Image Recognition in Autonomous Driving Based on Improved Swin Transformer

Yequan Bie Hao Tan

South China University of Technology

ABSTRACT

Traffic image recognition is one of the most important phases in the field of autonomous driving, including the classification of real-time periods and the detection of pedestrians, vehicles, etc. on the road. In this paper, we proposed an end-to-end classifcation and object detection method based on Swin Transformer with improved cascade RoI heads. Our method focuses on the scale problem from language to vision field in traditional Transformer model and the mismatch problem of bounding box regression in previous object detection methods (e.g. Faster R-CNN). A modified Swin Transformer architecture with multiple RoI heads is adopted in the proposed model to perform classification and object detection, meanwhile improved optimization strategies are used. We applied the model to SODA10M, an autonomous driving dataset released by Huawei, and finally attained a classification accuracy of 95.3% and a detection mAP of 91.9%, both achieving state-of-the-art.

INTRODUCTION & DATASET

In this paper, we propose a more innovative method for image classification and object detection on the dataset SODA10M. For the classification problem, we use an improved Swin Transformer to classify the period of the day in which the image scene is located. For the object detection problem, we aim at the scale problem from language to vision field, that is, tokens are of a fixed scale but vision applications like object detection are not, combine the two models of Swin Transformer and Cascade R-CNN with appropriate modification, and successfully achieve high performance.

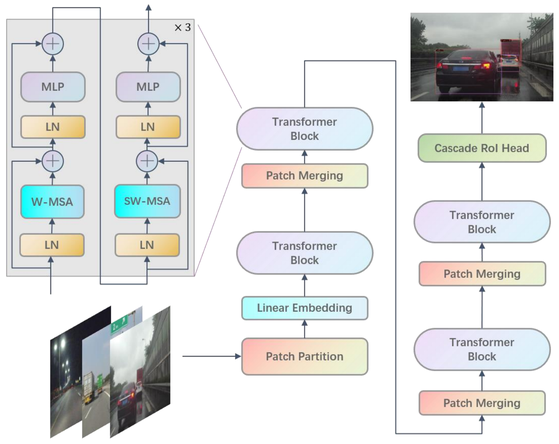
SODA10M is a new large-scale 2D dataset released by HUAWEI Noah's Ark Lab in 2021, which contains 10M unlabeled images and 20k labeled images with 6 representative object categories. This dataset is the largest 2D autonomous driving dataset until now which is ten times larger than Waymo dataset. The rich diversity ensures its generalization performance as self-supervised pre-training data set and semi-supervised additional data in downstream autonomous driving tasks.

The image collection task is distributed to the tens of thousands taxi drivers in the form of crowdsourcing. They are required to use the mobile phone or driving recorder to obtain images every ten seconds per frame. The horizon needs to be kept at the center of the image and the occlusion inside the car should not exceed 15% of the whole picture. To achieve more diversity, suppliers are required to obtain images in diverse weather conditions, periods, locations and cities.

METHOD

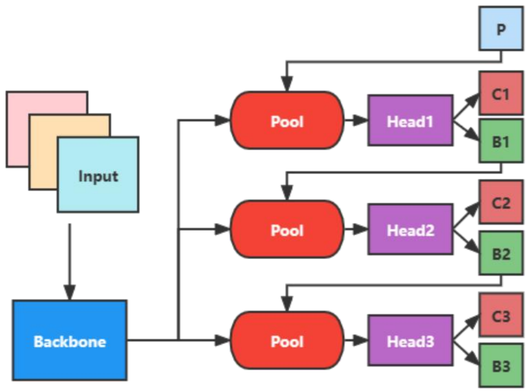
A. Architecture

We first split RGB images of SODA10M dataset into patches which are treated as " tokens" in transformer model by patch partition module. Then we use a linear embedding layer to change the dimension. Several Swin Transformer blocks are applied on these tokens together with Patch Merging module. Finally, we propose an architecture with improved cascade RoI heads to help performing object detection.



B. Improved RoI Head

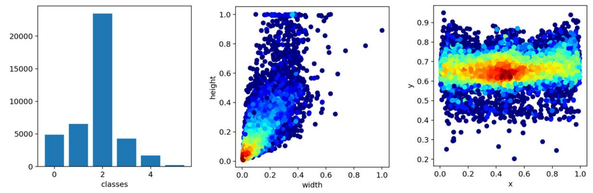
In the object detection task, we know that the Faster R-CNN has a wide range of applications as a classic model. However, it is very difficult to ask a single regressor which is used in Faster R-CNN to perform perfectly uniformly at all quality levels. Therefore, we can decompose the difficult regression task into a sequence of simpler steps, which is also in line with the idea of transformer. In order to apply Swin Transformer to task of object detection, we combine Cascade R-CNN with it. Specifically, in the downstream task of object detection, this paper uses the improved Swin Transformer as the backbone, and uses the cascade RoI head for object detection bounding box regression and object classification. In addition, different from the regression loss used by Cascade R-CNN, we use GIoU Loss for bounding box regression optimization.



