CNN-based Driver Activity Understanding: Shedding Light on Deep Spatiotemporal Representations

Alina Roitberg Monica Haurilet Simon Reiß Rainer Stiefelhagen

Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
{firstname.lastname}@kit.edu

***Abstract*— While deep Convolutional Neural Networks (CNNs) have become front-runners in the field of driver observation, they are often perceived as black boxes due to their end-to-end nature. *Interpretability* of such models is vital for building trust and is a serious concern for the integration of CNNs in real-life systems. In this paper, we implement a diagnostic framework for analyzing such models internally and shed light on the learned spatiotemporal representations in a comprehensive study. We examine prominent driver monitoring models from three points of view: (1) visually explaining the prediction by combining the gradient with respect to the intermediate features and the corresponding activation maps, (2) looking at what the network has learned by clustering the internal representations and discovering, how individual classes relate at the feature-level, and (3) conducting a detailed failure analysis with multiple metrics and evaluation settings (*e.g.* common versus rare behaviors). Among our findings, we show that most of the mistakes can be traced back to learning an object- or a specific movement bias, strong semantic similarity between classes (*e.g. preparing food* and *eating*) and underrepresentation in the training set. Besides, we demonstrate the advantages of the Inflated 3D Net compared to other CNNs as it results in more discriminative embedding clusters and in the highest recognition rates based on all metrics.**

I. Introduction and Related Work

The lack of transparency and the inability to efficiently visualize internal decision processes resulted in Convolutional Neural Networks (CNNs) being labeled as black boxes, considerably slowing down their integration in industrial systems. In contrast to conventional feature-based methods [1], [2], [3], intermediate representations of such end-to-end architectures, are not defined by hand but *learned* together with the classifier. While 3D CNNs have demonstrated impressive results in driver activity understanding [4], [5], [6], the analysis of what such models have learned is considerably harder due to their end-to-end nature. As failures of such networks might lead to catastrophic results in real-life applications, studying how such architectures function internally becomes increasingly important for overcoming biases [7], identifying most relevant data [8] and explaining failure cases [9].

In this work, we make the first step towards transparency behind spatiotemporal CNNs for driver monitoring, and implement multiple methods to systematically examine the learned representations. With our diagnostic framework, we are able to gain insight into (1) *where did the network look* (Figure 1), *i.e.* which video regions have guided the current decision in cases of both, success and failure (2) *what did the network learn i.e.* exploring the intermediate layer representations with unsupervised learning and detecting relationships between different behaviors, and (3) a detailed performance analysis focused on common misclassifications of the individual classes and the relation to data scarcity.

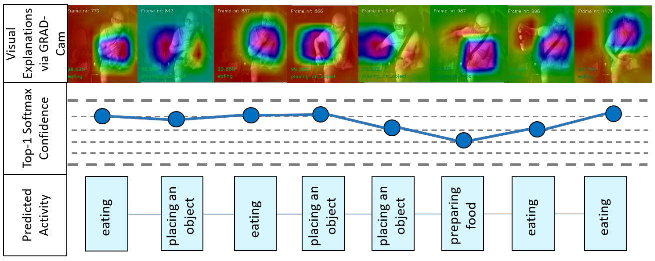


Fig. 1: Gaining insight into spatiotemporal CNNs for driver action recognition. We leverage the gradient for the predicted class with respect to the intermediate feature maps of the convolutional layer to obtain video regions which have driven the neural network decision and understand misclassifications.

Why automatic driver behavior understanding? Understanding the human behind the steering wheel makes human-vehicle cooperation more intuitive and safe. Rising levels of automation increase human freedom, leading to drivers being engaged in distractive behaviors more often. We recognize four major use-cases for applications of driver activity recognition models in practice. (1) As the current activity directly affects the cognitive workload [10], driver monitoring makes automated vehicles safer by assessing the level of alertness. Therefore, the key application of such algorithms at SAE levels 0 to 3 [11] is the assessment of human distraction levels and reacting accordingly, for example, with a warning signal. (2) With the automation rising to SAE levels 4 and 5, increasing driver comfort becomes the most important use-case. For example, movement dynamics might automatically adjust depending on the detected activity (*e.g.* softer driving if the person is drinking tea or sleeping). (3) Activity recognition task is also highly related to the problem of gesture recognition [12], which can be used as a novel intuitive communication interface inside the vehicle. (4) A further safety-related application of activity recognition during manual driving is intention prediction. As the majority of traffic fatalities is caused by human errors, a system capable of foreseeing such maneuvers might intervene, before it is too late.

