Towards Efficient and Explainable Distracted Driver Detection

Thanh Tran[1] , **Dung D. Le**[1] ,
[1]College of Engineering and Computer Science, VinUniversity
{21thanh.tq, dung.ld}@vinuni.edu.vn

Abstract

Distracted driver detection systems help keep people focused on the road. Previous studies on this topic utilized convolutional neural networks (CNNs) and recurrent neural networks (RNNs) while recent approaches use attention mechanism and vision transformer to achieve state-of-the-art (SOTA) results. This paper shows that, by carefully tuning the parameters, training a ResNet50 in 8 minutes can yield comparable performance with SOTA for detecting driver distracted behaviors.

1 Introduction

Driver-assist technologies have become increasingly popular recently. These technologies make driving more pleasant, but they can contribute to distracted driving, one of the prime factor of traffic accidents. It is suggested in [McFarland, 2022] that autopilot systems need to make sure the driver's eyes are on the road and their hands are ready to grab the wheels at any time. As a result, companies are investing in driver-monitoring technologies to comply with the regulations.

In literature, deep learning and computer vision techniques were used in state-of-the-art classification models to accurately identify drivers' distractions. Recent research focused on developing specialized neural networks to better capture the posture of drivers. In this study, we opted for the opposite direction and investigated the performance of a classic architecture, namely ResNet [He *et al.*, 2015]. I will show that a modified ResNet with proper training procedure can be on-par with SOTA techniques such as vision transformer in detecting distracted drivers.

The decision to investigate ResNet thoroughly was inspired by [Bello *et al.*, 2021], in which the authors demonstrated that using better training procedure, ResNet can achieve similar accuracy as EfficientNet [Tan and Le, 2020], while being 1.7-2.7x faster.

In general, the inputs are images of drivers; a ResNet model then classifies whether the drivers are driving safely.

2 Dataset and Features

The dataset used in this study is the American University in Cairo (AUC) - Distracted Driver Dataset v2 [Abouelnaga *et al.*, 2018; Eraqi *et al.*, 2019]. The second version contains more images with more drivers, more precise labeling, and better sampling per class. Most importantly, training and testing are split based on drivers. Consequently, the performance on the second version is usually worse than on the first version because it requires better generalization.

Distracted Driver v2 consists of 44 drivers, mostly from Egypt. There are a total of 14,478 frames distributed over 10 classes: Safe Driving (2,986), Phone Right (1,256), Phone Left (1,320), Text Right (1,718), Text Left (1,124), Adjusting Radio (1,123), Drinking (1,076), Hair or Makeup (1,044), Reaching Behind (1,034), and Talking to Passenger (1,797).

Out of 44 drivers in our dataset, the training contains 38 drivers (12,555 samples), and the test data contains 6 drivers (1,923 samples). The sample of each class is shown in Fig. 1.

3 Related work

The first dataset with ten distracted driver actions was released in 2016 through a Kaggle competition by State-Farm [Kaggle, 2016], but this dataset is not available for research purposes. In 2017, Abouelnaga et al. [Abouelnaga *et al.*, 2018] released a new dataset called AUC Distracted Driver Dataset v1 with the same ten actions as the StateFarm dataset (and version 2 in 2019 [Eraqi *et al.*, 2019]).

For the rest of this paper, readers can assume that the results from related papers and this paper are for AUC-DDDv2 unless stated otherwise explicitly.

The authors of AUC-DDDv2 dataset proposed a genetically weighted ensemble of InceptionV3 to achieve a 95.98% classification accuracy on v1 and 90% on v2. [Mase *et al.*, 2020] used InceptionV3 and stacked Bidirectional LSTM to reach 92.7%. [Koay *et al.*, 2021] combined CNNs predictions of pose estimation images and original images to achieve 94.28%. To my knowledge, the current state-of-the-art model for AUC-DDDv2 is a highly complex and specialized neural network called AG-Net [Bera *et al.*, 2021], obtaining 96.65% classification accuracy. AG-Net uses Gaussian Mixture Model to identified regions that contains the most relevant information and applies attention mechanism on them.

Despite the promising results, all of these works use relatively complex pipelines, leading to slow training and inference time. This work aims to create an architecture that can balance between accuracy and inference time.

