Attention-based Convolutional Architectures Improve Transfer Learning for Distracted Driving Classification

Casey Hirschmann
chirschmann3@gatech.edu
Georgia Institute of Technology
Atl, GA

Abstract

According to the CDC, distracted driving claims 3,000 deaths each year and over 20% of injuries in car accidents are attributed to it [2], thus a deployable model that can be quickly trained to categorize distracted driving is useful for insurance companies and authorities to use for prevention. Transfer learning was explored using a classic CNN architecture, DenseNet161, and improved upon using a modern, attention based CNN architecture, ConvNeXt. Along with data augmentation, hyperparameter tuning, and optimizer exploration, an improvement from 45.23% to 68.62% accuracy was observed. Average class accuracy improvement of 67% was also seen.

1. Introduction/Background/Motivation

To help authorities and insurance companies decrease distracted driving (DD), it is important to have models that can be trained quickly based on various camera locations and car interior set-ups. To do this, transfer learning where only the classification layer is trained is important. Thus, the experimental objectives were to increase the current best-in-class, transfer learning testing accuracy using more modern architectures, data augmentation, regularization, and optimizers.

Success has been seen using transfer learning on various CNN based architectures, however, around 57% is the highest accuracy obtained when all weights are frozen except for the classifier layer. Although 88% accuracy can be obtained with full end-to-end training, the required training time is too high for high model adoption [7].

Other approaches have achieved up to 90% accuracy, but they require hand manipulation of the images which is impractical for general usage [3].

The issue is in overfit due to the heavy similarity of the images. However, new CNN architectures, namely ConvNeXt, have been developed that are more likely to perform well on the DD data due to their incorporation of a vision transformer-style attention mechanism. This architecture still retains the efficiency and interpretability of a CNN but also incorporates ViT aspects as follows:

- Multi-scale fusion as a type of attention gating which allows the model to attend to different scales of features
- Cross-scale connections that enable information flow between different scales of feature maps
- Channel attention which selectively attends to different channels of feature maps using SE (squeeze-and excitation) blocks that weight channels based on their task contribution [5]

This makes ConvNeXt preferable to ViT architectures since ViTs have a much higher number of hidden parameters than traditional CNN based architectures, and thus typically take longer to train, even with transfer learning [4].

Finding a model that can achieve high accuracy without manual data manipulation or long training times can speed the adoption of algorithms that can help to decrease the prevalence of DD and save lives.

The DD data set provided by State Farm on Kaggle was used. The data consists of fully labeled images from 10 categories ranging from safe driving to texting - left, texting - right, hair and makeup, drinking, etc. all from the same perspective in the car.

The data set contains 22,424 total instances of 81 different people performing each of the categories. Categories are well-balanced, with $\sim$ 2300 instances of each class. Because of this, the train and validation sets were split based on subject number to ensure all categories were represented in each data set and $\sim$70% of the data was used in training as recommended by [7].

2. Approach

All models were trained with all weights frozen except for the replaced final, FC classification layer and CE loss was used. Data was prepped with normalizing by batch.

Initially, the DenseNet161 architecture was grid-search tuned and utilized as the CNN base case since it was the highest performer in [7]. Five epochs were used for training since [7] showed that minimal gains happened in later epochs. Learning curves further confirmed this. Drop out and L2 regularization were also explored to help with overfit.

[1] notes that "a pre-trained model may be very good at identifying a door but not whether a door is closed or open". As seen in Figure 1, the DD images fall into this as they are extremely similar with only minor differences in hand and face location to delineate the categories, thus CNNs tend to heavily overfit with transfer learning alone. ConvNeXt was trained and grid-search tuned as a new approach to overcome this overfit tendency as a way to better distinguish the nuances of the photos based in its added attention mechanism.

(a) Safe Driving (b) Hair and Makeup

Figure 1: Example images from DD dataset.

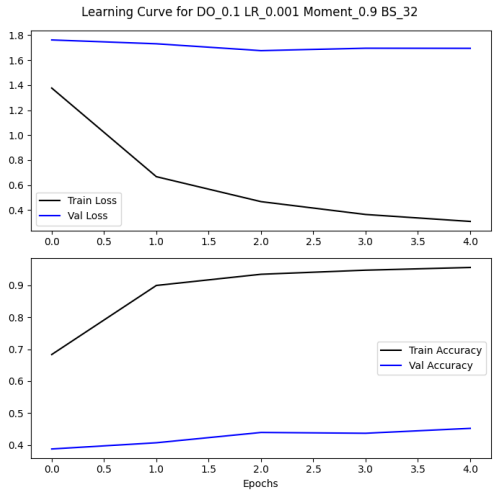
Overfit was still anticipated with the new architecture, so the need for experimentation with regularization and data augmentation was anticipated. TrivialAugment was planned due to its simplicity (fewer hyperparameters) and outperformance of more complex augmentation techniques [6]. TrivialAugment applies random augmentation to random images based on the selected "intensity" hyperparameter which determines number and intensity of augmentations. Regularization would be trialed based on the results from the DenseNet161 experiments.

