Evaluation of Video Masked Autoencoders' Performance and Uncertainty Estimations for Driver Action and Intention Recognition

Koen Vellenga[*†], H. Joe Steinhauer[*], Göran Falkman[*], and Tomas Björklund[†]
[*]University of Skövde, Sweden
[†]Volvo Car Corporation, Sweden

Abstract

Traffic fatalities remain among the leading death causes worldwide. To reduce this figure, car safety is listed as one of the most important factors. To actively support human drivers, it is essential for advanced driving assistance systems to be able to recognize the driver's actions and intentions. Prior studies have demonstrated various approaches to recognize driving actions and intentions based on in-cabin and external video footage. Given the performance of self-supervised video pre-trained (SSVP) Video Masked Autoencoders (VMAEs) on multiple action recognition datasets, we evaluate the performance of SSVP VMAEs on the Honda Research Institute Driving Dataset for driver action recognition (DAR) and on the Brain4Cars dataset for driver intention recognition (DIR). Besides the performance, the application of an artificial intelligence system in a safety-critical environment must be capable to express when it is uncertain about the produced results. Therefore, we also analyze uncertainty estimations produced by a Bayes-by-Backprop last-layer (BBB-LL) and Monte-Carlo (MC) dropout variants of an VMAE. Our experiments show that an VMAE achieves a higher overall performance for both offline DAR and end-to-end DIR compared to the state-of-the-art. The analysis of the BBB-LL and MC dropout models show higher uncertainty estimates for incorrectly classified test instances compared to correctly predicted test instances.

1. Introduction

Traffic fatalities continue to rank among the top 20 leading death causes worldwide [11]. One of the factors listed to reduce injuries and fatalities in the future is car safety [48]. Advancements in onboard computing power, available data, artificial intelligence (AI) and the increased number of sensors mounted on vehicles have enabled the development of Advanced Driver Assistance Systems (ADAS) that aim to continuously minimize human errors and prevent accidents from happening [1]. To support a driver of an ego-vehicle to drive safely, it is critical for an ADAS to timely recognize what that driver currently does, or aspires to do. However, supporting a driver can be difficult due to irrational human behavior [12] or unlikely and unseen complex road scenarios [29]. Therefore, this paper focuses on evaluating driver action and intention recognition model performance and uncertainty estimations.

The main difference between driver action and intention recognition is that the driving actions are observable, while the intentions are not [45]. Intentions denote what the driver aspires to do in the near future (e.g., perform an overtake or turn). Driver intention recognition (DIR) can be used for assessing whether it is safe to pursue the intention. To evaluate whether future driving maneuvers are safe, one has to consider what the driver is currently doing. Previous ego-vehicle driver action recognition (DAR) and DIR studies use observations from vehicle dynamics sensors (e.g., velocity or yaw-rate), a driver monitoring system (e.g., head pose estimation, or gaze estimation), or driving-scene observations (e.g., road user detection, lane detection, or traffic sign detection) as inputs to a deep neural network (DNN) to recognize driving actions and to infer driving intentions (e.g., [3, 16, 17, 21, 30, 34, 35, 46, 49]).

Since the rise of deep learning (DL) methods, handcrafted features have been mostly exchanged for neural networks that learn to represent the input data. For example, Tong et al. (2022) [40] employ a self-supervised video pre-training (SSVP) to learn latent video representations by first learning to reconstruct the input data. After the SSVP is completed, the model is fine-tuned for a downstream task. Although this approach achieved state-of-the-art (SOTA) performance on the Kinetics-400 action recognition benchmark, the decision process and learned representations are incomprehensible for humans [36]. The lack of interpretability and transparent decision making is an existing challenge for deploying AI systems in a safety-critical environment [19]. For a future ADAS to be safely integrated in a car on the road, it must be able to express when it is unable to produce a reliable result.

Uncertainties in machine learning (ML) models come from multiple sources (e.g., measurement errors, absent or contradictory data, the impact of regularization, or errors in making inferences) [7]. Regular DNNs produce a single-point estimate, but there are multiple methods that enable a DNN to estimate a distribution instead (e.g., [4, 14, 28]). These probabilistic DL methods commonly require multiple inferences for a single instance, which is undesirable for video based end-to-end DAR and DIR given that producing a single-point estimate is already computationally intensive.

