End-to-End Object Detection with Transformers

Seungmin Lee (profile2697@gmail.com; 2020-20866),
Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction and Motivation

Current object detection methods address object detection as a sequential task that predicts object classes after proposing the candidate bounding boxes (bbox). Unfortunately, this kind of approach makes the object detection pipeline complex, requiring a lot of heuristics such as Non-Maximal Suppression or a complicated training scheme [1].

To address this problem, this paper treats object detection as a direct set prediction where the set is a collection of the (class, bbox) pairs. This view simplifies the object detection process by removing the task-specific heuristics. The authors implement it using a transformer that predicts (class, bbox) pairs and bipartite matching loss.

2. Method

2.1. Architecture: DEtection TRansformer (DETR)

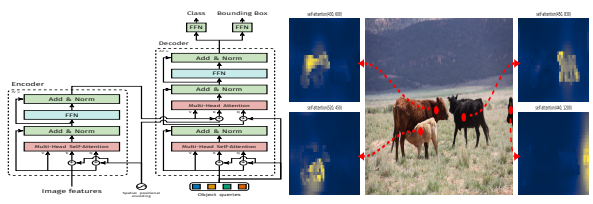
The proposed architecture named DETR consists of a CNN backbone, an encoder-decoder transformer that takes the CNN features and transforms them, and feed-forward networks that predict classes and bounding box coordinates for each token.

When we set N as a fixed prediction size, the output will be $\hat{y} = \{\hat{y}_i\}_{i=1}^{N}$ where $\hat{y}_i = \{\hat{p}_i, \hat{b}_i\}$ is a pair of the predicted class probability ($\hat{p}_i$) and box coordinates ($\hat{b}_i$).

The transformer has the typical form as proposed in [3] except they use additional positional encodings called *object queries* on the decoder. The detailed architecture is shown in Figure 1 (a).

2.2. Set Prediction Loss

The proposed method calculates the loss through two steps: finding the optimal one-to-one matching between the ground truth and the set prediction, calculating loss using the found matching.



(a) The transformer architecture

(b) The encoder resolves instance occlusions

Figure 1. (a) The proposed encoder-decoder transformer architecture, (b) The encoder's attention maps: the encoder separate each instance even though those are occluded.

2.2.1 Step 1: Finding the Optimal Matching

Suppose that $y = \{c_i, b_i\}$ is the ground truth of a given image. Note that we extend y to has the size N by appending $\emptyset$. Thus, we can match each element of y and $\hat{y}$ one-to-one. Among the possible permutation of N elements (P_N), we select the optimal permutation ($\rho^\star$) using the following equation:

$$\rho^\star = \arg\min_{\rho \in \mathrm{P}_N} \sum_{i=1}^{N} -\mathbb{1}_{\{c_i \neq \emptyset\}} \hat{p}_{\rho(i)}(c_i) + \mathbb{1}_{\{c_i \neq \emptyset\}} \mathcal{L}_{\mathrm{box}}(b_i, \hat{b}_{\rho(i)}),$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function, and $\hat{p}_{\rho(i)}(c_i)$ is the predicted confidence of the ground truth class c_i. The authors utilize the Hungarian algorithm [2] to solve the equation effectively.

2.2.2 Step 2: Calculating the Loss

The proposed method uses the following loss:

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{N} -\log \hat{p}_{\rho^\star(i)}(c_i) + \mathbb{1}_{\{c_i \neq \emptyset\}} \mathcal{L}_{\mathrm{box}}(b_i, \hat{b}_{\rho(i)}).$$

3. Results

The method shows comparable or higher performance on many metrics except on the small objects. However, considering the simplicity, the performance drop seems tolerable. As we can see in Figure 1, the encoder attention successfully separates occluded instances. Based on the result, this paper extends their method to instance segmentation.

4. Personal Note

The main contributions of this paper are two-fold: object detection as a direct set prediction, integrating transformer. The first contribution seems critical because it made the object detection pipeline simple. However, it is unclear which part of the contributions is more crucial to the performance. Therefore, it would be better to compare the proposed method with a CNN model that predicts a class-box set.

References

[1] S. Ren, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[2] R. Stewart, et al. End-to-end people detection in crowded scenes. In *CVPR*, 2016. 1

[3] A. Vaswani, et al. Attention is all you need. In *NeurIPS*, 2017. 1