Deep Learning Course at IGP TUBS

Segmentation of Vaihingen dataset including DSM data

Yu-Chuan, Cheng
yu-chuan.cheng@tu-braunschweig.de
Fabian, Linkerhägner
f.linkerhaegner@tu-braunschweig.de

July 2022

Abstract

Use this chance to compare and contrast DeepLabv3 and HRNet in terms of segmentation and semantics as well as to comprehend the distinction between focal loss and other loss functions.

Keywords: Vaihingen dataset, DSM, DeepLab3+,HRNet

1 Introduction

In order to determine whether HRNet is commonly used in semantic segmentation and whether human posture estimation can still produce high accuracy on satellite images, I want to compare it to other models I've learned in the course. This is because I found that HRNet is the most accurate model during the process of in-depth study and segmentation. To examine which experimental approach is superior, we wish to employ DeepLab3+ with Focal Loss and HRNet [2], [3]with Standard SGD and Cross Entropy.

2 Motivation

In training CNN for semantic segmentation, there is often just 2D input, but this issue contains a third dimension from DSM, and we want to learn more about Focal Loss, therefore this topic will make a suitable learning object.

3 Materials and Methods

The data collection consists of 33 patches (of various sizes), each of which is made up of a DSM and a true orthophoto (TOP) that was taken from a larger

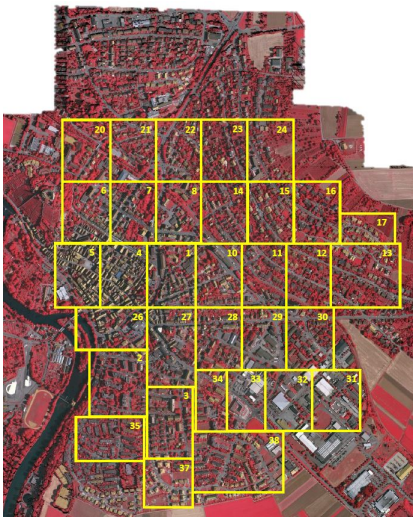

Figure 1: Vaihingen. Numbers refer to the individual patch numbers as written at the filename endings. source: `https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx`

TOP mosaic (see Figure 1). The TOP and the DSM both use a 9 cm ground sampling distance.

4 Evaluation

We anticipate that HRNet with a cross entropy loss function and an SGD optimizer will achieve comparable results to DeepLabv3+ with a focused loss function and an Adam optimizer. On the cityscapes dataset, HRNet approaches 81.6 percent mIOU, but DeepLabv3 achieves 81.3 percent mIOU [1] [2]. In addition, we anticipate that the assessment will be superior than the one we conducted without DSM data. mIOU is the primary measure we use.

$$IOU = \frac{1}{S} \sum_{s \in S} \frac{TP_s}{TP_s + FP_s + FN_s}$$

Furthermore, we will consider training time and the rate at which their trainable parameters will converge.

References

[1] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Rethinking atrous convolution for semantic image segmentation liang-chieh. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 2018.

[2] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-Resolution Representations for Labeling Pixels and Regions. 2019.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3349–3364, 2021.