DETR with Modulated Object Queries For Object Detection

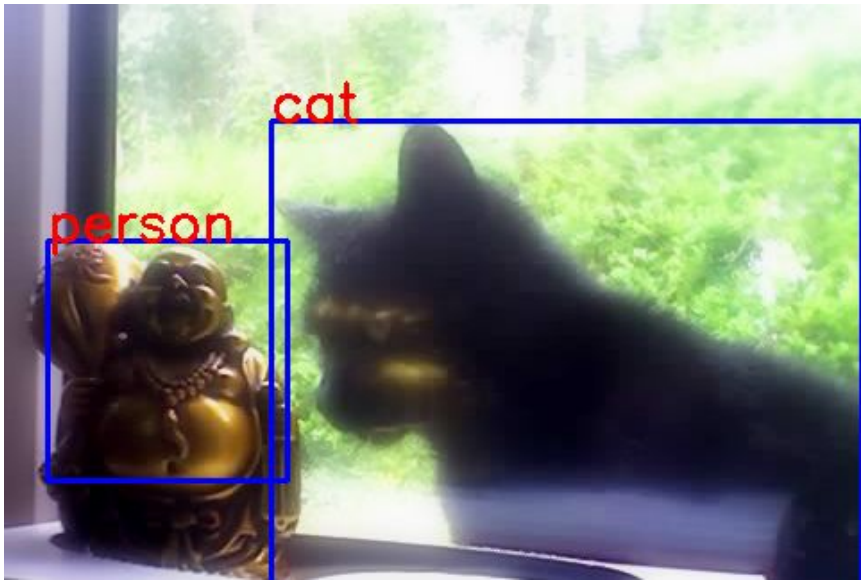
Sudeep Narala

Introduction

In the problem of object detection, the input to the model is an image and the output is a set of bounding boxes on the image with class designations for each box.

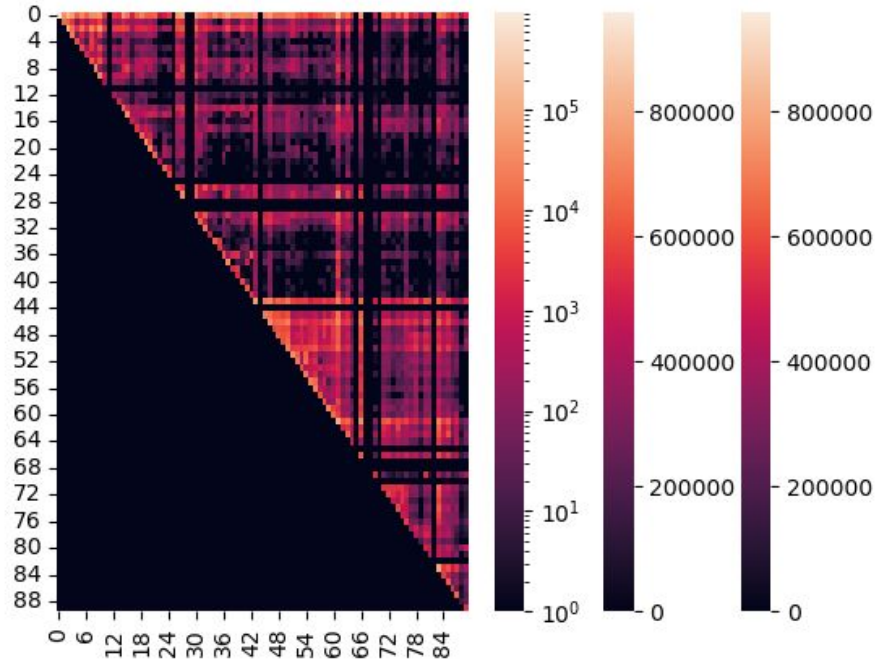
We are building on top of DETR (Detection with Transformers) which removed the need of a lot of hand-designed components of the model (such as anchor generation and NMS) introduced by R-CNN. DETR makes use of transformers and set matching.

