CamVid Dataset Semantic Segmentation Project Report

Group Members:

Member Name A (email@example.com)Member Name B (email@example.com) Member Name C (email@example.com) Member Name D (email@example.com)

November 2, 2024

Contents

Abstract

This project focuses on the semantic segmentation of urban scenes using the CamVid dataset, which is widely used in computer vision and autonomous driving research. We experimented with two models: U-Net and FCN-ResNet50, ultimately selecting FCN-ResNet50 for its superior class recognition capabilities. The model was trained to classify each pixel into one of 32 semantic classes, including cars, pedestrians, roads, and buildings. Through iterative training and validation, our model achieved a high degree of accuracy in segmenting urban scene components, contributing to the development of robust scene understanding models for real-world applications.

1 Introduction

1.1 About the Dataset

The Cambridge-driving Labeled Video Database (CamVid) provides pixel-level annotations for each frame, with each pixel labeled according to one of 32 semantic classes. This dataset is widely used in semantic segmentation research, particularly for tasks that involve real-time segmentation in complex urban scenes. It includes labeled instances of various objects like vehicles, pedestrians, road signs, and buildings.

The dataset is divided into three sets:

- 367 training image pairs,
- 101 validation image pairs, and
- 233 test image pairs.

These splits are standard across research papers in semantic segmentation (e.g., Brostow et al., 2008; Sturgess et al., 2009; Badrinarayanan et al., 2017) to ensure comparability of results.

1.2 Objective and Goals

The goal of this project is to develop a segmentation model that accurately assigns a semantic label to each pixel in a scene. This task is crucial for applications like autonomous driving, where real-time scene understanding is essential. Specifically, we aimed to improve segmentation accuracy by effectively identifying each urban component (e.g., roads, pedestrians, buildings).

2 Dataset Details

The CamVid dataset was obtained from The University of Cambridge. It comprises high-resolution images with associated pixel-wise annotations that represent urban scene objects across 32 semantic classes. These labels are encoded as RGB values, where each color corresponds to a specific class.

Each frame in the dataset is annotated to highlight different classes within an urban environment, providing a dense representation of a cityscape. The class labels and RGB mappings were used to construct clear, color-coded visualizations of both ground truth and model predictions.

3 Methodology

3.1 Data Preparation

The CamVid dataset provides images and pixel-wise ground truth annotations for each frame, which allows for detailed scene segmentation. To utilize this data effectively, we converted each RGB mask into a single-channel class ID mask. This step enabled efficient training by simplifying multi-class segmentation tasks.

Images were transformed into tensors, normalized, and resized for the model. Each pixel in the mask was mapped to a class ID, facilitating visualization and evaluation during model training and testing.

3.2 Model Selection: Transition from U-Net to FCN-ResNet50

Initial Model: U-Net

We initially selected U-Net, known for its effective encoder-decoder structure with skip connections, as our base model. While U-Net performed well in boundary detection and segmenting objects, it was limited in distinguishing complex classes accurately. This limitation is partly due to U-Net's architecture, which is less effective at precise class identification without pre-training.

Final Model: FCN-ResNet50
To overcome U-Net's limitations, we transitioned to FCN-ResNet50, which leverages a pre-trained ResNet50 backbone. The ResNet50 architecture enhances feature extraction, allowing the model to better recognize diverse object classes. The model's deep convolutional layers capture intricate details in the image, significantly improving pixel-wise class recognition accuracy.

3.3 Training Process

Training Details: The FCN-ResNet50 model was trained over 25 epochs with a batch size of 4, using the Adam optimizer at an initial learning rate of 0.0001. A learning rate scheduler (StepLR) with a step size of 5 and decay factor of 0.5 was implemented to gradually reduce the learning rate, promoting model convergence as training progressed. Mixed-precision training was used to enhance computational efficiency without sacrificing model accuracy.

Loss Function: Cross-entropy loss was selected as the primary loss function, as it is well-suited for multi-class classification tasks, including pixel-wise segmentation where each pixel must be assigned a class label. Cross-entropy loss measures the divergence between the predicted probability distribution and the ground truth, penalizing incorrect class predictions proportionally to their likelihood. This choice of loss function ensures that the model learns precise class boundaries for accurate segmentation.

Each epoch was followed by validation on a separate validation set to monitor overfitting and make necessary adjustments to hyperparameters. We logged training and validation loss, accuracy, and intersection-over-union (IoU) scores for each epoch, providing insight into model performance.