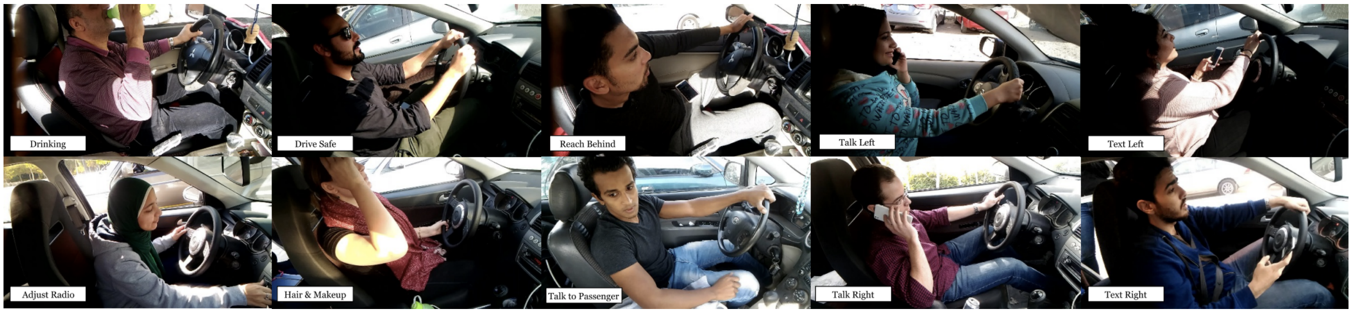

Figure 1: Ten Classes of Driver Postures. [Abouelnaga *et al.*, 2018]

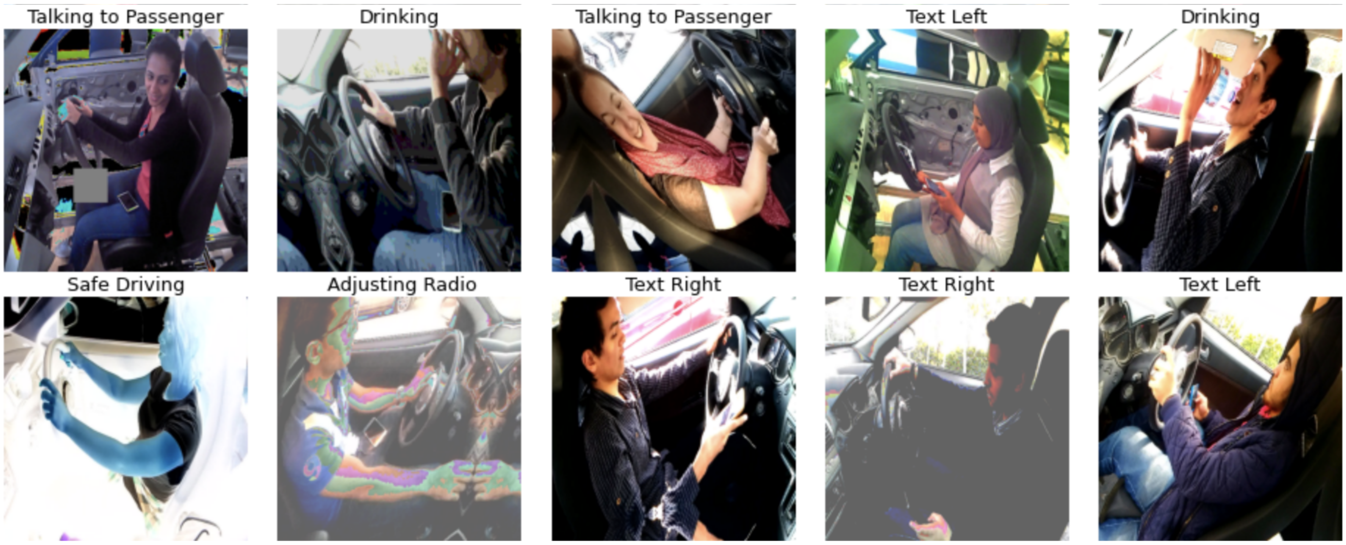

Figure 2: Samples augmented by RandAugment.

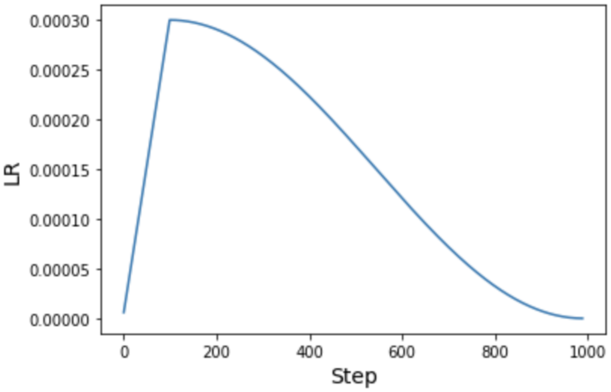

Figure 3: Cosine learning rate decay with linear warmup.

