Pix2seq: A Language Modeling Framework for Object Detection

Reyva Babtista

Pix2Seq is a task-agnostic (generic) framework for [object detection](id:667882a1-fc54-4f84-8eac-f1f009da5aba). It cast object detection as a language modeling task, conditioned on the observed pixel input. Pix2Seq was proposed as a general intelligence system, in contrast to most existing methods that are specialized and difficult to integrate to larger system. The model provides a language interface by producing a sequence of discrete tokens that correspond to object descriptions. The objective function defined in the model is simply the maximum likelihood of tokens given the pixel inputs and the preceding tokens. The model can be configured to task-specific prior knowledge with a sequence augmentation technique. Based on the reported experiments, Pix2Seq has a comparable result to SOTA object detection models such as [Faster R-CNN](id:e840c4b3-e08a-40f9-85bd-b31e56e30473).

Pix2Seq framework was built upon four main components:

1. Image augmentation for training.
2. Sequence construction and augmentation for object description or annotations.
3. An encoder-decoder model architecture.
4. Objective (loss) function is to maximize the log likelihood of tokens conditioned on preceding tokens and the input image (with softwmax cross-entropy loss).

Pix2Seq express sets of bounding boxes and class labels of objects exist in images as sequences of discrete tokens. Object's bounding box (top-left bottom-right or center and height/width)'s continous numbers of , are discretize using image bins. Thus, an object is represented by five discrete tokens, i.e. , with the as the class index. EOS token was used to incorporate the end of sequence.

The encoder used in Pix2Seq can be a ConvNet, Transformer, or their combination, while the decoder used is a Transformer based. The decoder generates one token at a time, conditioned on the preceding tokens and the encoded image representation.

The maximum likelihood loss was used for Pix2Seq objective function, that is defined as

(formula details are on the paper).

At inference time, Pix2Seq sample tokens from model likelihood from a join vocabulary for the bins and classes using sampling or other stochastic techniques. However, based on the experiments presented, nucleus sampling leads to a higher recall than sampling.

There is a downside to the task-agnostic model of Pix2Seq, that is the model tends to predict EOS token too soon, thus missing some objects left undetected. This problem is most likely because of the annotation noise and uncertainty in recognizing and localizing an object. Although this problem only affects the precision score by 1-2%, delaying the EOS token sampling by decreasing its likelihood resulting in a higher recall score with the trade off of noisy and duplicated predictions (precision is decreased more).

This problem of precision and recall trade-off can be solved by letting the model learn of a prior knowledge of the task using sequence augmentation with synthetic noise objects. Sequence augmentation improved the recall and delaying the EOS token prediction without increasing the frequency of noisy and duplicated predictions.

Although Pix2Seq has a competitive results as compared to other object detection models, there are limitations identified. First, autoregressive modeling is expensive for long sequences. Second, input data used for training presented in the paper are fully dependent on human annotation.