DeepLabV3+ and SegFormer robustness analysis

Muhammad Huzaifa [1] Anees Hashmi [2] Mohamed Boudjoghra [3]

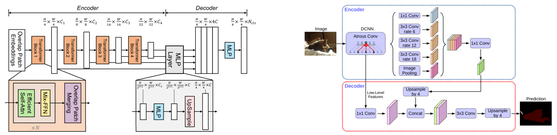
[1]Mohamed Bin Zayed University of Artificial Intelligence

Abstract

In this project, we conducted a comparative analysis of Segformer, a ViT-based semantic segmentation model, and DeeplabV3+, a CNN-based semantic segmentation model, using various image perturbations (e.g., image patch shuffling and removal) and noise addition (e.g., salt and pepper noise). Previous studies have compared these models' performance on segmentation of natural images without any noise, however, the performance of these models on perturbed and noisy images was unknown. Segformer was trained on ImageNet dataset which does not have any *person* class and fine-tuned on ADE20k, whereas, DeeplabV3+ was trained on ADE20k dataset directly. We considered the *person* class segmentation only to avoid any effect of ImageNet pre-training of Segformer. Results suggest that Segformer can tolerate more perturbation than DeeplabV3+ and can perform well on noisy images, however, both models' performance drop significantly as we increase the noise.

Introduction

Segmentation is a pixel-level class prediction instead of image-level class prediction as in classification. In this study, two state of the art (CNN and Transformer based) techniques of segmentation were evaluated. DeeplabV3+ is an improved version of DeeplabV3 which is constructed following a CNN-based encoder-decoder architecture, with a light weight decoder structure, and an encoder that uses atrous convolution to extract dense feature maps.
SegFormer, contains a positional-encoding-free hierarchical transformer encoder and a lightweight All-MLP decoder. The hierarchical structure allows generating a CNN-like multi-level features, producing high-resolution coarse features and low-resolution fin-grained features that improve the performance of semantic segmentation. Since the recent success of transformers in different vision tasks, an in depth comparison on CNN and transformer based segmentation model on perturbed and original data set is needed.



(a) Segformer (b) DeepLabV3+

Figure 1. Architecture Block Diagram

Objectives

1. To understand and evaluate (the performance) of Segformer and DeeplabV3+
2. To perform an unbiased comparison of both models by alleviating the effect of pre-training of ViT on ImageNet
3. To compare the robustness of both models against noise addition and image perturbation

Methodology

We started by generating different perturbed dataset from the person dataset to generate other perturbed ones to observe the performance of the two models.

- **Contextual information disturbance:** we added a weighted image with different context to the original image. The generated ones gradually loose some textures which define a person (e.g. color), when increasing the weight (denoted α). The motivation is to see if the models can segment a shape similar to a human.
- **Random patch shuffling:** We generated five different datasets by shuffling patches of the original set, with different grid sizes from 2 to 32. This test is made to see if the Mix-FNN of the SegFormer can recover the positional information even after the shuffling

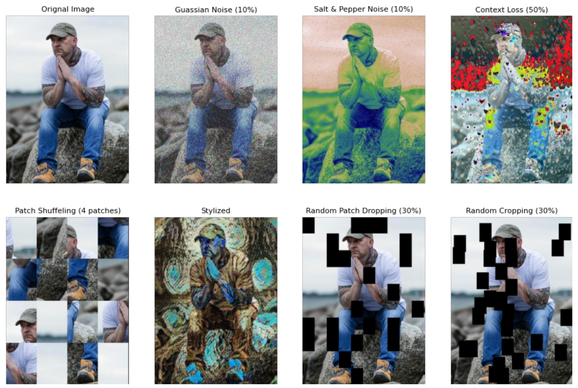


Figure 2. The generated images

- **Stylization:** The stylized images are generated by feeding a Picasso drawing and the original dataset to an arbitrary image stylization model available on tensorflow_hub. After testing with dataset, we can conclude weather attention or convolution is better in understanding texture-less images.
- **Random cropping** Randomly cropping 30%, 50% 70% of original image irrespective of the patch borders. This helps to assess the effect of fake edges on both models.
- **Random patch dropping** Randomly dropping 30%, 50% 70% of original image patches.
- **Gaussian Noise:** In this part, we generated three noisy datasets after adding gaussian noise with different intensity of 1.0, 2.5, and 5.
- **Salt and pepper noise:** We applied salt and pepper noise with four different occurrence probabilities, 0.001, 0.01, 0.1, and 0.3 to generate four noisy datasets.

