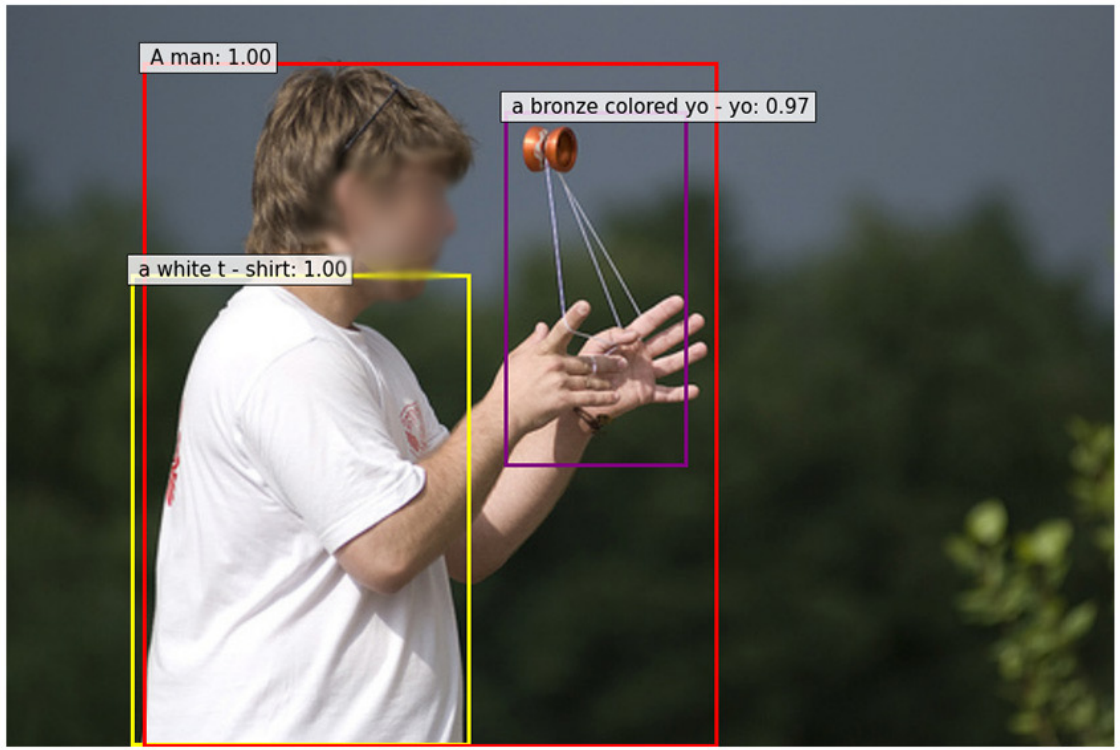
Image Grounding using Attention based Transformer

Aaditya Khant (group 19)

Task of Image Grounding

- Match phrases from caption with segments in image.



“A man in a white t-shirt does a trick with a bronze-colored yo-yo”

Approach

- MDETR is state of an art Image grounding model based on transformers.
- CLIP is transformer based neural network, which can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the “zero-shot”.
- Combine this two models to increase accuracy on the task of image grounding by overcoming limitations.

