A Comparative Study of DeeplabV3Plus and U-Net for Semantic Segmentation in Nighttime Motorcycle Ride Images

Heewon Seo
Aiffel
Seoul, South Korea
a1e0heewon@gmail.com

Abstract

With the advancement of autonomous vehicles, the significance of visual information processing technology has been steadily increasing. Nighttime driving environments, in particular, pose unique challenges, necessitating precise scene analysis. This research conducts a comparative study of DeeplabV3Plus and U-Net, two prominent deep learning models, for Semantic Segmentation tasks on nighttime motorcycle ride images.

The study utilizes the "Motorcycle Night Ride Dataset," comprising 200 original images paired with their corresponding labeled images. The labeling process was carried out using SuperAnnotate, resulting in six distinct classes: "Undrivable," "My bike," "Road," "Rider," "Lanemark," and "Movable."

Data preprocessing encompasses one-hot encoding of label images, resizing of images, and data augmentation techniques. An experimental setup was established to train and evaluate both models, facilitating a comprehensive performance comparison.

The comparative analysis of Semantic Segmentation performance demonstrates that both DeeplabV3Plus and U-Net achieve high segmentation accuracy. However, DeeplabV3Plus consistently outperforms U-Net, particularly in terms of the average Intersection over Union (IoU) metric. These findings underscore the effectiveness of DeeplabV3Plus in high-resolution image segmentation tasks, highlighting its applicability in autonomous driving and road safety applications.

1 Introduction

Ensuring safe nighttime driving has always been a paramount concern on the roads, with the need for innovative technologies becoming increasingly apparent. Motorcyclists, in particular, face unique safety challenges when navigating roads at night. Recognizing visual obstacles and taking timely actions are critical factors for motorcycle riders and overall traffic safety.

Recent advancements in computer vision and deep learning technologies offer promising opportunities to enhance road safety by employing image analysis and segmentation techniques. In the context of nighttime motorcycle rides, semantic segmentation technology can significantly contribute to improved road understanding and the early detection of potential hazards.

This study delves into the application of two prominent deep learning models, DeeplabV3Plus and U-Net, for semantic segmentation tasks in nighttime motorcycle ride scenarios. Both models are expected to accurately distinguish objects from backgrounds within images, facilitating precise object localization. The ultimate goal is to enhance safety for motorcycle riders and contribute to the advancement of autonomous driving technology.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

This paper conducts a comprehensive comparative study between DeeplabV3Plus and U-Net to assess their performance in semantic segmentation. The results validate the efficacy of these technologies in nighttime motorcycle riding scenarios.

2 Data Preprocessing

In this study, the "Motorcycle Night Ride Dataset" was employed, consisting of scenes extracted from nighttime motorcycle riding videos. The dataset comprises 200 original images, each paired with a corresponding labeled image. The labeling process was carried out using SuperAnnotate and resulted in six distinct classes: "Undrivable," "My bike," "Road," "Rider," "Lanemark," and "Movable."

Data preprocessing involved the following steps:

2.1 One-Hot Encoding

Label images were transformed into one-hot encoding format. Each pixel within the label images was compared to the colors representing the six classes. Pixels were encoded as binary values corresponding to their respective classes, effectively converting the images into segmentation maps.

2.2 Image Resizing

Original images, initially provided at a resolution of 1920x1080 pixels, were resized to 256x256 pixels. The resize operation was conducted while preserving the aspect ratio of the images and ensuring that content was maintained without distortion.

3 Method

3.1 Model Architectures

In this study, we employed two distinct model architectures, DeeplabV3Plus and U-net, for the task of semantic segmentation of nighttime motorcycle ride scenes.

3.1.1 DeeplabV3Plus Architecture

DeeplabV3Plus is a state-of-the-art deep learning architecture specifically designed for semantic segmentation tasks. It combines both optical and spatial streams to achieve accurate segmentation results.

The *optical stream* of DeeplabV3Plus is based on the Xception backbone network, a highly effective feature extractor. In addition, it utilizes the ASPP (Atrous Spatial Pyramid Pooling) module, which enables multi-scale feature extraction by applying different dilation rates to convolutional filters. This allows the model to capture information at various object scales.

The *spatial stream* complements the optical stream and is responsible for refining segmentation details. It fuses high-resolution features from the optical stream with the multi-scale features obtained from the ASPP module. This integration leads to the final semantic segmentation output.

3.1.2 U-net Architecture

U-net is a classic architecture that has proven to be effective for semantic segmentation tasks. It is characterized by its U-shaped architecture, consisting of an encoder and a decoder, connected by skip connections.

The *encoder* part of the U-net downsamples the input image, extracting abstract features using a series of convolutional and pooling layers. This process reduces the spatial dimensions while capturing hierarchical features.

The *decoder* part of U-net takes the encoded features and progressively upsamples them to match the original input resolution. Skip connections between the encoder and decoder are crucial for preserving fine-grained spatial information. These connections help in combining high-resolution details with contextual information, ultimately generating the segmentation map.

By utilizing both DeeplabV3Plus and U-net architectures, we aim to compare their performance on the task of semantic segmentation of nighttime motorcycle ride scenes. The next sections will detail the experimental setup, results, and discussion of our findings.

3.2 Hyperparameter Configuration

In our experiments, we carefully configured the following hyperparameters:

3.2.1 Learning Rate

The learning rate determines the step size during optimization. We set the learning rate to **1e-3**.

3.2.2 Batch Size

The batch size specifies the number of samples processed in each forward and backward pass. We utilized a batch size of **32**.

3.2.3 Number of Epochs

The number of training epochs indicates how many times the entire dataset was processed during training. Our models were trained for **60** epochs.