Use Coco Dataset.

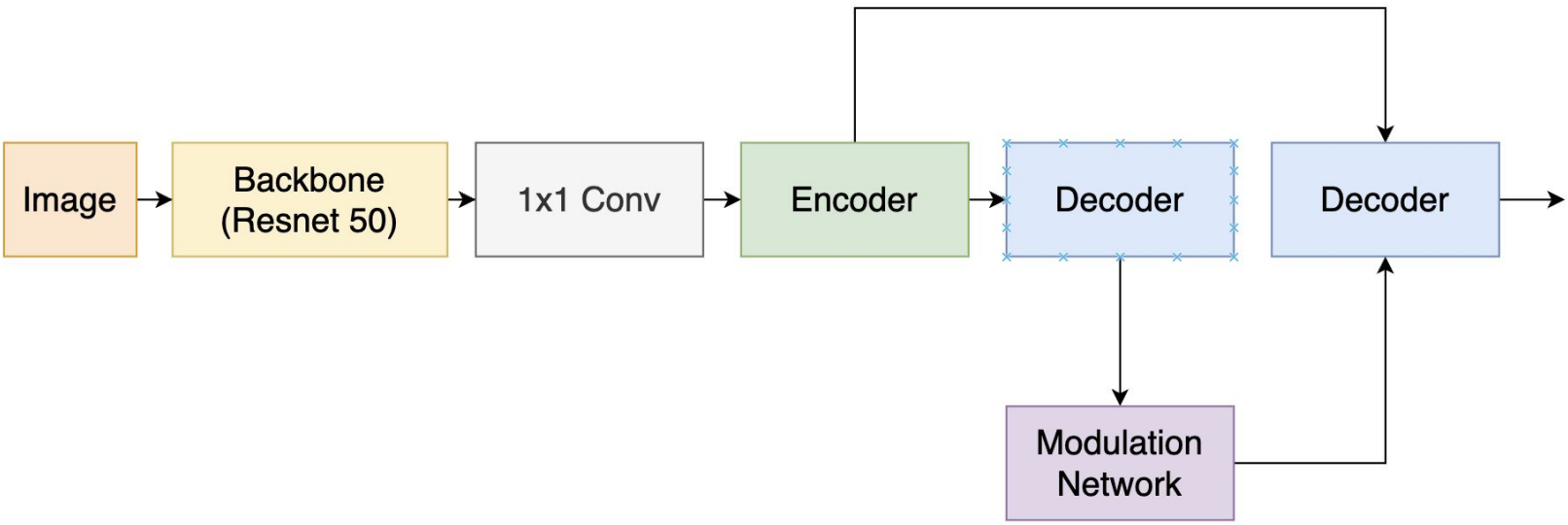


Motivation

- Prior work has shown that carefully constructed object queries can boost performance and reduce convergence time.
- Object queries are not taking the semantics of this specific image into account. Instead, they are learned and fixed.
- Two pass decoding can prove to be useful in this situation to give decoder a more global image view from the beginning