3.1 Transfer learning

Training a model from scratch requires careful search of hyperparameters and a sufficiently large dataset. I find out transfer learning consistently yields better performance and takes less training resources for AUC-DDDv2. The model used in this study is BiT-R50x1-M [Kolesnikov *et al.*, 2020], which was trained on the full ImageNet-21k dataset (14.2 million images and 21k classes). The only architectural change of BiT-R50x1-M from ResNet50 is the replacement of Batch Normalization with Group Normalization and Weight Standardization. In general, BiT-R50x1-M simplifies hyperparameter tuning.

4 Methods

4.1 Data Augmentation

Researches have shown data augmentation can significantly improve performance of deep learning models and are critical to the successes of SOTA models. Because of the small size of AUC-DDD, data augmentation techniques help the models generalize and act as a form of regularization. In this study, I empirically found the effectiveness of RandAugment[Cubuk *et al.*, 2019]. RandAugment can be represented by Figure 2. We used two parameters N = 2, M = 9 as suggested in [Cubuk *et al.*, 2019].

4.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

In this study, Grad-CAM was used to explain how BiT-R50x1-M classifies drivers distracted actions. Grad-

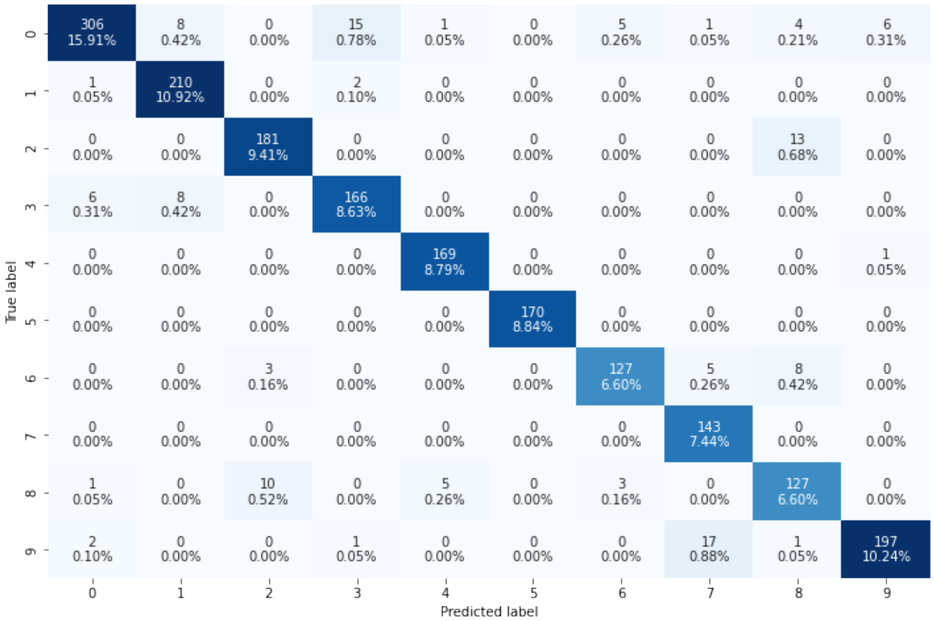

Figure 4: Confusion matrix generated by BiT-R50x1

CAM [Selvaraju *et al.*, 2019] makes CNNs models more transparent by visualizing the regions of input that are ”important” for predictions from these model. Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer of a CNN to produce a coarse localization map of the important regions in the image.

5 Experiments

5.1 Selection of Hyperparameters

All experiments were conducted on cloud TPUv3 using the Tensorflow 2 framework. We empirically chose parameters base on accuracy and loss in validation batch size of 128, image resolution 512x512. Cosine learning rate decay with warmup helps stabilize the models in the first few epochs. Adam optimizer gives great results with just 10 epochs (each epoch has 100 steps).

5.2 Metrics and Results

In this study, I follow a standard set of metrics. Accuracy is used to evaluate model performance and models are trained with the categorical cross-entropy loss.