Previous DAR and DIR studies have used a Transformer architecture (e.g., [27, 46]), but, to the best of our knowledge, have not considered SSVP video masked autoencoders (VMAEs). Therefore, in this paper we focus on the ability of SSVP video transformers to recognize driving actions and intentions. We demonstrate that SSVP VMAEs achieve SOTA performance on both the DIR Brain4Cars benchmark [21] and the DAR Honda Research Institute Driving Dataset (HDD) [34]. To assess the ability of SSVP VMAEs to express uncertainty about the produced predictions, we analyze two probabilistic DL methods and compare the uncertainty estimations for correct and incorrect predictions. Furthermore, we evaluate the effect of multiple strategies to combine the in-cabin and external video streams for the DIR Brain4Cars dataset, and review the performance over time to assess ability of the VMAEs to timely recognize driving actions and intentions.

2. Related work

2.1. Driver action and intention recognition

Various methods have been applied to recognize ego-vehicle actions and intentions. For example, Tran et al. (2015) [43] use a hidden Markov model to capture the unobservable transition probabilities between states (driving maneuvers in this case). However, the majority of recent studies relies on DL approaches to implicitly extract information about the actions and intentions from the sensor observations [45]. Ramanishka et al. (2018) [34] use a combination of a CNN to obtain a feature representation of every video frame and an LSTM to recognize the driving actions based on the sequential inputs, Wang et al. (2021) [46] proposed a framework using CNN for frame level feature extraction, and a Transformer encoder-decoder setup to capture interactions. Noguchi and Tanizawa (2023) [30] construct spatial-temporal graphs based on object detection and tracking and a semi-supervised contrastive learning framework for training a graph convolutional network to recognize driving actions.

Jain et al. (2016) [21] extracted the motion of the driver's head combined with external features about the current lane, number of lanes and upcoming intersections. The in-cabin and external information was fused and fed into an LSTM to predict the driving intention. Gebert et al. (2019) [16] employed an end-to-end approach and used optical flow to encode the in-cabin videos followed by a 3D Resnet to extract features from the estimated flow. Rong et al. (2020) [35] extracted the flow of external videos and used a ConvLSTM encoder–decoder architecture to include future flow prediction to enhance the intention recognition performance. Ma et al. (2023) [27] introduced the Cross-View Episodic Memory Transformer to efficiently learn unified memory representations combined with a context-consistency loss to improve the intention recognition performance.

2.2. Probabilistic deep learning

In this study, we use the Bayes by Backprop (BBB, [4]) and Monte-Carlo (MC) dropout [14] probabilistic DL methods based on their practical implementation (refer to Gawlikowski et al. (2021) [15], or Jospin et al. (2022) [22] for a comprehensive overview of probabilistic DL methods). BBB replaces the weights of a DNN with variational distributions to approximate the posterior. MC dropout randomly drops weights for each inference, which results in a different network constellation for every forward pass.

3. Methods

3.1. Problem formulation

Suppose that a driver currently performs (action recognition) or forms an intention to perform a driving maneuver (e.g., lane change or turn) $Y \in \{y_1, \ldots, y_n\}$, a set of sensor observations $X_{i,t} = \{x_{1,1}, \ldots, x_{m,k}\}$ is collected from m modalities for k time steps, then the learning task for our model is to recognize the driving maneuver Y based on the observed sequences $X_{i,t}$.

3.2. Video Transformers

The Transformers architecture was originally introduced for language translation [44], but has also been adapted to other tasks, such as video recognition [37]. Essential components of a Transformer are the input pre-processing and the self-attention (SA) operation. The input pre-processing of videos in the context of Transfomers requires converting raw video frames into a continuous and dense representation. The tokenization process covers how the input sequence is created, for example, a video frame can be divided into a set of multiple regions (2D patch tokenization, [10]) or 3D patches (also referred to as cubes), which allow for capturing motion features from the video [2]. To transform the divided input patches of the videos into embeddings, one can use a few fully-connected or convolutional layers. Alternatively, an embedding network can be used to encode and tokenize frames or clips as input tokens [37].