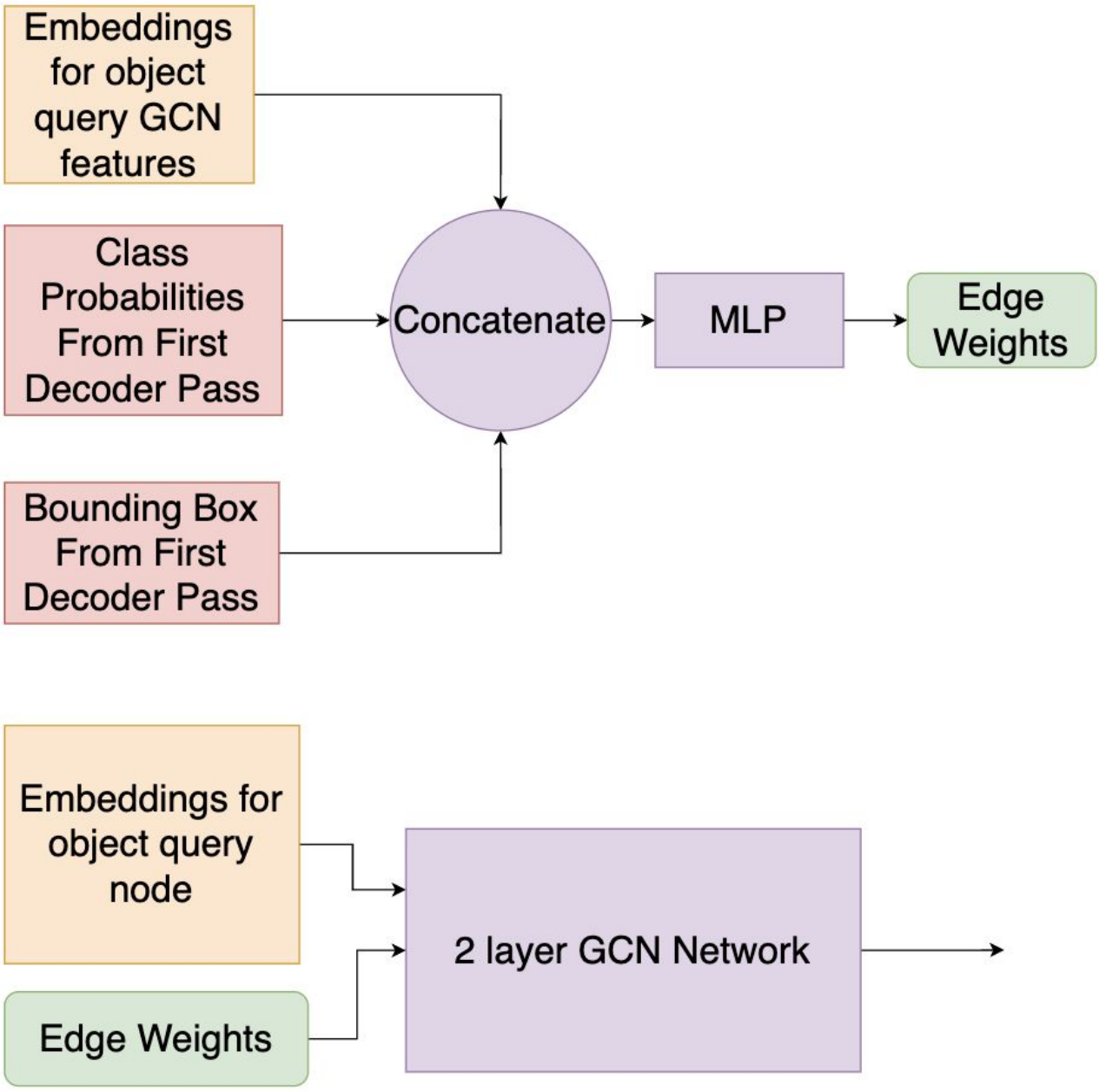
Object Query Modulation Model Overview



- Memory from encoder flows to both passes of the decoder

- The first pass of the decoder is responsible for modifying the object queries based on image semantics