Results

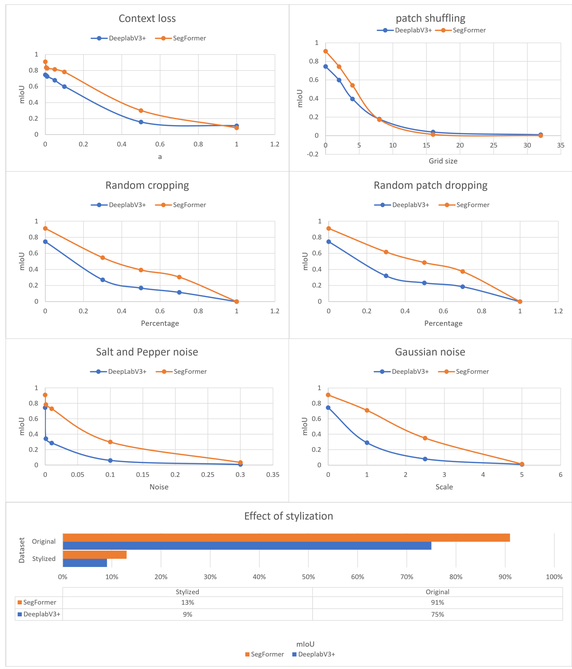


Figure 3. Test results

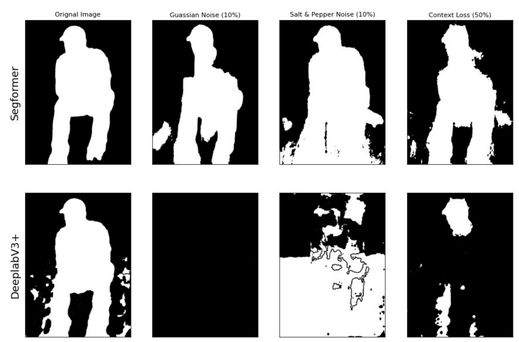


Figure 4. Segformer predicted masks (top) and DeeplabV3+ predicted masks (bottom)

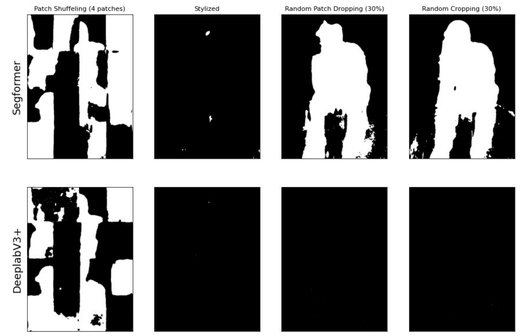


Figure 5. Segformer predicted masks (top) and DeeplabV3+ predicted masks (bottom)

Interesting observations

- Segformer outperforms DeeplabV3+ on all the available types of images (original and noisy images).
- Segformer is more robust to Gaussian noise addition particularly, but it too fails as we increase the noise beyond a certain level; hence, both models are unable to segment noisy images properly.
- Segformer can tolerate a significant amount of Gaussian noise, in particular.
- Unlike other experiments, Segformer fails as image patches are shuffled, which suggests its vulnerability to shuffled images in particular.

Conclusion

Segformer provides better segmentation results compared to DeepLabV3+ because of the attention mechanism that looks at the global context of the image. Its robustness indicates its practicability in applications where noise is present in the data, and should not influence the prediction (e.g. medical imaging, autonomous driving). Analyzing the robustness of ViT and CNN based segmentation models on different weather conditions would be an interesting future research topic to explore for more in-depth analysis of these architectures.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[2] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[3] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.