Sky Segmentation

Liticia Touzari

January 28, 2023

1 Introduction

Segmentation has been a fundamental problem in computer vision since the early days of the field. An essential component of many visual understanding systems, it involves partitioning images (or video frames) into multiple segments and objects and plays a central role in a broad range of applications, including medical image analysis, autonomous vehicles, video surveillance, and augmented reality to name a few.
Image segmentation can be formulated as the problem of classifying pixels with semantic labels (semantic segmentation), or partitioning of individual objects (instance segmentation), or both (panoptic segmentation). Semantic segmentation performs pixel-level labeling with a set of object categories for all image pixels; thus, it is generally a more demanding undertaking than whole-image classification, which predicts a single label for the entire image.

The goal of this study is to propose a solution in order to automate the creation of segmentation masks for sky in images using pretrained deeplearning models like U-net and Deeplab.

2 Proposed models

2.1 U-net

U-Net is a convolutional neural network originally developed for segmenting biomedical images. Its architecture is made up of two parts, the contracting path and the expansive path. The purpose of the contracting path is to capture context while the role of the expansive path is to aid in precise localization. This process is completed successfully by the type of architecture built. The main idea of the implementation is to utilize successive contracting layers, which are immediately followed by the upsampling operators for achieving higher resolution outputs on the input images. With this U-Net architecture, the segmentation of images of sizes 512X512 can be computed with a modern GPU within small amounts of time.

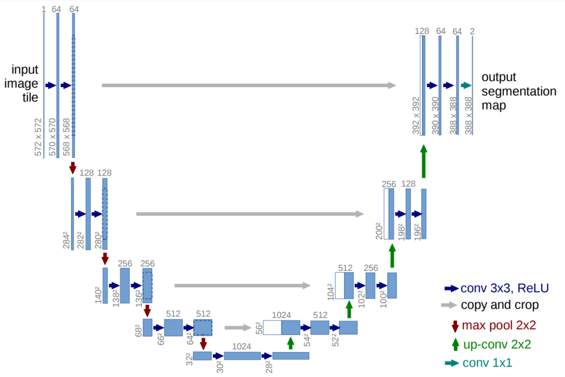


Figure 1: U-Net architecture

2.2 DeepLabV3

Dilated convolution (Atrous convolution) introduces to convolutional layers another parameter, the dilation rate, to extract more dense features where information is better preserved given objects of varying scale. For example, a 3 × 3 kernel with a dilation rate of 2 will have the same size receptive field as a 5 × 5 kernel while using only 9 parameters, thus enlarging the receptive field with no increase in computational cost.

Dilated convolutions have been popular in the field of real-time segmentation, and many recent publications report the use of this technique. Some of the most important include the DeepLab family.
The authors compare and combine several methodologies to create their finalized approach: Atrous Spatial Pyramid Pooling (ASPP). The first is cascading convolutions, which is simply convolutions proceeding each other. The second is a multi-grid method, a hierarchy of grids of different sizes. The authors also include batch normalization and image-level features. They do this by applying a global average pooling on the last feature map.

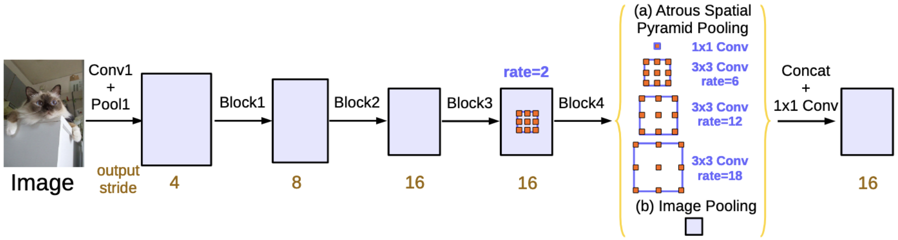


Figure 2: DeepLabv3 architecture

3 Datasets

3.1 ADE20K

The ADE20K semantic segmentation dataset contains more than 20K scene-centric images exhaustively annotated with pixel-level objects and object parts labels. There are totally 150 semantic categories, which include stuffs like sky, road, grass, and discrete objects like person, car, bed.

