

꼼꼼한 딥러닝 논문 리뷰와 코드 실습

Deep Learning Paper Review and Code Practice

나동빈(dongbinna@postech.ac.kr)

Pohang University of Science and Technology

DETR: End-to-End Object Detection with Transformers (ECCV 2020)

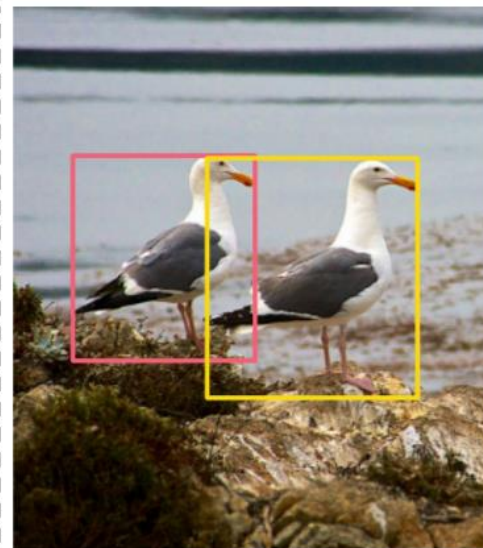
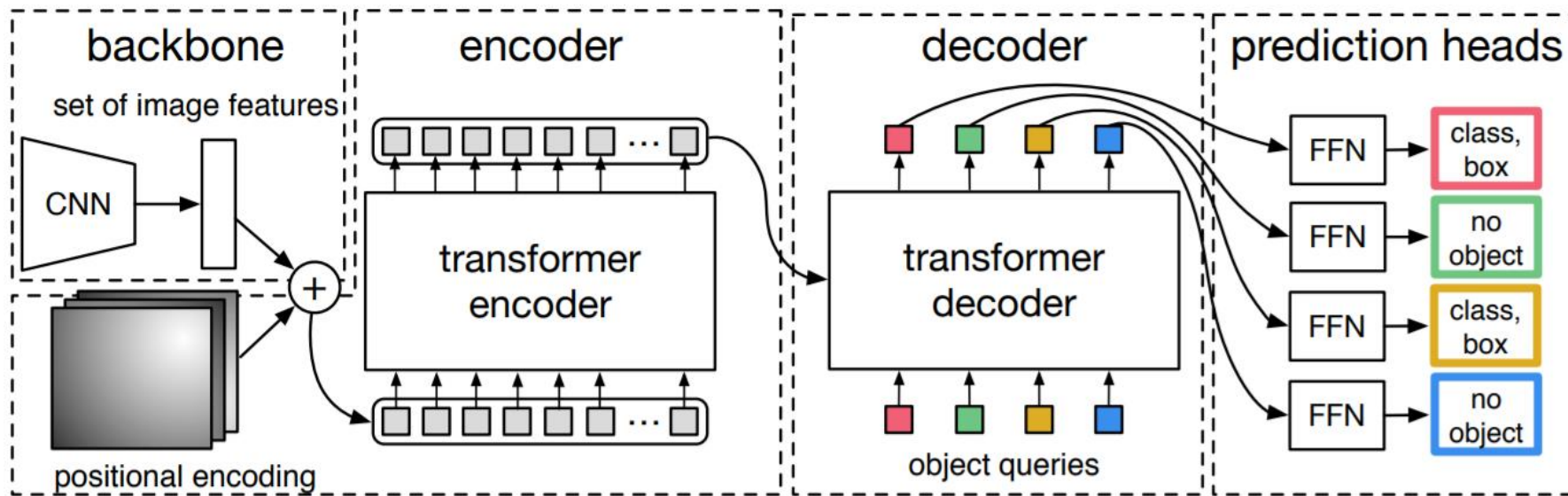
- 이 논문의 장점은 무엇인가요?

기존의 객체 탐지(object detection) 기술과 비교했을 때

매우 **간단하며** 또한 **경쟁력 있는 성능**을 보입니다.

