On the Analysis of Training Curves and Optimal Weight Selection in U-Net

Sungchan Maeng
aodtjdcks@naver.com

March 24, 2025

Abstract

In general, deep learning models are trained and optimized by monitoring overfitting and selecting the model weights from an appropriate epoch. However, it is often unclear which epoch yields the most optimal weights in terms of generalization performance. This ambiguity arises from various factors such as irregular training curves, difficulty in identifying trends, and overfitting to the validation set.

To address these issues, we propose a method for better identifying optimal model weights. When the training and validation curves are highly unstable, we apply smoothing using moving averages to the loss and metric graphs. Based on the smoothed curves, we select four candidate epochs and evaluate their performance through both quantitative (Mean IoU) and qualitative (visual comparison) assessments on the test set.

Through these evaluations, we identify the most generalizable model among the selected checkpoints. The results demonstrate that certain weight selections, particularly those based on validation loss minima, can serve as a practical reference for model selection. Detailed results are discussed in the "Results" section.

1 Introduction

The advancement of deep learning technologies has led to significant performance improvements across various domains, particularly in image segmentation tasks where the U-Net model has been widely adopted. Originally proposed in 2015 for biomedical image segmentation, U-Net has become a standard architecture based on its encoder-decoder structure and skip connections. It is known for delivering high performance even with relatively small datasets, and numerous variants have since been developed.

However, one persistent challenge in training deep learning models—including U-Net—is the question of how to define and select the optimal model weights. Typically, a model is evaluated based on validation loss or a chosen metric, and weights are selected from the epoch just before overfitting begins. Yet in practice,

training curves for loss and metrics often exhibit high variance or instability, making it difficult to determine which checkpoint yields the best generalization performance on the test set.

This issue arises not only from training instability but also from various factors such as the composition of the validation set, ambiguity in evaluation criteria, and inherent model uncertainty. In particular, overfitting to the validation set can lead to performance degradation on unseen data, raising the fundamental question: "Does strong validation performance guarantee good generalization?"

Motivated by these concerns, this study explores a more reliable approach to analyzing training curves from a U-Net model and identifying the optimal weight selection point. To mitigate the irregularity of the training graphs, we apply a smoothing technique using moving averages. Based on the smoothed curves, we select specific epoch points and evaluate their performance through both quantitative and qualitative methods. This process aims to determine the model checkpoint that achieves the best generalization, ultimately providing a practical reference for weight selection in real-world applications.

2 Method

In this study, we perform binary semantic segmentation of road areas using a deep learning model based on U-Net. The main focus of the experiment is to analyze training curves after model training and identify the model checkpoint that achieves the best generalization performance. To this end, we describe the process in stages: dataset composition, preprocessing and augmentation, model architecture and training setup, and the analysis of training curves with selected evaluation points.

2.1 Dataset Composition and Preprocessing

We utilized the KITTI segmentation dataset for this study. Out of a total of 200 images, 168 were used as the training set, 16 as the validation set, and the remaining 16 as the test set. The segmentation task was formulated as binary classification, focusing solely on the road class labeled as 7.

All images were resized to 224×224 pixels, and data augmentation was applied only to the training set using the `albumentations` library. The augmentation techniques included:

- `HorizontalFlip(p=0.5)`: 50% probability of horizontal flipping
- `RandomSizedCrop(p=0.5)`: 50% probability of random cropping

No augmentation was applied to the validation or test sets.

2.2 Model Architecture and Training Setup

The model follows the U-Net architecture, where the input is in the format of `(batch_size, 224, 224, 3)` representing RGB images, and the output is

`(batch_size, 224, 224, 1)`, predicting a binary mask. The final output layer is a 1×1 convolution followed by a sigmoid activation function to predict per-pixel probabilities of road regions.

The training setup is as follows:

- **Optimizer**: Adam
- **Learning Rate**: 1e-4
- **Loss Function**: Binary Cross Entropy
- **Evaluation Metric**: Intersection over Union (IoU)
- **Batch Size**: 8
- **Epochs**: 400

2.3 Training Curve Analysis and Checkpoint Selection

The training loss and metric curves showed significant fluctuations and instability, making it difficult to determine an optimal checkpoint based solely on raw values. To enable a more stable interpretation of the curves, we applied the **Moving Average** technique. The analysis was conducted with the following configurations:

- **Short-Term Smoothing**: Moving average with window size $n = 10$, to detect local fluctuations
- **Long-Term Smoothing**: Moving average with window size $n = 30$, to capture overall trends