3.2.4 Optimizer

For optimization, we employed the **Adam** optimizer.

3.2.5 Loss Function

The loss function is a crucial component of the training process. We used **Categorical-Crossentropy** as our loss function.

The remaining hyperparameters were set to the default values provided by Keras.

These hyperparameter settings served as the basis for our experiments, ensuring a consistent and well-controlled training process. The subsequent sections will detail our experimental setup, results, and discussion.

3.3 Data Split

In our experiments, we partitioned the dataset into three distinct subsets for training, validation, and testing purposes, following a standard practice in machine learning:

- **Training Data (Train)**: We allocated a total of **140** images to the training dataset. This subset was used to train our models and enable them to learn from the available data.
- **Validation Data (Validation)**: We set aside **30** images for the validation dataset. This subset played a critical role in hyperparameter tuning and model evaluation during training.
- **Test Data (Test)**: For the final evaluation of model performance, we reserved **30** additional images in the test dataset. This subset was not used during training or validation and served to assess the model's generalization to unseen data.

The division of data into these subsets allowed us to conduct comprehensive experiments and evaluate the effectiveness of our models.

3.4 Data Augmentation

To enhance the diversity of our training dataset and improve the generalization capability of our models, we applied data augmentation techniques. The data augmentation process involved the following transformations, which were randomly applied to the training dataset:

Horizontal Flip We horizontally flipped a randomly selected subset of images. This augmentation leverages the left-right symmetry in the data and effectively increased the dataset size.

Vertical Flip Another subset of randomly chosen images was subjected to vertical flipping. This augmentation further diversified the dataset and contributed to more robust model training.

Brightness Adjustment We applied random brightness adjustments to a separate set of images. This augmentation aimed to improve the model's performance under varying lighting conditions.

Through these data augmentation strategies, we substantially expanded the training dataset, increasing its size from **140** to **247** images. This augmented dataset was instrumental in training more effective and robust models.

3.5 Model Training and Evaluation

In this section, we present the training and evaluation results for our models, DeeplabV3Plus and U-Net. The evaluation metrics for both models are as follows:

- DeeplabV3Plus:
  - Validation Loss: 0.4206
  - Validation Accuracy: 0.8945
- U-Net:
  - Validation Loss: 1.0041
  - Validation Accuracy: 0.7770

The validation accuracy results indicate that there is not a significant difference in performance between the two models. This suggests that both models achieved comparable accuracy on the validation dataset.

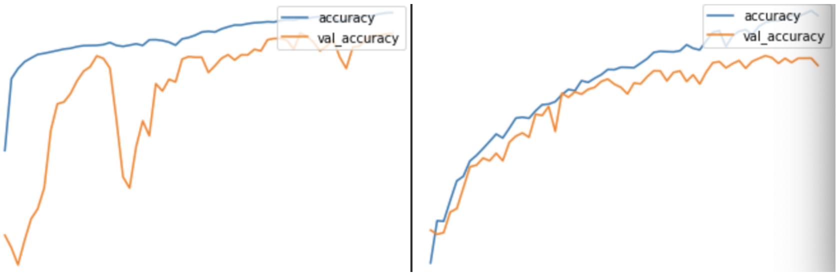

Figure 1: Training Graph (Figure 1)

4 Results

In this section, we present the evaluation results of our models using Intersection over Union (IoU) as a critical metric. IoU measures the overlap between the predicted and ground truth masks, providing insights into the segmentation accuracy.

For our evaluation, we computed the average IoU for both DeeplabV3Plus and U-Net on the validation dataset, which consists of 29 images. The results are as follows:

- **DeeplabV3Plus**: Average IoU of 0.548
- **U-Net**: Average IoU of 0.520

The IoU scores offer valuable insights into the segmentation performance of the two models. DeeplabV3Plus achieved a higher average IoU of 0.548, indicating its superior ability to accurately capture object boundaries and shapes in the images. U-Net, with an average IoU of 0.520, also demonstrated competitive performance, although slightly lower than DeeplabV3Plus.

These results suggest that both models exhibit strong segmentation capabilities, with DeeplabV3Plus showing a slightly better performance in terms of IoU. The choice between the two models should be guided by specific application requirements and the importance of segmentation accuracy.

5 Conclusion

In this study, we conducted a comparative analysis of two semantic segmentation models, DeeplabV3Plus and U-Net, in the context of motorcycle night ride scene understanding. The evaluation was based on a validation dataset consisting of 29 images, where we used Intersection over Union (IoU) as a key metric to assess segmentation accuracy.

Our findings indicate that both models exhibited strong segmentation capabilities, with DeeplabV3Plus achieving a slightly higher average IoU of 0.548 compared to U-Net's 0.520. These results suggest that DeeplabV3Plus excels in capturing object boundaries and shapes, making it an effective choice for semantic segmentation tasks in low-light conditions.

However, it is worth noting that U-Net, despite its lower IoU score, demonstrated competitive performance in terms of validation loss and accuracy. Additionally, its training dynamics, as depicted in Figure 1, reveal the model's adaptability and ability to generalize effectively.

The choice between DeeplabV3Plus and U-Net should be made based on specific project requirements and trade-offs. If the highest segmentation accuracy is crucial, DeeplabV3Plus may be preferred. On the other hand, if training stability and computational efficiency are primary concerns, U-Net offers a competitive alternative.

In conclusion, our study provides valuable insights into the strengths and trade-offs of these segmentation models, offering guidance for researchers and practitioners working on semantic segmentation tasks, particularly in challenging low-light conditions.

References

[1] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer.

[3] 5. Motorcycle Night Ride Dataset. Retrieved from kaggle motorcycle-night-ride-semantic-segmentation