CS 766 Mid-Term Report
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. In addition, several loss functions are considered in the model formulation to improve the model performance as well as tackle the data imbalance issue. 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 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. 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. 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 the self-attention mechanism, has been widely promoted as a powerful alternative for computer vision problems [1]. 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 [2]. Besides, the U-Net architecture, which decodes that up-samples features using transposed convolution corresponding to each downsampling stage, presented a good performance for medical image segmentation tasks [3].

While U-Net is capable of capturing localized features, TNT focuses more on global information extraction. In this project, a novel model structure incorporating TNT and U-Net is explored to realize better semantic segmentation performance for autonomous driving.

Datasets

In order to apply our proposed model on autonomous driving, we plan to try it on two street-view datasets, KITTI [4] and CityScapes [5]. 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.

To find the most suitable loss function for the model, four kinds of loss functions were adopted. First, Binary Cross-Entropy loss [6] is the most common and basic for solving classification problems. The equation is shown below:

$$BCE(p) = -\frac{1}{N}\sum_{i=1}^{N} y_i \times log(p(y_i)) \times log(1 - p(y_i))$$

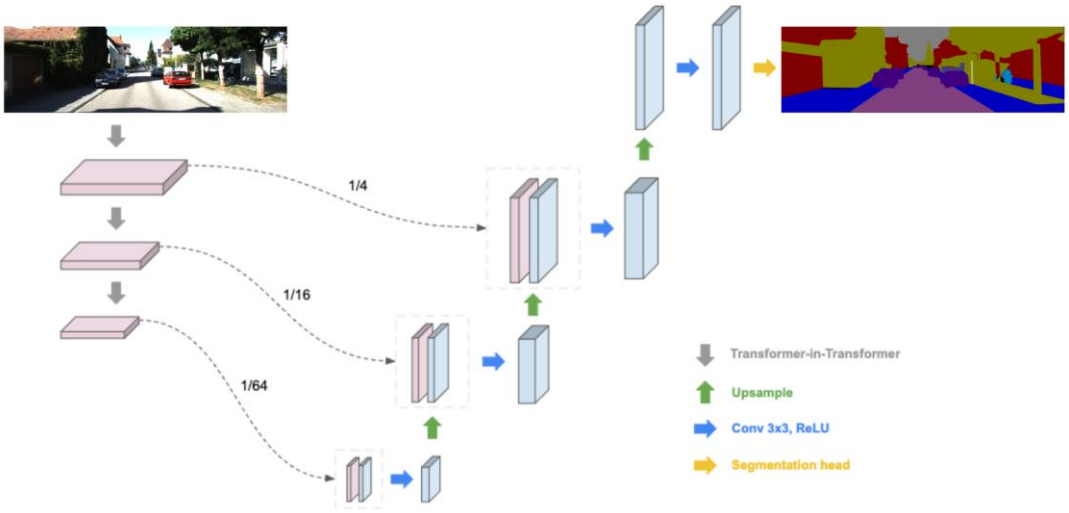


Figure 3. The model structure of TNTU-net

Data imbalance issues almost happen in every deep learning model training for the image recognition task, because it is hard to control the data distribution when we collect data. Moreover, the areas of objects of interest usually constitute small proportions of the images. It is noticed that both KITTI and Cityscapes datasets have the data imbalance issue. As a result, the following loss functions aim to solve this issue. Second, Focal loss [7] aims to solve the data imbalance issue of the object detection task by adding a modulating factor γ ($\gamma \geq 0$). Let p denote the probability from the predictions. The focal loss can be expressed as:

$$FL(p) = -(1-p)^{\gamma} log(p)$$

Third, dice loss [8] aims to solve the data imbalance problem of semantic segmentation problem, commonly being used in recognizing the medical image. Let p denote the probabilities from the predictions, and g represent the probabilities for the ground truth. Then the dice loss can be shown as:

$$Dice(p,g) = \frac{2\sum_{i=0}^{N} p_i g_i}{\sum_{i=0}^{N} {p_i}^2 + \sum_{i=0}^{N} {g_i}^2}$$

Last, we combine focal loss and dice loss with half and half weight respectively. The equation is shown below:

$$Loss(p,g) = 0.5 \times FL(p) + 0.5 \times Dice(p,g)$$

Outcome and Performance Evaluation

The model performance is evaluated 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 performs better, and try to figure out how we can improve it.

