ToMe: DETR But Faster

Team 8: 312551043 蔡佩君 312551048 王凱俐 312551082 謝惠喻

Project page

Introduction

ToMe [1] (Bolya et al., ICLR 2023)

A method to increase the throughput of ViT models

- Gradually combines similar tokens using a lightweight matching method.
- Delivers speed comparable to pruning with superior accuracy.
- Effective with both re-training and non-re-training scenarios.

Our Objectives

- ToMe's current applications are primarily based on the ViT architecture.
- We aim to expand ToMe's technology to additional tasks, especially within the **object detection** framework.
- To improve the inference throughput of the **DETR** [2] model.

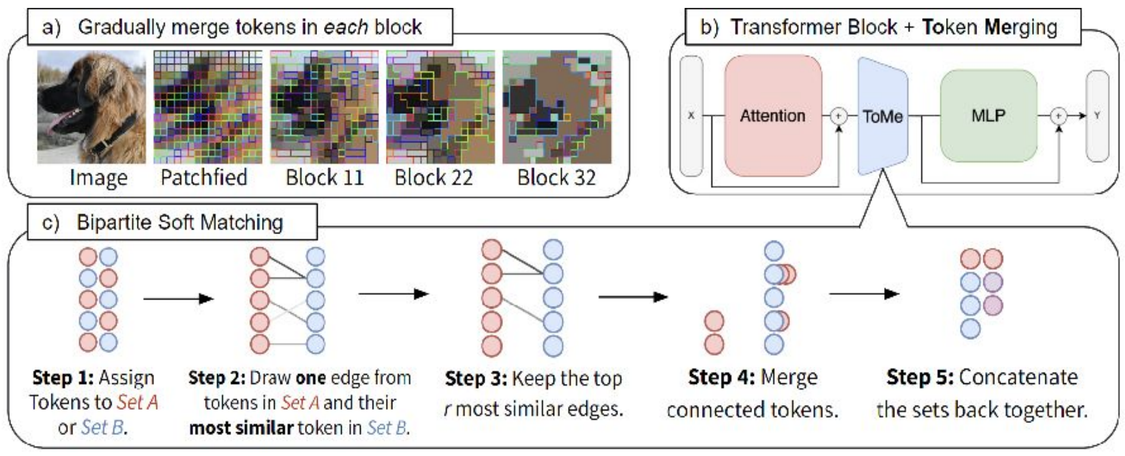

Figure 1: ToMe token merging process [1]

