INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

LaMI-DETR: Open-Vocabulary Detection with Language Model Instruction(ECCV 2024)

Undergraduate Researcher at CVLab

Lee Dohyeong

2025.5.7

Contents

Introduction

- VLM (Vision-Language Model)
- Open-Vocabulary

- Zero-shot vs Open-Vocabulary

Introduction

- T5(Text-To-Text Transfer Transformer)
- CLIP(Contrastive Language–Image Pretraining)

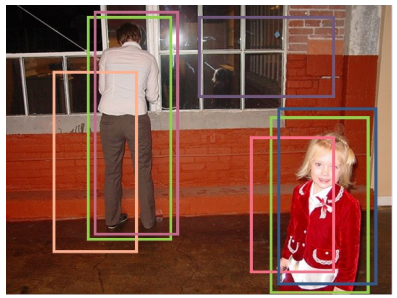
  Advantages : Competitive Performance

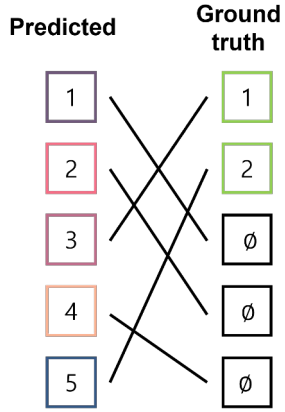
  1. Bipartite(이분) Matching : Find the optimal one-to-one correspondence between two sets
  2. Transformer : Capture the overall context of the image

  -> Accuracy and performance level similar to Faster R-CNN.

Related Work

- Bipartite Matching(이분매칭)
  - 1:1 matching, Slove set prediction problem
    - Set prediction problem : Predicting sets without imposing any order.





	1	2	ø	ø	ø
1	12	11	1	1	1
2	4	2	8	5	9
3	1	3	5	7	8
4	2	5	6	7	4
5	2	1	9	10	6

Permutation = [3, 4, 1, 5, 2]
Matching score = 1 + 5 + 1 + 4 + 1 = 12

출력 개수 고정: $N = 6$

예측 결과

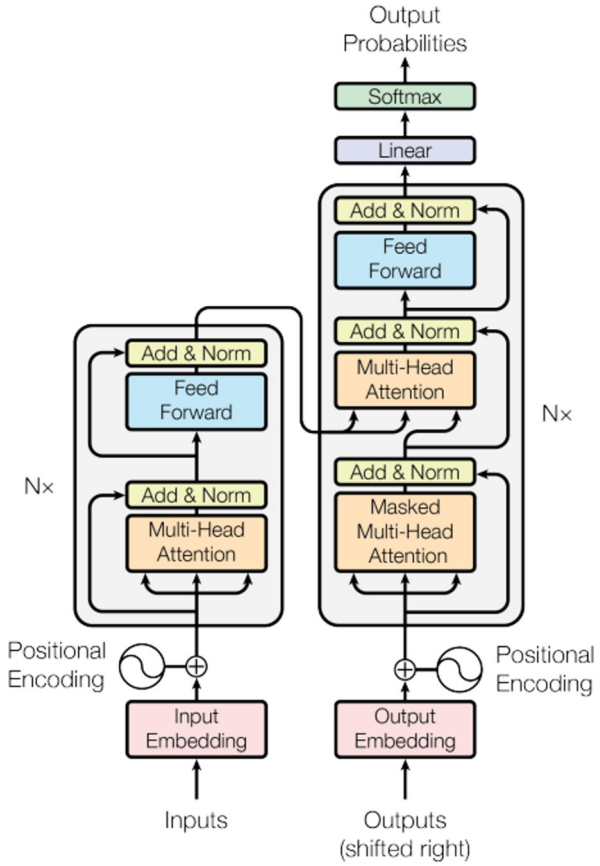
$(c_0 = \emptyset, b_0)$
$(c_1 = bird, b_1 = (180, 180, 150, 240))$
$(c_2 = \emptyset, b_2)$
$(c_3 = bird, b_3 = (120, 150, 100, 150))$
$(c_4 = \emptyset, b_4)$
$(c_5 = dog, b_5)$

