Winning the ICCV 2019 Learning to Drive Challenge 🏆🚘

Michael Diodato, Yu Li, Manik Goyal, Iddo Drori

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Models and Code: github.com/mdiodato/cu-iccv-2019-learning-to-drive-winners

Data and Model

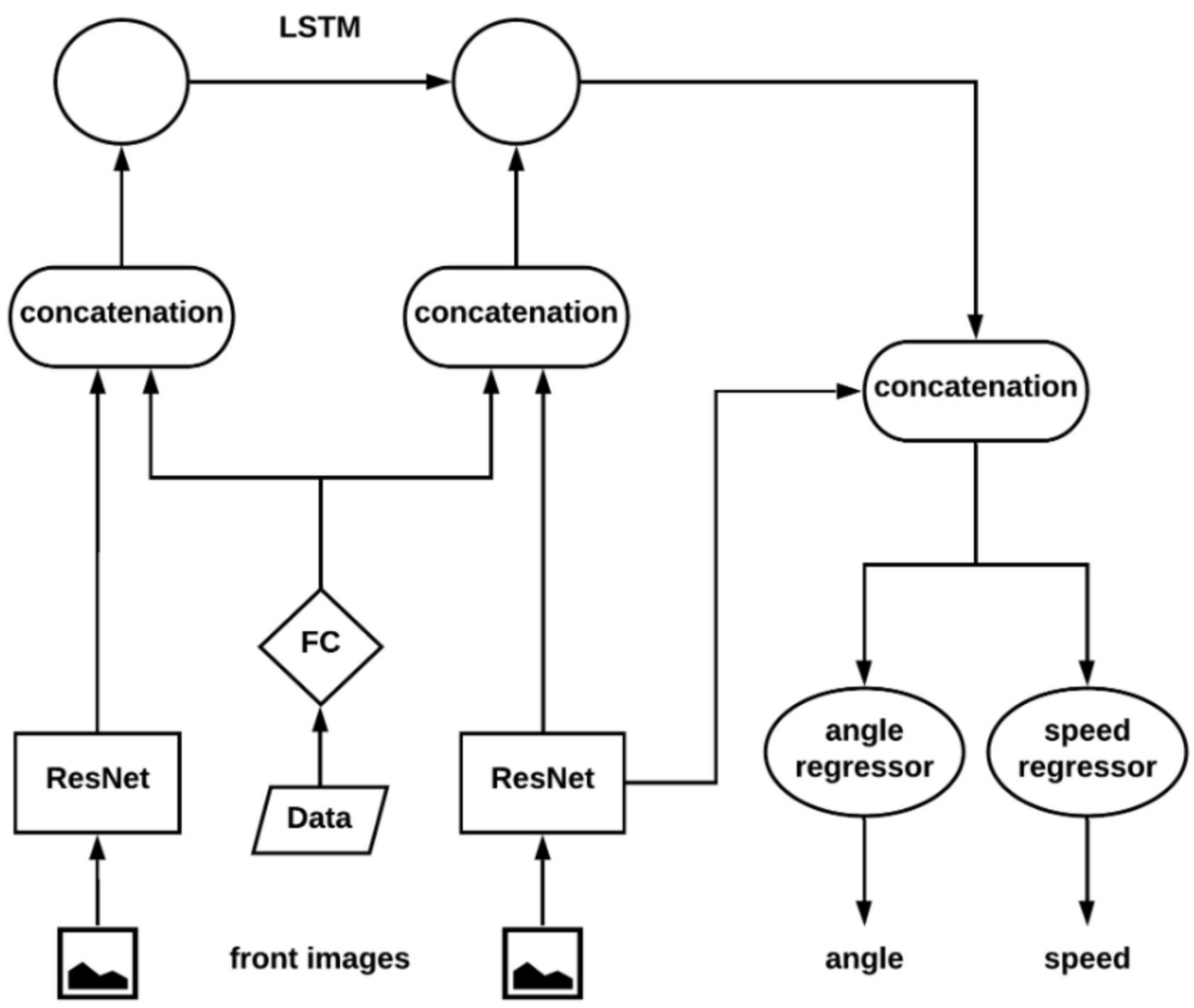
Predicting vehicle trajectories, angle and speed, is important for safe and comfortable driving. This work focuses on fusing inputs from camera sensors and visual map data which lead to significant improvement in performance and plays a key role in winning the challenge. We use pre-trained CNN's for processing image frames, a neural network for fusing the image representation with visual map data, and train a sequence model for time series prediction. We demonstrate the best performing MSE angle and best performance overall, to win the ICCV 2019 Learning to Drive challenge.

Example training images

Example test images

hereMmLatitude	hereSegmentExitHeading
hereMmLongitude	hereSegmentEntryHeading
hereSpeedLimit	hereCurvature
hereSpeedLimit_2	hereCurrentHeading
hereFreeFlowSpeed	here1mHeading
hereSignal	here5mHeading
hereYield	here10mHeading
herePedestrian	here20mHeading
hereIntersection	here50mHeading
hereMmIntersection	hereTurnNumber

Semantic map features used



Deep neural network architecture: pre-trained ResNet and fully connected network that feeds into an LSTM model. This and the output of the ResNet model on the current image are fed into an angle and speed regressor.

Results

Model	MSE Angle	MSE Speed
Overall	831.5	4.5
Zone30	2,981.1	0.3
Zone50	1,353.4	6.0
Zone80	168.6	4.1
Right	1,928.4	1.3
Straight	821.7	4.6
Left	833.6	2.6
Pedestrian	3,722.1	4.1
Traffic Light	329.2	5.3
Yield	1,818.9	2.8

Performance across various zones: the ensemble method did worst in the pedestrian sections and best in Zone80. Presumably, pedestrian sections would be hardest to train due to the unpredictability of cities and people. Zone80 sections are likely straighter and require less change in speed and steering angle and would probably be easier to train. Likewise, Right and Left sections, which would require learning a turn, would be difficult, but Straight segments are easier to train and performed better.

Model	CNN	Dimensions	Semantic Map	Batch	Epochs	MSE Angle	MSE Speed	Combined
1	ResNet34	320x180	No	64	2	1111.437	5.866	1117.303
2	ResNet152	320x180	No	8	1	1211.434	5.461	1216.895
3	ResNet34	160x90	20 Features	8	1	897.489	6.664	904.153
					2	**883.501**	**6.403**	**889.904**
					3	931.689	6.445	938.134
					4	970.96	6.714	977.674
					5	956.262	6.576	962.838
4	ResNet34	320x180	20 Features	32	1	995.42	5.316	1000.736
					2	946.516	5.337	951.853
					3	989.013	5.519	994.532
					4	965.791	5.706	971.497
					5	987.572	5.846	993.418
5	ResNet34	320x180	47 Features	64	1	900.407	5.571	905.978
Ensemble						**831.504**	**4.543**	**836.047**

Parameters and results for 5 different models and the result of the ensemble. The best overall result is an ensemble. Individually, we note that inclusion of the semantic map reduces the MSE by about 300, comparing models 1 and 4. Models 3 and 4 likely suffer from overfitting as evidence by the increasing test MSE, although the training loss decreased. Best standalone and overall models are in bold.

Acknowledgements: We would like to thank the Columbia University students of the Fall 2019 Deep Learning class for their participation in the challenge. Specifically, we would like to thank Xiren Zhou, Fei Zheng, Xiaoxi Zhao, Yiyang Zeng, Albert Song, Kevin Wong, and Jiali Sun for their participation.