Comparative Study of Semantic Segmentation Using FCN, Unet, and DeepLabV3

Yeowon Youn
Aiffel
South Korea
dudnjsckrgo@gmail.com

February 2,2023

1 Abstract

Semantic segmentation plays a pivotal role in computer vision, facilitating the understanding of images at a pixel level. This paper presents a comparative analysis of three leading semantic segmentation models: Fully Convolutional Networks (FCN), Unet, and DeepLabV3. Utilizing the PASCAL VOC 2012 dataset, we explore each model's efficacy, underscored by a preprocessing regimen that includes integer encoding of segmentation masks and strategic data augmentation via horizontal flipping. Our study evaluates the performance of these models through a comprehensive lens, examining their strengths and weaknesses in various semantic segmentation tasks. The insights garnered from this comparison aim to guide future research and application of semantic segmentation models in real-world scenarios.

2 Introduction

Semantic segmentation, the process of assigning a class label to each pixel in an image, is a cornerstone task in the field of computer vision. It underpins a variety of applications, from autonomous driving to medical image analysis, by enabling detailed understanding and interpretation of visual data. The evolution of deep learning has significantly advanced the state of the art in semantic segmentation, giving rise to models that can accurately parse and segment images with high precision.

Among the myriad of models developed for this task, Fully Convolutional Networks (FCN), Unet, and DeepLabV3 have emerged as benchmarks, each introducing novel architectural features and techniques to enhance segmentation performance. However, the effectiveness of these models can be contingent on the preprocessing steps applied to the data and the specific challenges presented by the dataset in use.

In this study, we focus on the PASCAL VOC 2012 dataset, a widely recognized benchmark in semantic segmentation. We detail our preprocessing approach, which includes the integer encoding of segmentation masks and the use of horizontal flipping for data augmentation, to prepare the data for model training and evaluation. Through a comparative analysis of FCN, Unet, and DeepLabV3, we aim to elucidate the relative advantages and limitations of these models, providing a nuanced understanding of their performance in semantic segmentation tasks.

Our investigation not only sheds light on the current landscape of semantic segmentation models but also seeks to inform the selection and optimization of these models for future research endeavors and practical applications in the field.

3 Related works

The concept of semantic segmentation has evolved significantly with the advent of deep learning. Early attempts were based on hand-crafted features and graphical models. The introduction of FCN marked a paradigm shift, enabling end-to-end training for pixel-wise classification. Following FCN, architectures like Unet and DeepLabV3 have built upon and refined the approach to semantic segmentation, introducing innovations such as skip connections and atrous convolutions, respectively.

3.1 Fully Convolutional Networks (FCN)

FCNs revolutionized semantic segmentation by adapting conventional CNNs for pixel-wise prediction without fully connected layers. This adaptation allows for segmentation maps of varying resolutions and has been foundational for subsequent models.

3.2 Unet

Unet further specialized the approach towards biomedical image segmentation, introducing a symmetric expanding path that improves localization through precise boundary delineation, critical in medical imaging.

3.3 DeepLabV3

DeepLabV3, with its atrous convolutions and spatial pyramid pooling, has enhanced the model's ability to capture multi-scale information, making it particularly effective for segmenting objects at various scales within a single scene.

4 Method

4.1 Dataset and Preprocessing

We utilized the PASCAL VOC 2012 dataset for our study, a staple in semantic segmentation research, divided into a training set with 1,464 images and a validation set with 1,449 images. Each dataset comprises pairs of images and their corresponding segmentation masks, essential for conducting detailed semantic analyses.

A key step in our preprocessing was the encoding of segmentation masks into an integer format. This process involved the development of a mask dictionary, mapping each unique object class present in the segmentation masks to a distinct integer value. This encoding facilitates the efficient representation and handling of categorical labels during the training and evaluation of our models.

To augment our dataset, we implemented horizontal flipping on both images and masks, effectively doubling our dataset size and introducing variability that aids in the model's generalization capabilities across varied orientations of objects.