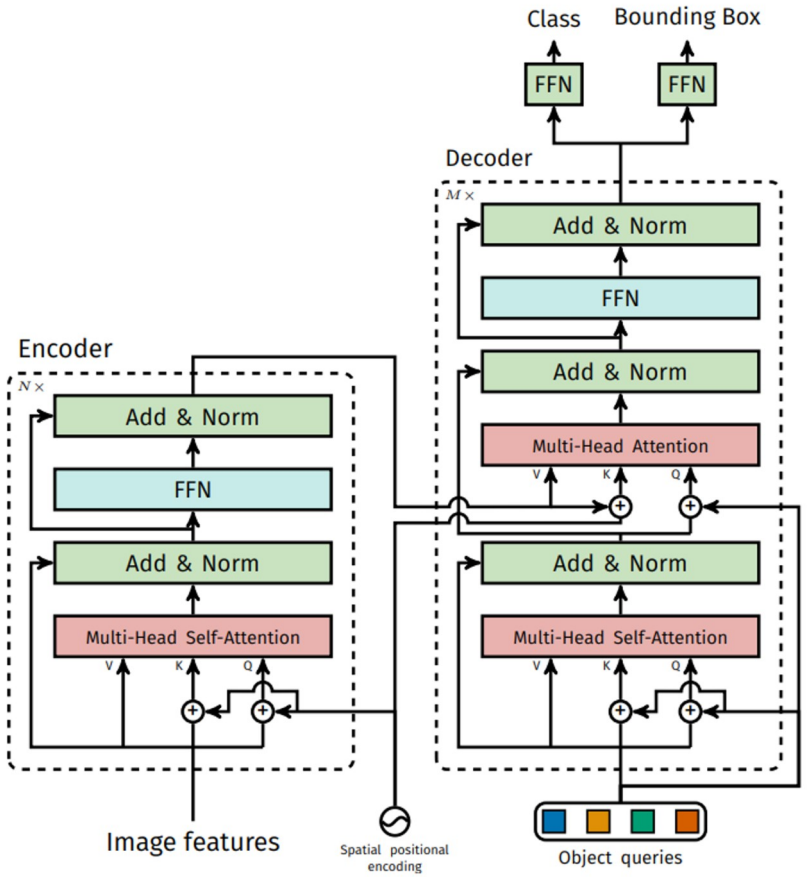
$(c_0 = bird, b_0 = (122, 151, 100, 150))$
$(c_1 = bird, b_1 = (182, 180, 148, 238))$
$(c_2 = \emptyset, b_2)$
$(c_3 = \emptyset, b_3)$
$(c_4 = \emptyset, b_4)$
$(c_5 = \emptyset, b_5)$

실제 값

Related Work

- Transformer

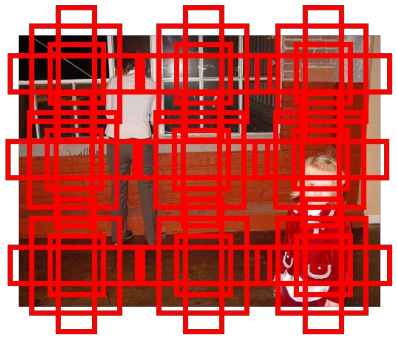




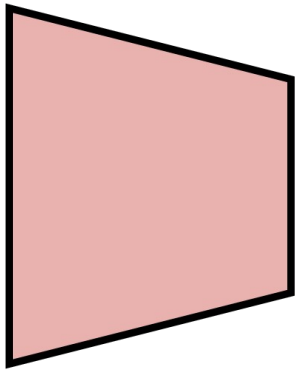
항목	기본 Transformer	DETR(Detection Transformer)
입력 데이터	단어(텍스트 시퀀스)	이미지 (CNN을 통한 Feature Map)
출력 데이터	다음 단어(확률 분포)	객체의 클래스 및 위치(BBox)
Encoder 역할	문장의 의미 표현(컨텍스트 학습)	이미지 전체에서 객체의 특징 학습
Decoder 역할	단어 예측 (단어 순서 중요)	객체 위치 및 클래스 예측 (순서 불필요)
Self-Attention 방식	단어 간의 관계 학습	픽셀(Feature) 간의 관계 학습
Positional Encoding	단어 순서 표현	이미지의 공간적 정보 표현
출력 방식	Auto-Regressive(순차 예측)	Parallel(한 번에 전체 객체 예측)