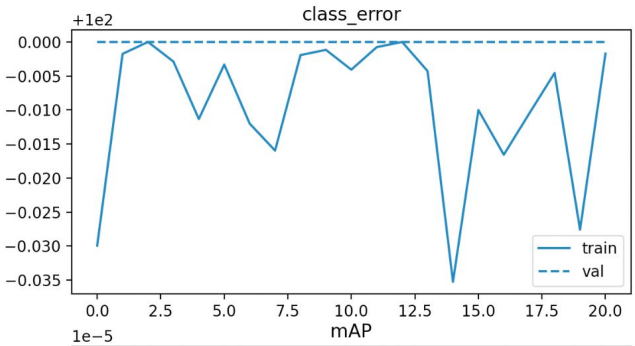
GCN Model

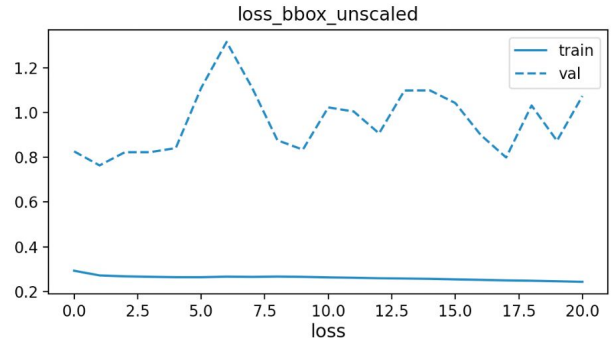


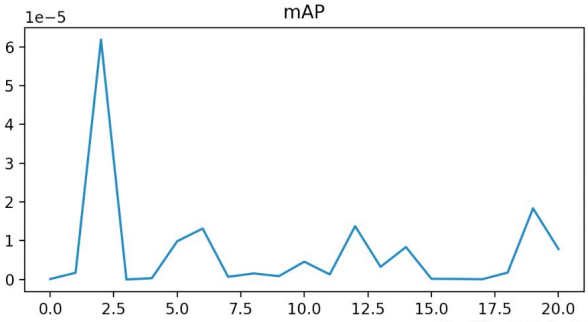
- Also tried a simpler model without GCNs but not included here for brevity

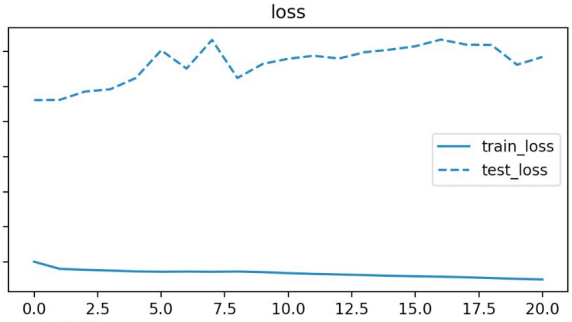
Experimentation + Results

Base Model From Scratch (Sped Up Training Schedule)

