Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

Anantapadmanaabha Prasannakumar
University of Central Florida
Orlando, FL
anantalp@knights.ucf.edu

ABSTRACT:

Spatial pyramid pooling module or encoder-decoder structure are used in deep neural networks for semantic segmentation task. The former networks are able to encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and multiple effective fields-of-view, while the latter networks can capture sharper object boundaries by gradually recovering the spatial information. In this paper, the authors propose to combine the advantages of both methods. Specifically, the proposed model, DeepLabv3+, extends DeepLabv3 by adding a simple yet effective decoder module to refine the segmentation results, especially along object boundaries. they further explore the Xception model and apply the depthwise separable convolution to both Atrous Spatial Pyramid Pooling and decoder modules, resulting in a faster and stronger encoder-decoder network.

INTRODUCTION:

Deep neural networks such as convolutional neural networks, in a supervised learning environment, have achieved excellent breakthroughs in many Computer Vision domains such as image segmentation, image classification, and object detection. However, creating the necessary labels for supervised tasks is a laborious endeavor. In the past, self-supervised and unsupervised learning has been intently explored because of its affinity to no or very few labels and very manageable model training requirements.

The present work extends DeepLabv3 by adding a simple yet effective decoder module to recover the object boundaries, as illustrated in Fig. 1. The rich semantic information is encoded in the output of DeepLabv3, with atrous convolution allowing one to control the density of the encoder features, depending on the budget of computation resources. Furthermore, the decoder module allows detailed object boundary recovery.

The paper proposes a novel encoder-decoder structure that employs DeepLabv3 as a powerful encoder module and a simple yet effective decoder module. In other structures, one can arbitrarily control the resolution of extracted encoder features by atrous convolution to trade-off precision and runtime, which is not possible with existing encoder-decoder models. The paper adopts the Xception model for the segmentation task and applies depthwise separable convolution to both the ASPP module and the decoder module, resulting in a faster and stronger encoder-decoder network. The re-implementation shows state-of-the-art results by making use of Resnet and XceptionNet. Further, improvements were tried by making use of MobileNet.

RELATED WORK

The encoder-decoder networks have been successfully applied to many computer vision tasks, including human pose estimation, object detection, and semantic segmentation. Typically, the encoder-decoder networks contain (1) an encoder module that gradually reduces the feature maps and captures higher semantic information and a decoder module that gradually recovers the spatial information. Building on top of this idea, we propose to use DeepLabv3 as the encoder module and add a simple yet effective decoder module to obtain sharper segmentations.

Depthwise separable convolution or group convolution, a powerful operation to reduce the computation cost and the number of parameters while maintaining similar (or slightly better) performance. This operation has been adopted in many recent neural network designs. In particular, this paper explores the Xception model and show improvement in terms of both accuracy and speed for the task of semantic segmentation.

METHOD

Atrous convolution, a powerful tool that allows us to explicitly control the resolution of features computed by deep convolutional neural networks and adjust filter's field-of-view in order to capture multi-scale information, generalizes standard convolution operation. In the case of two-dimensional signals, for each location on the output feature map and a convolution filter, atrous convolution is applied over the input feature map.

Depthwise separable convolution, factorizing a standard convolution into a depthwise convolution followed by a pointwise convolution (i.e., 1 × 1 convolution), drastically reduces computation complexity. Specifically, the depthwise convolution performs a spatial convolution independently for each input channel, while the pointwise convolution is employed to combine the output from the depthwise convolution. In this work, the paper refers to the resulting convolution as atrous separable convolution and found that atrous separable convolution significantly reduces the computation complexity of the

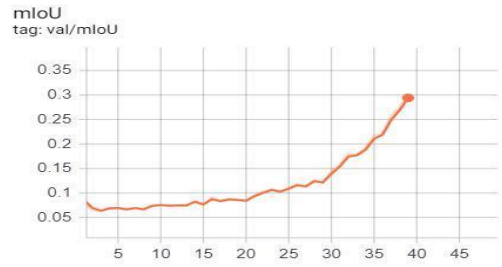
proposed model while maintaining similar (or better) performance. Additionally, DeepLabv3 augments the Atrous Spatial Pyramid Pooling module, which probes convolutional features at multiple scales by applying atrous convolution with different rates.

