Pascal VOC 2012 and Deep Learning: A Segmentation Model Comparison at DL THON

LeeSeungJe
AIFFEL
South Korea
happybin2013@gmail.com

Abstract

This study conducts a comparison of segmentation performance among deep learning-based models, including FCN, U-Net with EfficientNet, U-Net with MobileNet, and DeepLab, using the Pascal VOC 2012 dataset. Through experimentation and analysis, it demonstrates a progressive improvement in segmentation accuracy across the models, with DeepLab, in particular, showing superior performance. Please check our GitHub for the details of the experiment. https://github.com/aiffel-smile-maker/aiffel_DLThon_RS7/tree/main.

1 Introduction

In this study, we embark on a comprehensive examination of the forefront in image segmentation technologies by leveraging the Pascal VOC 2012 dataset, a benchmark that has been pivotal in advancing computer vision research. The focus is on a comparative analysis between models employing Deeplab and U-Net architectures, both of which are renowned for their exceptional performance in semantic segmentation tasks. These models are further enhanced by integrating various backbone networks, aiming to explore how these combinations influence overall segmentation accuracy and efficiency.

2 Approach

2.1 Pascal VOC 2012 Dataset for Segmentation

This research aims to harness the capabilities of deep learning for advancing image segmentation, utilizing the Pascal VOC 2012 dataset as a primary resource. Recognized for its diversity and complexity, the dataset comprises a wide range of images across different categories, making it an ideal benchmark for evaluating segmentation models. Our approach involves applying sophisticated segmentation techniques to this dataset to extract meaningful patterns and insights, which can contribute significantly to the development of more accurate and efficient segmentation solutions.

2.2 Model Selection and Implementation

Given the necessity to obtain results within a short timeframe, our strategy incorporates the use of U-Net architecture combined with pre-trained EfficientNet and MobileNet backbones sourced from ImageNet. These models were selected for their proven efficiency and effectiveness in handling complex image segmentation tasks. The EfficientNet backbone is chosen for its scalability and balance between accuracy and computational resources, while MobileNet offers advantages in terms of speed and size, making it suitable for applications requiring rapid processing.

In addition to these, we plan to implement and evaluate the performance of the Deeplab model, which has demonstrated superior results on the Pascal VOC 2012 test dataset. Deeplab's success is attributed to its advanced features, such as atrous convolution and spatial pyramid pooling, which enable precise segmentation of objects at various scales. By comparing these models, we aim to identify the most effective approach for achieving high-quality segmentation results promptly.

This multi-faceted approach allows us to not only assess the performance of different architectures and pre-training strategies but also to explore their synergistic effects on the segmentation task. Our goal is to determine the optimal model configuration that can deliver the best balance of speed, accuracy, and computational efficiency, ultimately enhancing the applicability of deep learning techniques in real-world segmentation scenarios.

3 Methods

3.1 Preprocessing

In the initial phase of our research, an extensive evaluation of the available dataset was conducted to ascertain the feasibility of conducting our segmentation experiments. The Pascal VOC 2012 dataset, which was initially considered for this study, presented a significant challenge due to the limited quantity of training data, comprising only 1,464 images. To overcome this limitation and enhance the robustness of our model training, we opted to augment the training dataset by incorporating the test dataset, consisting of 1,449 images, into our training procedure. This amalgamation resulted in a consolidated dataset, which was then strategically divided into training, validation, and test sets with proportions of 70%, 20%, and 10%, respectively. This distribution was carefully chosen to maximize the training data while still retaining sufficient data for validation and independent testing of the model's performance.

A notable challenge encountered during the dataset preparation was the absence of explicit class information within the dataset. To address this, we devised a method to derive class labels from the color coding present in the segmentation masks. This innovative approach enabled us to effectively create a mapping between mask colors and corresponding class labels, facilitating the accurate classification of different objects within the images.

3.2 Model

Our study employs three distinct models to evaluate their performance in image segmentation tasks on the enhanced Pascal VOC 2012 dataset:

Deeplab: Known for its advanced segmentation capabilities, Deeplab utilizes atrous convolution and spatial pyramid pooling to efficiently segment objects at various scales. Its architecture is designed to capture rich contextual information, making it highly effective for semantic segmentation tasks.

U-Net with EfficientNet Backbone: This model combines the proven architecture of U-Net with the efficiency and scalability of EfficientNet as its backbone. The integration of EfficientNet allows for a more refined feature extraction process, contributing to the model's enhanced ability to delineate complex image content.

