Real-Time Classification and Detection of Distracted Driver using Deep Learning

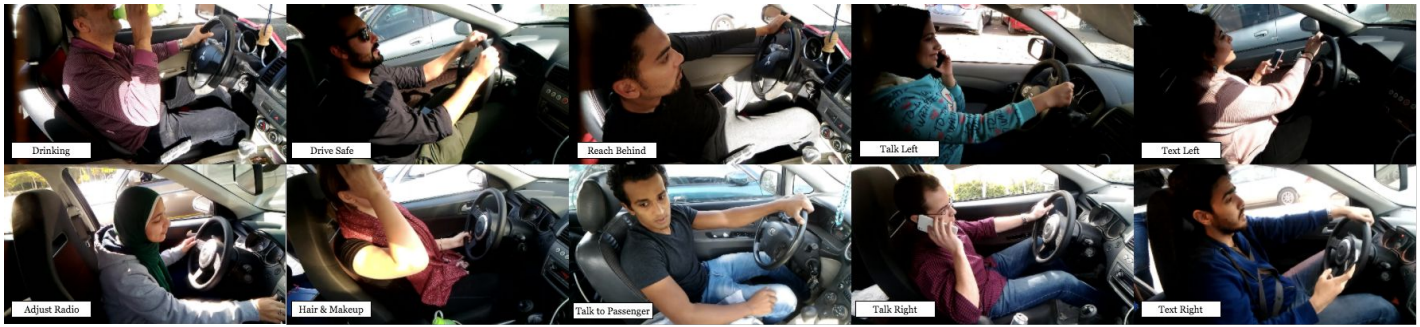
COMPUTER VISION PROJECT PRESENTATION

Group Members : 1) Suyashi SInghal
2) Harshita Gupta
3) Shreya Tomar

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Problem Statement

- Any activity that takes away the driver's attention from the road is distracted driving. The Centre for Disease Control and Prevention(CDC) provided a more precise definition of distracted driving by classifying it into three types
  - visual (not being visually attentive on the road ahead)
  - cognitive (mentally distracted by eye gaze is on the road)
  - manual (physically distracted like driver's hands off the wheel)
- Aim is to develop a deep learning-based model using computer vision that can detect the distracted manual activities of the driver.



Problem Statement

- We have a dataset of images captured from a camera mounted on the driver's dashboard. We form this dataset by combining two datasets to obtain a more extensive training set and thereby aim to avoid overfitting.

- In our dataset, we will perform basic preprocessing of images like creating validation sets, resizing and normalization of the images. The major preprocessing will go into merging the two datasets into one.

- In a real-world scenario, the camera will capture a video of the driver while driving. We can concurrently segment the obtained video into various frames in real-time that can be used as an input to the model to detect the distracted activity of the driver. This model can be ingrained with each driving vehicle system to alert the driver in case of any distractions and prevent accidents.

Dataset Description

- Two datasets are combined to form a more extensive dataset and to avoid overfitting.
  - The first dataset is the State Farm Distracted Driver Dataset, available on the following website. In this data, the training set contains 22.4K images that have been labelled into ten classes.
  - The American University in Cairo Distracted Driver Dataset. It contains images in the form of frames from videos captured while driving.
- The combined dataset can help minimize the problem of overfitting seen in many studies on this topic. This also forms one of the novel contributions of our project. Our algorithm will detect the distracted activity of the driver and provide the most appropriate label as the output.

Class	Description
0	Normal Driving
1	Texting while driving using right hand
2	Talking to somebody on the phone using right hand
3	Texting while driving using left hand
4	Talking to somebody on the phone using left hand
5	Reaching the dashboard to operate the radio
6	Drinking or eating
7	Reaching behind
8	Fixing hair and makeup
9	Talking to passenger