Conventional driver activity recognition Feature-based methods, which have dominated the field for decades, follow the classical machine learning pipeline comprising two phases. First, a feature vector representing the input data is estimated. The way the data is processed in this step is manually defined by human experts and is often based on the body pose [1], [3], eye gaze [13], [14], hand location [15], [1], head pose [14], [2], [15], detected objects [16] or vehicle dynamics [17], [2]. The resulting feature is then passed to a machine learning framework based on *e.g.* Support Vector Machines [14], Hidden Markov Models [2] or Recurrent Neural Networks [16], [3], [2]. Given the controlled nature of the first phase (*i.e.* , human designed feature calculation and -selection), the decision pathways of such methods tend to be easier to interpret.

End-to-end driver observation with CNNs In end-to-end networks, feature extraction and classification merge *into one global model*. Deep CNNs *operate directly on the input video* and the intermediate representation is not defined but *learned* through convolution filters. Such models have been recently utilized for driver activity recognition [4], [18], [19], intention prediction [5], posture classification [6] and gaze zone estimation [20] with great success. As activity understanding extends image recognition with the temporal dimension, most approaches leverage 3D convolutions. For example, in Inflated 3D Net [21], [4], weights of hierarchically stacked $3 \times 3 \times 3$ convolution filters are learned together with the classification layer to obtain discriminative spatiotemporal representations. In contrast to *e.g.* skeleton-based methods, intermediate representations of deep CNNs are an enigma to the naked eye. Even more so, demystifying the internal decision processes and analyzing failure cases is vital for applications and is a growing area in general image recognition [8], [22]. While recent works aim to *quantify* the uncertainty of driver activity classification models [23], [24], [25], diagnostic tools for tracing back the root causes for the classification outcome have been overlooked in this field.

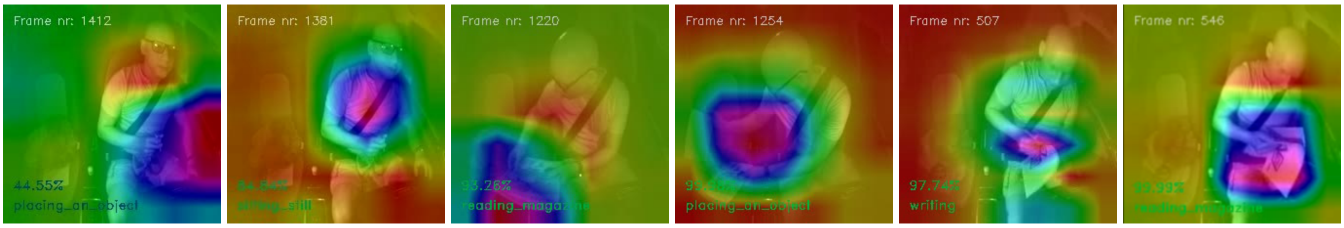
Contributions and Summary Given the black box nature of CNNs, we argue, that shedding light on the learned spatiotemporal representations is crucial for their applications in real driver monitoring systems, where lack of trust in such data-driven models remains an obstacle. This work develops a diagnostic framework for interpreting decisions of such networks and can be summarized in three major contributions. (1) First, we aim for *visual explanations* of the internal decisions and analyze, where the network attended when it predicted the specific behavior. To this end, we set our target as the predicted class and backpropagate the gradient to the last convolution layer, building on the method of [8] and extending it to the temporal dimension. We then weigh the individual feature activation maps at that particular layer based on the gradient. We examine the resulting heat-maps which indicate the image regions directing the specific decision (Figure 1) and compare the focus of the network in cases of success and incorrect predictions. (2) We then consider the *representation point of view* and examine, what the network has learned internally for three different models previously used for driver monitoring. To interpret hundreds of neurons of the last network layer, we reduce the dimensionality using t-SNE [26] and examine the resulting clusters, which are far more discriminative for the Inflated 3D Net. We further identify relationships between the learned representations of individual classes by using Ward's hierarchical agglomerative clustering. (3) Finally, we conduct a comprehensive study of the model performance, by analyzing *e.g.* the top-5 generalization and the most common confusion of the individual classes. Our findings indicate, that the main failure cases can be traced back to either semantic similarity combined with underrepresentation in the training set (*e.g. closing* versus *opening bottle*) or a learned movement-, object- or position bias (*e.g.* misclassification as *reading magazine* if a magazine is somewhere in the scene), highlighting the need of more diverse object placement in the datasets. The experiments using our diagnostic framework show encouraging evidence, that deep CNNs for driver observation have the potential to become more interpretable.

