CS 766 Project Proposal
TNTU-Net: A Semantic Segmentation Model for Autonomous Driving

Jinwen Sun: jsun279@wisc.edu
Wei Han: whan59@wisc.edu

Overview

In this project, a novel semantic segmentation model based on Transformer-iN-Transformer (TNT) and U-Net is proposed to realize better scene understanding for autonomous driving. This model leverages both two structures to create a more precise localization and a better understanding of global information. Experiments on two benchmark datasets will be conducted to demonstrate the effectiveness of the proposed TNTU-Net architecture by comparing it with several pioneering algorithms.

Background

Semantic segmentation is becoming increasingly popular and important in the field of computer vision. It refers to the process of classifying each pixel of an image into semantically similar labels. The importance of semantic segmentation is highlighted by the fact that the inferred knowledge from imagery enables many applications such as virtual reality [1], human-computer interaction [2] and autonomous driving [3]. Traditional machine learning and computer vision techniques have been utilized to address such problems in the past, but with the emergence of deep learning, especially Convolutional Neural Network (CNN), the accuracy and efficiency of the approach has increased exponentially. The fully convolutional network (FCN) with an encoder-decoder architecture has been the popular paradigm for semantic segmentation [4]. However, one weakness of the pure convolution architecture is that the global context is unavoidably not well modeled.

Recently, the new kind of neural architecture transformer, which can provide the relationships between different features based on self-attention mechanism, has been widely promoted as a powerful alternative for computer vision problems [5]. Specifically, Han et. al proposed the Transformer-iN-Transformer network architecture which takes into account the attention inside the local patches of images and achieved better accuracy on the ImageNet benchmark [6]. Besides, the U-Net architecture, which decodes that up-samples features using transposed convolution corresponding to each downsampling stage, presented good performance for medical image segmentation tasks [7]. In this project, we will focus on exploring a novel model structure based on TNT and U-Net to realize better semantic segmentation performance for autonomous driving.

As an interesting and practical topic, the semantic segmentation for autonomous driving has attracted extensive studies. The majority of the algorithms are developed based on deep learning techniques. Visual Geometry Group-16 (VGG-16) introduced by the University of Oxford is a pioneering deep CNN model [8]. It is composed of 16 weight layers and uses a stack of convolution layers with small receptive fields in the first layer. In addition, ResNet18 is another advanced benchmark model which incorporates the usage of residual blocks that directs the network toward learning the residual representation on identity mapping [9]. In this study, the proposed model will be compared with these state-of-the-art models.

Datasets

In order to apply our new proposed model on autonomous driving, we plan to try it on two street-view datasets, KITTI [10] and CityScapes [11]. KITTI dataset consists of 11 categories: building, tree, sky, car, sign, road, pedestrian, fence, pole, sidewalk, and bicyclist. The dataset contains 200 training data and 200 testing data. We can only use the training dataset for training and self-evaluation because the website does not provide annotations of testing data. But it provides a submission application to evaluate our results.

(a) (b)

Figure 1. (a) Image and (b) annotation sample from KITTI dataset

CityScapes dataset contains 30 classes shown in figure 2 (a), and figure 2 (b) shows the sample in the dataset. It provides 5000 data for training and validation. We will split the dataset into a 7:3 ratio of training versus testing.

Group	Classes
flat	road · sidewalk · parking[+] · rail track[+]
human	person[*] · rider[*]
vehicle	car[*] · truck[*] · bus[*] · on rails[*] · motorcycle[*] · bicycle[*] · caravan[*+] · trailer[*+]
construction	building · wall · fence · guard rail[+] · bridge[+] · tunnel[+]
object	pole · pole group[+] · traffic sign · traffic light
nature	vegetation · terrain
sky	sky
void	ground[+] · dynamic[+] · static[+]

(a) (b)

Figure 2. (a) Classes and (b) sample of CityScapes dataset.

Method

In recent years, transformer mechanisms are prevailing in the computer vision domain. Transformer-in-Transformer is proposed as a new mechanism based on the transformer. Also, U-Net has an interesting and state-of-the-art semantic segmentation model. As a result, we plan to build a new semantic segmentation model, TNTU-Net. We leverage the precise localization of U-Net and global understanding of Transformer-in-Transformer (TNT) to create TNTU-Net to see if it will perform better than recent semantic segmentation models. By using this model, we are able to recognize the objects in the datasets with precise pixel localization and high accuracy. Figure 3 below illustrates the architecture of the model. We chose to use semantic segmentation because it is a relatively new technique and also allows us to apply the concepts we used in class. We feel that through this project, we will get a better understanding of this model.

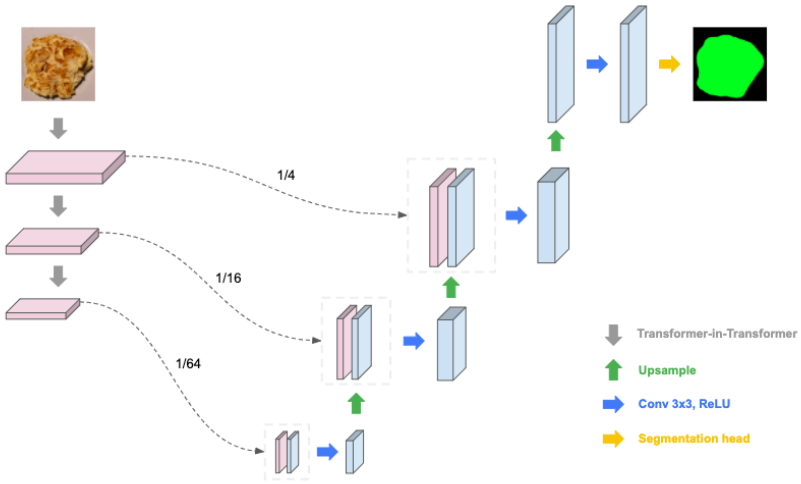

Figure 3. The model structure of TNTU-net

Outcome and Performance Evaluation

We will evaluate the model performance by the precisions and mean IoU of all the categories of interest as well as plotting a confusion matrix to better understand the performance of each category. Finally, we will compare our results with benchmark models to see if our model is better or worse, trying to figure out what else we can improve.

Project plan

In the following chart, we have outlined the tasks we will perform to complete the project and the deadlines for the completion of each task. As the semester continues, we will check this timeline to ensure that we manage our time effectively and keep the project on schedule.

Date	**Objective**
2/24	Project proposal due
2/24 – 3/10	Benchmark model review and construction
3/10 – 3/31	Proposed model construction and evaluation
3/31	Project mid-term report due
3/31 - 4/22	Model fine-tuning and evaluation
4/22 - 5/6	Final write-up, presentation preparation, website completion

References

[1] Perry, J., & Fernandez, A. S. (2020, August). Eyeseg: Fast and efficient few-shot semantic segmentation. In *European Conference on Computer Vision* (pp. 570-582). Springer, Cham.

[2] Benitez-Garcia, G., Prudente-Tixteco, L., Castro-Madrid, L. C., Toscano-Medina, R., Olivares-Mercado, J., Sanchez-Perez, G., & Villalba, L. J. G. (2021). Improving real-time hand gesture recognition with semantic segmentation. *Sensors*, *21*(2), 356.

[3] Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F., Schuberth, P & Hochreiter, S. (2016). Speeding up semantic segmentation for autonomous driving.

[4] Briot, A., Viswanath, P., & Yogamani, S. (2018). Analysis of efficient cnn design techniques for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 663-672).

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[6] Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*.

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[8] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[10] Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., & Rother, C. (2018). Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision, 126*(9), 961-972.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.