Survey

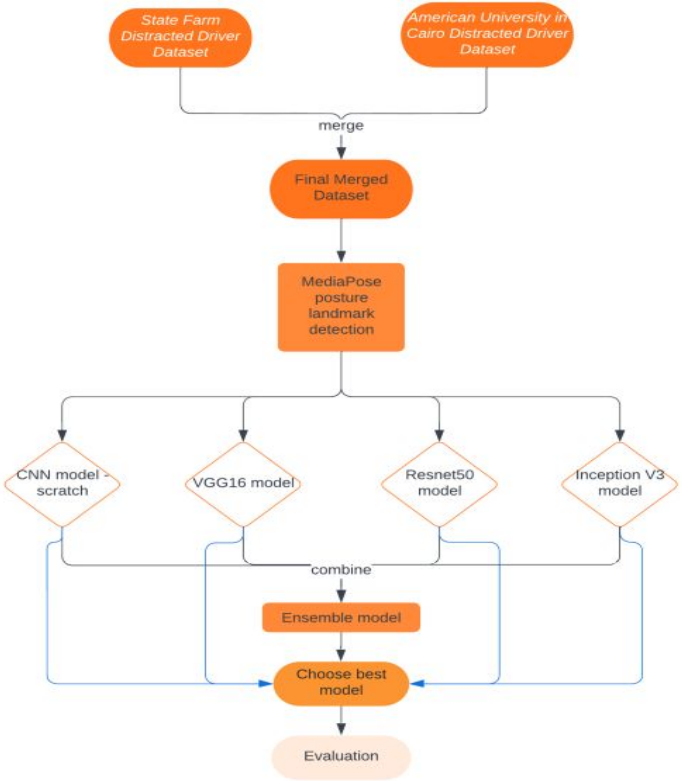
- **Paper 1:**
  https://www.researchgate.net/profile/Pramila-Chawan/publication/337275106_Distracted_Driver_Detection_and_Classification/links/5dce5afc299bf1b74b426c58/Distracted-Driver-Detection-and-Classification.pdf
  The best ensemble was created after averaging the probabilities generated by the models, VGG-16, VGG-19, and InceptionV3. The final log loss value obtained was 0.795. CPU resources provided by Google Cloud Platform were used.
- **Paper 2:**
  https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w14/Baheti_Detection_of_Distracted_CVPR_2018_paper.pdf
  A CNN-based system was implemented to detect a distracted driver with the specific reason for that distraction. They modified the VGG-16 architecture by applying several regularization techniques to prevent the overfitting of the training data. The accuracy of the model was 96.31%.

Survey

- **Paper 3:**
  https://sci-hub.hkvisa.net/10.1016/j.asoc.2020.106657
  A publicly available dataset was used. Methods such as transfer learning and fine tuning were used to build a highly accurate vision based detection model. The model obtained on the public dataset was applied to the vision part of the multimodal dataset. The predictions of vision and sensor based modalities with two different fusion approaches to map the inputs to one of the ten actions.
- **Paper 4:**
  https://link.springer.com/article/10.1007/s11760-019-01589-z
  A method is proposed to combine the three deep learning models, the residual network, the Inception module and the hierarchical recurrent neural network to improve the performance of detecting a distracted driver's behavior. The approach was then evaluated with the help of two types of datasets, namely, State Farm Distracted driver dataset and AUC distracted driver dataset.

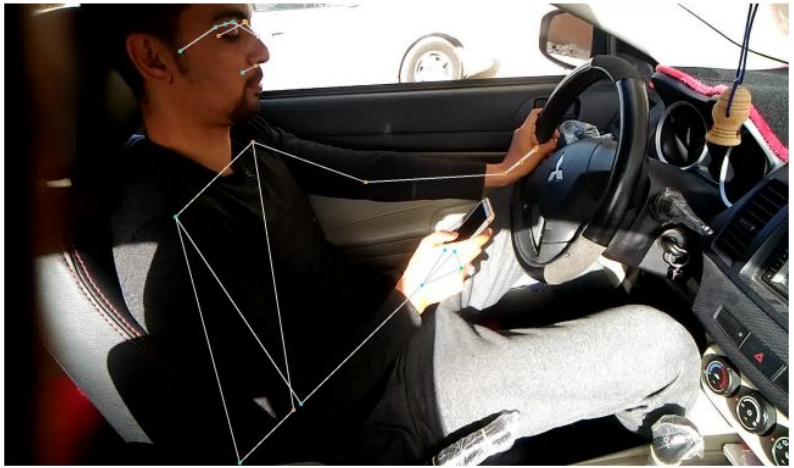
Methodology

- The pipeline of our project is depicted in the given figure.