Based on this analysis, we selected a total of five candidate epochs:

1. **Epoch 80**: The point where train loss and validation loss begin to diverge, based on 10-point moving average (see Figure 2)
2. **Epoch 320**: The minimum value of raw validation loss (see Figure 3)
3. **Epoch 390**: The maximum value of raw validation IoU (see Figure 3)
4. **Epoch 350**: The lowest point of validation loss based on 30-point moving average (see Figure 4)
5. **Epoch 200**: A midpoint between Epoch 80 and 320, selected as a comparative baseline

These points were not chosen merely based on local minima or maxima but were strategically selected by considering the *patterns and trends* in the training curves. For each selected epoch, the corresponding model weights were evaluated on the test set using both quantitative and qualitative methods, aiming to identify the checkpoint with the best generalization performance.

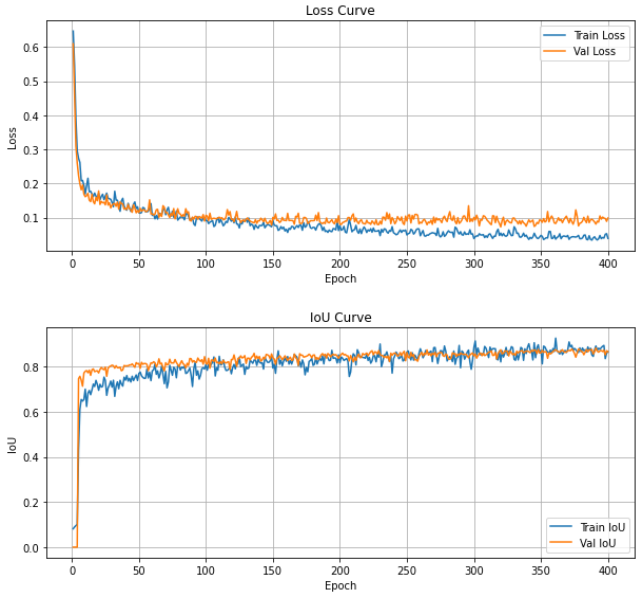

Figure 1: Training and validation loss (top) and IoU (bottom) over 400 epochs.

3 Result

In this section, we present the quantitative and qualitative evaluation results based on the five selected epoch checkpoints using the test set.

3.1 Quantitative Evaluation

The following table summarizes the **Mean IoU** values obtained on the test set for each model checkpoint.

Table 1: Mean IoU of Selected Epoch Checkpoints on Test Set

Epoch	**Description**	**Mean IoU**
80	Train/Val Loss divergence point	0.8001
200	Midpoint between Epoch 80 and 320	0.8459
320	Minimum raw Validation Loss	**0.8743**
350	Minimum Val Loss with Moving Average (n=30)	0.8613
390	Maximum raw Validation IoU	0.8656

Among the evaluated checkpoints, the **Epoch 320** model showed the highest Mean IoU (0.8743) on the test set, indicating the best generalization performance. In contrast, the model at Epoch 80 demonstrated the lowest performance, making it an unsuitable choice for optimal weights. While the models at Epochs 200, 350,

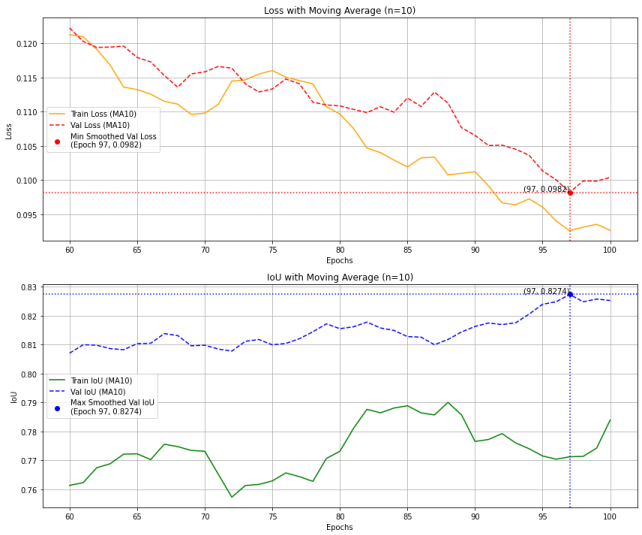

Figure 2: Training and validation curves smoothed with a moving average window of $n = 10$. Epoch 97 shows the minimum validation loss and maximum validation IoU in this setting.

and 390 also showed relatively high Mean IoU values, the differences among them were not significant, suggesting the need for **further qualitative analysis**.

