Real-Time Street View Segmentation Using Transfer Learning SegFormer

Alvaro Avalos-Morales, Jason Henry

Department of Computer Science, University of Nevada, Las Vegas

ABSTRACT

Semantic segmentation plays a critical role for self-driving cars to understand their environment. In this work, we investigate the effectiveness of transfer learning on SegFormer models by fine-tuning pretrained Cityscapes architectures using the higher-resolution, 65-class Mapillary Vistas dataset. To explore whether attention structure impacts performance, we introduce hybrid transformer variants that replace global attention with local window attention and, in some cases, expand encoder depth. Five model configurations, including modified B0 and B2 variants, were trained and evaluated using Dice, IoU, and loss metrics. Results show that these architectural changes do not yield significant improvements over baseline SegFormer models, although the larger B2 consistently outperforms B0. Our findings highlight the robustness of the original SegFormer design for real-time street-view segmentation.

1. INTRODUCTION

Understanding the semantics of the space around a vehicle is critical to self-driving, especially for those that rely on image input. Self-driving vehicle models can leverage abundant image data from sources such as cameras. To extract meaningful relationships between objects in an image, semantic segmentation can be used. Using a transformer-based segmentation model makes segmentation more accurate by leveraging attention mechanisms [4], which allow pixel classifications to be informed by other pixels.

SegFormer [1] is a family of image segmentation models introduced by Nvidia in 2021, ranging from 3.9 million to 85 million parameters. The SegFormer models we experimented with were trained on the Cityscapes [2] dataset, but Cityscapes may be limited in its practical usefulness for self-driving cars and contains only 30 classifications. By training SegFormer using transfer learning, accuracy can be increased, and convergence can be achieved faster. This would allow SegFormer to produce functional segmentations for self-driving applications.

The dataset chosen for transfer learning with SegFormer was the Mapillary Vistas [3] dataset due to its higher class count (65) and better road scenes. Attention blocks within the models' transformers were also modified to be global-local hybrids, so the model

could better understand the relationship between the road and the world.

2. METHODS

2.1 Mapillary Vistas Dataset

The Mapillary Vistas Dataset is a large street-level image dataset containing 20,000 high-resolution images with 65 annotated semantic classes, covering roads, buildings, vehicles, pedestrians, vegetation, and other urban scene elements. Designed for semantic segmentation, it provides high-quality polygon annotations collected from diverse geographic locations, making it a strong benchmark for real-world semantic segmentation tasks [3]. The dataset and official documentation are available at: https://www.mapillary.com/dataset/vistas.

2.2 Dataset Pre-processing

Because the Mapillary Vistas Dataset is extremely large and contains high-resolution images, it is impractical to train on the raw data due to our hardware limitations. To make training feasible, all photos are preprocessed. First, the shorter side of each image is scaled to 512 pixels while preserving the aspect ratio. The longer side scales proportionally. Then, a center crop of 512×512 is applied to produce fixed-size inputs suitable for training. This preprocessing significantly reduces disk space, speeds up data loading, and eliminates the need for on-the-fly resizing during training. The full preprocessed dataset used in this work can be accessed here: https://drive.google.com/file/d/11ZPJbu9ZVcWOaSFcyV6tUnwuW107nrM3/view?usp=sharing

2.3. Transfer Learning

For this project, we used SegFormer models pretrained on the Cityscapes [2] dataset as the starting point. Cityscapes provides urban street-scene imagery and serves as a strong initialization for segmentation tasks involving roads, vehicles, and pedestrians. We selected two architectures from the SegFormer family: SegFormer B0 (smallest model with 3.9 million parameters, designed for real-time performance) and SegFormer B2 (a mid-sized 24-million-parameter model offering improved accuracy). To adapt these pretrained models to the larger class label of the Mapillary Vistas dataset, we replaced the original decoder head with a newly initialized classification layer matching the target number of classes. The encoder weights were preserved, and the models were then fine-tuned end-to-end on our preprocessed 512×512 dataset.

2.4. Transformer Modification

Transformers in the SegFormer architecture provide attention [4], which is the ability for the model to understand the context of a token, in our case, classifying pixels in relation to an image. SegFormer uses global [1] attention by default, which is appropriate for this use case, but we wanted to experiment to see if accuracy or other metrics could improve if we replace half of the transformer's attention blocks with local [5] attention.

2.5. OpenCV Demo

After training, each SegFormer model and its fine-tuned weights are saved for later use. The included OpenCV demo script loads the chosen SegFormer architecture and its trained weights, and performs real-time semantic segmentation using a standard webcam. Incoming frames are preprocessed, fed to the model, converted into colorized masks, and displayed alongside the original and overlaid views, enabling a live demonstration of the model's segmentation performance.

3. EXPERIMENTS

We conducted experiments on five different variants of the SegFormer architecture. These models used in this study are:

1. **SegFormer B0**: The baseline, lightweight, pre-trained Cityscapes SegFormer-B0 architecture (3.9M parameters), fine-tuned on the Mapillary Vistas dataset.

