04|07|06|2023

Topic: InternImage: Exploring Large scale vision foundation models with deformable convutions

Abstract:

Compared to the great progress of large scale vision transformers (ViTs) in recent years, large scale models based on convolutional neutral networks (CNNs) are still in an early state. This work presents a new large-scale CNN based foundation model, termed Intern Image, which can obtain the gain from increasing parameters and training data like ViTs. Different from the recent CNNs that focus on large dense kernals, Intern Image takes deformable convulations as the core operator, so that our model not only has the large effective receptive field required for down-stream tasks such as detection and segmentation, but also has the adaptive spatial aggregation condit-ioned by input and task information. As a result, the proposed InternImage reduces the strict inductive bias of traditional CNNs and makes it possible to learn stronger and more robust patterns with large scale parameters from massive data like

ViTs. The effectiveness of our model is proven on challenging benchmarks including ImageNet, COCO and ADE20K. It is worth mentioning that InternImage-H achieved a new record 65.4 mAP on COCO test-dev and 62.9 mIoU on ADE20K outperforming current leading CNNs and ViTs.

Conclusion:

We introduce InternImage, a new large scale CNN based foundation model that can provide strong representations for versatile vision tasks, such as image classification, object detection and semantic symentation. We tune the flexible DCNv2 opeator to satisfy the requirement of foundation models and develop a series of blocks, stacking and scaling rules centered on the core operator. Extensive experiments on object detection and semantic segmentation benchmarks verify that our InternImage can obtain comparable or better performance than well defined large scale vision transformers trained with massive data showing that CNN is also a considerable choice for large scale visions foundation model research. Nonetheless, latency remains an issue for DCN based operators adapting to down stream requiremen tasks with high speed

Also, large scale ~~requirements~~ CNNs are still in their early stages of development, and we hope Intern Image can serve as a good starting point.