Related Work

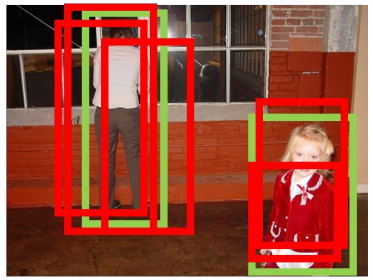
Input image



Pre-defined anchors

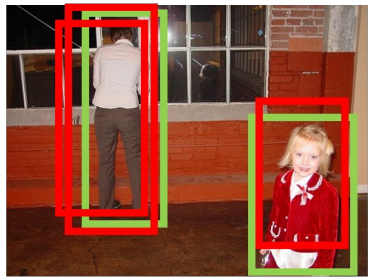


Object detetor



Predicted bboxes

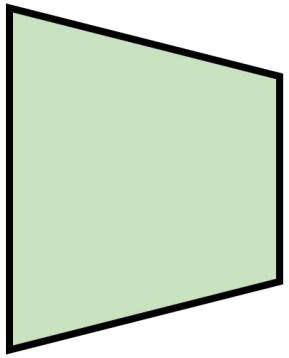
NMS



Post-prcessed bboxes

End-to-end

Input image



DETR

Predicted bboxes

Related Work

- Existing object detection methods
  - Remove require prior knowledge -> end to end :

- Expected shape of the bounding box (example : Train -> the box need to be long)

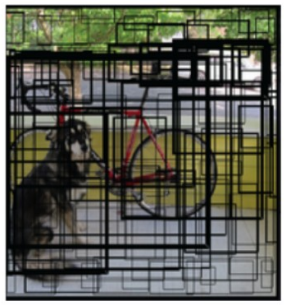
- Predefined rules for handling overlapping bounding boxes

  - Removes NMS and anchor generation -> end to end :

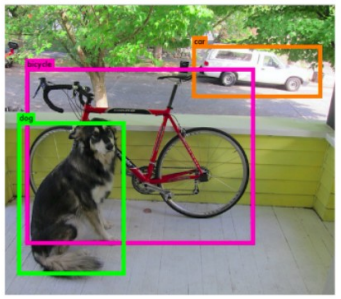
NMS(non-maximum suppression) : Eliminates duplicate bounding boxes

