Documentation

Vision Transformers for Image Segmentation

# Introduction

Our chosen topic was Vision transformers for Image Segmentation. Vision Transformers are specifically designed for computer vision tasks and uses the transformer architecture originally designed for natural language processing. It has an advantage against convolutional networks which were mostly used in the past for image segmentation, that it has global context. While CNNs can only focus on local contexts, Vision Transformers use self-attention for global context of the dataset.

# Dataset

For our dataset, we selected a road layout segmentation dataset from Kaggle. These are top-down-view pictures of maps with their masks to train the neural network.

A képen épület, Városépítészet, ház, Lakóövezet látható

Automatikusan generált leírás A képen szimbólum, Szimmetria, fekete, fehér látható

Automatikusan generált leírás

The dataset contains 6226 images for training, with additional for validation and testing.

# Methods of training

For the baseline models, we used convolutional neural networks. The first one is a U-Net architecture. This architecture uses an encoder and a decoder and extracts the information through a bottleneck. It also uses skip-connections.

The other one is a DeeplabV3 pretrained model from pytorch. DeeplabV3 uses dilated convolution, which can increase the receptive field without increasing the number of parameters.

For the Vision Transformer model, we used a Swin-Unet model, which combines the advantages of both the Swin and U-Net. The Swin Transformer architecture offers hierarchical feature representation and a window attention mechanism while the U-Net architecture gives the encoder-decoder structure and skip-connections, which means passing input into a deeper layer, skipping intermediate layers.

We followed the model trainings on WandB. Optimized the hyperparameters manually on low epoch, and then

# Evaluation

To evaluate the models, we used multiple different key metrics.

Since in most images, the pixels classified as not road far exceed the pixels classified as road, accuracy is not too effective to describe the difference between model’s performance. A better variable is Intersection over Union, which means the intersections of the prediction and the reality divided by the union. Another useful evaluation criteria which we used is DICE coefficient, which measures the overlap between the predicted value and the truth.

The task also specified to compare the models on inference time. The results indicate that the U-Net model is much faster than the DeepLabV3, evaluating in less than a second compared to approximately 10 seconds.

# Visualization

We made a few examples of each model with the original image, the mask and the model’s prediction for the image. Additionally, we created a frontend UI using Gradio.

On the UI, you can choose a random image from the dataset, and model to evaluate. It shows the models prediction to the image, the intersection and

# Conclusion

# References

Road layout image segmentation: <https://www.kaggle.com/datasets/balraj98/deepglobe-road-extraction-dataset/data>

WandB:

DeeplabV3: <https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/>

Swin-Unet: <https://arxiv.org/abs/2105.05537>

Swin-Unet model: <https://github.com/HuCaoFighting/Swin-Unet/tree/main>

Gradio: <https://www.gradio.app/>