Results

We train the model with four different loss functions (binary cross-entropy loss, focal loss, dice loss, and focal loss + dice loss) for 100 epochs, and below are the prediction results. As shown in the figure below, the models with binary cross-entropy loss and focal loss focus less on the small proportion of objects (e.g. signs). And their prediction performances are better for the bigger proportion of objects (e.g. cars) because the model will not be affected by the small objects. As for dice loss, the model focuses more on the small objects. Moreover, the model with the combination of focal loss and dice loss functions performs almost the same as the model with the focal loss function, because the focal loss is still dominant in the combined loss function. Thus, in the future, we will adjust the weights of focal loss and dice loss to improve the performance by focusing more on the small objects.

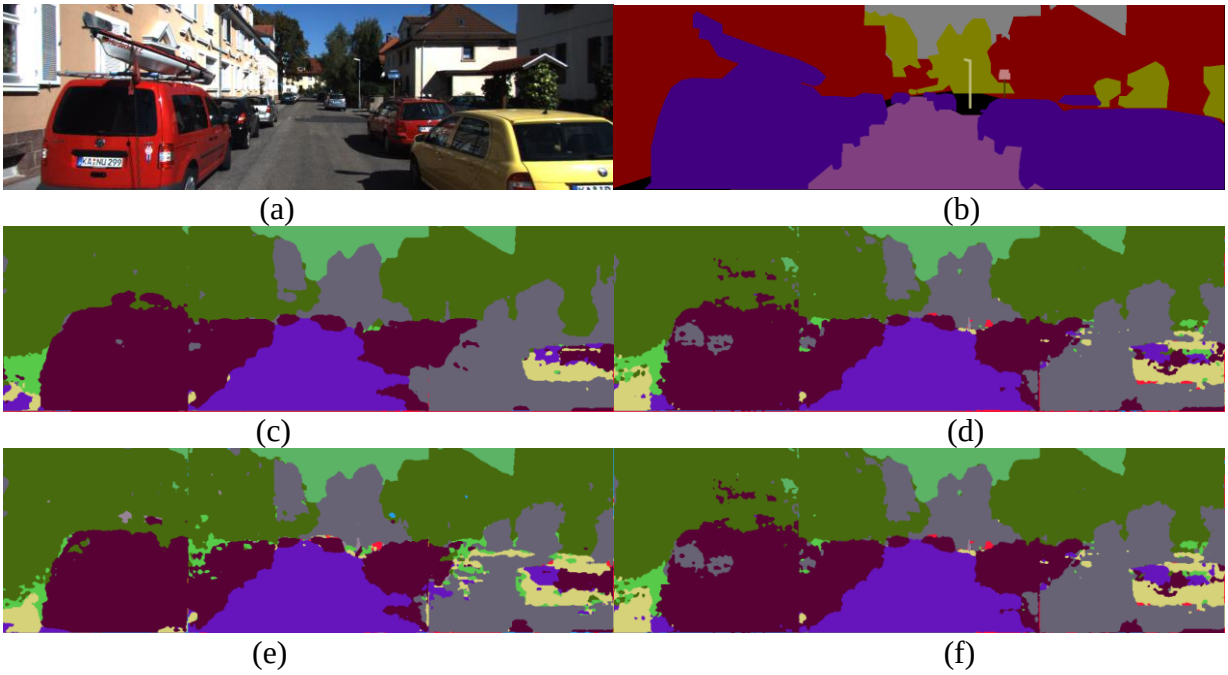


Figure 4. Prediction outputs with different models: (a)Data image, (b) Ground truth label, and the predictions of the model train with (c) Binary cross-entropy loss (d) focal loss, (e) dice loss, (f) focal loss + dice loss

The confusion matrices of the model trained with different loss functions are shown below. These confusion matrices indicate that the models did not pay enough attention to the small objects (e.g. signs, people, and cyclists). We plan to separate the small and big objects into two different datasets to train two different models and combine the results.