Anchor generation : Defines bounding box anchors in advance (such as Faster R-

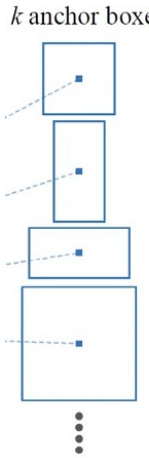
(



Multiple Bounding Boxes

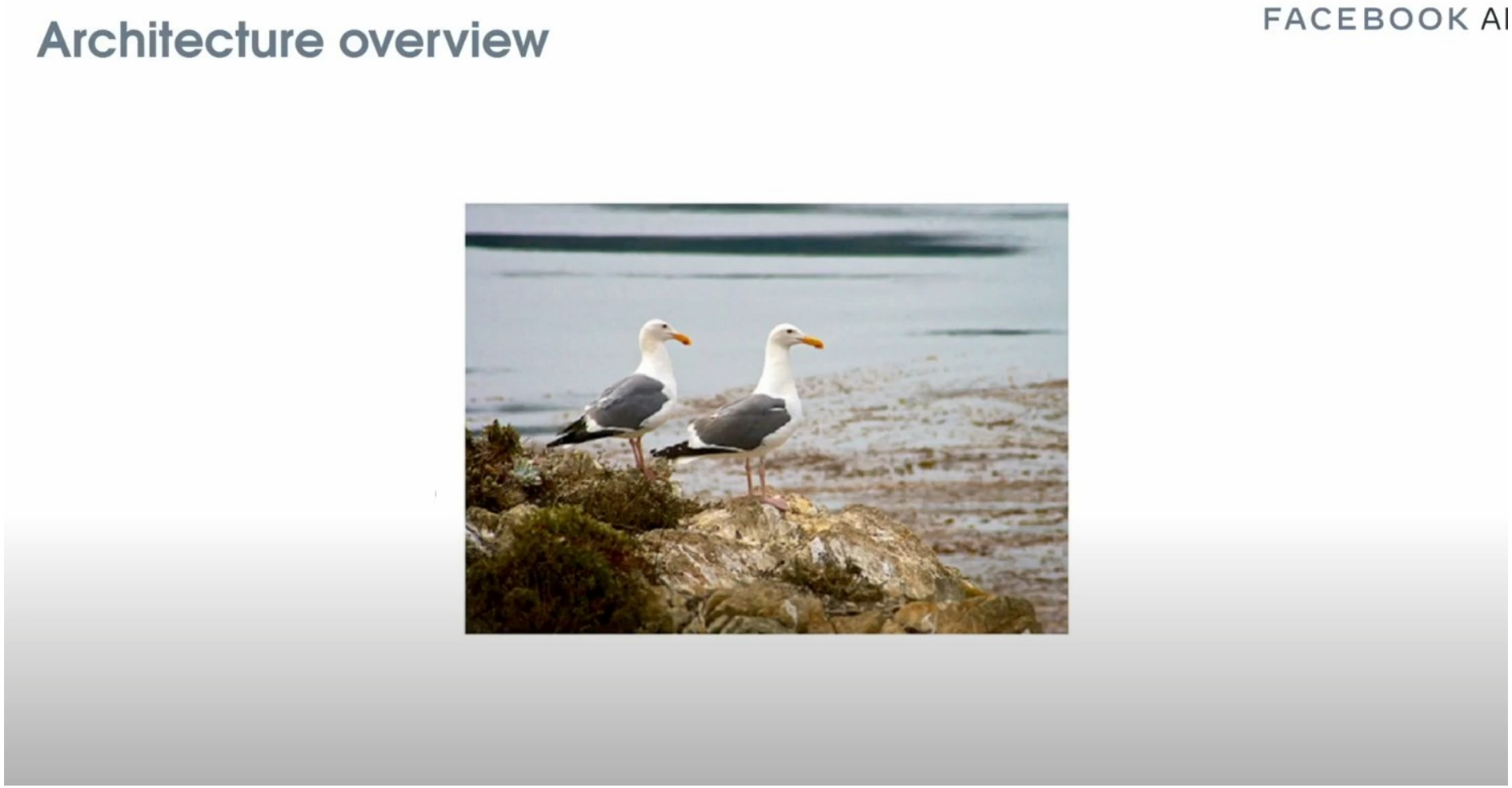


Final Bounding Boxes



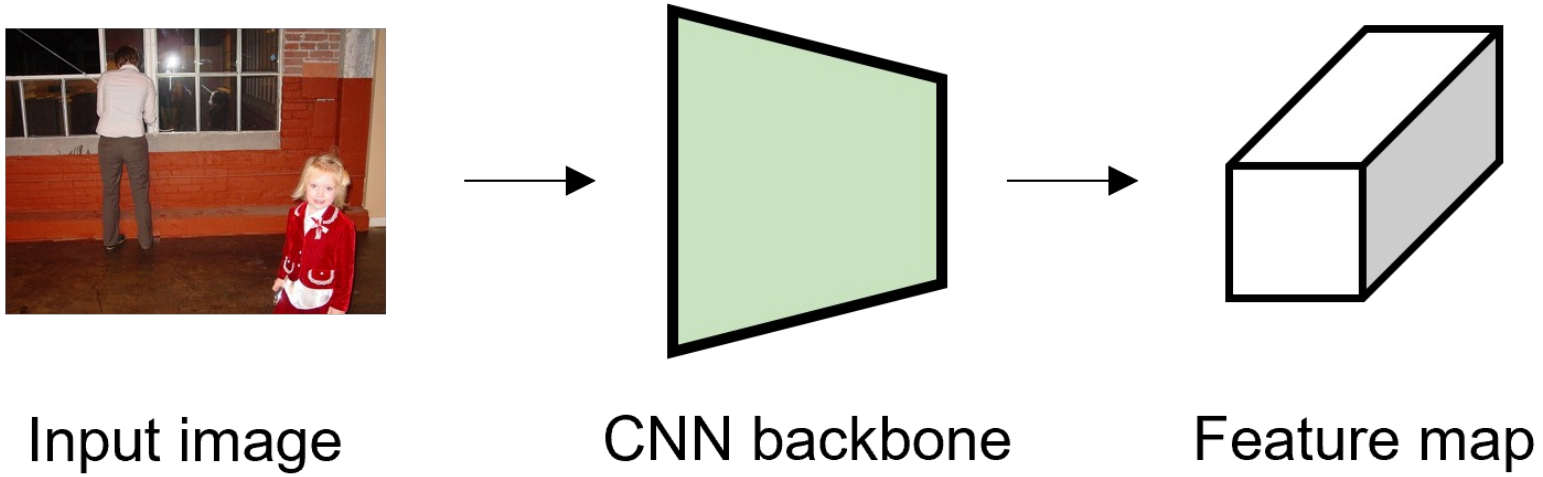
Method

- Architecture



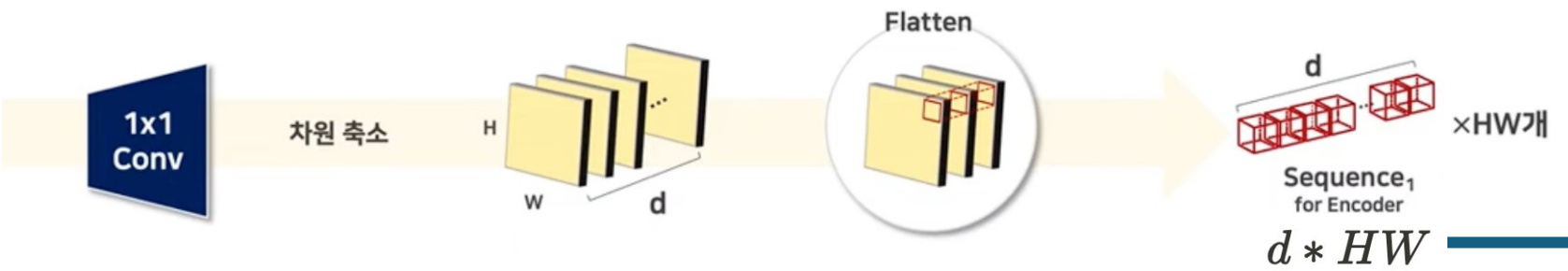
Method

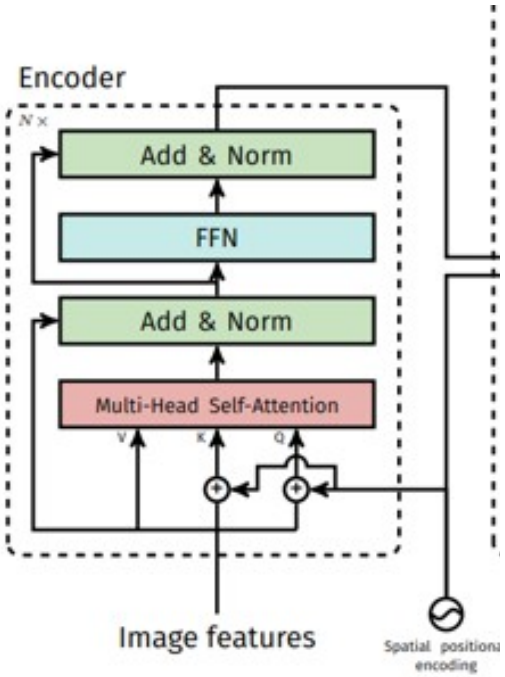
- Extract feature map by CNN backbone
  - **Input** : image $x_{img} \in \mathbb{R}^{3\times H\times W}$