Intuitively, the SA operation computes for any token in a sequence how much relevant information the other tokens

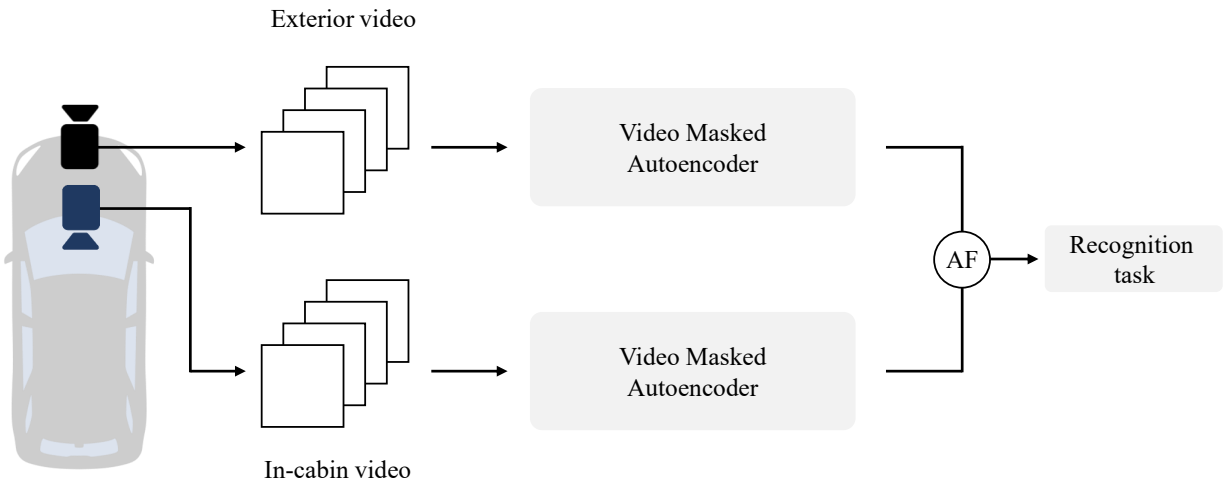


Figure 1. Overview of the multi-video setup. Video representations are learned by two video masked auto encoders, after which attention fusion (AF) is performed to create a joint embedding.

in that sequence contain. To compute the relevance for a token in sequence, the SA operation requires positional information for each token. Positional embeddings contain the temporal information of each token, and are added to the patch, frame or clip embedding.

3.3. Video Masked Autoencoder

After the success of self-supervised learning in natural language processing [9] and for images classification [18], Tong et al. (2022) [40] showed that video masked auto encoders are capable of efficiently learning relevant representations when using SSVP. To overcome the temporal redundancy aspect of videos (i.e., consecutive frames vary slowly), a higher masking ratio is required compared to image masking to make the reconstruction pre-training task difficult enough. To avoid information leakage between succeeding masked frames, Tong et al. [40] introduced tube masking to ensure that the same patches are masked for all frames of a video.

3.4. Model Architecture

Dependent on the number of available video streams, we use a setup where we fine-tune a Kinetics-400 pre-trained VMAE for each of the available video streams (see Figure 1). The benefit of this setup is that it does not require extracting additional features, such as computing the optical flow between succeeding frames as performed by Gebert et al. [16] and Rong et al. [35]. When both the in-cabin and external video streams are available, we fuse the learned representation of the VMAEs with an attention fusion layer before a fully connected layer is used to perform the recognition task. Within the pre-trained VMAEs, only the attention layers are fine-tuned [41].

3.5. Representation fusion

A challenging aspect of learning a joint representation for multiple sensor observations (modalities) is to learn a shared representation that reflects the interaction across the modalities [25]. Commonly, fusion strategies are categorized as early, intermediate and late fusion [33]. Early fusion refers to learning patterns from combined raw low-level features. Intermediate fusion first learns a representation and combines the modalities at a later stage. For the late fusion strategy, each modality is trained individually to learn uni-modal representations from which the predictions are made. The uni-modal predictions are then combined into a single prediction. In our setup, we employ an intermediate attention fusion mechanism [20] to combine the learned video representations.

3.6. Uncertainty quantification

Typically uncertainty is decomposed into aleatoric (randomness of the observations) and epistemic (the lack of knowledge or observations) [8]. The decomposition of uncertainty can help to address what component requires improvement. However, we only assess the difference in total predictive uncertainty estimates for correct and incorrect test instances. To quantify the uncertainty, we use the predictive entropy (see Equation 1) [38]. Intuitively, for an instance x the predictive entropy reaches it maximum value when the predicted probabilities for all classes are exactly the same [13]. The higher the predicted probability is for one class, the lower the predictive entropy, and the more certain the model is about the predictions.

$$H[P(y|x)] = -\sum_{y \in \mathcal{Y}} P(y|x) \log P(y|x) \quad (1)$$

Where:

$H[P(y\|x)]$	=	the entropy of the predictive distribution
$P(y\|x)$	=	conditional probability distribution over some discrete set of outcomes $\mathcal{Y}$
$\mathcal{Y}$	=	set of N stochastic predictions

4. Experiments

4.1. Datasets

For the experiments, we use two open-source driving datasets. The Honda research institute Driving Dataset (HDD) [34] contains driving actions labels and the Brain4Cars dataset [21] driving intention labels. The sections provide detailed descriptions of each dataset.