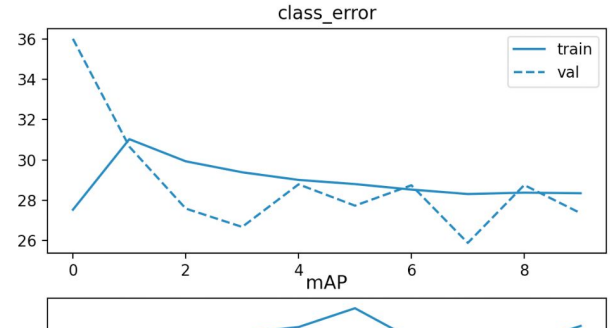


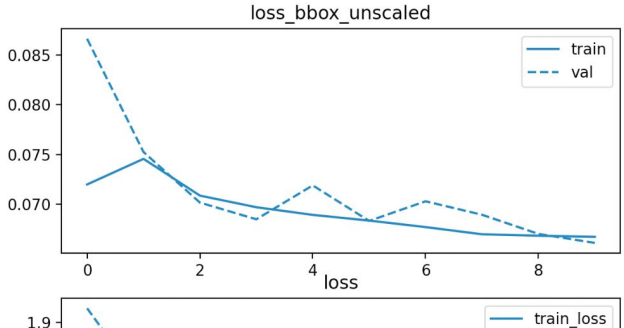


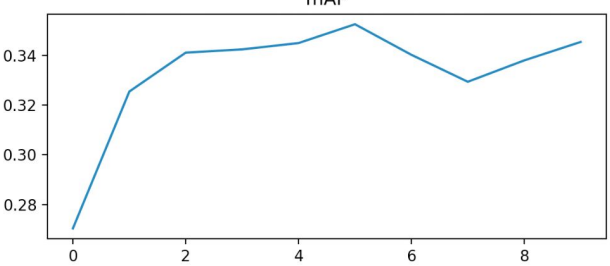


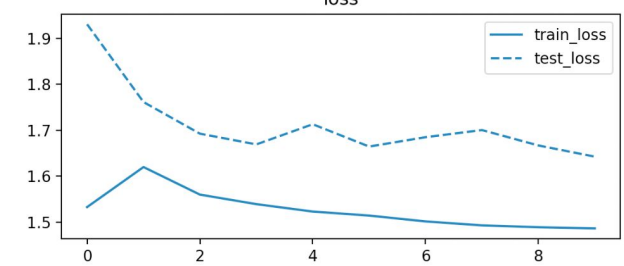


Modified Object Queries Through GCNs

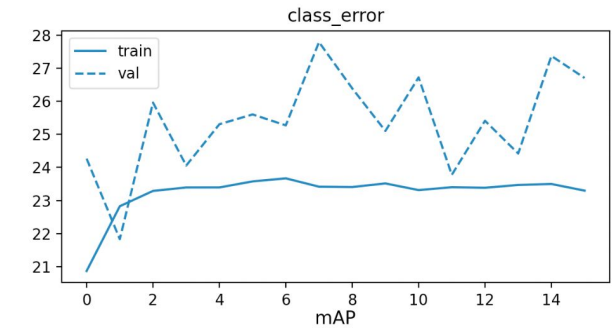


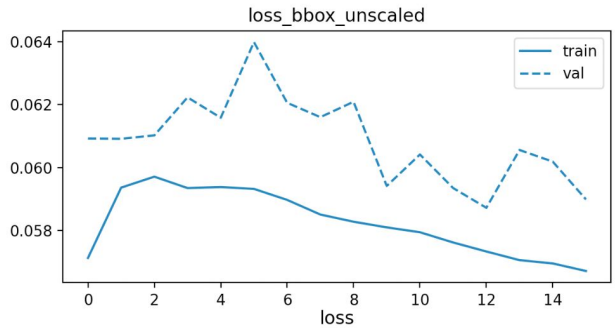


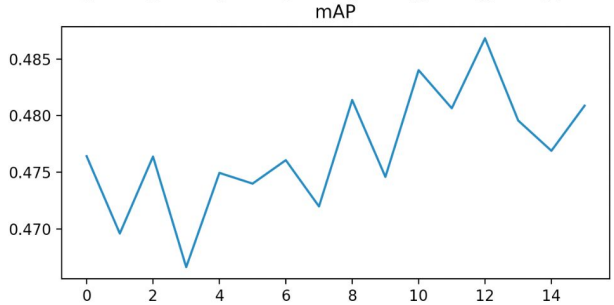


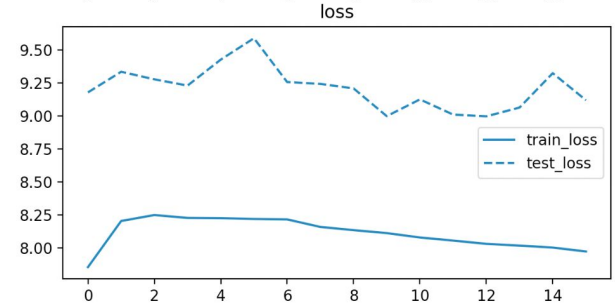


Modified Object Queries Without Spatial Awareness









Conclusions:

- ➔ Need to analyze computational costs of training a specific model relative to the time and compute one has access to
- ➔ Might benefit from an alternating training approach where the modulation network is essentially trained alternating with the rest of the network (freeze one and train the other) with augmented losses
- ➔ In general, might also benefit by training from scratch instead of getting model to “unlearn” from its stable state after 500 epochs

Next Steps:

- ➔ For GCN technique, use a pairwise scoring function to learn edge strength
- ➔ Spend more time training with the spatially unaware modulation technique because there are signs of promise from the loss curve
- ➔ Try the alternating training approach
- ➔ Consider more approaches where just the encoder values are used to modulate the object queries so we don’t need 2 passes through the decoder