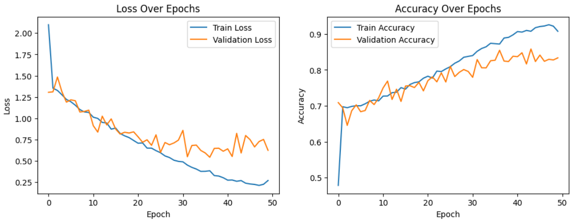

Figure 1: FCN history

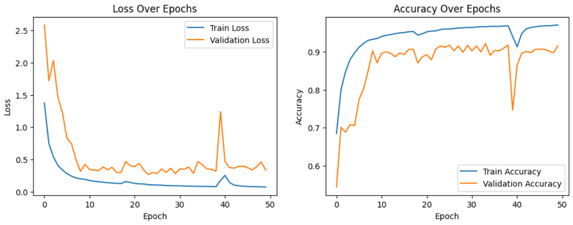

Figure 2: Deeplabv3 history

4.2 Training Approach

In our study, the Unet model plays a crucial role due to its prominence and efficacy in semantic segmentation tasks, particularly in medical image analysis. To adapt Unet for our specific requirements and the PASCAL VOC 2012 dataset, we employed a two-phase training strategy aimed at optimizing model performance while ensuring robustness and generalizability.

4.2.1 Initial Training Phase

Initially, we set the Unet model's layers to be non-trainable (*Trainable=False*). This approach allows the model to be used in a transfer learning context, where the pre-trained weights are not altered in the early stages of training. This strategy is particularly beneficial for quickly gauging the model's performance on the task without the risk of overfitting or significant computational expense. We proceed with training under this configuration, closely monitoring the model's performance through validation loss and accuracy metrics.

If the initial results, as visualized in our monitoring graphs, indicate satisfactory performance without any significant issues, we continue the training process under this fixed-weights regime. This decision is based on the premise that the pre-trained aspects of the model sufficiently capture the relevant features for semantic segmentation on our dataset.

4.2.2 Adaptive Training Phase

Should we encounter performance plateaus or degradation, indicating that the model could benefit from further adaptation to our specific dataset, we switch to an adaptive training phase. In this phase, we set the Unet model's layers to be trainable (*Trainable=True*), allowing the weights to be updated during training. This flexibility enables the model to fine-tune its parameters specifically to the nuances and intricacies of the PASCAL VOC 2012 dataset, potentially overcoming the limitations observed in the initial training phase. Additionally, we employ a dynamic approach to adjusting the learning rate throughout the training process. By carefully modulating the learning rate, we

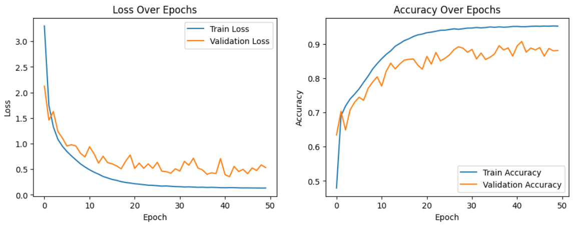

Figure 3: Unet with EffecientNet history

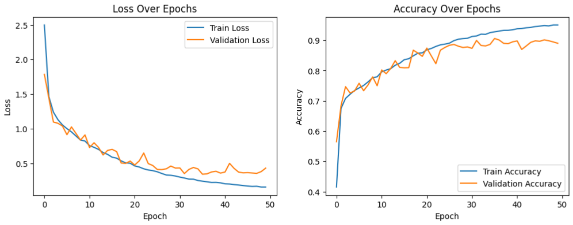

Figure 4: Unet with MobileNet history

aim to find an optimal balance that allows for steady improvement in performance without causing instability in the training dynamics. The learning rate adjustments are informed by the model's performance on the validation set, with reductions applied to mitigate overfitting or to rejuvenate progress when training plateaus are encountered.

4.3 Evaluation Metrics

To thoroughly assess the performance of the Unet model, as well as the Fully Convolutional Networks (FCN) and DeepLabV3 models, in our semantic segmentation tasks, we employ two key metrics: mean Intersection over Union (mIoU) and Pixel Accuracy (PA). These metrics are instrumental in providing a comprehensive evaluation of the models' segmentation accuracy and overall effectiveness.

4.3.1 Mean Intersection over Union (mIoU)

The mIoU metric measures the average overlap between the predicted segmentation and the ground truth across all classes. It is defined as the ratio of the intersection and the union of the predicted and ground truth masks for each class, averaged over all classes. mIoU is particularly valuable in semantic segmentation as it accounts for the balance between precision and recall, offering insight into how well each class is segmented, especially in scenarios with class imbalance.

