2024 Spring Computer Vision VIVOTEK Final Project

CHIH-HAO LIAO[1,2,*], YI-HAN LEE[2,+], and HSIN-TZU LI[2,+]

[1]School of Forestry and Resource Conservation, National Taiwan University, Taipei, 106319, Taiwan
[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, 106319, Taiwan
[*]R11625015r@ntu.edu.tw
[+]these authors contributed equally to this work

ABSTRACT

Automated Passenger Counters (APCs) are essential for analyzing passenger behavior in public transportation, aiding in the enhancement of operational efficiency and service optimization. This final project introduces a Transfer Learning framework utilizing the Swin Transformer V2 model to detect door status (open or closed) through camera images. By implementing a pre-trained model on the camera, the system can accurately determine door status in real-time, achieving up to 99.6% accuracy. The findings demonstrate that incorporating a pre-trained door status detection model into APC systems on cameras is a viable solution.

Introduction

This report serves as the final project for Prof. Shao-Yi Chien's spring course on Computer Vision: from Recognition to Geometry at National Taiwan University, and the final competition is sponsored by Vivotek. The final competition involves developing a system for real-time monitoring of door status in public transit systems. In this project, we developed a fine-tuned Swin Transformer V2 model capable of accurately predicting the status of doors in public transit systems. The model can detect and localize doors using video from cameras, and it provides precise monitoring of door statuses, including Opening, and Closing statuses.

Background

Automated Passenger Counters (APCs) are critical devices installed on public transit vehicles, such as buses and trains, to accurately record the times and locations of passengers boarding and disembarking. This data is pivotal for analyzing travel patterns, optimizing routes, and enhancing the overall efficiency of transportation services. Traditionally, APC systems depend on real-time signals from vehicle doors to determine when they open and close, which triggers the counting process. However, integrating APC systems with the door status signals in older public transit vehicles presents significant challenges[1] due to the difficulty of accessing and correctly connecting the required wiring.

To address this issue, there is a growing need for a vision-based automatic door status monitoring technology that eliminates the reliance on external wiring for door status signals. The goal of this project is to integrate a vision-based door status detection model into APC systems, demonstrating its feasibility and effectiveness in real-world applications. This approach not only simplifies the installation process on older vehicles but also enhances the reliability and accuracy of passenger counting, ultimately contributing to better data collection and improved transit service management.

Methods

Dataset

The dataset utilized in the final project was constructed from video frames, which were extracted using OpenCV with its default frames per second (FPS) setting. This process ensured a consistent and automated extraction of frames from each video, providing a reliable and uniform input for model training. The videos used in this study were provided by Vivotek, encompassing three different types of scenarios and environments to ensure a comprehensive dataset. The videos were selected to cover various conditions, including different window lighting, types of doors, and dynamic activities, to create a robust and diverse dataset suitable for training a generalized model.

The final dataset comprises a total of 4,152 images, organized in a structured format. The images were randomly divided into training and validation sets. The labels for the images are either "open" or "closed," and there are 808 images labeled "Open" and 2347 images labeled "Closed" in the training set, while images labeled "Open" and "Closed" in the validation set are 255 and 742 respectively.

Data processing

Introducing variability during the training process can enhance model robustness and generalization while ensuring consistency and standardization across datasets[2], thereby improving overall model performance across different datasets and targets. In this project, the process begins by resizing each image to match the dimensions expected by the feature extractor. Then, converting them to tensors, and normalized them using the mean and standard deviation values from the feature extractor of the pre-trained model. This normalization standardizes the pixel values across all images, ensuring consistent input for the model during training. The difference between the training set and the validation set is that there are no data augmentation techniques applied to the validation set, focusing solely on preparing the images for evaluation without altering their content. The data augmentation techniques applied to the training set include random horizontal flipping and random rotation (up to 15 degrees) since the traffic door has a symmetric structure and the door may appear at any position in the video.

Swin Transformer V2

The Swin Transformer is a type of Vision Transformer designed for image classification and other vision tasks[3]. It introduces a hierarchical architecture that processes images at multiple scales using shifted windows for efficient computation. This approach allows the model to handle high-resolution images and capture local context effectively. It adopts a hierarchical feature map architecture similar to convolutional neural networks (CNN), enabling the creation of feature maps with varying downsampling rates (4x, 8x, 16x, see Figure 1), which facilitates tasks sucks as object detection and instance segmentation. This contrasts with the Vision Transformer, which applies a uniform 16x downsampling from the beginning and maintains it throughout.