  - **Output** : feature map $f \in \mathbb{R}^{C\times h\times w}$ $\quad C = 2048, H, W = \frac{H_0}{32}, \frac{W_0}{32}$



Method

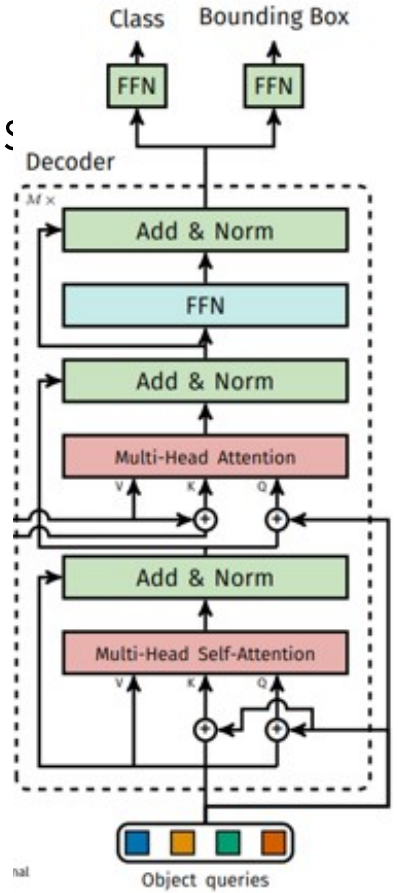
- Transformer Encoder
  - Before Encoder : 1 * 1 Convolutior → C(2048) → d(256) → $z_0 \in \mathbb{R}^{d \times H \times W}$ → $d * HW$
  - Positional Encoding :
    - Preserve spatial location information, Adding learnable position embeddings
  - Multi-Head Self-Attention (MHSA)
    - Correlation between different pixels
  - FFN(Feed Forward Network) : FC Layer + ReLU (MLP)
    - Maximizing the model's representational