3.4 Evaluation and Metrics

To comprehensively evaluate our model's segmentation performance, we used the following key metrics:

- **Pixel Accuracy**: Pixel accuracy measures the percentage of pixels in the image that the model correctly classifies. It is calculated as the ratio of correctly predicted pixels to the total number of pixels in the image. While pixel accuracy is a straightforward metric, it can be less informative for imbalanced datasets where some classes dominate.

- **Intersection over Union (IoU)**: IoU, also known as the Jaccard Index, evaluates the model's ability to correctly identify the boundaries of each class. IoU is calculated as the area of overlap between the predicted and ground truth masks divided by the area of their union:

$$\text{IoU} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{1}$$

We compute IoU for each class and average the scores to obtain a mean IoU value. Higher IoU values indicate better segmentation, as they capture both correct predictions and boundary accuracy, especially in complex scenes with multiple classes.

- **Confusion Matrix**: The confusion matrix provides a detailed breakdown of the model's predictions across all classes. Each row in the confusion matrix represents the true class, while each column represents the predicted class. Diagonal elements indicate correctly classified instances, while off-diagonal elements reveal misclassifications between specific classes. Analyzing the confusion matrix helps identify common misclassifications and areas for improvement.

These metrics, combined, give a comprehensive view of the model's performance. Pixel accuracy provides an overall view, IoU highlights boundary precision and segmentation quality, and the confusion matrix offers insights into specific class-based performance, allowing for targeted model refinements.

3.5 Visualization and Interpretation

For a comprehensive evaluation, we visualized:

- the original input image,
- the ground truth mask with class labels,
- the model's predicted mask, and
- the predicted mask with overlaid class labels.

These visualizations allowed us to qualitatively assess segmentation quality, identify segmentation errors, and ensure consistent performance across various object classes.

4 Results and Analysis

4.1 Training and Validation Loss over Epochs

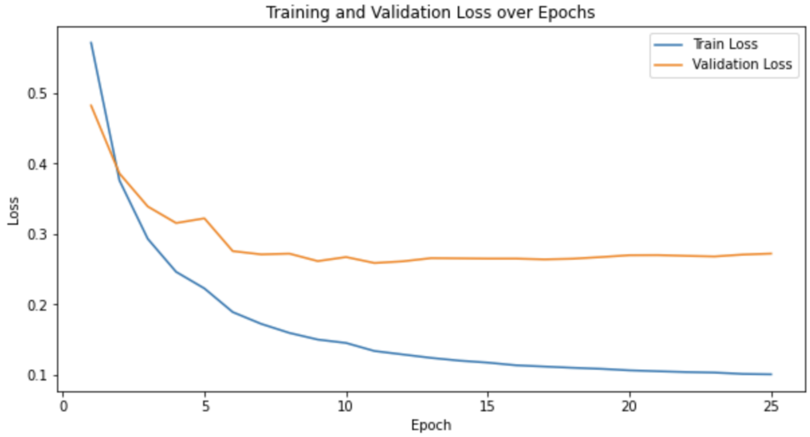

Figure 1: Training and Validation Loss over Epochs

Explanation: This line graph shows the model's loss for each epoch during the training and validation phases. The loss function here, typically Cross-Entropy Loss for segmentation tasks, is a measure of how well the model is learning to predict pixel classes accurately.

Key Observations:

- The training loss steadily decreases over time, which is a good indication that the model is successfully learning the features of each class in the dataset.
- The validation loss also decreases initially, which suggests that the model is generalizing well to data it hasn't seen before. After a certain point, the validation loss stabilizes or slightly increases, indicating that the model has likely reached its best performance on the validation set.

Significance: A significant gap between training and validation loss could indicate overfitting, but in this case, both curves converge reasonably well. This indicates a well-trained model without severe overfitting, with good potential for generalization on new data.

4.2 Training and Validation Accuracy over Epochs

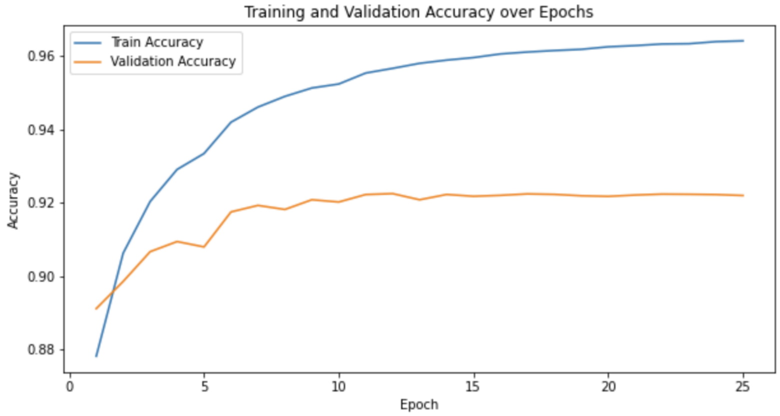

Figure 2: Training and Validation Accuracy over Epochs

Explanation: This graph illustrates the pixel accuracy of the model on the training and validation sets as it learns over each epoch. Pixel accuracy is the percentage of correctly classified pixels, a simpler but important metric in semantic segmentation.