4.1.1 HDD

The HDD [34] dataset consists of 104 hours of naturalistic driving data collected in 137 sessions from February to October in 2017 in the San Francisco Bay Area. The data consists of external video data that was captured by a forward facing camera at 30 frames per second (FPS) and vehicle dynamics sensor data (e.g., throttle angle, brake pressure, steering angle, yaw rate and speed). The following eleven driving action labels are provided on a frame level: *intersection passing, left turn, right turn, left lane change, right lane change, left branch, right branch, crosswalk passing, railroad passing, merge, and u-turn*.

4.1.2 Brain4Cars

The Brain4Cars dataset [21] is an open-source DIR dataset that consists of 124 left lane changes, 58 left turns, 123 right lane changes, 55 right turns, and 234 driving straight maneuvers with a five fold 80/20 train/test split. The in-vehicle camera operated at 25 FPS, and the outside-facing camera at 30 FPS. Five driving maneuver intention labels are provided for every video: *go straight, left lane change, left turn, right lane change, right turn*.

4.2. Implementation details

For our experiments, we use an SSVP VMAE [40] with a ViT-B/16 backbone [10]. Due to the relatively small size of the datasets, and similar to previous DAR and DIR studies (e.g., [16, 24, 35]), we use a Kinetics-400 fine-tuned pre-trained model. We sample 16 frames from each video and resize the frames to a 224 x 224 resolution. Videos with less than 16 frames are extended with zero-padding at the end. As a data augmentation strategy, we divide every video into four segments from which we randomly sample four consecutive frames, apply random cropping, and horizontal flips. For in-cabin videos, we adopt the method from Rong et al [35] to randomly crop the videos towards the driver's side of the footage. The horizontal flip augmentation also results in flipping the label if necessary (e.g., for the lane change, turn and branch maneuvers). We use an AdamW [26] optimizer with a weight decay of 0.05, a linear learning rate scheduler, and 100 warm-up steps with a 5×10^{-5} base learning rate. Since the datasets are relatively small, we only fine-tuned the attention layers of the pre-trained VMAEs for 200 epochs with a 25 epochs early stopping mechanism. The models are implemented in PyTorch [31] and BayesianTorch [23], trained on multiple servers with NVIDIA T4 GPUs and use the Accelerate's [39] gradient accumulation module to achieve a batch size of 8.

For the HDD dataset, we perform DAR in an offline setting, which means that an action label is predicted for a short video instead of classifying every newly incoming video frame separately. Therefore, we only compare the DAR results to studies that also perform offline DAR based on external video data. Similar to previous studies, we use the average precision (AP) per driving action and the mean AP (mAP) as an indication for the overall performance, and use the same train/test split as provided by [34]. For the Brain4Cars dataset, we employ a five fold cross-validation and use accuracy and the F1 score in order to be consistent with previous studies. The results from Jain et al. [21] are excluded from the comparison, due to that the open-source dataset consists of less data compared to what was used in the original study.

To analyse if probabilistic variants of the VMAE based models produce different uncertainty estimates for correct and incorrect test instances, we use MC dropout and BBB. To simplify the learning of the BBB model, we only replace the last-layer (LL) weights with distributions [47]. We perform MC sampling (N=25) to produce a predictive distribution for each test instance. Subsequently, we compute the predictive entropy (Eq. 1) to analyse if the estimated uncertainty is higher for the incorrectly predicted test instances compared to the correctly predicted test instances for both datasets.

4.3. Comparisons with State-of-the-art

Table 1 presents an overview of previous offline DAR results alongside the results of the fine-tuned single exterior video stream SSVP VMAE for the eleven driving actions. Except for the recognition of U-turns, the VMAE outperforms all previous methods and achieves a mAP of 85.6%. Similar to previous studies, the *U-turn, railroad passing and right lane branch* driving maneuvers are most difficult to correctly recognize. The poor performance for these classes is most likely due to the low number of avail-

Table 1. Performance comparison to previous state-of-the-art offline driving action recognition for the HDD dataset. Best performance is highlighted in bold.