3.2 ADE20K-Outdoors

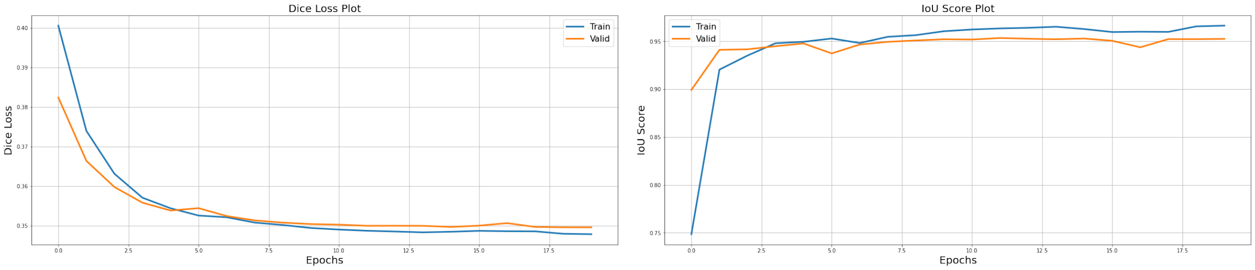
In this study, we choose to use a smaller dataset containing only outdoors images in order to have a dataset with a larger percentage of images with a sky. ADE20K-Outdoors [1] dataset is a 5,000-image subset of the 20,000-image ADE20K challenge. The selection was performed using rough class analysis with some outliers. The labels (masks) are then transformed to keep only two classes: "sky" and "other" with the RGB values [255, 255, 255] and [0, 0, 0] respectively.
We split the dataset into Train, Valid and Test sets containing respectively 60%, 20% and 20% images from the original dataset.

4 Experiments

4.1 Unet

We use a UNet segmentation model with pretrained ResNet50 encoder on imagenet dataset and finetune it on our ADE20K-Outdoors dataset to perform a sky segmentation (a binary classification of each pixel).

[1] hhttps://www.kaggle.com/datasets/residentmario/ade20k-outdoors



(a) Dice loss on Train and Valid sets

(b) IoU score on Train and Valid sets

Figure 3: Unet results on ADE20K-outdoor

The Above plots 3a and 3b show good results of Dice loss on train and validation sets around 0.35 and 0.35 respectively and an IoU score of 97.50% and 96.64% after 20 epochs.

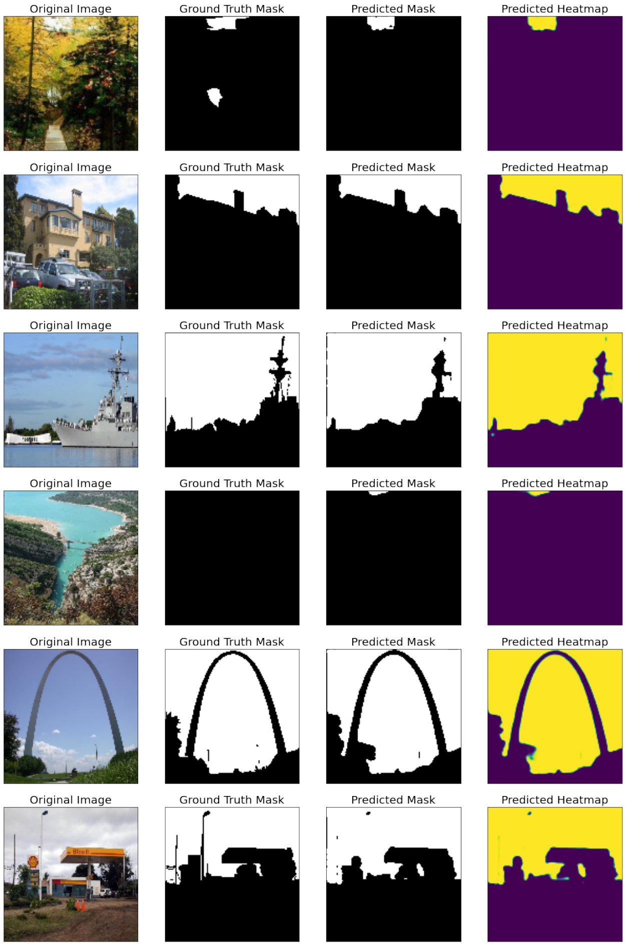
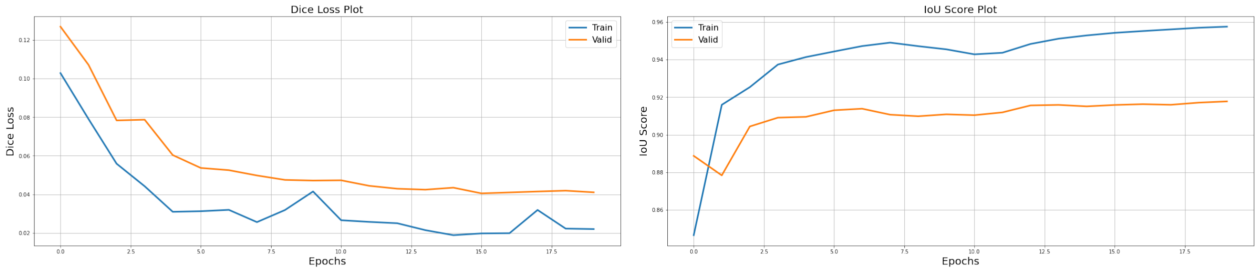