II. Analysis of Deep Spatiotemporal Representations for Driver Monitoring

A. Evaluated CNNs and Dataset for Driver Observation

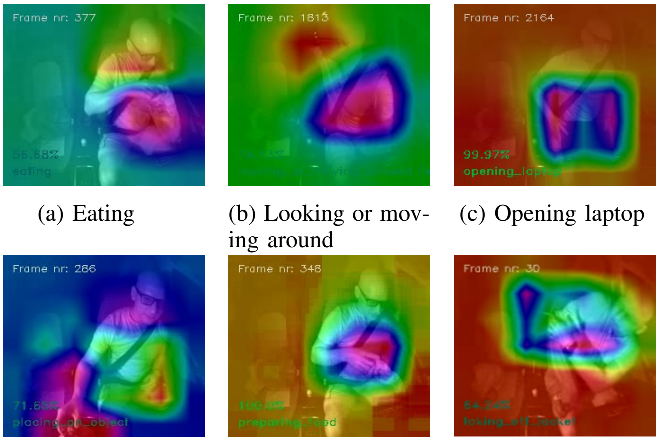
Testbed We use Drive&Act [4], the largest available testbed for driver activity recognition and extend it with our diagnostic framework. Following the original work [4], we use the three Drive&Act splits, which have different people for training (10 people), test (3 people) and validation (2 people). We focus on the 34 fine-grained activities and the frontal near-infrared camera view as our evaluation setup.

Neural Architectures We consider three published approaches based on spatiotemporal CNNs, which were initially developed for standard video classification and have recently become front-runners in driver observation [4]: C3D [27], Inflated 3D ConvNet [21] and Pseudo3D ResNet [28]. All these architectures directly operate on the video data and learn the intermediate embeddings together with the classifier layers in an end-to-end fashion. C3D and Inflated 3D ConvNet deal with the spatial and temporal dimensions of our input by leveraging hierarchically stacked 3D-convolution and -pooling kernels with the size of $3 \times 3 \times 3$ for most layers. P3D ResNet, on the other hand, mimics 3D convolutions by applying a filter on the spatial domain ($3 \times 3 \times 1$) followed by one in the temporal dimension ($1 \times 1 \times 3$). While we analyze all three models in Section II-C and Section II-D, we choose the Inflated 3D Net for the visual explanations in Section II-B, as it has shown the best recognition results in previous work.



(a) Prediction: **placing object** (incorrect ✗) (b) Prediction: **sitting still** (correct ✓) (c) Predict.: **reading magazine** (incor. ✗) (d) Prediction: **placing object** (correct ✓) (e) Prediction: **writing** (incorrect ✗) (f) Predict.: **reading magazine** (corr. ✓)

Fig. 2: **Correct vs. Misclassified Predictions**: Analysis of video segments using gradient weighted class activation maps, where samples were close to each other and comprised the same behavior, but resulted in different predictions.



(a) Eating (b) Looking or moving around (c) Opening laptop

(d) Placing an object (e) Preparing food (f) Taking off jacket

Fig. 3: Activation maps of the last Inflated 3D ConvNet convolution layer weighted by the gradient. Heat-map overlays illustrate, which region has driven the network's decision.

