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. Thus, these methods tackle object detection as a combination of a surrogate regression task for boxes proposals and a classification task on the selected proposals. 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 [].

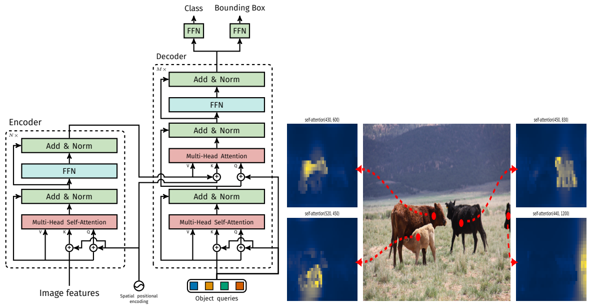
To address this problem, this paper view object detection as a direct set prediction where the set is a collection of the (class, bounding box coordination) pairs. The changed view simplifies the object detection process by removing the heuristics that encode the human's prior knowledge about object detection.

The authors implement the set predictor using an encoder-decoder transformer and bipartite matching loss. The transformer takes features extracted from a traditional convolutional neural network (CNN) as input tokens and outputs a set of (class, bounding box)s.

2. Method

2.1. Architecture

The proposed architecture consists of three main components: a CNN backbone that extracts features used as input tokens for transformer, an encoder-decoder transformer that takes CNN features and transforms the features, and two feed-forward networks that predict classes and bounding box coordinations, respectively.



(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.

The transformer has the typical form as proposed in [] 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

3. Results

4. Personal Note

The main contributions of this paper are two-fold. First, it views object detection as a class-box set prediction rather than a sequential task that classifies object classes in the boxes after predicting boundary boxes. Second, it integrates encoder-decoder transformers into their architecture. The first contribution seems critical because it made the object detection pipeline simple. However, I think 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