EXPERIMENTS

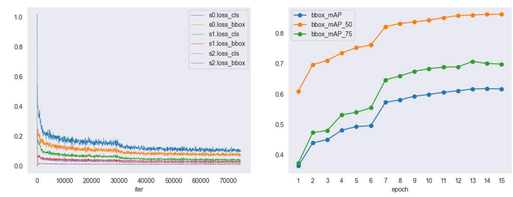
A. Exploratory Data Analysis

The distribution of the labeled boxes of the training data is shown in the figure below. It is obvious that there are very few samples of the " tricycle" category, meaning that there is a data imbalance problem, which will definitely affect the performance of the model. Thus we take appropriate resampling strategy on the "tricycle" category, and it is verified that the model mAP increases by about 2.6% after resampling.



B. Model Settings

We compared models of different complexity and give the best model on SODA10M based on experimental results. The initial embedding dimension is set to 128 and 4 swin transformer blocks are stacked 2, 2, 16 and 2 times respectively. The number of heads in multi-head self attention is 4, 8, 16 and 32. For the detection head, a total of 3 cascade RoI heads are used. The weights of loss are 1, 0.5 and 0.25 respectively. Moreover, we adopt GIoU Loss instead of smooth L1 Loss in bounding box regression.



C. Implementation Details

Our network is implemented based on PyTorch. We complete all the experiments on Nvidia GeForce RTX 2080 Ti. Training for the proposed network takes about 9 hours. For classification task, we apply Swin Transformer as feature exactor with MLP head. The loss function we adopt is multi-class cross entropy. And the optimization method is SGD. The weight decay and momentum are set to 0.0001 and 0.9, respectively. The weights we use are pretrained from ImageNet-1K. After 10 epochs of training, the model converges rapidly. The top-1 accuracy achieves 95.3% in the end. It can be observed that using Swin Transformer as backbone, the convergence speed and model accuracy both are excellent on the classfication task. As for the detection task, we apply the AdamW optimization algorithm with learning rate policy of linear warmup and step decay. The learning rate will increase to 0.0001 in 500 iterations and then divided by 10 after the 6th and the 12th epoch. For the best results, we ultimately decline the learning rate by 10 times in the last epoch. The weight decay is set to 0.05 to alleviate overfitting. On top of that, we adopt multi-scale training for better performance. The training of models based on CNN (i.e. YOLO and RetinaNet) follows the default settings.

RESULT ANALYSIS

Comparison between different methods including SOTA on SODA10M validation set.

Methods	pedestrain	cyclist	car	truck	tram	tricycle	mAP_{50}
RetinaNet [18]	58.92	61.83	66.14	72.43	62.92	38.82	60.2
YOLOv5 [19]	85.37	84.29	90.12	91.36	86.63	68.32	84.3
YOLOR [20]	83.92	87.48	92.36	91.85	87.32	64.92	84.6
YOLOX [21]	85.37	89.65	93.12	91.23	89.36	66.12	85.8
Ours	89.45	95.49	96.84	94.37	95.77	79.43	**91.9**

Our model outperforms most popular one-stage object detection methods. The mAP of validation set reaches 91.9%. Our model performs particularly well in the recognition of "car" and "cyclist" (95.49% and 96.84%, respectively), especially cyclist category, which is significantly better than YOLO series. This may be due to the self-attention mechanism in Swin Transformer, which can quickly lock the region with obvious features in the image. The addition of the shifted window module ensures that there is continuity between the different windows, making it more helpful for big targets that can easily be shelled. The main constraint on the model performance is "tricycle" (only 79.43%AP). But even so, our method significantly outperforms RetinaNet by 40.61%.

CONCLUSION

In this paper, an end-to-end classification and object detection method based on Swin Transformer with improved cascade RoI head is proposed and is applied to autonomous driving dataset. Our model focuses more on the scale problem from language to vision field in traditional Transformer model through self-attention and shifted window scheme, and applied cascade RoI head with different weights to decompose the bounding box regression task into a sequence of smaller steps, which is also in line with the idea of Transformer as a sequence model. We also improve the optimization method by modifying the loss function and training strategy.

Our model performs both classification and object detection tasks in SODA10M and achieves state-of-the-art, which attains 95.3% (top-1 accuracy) and 91.9%, respectively. We believe that our model can also be used and performs well in other fields besides traffic image recognition.