BiT-R50x1-M takes only 8 minutes to train 10 epochs and can reach 95.58% accuracy on validation set of AUC-DDDv2. The previous best performance by a ResNet50 is 87.7% classification accuracy. This result shows how subtle changes in training procedure can dramatically improve performance and reduce training time. The results on AUC-DDDv2 valididation set is illustrated in 4.

5.3 Model explainability

GradCam is used to to discover how CNNs models decide. We manually investigated results from GradCAM and found that the models successfully learned to focus key features like eyes, face, phones, position of hands. Some example results from GradCam is shown in Figure 5.

5.4 Importance of components

In [Bello *et al.*, 2021], the authors suggested that training and scaling strategies may matter more than architectural changes. Therefore, in this section, we would verify this claim by investigating how each component affects the performance. Besides components discussed here, the choice of components such as the Adam optimizer and the Cosine learning rate decay depends on specific tasks and models. The goal of this section is to give a general view on effective training techniques.

Architectures. It is natural to try architectures other than ResNet given the recent success of vision transformers and EfficientNet. We replaced ResNet50 backbone with EfficientNetV2-M [Tan and Le, 2021] and ViT-S16 [Dosovitskiy *et al.*, 2021] and kept the same training procedure. Overall, we found that this training pipeline works well for CNNs without overfitting to ResNet and that architectural changes are not too impactful compared to other components.

Another CNNs model, EfficientNet achieved 96.05% accuracy. Even when EfficentNetV2-M is twice as large as ResNet50, we believe the 0.5% performance boast from ResNet50 is insignificant because:

- The size of AUC-DDDv2 validation set is too small.
- The average accuracy after training 10 times of EfficientNet and ResNet50 is both around 94.5%.

As for ViT-S16, we were only able to achieve 84.14% accuracy. However, it may be true that the chosen parameters are not suitable for a vision transformer. Vision transformers are especially suitable for the Distracted Driver Project

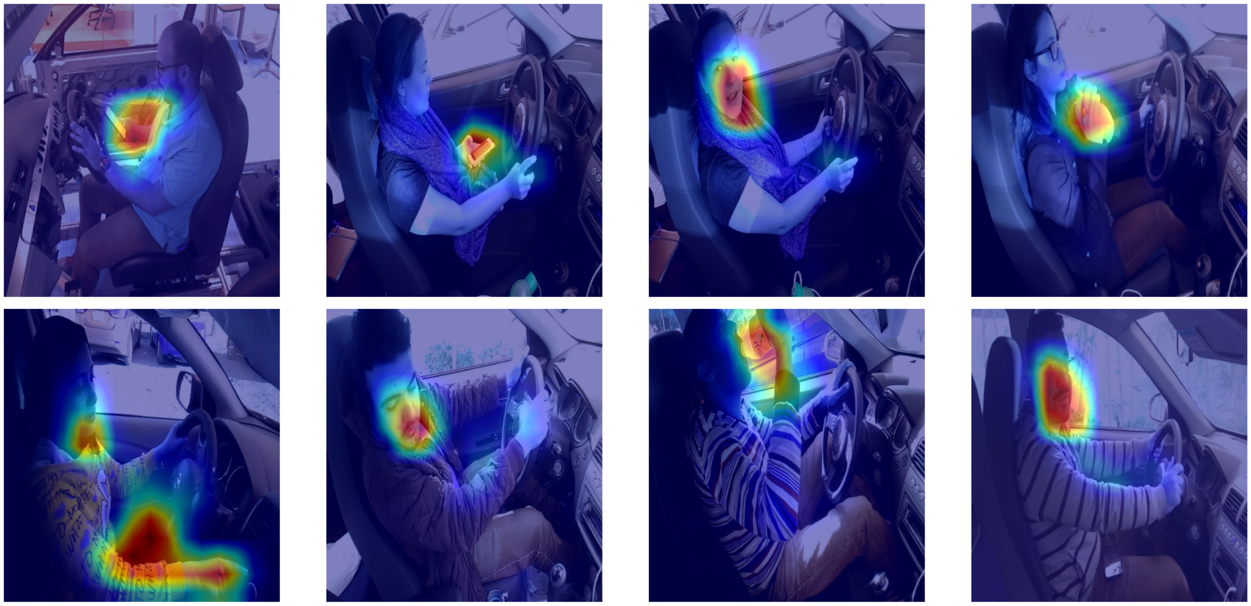

Figure 5: Superimposed images using heatmap from GradCam.

because they inherently give visual explanation through attention mechanism. Therefore, future work can investigate how to effectively fine tune ViT on small datasets.