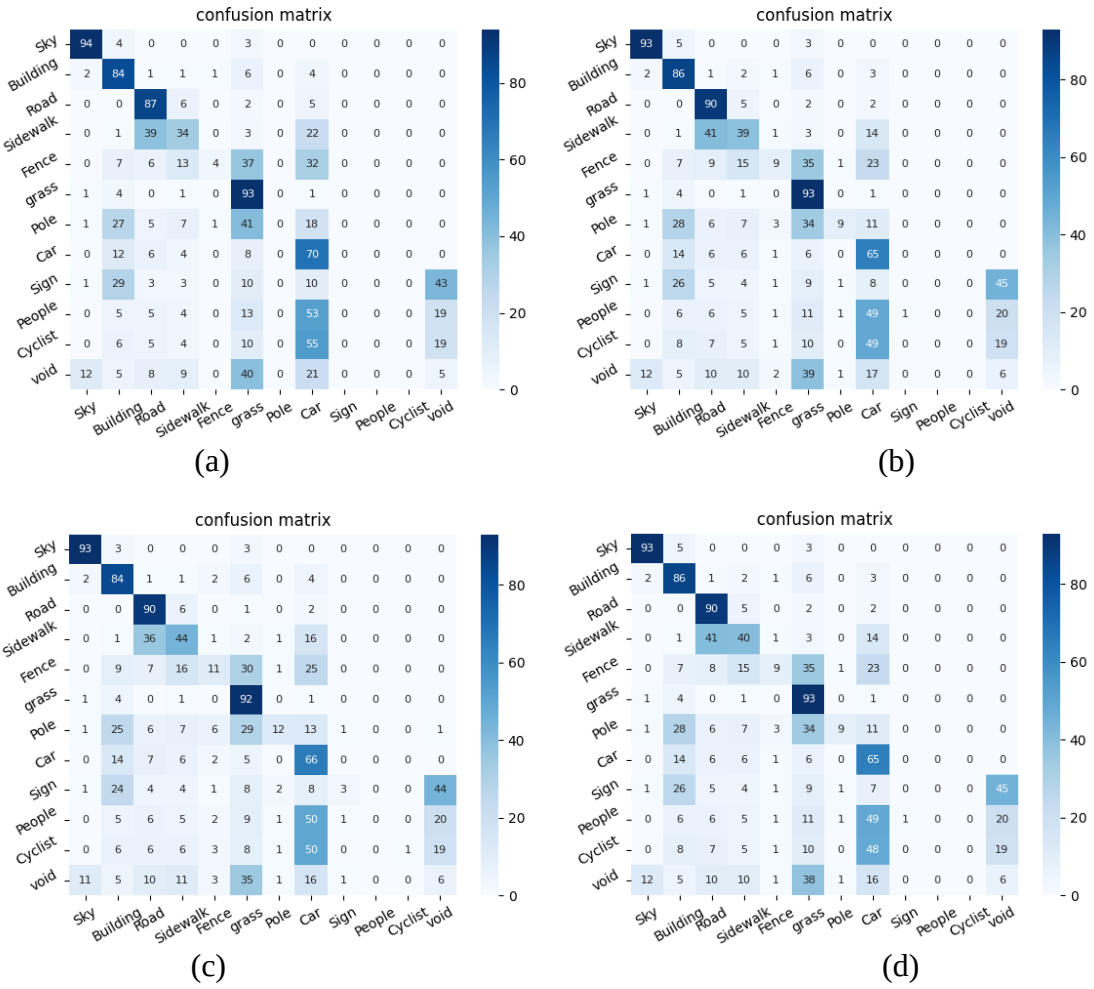
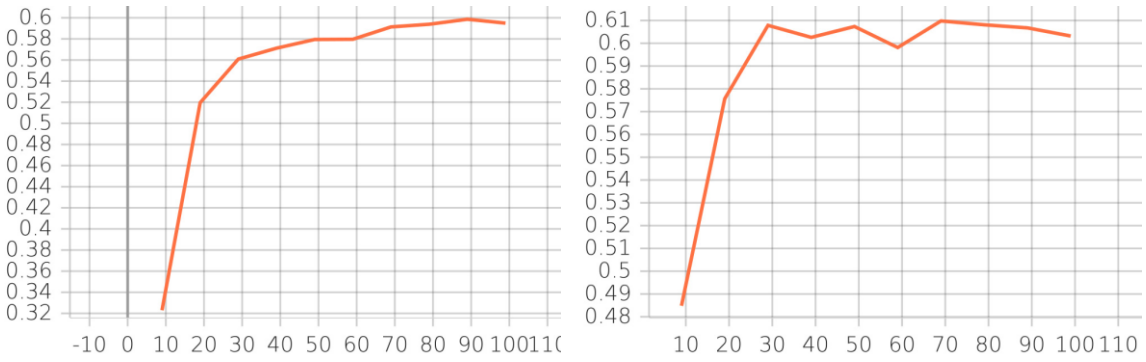
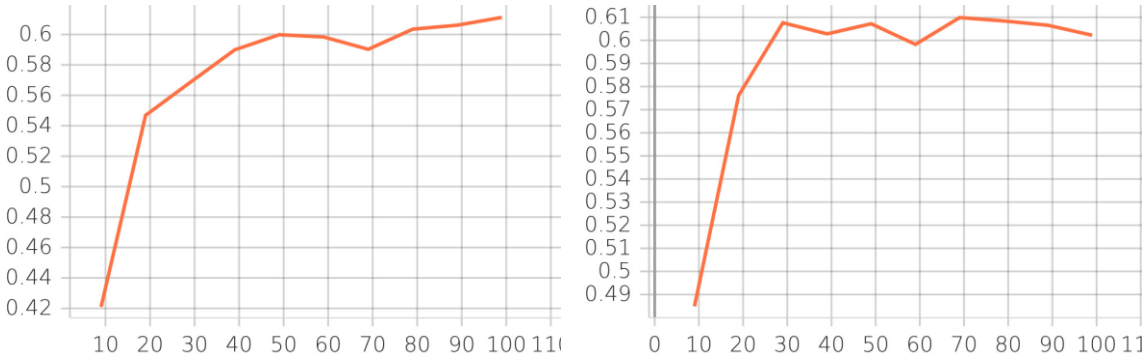