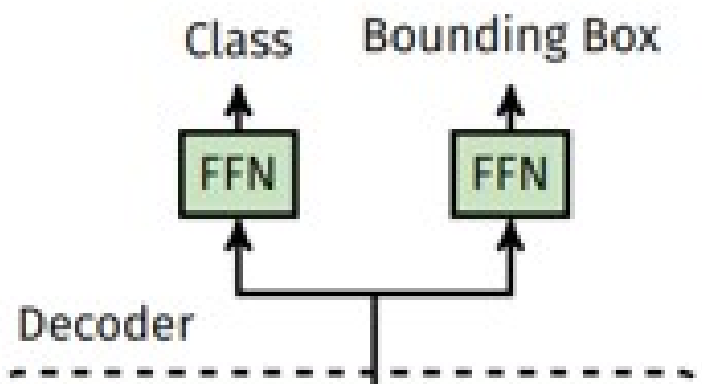
Method

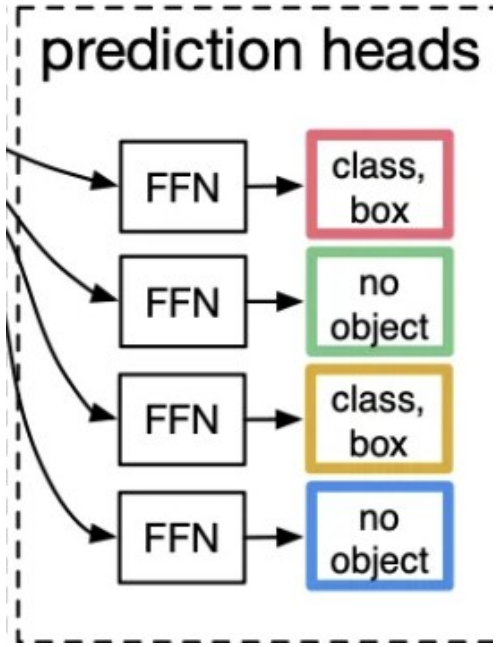
- Transformer Decoder
  - Self-Attention : Distributes role, learn optimal one-to-one matching
  - Encoder-Decoder Attention: Determines which parts of the image to focus
  - Object Queries
    1. Random vectors with no meaningful information.
    2. Update Object queries through self-attention
    3. Query can detect objects in which location
    4. FFN: Assists training and normalizes the output.
  - Decoder output : Object Query final embedding vector.
    - Feature vector representing N object candidates



Method

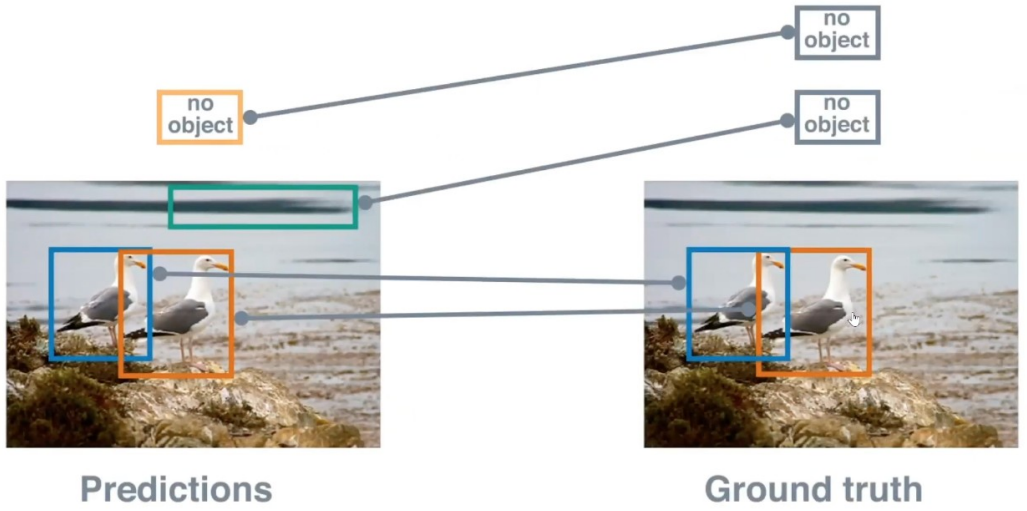
- FFNs(Prediction Feed-Forward Networks)
  - Class : Softamax + No object class
  - Bounding Box : Sigmoid + localization(0~1)





Method

- DETR
  - Direct Set Prediction :
    - Detect multiple objects without duplicates, using a fixed set of predictions.
    - Order does not matter
    - Determine how each predicted box should be matched to a ground-truth object.