MDETR - Architecture

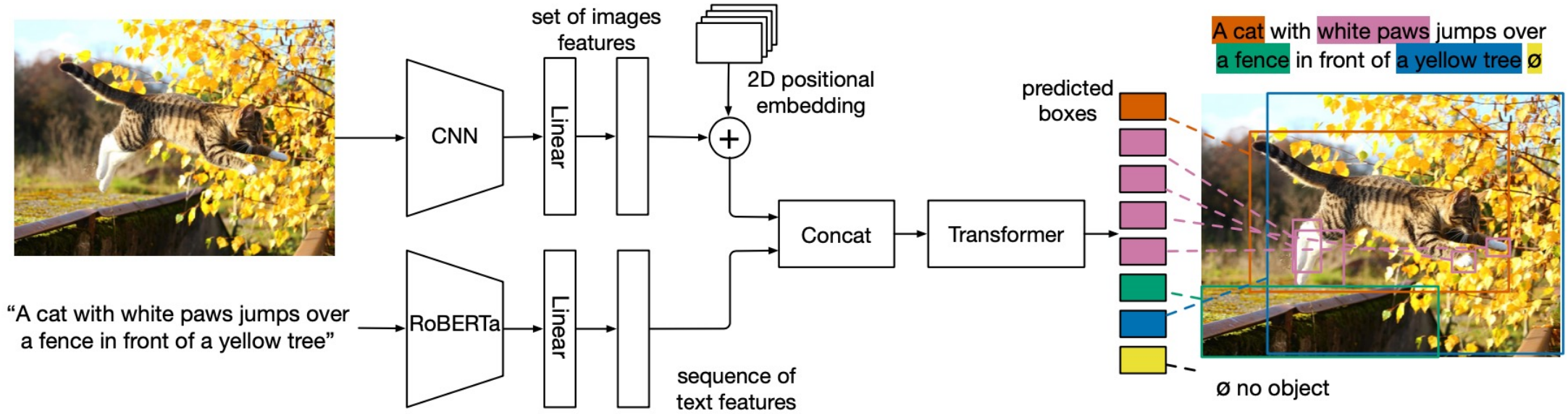
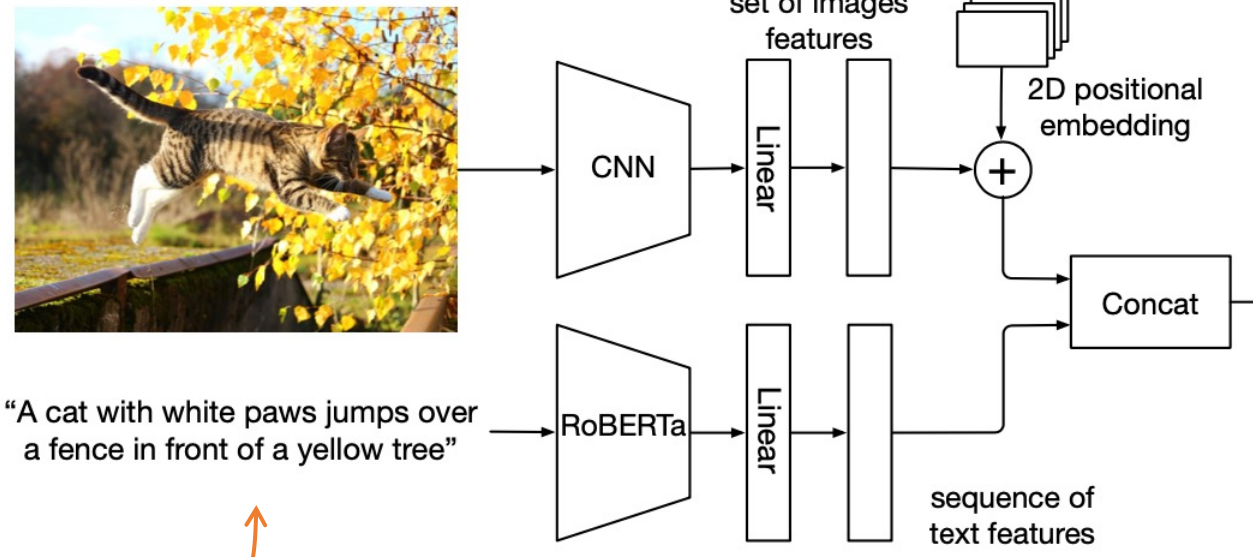


