Title
Predicting Driver Attention in Critical Situations

Authors
Xia, Ye
Zhang, Danqing
Kim, Jinkyu
et al.

Publication Date
2019

DOI
10.1007/978-3-030-20873-8_42

Peer reviewed

Training a network to attend like human drivers saves it from common but misleading loss functions

Ye Xia[1], Danqing Zhang[1], Alexei Pozdnukhov[1], Ken Nakayama[1, 2], Karl Zipser[1], and David Whitney[1]

[1] *University of California, Berkeley*
[2] *Harvard University*
{yexia,danqing0703,alexeip,karlzipser,dwhitney}@berkeley.edu},nakayama@g.harvard.edu

November 20, 2017

Abstract

We proposed a novel FCN-ConvLSTM model to predict multi-focal human driver's attention merely from monocular dash camera videos. Our model has surpassed the state-of-the-art performance and demonstrated sophisticated behaviors such as watching out for a driver exiting from a parked car. In addition, we have demonstrated a surprising paradox: fine-tuning AlexNet on a largescale driving dataset degraded its ability to register pedestrians. This is due to the fact that the commonly practiced training paradigm has failed to reflect the different importance levels of the frames of the driving video datasets. As a solution, we propose to unequally sample the learning frames at appropriate probabilities and introduced a way of using human gaze to determine the sampling weights. We demonstrated the effectiveness of this proposal in human driver attention prediction, which we believe can also be generalized to other driving-related machine learning tasks.

1 Introduction

It is amazing how human visual attention enables drivers to locate, within a small fraction of a second, potential risks across the visual field, such as a darting-out pedestrian, a changing traffic light or a likely incursion of a nearby cyclist. Even though in recent years we have seen exciting progress in a variety of approaches to autonomous driving, no machine agent has even been close to the ability of human visual attention in understanding which parts of the scene may bring risks or demand a control action, not to mention the fact that human attention can achieve this with visual input alone. Most recently, a few Deep Neural Networks end-to-end trained on large datasets have shown the ability to predict car ego-motion or control signals from visual input alone [3, 27, 12, 9]. But, the rules and features of these Deep Neural Networks are not transparent, so it is uncertain which level of understanding of the visual scene these models have achieved. Other Deep Neural Networks have been trained to predict human driver attention [21, 20, 25]. However, these models were trained based on a dataset of an individual driver's eye movements. This is limiting because an individual driver can only gaze at one location at a time, due to biological constraints. Yet, there are usually multiple objects that

simultaneously need to be attended in a given scene. As a consequence, these previous models incorrectly predict only one attention focus for a scene, which usually only covers the center of the road or the car immediately ahead, but not many other important objects.

Compared to learning where to attend like human drivers, the importance of learning when to attend like human drivers is less acknowledged. Here, by learning when to attend like human drivers, we mean acknowledging, like human drivers, the fact that not every moment during driving is equally important. Human drivers are not constantly attentive during driving because they know that in many cases it may not matter much whether one drives a little faster or slower, whereas, at some crucial moments a small mistake may have fatal consequences [24]. In contrast, driving-related Deep Neural Networks are commonly trained to minimize the prediction error equally averaged over the frames/examples of the dataset. The problem, though, is that safe driving is not about how well you do on average.

In this context, we collected human eye movement data and generated multi-focal human gaze maps for a large number of crowd-sourced driving videos covering a large variety of situations. We used gaze maps as a proxy for human attention maps, as the gaze is tightly (though not always) linked to attended locations [23]. For consistency with prior literature in computer vision[21, 20, 25], which uses the phrase "attention map," we will use "attention map" and "gaze map" interchangeably in this paper. We have also built a model that can predict where human drivers would look. The model has surpassed the state-of-the-art performance ([20]) and showed sophisticated behaviors like watching out for a driver who is exiting from a parked car. The fact that our model can predict human drivers' attention maps suggests that our model learned some high-level understanding of the scene in the context of driving. Our model may not only have applications in assistant driving systems that warn drivers of situations demanding caution, but also in autonomous driving where understanding scenes within a driving specific context is greatly useful.

