MSc Project - Reflective Essay

Project Title:	Pedestrian Action Recognition in Self-driving System Using Video Vision Transformer with Variable Frame Lengths
Student Name:	Hyowon Lee
Student Number:	230861592
Supervisor Name:	Georgios Tzimiropoulos
Programme of Study:	MSc Artificial Intelligent

This research suggests a training method approach for the model recognise the pedestrian’s action. I trained the model with three different training methods to compare the improvement. Firstly, training the model without any data augmentation. In other words, train the model only with original dataset. Second, training the model with data augmentation. I extend the dataset as 4 times of original dataset. Lastly, training the model with various frame numbers. Other two experiments are used 16 frames which is a fixed frame number. On the other hand, last experiment trains the model various random number from 16 to 32 frames.

1. Strengths

1.1 Data regularisation and data augmentation

By slicing the frames within a random range between 16 to 32 frames, data augmentation is achieved. Furthermore, one video is sliced four times with different frames, leading to data generalisation.

1.2 Training with not enough dataset

I utilised a pre-trained model provided by Hugging Face, meaning the model had already been trained on a vast dataset. As a result, it could perform well even with a relatively small amount of data.

2. Weaknesses

2.1 Video quality

The model's performance is contingent upon the quality of the video. I trained the model using two datasets: JAAD and PIE. The JAAD dataset is of lower quality than the PIE dataset, even though they are of similar size. This lower quality leads to reduced model performance, despite attempts to enhance video quality using the OpenCV library.

2.2 Handling padding

It was challenging to adjust the model settings for each video. To standardise the number of frames per video, I used padding, processing it within the model. However, because I employed a pre-trained model for transfer learning, I was unable to modify the model architecture to handle padding directly.

2.3 Processing the video only considering frame

If pedestrian actions vary within a single video, it becomes difficult for the model to accurately process all actions since this is a recognition model. The model is trained with one label per video. Additionally, using an open dataset can result in excessively short videos when sliced according to each action label. Therefore, in this research, it was deemed better to use the entire video for training. However, if the actions vary within a video, the model may be trained incorrectly.

3. Presentation of possibilities for further work

3.1 Processing padding

Due to the use of a pre-trained model architecture, it was challenging to process padding directly. In other words, it was difficult to modify the model architecture. However, if the Video Vision Transformer model architecture could be altered, employing masked self-attention could be a potential solution for processing padding. According to Grisi et al. (2024), a Vision Transformer model using masked self-attention enhances performance by removing unnecessary patches, providing more accurate and clinically significant heatmaps. (Grisi. C, et al, 2024) Since the Video Vision Transformer is based on the Vision Transformer architecture, it is anticipated that masked self-attention could process padding directly, potentially improving model performance in future work.

3.2 processing each label of pedestrian

In this study, each pedestrian label was not processed separately. That is, even if a pedestrian has multiple labels within one video, the model only selects and processes the first label, treating the video as having a single label. The training dataset frames are relatively short, so the dataset was created using the first label of each video. However, if each pedestrian label is processed separately, a more accurate labelled dataset could be created for training. This approach is expected to enhance model performance through training with more precise labels.

3.3 extending domain

This study focuses on recognising pedestrian actions. Specifically, the model detects pedestrian actions using the JAAD and PIE datasets. However, this proposed method could be extended to other domains if the dataset is a video dataset. In other words, with a video dataset, the model's performance could be improved by training with a variable number of frames.

4. Critical analysis of the relationship between theory and practical work produced

This study deals with video data and focuses on pedestrian behaviour recognition, a field that has not been extensively studied compared to other areas, likely due to the more complex nature of image data. Despite this, the market for pedestrian behaviour recognition within artificial intelligence has been steadily growing. Fortunately, this project had access to numerous open datasets related to pedestrian behaviour, making data collection easier. Among the open datasets, the JAAD and PIE datasets were used to compare and analyse the impact of data augmentation, as well as the effects of using a fixed versus variable number of frames on model learning. The primary goal was to improve the performance of the Video Vision Transformer model with image data.

Before commencing this project, I reviewed theories related to pedestrian action recognition in CNN and Vision Transformer models, which are prominent in computer vision. Since these models are designed for image learning, methods such as learning temporal elements with time models like LSTM were employed. During this process, I learned about the Video Vision Transformer model, which can learn video data by capturing temporal information from Vision Transformer. This model essentially adds a layer to learn temporal information based on the Vision Transformer architecture.

The project began with the hypothesis that varying the temporal information (i.e., frame length) could improve performance. However, according to Vision Transformer studies, it was challenging to directly implement the structure to learn from scratch due to insufficient data. Therefore, I opted for transfer learning using a pre-trained model from Hugging Face. Theoretically, it was determined that images with a varying number of frames could be learned by modifying time-related parameters, such as frame length. However, this method increases the model's complexity, as it necessitates the use of numerous 'for' loops to modify the frame length parameter for each batch. Additionally, due to the use of the collate function, the frames must be applied one by one. Therefore, this approach does not suit practical applications. As a result, padding was applied to process the different frame numbers by standardising the video dataset to the same number of frames.

Through this project, I gained a clearer understanding of the Video Vision Transformer model's structure, as well as the advantages and disadvantages of using a pre-trained model. While pre-trained models offer the benefit of learning without requiring extensive datasets, the Vision Transformer structure still demands a substantial amount of data, even when using a pre-trained model.

5. Awareness of Legal, Social Ethical Issues and Sustainability

5.1 Awareness of Legal and Ethical Issues

This project required pedestrian videos that included pedestrian actions in traffic situations. Therefore, it was crucial to handle personal data, such as facial images, while avoiding ethical issues and complying with laws, particularly the General Data Protection Regulation (GDPR). However, collecting pedestrian video data while protecting personal privacy proved difficult. Additionally, even if data collection is successful, blurring faces to protect personal data may reduce the model's accuracy. Consequently, I decided to use publicly available open datasets, minimising ethical issues and privacy concerns.

However, for future work, open datasets may not suffice to train the model with precision and accuracy. In such cases, custom datasets would be needed. If custom datasets are used, they must be processed in compliance with GDPR and ethical frameworks. It is well understood that protecting personal information can enhance the credibility of the research.

5.2 Social Issues and Sustainability

As the self-driving system market grows, this project has significant social implications, as it contributes to improving traffic safety and protecting pedestrians. On the other hand, this could also raise several social issues. According to a survey by Chougule et al. (2023), 50% of people tend not to trust self-driving cars regarding safety. This reflects a general anxiety about the accuracy of self-driving cars in recognising pedestrians, as well as concerns about accidents involving pedestrians or other vehicles. Furthermore, if a self-driving vehicle were to crash, it could spark controversy over liability. For sustainable development, this project must not only focus on technological advancements but also integrate with diverse social contexts, legal frameworks, and communication with stakeholders.

6. Reference

Grisi, C., Litjens, G., & van der Laak, J. (2024). Masked Attention as a Mechanism for Improving Interpretability of Vision Transformers. *arXiv preprint arXiv:2404.18152*.

Chougule, A., Chamola, V., Sam, A., Yu, F. R., & Sikdar, B. (2023). A comprehensive review on limitations of autonomous driving and its impact on accidents and collisions. IEEE Open Journal of Vehicular Technology.