The final model was trained to 15 epochs to determine if any marginal gains could be made from extended training.

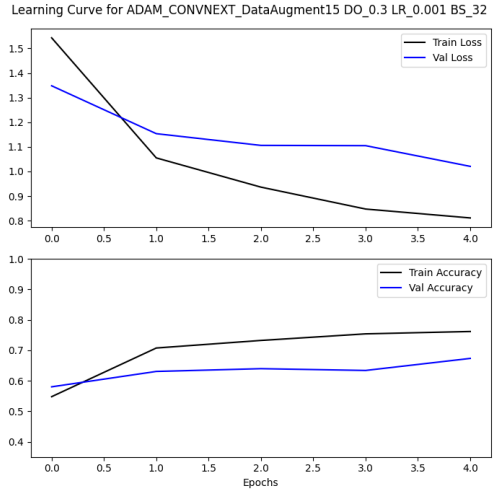
3. Experiments and Results

Success was determined base on an increase in (1) overall accuracy and (2) average by-class accuracy between the DenseNet161 model and the final ConvNeXt model. Loss was also tracked to yield insight into learning but wasn't used as a measure of success. Final results and hyperparameters used can be seen in Table 1.

Learning rate, momentum, and batch size were tuned to 0.001, 0.9, and 32 respectively with DenseNet161 which resulted in overall accuracy of 0.452265 and average class accuracy of 0.45. Important findings were as follows:



(a) DenseNet161 Tuned Model



(b) ConvNeXt Tuned Model

Figure 2: DenseNet161 learning curves show minimal learning, large variance and overfit vs the ConvNeXt curves.

- **Learning rate**: although higher learning rate (lr) achieved slightly higher accuracy, no learning was seen in the loss curve (training loss decreased while validation loss stayed constant), indicating that the aggressive lr found a local minima that couldn't be escaped, thus a lower lr was used [8]
- **Momentum**: a higher momentum was better, confirming that it is necessary for local minima escape
- **Batch size**: larger batch sizes were better, further confirming that local minima are an issue [9]

Drop out and L2 regularization were explored to reduce this yielding a model with 0.452265 overall and 0.46 average class accuracy. Dropout of 0.1 and no L2 regularization was chosen and revealed the following:

Architecture	Hyperparameters	Overall Accuracy	By-class Avg Accuracy
DenseNet161 - Base	LR=0.001, Moment=0.9, BS=32	0.450809	0.45
DenseNet161 - Dropout	DO=0.1 LR=0.001, Moment=0.9, BS=32	0.452265	0.46
ConvNeXt - Base	LR=0.001, Moment=0.9, BS=32	0.605502	0.60
ConvNeXt - Dropout	DO=0.3, LR=0.001, Moment=0.9, BS=32	0.599029	-
ConvNeXt - Data Augment	NMB=5, LR=0.001, Moment=0.9, BS=32	0.619741	0.62
ConvNeXt - Dropout Augment	NMB=5, DO=0.3, LR=0.001, Moment=0.9, BS=32	0.614725	-
ConvNeXt - Final	Epochs= 15, NMB=5, DO=0.3, LR=0.001, Moment=0.9, BS=32	0.686246	0.69

Table 1: Final results of various models and hyperparameters.

- **Dropout**: lower dropout was better as anticipated for heavily overfit data [10]
- **L2 regularization**: adding and increasing weight decay decreased model performance which can likely be attributed to its potential to make models simpler via shrinking weights which omits complex patterns [11]. This is particularly impactful with the DD data where minor pattern alterations are crucial to capture.

Overall, it was seen that fine-tuning the DenseNet161 model yielded minimal increase in accuracy, further highlighting the inability for traditional CNN architectures to learn small differences between DD classes without full end-to-end training.

The same hyperparameters were utilized with ConvNeXt except for lr as it was anticipated less overfitting would occur with ConvNeXt. However, higher lr still revealed no learning was occurring, thus the same parameters were used from the DenseNet161 model. An immediate increase to 0.605502 overall and 0.60 average class accuracy was observed. Learning rate curves also showed a vast decrease in variance and overfit as seen in Figure 2.

This confirms that the attention mechanisms of ConvNeXt have a large impact on the ability to "learn" minor alterations in classes without end-to-end training or manual image intervention.

Next, dropout was explored as it improved the DenseNet161 model. Dropout decreased model performance to 0.599029, but revealed less variance in the learning curves and showed promise to increase with increased epochs, thus dropout of 0.3 was carried forward.

To further reduce variance, TrivialAugmentation of various number of magnitude bins (NMB) was explored. Image augmentation has been shown to help reduce overfitting. Accuracy increased to 0.619741 overall and 0.62 per class. Experiments revealed that a lower NMBs yielded slightly lower accuracy, but less overall variance, so 5 NMBs were used. This makes sense as too much augmentation via affine transformations can crop some of the key indicators out of images (e.g. visor, cell phone, etc.).

The final attempt to reduce variance was incorporating dropout. This actually lead to a decrease in accuracy (0.614725), but showed lower variance and a higher potential for continued learning with extended epochs.

