Machine Learning Algorithms for Semantic Segmentation with Convolutional Neural Networks (CNN)

STUDENTS / UNIVERSITIES

Arda Akça Büyük — **Işıl Dereli**
Ceyda Ömür — **Sena Korkut**
Gökberk Yar — **Oğuz Çelik**
Hasan Ocak

SUPERVISOR(S)

Assoc. Prof. Dr. Mehmet Keskinöz

Sabancı Üniversitesi

PURE PROGRAM FOR UNDERGRADUATE RESEARCH

ABSTRACT

An image is a set of different pixels and each pixel has many different characteristics such as color, intensity and texture. Image segmentation is a process of partitioning a digital image into multiple segments that share similar attributes. It is typically used to locate objects and boundaries in images. Pixels that are nearby to each other and share the same color or pattern or gentle gradient of brightness are grouped into a single object. In that way we create a pixel-wise mask for each object in the image to identify the shape and boundary of each object. In our project, the aim is to perceive the impact of training datasets in human segmentation and compare the accuracy of existing models trained with appropriate datasets.

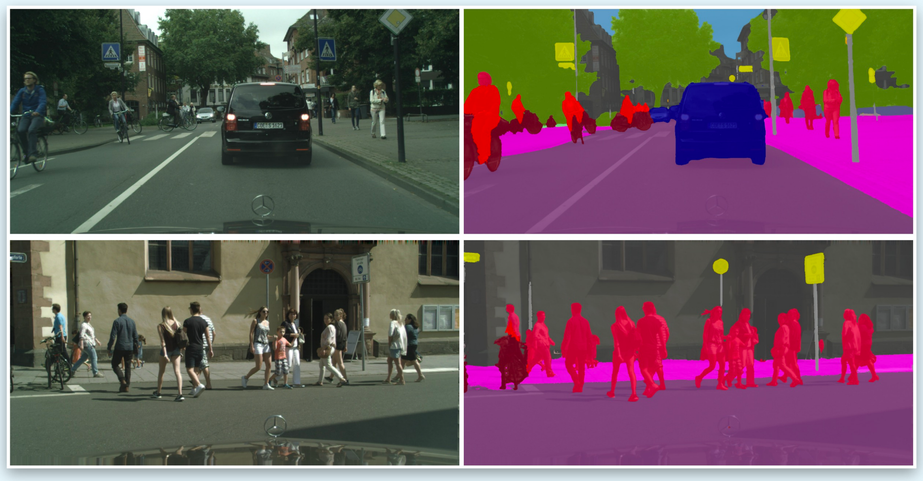

FIGURE 1: Semantic segmentation on the Cityscapes dataset [1].

METHOD

- To observe the accuracy rate for image segmentation in one model trained with different datasets, we used PSPNet101 (Pyramid Scene Parcing Network 101) with the training datasets VOC2012 (Visual Object Classes 2012) and cityscape.
- To analyze the accuracy rate in image segmentation in various depths of same model, we used PSPNet101(trained with VOC2012) and PSPNet50.
- To analyze the accuracy rate in image segmentation in different models trained on same dataset (VOC2012), we used PSPNet 101 and FCN (Fully Convolutional Network).

MODELS

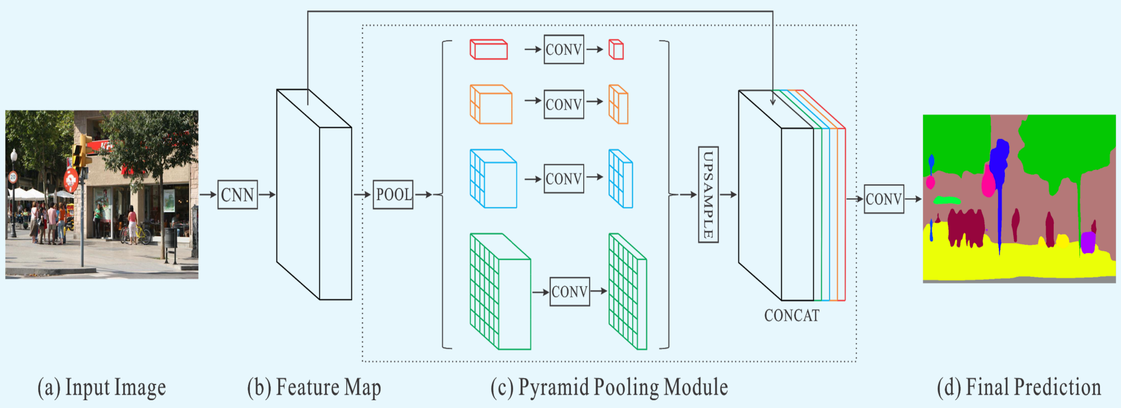

FIGURE 2: Pyramid Scene Parcing Network [2].