- 이 논문은 무엇을 제안했나요?

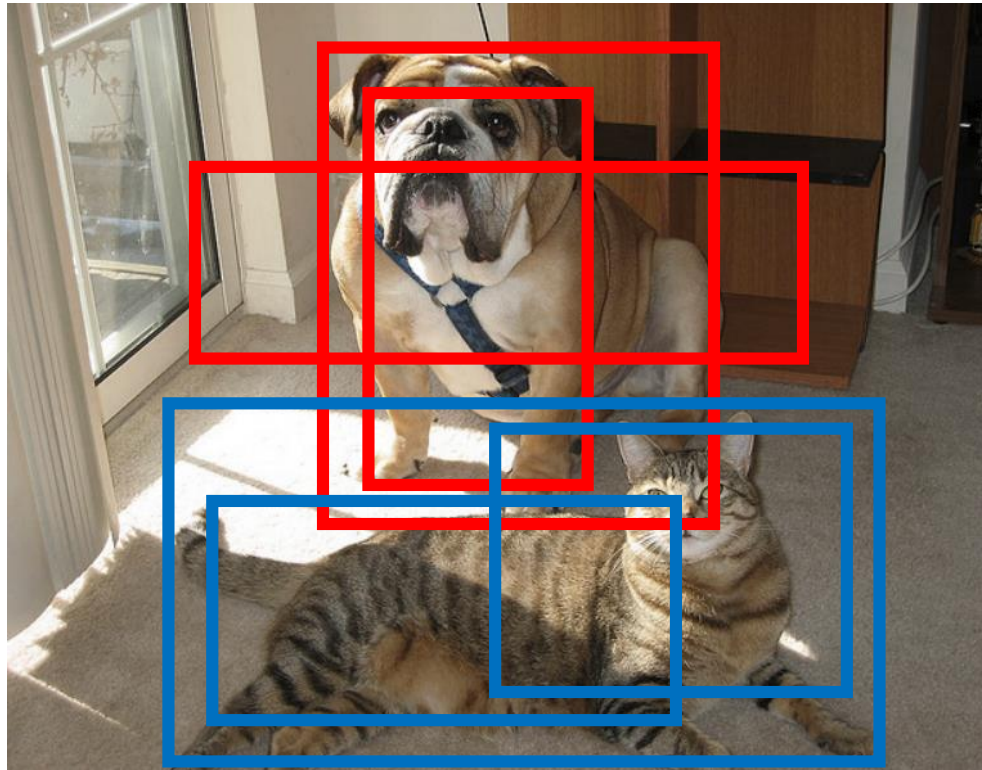
- DETR (DEtection TRansformer):** ① 이분 매칭 손실 함수 + ② Transformer