Finally, to decrease the bias of the model, an alternative optimizer was used - Adam. To reduce the requirement for further tuning of momentum, and encourage learning with a dynamic learning rate, Adam had the potential to outperform stochastic gradient descent. Along with extended epochs, the final model obtained a 0.686246 overall and 0.69 by-class accuracy as seen in Figure 3.

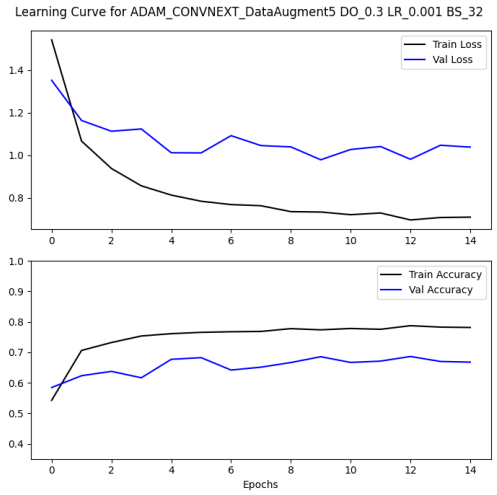
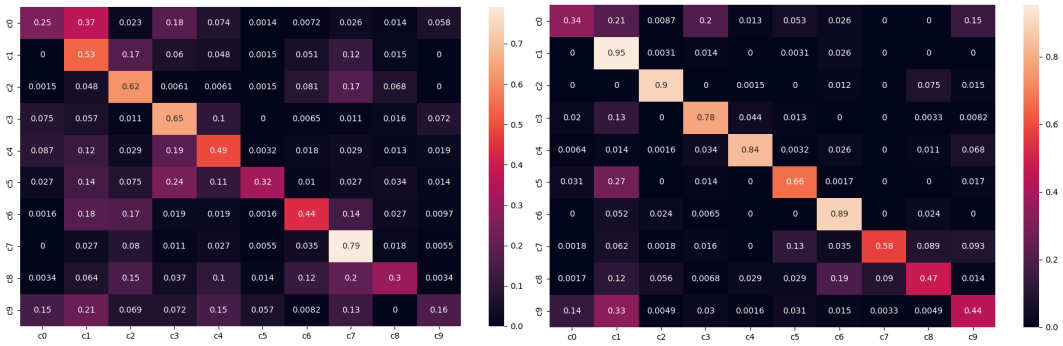


Figure 3: Final ConvNeXt model learning curves.

By class, improvement up to 175% was seen and was and average of 67% improvement by class, which indicates the true power of this attention-based architecture. Confusion matrix comparison between the models can be seen in Figure 4.



(a) DenseNet161 Confusion Matrix

(b) ConvNeXt Confusion Matrix

Figure 4: By-class accuracy is higher and more consistent for the final ConvNeXt model compared to the tuned DenseNet161 model as seen by the confusion matrices.

Something to not overlook is the time to train each model. DenseModel161 takes 38 mins to run where ConvNeXt takes 50 mins. Although this is an increase, it is much faster than end-to-end training or ViT training.

4. Future Work

The model still struggles with the "Safe Driving", "Hair and Makeup", and "Talking to Passenger" categories. Incorporating some object-detection algorithm like YOLO could help highlight things like the visor or makeup in hands, as well as further improve identification of a phone for "Texting" or "Talking on the Phone" categories.

References

[1] Neptune AI. Transfer learning guide: Examples for images and text in keras, 2021. 2

[2] Centers for Disease Control and Prevention. Distracted driving. https://www.cdc.gov/transportationsafety/distracted_driving/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fmotorvehiclesafety%2Fdistracted_driving%2Findex.html, April 26, 2022. Accessed: April 27, 2023. 1

[3] JACOBKIE. State Farm Distracted Driver Detection. https://www.kaggle.com/competitions/state-farm-distracted-driver-detection/discussion/22906, 2016. Accessed: April 27, 2023. 1

[4] Xuefeng Liu, Xianglong Liu, and Zhichao Feng. How to fine-tune transformers (almost) efficiently. *arXiv preprint arXiv:2106.13700*, 2021. 1

[5] Z. Liu, M. Li, Y. Chen, Y. Gao, X. Wang, and S. Liu. A convnet for the 2020s. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1905–1919, 2021. 1

[6] S. Muller, S. Soleymani, A. Daghagheleh, H. Falsafi, and S. Escalera. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7129–7138, 2021. 2

[7] J. C. Quiroz, J. P. Aguirre, and R. Pizarro. Comparing approaches for distracted driver detection using computer vision techniques. In *2018 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 29–35, 2018. 1, 2

[8] Leslie N. Smith. Cyclical learning rates for training neural networks. *arXiv preprint arXiv:1506.01186*, 2017. 2

[9] Leslie N. Smith and Nicholay Topin. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2018. 2

[10] Stefan Wager, Sida Wang, Percy Liang, and James Zou. Understanding dropout. *arXiv preprint arXiv:1312.6197*, 2013. 3

[11] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 3