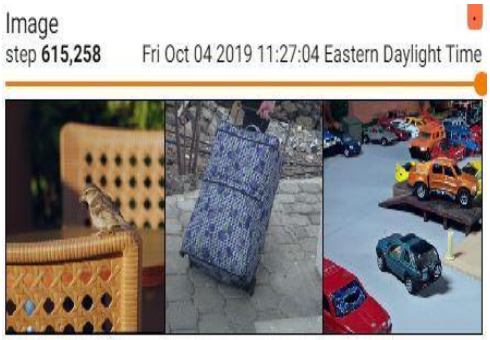
EXPERIMENTS:

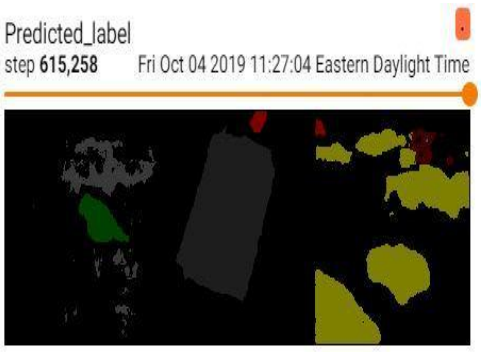
The proposed model is employed on ImageNet-1k, with pretrained ResNet-101 or modified aligned Xception being the backbone, to extract dense feature maps by atrous convolution. Further, Pretrained MobileNet is employed to test the similarities between Xception and MobileNet. The performance is measured in terms of pixel intersection-over-union averaged across the 21 classes (mIOU). Below is a brief visualization of the performance of all the 3 models used:

For ResNet101 backbone:

mIOU for ResNet101 backbone, an image sample and its prediction

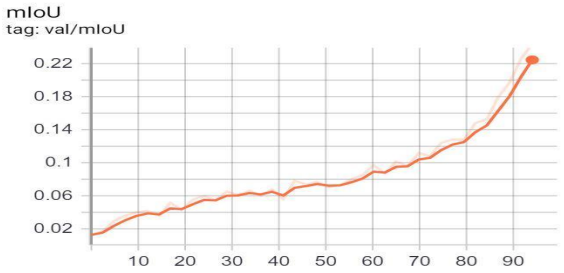


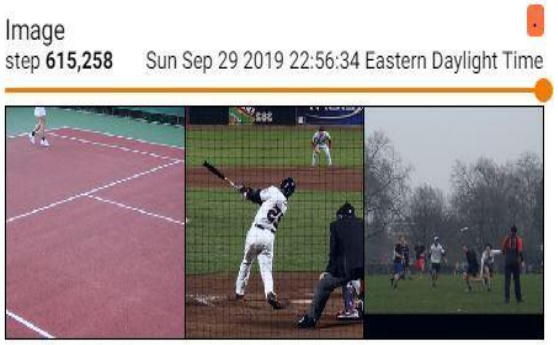




For Xception backbone:

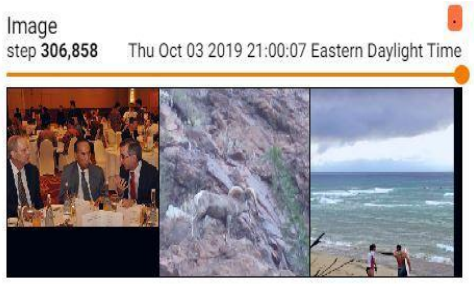
Shown below mIOU, image sample and prediction





For MobileNet backbone:

Shown below: mIOU, image sample and prediction



CONCLUSIONS:

The proposed model DeepLabv3+ employs the encoder-decoder structure where DeepLabv3 is used to encode the rich contextual information and a simple yet effective decoder module is adopted to recover the object boundaries. The atrous convolution to extract the encoder features at an arbitrary resolution can be applied too. The re-implementation also explores the Xception model, Mobilenet and atrous separable convolution to make the proposed model faster and stronger. Finally, the experimental results show that the proposed model sets a new state-of-the-art performance on PASCAL VOC 2012.

REFERENCES:

[1] Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)

[2] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR.

(2017)

[3] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)

[4] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR. (2017)

[5] Sifre, L.: Rigid-motion scattering for image classification. Ph.D. thesis (2014)

[6] Vanhoucke, V.: Learning visual representations at scale. ICLR invited talk (2014)

[7] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)

[8] Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR. (2018)

[9] Qi, H., Zhang, Z., Xiao, B., Hu, H., Cheng, B., Wei, Y., Dai, J.: Deformable convolutional networks – coco detection and segmentation challenge 2017 entry. ICCV COCO Challenge Workshop (2017)

[10] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2017)

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam: Deeplabv3+: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. TPAMI (2019)