Unet Mixed-and-Mingled Further for Semantic Segmentation

First Author
Inseop Byeon
Institution 1 address
isbyeon@gmail.com

Second Author
ChatGPT
Institution 2 address
chatgpt@openai.com

Abstract

We enhanced our autonomous driving system's semantic segmentation by replacing U-Net with UNet++. This model uses dense connections to improve feature propagation and segmentation accuracy. Additionally, deep supervision was incorporated to address the vanishing gradient problem and improve performance. By combining Binary Cross-Entropy and Dice Loss, we effectively handled class imbalance and pixel-wise discrepancies. UNet++ showed notable improvements over U-Net, though further adjustments and more extensive testing are needed for thorough evaluation.

1. Introduction

In the early stages of autonomous driving, we have been using U-Net for semantic segmentation. However, achieving significant performance improvements has been challenging. Therefore, we are considering exploring alternative approaches to enhance performance.

2. Method

In this study, we employ the UNet++ architecture [Figure 1], which offers several advantages over the traditional UNet model. UNet++ enhances the original UNet by redesigning the skip pathways to include dense connections, thereby improving feature propagation and reducing semantic gaps between encoder and decoder features. This architectural enhancement leads to more accurate segmentation outcomes, particularly in complex tasks.

Additionally, we utilize deep supervision in UNet++, which involves incorporating auxiliary outputs at multiple levels of the network. Deep supervision helps in mitigating the vanishing gradient problem and promotes the learning of more discriminative features across different layers. By leveraging deep supervision, our model is expected to achieve better convergence and improved performance.

We make use of a combination of Binary Cross-Entropy (BCE) and Dice Loss. BCE loss is effective in handling binary classification tasks by measuring the pixel-wise difference between the predicted and ground truth masks. Dice Loss, on the other hand, is particularly useful for handling class imbalance and ensuring better overlap between the predicted and actual segmented regions. The combination of these two loss functions provides a comprehensive objective for training, enhancing the overall performance of the segmentation model.

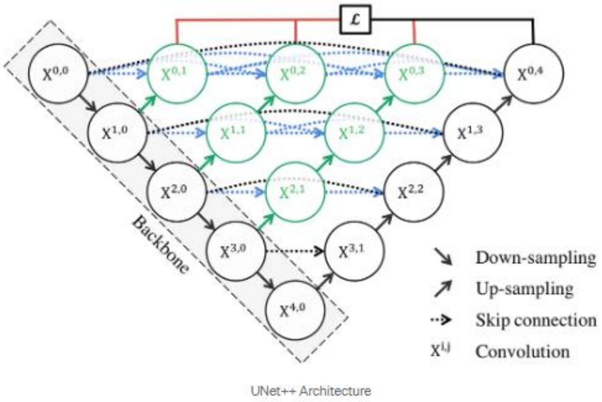

Figure 1: Unet++ structure

2.1. Deep Supervision

Deep Supervision is a technique used in deep learning models, particularly in architectures like U-Net, where additional loss functions are applied at intermediate layers of the network. This method helps the model learn more useful features at various levels and can aid in training deeper networks more effectively.

[Formula for Deep Supervision]

Let $f(x)$ represent the output of the network, and

$f_i(x)$ represent the predictions at intermediate layer i. The total loss function of the model with deep supervision is defined as:

$$L_{total} = L_{final}(f(x), y) + \sum_{i=1}^{n} \lambda_i \, L_{intermediate}(f_i(x), y)$$

where:

$L_{final}(f(x), y)$ *is the loss function calculated from the final output of the network.*

$L_{intermediate}(f_i(x), y)$ *is the loss function calculated at intermediate layer i.*

λ_i *is the weight associated with the loss from intermediate layer i.*

y *is the ground truth label.*

In this formula, the loss functions from intermediate layers are added to the overall loss function, o the network to learn useful features from these layers as well. Deep supervision can make the training process **more** stable and efficient, potentially leading to better performance, especially in deeper networks.

2.2. BCE Dice Loss

Deep Combining BCE Loss with Dice Loss, known as BCE-Dice Loss, allows the model to benefit from both pixel-wise accuracy and overall region overlap. BCE Loss provides detailed pixel-wise error metrics, while Dice Loss emphasizes the quality of segmentation in the presence of class imbalance. This combination helps in training a more robust model that performs well on both individual pixels and larger segmented regions.

The BCE-Dice Loss can be expressed as:

$$\text{BCE-Dice Loss} = \text{BCE Loss} + \lambda \times \text{Dice Loss}$$

Where λ is a weighting factor that balances the contributions of BCE and Dice Loss.

2.3. Conditions

In this experiment, we trained three models: U-Net, U-Net++, and U-Net++ with deep supervision. For all models, the training parameters were set to 50 epochs and a learning rate of 1e-4. We utilized the ModelCheckpoint callback with the save_best_only option enabled to save the best-performing model during training. The Intersection over Union (IoU) metric was used to evaluate the performance of the models.

IoU, also known as the Jaccard Index, is a metric used to evaluate the accuracy of an image segmentation model. It measures the overlap between the predicted segmentation and the ground truth segmentation. The IoU is calculated as the area of overlap divided by the area of union between the predicted and ground truth segmentations. An IoU score ranges from 0 to 1, where 1 indicates perfect overlap and 0 indicates no overlap.

3. Result

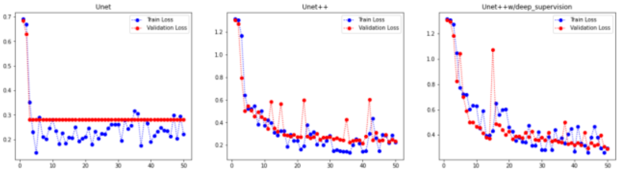

Figure 2: Loss along with epochs for models

When observing the loss curves for the models [Figure 2], it was found that U-Net showed signs of saturation at epoch 3. In contrast, both U-Net++ and U-Net++ with deep supervision demonstrated a stable convergence throughout the training process.

The IoU scores for the models [Figure 3] were as follows: U-Net achieved an IoU of 0.42, U-Net++ had an IoU of 0.90, and U-Net++ with deep supervision reached an IoU of 0.89. While U-Net++ showed significantly better results compared to U-Net, incorporating deep supervision into U-Net++ did not lead to any further improvement.

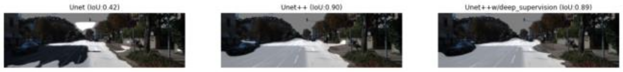

Figure 3: Semantic segmentation for road

4. Conclusion

Using U-Net++ has shown noticeable improvements in semantic segmentation and allowed us to observe the benefits of employing a more complex model. However, there are some areas for improvement in this analysis. Firstly, U-Net exhibited saturation at epoch 3, suggesting that adjusting the learning rate to a larger value might yield better results. Additionally, to obtain more objective results, it would be beneficial to calculate the average IoU values by repeating experiments across various images for all three models.

References

[1] Asked for answers to ChatGPT back and forth
[2] Etc, if any