Model	Intersection passing	L turn	R turn	L lane change	R lane change	L lane branch	R lane branch	Crosswalk passing	Railroad passing	Merge	U-turn	Overall mAP (↑)
C3D [42]	82.4	77.4	80.7	67.9	56.9	59.7	5.2	17.4	3.9	20.1	29.5	45.5
ID3 [5]	85.6	79.1	78.9	74.0	62.4	59.0	14.3	29.8	0.1	20.1	41.4	49.5
GCN [24]	85.5	77.9	79.1	76.0	62.0	64.0	19.8	29.6	1.0	27.7	39.9	51.1
SCL [30]	98.3	94.1	95.8	62.6	67.3	53.4	28.4	78.0	1.2	22.2	**60.0**	60.1
GCL [30]	98.4	93.9	95.5	64.2	69.0	55.8	34.5	73.4	24.4	42.4	30.0	62.0
VMAE	**99.2**	**98.0**	**99.8**	**99.0**	**94.5**	**95.8**	**69.2**	**94.8**	**45.4**	**90.6**	55.6	**85.6**

Table 2. Comparison of the fine-tuned VMAE models to previous state-of-the-art approaches for the Brain4Cars dataset. Best performance is highlighted in bold.

Data Source	Method	Acc (↑)	F1 (↑)
In-cabin	Gebert et al. [16]	83.00 ± 2.50	81.70 ± 2.60
	Rong et al. [35]	77.40 ± 0.02	75.49 ± 0.02
	Ma et al. [27]	84.47 ± 5.98	**82.66 ± 5.40**
	VMAE	**85.48 ± 3.06**	80.22 ± 4.13
External	Gebert et al. [16]	53.20 ± 5.00	43.40 ± 9.00
	Rong et al. [35]	60.87 ± 0.01	66.38 ± 0.03
	Ma et al. [27]	64.75 ± 2.82	66.31 ± 2.19
	VMAE	**86.50 ± 1.75**	**86.57 ± 2.54**
In-cabin & External	Gebert et al. [16]	75.50 ± 2.40	73.20 ± 2.20
	Rong et al. [35]	83.87 ± 0.01	84.30 ± 0.01
	Ma et al. [27]	85.37 ± 2.95	87.09 ± 0.23
	VMAE - (Attention Fusion)	**93.45 ± 2.13**	**92.74 ± 1.64**

able training instances and the distinctiveness of the action. To illustrate, for the *left and right turn* we have 1159 and 1124 video clips respectively. For the *U-turn, railroad passing, right lane branch* maneuvers, we have 66, 71, and 96 clips, respectively, in the training set.

Table 2 shows a comparison of the Brain4Cars performance of multiple end-to-end video based DIR approaches. The SSVP VMAE architecture that uses attention fusion to combine both video streams achieves the highest accuracy of 93.45%, and an F1 score of 92.74%. Both Gebert et al. [16] and Rong et al [35] also fine-tuned a Kinetics-400 pre-trained model, whereas Ma et al. [27] used an Imagenet pre-trained backbone. When only the external video data is used to recognize the driving intentions, we observe that the VideoMAE achieves an accuracy of 86.50%, whereas the previous SOTA achieved an accuracy of 64.75%. When recognizing the intentions of a driver solely based on the in-cabin videos, we observe no clear performance difference between the VMAE model and the results from Gebert et al. [16] or Ma et al. [27].

Table 3. Results of the Monte-Carlo (MC) sampling (N=25) for the MC dropout and Bayes-by-Backprop (BBB) – Last Layer probabilistic deep learning models. Drop rate probabilities (p) range from 0.05 – 0.25.