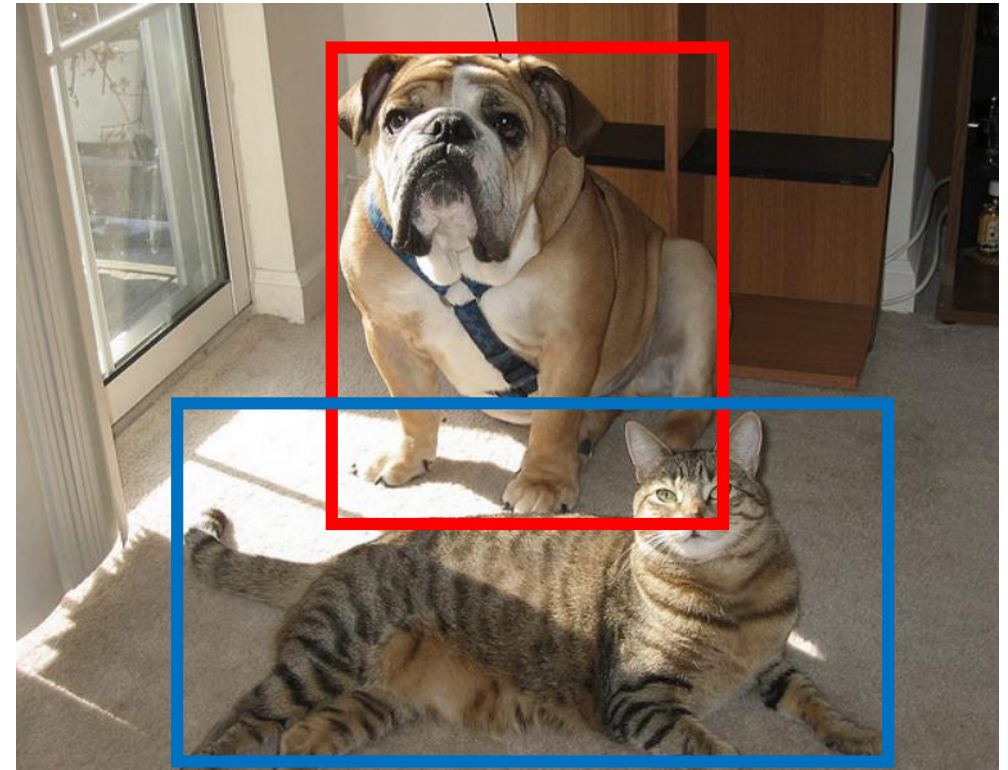
연구 배경: 기존 방법의 문제점

- 기존 객체 탐지(object detection) 방법들은 너무 **복잡**하며 다양한 라이브러리를 활용합니다.
 - 사전 지식(prior knowledge) 요구: bounding box의 형태, bounding box가 겹칠 때의 처리 방법, ...

■ : dog ■ : cat

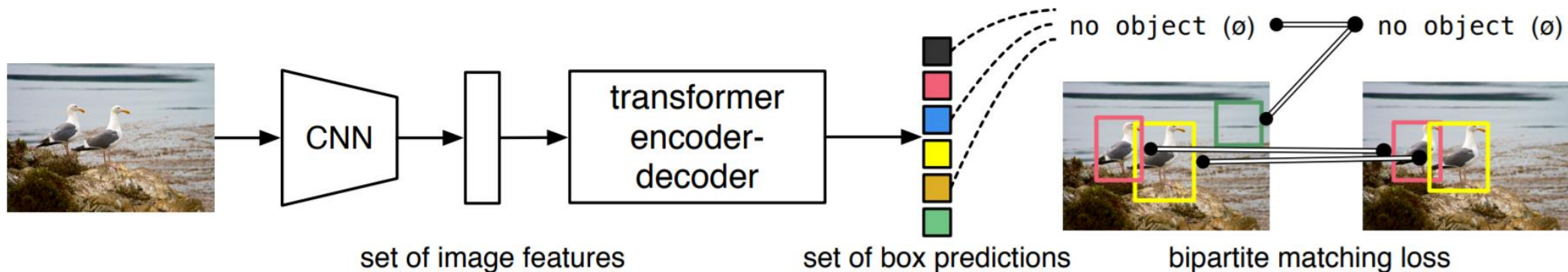


NMS



본 논문의 핵심 아이디어: 이분 매칭

- 이분 매칭(bipartite matching)을 통해 set prediction problem을 직접적으로(directly) 해결합니다.



- 학습 과정에서 이분 매칭을 수행함으로써 인스턴스가 중복되지 않도록 유도합니다.

