CS323 Reading Assignment 4: Transformers for Object Detection and Object Queries in DETR

David Felipe Alvear

April 24, 2023

1 Introduction

Extending the capabilities and scalability of Transformers used in Natural Language Processing some authors tried to use the same system to apply to visual computing. We have seen that transformers provide computational efficiency and flexibility while maintaining the capacity of training on large models and larger datasets. Image transformers have arisen to extend the benefits of NLP transformers, some early applications used transformers with convolutional neural networks. The authors of the ViT paper introduce the application of Transformers for object classification without the need for CNN, in the same way, the DETR paper extends the use of image transformers to propose an end-to-end model for object detection. In the next subsections, we discuss briefly the contributions of each study.

1.1 ViT: An Image is Worth 16x16 Words

In the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Alexey Dosovitskiy et al., The authors introduce the Vision Transformer (ViT), an implementation of Transformers for image recognition. They demonstrated that transformers can achieve comparable performance to CNN models when the ViT models are trained on large datasets (14M to 300M) images, beating the current state of the art in image

recognition tasks. ViT models have a lack of inductive bias that aims to not generalize well, compared to CNN which contains translational equivariance and locality. However, for training in large datasets, the ViT model can learn spatial relations.

ViT models follow the original transformer architecture due to simplistic setup and scalable NLP transformer architectures. The key insight of the paper is to treat the images with size $(C * H * W)$ as sequences of tokens or "words" based on the original implementation. The tokens are created by transforming the image in a sequence of patches ($N = H * W/P^2$, $p : patchsize$), then projecting the data to a lower dimensional space using a learnable linear projection to have the output patch embeddings. Then is added a learnable positional embedding to the patches to aim the model to encode distance in the image to finally pass the tokens to the classification head based on transformers. The self-attention in the transformer architecture helps to integrate information across the whole image from the lowest layers, it uses the attention distance which has a similar meaning to the receptive field in CNNs to show the ability to integrate information globally. The ViT model attends to regions that are relevant for the classification task, focusing on parts of the image that are important for identifying the object.

Image transforms demonstrated that can outperform state-of-the-art convolutional neural networks (CNNs) in image classification tasks when pretrained in large-scale datasets, also having a lower computational cost and promising self-supervision training. Additionally, the paper highlights the scalability of the transformer architecture, showing that the ViT can be further improved by increasing dataset size and model capacity.

1.2 DETR: End-to-End Object Detection with Transformers

In the paper "End-to-End Object Detection with Transformers" by Nicolas Carion et al., the authors introduce an object detection model DETR (Detection Transformer), that uses the transformer architecture for object detection tasks. Object detection models make use of hand-designed components such as non-max suppression and anchor boxes. The initial guesses of anchor boxes or box proposals, as well as, post-processing techniques (NMS) affect

the performance of the object detection task. DETR model proposes end-to-end object detection replacing traditional handcrafted components with a direct set of predictions without complex pipelines. The authors demonstrated that DETR can match the performance of current object detection models such as Faster-RCNN while maintaining a simpler, more interpretable architecture.

The paper focuses on two main ideas to generate direct predictions, design a prediction loss to enforce unique predictions, and an architectural design that in a single pass the model predicts objects and models their relation. The DETR model can infer N predictions which can be more than the objects present in the image, for this the model uses a loss function to produce optimal bipartite matching between the prediction and the ground truth, the objective is to find a one-to-one matching for direct set predictions without having duplicates. The loss function is based on the Hungarian loss and combines a loss for the class prediction and a bounding box loss that is a linear combination of L_1 loss and a scale-invariant generalized IoU loss.

The DETR architecture is composed of a backbone model that can be either a ResNet-50 or ResNet-101 pre-trained on ImageNet that passes the image and compute a set of image features flattening the map features from the backbone. A positional encoding is added to the image features to pass to the transformer encoded composed by multi-head attention layers and FFN, the encoder will capture the spatial relationships between the objects in the image. The transformer decoder processes a fixed number of learnable object queries that represent the possible objects that can be detected in the image, the decoder can process parallel the object queries and attend the encoder output, leveraging both self-attention mechanism and cross-attention, this allows the decoder to generate a set of predictions for object classes and bounding boxes per each object query. The prediction is performed by a Feed Forward Network applied to each of the object queries.

2 Transformers for Object Detection

Transformer architecture can be used for object detection adapting the input to handle visual information. The general approach to solving the object

detection task using transformers is described in the following steps:

- Backbone: Transformers for object detection use a Convolutional Neural Network to compute feature maps from an input image. The objective is to capture spatial relations and low-level features of the image. The backbone of the DETR model is composed of ResNet-50 or ResNet-101.

- Image features: Similar to the transformer architecture the image transformer creates a sequence of tokens. The output of the CNN is flattened and linearly embedded to create a sequence of tokens. Additionally, given that the transformer can capture the position in the architecture, a positional encoding is added to retain spatial information.

- Transformer encoder: The transformer encoder uses self-attention mechanisms to process the input tokens. It produces an encoded representation that captures the contextual information of the image, in detail, the encoder is important to disentangling objects present in the image due to its capacity for global reasoning, this helps to understand relationships between parts of the image.

- Transformer decoder: The transformer decoder is the last step in the object detection task, this is responsible for predicting the object class and bounding boxes. The decoder takes learnable object queries to perform self-attention mechanisms and cross-attention with the encoder output. The object queries guide the decoder to generate the set of object detections.

Image transformers eliminate the need for handcrafted modules such as anchor boxes, region proposals, and non-max suppression. Transformers allow an end-to-end approach to object detection. Additionally, this offers a simple, interpretable architecture compared to traditional object detection approaches which rely on complex pipelines and heuristics. Transformers have demonstrated scalability and lower computational cost in NLP tasks, as well as, the transformers can model global context and long-range dependencies between different parts of the image. The application of object detection using transformers is expected to improve as the training dataset increases and the models are more complex.

3 Object Queries in DETR's Decoder

The transformer decoder processes the object queries while attending to the encoder output. The object queries are learnable representations that guide the decoder to detect and localize the object in the image. They are fixed in number by a hyper-parameter that depends on the dataset and is initialized randomly. In the DETR model, each object query is responsible for predicting the class and bounding box of the object, in training, the model learns how to assign to each object query a ground truth object or a "no object" class. This allows to process in a parallel way the object queries without the need for anchor boxes or region proposal, each object query is capable of representing the object present in the image.

The decoder uses as input the object queries to iteratively refine the representation of the objects or regions in the input image. The self-attention layers allow the object queries to model the relationship between objects in the image, and the cross-attention enables the object queries to attend the encoded representation of the image from the transformer encoder. In the end, the object queries are passed to a feed-forward network to predict the object class and the bounding box of the object. This allows the model to process each object in parallel and perform end-to-end detection without the need for pre-processing techniques or initial guesses.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929v2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, *End-to-End Object Detection with Transformers*. arXiv:2005.12872