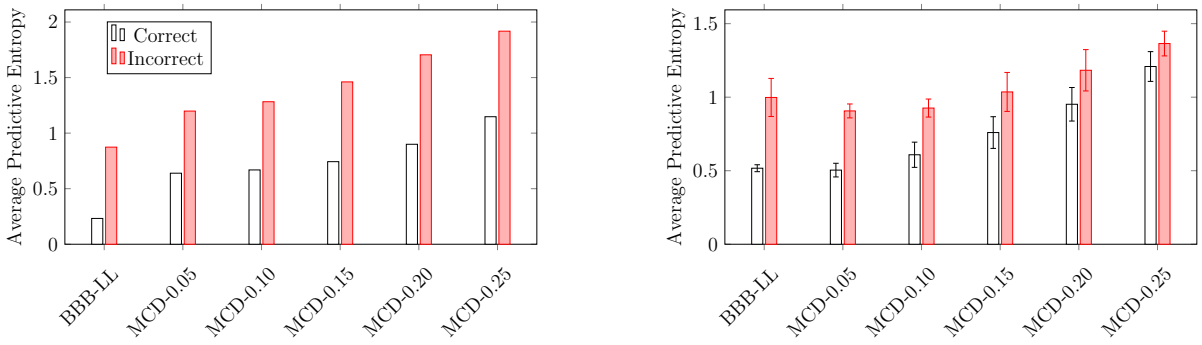
Method		HDD	Brain4Cars	
		mAP (↑)	Acc (↑)	F1 (↑)
Deterministic		**85.6**	**93.45 ± 2.13**	**92.74 ± 1.64**
BBB – Last Layer		83.4	91.71 ± 3.66	90.70 ± 3.34
MC dropout	p=0.05	85.0	90.83 ± 3.55	90.67 ± 3.18
	p=0.10	83.0	86.44 ± 3.87	87.78 ± 2.96
	p=0.15	78.5	78.80 ± 6.94	80.36 ± 5.33
	p=0.20	71.2	67.36 ± 9.69	69.11 ± 7.15
	p=0.25	62.9	55.64 ± 12.43	56.42 ± 7.89

Table 4. Comparison of different fusion methods for the fine-tuned VMAE models for the Brain4Cars dataset using both the in-cabin and external video streams.

Data Source	Fusion Method	Acc (↑)	F1 (↑)
In-cabin & External	Attention	**93.45 ± 2.13**	**92.74 ± 1.64**
	Concatenation	92.87 ± 3.12	91.68 ± 1.91
	Late - averaging	89.10 ± 2.43	89.07 ± 2.46

4.4. Uncertainty estimation analysis

Figure 2 and Table 3 present the test results for the probabilistic DL methods. From a performance perspective, the BBB-LL and MC dropout with a drop rate of 0.05 are slightly worse than the deterministic approach for both datasets. The performance of MC dropout models gradually decreases for higher dropout rates. Figure 2 shows the average predicted entropy for correctly and incorrectly classified test instances for both datasets. The BBB-LL and MC dropout produce lower uncertainty estimates for the correctly classified test instances compared to the incorrectly classified instances (for DIR only with a drop rate of 0.05 and 0.10). For the MC dropout models with higher drop rates, the average predictive entropy between the groups becomes more similar for the Brain4Cars results, but not for the HDD test instances.



(a) HDD

(b) Brain4Cars (five folds)

Figure 2. The average predictive entropy for correctly and incorrectly classified test instances based on Monte-Carlo sampling (N=25) for the Bayes-by-Backprop last layer (BBB-LL) model and Monte-Carlo dropout (MCD) with different drop rates.

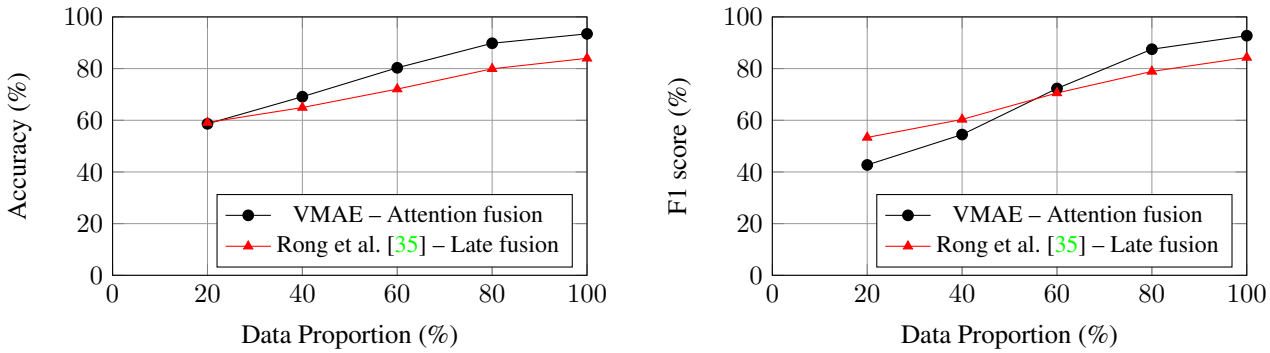


Figure 3. Performance comparison of driver intention recognition for different input time windows. Results are the average performance over five folds and based on both the in-cabin and external video stream.