출력 개수 고정: $N = 6$

예측 결과

$(c_0 = \emptyset, b_0)$

$(c_1 = bird, b_1 = (180, 180, 150, 240))$

$(c_2 = \emptyset, b_2)$

$(c_3 = bird, b_3 = (120, 150, 100, 150))$

$(c_4 = \emptyset, b_4)$

$(c_5 = dog, b_5)$



$(c_0 = bird, b_0 = (122, 151, 100, 150))$

$(c_1 = bird, b_1 = (182, 180, 148, 238))$

$(c_2 = \emptyset, b_2)$

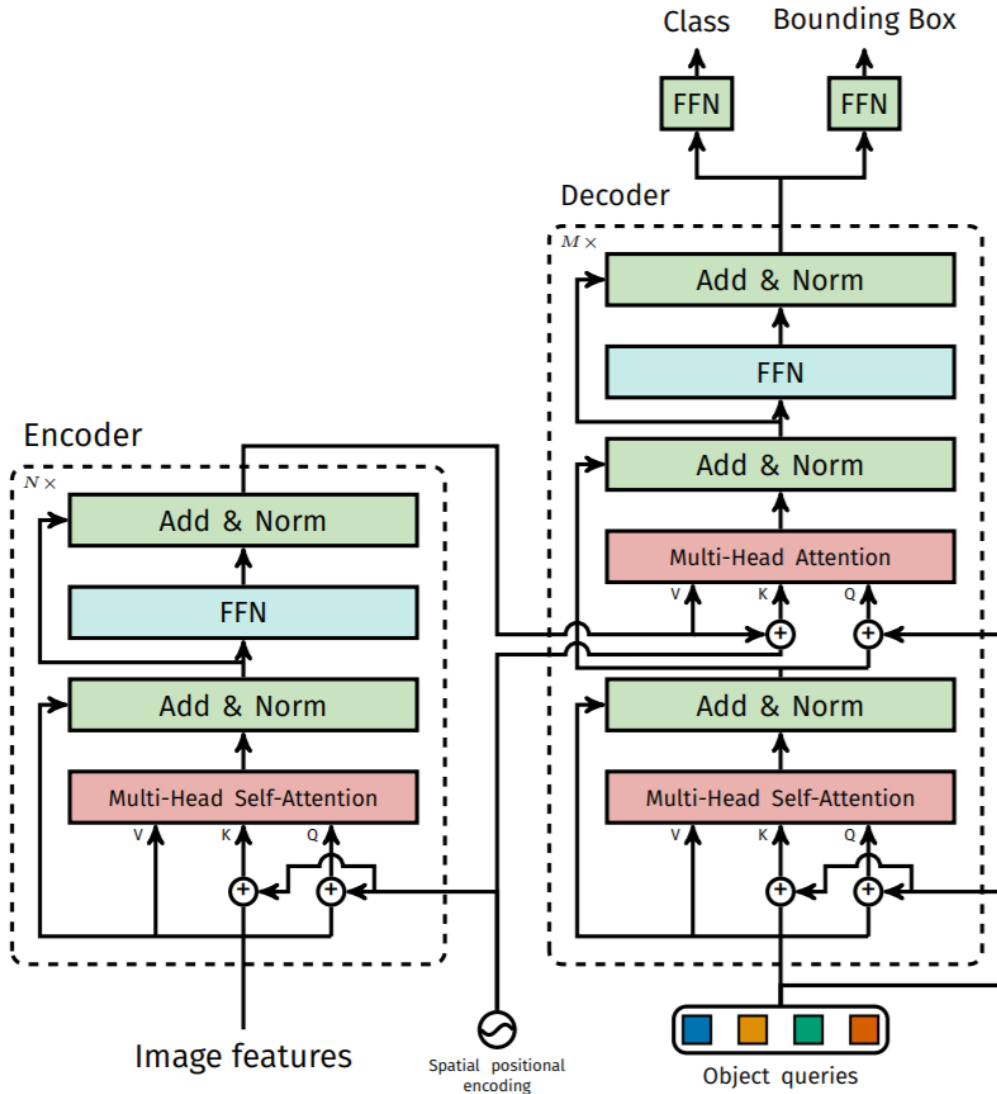
$(c_3 = \emptyset, b_3)$

$(c_4 = \emptyset, b_4)$

$(c_5 = \emptyset, b_5)$

실제 값

본 논문의 핵심 아이디어: Transformer



Transformer

- Attention을 통해 전체 이미지의 문맥 정보를 이해
- 이미지 내 각 인스턴스의 상호작용(interaction) 파악 용이
- 큰 bounding box에서의 거리가 먼 픽셀 간의 연관성 파악 용이