Image and Text Feature Extraction

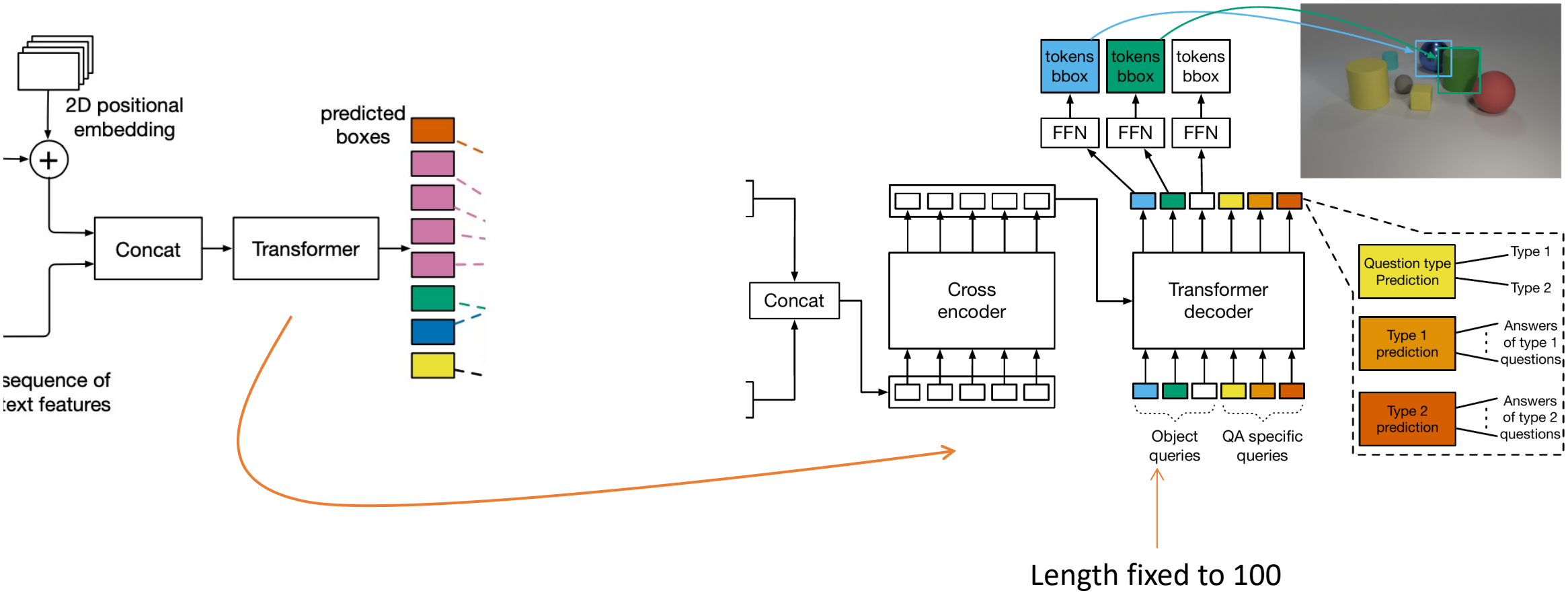
RESNET-101 (pretrained on ImageNet)
Or
EfficientNet family



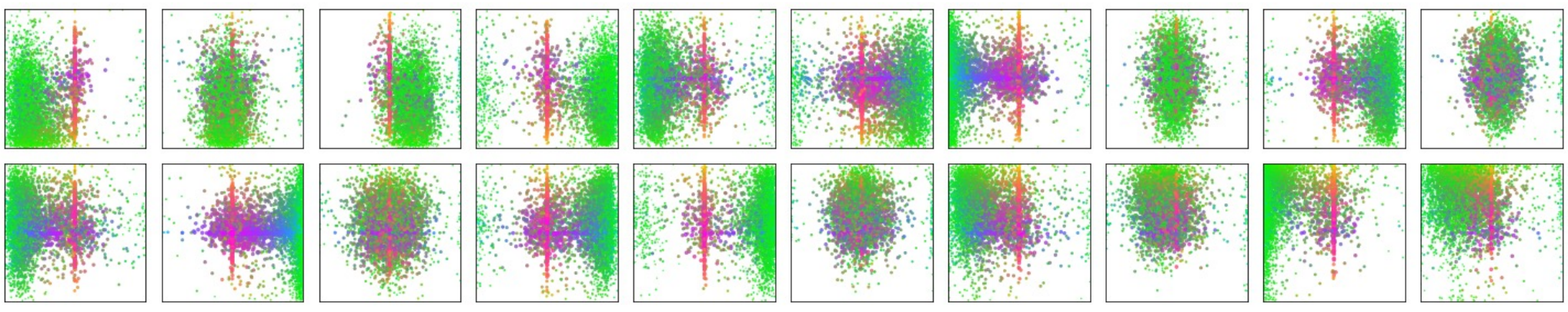
Because Attention-based transformer

Fixed length of 256 tokens

Encoding-Decoding



Object Queries

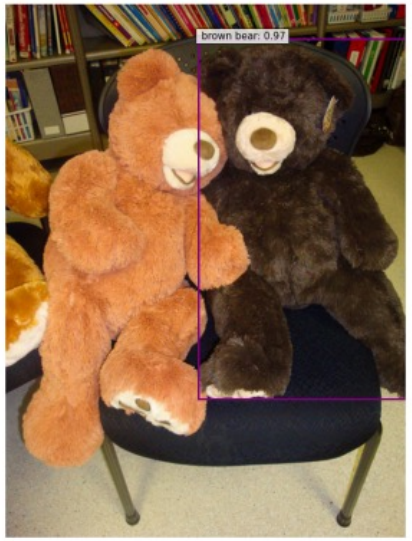


Output

- Pred_Box : bounding box corresponding to each query (100 x 4)
- Pred_logits : probability distribution over caption tokens for each query (100 x 257)

Dataset

- 1.3M image-text pair for training and testing.
- Combined dataset of Flicker30k, MS COCO, CLEVER and Visual-Genome



(a) "brown bear"

(b) "zebra facing away"

(c) "the man in the red shirt carrying baseball bats"