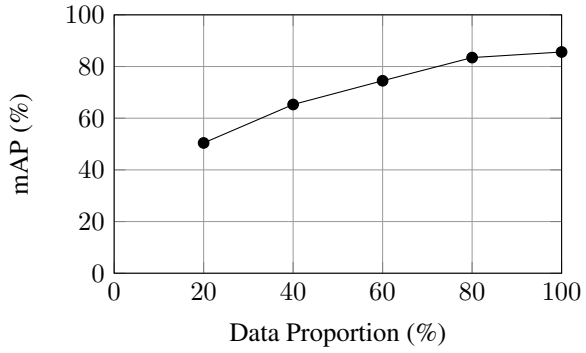


Figure 4. Driver action recognition performance for different input time windows.

4.5. Ablation study

Impact of fusion operations. Several strategies can be used to combine the learned video representations to perform DIR. To understand the impact of the fusion operations, we compare a late fusion and a concatenation of the intermediate video embeddings to the attention fusion approach. Table 4 shows a performance overview of the fusion strategies on the Brain4Cars DIR performance. The intermediate attention fusion yields the best average result for the five fold evaluation.

Prediction performance over time. Both DAR and DIR benefit from quick accurate recognition to allow more time to evaluate and safely anticipate on intended driving maneuvers based on the current traffic situation. Therefore, we examine the performance of the VMAEs over time. Figures 3 and 4 show the performance when the video footage is shortened. The original videos are shortened by steps of 20%, but in the case of a short video clip, we always make sure there is at least one frame to predict on.

For the DIR evaluation over time in Figure 3, we also include the results for different time periods from Rong et al. [35]. While it is not a perfect comparison, we do observe that the F1 score when predicting four seconds (20%) or three seconds (40%) ahead of the intention execution is lower for the VMAEs attention fusion model. Similar to Rong's observations, we see the highest increase in DIR performance after using more than 40% of the input sequence length.

Figure 4 shows the benefits of the offline classification setting that we used for recognizing the driving actions. The online DAR setting, where an action recognition method infers an action label for every new incoming video frame, is much harder because it is difficult to learn temporal information from a single or a few frames [49].

5. Discussion and conclusion

Employing AI in safety-critical environments including human behavior requires a cautious strategy. In this paper, we demonstrate state-of-the-art (SOTA) performance for end-to-end driver action and intention recognition based on raw-videos. We observe that without extracting additional features, such as optical flow, self supervised video pre-trained (SSVP) video masked autoencoders (VMAEs) outperform both offline driver action recognition (DAR) SOTA approaches for all except one driving action, and existing end-to-end driver intention recognition (DIR) methods for single and multi-video data. A true comparison to Ma et al. [27] for the Brain4Cars dataset is tricky, because they used a backbone that was pre-trained on another dataset, used a different loss function, and applied different data augmentations. Similarly for the HDD dataset, Noguchi and Tanizawa (2023) [30] used a graph-based framework, which relies on first detecting and tracking road users, whereas our setup learns from end-to-end raw video footage and employs different data augmentations.

The VMAE end-to-end video recognition setup employs a cold cognition approach [6], which means that no explicit form of reasoning is included, instead it is assumed to be learned implicitly [32]. The problem with an implicit learned form of reasoning for a safety-critical AI application that is used to estimate human behavior, is that it is difficult to verify the learned reasoning. Therefore, it would be beneficial to extend the approach with a form of explicit reasoning that allows for inspecting how the model produces a prediction.

All probabilistic variations of the VMAE perform slightly worse compared to the deterministic approach, but we do observe a difference in uncertainty estimations between correctly and incorrectly predicted test instances for the Bayes-by-Backprop last layer (BBB-LL) and Monte-Carlo (MC) dropout models (only the MC dropout models with a drop rate of 0.05 and 0.10 for DIR). However, the probabilistic models require MC sampling, which inherently increases computations and makes the real-time use impractical.

Moreover, we analysed the effects of different information fusion strategies and observed that an attention fusion of the in-cabin and external video embeddings yielded the highest overall DIR performance. Lastly, we showed the performance effects on both DAR and DIR when using less input data. For DAR this highlights the challenge of action recognition in an online setting, whereas for DIR it showed a similar trend compared previous work [35].