Encoder

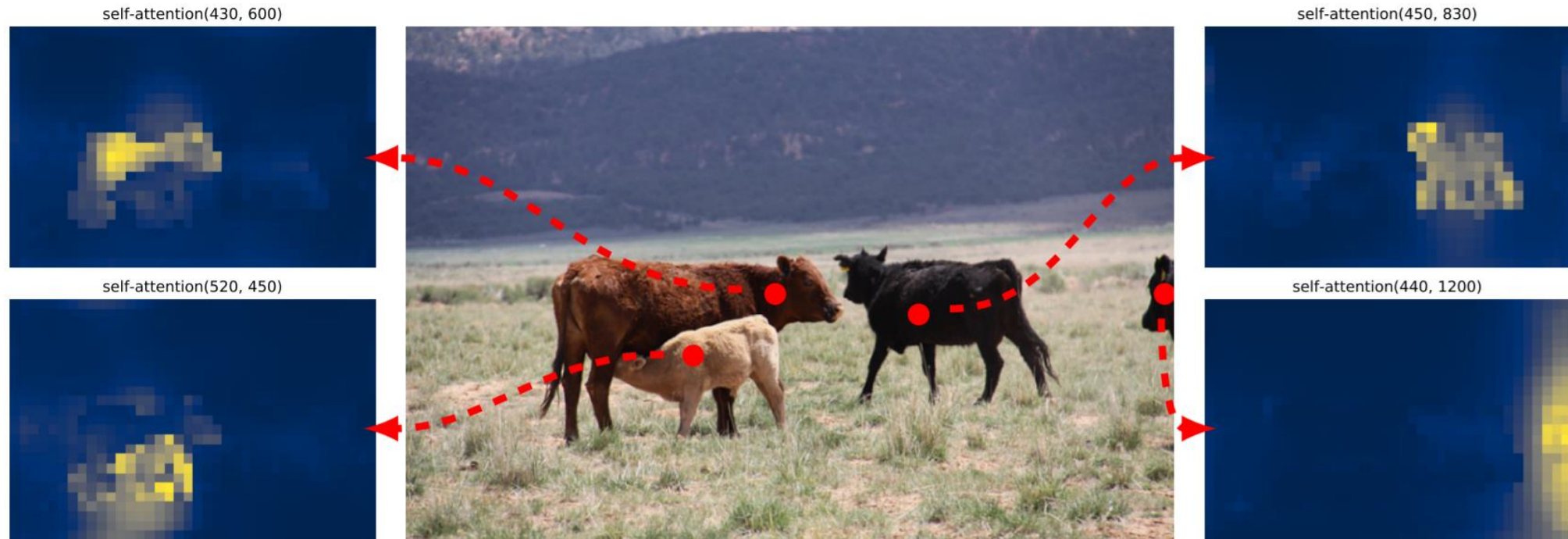
- 이미지의 특징(feature) 정보를 포함하고 있는 각 픽셀 위치 데이터를 입력받아 인코딩 수행

Decoder

- N 개의 object query를 초기 입력으로 받으며 인코딩된 정보를 활용
- 각 object query는 이미지 내 서로 다른 고유한 인스턴스를 구별

본 논문의 핵심 아이디어: Transformer (Encoder)