Method

- Hungarian Algorithm

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

$$\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) = -\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)})$$

$$-\log \hat{p}_{\sigma(i)}(c_i)$$

클래스 예측 Cost — 박스 좌표 예측 Cost

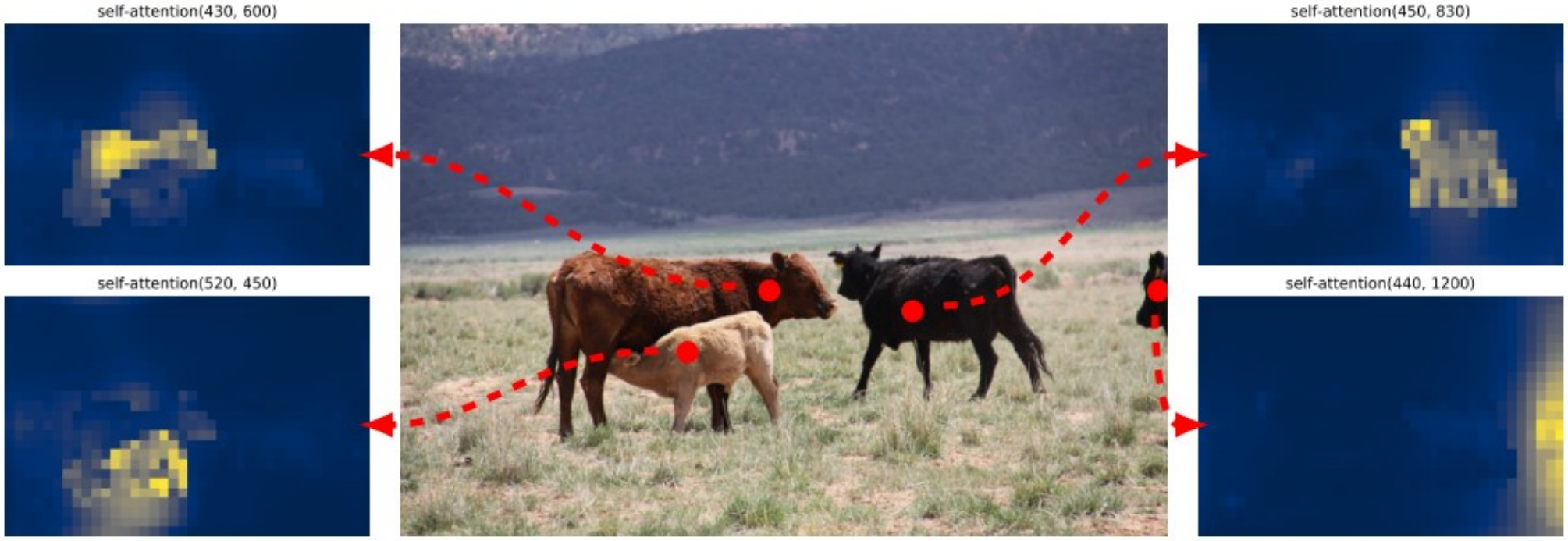
$\hat{p}_{\sigma(i)}(c_i)$ 순열 σ의 i번째 요소가 해당하는 GT의 클래스를 예측한 확률

$\mathcal{L}_{box}(b_i, \hat{b}_{\sigma(i)})$ 순열 σ의 i번째 요소의 예측 박스 좌표와 해당하는 GT의 박스 좌표 간 Loss

$$\mathcal{L}_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{N} [-log\hat{p}_{\hat{\sigma}(i)}(c_i) + 1_{\{c_i \neq \varnothing\}} \mathcal{L}_{box}(b_i, \hat{b}_{\hat{\sigma}}(i))]$$