Proposed Method

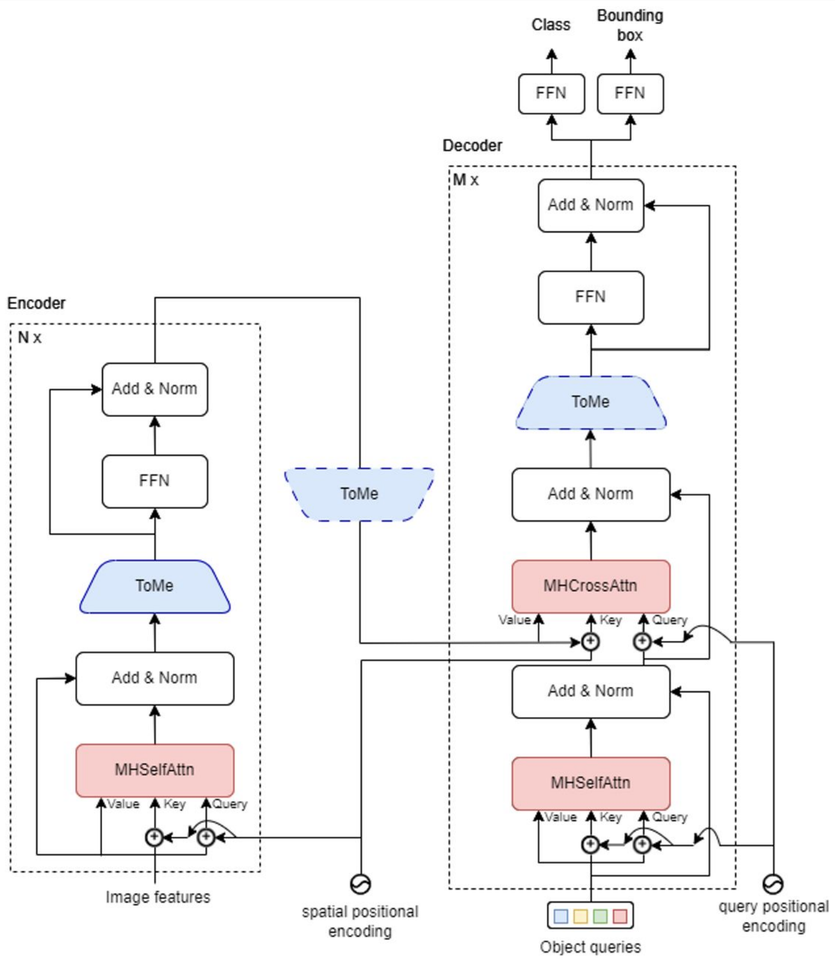

Figure 2: Model Architecture

Model Architecture

- Figure 2 illustrates our model architecture, which is based on DETR. We insert the token merging step **between the attention module and FFN** of each encoder and decoder layer.
- Both the encoder and decoder consist of 6 identical layers, primarily composed of attention modules, ToMe modules, and FFNs, connected by Add & Norm layers. The attention module's query and key incorporate spatial/query positional encoding. Additionally, the encoder's memory passes through a ToMe module before being sent to the decoder.
- Finally, the output of the decoder passes through two separate FFNs to predict the bounding box positions and object classes.

ToMe Module

- We first define the hyperparameters **er**, **mr**, and **dr**, representing the **number of encoder tokens**, **memory tokens**, and **object queries to be reduced** in each ToMe module.
- In each layer, the number of tokens is gradually reduced as shown in Figure 1. By the final layer, a total of 6*(er+mr+dr) tokens will have been reduced.
- In DETR, tokens are combined with positional encoding. To avoid dimension mismatches, we **merge the corresponding positional encodings** similarly to Figure 1. However, in the fourth step, we replace the weighted average with **the positional encoding of set B**.

Experiment Results

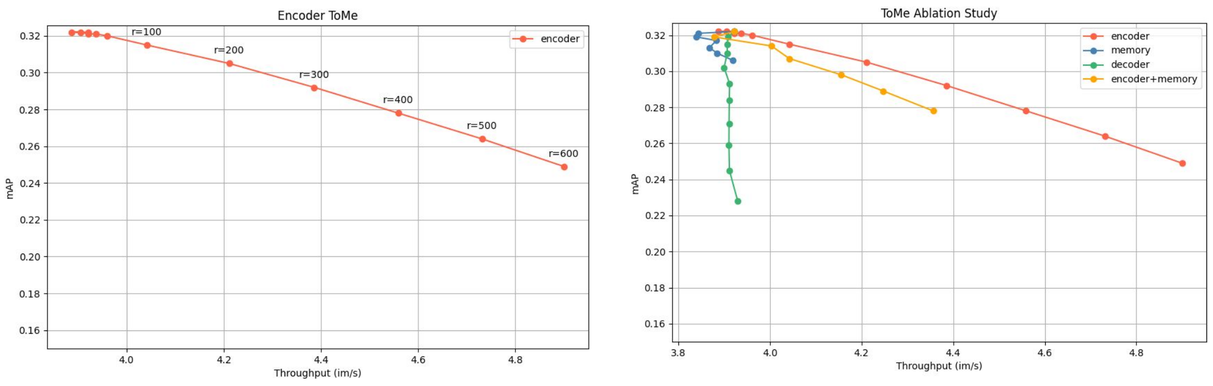

Figure 3: Apply ToMe Results

Figure 4: Image Visualization

Conclusion

- According to our experimental results, applying ToMe to the DETR model accelerates inference. However, increased acceleration comes with a trade-off in accuracy.
- Through comparative experiments, integrating ToMe into the encoder yielded the most optimal results.
- Compared to the original paper, the effectiveness of ToMe when applied to DETR is less pronounced. We hypothesize that this is because object detection models need to account for more positional information, resulting in greater information loss during the merging process.

[1] Bolya, Daniel, et al. "Token merging: Your vit but faster." arXiv preprint arXiv:2210.09461 (2022).
[2] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." arXiv preprint arXiv:2010.04159 (2020).