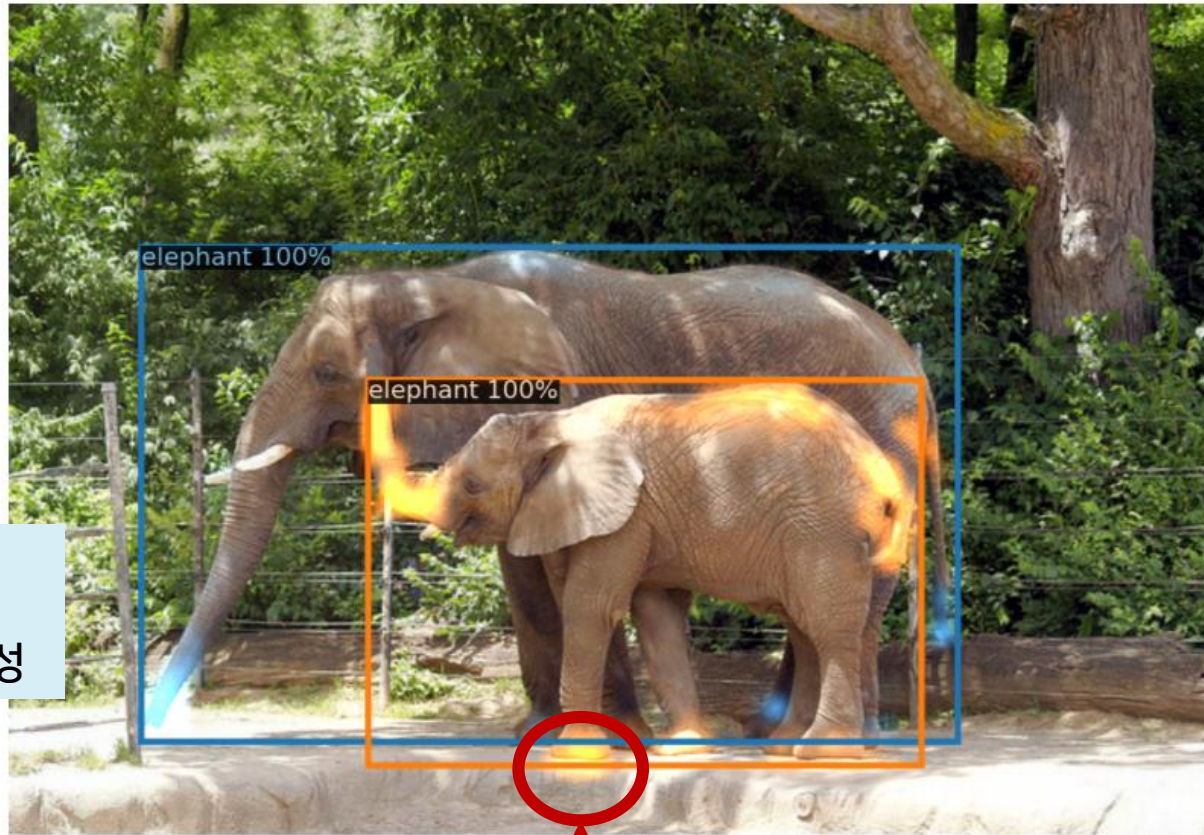
- 인코더(Encoder)는 $d \times HW$ 크기의 연속성을 띠는 feature map을 입력으로 받습니다.
 - 이때 d 는 image feature를 의미하고 HW 는 각각의 픽셀 위치 정보를 담고 있습니다.
- 인코더의 self-attention map을 시각화 해보면 개별 인스턴스를 적절히 분리하는 것을 확인할 수 있습니다.



[Table] Attention maps of the last encoder layer of a trained model

본 논문의 핵심 아이디어: Transformer (Decoder)

- N 개의 object query(학습된 위치 임베딩)를 초기 입력으로 이용합니다.
- 인코더가 global attention을 통해 인스턴스를 분리한 뒤에 디코더는 각 인스턴스의 클래스와 경계선을 추출합니다.



각 인스턴스에서
말단(Extremities) 부분의
Attention Score 값이 높게 형성