Figure 5. The confusion matrices of the model train with (a) binary cross-entropy loss (b) focal loss, (c) dice loss, (d) focal loss + dice loss

The validation accuracies for different models trained with different loss functions are shown below. The accuracy curve of the model with dice loss function still goes up at the end of the training process. Thus, we plan to train the model for 200 epochs to figure out how many epochs are enough for this model.





(c) (d)

Figure 6. The confusion matrices of the model train with (a) binary cross-entropy loss (b) focal loss, (c) dice loss, (d) focal loss + dice loss

Project Progress and Future Plan

In the past weeks, we have successfully established and evaluated the proposed TNTU-Net model for semantic segmentation. To tackle the data imbalance issue and improve the model performance, we have also tested several different loss functions for the model. For the benchmark models, we have built up and evaluated some fundamental models, and the overall comparison between the proposed model and the benchmark models will be presented in the final report.

In the following chart, we have outlined the tasks that we have completed as well as the tasks that we will perform in the project. 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 – 3/20	Proposed model construction and evaluation
3/20 – 4/4	Model improvement considering data imbalance issue
4/5	Project mid-term report due
4/5 - 4/22	Model fine-tuning and evaluation. Comparison with benchmark models
4/22 - 4/28	Final write-up, presentation preparation, website completion

References

[1] 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*.

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

[3] 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.

[4] 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.

[5] 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.

[6] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[7] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26(3)*, 297-302.

[8] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).