4.3.2 Pixel Accuracy (PA)

Pixel Accuracy, on the other hand, measures the proportion of correctly classified pixels across the entire image. This metric is calculated by dividing the number of correctly predicted pixels by the total number of pixels. PA provides a straightforward indication of a model's overall segmentation accuracy but may be less informative in cases of significant class imbalance, as it could be skewed by the predominance of a particular class.

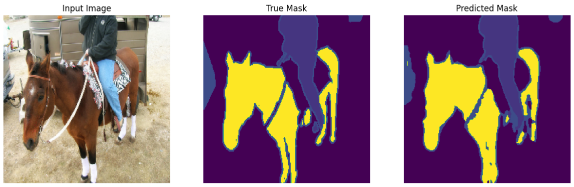

Figure 5: Deeplab prediction visualize

4.4 Evaluation Process

Our evaluation process involves applying the trained models to the validation subset of the PASCAL VOC 2012 dataset and computing both mIoU and PA metrics. This dual-metric approach allows us to capture both the detailed class-wise segmentation performance through mIoU and the general accuracy of the segmentation via PA.

The models' performances are benchmarked against these metrics to ascertain their strengths and limitations in accurately segmenting various objects within diverse scenes. Such an evaluation not only highlights the comparative effectiveness of FCN, Unet, and DeepLabV3 in semantic segmentation but also guides potential improvements and adjustments to the models and training procedures.

4.5 Interpretation of Results

The results obtained from this evaluation will be critically analyzed to understand the implications of each model's performance metrics. A higher mIoU indicates a model's superior capability in handling the intricacies of different classes, while a higher PA suggests strong overall segmentation performance. By interpreting these results in the context of our dataset and preprocessing choices, including the impact of integer mask encoding and data augmentation, we aim to derive meaningful insights that can inform future research directions and practical applications in semantic segmentation.

5 Results

In our study, we compared the performance of FCN, DeepLabV3, and Unet with two different backbones: MobileNet and EfficientNet, across various classes for semantic segmentation tasks. Our results, detailed in Table 5, demonstrate that DeepLabV3 leads with an outstanding mean IoU (mIoU) of 63.73%, showcasing its superior ability to handle diverse segmentation scenarios. Unet, when augmented with EfficientNet as a backbone, achieves a mIoU of 24.90%, whereas with MobileNet, it records a slightly higher mIoU of 44.58%. FCN, a foundational model for semantic segmentation, shows a mIoU of 24.12%.

These results highlight the impact of backbone architectures on Unet's performance. While Unet with MobileNet outperforms the version with EfficientNet in terms of mIoU, both configurations fall short of the benchmark set by DeepLabV3. The class-wise IoU scores further reveal that specific classes such as "aeroplane", "dog", and "motorbike" benefit more from the advanced features captured by the DeepLabV3 model.

