image-conditioned detection. Images in the middle are used as queries; the respective detections on the target image are shown on the left and right. text-based querying detects the correct species only for the top example (“swallowtail butterfly”) but not for the bottom (“luna moth”).

Training and model design

- Image encoder - ViT B/32 – ViT Base with patch size 32
- Text encoder - Transformer architecture similar to the image model

Contrastive Pre-Training

- The image and text encoder are pretrained contrastively using 3.6 billion image –text pairs.
- Image representation - Multihead attention pooling (MAP) to aggregate token representation
- Text representation - The final end-of sequence (EOS) token of the text encoder

Training the detector

- The pre-trained model is fine tuned for object detection
- Token pooling is removed and detection heads are added.
- Since not all object categories are found in every image, the queries provide both positive (present) and negative (known to be absent) annotations for each image.
- Publicly available detection datasets were used for object-level fine tuning, with a total of around 2 million images (OpenImages V4 , Objects 365 (O365) , and and/or Visual Genome (VG))

Open vocabulary detection performance

	Method	Backbone	Image-level	Object-level	Res.	AP^{LVIS}	AP^{LVIS}_{rare}
LVIS base training:							
1	ViLD-ens [12]	ResNet50	CLIP	LVIS base	1024	25.5	16.6
2	ViLD-ens [12]	EffNet-b7	ALIGN	LVIS base	1024	29.3	26.3
3	Reg. CLIP [45]	R50-C4	CC3M	LVIS base	?	28.2	17.1
4	Reg. CLIP [45]	R50x4-C4	CC3M	LVIS base	?	32.3	22.0
5	OWL-ViT (ours)	ViT-H/14	LiT	LVIS base	840	35.3	23.3
6	OWL-ViT (ours)	ViT-L/14	CLIP	LVIS base	840	34.7	25.6
Unrestricted open-vocabulary training:							
7	GLIP [26]	Swin-T	Cap4M	O365, GoldG, ...	?	17.2	10.1
8	GLIP [26]	Swin-L	CC12M, SBU	OI, O365, VG, ...	?	26.9	17.1
9	OWL-ViT (ours)	ViT-B/32	LiT	O365, VG	768	23.3	19.7
11	OWL-ViT (ours)	R26+B/32	LiT	O365, VG	768	25.7	21.6
10	OWL-ViT (ours)	ViT-B/16	LiT	O365, VG	768	26.7	23.6
12	OWL-ViT (ours)	ViT-L/16	LiT	O365, VG	768	30.9	28.8
13	OWL-ViT (ours)	ViT-H/14	LiT	O365, VG	840	33.6	30.6
14	OWL-ViT (ours)	ViT-B/32	CLIP	O365, VG	768	22.1	18.9
15	OWL-ViT (ours)	ViT-B/16	CLIP	O365, VG	768	27.2	20.6
16	OWL-ViT (ours)	ViT-L/14	CLIP	O365, VG	840	34.6	31.2

Open-vocabulary and zero-shot performance on LVIS v1.0 val.

Method	Backbone	Image-level	Object-level	Res.	AP^{COCO}	$AP50^{COCO}$	AP^{O365}	$AP50^{O365}$
ViLD [12]	ResNet50	CLIP	LVIS base	1024	36.6	55.6	11.8	18.2
Reg. CLIP [45]	R50-C4	CC3M	COCO base	?	-	50.4	-	-
Reg. CLIP [45]	R50x4-C4	CC3M	COCO base	?	-	55.7	-	-
GLIP [26]	Swin-T	Cap4M	O365, GoldG, ...	?	46.7	-	-	-
GLIP [26]	Swin-L	CC12M, SBU	OI, O365, VG, ...	?	49.8	-	-	-
Detic [46]	R50-C4	CLIP, COCO-Cap	COCO base	1333	-	45.0	-	-
Detic [46]	Swin-B	CLIP, I21K	LVIS base	869	-	-	21.5	-
OWL-ViT (ours)	ViT-B/32	CLIP	OI, VG	768	28.1	44.7	-	-
OWL-ViT (ours)	ViT-B/16	CLIP	OI, VG	768	31.7	49.2	-	-
OWL-ViT (ours)	ViT-L/14	CLIP	O365, VG	840	43.5	64.7	-	-
OWL-ViT (ours)	ViT-B/32	LiT	OI, VG	768	28.0	44.4	9.4	15.2
OWL-ViT (ours)	ViT-B/16	LiT	OI, VG	768	30.3	47.4	10.7	17.0
OWL-ViT (ours)	R26+B/32	LiT	OI, VG	768	30.7	47.2	11.1	17.4
OWL-ViT (ours)	ViT-L/16	LiT	OI, VG	672	34.7	53.9	13.7	21.6
OWL-ViT (ours)	ViT-H/14	LiT	OI, VG	840	36.0	55.3	15.5	24.0
OWL-ViT (ours)	ViT-H/14	LiT	O365, VG	840	42.2	64.5	-	-

Open-vocabulary and zero-shot performance on COCO and O365 datasets.

Image-Conditioned Detection performance

	Method	Split 1	Split 2	Split 3	Split 4	Mean
Seen	SiamMask [30]	38.9	37.1	37.8	36.6	37.6
	CoAE [16]	42.2	40.2	39.9	41.3	40.9
	AIT [7]	**50.1**	47.2	45.8	46.9	47.5
	OWL-ViT (ours)	49.9	**49.1**	**49.2**	**48.2**	**49.1**
	OWL-ViT ($k = 10$; ours)	54.1	55.3	56.2	54.9	55.1
Unseen	SiamMask [30]	15.3	17.6	17.4	17.0	16.8
	CoAE [16]	23.4	23.6	20.5	20.4	22.0
	AIT [7]	26.0	26.4	22.3	22.6	24.3
	OWL-ViT (ours)	**43.6**	**41.3**	**40.2**	**41.9**	**41.8**
	OWL-ViT ($k = 10$; ours)	49.3	51.1	42.4	44.5	46.8

One- and few-shot image-conditioned detection performance on COCO AP50.
Note: Note the improvements as the number of conditioning queries is increased to k = 10.

Summary

Similarities

- Both the models are based on visual grounding
- Both the models are highly capable of scaling up detection vocabulary.

Differences

- GLIP uses region proposal method to extract region features of the objects while OWL- ViT divides the image into image embeddings.
- GLIP is only one level of pretraining (image level) while OWL- ViT has two levels of pretraining (image level and detection level).

Questions?

Thank you!