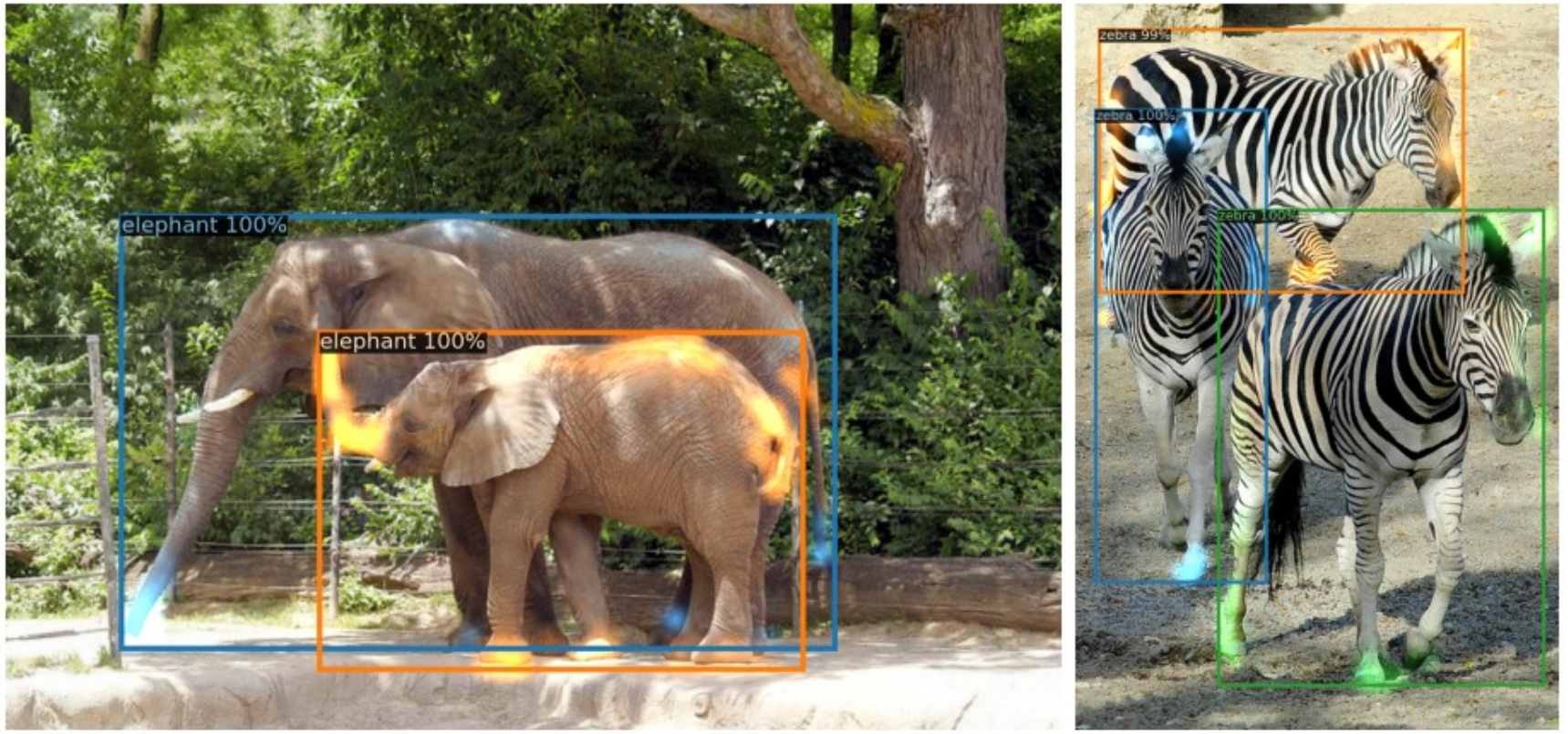
Experiments

- Attention Map(After Encoder)
  - The encoder already partially separates objects.



Experiments

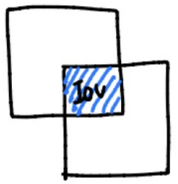
- Decoder Attention Visualization
  - High attention scores appear near object extremities.



Experiments

- Comparison with Faster R-CNN
  - GFLOPS : 부동소수점 연산량
  - FPS : 1초에 처리할 수 있는 프레임 수
  - AP = mAP : 평균 정밀도

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

Experiments

- Encoder Size
  - Choose an appropriate number of layers.

#layers	GFLOPS/FPS	#params	AP	AP_{50}	AP_S	AP_M	AP_L
0	76/28	33.4M	36.7	57.4	16.8	39.6	54.2
3	81/25	37.4M	40.1	60.6	18.5	43.8	58.6
6	86/23	41.3M	40.6	61.6	19.9	44.3	60.2
12	95/20	49.2M	41.6	62.1	19.8	44.9	61.9

Conclusion

- Object Detection -> Direct set prediction Task
  - Remove NMS or Anchor
  - End to end pipeline
- Transformer and bipartite matching
  - Predict all objects at once
- Superior results on large objects