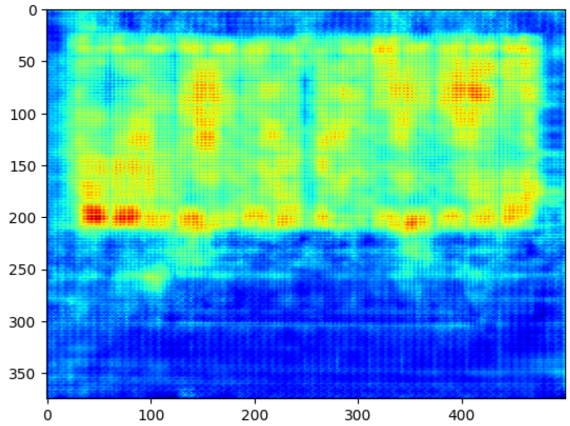

Figure 6: Unet Class Activation Map

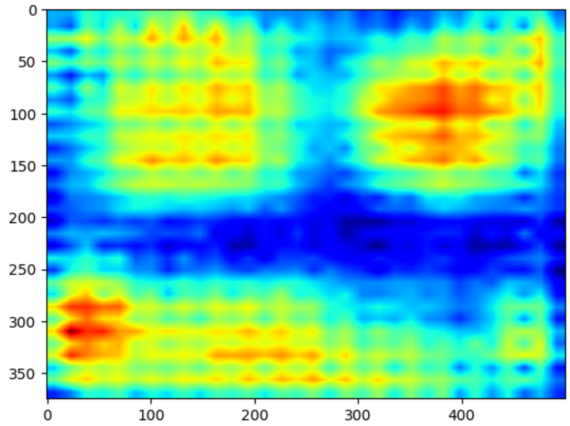

Figure 7: FCN Class Activation Map

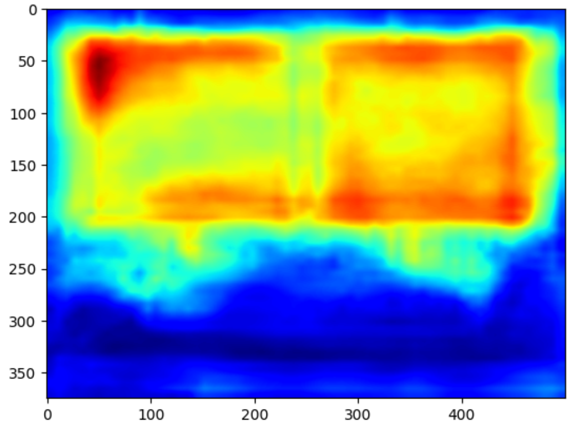

Figure 8: DeepLab Class Activation Map

Class	FCN (%)	Unet with MobileNet (%)	Unet with EfficientNet (%)	DeepLabV3 (%)
Background	86.92	88.46	90.87	91.56
Aeroplane	24.88	29.46	24.31	53.42
Person	33.47	27.90	32.69	37.58
TV Monitor	19.34	13.23	17.95	50.81
Dog	13.81	23.04	12.81	23.21
Chair	17.98	23.84	12.52	58.20
Bird	11.31	4.22	9.92	21.30
Bottle	13.76	10.00	16.16	37.86
Boat	7.30	20.95	8.18	12.29
Dining Table	9.26	18.50	14.29	44.27
Train	12.55	9.15	11.52	31.22
Motorbike	25.69	19.79	33.85	60.66
Horse	17.57	15.23	28.68	42.61
Cow	14.63	9.07	7.86	44.06
Bicycle	15.43	5.59	8.42	39.28
Car	2.71	1.98	4.63	15.08
Cat	25.31	18.00	16.71	43.49
Sofa	22.73	23.11	28.39	61.61
Bus	6.68	7.77	14.76	22.51
Potted Plant	20.30	19.88	28.64	32.38
Sheep	10.87	6.04	10.08	17.48
mIoU	24.12	44.58	24.90	63.73

Table 1: Class-wise IoU Scores and mIoU for Different Models

Model	mIoU (%)	PA	Dataset
FCN	24.12	84.8	PASCAL VOC 2012
Unet with effecientnet	24.90	89.1	PASCAL VOC 2012
Unet with mobilenet	44.58	89.9	PASCAL VOC 2012
DeepLabV3	63.73	91.7	PASCAL VOC 2012

Table 2: Comparison of mIoU, PA Performance

6 Conclusion

This comparative study elucidates the crucial role of backbone networks in enhancing the performance of Unet for semantic segmentation tasks. DeepLabV3 emerges as the most effective model, owing to its comprehensive architecture that adeptly captures multi-scale information and delivers high accuracy across a broad spectrum of classes.

The disparity in performance between Unet models utilizing MobileNet and EfficientNet backbones underscores the importance of selecting an appropriate backbone that complements the segmentation task at hand. MobileNet, with its balance of efficiency and accuracy, provides a significant boost to Unet, making it a strong contender for applications requiring both speed and precision.

EfficientNet, while slightly lagging behind in this comparison, still offers valuable contributions, especially in scenarios where model scalability and fine-grained feature representation are critical. The performance of FCN, though not at the forefront, remains commendable, highlighting its enduring relevance in the field of semantic segmentation.

Our findings advocate for a nuanced approach to model selection, where the specific characteristics of the segmentation task dictate the choice of both the segmentation framework and its backbone network. Future research will focus on exploring further combinations of backbones and segmentation models to unlock new levels of accuracy and efficiency, catering to the ever-evolving demands of computer vision applications.

In conclusion, the integration of Unet with different backbone networks presents a versatile framework for semantic segmentation, offering a balance between accuracy and computational demands. This study provides a foundation for future explorations into optimizing segmentation models for a wide array of applications.

References

[1] Olaf Ronneberger, Philipp Fischer, Thomas Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation.* arXiv:1505.04597 (2015)

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam. *Rethinking Atrous Convolution for Semantic Image Segmentation*

[3] Jonathan Long, Evan Shelhamer, Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*