Additionally, Swin Transformer introduces a novel mechanism called Shifted Windows Multi-Head Self-Attention (SW-MSA) to extract local features by shifted windows (see Figure 2), followed by performing self-attention within each window. This approach not only enables cross-window information flow, and enhances the model's representational power, but also reduces the computational costs, especially for large feature maps, compared to the Vision Transformer's global self-attention. As illustrated in Figure 3, the architecture of the Swin Transformer consists of four stages, including patch partition, linear embedding, patch merging, and Swin Transformer block.

In this project, we utilized the Microsoft Swin Transformer v1 as our baseline model to predict the status of the door from each video frame. The specific model architecture implemented with PyTorch is the Swim Transformer v2 base[4], and the model was then further improved by applying transfer learning from the pre-trained model from Hugging Face[5], which is inherited from Microsoft swinv2-tiny-patch4-window8-256[4].

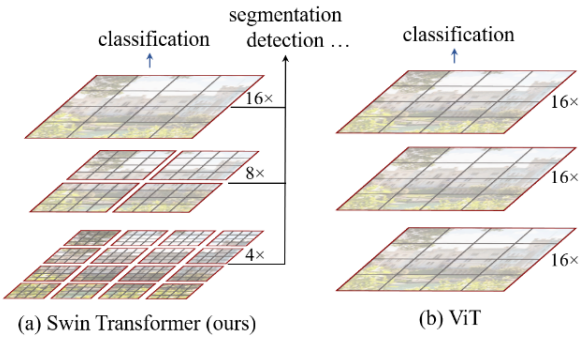

Figure 1. (a) Swin Transformer builds hierarchical feature maps by merging image patches; (b) In contrast, Vision Transformers produce feature maps of a single low resolution.

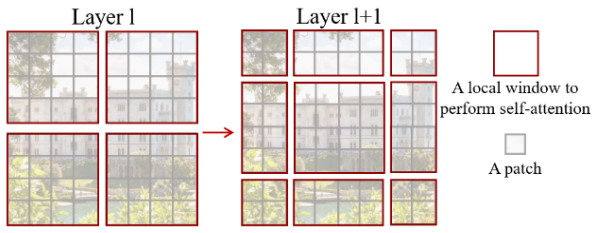

Figure 2. An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architecture.

Algorithm for Guessing Logic

Due to certain constraints within our dataset (to be elaborated in the following section), our methodology for estimating the gate opening time employs an approach analogous to identifying the start and end positions of the Longest Substring Without Repeating Characters question, with a requirement for redundancy of at least two video frames. We operate under the assumption that the video captures a singular occurrence of the gate operation, allowing us to deduce the earliest and latest possible times as the opening and closing instances, respectively. This forms the foundational framework of our predictive algorithm, distinct from direct reliance on outputs generated by a pre-trained model.

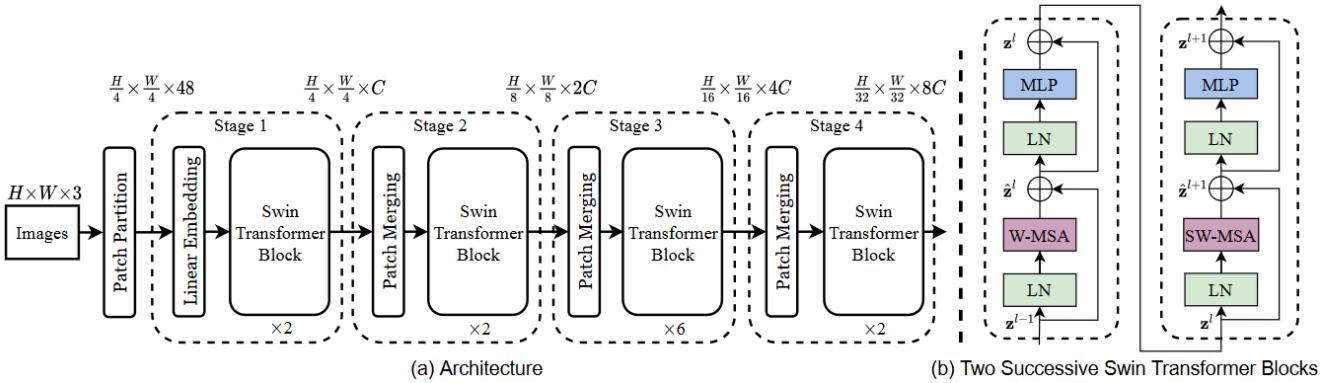

Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks.