References

[1] Arash Hosseinian Ahangarnejad, Ahmad Radmehr, and Mehdi Ahmadian. A review of vehicle active safety control methods: From antilock brakes to semiautonomy. *Journal of Vibration and Control*, 27(15-16):1683–1712, 2021. 1

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

[3] Abdelmoudjib Benterki, Moussa Boukhnifer, Vincent Judalet, and Choubeila Maaoui. Artificial intelligence for vehicle behavior anticipation: Hybrid approach based on maneuver classification and trajectory prediction. *IEEE Access*, 8:56992–57002, 2020. 1

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[6] Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. Knowing me, knowing you: theory of mind in AI. *Psychological medicine*, 50(7):1057–1061, 2020. 7

[7] Michael C Darling. Using uncertainty to interpret supervised machine learning predictions. 2019. 2

[8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural safety*, 31(2):105–112, 2009. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2, 4

[11] Ronald Fisa, Mwiche Musukuma, Mutale Sampa, Patrick Musonda, and Taryn Young. Effects of interventions for preventing road traffic crashes: an overview of systematic reviews. *BMC public health*, 22(1):513, 2022. 1

[12] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. 1

[13] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3):4, 2016. 3

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 2

[15] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 2

[16] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 969–974. IEEE, 2019. 1, 2, 3, 4, 5

[17] Yingshi Guo, Hongjia Zhang, Chang Wang, Qinyu Sun, and Wanmin Li. Driver lane change intention recognition in the connected environment. *Physica A: Statistical Mechanics and its Applications*, 575:126057, 2021. 1

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[19] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021. 1

[20] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4193–4202, 2017. 3

[21] Ashesh Jain, Hema S Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. 2016. 1, 2, 4

[22] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022. 2

[23] Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. Bayesian-torch: Bayesian neural network layers for uncertainty estimation. https://github.com/IntelLabs/bayesian-torch, Jan. 2022. 4

[24] Chengxi Li, Yue Meng, Stanley H Chan, and Yi-Ting Chen. Learning 3d-aware egocentric spatial-temporal interaction via graph convolutional networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8418–8424. IEEE, 2020. 4, 5

[25] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2018. 4

[27] Yunsheng Ma, Wenqian Ye, Xu Cao, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, and Ziran Wang. CEMFormer: Learning to predict driver intentions from in-cabin and external cameras via spatial-temporal transformers. *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) (Upcoming)*, 2023. 2, 5, 7

[28] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019. 2

[29] Osama Makansi, Özgün Çiçek, Yassine Marrakchi, and Thomas Brox. On exposing the challenging long tail in future prediction of traffic actors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13147–13157, 2021. 1

[30] Chihiro Noguchi and Toshihiro Tanizawa. Ego-vehicle action recognition based on semi-supervised contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5988–5998, 2023. 1, 2, 5, 7

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[32] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227. PMLR, 2018. 7

[33] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017. 3

[34] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1, 2, 4

[35] Yao Rong, Zeynep Akata, and Enkelejda Kasneci. Driver intention anticipation based on in-cabin and driving scene monitoring. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2020. 1, 2, 3, 4, 5, 6, 7

[36] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 1

[37] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[38] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018. 3

[39] Thomas Wolf Philipp Schmid Zachary Mueller Sourab Mangrulkar Marc Sun Benjamin Bossan Sylvain Gugger, Lysandre Debut. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022. 4

[40] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 1, 3, 4

[41] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. In *European Conference on Computer Vision*, pages 497–515. Springer, 2022. 3

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5

[43] Duy Tran, Weihua Sheng, Li Liu, and Meiqin Liu. A hidden markov model based driver intention prediction system. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 115–120. IEEE, 2015. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing systems*, pages 5998–6008, 2017. 2

[45] Koen Vellenga, H. Joe Steinhauer, Alexander Karlsson, Göran Falkman, Asli Rhodin, and Ashok Chaitanya Koppisetty. Driver intention recognition: state-of-the-art review. *IEEE Open Journal of Intelligent Transportation Systems*, 2022. 1, 2

[46] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575, 2021. 1, 2

[47] Joe Watson, Jihao Andreas Lin, Pascal Klink, Joni Pajarinen, and Jan Peters. Latent Derivative Bayesian Last Layer Networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1198–1206. PMLR, 2021. 4

[48] WHO. Global health estimates: Leading causes of death, 2020. 1

[49] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *Proceedings of the IEEE International Conference on Computer V*. 1, 7