Human Body Segmentation: A Comparison of DeepLabV3, DeepLabV3+, and U-Net

Introduction

This project compares the performance of three semantic segmentation models: **DeepLabV3**, **DeepLabV3+**, and **U-Net**, in the task of **human body segmentation**. The models are trained on a dataset of human body images, and their performance is evaluated based on segmentation accuracy metrics. The project explores different loss functions, including the **Dice Loss** and a combination of **Dice Loss** and **Cross Entropy Loss**.

Model Architectures

DeepLabV3

DeepLabV3 uses **atrous (dilated) convolutions** to capture context at multiple scales without reducing the spatial resolution. The key component is the **Atrous Spatial Pyramid Pooling (ASPP)** module, which applies convolutions with different dilation rates r. The output of the ASPP is given by:

$$y_i = \sum_k w_k \cdot x_{i+r \cdot k}$$

where:

- x_i is the input feature map.
- w_k are the convolutional weights.
- r is the dilation rate.

The use of different dilation rates allows the model to capture features at various receptive field sizes.

DeepLabV3+

DeepLabV3+ extends DeepLabV3 by adding a **decoder** module to refine segmentation results. The decoder upsamples the coarse segmentation map and

fuses it with low-level features from the encoder. Mathematically, the decoder can be represented as:

$$\hat{y} = \text{Upsample}(y_{\text{coarse}}) + f_{\text{low-level}}$$

where:

- y_{coarse} is the coarse output from DeepLabV3.
- $f_{\text{low-level}}$ are the low-level features from early layers.

This fusion helps improve the spatial accuracy of the segmentation map.

U-Net

U-Net has a **U-shaped architecture** with an encoder-decoder structure and **skip connections**. The encoder applies successive convolutions and pooling to reduce spatial dimensions, while the decoder upsamples the feature maps. The skip connections combine encoder and decoder features:

$$\hat{y}_i = f_{\text{decoder}}(\hat{y}_{i+1}) + f_{\text{encoder}}(x_i)$$

where:

- f_{decoder} is the upsampled decoder feature map.
- f_{encoder} is the corresponding encoder feature map.

These skip connections help preserve fine-grained spatial details.

Loss Functions

Dice Loss

The **Dice Loss** is based on the **Dice Coefficient**, a measure of overlap between the predicted segmentation p_i and ground-truth mask g_i:

$$\text{Dice Coefficient} = \frac{2\sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2}$$

The **Dice Loss** L_{Dice} is:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i p_i g_i + \epsilon}{\sum_i p_i^2 + \sum_i g_i^2 + \epsilon}$$

where ϵ is a small constant to avoid division by zero.

Cross Entropy Loss

The **Cross Entropy Loss** for a binary segmentation task is defined as:

$$L_{\text{CE}} = -\sum_i \left[g_i \log(p_i) + (1 - g_i) \log(1 - p_i) \right]$$

Combined Loss

To balance the advantages of both losses, a combined loss function is used:

$$L = \alpha L_{\text{Dice}} + \beta L_{\text{CE}}$$

where α and β are weighting factors (e.g., $\alpha = 0.5, \beta = 0.5$).

Model Training

The models were trained with the following parameters:

- **Optimizer**: Adam
- **Learning rate**: $\eta = 0.001$
- **Batch size**: 16
- **Epochs**: 50

The training objective was to minimize the combined loss function using **backpropagation**.

Evaluation Metrics

Model performance was evaluated using:

- **Dice Coefficient**:
$$\text{Dice} = \frac{2 \sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2}$$
- **Intersection over Union (IoU)**:
$$\text{IoU} = \frac{\sum_i p_i g_i}{\sum_i (p_i + g_i - p_i g_i)}$$
- **Pixel Accuracy**:
$$\text{Accuracy} = \frac{\sum_i \mathbb{1}(p_i = g_i)}{N}$$
where N is the total number of pixels and $\mathbb{1}$ is the indicator function.

Results and Insights

- **DeepLabV3** provided good segmentation results but struggled with boundary details due to the lack of a decoder.
- **DeepLabV3+** achieved better performance by refining the segmentation maps using its decoder module.
- **U-Net** excelled in preserving fine details, thanks to its skip connections.

Training with the combined **Dice Loss and Cross Entropy Loss** improved performance by balancing overlap maximization and classification accuracy.

Conclusion

This project demonstrated the strengths and weaknesses of DeepLabV3, DeepLabV3+, and U-Net for human body segmentation. The choice of loss function significantly impacted performance, with the combined Dice Loss and Cross Entropy Loss yielding the best results. These insights are valuable for selecting and optimizing models for real-world segmentation tasks.