RandAugment. Given the small size of dataset, data augmentation techniques are central to the training procedure. To test the effectiveness of RandAugment, we trained BiT-R50x1 with different data augmentations strategies:

- No augmentations at all: Performance varies greatly from 86.69% to 94.38%.
- A simple color distortion of randomizing images brightness, hue, saturation, contrast: Performance varies from 89.86% to 94.18%.
- RandAugment: Performance varies from 92.67% to 95.58%.

Training time using RandAugment is even 30% less than that using simple augmentation pipeline. That is because RandAugment actually use less operations (N = 2), while providing more variations.

RandAugment acts as a regularizes for overfitting problems. With no augmentations, the train set accuracy reaches 99% after only 2 epochs, while validation set accuracy is less than 80%. With RandAugment, the gap between two sets is smaller as the train set accuracy reaches 99% after approximately 8 epochs.

We also tested EfficientNetV2-M without data augmentation, and the accuracy achieved was only 87.89%. This is understandable, because RandAugment was identified in previous studies as a key component of EfficientNet.

These experiments imply that proper data augmentation can improve the performance and robustness of models.

Image size. The size of images used in this study is 512x512, which is relatively large in computer vision. Most of related works on distraction detection, including SOTA [Bera *et al.*, 2021], used an image size of 224x224.

When using image input size of 224x224, BiT-M50x1 and EfficientNet classification accuracy dropped by 3% and 2% respectively. We believe there are two primary reasons:

- AUC-DDDv2 dataset demands models to look at small but important regions such as eyes, face, hands, steering wheel. However, CNNs models used in this study were pretrained on Imagenet, in which information is sparsely concentrated all over the image. This difference makes transfer learning from Imagenet difficult.
- CNNs are not good at capturing information from small and fine-grained regions due to the nature of convolutional layers.

As for ViT, we found that the changes in image size do not affect performance. ViT is designed to mitigate the aforementioned drawback of CNNs. Through the attention mechanism, ViT can focus solely on important and fine-grained regions of images.

6 Future Work

This study suggests that researchers could use simple CNNs such as ResNet50 and a proper training procedure as baseline before trying more complicated architectures. In this study, we also give a relatively good pipeline for training CNNs future extensions.

For future work, vision transformers can be investigated more thoroughly. ViT not only achieves state-of-the-art vision tasks, but also inherently gives visual explanation, making ViT ideal for this Distracted Driver Project. However, training vision transformer requires careful parameter search and a large dataset. If we can successfully transfer learning ViT to AUC-DDDv2, that training pipelines can be applied to other small datasets.

References

[Abouelnaga *et al.*, 2018] Yehya Abouelnaga, Hesham M. Eraqi, and Mohamed N. Moustafa. Real-time distracted driver posture classification, 2018.

[Bello *et al.*, 2021] Irwan Bello, William Fedus, Xianzhi Du, Ekin D. Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies, 2021.

[Bera *et al.*, 2021] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. Attend and guide (ag-net): A keypoints-driven attention-based deep network for image recognition. *IEEE Transactions on Image Processing*, 30:3691–3704, 2021.

[Cubuk *et al.*, 2019] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space, 2019.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[Eraqi *et al.*, 2019] Hesham M. Eraqi, Yehya Abouelnaga, Mohamed H. Saad, and Mohamed N. Moustafa. Driver distraction identification with an ensemble of convolutional neural networks, 2019.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[Kaggle, 2016] Kaggle. State farm distracted driver detection, 2016.

[Koay *et al.*, 2021] Hong Vin Koay, Joon Huang Chuah, Chee Onn Chow, Yang-Lang Chang, and Bhuvendhraa Rudrusamy. Optimally-weighted image-pose approach (owipa) for distracted driver detection and classification. *Sensors*, 21:4837, 07 2021.

[Kolesnikov *et al.*, 2020] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2020.

[Mase *et al.*, 2020] Jimiama Mafeni Mase, Peter Chapman, Grazziela P Figueredo, and Mercedes Torres Torres. A hybrid deep learning approach for driver distraction detection. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1–6. IEEE, 2020.

[McFarland, 2022] Matt McFarland. Driver monitoring gets spotlight amid safety questions, 2022.

[Selvaraju *et al.*, 2019] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 10 2019.

[Tan and Le, 2020] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[Tan and Le, 2021] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.