B. Where did the network look? Visualizing Internal Decisions with Gradient-weighted Class Activation Mapping

We implement a three-dimensional version of the gradient-weighted class activation map technique [8] and provide visual explanations of spatiotemporal CNNs for the first time. Given an input video, we first conduct a conventional forward pass and obtain the predicted class c *i.e.* the class with the highest activation. Then, we estimate the gradient over y_c (the output before the softmax layer) with respect to each individual value in the kth feature map A_k of a layer in the CNN. This is used to obtain the *feature importance* w_c^k for each individual feature map k by averaging the gradients over all its n values:

$$w_c^k = \frac{1}{n} \sum_{i,j,t} \Big(\frac{\partial y_c}{\partial A_k^{i,j,t}} \Big), \tag{1}$$

where $A_k^{i,j,t}$ is the activation at position in space i, j and time t. In each location (i, j, t) we linearly combine the values in the feature map by the importance estimate w_c^k. The final weights $V_c^{i,j,t}$ are obtained by passing the computed values to a ReLU function to remove negative values, as we are only interested in pixels that increase y_c. More formally, we calculate the final weights as follows:

$$V_c^{i,j,t} = \text{ReLU} \Big(\sum_k w_c^k A_k^{i,j,t} \Big). \tag{2}$$

To be able to visualize the resulting explanation as images, we average the resulting heat-maps over the time dimension. We provide the resulting visual explanations of the Inflated 3D ConvNet decisions for different classes in Figure 3, while Figure 2 illustrates key differences between correct and failed predictions. For example, the network features characteristic for *eating* are focused around both, hands and head (probably due to chewing, Figure 3a), while *preparing food* is linked to the hands only (Figure 3e). The network attention is different depending on the activity, but in general, we observe increased focus on human hands and head. There is also a visible object bias, which is useful in many cases (*e.g.* a laptop or a newspaper in the scene increases the chances of an activity involving these objects). However, such object bias might lead to mistakes, if *e.g.* the human is only *placing* a magazine but *reading magazine* is predicted (Figure 2c). Figure 2e reveals, that a specific hand movement leads to the network predicting *writing*, while the person is actually *reading*. While in most cases the network seems to make the predictions for the right reasons, specifically looking at uncertain cases helps us to draw useful conclusions for improvement. For example, our analysis highlights the need for diversification of training data in terms of object placement, so that the network predicts object-related activities if the human interacts with them, and not, if they are simply present in the scene.

C. What did the network learn?

We now gain insight into the intermediate features of the CNNs, to verify whether they provide good generic representations of driver behavior. We use the first validation split of Drive&Act and extract the features of the fully connected layer of the C3D, Pseudo 3D ResNet and Inflated 3D ConvNet. To make sense of hundreds of neurons, we first reduce the dimensionality using t-SNE [26]. We then visualize each video clip in two-dimensional space in Figure 4, marking behavior classes with different colors. We qualitatively observe that Inflated 3D Net captures the nature of activities better, as its features form far more discriminative clusters. Still, samples of the same activities also shape visible groups for C3D and Pseudo 3D ResNet, but the boundaries are far less concise.

We now examine how different behaviors are connected from the CNN point of view. First, we compute the class

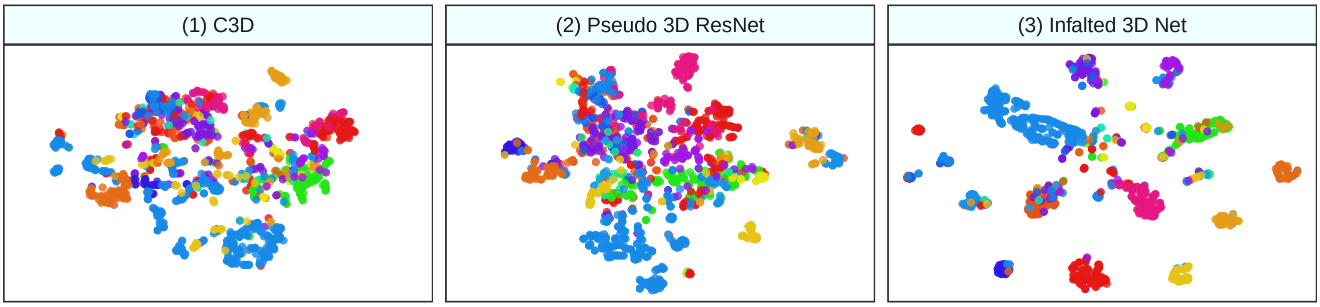


Fig. 4: Visualizations using t-SNE of the intermediate representations learned by different CNN models. Different behavior classes are marked with different colors. While all models have clear correlations of the embedding values and the activity, such "class-specific cluster" are much more discriminative for the Inflated 3D Net.