Figure 4: Unet results on Testset

The above figures 4 show the original, groud truth, predicted mask and predicted hitmap of some samples on testset. The model seems to detect well specific shapes in the image regardless of their location but details such as thin shapes and sharp tips are missing in the segmentation.

4.2 DeepLabv3

For the second experiment, we use a pre-trained DeepLabv3 segmentation model with ResNet50 as a backbone. In order to perform a sky segmentation we replace the classifier module of the model with a new DeepLabHead with a new number of output channels each corresponding to the classes "sky" and "other".



(a) Dice loss on Train and Valid sets

(b) IoU score on Train and Valid sets

Figure 5: Deeplabv3 results on ADE20K-outdoor

After training the model for 20 epochs we obtain the results shown in 5a and 5b with a Dice loss on train and validation sets around 0.04 and 0.02 respectively and an IoU score of 95.75% and 91.70% after 20 epochs.

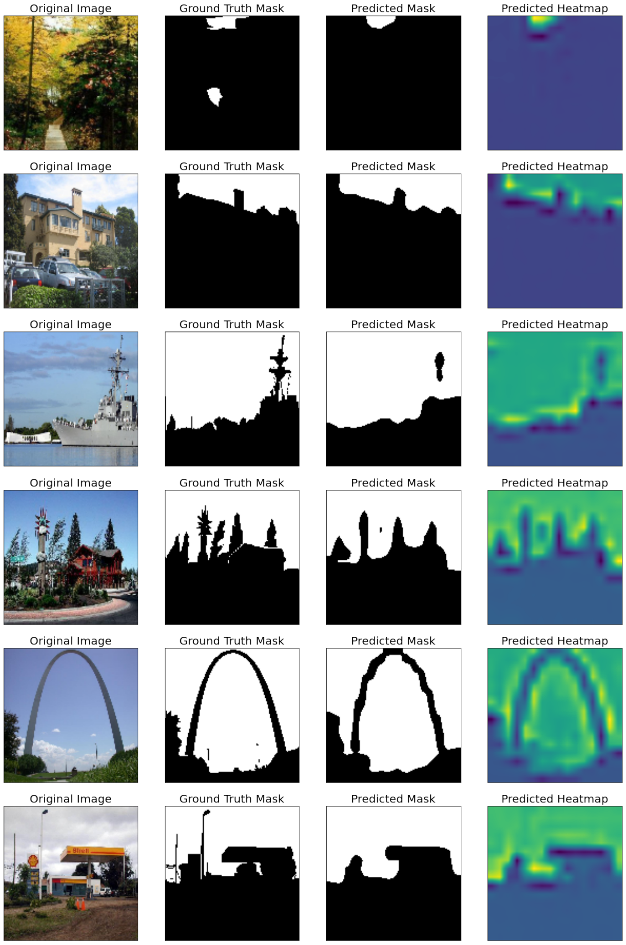


Figure 6: Deeplabv3 results on Testset

The above figures 4 show the original, groud truth, predicted mask and predicted hitmap of some samples on testset. We notice that this model detects less well specific shapes than the previous one especially sharp tips and edges.

4.3 Comparaison

Model	Auroc Score	IoU Score
Unet	80.16%	94.89%
DeepLabv3	80.02%	93.92%

Table 1: Comparison of the Results of Unet and DeepLabv3 on Testset

The scores of the models are close but Unet is slightly better at detecting complex shapes and edges. The pretrained Deeplabv3 with resenet50 backbone has only been trained for 20 classes on MS COCO dataset which explains its difficulty to detect some shapes. Using a resenet101 backbone would inprove the results.

5 Conclusion

U-Net is a widely used deep learning architecture in extracting some of the complex image-derived features that are helpful for accurate image segmentation. Addressing some of the issues mentioned before by either modifying the architecture or changing the training scheme could improve the network capability in object segmentation. DeepLabv3 is a good model as well and with DeepLabv3+ that extends DeepLabv3 by adding a simple yet effective decoder module to refine the segmentation results we obtain better results especially with Xception-71 as a backbone.