During our training experiments, we also demonstrated an important paradox: fine-tuning AlexNet on a large dataset of driving videos, under the typically used training paradigm of minimizing average loss, degraded its ability to register pedestrians. Here, we propose a new paradigm of training the model with weighted sampling. We demonstrated a way of using the human gaze data to determine appropriate sampling weights for the training frames. Our model, trained with weighted sampling, showed better performance than our model trained under the typical paradigm. We suggest that the performance degradation due to the typically-used equally weighted training is a concern for driving-related machine learning in general, and we believe the effectiveness of weighted sampling for human attention prediction can be generalized to other driving-related learning tasks.

2 Literature Review

Saliency prediction is a task that estimates the probability that a certain location in a scene will attract human visual attention, and is therefore highly related to our work of predicting human driver attention. In the literature on image saliency prediction using Deep Learning, models built upon pretrained AlexNet and VGG Net are the mainstream [15, 16]. As for video saliency prediction, LSTM- or ConvLSTM-based models are popular and effective [2, 8, 19]. These models have achieved the start-of-the-art performance on visual saliency benchmarks collected mainly when human subjects were doing the free-viewing task. It has been shown that human attention pattern greatly depends on the task [17]. Therefore, models that are specifically trained for predicting the attention of drivers are still needed.

As for driver attention prediction, several deep networks have been proposed to predict where human drivers would look. [1, 21, 20]. These existing models were trained on eye movement

data of individual drivers. Since a driver can only look at one place at a time, these models tend to predict one single attention focus instead of a multi-focal attention map, which is unrealistic [5]. Also, the driving video dataset that these models used consisted of just 74 trials of driving. Even if the total number of frames in the dataset is large, the richness and diversity of the data would still be limited, which may strongly affect the generalizability of their models.

Existing autonomous driving systems can be classified into two major thrusts: mediated perception approaches and end-to-end training approaches. For the mediated approach [26, 4, 18], the task is decomposed into more tractable sub-tasks of recognizing driving-relevant objects, such as road lanes, traffic signs, pedestrians, etc. More recently, there is growing interest in deep learning for the end-to-end training of self-driving cars [27, 3, 9, 6]. These models have demonstrated impressive performance in predicting vehicle ego-motion or control signals from video images. However, since their models are not transparent, it is not certain what features these model have learned. More importantly, these models were trained to minimize the average prediction error equally over the frames of their dataset, which does not reflect the varying level of importance in each of these frames. In real driving, the moments where a small error may have severe consequences are relatively rare, and therefore also rare in the datasets. The models that showed very good performance on average over the frames of the testing dataset [13, 27, 3, 9, 6] may actually make fatal errors in those relatively rare and crucial frames.

One may note that studies of importance sampling for Deep Neural Network training [11, 10] sound very similar to our proposed weighted sampling training paradigm. In fact, these studies address a problem that is substantially different from ours. They try to speed up the learning of the model, so they oversample the examples where the model is make large prediction errors. We try to solve the problem that in the context of driving the prediction error may not reflect the actual cost of making that error, so we seek for appropriate sampling weights that correlate with the cost.

3 Data collection

Instead of collecting eye movement from a driver in car during driving, we collected drivers' eye movement data in lab by having them watch recorded driving videos while doing a driving instructor task. The main motivation of this experiment design is that we can collect eye movement data from multiple drivers for the same scene. By aggregating multiple drivers' eye movement data, we can wash out noises and get a multi-focus attention map for each frame since different drivers won't happen to look at the same location at the very same moment. In addition, the accuracy of eye trackers in lab is much higher than wearable eye trackers operating in cars. Some details of our data collection experiment are discussed below.