Figure 5 shows the IoU values for individual test samples across the five selected model checkpoints.

Notably, for **Sample 7**, all models performed poorly; however, the Epoch 320 model still achieved the highest IoU among them. Overall, the Epoch 80 model consistently showed lower performance across all samples, whereas the other models—particularly the one at Epoch 320—exhibited strengths depending on the sample.

3.2 Qualitative Evaluation

The qualitative evaluation involved visually comparing the segmentation results predicted by each model at different epochs. The figure below displays a matrix in which each column represents a model (i.e., a specific epoch), and each row corresponds to a selected test sample (see Figure 6).

Focusing on **Samples 1, 4, 8, and 13**, the model at Epoch 320 clearly outperformed others in accurately identifying road areas. The predictions were more precise in terms of **boundary delineation**, **suppression of non-road regions**, and **removal of noise**, indicating that the Epoch 320 model provided the most stable and generalizable results.

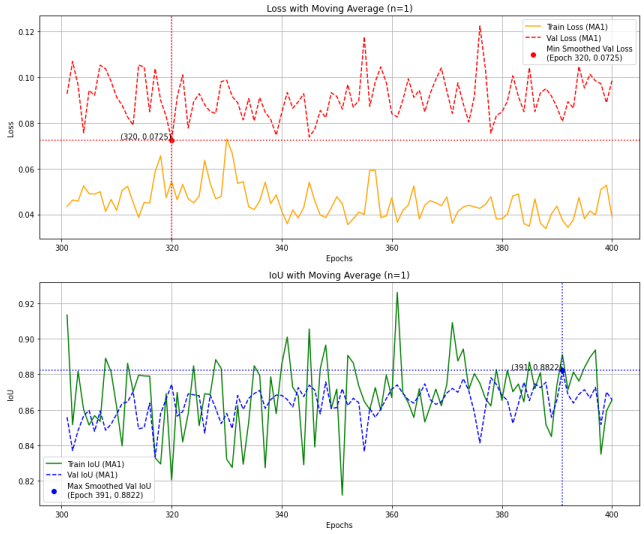

Figure 3: Raw curves without smoothing ($n = 1$). Epoch 320 and 391 were selected as the points with minimum validation loss and maximum IoU, respectively.

4 Discussion and Conclusion

4.1 Discussion

In this study, we conducted binary segmentation experiments using the U-Net model, focusing on addressing the instability of training curves by applying a smoothing technique and exploring various perspectives for selecting optimal model weights. Through both quantitative (Mean IoU) and qualitative (visual comparison) evaluations, we found that the model checkpoint at **Epoch 320**, corresponding to the **minimum validation loss**, yielded the best generalization performance on the test set.

This result suggests that using the *minimum validation loss* as a criterion may lead to better performance than the more commonly used approaches such as selecting the *highest validation metric* or the *last epoch before overfitting*. The application of a smoothing method to reveal the overall trend of the training process allowed for a clearer interpretation of the curves, and the strategy of selecting five candidate epochs based on this analysis provides a **more objective and systematic approach** to weight selection.

However, it is important to note that this conclusion was drawn under **specific conditions**—a particular model (U-Net), dataset (KITTI), and binary classification setting. The same trend may not necessarily hold in different scenarios such as multi-class segmentation, alternative backbone architectures, or larger and more diverse datasets.

Nonetheless, this experiment highlights the importance of a **multi-faceted**

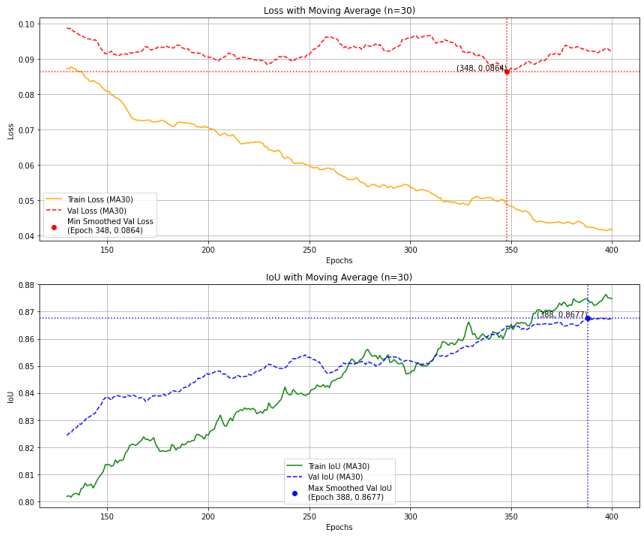

Figure 4: Curves smoothed with a larger moving average window of $n = 30$. Epoch 348 and 388 show the extreme values for validation loss and IoU in this configuration.

evaluation approach that combines both quantitative metrics and qualitative insights, rather than relying solely on validation performance, when selecting optimal model weights in deep learning.

