Deep Learning on Image Classification: Hybrid Model for Distracted Driver Detection

Lok Ping Joanna Wang, Mo Zhou

Introduction

According to the National Highway Traffic Safety Administration (2024), 3,308 people were killed and 289,310 people injured in motor vehicle crashes because of distracted drivers in 2022. The estimated economic cost of distracted-driving traffic crashes is $98 billion. By detecting and alerting drivers when they are distracted, technology can help prevent accidents, saving lives and reducing injuries.

Objectives

This project proposes to combine CNN with different transformer architectures by adding a transformer layer on top of a CNN pre-trained model serving as the feature extractor. The project will compare the novel architecture to previous solutions, such as comparing hybrid models composed of transformers and CNN to pre-trained CNN models including ResNet and VGGNet, and transformers including ViT and Swin, as the baseline.

Dataset

The dataset we will use is State Farm Distracted Driver Detection from kaggle. It includes driver images taken in a car with a driver doing something in the car (texting, eating, talking on the phone, makeup, reaching behind, etc).

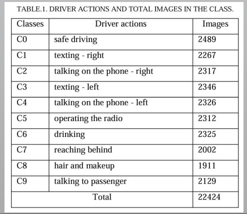


TABLE.1. DRIVER ACTIONS AND TOTAL IMAGES IN THE CLASS.

Classes	Driver actions	Images
C0	safe driving	2489
C1	texting - right	2267
C2	talking on the phone - right	2317
C3	texting - left	2346
C4	talking on the phone - left	2326
C5	operating the radio	2312
C6	drinking	2325
C7	reaching behind	2002
C8	hair and makeup	1911
C9	talking to passenger	2129
Total		22424

Related Work

1. Yan et al. (2016) used CNN on Southeast University Driving-posture Dataset (SEU dataset) and performed an accuracy of 99.47%. But our dataset includes more distracted driver image classes including operating car accessories, talking to passengers, reaching behind

2. Koesdwiady et al. (2017) fine-tuned pre-trained VGG-19 and reached an overall accuracy of 95% and an average accuracy of 80% per class on the test set. Koay et al. (2021) trained and compared a total of 17 CNN models and 16 Vision Transformers, and showed pre-trained CNN models outperform Vision Transformers. This demonstrates that pre-trained CNN models have a satisfying result and this project will use them as the baseline.

3. Li et al. (2022) proposed a ViT-Conv module that combines CNN with ViT and has shown it outperformed VGG-19. So we proposed hybrid models pre-trained CNN + transformer like RestNet with Swin might have better performance.

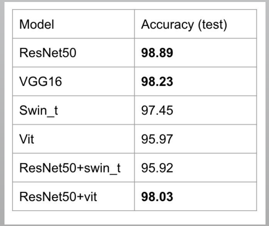
Methods

1. Choose 2 pre-trained models (RestNet and VGG) as baseline. Test their performance on this dataset, and fine-tuning them;

2. Train 2 transformer models (Vit and Swin) based on the pre-trained weights from 'ImageNet', and compare their performance with each other and CNN models in step1

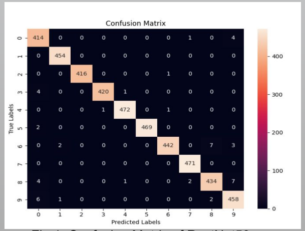
3. Train hybrid models with different combinations of transformer architectures and CNN-based models; Explore the possible improvement and deterioration and analyze the reason.

Results



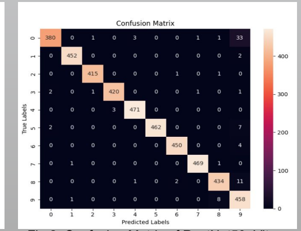
Model	Accuracy (test)
ResNet50	**98.89**
VGG16	**98.23**
Swin_t	97.45
Vit	95.97
ResNet50+swin_t	95.92
ResNet50+vit	**98.03**

Table 2. Accuracy of different model



	0	1	2	3	4	5	6	7	8	9
0	414	0	0	0	0	0	0	1	0	4
1	0	454	0	0	0	0	0	0	0	0
2	0	0	416	0	0	0	1	0	0	0
3	4	0	0	420	1	0	0	0	0	0
4	0	0	0	1	472	0	1	0	0	0
5	2	0	0	0	0	469	0	0	0	0
6	0	2	0	0	0	0	442	0	7	3
7	0	0	0	0	0	0	0	471	0	0
8	4	0	0	0	1	0	0	2	434	7
9	6	1	0	0	0	0	0	0	2	458

Fig 1. Confusion Matrix of RestNet50



	0	1	2	3	4	5	6	7	8	9
0	380	0	1	0	3	0	0	1	1	33
1	0	452	0	0	0	0	0	0	0	2
2	0	0	415	0	0	0	1	0	1	0
3	2	0	1	420	0	0	0	1	0	1
4	0	0	0	3	471	0	0	0	0	0
5	2	0	0	0	0	462	0	0	0	7
6	0	0	0	0	0	0	450	0	0	4
7	0	1	0	0	0	0	0	469	1	0
8	0	0	0	0	1	0	2	0	434	11
9	0	1	0	0	0	0	0	0	8	458

Fig 2. Confusion Matrix of RestNet50+Vit

Findings

1. ResNet50 and VGG16 performed the best, followed by ResNet50+vit. So transformer and hybrid models does not always perform better than pre-trained CNN. This suggests CNN models are good at capturing local features, which makes them more efficient for classification tasks;

2. Different performances of ResNet50 + Swin_t and ResNet50+vit suggest the way of combining CNN-based models and transformers in the transition layer will significantly influence the performance. A hybrid model with a transition that treats each feature map as an input image of the transformer helps improve the accuracy;

3. Both ResNet50 and ResNet50+vit are confused with class 0 (safe driving) and class 9 (talking to passengers). But ResNet50+vit has better recall than ResNet50. So hybrid model (ResNet50+vit) still has the potential to classify distracted driver images;

References

Hoang Ngan Le, T., Zheng, Y., Zhu, C., Luu, K., & Savvides, M. (2016). Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 46-53).

Hssayeni, M. D., Saxena, S., Ptucha, R., & Savakis, A. (2017). Distracted driver detection: Deep learning vs handcrafted features. Electronic Imaging, 29, 20-26.

Koay, H. V., Chuah, J. H., & Chow, C. O. (2021, December). Convolutional neural network or vision transformer? Benchmarking various machine learning models for distracted driver detection. In TENCON 2021-2021 IEEE Region 10 Conference (TENCON) (pp. 417-422). IEEE.

Koesdwiady, A., Bedawi, S. M., Ou, C., & Karray, F. (2017). End-to-end deep learning for driver distraction recognition. In Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14 (pp. 11-18). Springer International Publishing.

National Center for Statistics and Analysis. (2024, April). Distracted driving in 2022 (Research Note. Report No. DOT HS 813 559). National Highway Traffic Safety Administration.

Li, Y., Wang, L., Mi, W., Xu, H., Hu, J., & Li, H. (2022, May). Distracted driving detection by combining ViT and CNN. In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 908-913). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Yan, C., Coenen, F., & Zhang, B. (2016). Driving posture recognition by convolutional neural networks. IET Computer Vision, 10(2), 103-114.