Methodology

- Our dataset has performed basic preprocessing of images like creating validation sets, resizing and normalization of the images.
- The dataset contains images with the side view of the drivers only. In order to make the model more generalizable, MediaPipe Pose Detection model has been used for predicting and drawing the pose landmarks on the image of the drivers.
- Good performance can be achieved even when the driver video is taken from a different angle than those in the dataset.



Methodology

- Training CNN models with a few layers and an extensive training set can overfit. Increasing the number of layers and training the models from scratch is computationally expensive. Hence we employ the pre-trained models since they provide us with the best of both worlds in this case.
- Trained a CNN model from scratch and used transfer learning to fine tuning pre-trained models.
- Finally create an ensemble of the fine-tuned models to get our final results.
- Models used for implementing our project are as follows:
  - **CNN from scratch** - A neural network specifically used for images. The input layer comprises the individual pixel values of the resized photos from our combined dataset, while the output layer gives the probability of each class.
  - **VGG16** - Has 13 convolutional layers and two fully connected layers at the top, followed by the softmax output. The most distinctive feature of VGG16 is that, rather than having a massive number of hyper-parameters, they concentrated on having 3x3 filter convolution layers with a stride one and always utilized the same padding and max pool layer of 2x2 filter.

Methodology

- ➢ **Resnet 50** - The ResNet-50 model is a 50-layer convolutional neural network (CNN). ResNets tackle the problem of accuracy saturation and degradation after a point due to vanishing and exploding gradients by introducing the concept of skip connections.
- ➢ **Inception V3** - The symmetric and asymmetric construction components utilized in Inception v3 include convolutions, average pooling, max pooling, concatenations, dropouts, and completely connected layers. The loss is calculated using Softmax.

❖ **Ensemble Model:** We ensemble the above models and perform a weighted average of the results in order to make a more accurate and generalizable classifier. Each model is given an equal weightage while combining the results. The final model combines a few deep learning models with some of the most complex architectures to perform precise categorization.

Experimental Results

- Accuracy, precision, recall and F1-score were used as the evaluation metrics to test our models :
  - **Accuracy** - measures the overall efficiency of a classifier.
  - **Precision** - ratio of true positives to the total of the true positives and false positives.
  - **Recall** - ability of a classifier to categorize positively labeled data.
  - **F1 score** - harmonic mean between precision and recall, gives good tradeoff between them.
- The log-loss is used when the performance of a classification model with a prediction input of a probability value between 0 and 1 is to be measured.
- Our model aims to minimise this loss. We will use a set of predicted probabilities for each image to compute the loss and optimise the model's performance.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

Figure 2: Evaluation Metric (Log Loss)

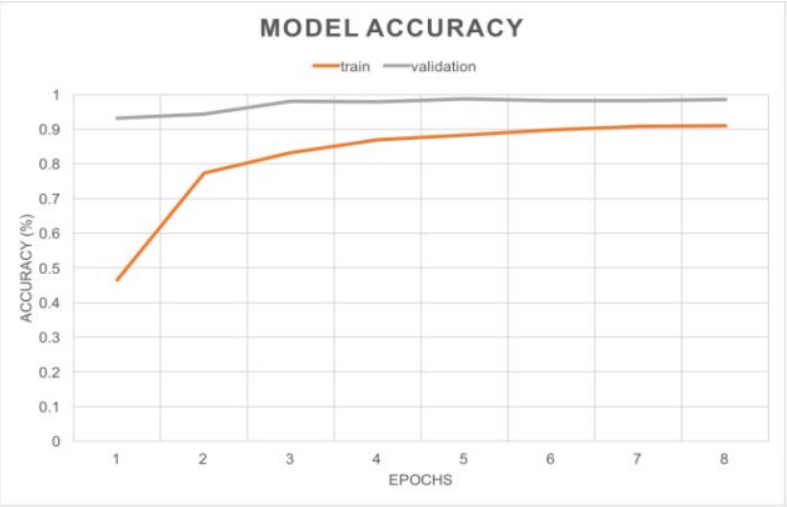
Experimental Results

- We observe that ResNet gives the least validation accuracy and poor performance. Hence we exclude it from our ensemble model and only use the remaining three models for ensemble.
-