45 observers participated in the experiment. All observers reported normal or corrected-to-normal vision. All observers had more than one year of driving experience. Eye movements were recorded with an EyeLink 1000 desktop-mounted infrared eye tracker (SR Research Ltd., Mississauga, Ontario, Canada), used in conjunction with the Eyelink Toolbox scripts for MATLAB [7]. The right eye was recorded for each subject at 1000 Hz. Observers were calibrated with a standard nine-point calibration procedure before completing each block of trials (average error $< 0.5°$).

Vehicle dashboard camera videos were selected from the Berkeley DeepDrive Video Dataset. To collect rich eye movement data as efficiently as possible, we chose videos in busy areas that included at least one braking response. We first selected videos during which the average speed of the car was slower than 10 mile/hour. We then located braking events within the videos, defined as vehicle decelerations greater than 1 m/s^2. Selected videos included 6.5 seconds

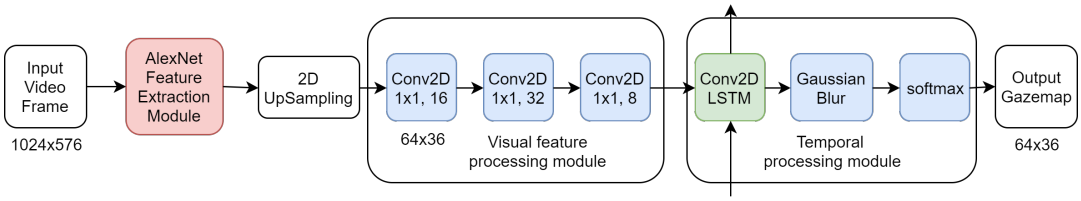

Figure 1: FCN-ConvLSTM structure of our model

prior to each braking event and 3.5 seconds after the braking event. 1,232 videos in total were collected following these procedures. The frame rate of the stimulus videos was 29 fps.

During the experiment, the display resolution of the videos was set to 1024×768 pixels, and the refresh rate to 60 Hz. Observers were asked to imagine that they were driving instructors sitting in the copilot seat and they were supposed to verify that their student driver was driving the car properly. They were asked to press the space key on the keyboard whenever they thought they needed to warn or correct the student driver. Each observer completed 200 trials divided into 3 blocks (60, 70, and 70 trials). Each block lasted 15 - 20 minutes.

The collected gaze data had a spatial resolution of 1024×768 pixels and was temporally downsampled to a 29 fps to match the frame rate of the videos. Each stimulus video was viewed by at least 4 observers. The gaze patterns made by these independent observers were aggregated and smoothed to make an attention map for each frame of the stimulus video. The smoothing was done by a spatial-temporal Gaussian kernel. The sigma of the Gaussian kernel in the x and y dimensions was 25 pixels and the sigma in the temporal dimension was 4 frames.

4 Model

Our goal is to predict attention map for a video frame given the current and previous video frames. Our model structure can be divided into a visual feature extraction module, a visual feature processing module, and a temporal processing modules (Figure 1).

4.1 Visual feature extraction module

The visual feature extraction module is a pre-trained dilated fully convolutional neural network and its weights are fixed during training. We used ImageNet pre-trained AlexNet [14] as our visual feature extraction module. We chose to follow [27]'s design to use the features from conv5 layer. One reason of this choice was to facilitate the comparison between using ImageNet pre-trained AlexNet and using AlexNet fine-tuned by [27], which will be discussed in details in the experiment section. In our experiment, the size of the input was set to 1024×576 pixels and the feature map by AlexNet was upsampled to 64×36 pixels and then fed to the following visual feature processing module.