(d) "the front most cow to the right of the other cows"

Losses

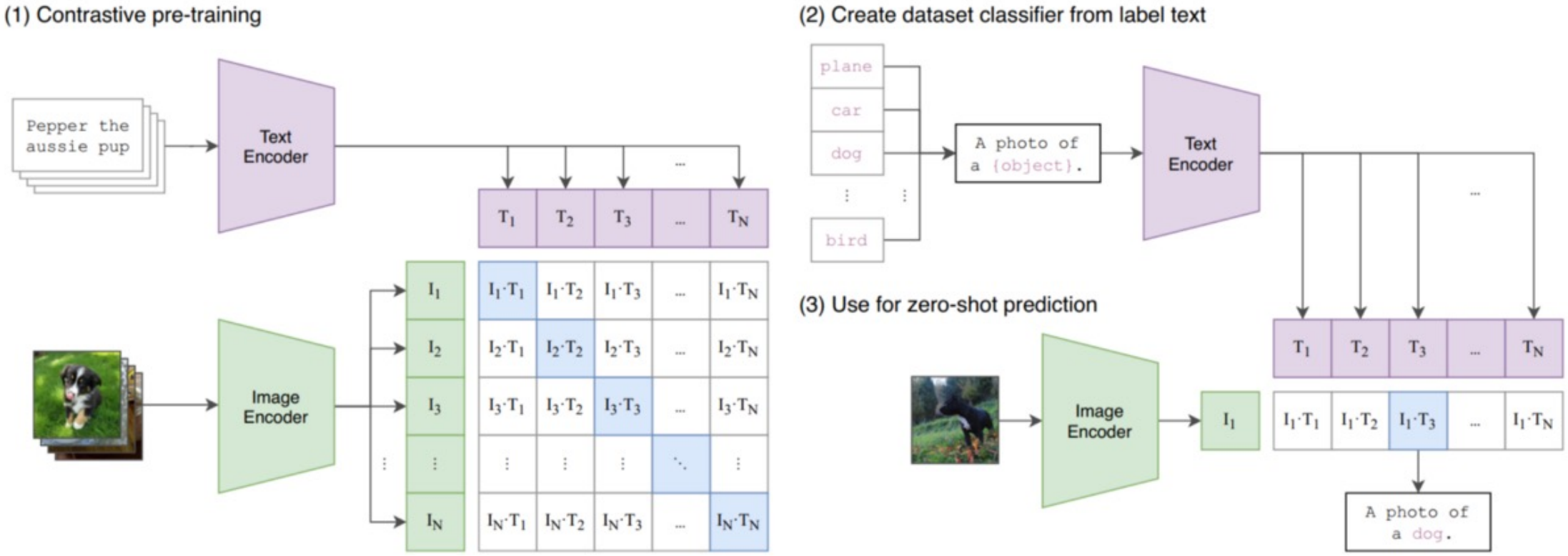
- **Contrastive alignment:** enforces alignment between the embedded representations of the object at the output of the decoder, and the text representation at the output of the cross encoder.
- **Soft token prediction**: predict the span of tokens from the original text that refers to each matched object

Bonus!

- We can use MDETR model for tasks other then phrase grounding such as, **Referring expression comprehension / segmentation** and **Visual Question Answering.**

Figure 5: MDETR provides interpretable predictions as seen here. For the question "What is on the table?", MDETR fine-tuned on GQA predicts boxes for key words in the question, and is able to provide the correct answer as "laptop". Image from COCO val set.