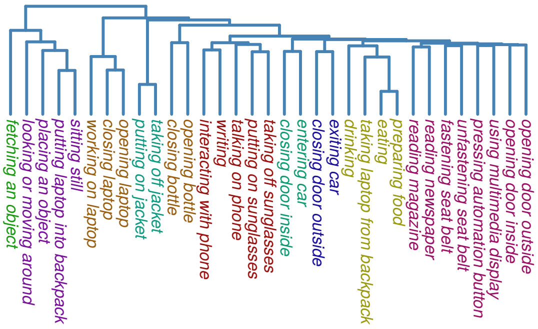


Fig. 5: Hierarchical clustering of the learned representations of the individual classes. We compute the mean vector of the intermediate Inflated 3D Net embedding for each activity and then use Ward's hierarchical agglomerative clustering to reveal learned relationships between them.

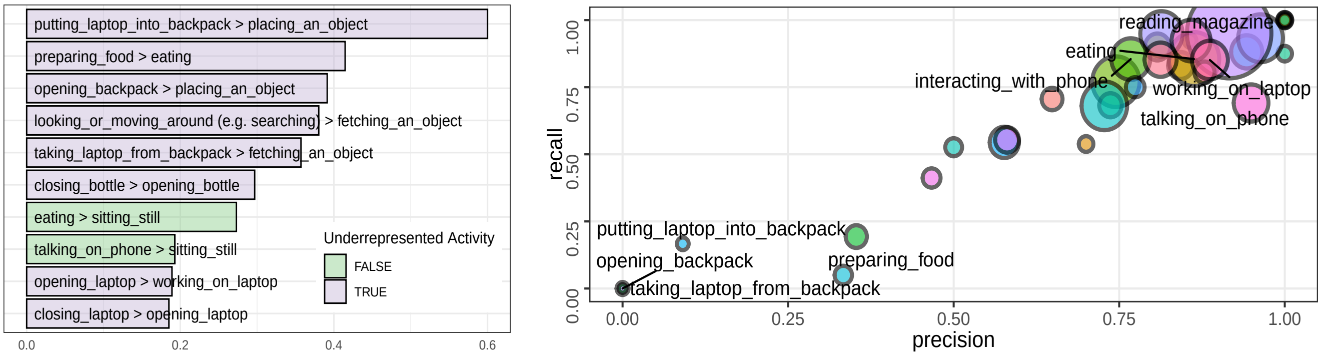
centroid vector by averaging the fully connected I3D features of each activity. We apply the Ward's Hierarchical Agglomerative Clustering method [29] on the class centroids. The resulting class hierarchy, illustrated in Figure 5, reveals how the classes are connected in the model internally. While most of the semantically related activities are also placed together in the cluster hierarchy (*e.g.* *opening* and *closing bottle*), such similar cases often lead to high confusion, as we will show quantitatively in the next section. We can also understand how the network operates by looking at these relations, *e.g.* the activities *writing*, *talking on phone* and *putting on sunglasses* all fall into the same red cluster (Figure 5), while they do not match semantically at first glance. As the network groups these behaviors, we infer that it has learned them as fine-grained hand-centric actions and makes its decisions based on the concise hand movements. This is confirmed by the visual explanation in Figure 2e, where the model inaccurately predicts *writing* by focusing on a very small area around the hand instead of the object. The model view of some activities is surprising, for example, *taking laptop from backpack* is connected to *eating*, *preparing food* and *drinking*. The quantitative analysis in our next section will uncover, that this action is indeed very poorly recognized. The way, the network interprets this behavior is therefore simply incorrect. We assume that the model has learned a certain place bias, as a lot of coarse movements in front of the torso is typical for these actions. Extending the dataset with more diverse examples of this action (*e.g.* taking out the laptop in other locations) might therefore be beneficial.

