THE UNIVERSITY OF SYDNEY

Twostageclip-former Swin transformer with redundant attention optimize

Ziyu Zhang(SID520318384) Ziyang Zeng(SID520596085)

Present by Team RTX4090 Working all Day

COMP5329 Deep Learning

Acknowledgement

I would like to extend my sincere gratitude to our esteemed mentors, particularly Professor **Xv Chang** and Tutor **Linwei Tao** and Tutor **Zebing Du**. Your expert guidance and unwavering support have been invaluable in this journey. You have inspired us with your passion, and your insightful advice has significantly contributed to the realization of this project. I am truly grateful for your mentorship and the knowledge you have imparted.

Thank you!

THE UNIVERSITY OF SYDNEY

Twostageclip-former Swin transformer with redundant attention optimize

Ziyu Zhang(SID520318384)   Ziyang Zeng(SID520596085)

Present by Team RTX4090 Working all Day

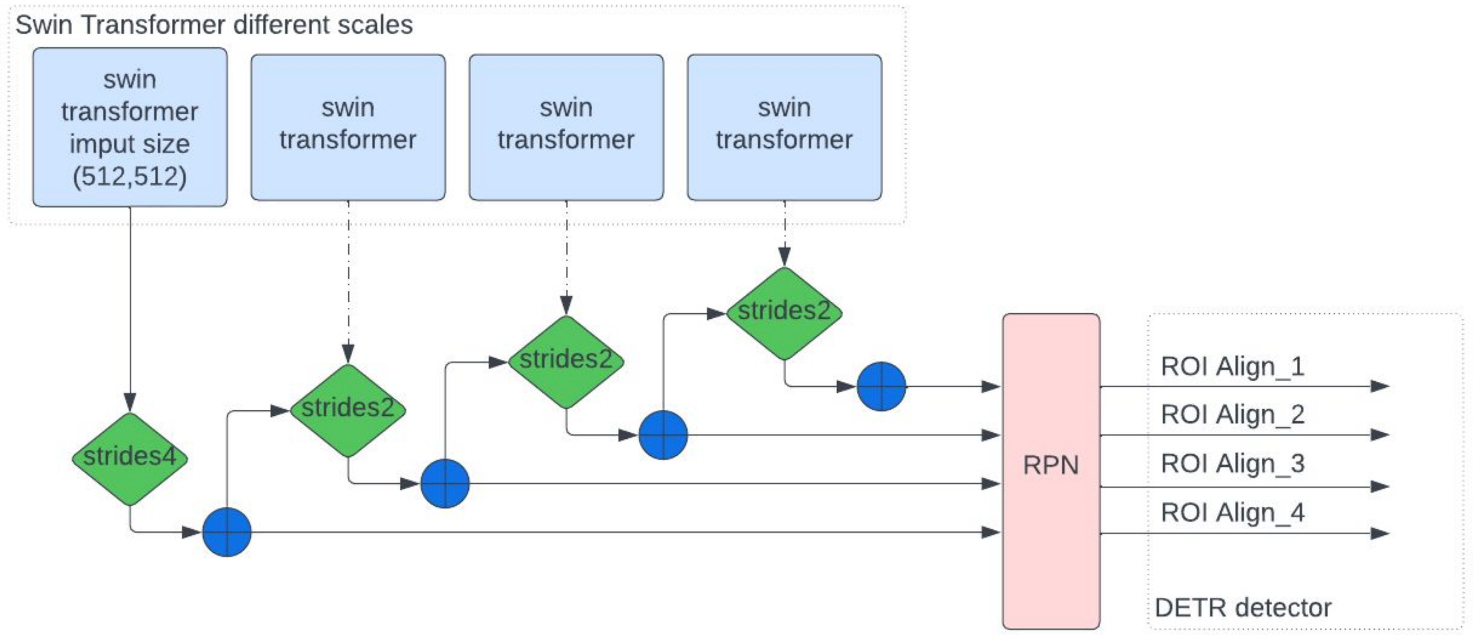
COMP5329 Deep Learning

Overview:

Our project focuses on enhancing the efficiency of transformer-based models used for image detection. We developed and compared two models: one using Mask RCNN and another using a DETR detector model with a novel TwoStageClipSwinTransformer backbone. Contrary to expectations, we did not observe a significant reduction in computational cost (measured in FLOPs). We also encountered memory leakage issues during the training process. Despite these challenges, the models demonstrated effective learning and competitive performance. Our findings serve as a valuable resource for further research and development in optimizing transformer-based models for image detection tasks. The code could be find from https://github.com/shiyuasuka/shiyuasuka-twostageclip_swintransformer

Methodology:

Our study centers around the application and comparison of two image detection models, each employing a unique combination of detector heads and transformer backbones. The first model utilized a Mask RCNN detector with a standard Swin Transformer backbone, while the second incorporated a DETR detector with a novel TwoStageClipSwinTransformer. The latter was specifically designed to retain only a percentage of the top patches based on attention scores. Despite computational and memory challenges encountered during training, the models demonstrated a competitive performance, providing a foundational step for future research in the field.



Approach:

Our approach to reducing FLOPs hinges on the application of a clipping mechanism in the Swin Transformer model. By selectively retaining a proportion of patches based on attention scores, we reduce the total number of patches N. Given that the self-attention operation's complexity scales quadratically with N, a reduction in patches theoretically leads to a quadratic decrease in total FLOPs. Nonetheless, the actual reduction might be influenced by other factors like implementation details and optimization techniques.

Results:

Model	Flops	params:	acc:
DETR	95.613G	45.433M	95.13%
MASK-RCNN	0.105T	47.795M	94.72%

Discussion and future improvement:

Our experimentation resulted in unexpected outcomes, particularly in terms of the anticipated reduction in FLOPs through the use of a two-stage clipping mechanism. This led us to reflect on potential aspects of the model construction that may have contributed to this. We speculate that our use of the original Swin.py script from the mmdetection platform, and a potentially lower level of understanding of the original model construction, might have resulted in patches that should have been pruned being retained with zero attention instead. This could have contributed to the overall computational cost and influenced the performance of our models. For future work, we propose to construct a fully custom Swin Transformer model, which would offer more flexibility and control over model design, and potentially ensure the efficient pruning of feature map patches.