U-Net with MobileNet Backbone: Similar to the U-Net with EfficientNet, this variant employs MobileNet as the backbone, offering a balance between speed and accuracy. The lightweight nature of MobileNet makes this model particularly well-suited for applications where computational resources are limited.

3.3 Segmentation Metric

To quantitatively assess the performance of our models, we utilized the mean Intersection over Union (mIoU) as the primary segmentation metric. mIoU provides a comprehensive measure of the model's ability to accurately segment various classes within an image by calculating the average ratio between the intersection and the union of the predicted and ground truth segmentation masks across all classes. This metric is critical for evaluating the precision of segmentation models, as it considers both the accuracy and the completeness of the predicted segmentation areas, making it an ideal standard for comparing the efficacy of different models in semantic segmentation tasks.

4 Results

In our study, we conducted segmentation using the models detailed from Figure 1 to Figure 3. Through this analysis, it was observed that the model with the EfficientNet backbone displayed the least fa-

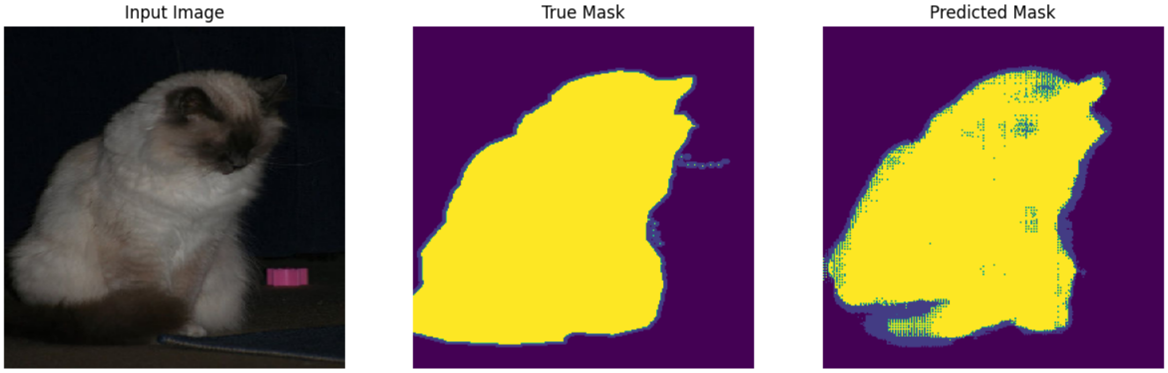

Figure 1: U-Net with EfficientNet Backbone

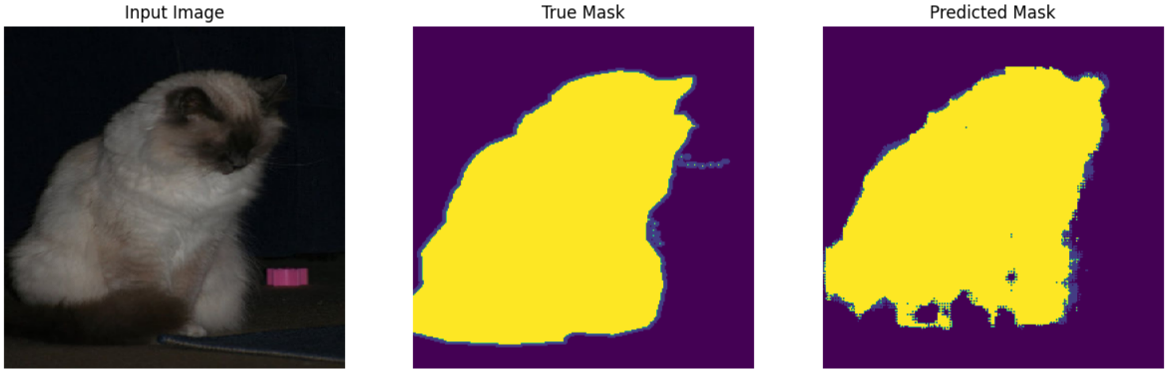

Figure 2: U-Net with Mobile Backbone

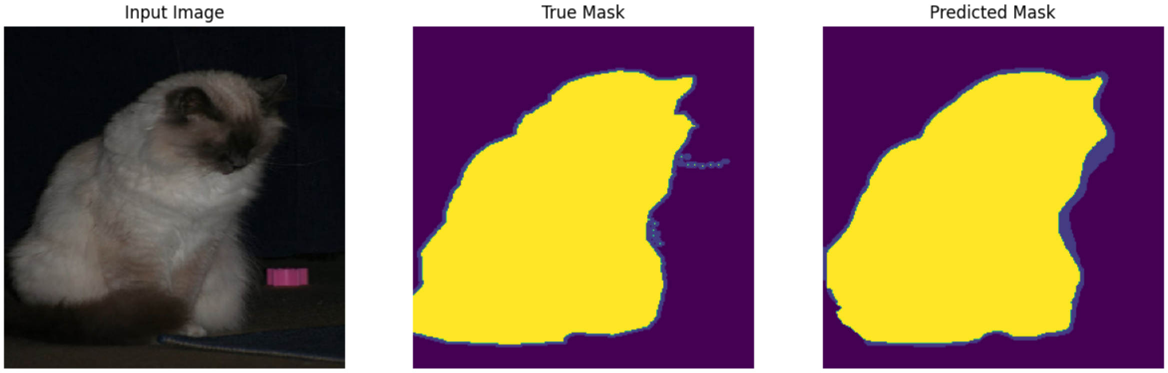

Figure 3: DeepLab

vorable performance, while the Deeplab model demonstrated the most superior results in terms of segmentation accuracy.

As evident from Table 1, we analyzed the class-wise Intersection over Union (IoU) for each model, revealing that the average segmentation performance of FCN and Deeplab models was commendable. However, it was observed that Deeplab surpassed all other models in overall performance. Additionally, the occurrence of many values below 0.1 in the remaining models suggests that they were not as effectively trained, indicating a significant disparity in the learning outcomes across the different models.

Finally, as illustrated in Table 2, by examining the mean Intersection over Union (mIoU) and Pixel Accuracy (PA) metrics, we observed a progressive improvement in performance from FCN to DeepLab models. This trend underscores the advancements in model architecture and training techniques, with DeepLab showcasing the pinnacle of segmentation efficacy. The data distinctly highlights the evolutionary path of segmentation models, with each subsequent model building on the strengths and