D. A Detailed Misclassification Analysis

Previous evaluation of CNN-based models inside the vehicle cabin has focused on the multi-class top-1 accuracy as a single performance metric [4]. This is an oversimplification as the prediction quality varies greatly depending on multiple factors, that we are going to uncover in this section. To examine the strengths and weaknesses of CNN-based algorithms, we extend the evaluation procedure of [4] with multiple settings and metrics. Drive&Act comprises 34 fine-grained activity classes, which, however are highly unbalanced. As CNNs are notably bad in learning from few examples, we sort the behaviors by their frequency in the dataset and divide them into *common* (top half of the classes) and rare (the bottom half). We subsequently evaluate the models in three modes: considering all activities, as it is usually done, using only the overrepresented- or only the rare classes. In addition to the conventional top-1 accuracy, we evaluate the top-5 accuracy, *i.e.* we consider the sample as correctly classified if any of the five classes with the highest probabilities match the ground truth. The top-5 accuracy might be useful if we want to overlook confusions of highly similar classes (*e.g.* *fastening* and *unfastening seatbelt*) and are only interested in coarse recognition. We further extend the original evaluation protocol with the Precision P, Recall R and $F1$ score of the individual classes. Formally, our metrics (including the balanced multi-class accuracy Acc) are defined as:

$$Acc = \frac{\sum_{i=1}^{n} \frac{A_i^{corr}}{A_i^{total}}}{n} \qquad P = \frac{A_i^{corr}}{A_i^{pred}}, \quad R = \frac{A_i^{corr}}{A_i^{total}} \quad F1 = 2 \times \frac{P \times R}{P + R} \qquad (3)$$

where n is the total number of classes, A_i^{pred} the total number of examples which were assigned the label i, A_i^{corr} is the number of correctly predicted instances of class i, and A_i^{total} depicts the total frequency of class i in the test set.



(a) Most common misclassifications of the I3D model. Color shows, whether the class was underrepresented

(b) Precision and recall of the individual classes. Circle size corresponds to the number of samples in the dataset (for readability only some activities are labeled)

Fig. 6: Misclassification statistics of the Inflated 3D ConvNet on the Drive&Act dataset

In Table I we compare different architectures in terms of their top-5 and top-1 accuracy for rare, overrepresented and all activity classes. While the Inflated 3D ConvNet outperforms other approaches in all metrics (63.64% top-1 test accuracy for all classes), C3D seems to be stronger than Pseudo 3D ResNet in terms of the top-5 accuracy, while the latter model is better in top-1 classification. C3D therefore is well suited for coarse classification but has issues discovering fine-grained structures. While it is expected, that the top-1 recognition rate is significantly lower than the top-5 results, this gap grows by a large margin for rare classes (*e.g.* this difference is 32.52% for uncommon- and 17.18% for common actions when considering the Inflated 3D ConvNet test setting). In general, activity recognition models seem to perform well for coarse behavior recognition (over 80% top-5 recognition rate in all settings for Inflated 3D ConvNet), while there is room for improvement in detecting fine-grained structures, especially for underrepresented classes (top-1 Inflated 3D ConvNet accuracy for rare categories under 50%). Still, identifying half of the actions which only had few training samples correctly is a good result, as CNNs are known for being data-hungry and the random baseline is only $100/34 = 2.94\%$, as we have 34 actions in total.

We now examine model performance *for the individual classes*, with exact precision, recall, F1-score and most common confusion provided in Table II. We see in Figure 6b, that while *all* of the very poorly recognized actions are underrepresented (frequency in the training set is illustrated through the circle size), well-recognized behaviors can be both: common and rare classes. The models are therefore *surprisingly tolerant to learning from few examples in case of highly discriminative actions*. For example, *closing door from outside* only has around 20 examples in the complete dataset (see [4] for the sample frequency statistics). However it is recognized correctly in 73% of the test cases (Table II), probably since the human is acting outside of the vehicle, which is easy to distinguish from the other activities. The combination of low discriminativeness and underrepresentation are fatal for a class: *e.g. taking laptop from backpack* and *preparing food*) recognized correctly in only 14% and 7% of the test set cases. In Figure 6a we summarize the most common Inflated 3D ConvNet confusions, disclosing that eight out of ten most frequent mistakes entail an underrepresented ground-truth class. Oftentimes, the confusion happens when two the behaviors are semantically very close and one of them is rare. In this case, the model tends to predict the more frequent class (*e.g. preparing food* classified as *eating* in 41% of cases). Another cause of confusion if one action being a special case of another: *putting laptop into backpack* is a specialization of *placing an object* and is classified as such 60% of times. Similarly, *taking laptop from backpack* is marked as *fetching an object* in 36% of the test set samples. This might be connected to the fact, that modern architectures downsample the image relatively fast to obtain large receptive fields and therefore focusing on classification of coarse structures. Developing models which fit well for *fine-grained* recognition would therefore be beneficial. Some of the common confusions in Table II are surprising and uncover potential biases. For example, most common confusion of *putting on sunglasses* is not *taking off sunglasses*, but *closing bottle*. The model has presumably learned a bias of concise hand-centric movements, which are the common pattern of all these actions. Expanding the training set with more diverse examples might be important for learning to predict these activities *for the right reasons*, such as a *combination* of typical hand location, -movement and the correct object being held.