Limitations

In this final project, the dataset is limited to three videos supplied by Vivotek. The testing footage may include content captured with fisheye cameras, resulting in elliptical or circular images rather than standard rectangular frames. Furthermore, the door positions within the footage are not consistent, and there is significant variation in door types. The training data we have only represents a limited range of scenarios, which is insufficient to cover all potential variations comprehensively. Consequently, a more extensive and diverse dataset is required to ensure robust performance across a broader range of possible conditions.

Results

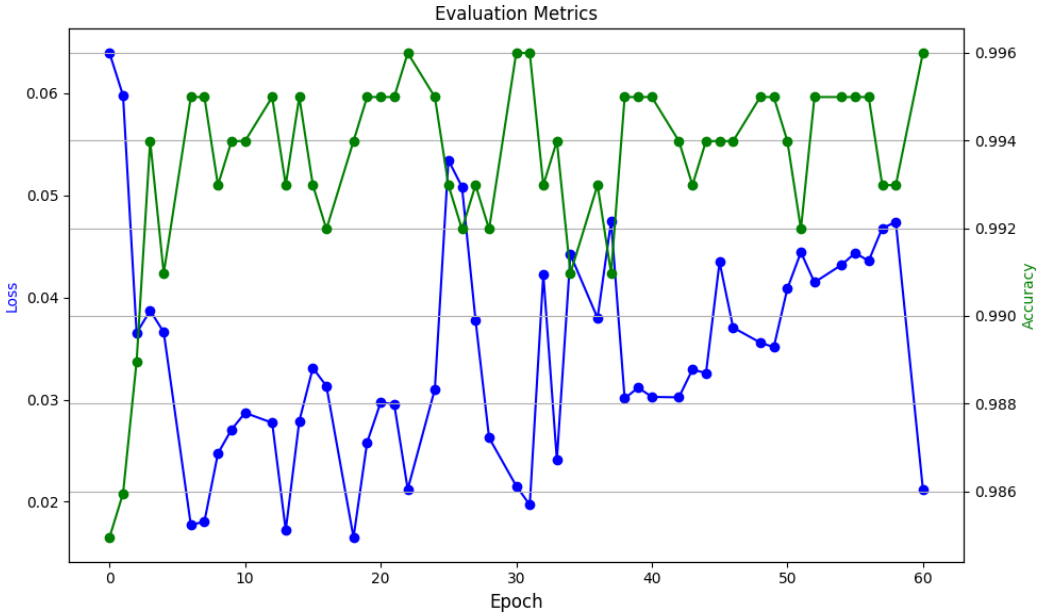

Figure 4. The loss curve and accuracy of the validation set.

Based on the training results detailed in Figure 4, Swin Transformer V2 achieved significant performance on the classification task, demonstrating robust accuracy metrics over epochs. The training regimen spanned 20 epochs, with periodic evaluations every epoch. The final epoch showed the highest accuracy, reaching 99.6%, accompanied by a corresponding low evaluation loss of 2.12

Throughout the training, the model consistently improved, as evidenced by the evaluation accuracy increasing steadily from 98.5% in the initial epoch to 99.60% by the final epoch. Concurrently, the evaluation loss decreased from 6.39% to 2.12% over the same period, indicating effective convergence and generalization capability.

In conclusion, the reported findings highlight the model's robust performance and efficiency in the classification task, substantiated by meticulous evaluation and logging practices throughout the training epochs. These results underscore the effectiveness of the chosen deep learning architecture and training strategy in achieving state-of-the-art performance on the given dataset.

Work Distribution

Each member of our team makes an equitable contribution to the project, demonstrating diligent effort.

References

1. Kotz, A. J., Kittelson, D. B. & Northrop, W. F. Novel vehicle mass-based automated passenger counter for transit applications. *Transp. Res. Rec.* **2563**, 37–43, DOI: 10.3141/2536-05 (2016). https://doi.org/10.3141/2536-05.

2. Rebuffi, S.-A. *et al.* Data augmentation can improve robustness (2021). 2111.05328.

3. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).

4. Liu, Z. *et al.* Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

5. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).