2. **SegFormer B0-1**: Modified B0 version where one global attention block in the encoder is replaced with local window attention.

3. **SegFormer B0-2**: Modified B0 version with expanded encoder depth (additional layers in stages 2 and 4) and full replacement of global attention with local window attention.

4. **SegFormer B2**: The baseline, pretrained, Cityscapes SegFormer-B2 architecture (24M parameters), fine-tuned on the Mapillary Vistas dataset.

5. **SegFormer B2-1**: Modified B2 version, mirroring the B0-2 modifications by adding extra encoder layers and replacing attention mechanisms with local window attention.

All models were trained on the preprocessed 512×512 Mapillary Vistas dataset for 20 epochs using a learning rate of 1×10^{-4}.

4. RESULTS

The results across the five model variants show that our architectural modifications did not yield meaningful performance gains. When examining the Dice and IoU charts, all modified models fluctuate around the baseline SegFormer performance, with no variant consistently outperforming the original architectures. While some modifications produced minor peaks in Dice or IoU, these changes were not stable across epochs, indicating that the added layers and local attention mechanisms did not provide a significant benefit over the pretrained baseline.

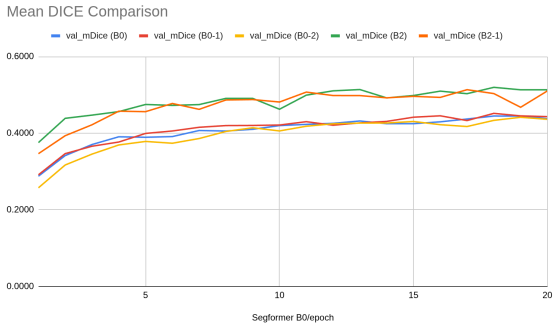

Figure 1

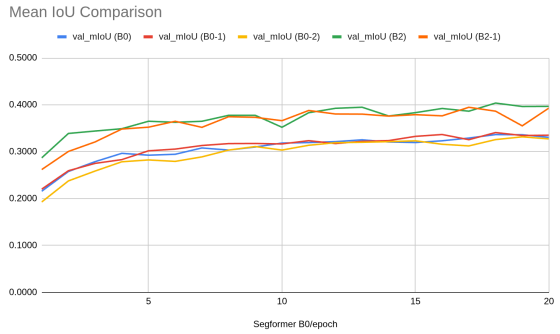

Figure 2

A similar pattern appears in the training and validation loss curves. The training loss decreases steadily across all models, while the validation loss remains flat, showing little to no downward trend.

Although this might initially suggest overfitting, the situation is more nuanced. Alongside the stagnant validation loss, the Dice scores continue to rise, demonstrating improved mask quality despite the unchanged loss value.

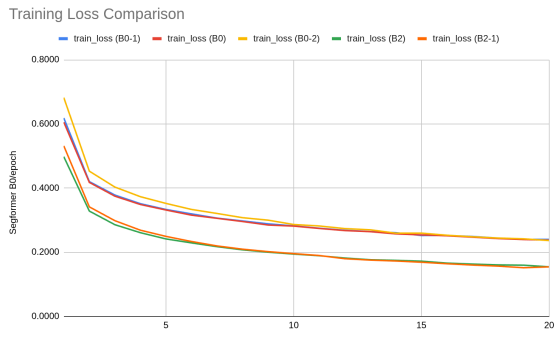

Figure 3

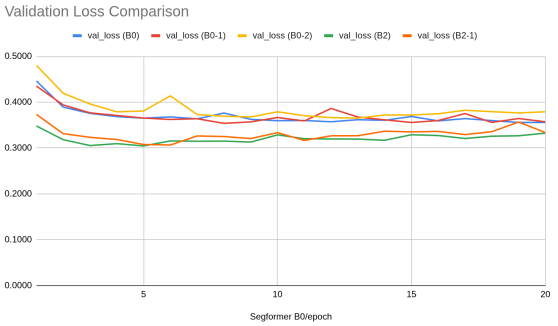

Figure 4

5. CONCLUSION

Although our modifications did not yield significant improvements, the results consistently show that the SegFormer B2 model outperforms the smaller B0 model across all metrics.

6. ACKNOWLEDGEMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

There exists no real or potential conflict of interest for any of the authors listed.

7. REFERENCES

[1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo,

"SegFormer: Simple and efficient design for semantic segmentation with transformers,"

Advances in Neural Information Processing Systems, vol. 34, 2021.

[2] M. Cordts et al., “The Cityscapes dataset for semantic urban scene understanding,”

in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),

2016, pp. 3213–3223.

[3] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder,

“The Mapillary Vistas dataset for semantic understanding of street scenes,”

in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 4990–4999.

[4] A. Vaswani et al., “Attention is all you need,”

Advances in Neural Information Processing Systems, vol. 30, 2017.

[5] N. Parmar et al., “Image transformer,”

in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 4055–4064.

[6] S. J. Pan and Q. Yang, “A survey on transfer learning,”

IEEE Transactions on Knowledge and Data Engineering,

vol. 22, no. 10, pp. 1345–1359, 2010.