Models	Training loss	Training accu-racy	Validation loss	Validation accu-racy
CNN	0.0339	0.9903	0.0337	0.9935
Inception	0.1038	0.9814	0.0841	0.9801
Resnet-50	4.6218	0.8750	3.7207	0.4913
VGG	0.1710	0.9548	0.3300	0.9137

Experimental Results

- Both the training and validation accuracies increase with the number of epochs until they reach a particular threshold, indicating that no more improvements can be made and training must be halted.
- Similarly, when the model learns, both validation and training losses decrease, eventually disappearing when the model begins to overfit.



Ensemble model's accuracy v epochs while training

Ensemble model's categorical cross entropy v epochs while training

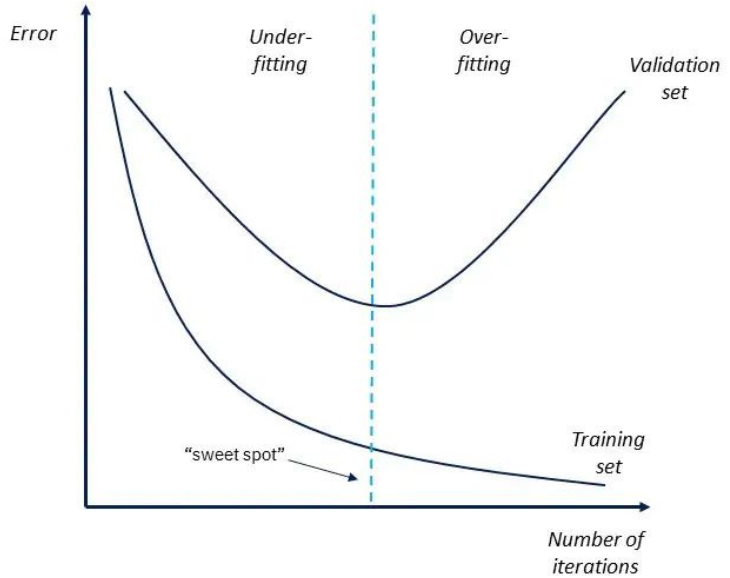
Experimental Results

- We can conclude that the training accuracy is significantly greater than validation accuracy. This leads to overfitting. Dropout layers are used to prevent overfitting. We see this behaviour because the model has a small number of layers for a big amount of training data. It causes the model to not fit accurately on the training set.
- The Precision, Recall and F1 scores are comparable to the accuracies of the model. For instance for the CNN scratch model, the F1 score was 99.33, precision and recall were approximately 99.3. Similar inferences can be drawn for them as well.

Demo Video

Challenges

- **Overfitting:-** Since these neural networks have very complex architectures with many parameters, more often than not, they overfit.
- **Model generalizability:-** Most researchers in this domain have mentioned this as a drawback of their model. This reduces the generalizability of the model and hence diminishes its performance.

