Lightweight Transformer for Multi-Modal Object Detection (Student Abstract)

Yue Cao, Yanshuo Fan, Junchi Bin, Zheng Liu

School of Engineering, The University of British Columbia, Kelowna, BC, Canada
caoyuecc@mail.ubc.ca, zheng.liu@ubc.ca

Abstract

It has become a common practice for many perceptual systems to integrate information from multiple sensors to improve the accuracy of object detection. For example, autonomous vehicles use visible light, and infrared (IR) information to ensure that the car can cope with complex weather conditions. However, the accuracy of the algorithm is usually a trade-off between the computational complexity and memory consumption. In this study, we evaluate the performance and complexity of different fusion operators in multi-modal object detection tasks. On top of that, a Poolformer-based fusion operator (PoolFuser) is proposed to enhance the accuracy of detecting targets without compromising the efficiency of the detection framework.

Introduction

Object detection has always been a vital task in the field of computer vision. The availability of low-cost sensors enables many vision systems in the real world to use multiple sensors to enhance the reliability of models in different environments. For example, in autonomous driving, both RGB and IR information are considered, since RGB images can provide more details on color and texture. RGB images, however, fail to provide adequate information in the dim light, and IR images are adopted to counteract the deficiency in this aspect. In the framework of multi-modal object detection, the distribution of attention on different modalities is an important factor that affects the performance of the detection model. In the previous work, Woo et al. (2018) proposes a compact attention model called CBAM, which uses pooling techniques to obtain attention maps from both channel and spatial perspectives. This design allows the model to learn the assignment of attention weights without increasing the complexity of the model, but this architecture does not take into account global information. On the other hand, Chitta et al. (2022) employs Transformer (Vaswani et al. 2017) as the fusion operator (TransFuser) to assign weights to different inputs. The multi-head attention architecture in the transformer (Vaswani et al. 2017) considers the global context between the inputs, thus improving the quality of data fusion. However, Transformer (Vaswani et al. 2017) results in large memory consumption and latency, because the computational complexity of self-attention is quadratic. In this study, inspired by the work presented by Yu et al. (2022), we develop PoolFuser, a Poolformer-based fusion operator that can be used in multi-modal object detection. The operator can learn global representations without increasing the parameters of the model or computation time. In experiments, Faster-RCNN (Ren et al. 2015) is chosen as the detection framework to evaluate the performance and complexity of aforementioned fusion operators on multi-modal object detection tasks. The overall architecture is shown in Figure 1.

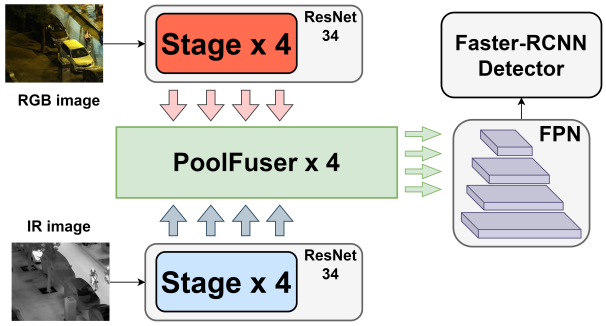

Figure 1: The overall architecture of the multi-modal object detection framework proposed in this study.

Methodology

Detection framework. In this study, Faster-RCNN (Ren et al. 2015) is adopted as the detection framework. The framework consists of two ResNet-34s (He et al. 2016) as backbone networks, which are mainly responsible for feature extraction of RGB and IR inputs. Then, the feature maps generated by the backbones are fed into the fusion blocks for attention allocation and form new weighted feature maps. Finally, FPN and Faster-R-CNN detection head (Ren et al. 2015) are applied to identify the object from the fused features.

PoolFuser. To tackle the problem of excessive memory usage caused by Transformer (Vaswani et al. 2017), a Poolformer-based (Yu et al. 2022) fusion operator referred to as PoolFuser is proposed in this study. PoolFuser uses the general structure of Transformer (Vaswani et al. 2017) as a

Fusion Operators	mAP	AP50	AP75	GPU Memory	Inference time
Mono IR	0.327	0.691	0.265	922 MB	0.0282 s
Mono RGB	0.195	0.514	0.100	922 MB	0.0291 s
Summation	0.371	0.714	**0.344**	**1323 MB**	**0.0307 s**
CBAM (Woo et al. 2018)	0.368	0.753	0.313	1378 MB	0.0437 s
TransFuser (Chitta et al. 2022)	0.369	0.780	0.297	1940 MB	0.0712 s
PoolFuser	**0.384**	**0.803**	0.317	1699 MB	0.0490 s

Table 1: Experiment results of different fusion operators.

weight allocator that consists of two normalization layers, an attention distributor, and a feed-forward network. However, pooling, a parameter-free operator, is employed instead of multi-head attention in this fusion operator as the attention distributor. PoolFuser first concatenates the feature maps produced by backbones and performs patch embedding. Then, the embedded inputs are fed into the above-mentioned weight allocator for weight allocation. The output of PoolFuser is generated by combining the two weighted feature maps using element-wise summation.

Experiments

Dataset. LLVIP (Jia et al. 2021) is a publicly available dataset that comprises a total of 15,488 pairs of RGB and IR images. The dataset is built on information gathered from dimly-lit roads, with all data aligned in time and space in a semi-manual way. All pedestrians in the image are annotated manually. During the training, the dataset is randomly divided into training and test sets based on 8:2 portions, and COCO evaluation metrics are applied for testing purposes.

Experimental Results. To verify the effectiveness of the fusion operator proposed in this paper, three fusion operators are implemented in this experiment as baseline models, including the CBAM, TransFuser (Woo et al. 2018; Chitta et al. 2022) aforementioned and a summation model, which directly combines the feature maps generated by two backbones. In addition, two mono-modality models are trained as the control groups of the previously mentioned fusion models. The experimental result in Table 1 proves that both mono-modality models struggle to obtain sufficient information from the data because of insufficient lighting, resulting in low AP50 and mAP scores. The summation method simply mixes all the information, thus slightly improving the accuracy of the model. Besides, the channel attention and spatial attention in CBAM (Woo et al. 2018) enable the detection model to identify the feature correlation between the two modalities more accurately, thus improving the AP50 to 0.735. Finally, TransFuser (Chitta et al. 2022) and PoolFuser further improve the performance of the detection model, obtaining 0.780 and 0.803 on AP50 respectively, with global information integrated during the training process. Besides evaluating accuracy, we also look into the impact of the four fusion methods on memory usage and inference time. Because of the property of multi-head attention, TransFuser (Chitta et al. 2022) has the longest latency and highest resource consumption compared to other models. On the other hand, calculating the attention weight by pooling can greatly reduce the parameters and computational complexity of the model, so PoolFuser shows a better performance than TransFuser (Chitta et al. 2022) in this regard. In general, PoolFuser has the best balance of detection accuracy and computational resource usage among all models in this experiment.

Conclusion

This paper evaluates the performance of fusion operators such as CBAM and TransFuser (Woo et al. 2018; Chitta et al. 2022) in object detection tasks and compares them with the PoolFuser proposed in this study. The experimental result suggests that PoolFuser can outperform other baseline models on AP50 and mAP without using excessive computational resources. In the future, PoolFuser will be applied to other detection frameworks to build a lightweight and stable object detection model.

References

Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; and Geiger, A. 2022. TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Jia, X.; Zhu, C.; Li, M.; Tang, W.; and Zhou, W. 2021. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3496–3504.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, volume 28, 1–9.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 1–11. Curran Associates, Inc.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–17.

Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. MetaFormer Is Actually What You Need for Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10819–10829.