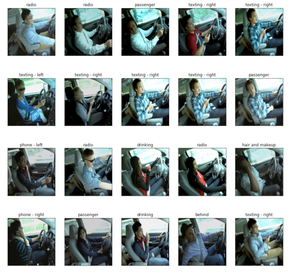
Eyes on the Road: Vision Transformers(ViT) for in-Cabin Driver Detection

Justin Lee, Aaron Li
justinl8@stanford.edu, aaronli9@stanford.edu

Motivation

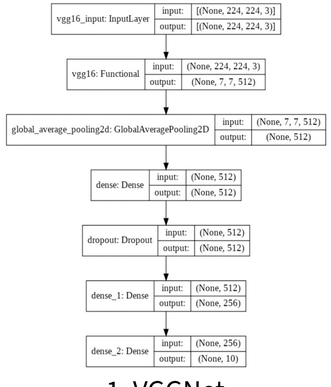
With the rollout of vehicles with Level-3 self-driving features, there has been a major push on not only R&D on advancing a vehicle's capability of monitoring external conditions but also a unique emphasis on in-cabin driver monitoring. As it is required by law for L3 vehicles to have a mechanism to monitor the driver and hand back control to the driver. There needs to be technology inside the cockpit that not only ensures a driver is available but also his or her ability to take control of the vehicle at any given time. With that in mind, we introduce EyesOnTheRoad: a vision transformer (ViT) that performs multi-class distracted driver detection. We also explore explainability in AI by exploring methods that demystify deep learning models.

Dataset



The dataset consists of 22422 images provided by State Farm. The image consists of a driver behind the wheel with one of ten classes: driving safely, texting right, texting left, talking on phone right, talking on phone left, operating radio, drinking, reaching behind, doing makeup, and talking to a passenger. The images were taken with 26 unique subjects in different cars, each image is 640 x 480 pixels. We did a 5-to-1 training-test split, follow by an 8-to-1 training-validation split, resulting in 16144 train, 2243 validation, and 4037 test images.

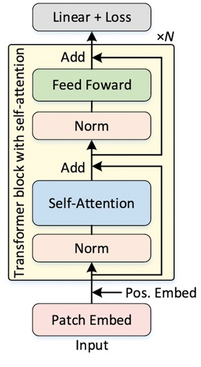
Methods



1. VGGNet



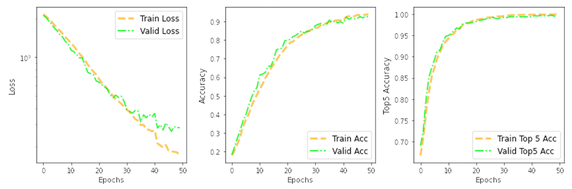
2a. ViT Patch Generation



2b. ViT Architecture

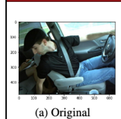
Result

Classifier	Training Acc.	Testing Acc.	Loss
MobileNet	72.45	-	8.23
VGG 16 Net	87.94	-	5.012
Vision Transformer	93.45	92.01	0.28



Vision transformers performed significantly better than VGGNet, with higher accuracy and lower loss. These results make sense due to some properties of the vision transformers algorithm that are advantageous over CNNs. CNNs, like VGGNet, represent images as a grid of pixels and create new feature maps by sliding a filter across the image grid, taking the dot product at each location. Consider two small batches of pixels that are not very close to each other in the image, but together, are crucial in classifying the image. In order for any sort of relationship to be encoded between these two areas, there would have to be enough convolutional layers to reduce feature maps such that these two areas eventually end up being part of the same dot product computation. Sometimes, as VGGNet is a deep neural network, but not limitlessly deep due to computational constraints, some important relationships between pixels can be missed. In vision transformers, pixels are represented as image patches or tokens. The relationship between each pair of these tokens is calculated using a self-attention method, indicating which pairs are more important than others. In the self-driving car dataset, which requires high-level feature selection like seeing that someone is holding a phone or a drink, having these relationships encoded is very beneficial for classification. However, because this self-attention mechanism needs to maintain weighted connections between every pair of image patches (tokens), vision transformers' training time is long. Even after normalizing our images to be 64 x 64, our vision transformers model still took 8 hours to train 50 epochs compared to the 2 hours needed for our VGGNet.

Dicussion



(a) Original

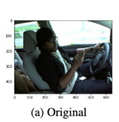


(b) Saliency Map



(c) Integrated Grad

Figure 7



(a) Original



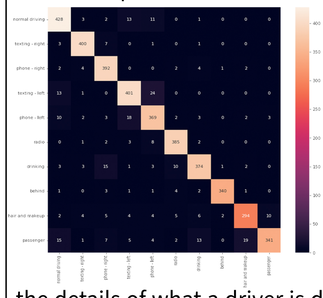
(b) Saliency Map



(c) Integrated Grad

Figure 8

Attribution methods allow us to quantify the contribution of particular variables in a model towards a particular behavior of the model. We have decided to compare between saliency map and integrated gradients. Visualizing the results on the left, it is apparent that Integrated Gradients in this case are better able to focus on the pixels corresponding class of concern. In figure 8, the Integrated Gradient can highlight the body of the driver turning backward while in Fig. 9 it was able to highlight the driver's hand and phone, explaining how the model makes its decision for classification based on the highlighted features. Integrated gradients are superior because they can address some issues with saliency maps where they reflect the importance of edges more accurately than vanilla saliency maps because they capture the gradient information in a relative global manner as shown in the comparison above.



We have also created a confusion matrix to visualize the performance of our Vision Transformer(ViT). Our model mostly predicted well across the board; however, there are a couple of classes that drew our attention. Since the images are all taken from the center of the front dash, we think this creates more mistakes because, on the left side, the details of what a driver is doing with his left hand are obscured by other parts of the image, such as the driver's head. This hypothesis is supported by the fact that the model mainly incorrectly predicted safe driving with the phone left and texting left classes, suggesting that any phone actions on the left side could sometimes be obscured.

Future Work

Create a semi-supervised learning method that allows us to use the remaining 70,000 unlabeled test images from State Farm to improve data robustness. Improve upon our data visualization and explainability methods, in order to better address our model performing significantly stronger in certain classes. Implement more sophisticated ensemble models to improve performance, and see whether CNN-based methods can compare with Vision Transformers, as VGG doesn't represent the most modern state-of-the-art CNN.