4.2 Conclusion

This study investigated the challenge of **selecting the optimal model weights** during the training process of a U-Net segmentation model. To address the instability of training and validation curves, we applied a moving average smoothing technique and selected five candidate epochs based on observable trends in the graphs. Each checkpoint was evaluated using both quantitative and qualitative methods. The results demonstrated that the model checkpoint at **Epoch 320**, corresponding to the lowest validation loss, achieved the best performance on the test set.

These findings can serve as a **useful reference** for future segmentation experiments or practical applications involving U-Net models. Furthermore, this study lays the groundwork for **future research** aimed at identifying **generalizable trends in optimal weight selection** across various architectures and datasets.

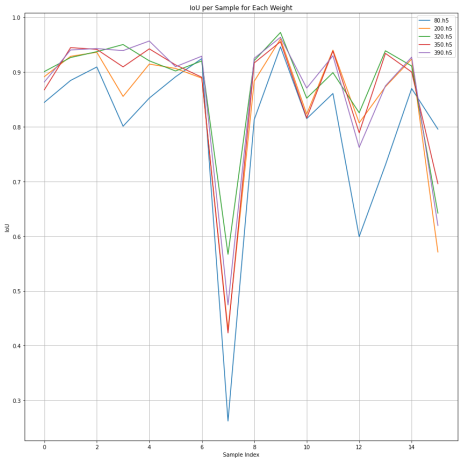

Figure 5: IoU scores for each test sample across five different model checkpoints (Epochs 80, 200, 320, 350, and 390). Notably, the model at Epoch 320 shows consistently strong performance across most samples, while Epoch 80 exhibits more fluctuation.

5 References

References

[1] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. Springer. DOI: 10.1007/978-3-319-24574-4_28

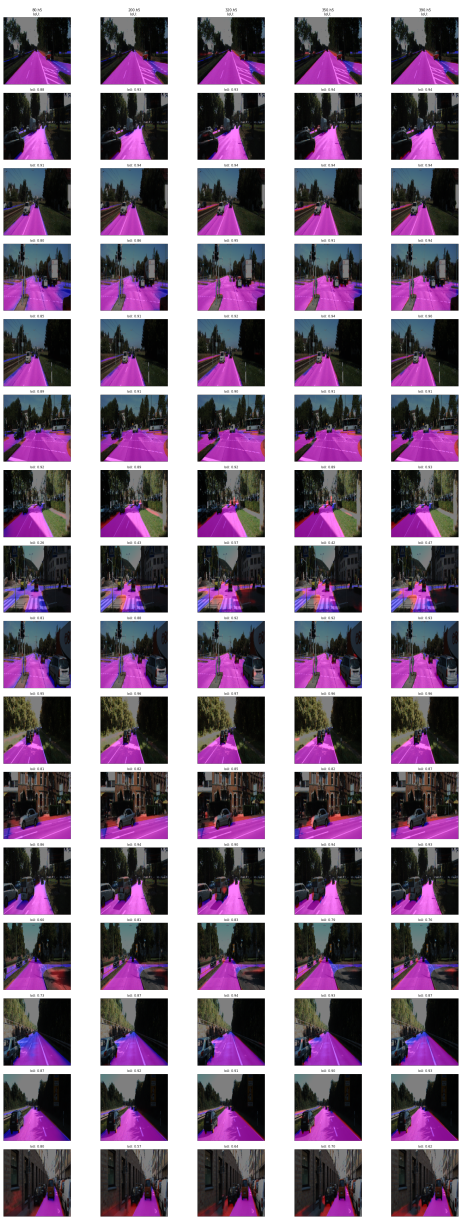

Figure 6: Qualitative comparison of segmentation results across selected model checkpoints. Each row represents a test sample, and each column corresponds to a model trained at a different epoch (80, 200, 320, 350, and 390). The model at Epoch 320 shows clearer boundary detection and less noise in multiple cases.