Key Observations:

- The training accuracy steadily increases, showing that the model learns and improves its pixel-wise classifications with each epoch.
- The validation accuracy also improves initially and then plateaus, indicating that the model is learning well without memorizing the training data.

Significance: This plot helps validate the effectiveness of the training process. A small gap between training and validation accuracy indicates that the model's performance on unseen data (validation set) is close to its performance on training data, which is desirable.

4.3 IoU per Class (Test Set)

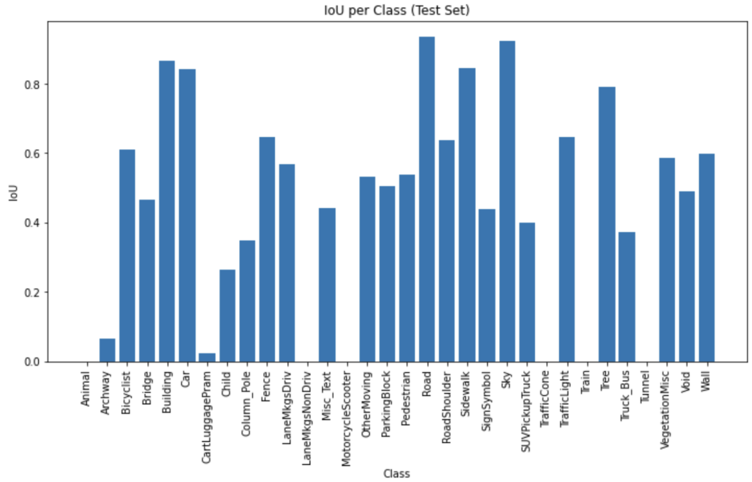

Figure 3: IoU per Class on the Test Set

Explanation: The Intersection over Union (IoU) metric is often used for evaluating segmentation models. IoU is calculated per class, representing the overlap between predicted and ground truth pixels for each class, divided by the union of the predicted and ground truth pixels.

Key Observations:

- Each bar represents the IoU score for a specific class, which tells us how accurately the model segments each category. High IoU values for certain classes mean that the model is accurately segmenting those categories, while low values suggest room for improvement.
- Classes like Building, Road, and Sky typically have higher IoU scores in segmentation tasks, as they often cover larger areas and have distinct boundaries.

Significance: This plot allows you to identify which classes the model segments well and which classes it struggles with. This can guide future improvements—such as data augmentation or model adjustments—to address specific classes where the model underperforms.