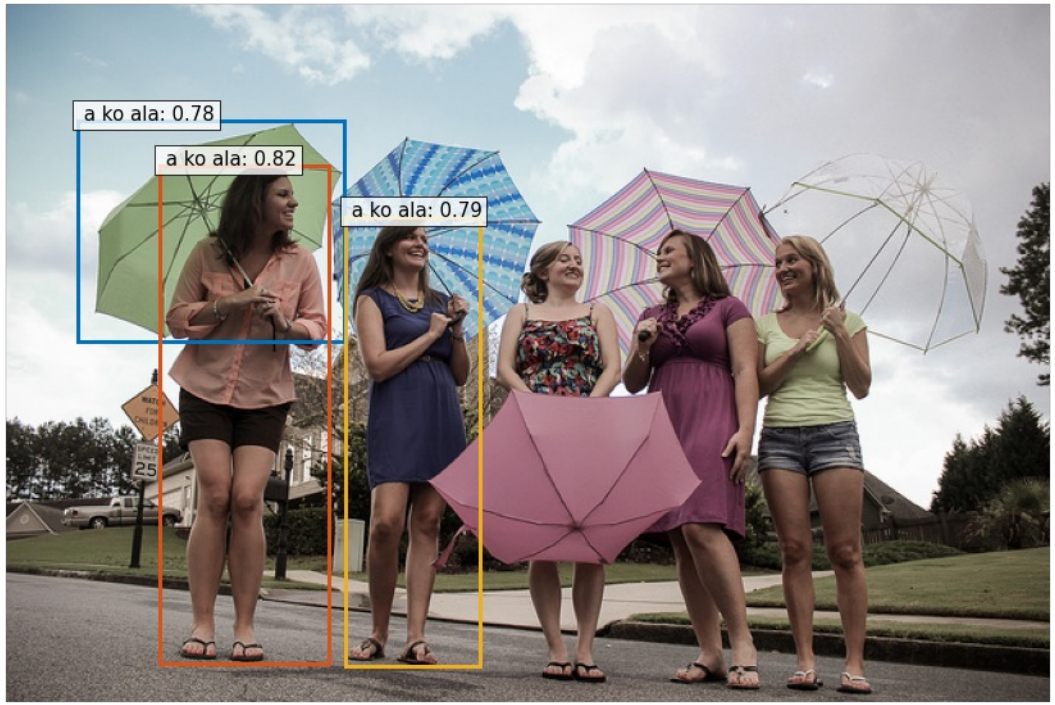
CLIP Architecture



Trained on 14 million images for 22,000 object categories!

MDETR – limitation ; CLIP - Advantage

- MDETR uses ‘few-shot’ approach, which results in matching of nonrelevant phrase with image segment. We can use ‘zero-shot’ approach of CLIP to improve accuracy.



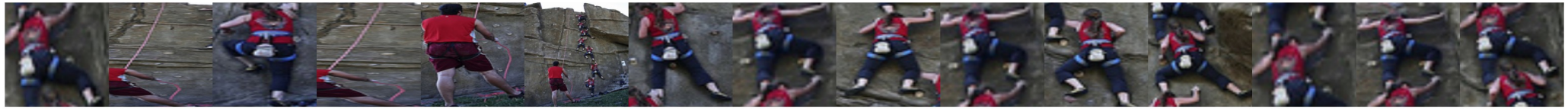
Methodology

- Retrieve all bounding boxes from MDETR, that doesn't have probability more than 0.3 for no object class, segment image using these BB. Parse caption for nouns and generate list of captions.
- Pass segmented images and captions through CLIP and generate prediction score.
- For evaluation only flickr30k dataset is used. Pretrained ViT-B/32 model of CLIP is used.

```
Caption: A cat with white paws jumps over a fence in front of a yellow tree
Generated captions
Image of a  cat
Image of a  paws
Image of a  jumps
Image of a  fence
Image of a  front
Image of a  tree
```

Results

Seven climbers are ascending a rock face whilst another man stands holding the rope .



a Seven	0.21	0.22	0.21	0.22	0.21	0.21	0.20	0.22	0.22	0.23	0.20	0.22	0.22	0.22	0.21
a climbers	0.25	0.27	0.22	0.24	0.29	0.30	0.24	0.24	0.25	0.26	0.26	0.27	0.23	0.24	0.26
a rock	0.21	0.27	0.19	0.27	0.21	0.22	0.20	0.21	0.22	0.22	0.22	0.23	0.22	0.22	0.21
a face	0.18	0.20	0.20	0.20	0.18	0.16	0.19	0.20	0.20	0.20	0.18	0.19	0.21	0.20	0.19
a man	0.20	0.20	0.22	0.20	0.20	0.18	0.20	0.22	0.21	0.23	0.18	0.20	0.22	0.22	0.19
a rope	0.21	0.26	0.20	0.24	0.24	0.22	0.21	0.22	0.23	0.23	0.22	0.22	0.22	0.22	0.22

Results

- Final submission will contain accuracy on the proposed model.

Backbone	Pre-training Image Data	Val R@1	Val R@5	Val R@10	Test R@1	Test R@5	Test R@10
Resnet-101	COCO+VG+Flickr	82.5	92.9	94.9	83.4	93.5	95.3
EfficientNet-B3	COCO+VG+Flickr	82.9	93.2	95.2	84.0	93.8	95.6
EfficientNet-B5	COCO+VG+Flickr	83.6	93.4	95.1	84.3	93.9	95.8

Evaluation of different CNN Backbone in MDETR for phrase grounding

Future Work

- We can eliminate passing of 2D queries. We can improve accuracy by training model to predict soft token prediction and bounding box for each token of caption by passing embedded text as queries.