- Overview of PSPNet: Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d). [2]

- The difference between PSPNet 101 and 50 is the depth. PSPNet 101 has more depth, therefore it is more complex than PSPNet 50. [2]

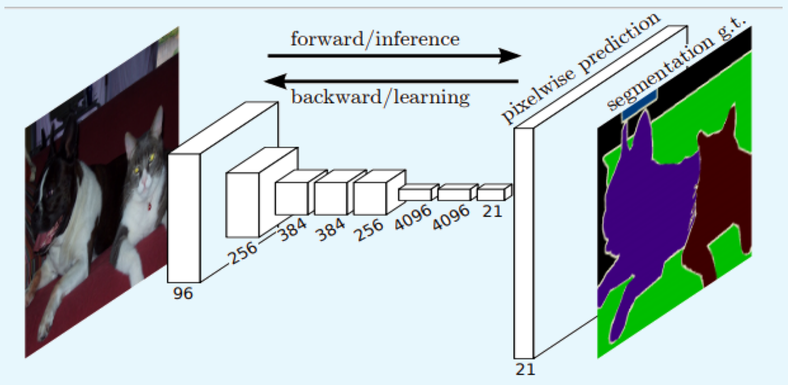

FIGURE 3: Fully Convolutional Network [3].

- Fully convolutional networks consists of consecutive convolutions (downsampling) and a following pixelwise prediction layer, therefore it can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation [3].

COMPARISON / FINDINGS

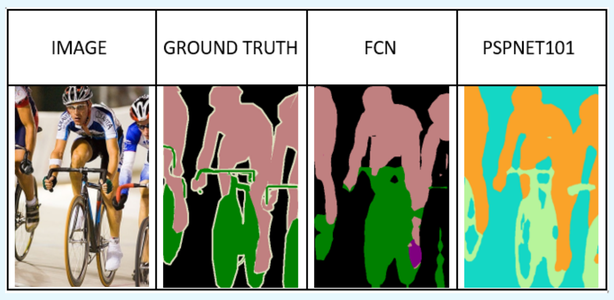

FIGURE 4: Difference between PSPNet and FCN on the VOC2012 dataset [4].

- PSPNet 101 has a better performance than FCN when they are both trained with the same dataset (VOC2012). PSPNet 101 resulted as a more accurate segmentation to the ground truth.

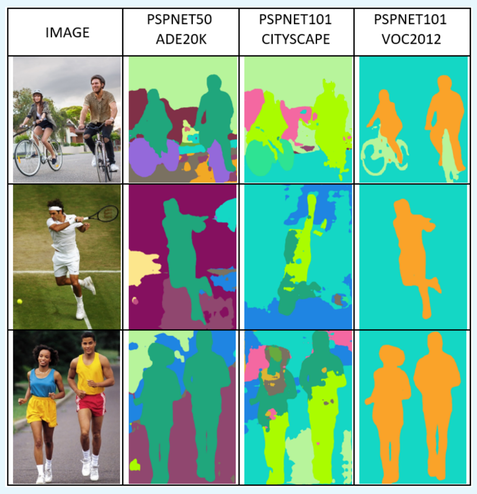

FIGURE 5: Difference between PSPNet 50 (trained with ADE20K) and PSPNet 101 model trained with two different dataset (Cityscape and VOC2012)

- PSPNet 101 has a higher accuracy when it is trained with VOC2012 dataset instead of Cityscape dataset.
- PSPNet 101 model which is trained by VOC2012 dataset has a higher accuracy than PSPNet 50 model trained with ADE20k dataset.
- PSPNet 50 model trained with ADE20k dataset has a higher accuracy than PSPNet 101 trained with Cityscape dataset.

Method	Mean IoU(%)	Pixel Acc.(%)
PSPNet(50)	41.68	80.04
PSPNet(101)	41.96	80.64

Table 1: Mean IoU (Intersection over Union) and pixel accuracy comparison of PSPNet(50) and PSPNet(101) [2].

- PSPNet 101 has a better accuracy than PSPNet 50 because deeper pre-trained model gets a higher performance. [2]

CONCLUSION

Different models trained with the same dataset (see Figure 4) and same models trained with different datasets (see Figure 5) perform distinctively. Also, as can be seen in Table 1, model depth makes no significant contribution to accuracy after a plateau point. Statistically speaking, distribution of data affects objective function's convergence degree to local minimum (under assumption all data comes from same distribution not distinguishable from global minimum) with also affecting convergence speed.

FUTURE WORK

- Try to develop our own model which has a better accuracy than the current ones by modifying old models focusing on segmenting only human figures.
- Apply our original model to an application in which human segmentation is intensely used such as portrait mode.

REFERENCES

1. A. Kundu, V. Vineet, and V. Koltun, "Feature Space Optimization for Semantic Video Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
2. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
3. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
4. aurora95, "aurora95/Keras-FCN," GitHub, 26-Jan-2018. [Online]. Available: https://github.com/aurora95/Keras-FCN. [Accessed: 01-Aug-2019].