TABLE I: Top-1 and top-5 accuracy for fine-grained activity recognition on the Drive&Act dataset, evaluated separately for classes over- and underrepresented during training.

Model	**Common**		**Rare**		**All classes**	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Validation						
C3D	54.44	87.53	45.70	75.82	50.07	81.67
Pseudo 3D ResNet	58.00	86.61	52.08	74.77	55.04	80.69
Inflated 3D Net	**80.62**	**95.83**	**58.50**	**87.88**	**69.67**	**91.85**
Test						
C3D	47.97	83.75	38.86	74.02	43.41	78.89
Pseudo 3D ResNet	52.43	84.05	38.20	65.09	45.32	74.57
Inflated 3D Net	**77.88**	**95.06**	**49.41**	**81.93**	**63.64**	**88.49**

TABLE II: Detailed test set performance of I3D . Mistakes often occur in semantically close activities or in cases, where one activity is a specialization of another one (*e.g. taking laptop from backpack* as a special type of *fetching an object*).

True Activity Class	Prec.	Recall	F1	Most Common Confusion	
	%	%	%	Class	%
close_bottle	0.57	0.47	0.51	open_bottle	0.30
close_door_inside	0.70	0.82	0.76	entering_car	0.06
close_door_outside	0.73	0.73	0.73	exiting_car	0.18
close_laptop	0.67	0.37	0.48	open_laptop	0.19
drinking	0.93	0.88	0.90	close_bottle	0.05
eating	0.76	0.59	0.67	sitting_still	0.27
entering_car	0.77	0.74	0.75	close_door_inside	0.11
exiting_car	0.83	0.80	0.82	close_door_outside	0.08
fastening_seat_belt	0.77	0.82	0.79	placing_an_object	0.04
fetching_an_object	0.64	0.67	0.65	placing_an_object	0.13
interact_with_phone	0.92	0.86	0.88	eating	0.04
looking_or_moving	0.14	0.04	0.06	fetching_an_object	0.38
open_backpack	0.14	0.09	0.11	placing_an_object	0.39
open_bottle	0.72	0.68	0.70	close_bottle	0.13
open_door_inside	0.65	0.57	0.60	close_door_inside	0.09
open_door_outside	0.89	0.89	0.89	exiting_car	0.11
open_laptop	0.54	0.51	0.53	working_on_laptop	0.19
placing_an_object	0.59	0.72	0.64	fetching_an_object	0.11
preparing_food	0.19	0.07	0.11	eating	0.41
pressing_button	0.89	0.98	0.93	using_mm_display	0.02
put_laptop_backpack	0.27	0.20	0.23	placing_an_object	0.60
putting_on_jacket	0.43	0.62	0.51	taking_off_jacket	0.15
putting_on_sunglasses	0.88	0.71	0.79	close_bottle	0.05
reading_magazine	0.89	0.88	0.88	reading_newspaper	0.08
reading_newspaper	0.79	0.90	0.84	placing_an_object	0.05
sitting_still	0.87	0.93	0.90	using_mm_display	0.02
take_laptop_backpack	0.40	0.14	0.21	fetching_an_object	0.36
taking_off_jacket	0.45	0.70	0.55	putting_on_jacket	0.15
taking_off_sunglasses	0.75	0.56	0.64	fetching_an_object	0.16
talking_on_phone	0.85	0.71	0.77	sitting_still	0.19
unfastening_seat_belt	0.84	0.68	0.75	putting_on_jacket	0.08
using_mm_display	0.87	0.98	0.92	sitting_still	0.01
working_on_laptop	0.90	0.76	0.82	fetching_an_object	0.06
writing	0.86	0.58	0.70	reading_newspaper	0.13