4.4 Confusion Matrix for Segmentation Accuracy

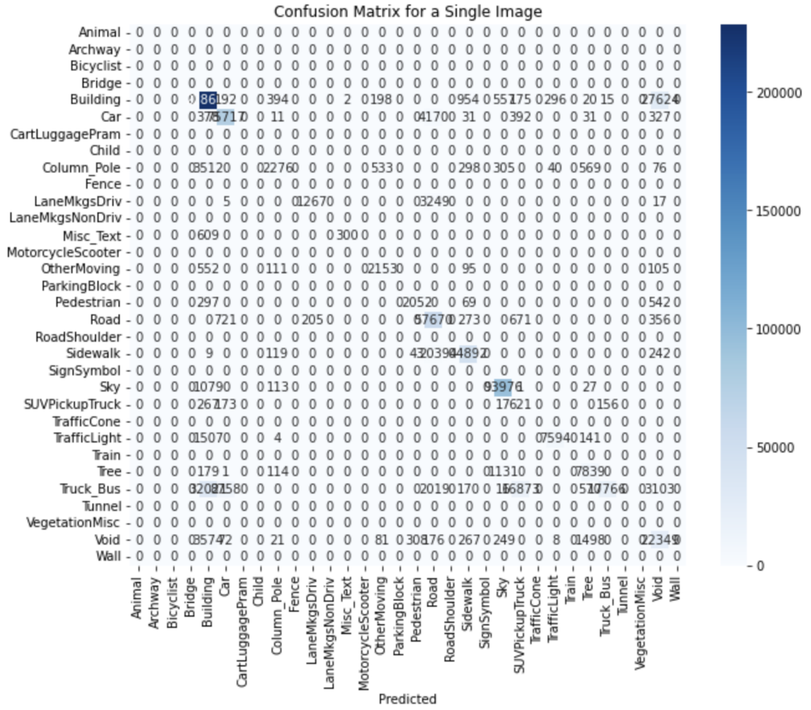

Figure 4: Confusion Matrix for Segmentation Accuracy on the Test Set

Explanation: The confusion matrix provides a comprehensive overview of the model's classification performance on a per-class basis by showing the true positive, false positive, and false negative rates for each class. Each cell in the matrix indicates the count of pixels where the model predicted a particular class versus the actual class in the ground truth.

Key Observations:

- Diagonal entries represent the pixels correctly classified by the model for each class. High values along the diagonal indicate strong predictive performance for those classes.
- Off-diagonal entries reveal common misclassifications between classes.

Significance: The confusion matrix is valuable for diagnosing specific weaknesses in the model. It provides insights into which classes are being confused, potentially due to similar visual features or low representation in the dataset. This information can guide targeted improvements, such as collecting more data for underrepresented classes or adjusting the model to better differentiate similar classes.

4.5 Visual Comparison of Predictions

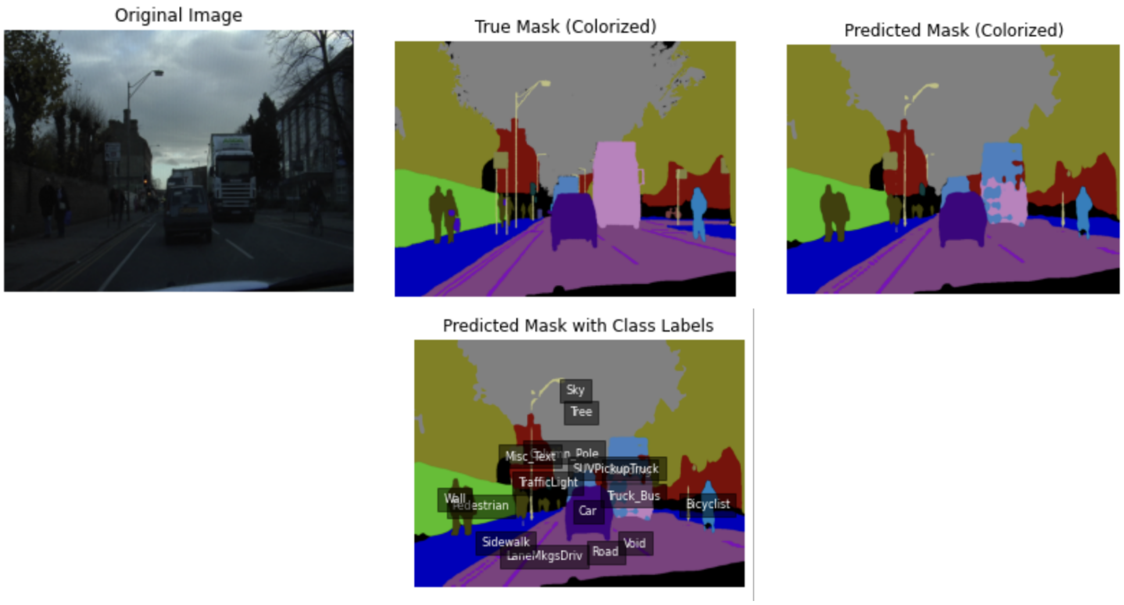

Figure 5: Visual Comparison of Predictions. (From left to right: Original Image, Ground Truth Mask, Predicted Mask, Predicted Mask with Class Labels)

Explanation: This visual comparison includes the original image, the ground truth mask (true labels), the predicted mask (predicted labels), and the predicted mask with class labels annotated. This comparison allows for a qualitative assessment of the model's segmentation abilities.

Key Observations:

- Observing the differences between the true and predicted masks provides insight into the model's strengths and limitations. For example, if the model consistently captures large objects like "Building" and "Road" but struggles with smaller objects like "Bicyclist," it highlights areas for improvement.
- The overlay of class labels in the final image also allows for easy verification of specific classes that were correctly or incorrectly identified.

Significance: This qualitative analysis provides a clear, visual representation of the model's segmentation quality. It helps verify whether the model captures complex scenes and distinct classes correctly, which is essential for real-world applications where pixel-perfect accuracy may not always be feasible.