Class	FCN	EfficientNet	MobileNet	DeepLab
Background	0.869	0.909	0.885	0.916
Aeroplane	0.249	0.243	0.295	0.534
Person	0.335	0.327	0.279	0.376
TV Monitor	0.193	0.179	0.132	0.508
Dog	0.138	0.128	0.230	0.232
Chair	0.180	0.125	0.238	0.582
Bird	0.113	0.099	0.042	0.213
Bottle	0.138	0.162	0.100	0.379
Boat	0.073	0.082	0.210	0.123
Diningtable	0.093	0.143	0.185	0.443
Train	0.126	0.115	0.092	0.312
Motorbike	0.257	0.338	0.198	0.607
Horse	0.176	0.287	0.152	0.426
Cow	0.146	0.079	0.091	0.441
Bicycle	0.154	0.084	0.056	0.393
Car	0.027	0.046	0.020	0.151
Cat	0.253	0.167	0.180	0.435
Sofa	0.227	0.284	0.231	0.616
Bus	0.067	0.148	0.078	0.225
Pottedplant	0.203	0.286	0.199	0.324
Sheep	0.109	0.101	0.060	0.175

Table 1: Class IoU Comparison Across Models

Model	mIoU (%)	Accuracy (%)
FCN	0.241	0.8445
Unet-EfficientNet	0.249	0.8756
Unet-Mobile	0.446	0.8936
DeepLab	0.637	0.9167

Table 2: mIoU of segmentation models

addressing the limitations of its predecessors, culminating in DeepLab's superior performance.

5 Conculusion

This research presents a thorough comparison of segmentation models on the Pascal VOC 2012 dataset, highlighting the performance of FCN, U-Net with EfficientNet, U-Net with MobileNet, and DeepLab models. Our findings indicate a progressive improvement in segmentation accuracy, culminating with DeepLab exhibiting the highest efficacy. This is evidenced by our analysis of class-wise Intersection over Union (IoU) and the mean IoU (mIoU) along with Pixel Accuracy (PA), where DeepLab consistently outperformed other models, demonstrating its superior ability to capture detailed contextual information and accurately segment various classes within an image.

The observed performance disparity among the models underscores the importance of selecting an appropriate architecture and backbone for specific segmentation tasks, with DeepLab's advanced features such as atrous convolution and spatial pyramid pooling playing a pivotal role in its success. Additionally, the inclusion of test data in training due to the initial scarcity of training samples, and the innovative approach to class label derivation from segmentation masks, were key factors in enhancing model training and segmentation accuracy.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, vol 9351. Springer, Cham, 2015, pp. 234–241.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, 2018, pp. 834–848.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[5] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017.