III. Conclusion

Safety critical systems with human lives at stake have to be robust in fulfilling their objective. Yet, if a system failure does arise its root cause has to be understood. With this notion in mind, we propose measures to overcome the deficiency in interpretability of CNN-based behavior recognition in passenger vehicles. With a thorough inspection of the automatically learned inner representations, we are able to reason about preferable decision boundaries drawn by different CNN architectures. With our extension of the gradient-weighted class activation maps into the temporal space, the visual inspection of spatiotemporal cues leading to failed predictions become much more tangible. With our diagnostic framework in place, narrowing down causes of failures enable testing pipelines to preemptively identify shortcomings in (1) the data-distribution (2) learned representations and (3) may provide guidance in eradicating bias.

References

[1] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

[2] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture," *arXiv preprint arXiv:1601.00740*, 2016.

[3] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhagen, "Body pose and context information for driver secondary task detection," in *Intelligent Vehicles Symposium*. IEEE, 2018, pp. 2015–2021.

[4] M. Martin*, A. Roitberg*, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles," in *ICCV*. IEEE, October 2019.

[5] P. Gebert, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks," in *Intelligent Vehicles Symposium*. IEEE, 2019.

[6] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," *NeurIPS Workshop on Workshop on Machine Learning for Intelligent Transportation Systems*, 2018.

[7] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *ECCV*. Springer, 2018, pp. 793–811.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[9] C. Feichtenhofer, A. Pinz, R. P. Wildes, and A. Zisserman, "What have we learned from deep representations for action recognition?" in *CVPR*, 2018, pp. 7844–7853.

[10] E. Wolf, M. Martinez, A. Roitberg, R. Stiefelhagen, and B. Deml, "Estimating mental load in passive and active tasks from pupil and gaze changes using bayesian surprise," in *ICMI-MCPMD*, 2018.

[11] S. International, "Automated driving: levels of driving automation are defined in new sae international standard j3016," 2014.

[12] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhagen, "Analysis of deep fusion strategies for multi-modal gesture recognition," in *IEEE CVPR-AMFG Workshop*, 2019.

[13] N. Akai, T. Hirayama, L. Y. Morales, Y. Akagi, H. Liu, and H. Murase, "Driving behavior modeling based on hidden markov models with driver's eye-gaze measurement and ego-vehicle localization," in *Intelligent Vehicles Symposium*. IEEE, 2019, pp. 949–956.

[14] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Head, eye, and hand patterns for driver activity recognition," in *ICPR*, 2014.

[15] S. Martin, E. Ohn-Bar, A. Tawari, and M. M. Trivedi, "Understanding head and hand activities and coordination in naturalistic driving videos," in *Intelligent Vehicles Symposium*, 2014, pp. 884–889.

[16] P. Weyers, D. Schiebener, and A. Kummert, "Action and object interaction recognition for driver activity classification," in *Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019.

[17] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán, "Exploiting map information for driver intention estimation at road intersections," in *Intelligent Vehicles Symposium*. IEEE, 2011, pp. 583–588.

[18] S. Reiß, A. Roitberg, M. Haurilet, and R. Stiefelhagen, "Deep Classification-driven Domain Adaptation for Cross-Modal Driver Behavior Recognition," in *Intelligent Vehicles Symposium (IV)*.

[19] ——, "Activity-aware Attributes for Zero-Shot Driver Behavior Recognition. ," in *CVPR VL-LL Workshop*. IEEE, June 2020.

[20] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 849–854.

[21] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*. IEEE, 2017.

[22] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *CVPR*, 2015, pp. 5188–5196.

[23] A. Roitberg, C. Ma, M. Haurilet, and R. Stiefelhagen, "Open Set Driver Activity Recognition," in *Intelligent Vehicles Symposium (IV)*. IEEE, June 2020.

[24] A. Roitberg, M. Haurilet, M. Martinez, and R. Stiefelhagen, "Uncertainty-sensitive Activity Recognition: a Reliability Benchmark and CARING models," 2020.

[25] A. Roitberg, Z. Al-Halah, and R. Stiefelhagen, "Informed Democracy: Voting-based Novelty Detection for Action Recognition," in *British Machine Vision Conference (BMVC)*, UK, 2018.

[26] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[28] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017, pp. 5533–5541.

[29] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, 1963.