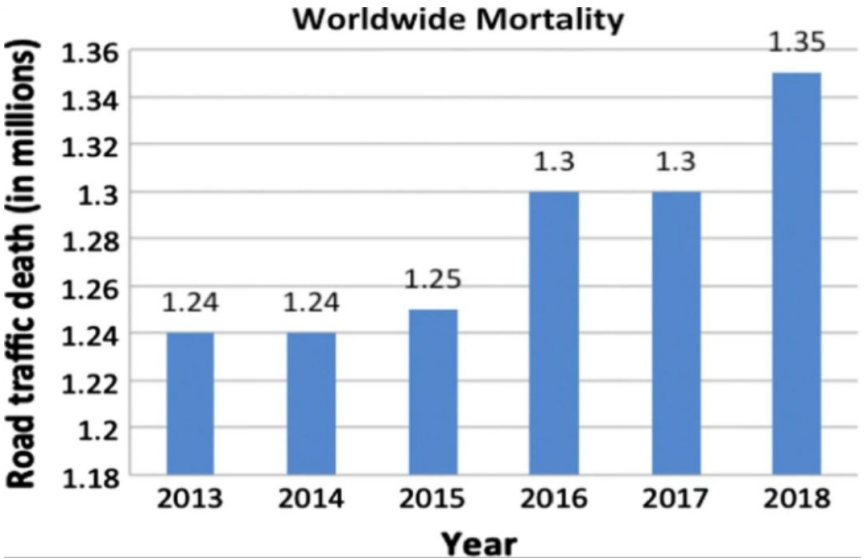


Challenges

- **Computational power:-** Training large neural networks from scratch requires a lot of computational power that might not be available readily.
- **Data leak:-** There is an added problem of data leakage encountered while using the dataset. With slight changes, multiple images of the same person within the training dataset leaked into the validation set. This can lead to the model being trained on the same information it would predict.
- **Blurring:-** Generation of blurry frames due to the vehicle's movement and problems of using cameras in two-wheelers.

Motivation

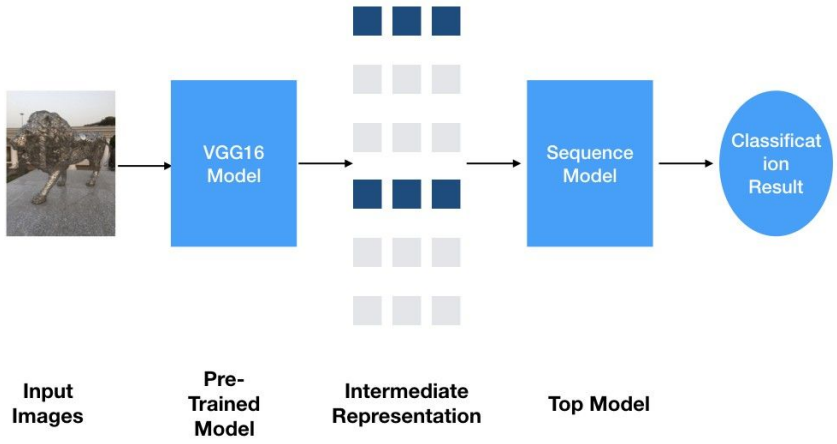
- ➔ The World Health Organization (WHO) published the 2015 Global Status Report, which revealed that about 1.25 million people die each year in car accidents around the world.
- ➔ Our idea can detect the driver's action using CNN model and can predict whether the driver is distracted in some activities or not.
- ➔ The government can use it in the form of surveillance analysis of the data.
- ➔ As a result, major accidents can be avoided, and our country can be a safer place.



Contributions

- Performing high-performing CNN training on a large number of images. Higher-level features are less but are suitable for fine-tuning.

- The accuracy of the distracted driver classification is improved by using transfer learning and combining various deep learning models to develop an ensemble of convolutional neural networks.



Contributions

Figure 1: Example Images of each class

- The data of the distracted drivers can be useful for the government to take control over reckless driving.
- Real-time alerting systems for the driver can be done.
- Images are classified as shown from 0-9 categories.
- Good accuracy is maintained using a Convolutional Neural Network-based approach.

Novelty

- **Combining two datasets** - This merged dataset would be used to improve the model's generalizability and decrease overfitting.
- **Real time data** - It would allow a video or image to be used as an input in real-time, and if the driver is distracted for a significant amount of time, they can be warned utilizing some signal.
- Hence, our model would work not only for the dataset being used here but can also be integrated with a camera that captures real-time video and photographs. We would employ some metrics to suggest whether to alert the driver or not by capturing multiple frames within a few minutes of driving and generating the signal accordingly.
- The police can use the data from such a model to impose penalties on people not following driving rules and regulations and prevent accidents in the long run.

THANK YOU