4.2 Visual feature processing module

The visual feature processing module is a fully convolutional neural network. It consists of three convolutional layers with 1×1 kernels and a dropout layer after each convolutional layer. It further processes the visual features from the previous extraction module and reduces the dimensionality of the visual features from 256 to 8. In our experiments we observed that without the dropout layers the model easily got stuck in a suboptimal solution which was simply

Table 1: Summary of the settings of the three experiments and the names of the corresponding versions of our model

	visual feature extraction module	training paradigm	version name
Experiment 1	ImageNet pretrained AlexNet	regular	default
Experiment 2	Alexnet fine-tuned on BDDV	regular	fine-tuned
Experiment 3	ImageNet pretrained AlexNet	weighted sampling	weighted sampling

predicting a central bias map, i.e. an attention map concentrated in a small area around the central of the frame.

4.3 Temporal processing module

The temporal processor is a convolutional LSTM network with a kernel size of 3×3 follwed by a gaussian smooth layer (sigma set to 1.5) and a softmax layer. It receives the visual features of successive video frames in sequence from the visual feature processing module and predicts an attention map for every new time step. Dropout is used for both the linear transformation of the inputs and the linear transformation of the recurrent states. We had also experimented with using an LSTM network for this module and observed that the model tended to incorrectly attend only to the central region of the video frames.

5 Experiments

We trained our model in different settings in three experiments. In experiment 1 we trained our model with a regular regime to test the default performance of our model and compare it with a baseline model and a state-of-the-art model. We will refer to this version of our model as the default version in the following. In experiment 2 we used AlexNet fine-tuned by [27] instead of AlexNet pre-trained on ImageNet as our visual feature extraction module. [27] fine-tuned AlexNet on a large driving video dataset with respect to the task of predicting the ego-motion of the car. Intuitively, compared with ImageNet pre-trained AlexNet, this AlexNet fine-tuned with respect to ego-motion prediction task (which we will simply refer to as the fine-tuned AlexNet in the following) should extract information that is more relevant to driving from the video, and it should therefore make our model predict human drivers' attention better. A comparison between the results of experiment 1 and experiment 2 will directly test this hypothesis. In experiment 3 we kept the same visual feature extraction module as experiment 1 (i.e. ImageNet pre-trained Alexnet) and trained our model under our proposed weighted sampling training paradigm. The goal was to test the effectiveness of weighted sampling. We will refer to this version of our model as the weighted sampling version in the following. The settings of these three experiments and the version names are summarized in Table 1.

In order to guarantee fair comparisons, except for the factors described above, the other training settings and parameters were the same for the three experiments. We used 926 training videos and 306 testing videos. Most of the videos were 10 seconds long and a few of them were longer. We downsampled the videos to 1024×576 pixels and 3Hz. After this preprocessing, we had about 30k frames in our training set and 10k frames in our testing set. We used cross-entropy between predicted attention maps and human attention maps as the training loss, along with Adam optimizer (learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$). Each training batch contained 10 sequences and each sequence had 6 frames. The training was done for 10,000 iterations. All the experiments showed stabilized testing errors by iteration 10,000.

Table 2: Test results obtained by the baseline model, the state-of-the-art model (Dreyeve) and the different versions of our model

Model	mean DKL	std. error
Baseline	1.495	0.007
Dreyeve	1.95	0.01
Ours (default)	1.274	0.007
Ours (fine-tuned)	1.300	0.007
Ours (weighted sampling)	1.436	0.008

5.1 Experiment 1: Performance of the default version of our model

To our knowledge, [21] and [25] are the two Deep Learning models that use dash camera videos alone to predict human driver's gaze. They demonstrated similar results and were shown to surpass other deep learning models or traditional models that predict human gaze in non-driving-specific contexts. We chose to replicate [21] to compare with our work and because their code is publicl. The model designed by [21] was trained on the Dreyeve dataset [1]. We will refer to [21]'s model as Dreyeve model in the following. We tested their pre-trained model directly on our testing dataset without any training on our training dataset. The videos of Dreyeve dataset were collected by a roof-mounted camera in urban environments similar to where the videos of our dataset were collected. Since the goal of the comparison was to test the effectiveness of the combination of model structure, training data and training paradigm as a whole, we think it is reasonable to test how well Dreyeve model performs on our dataset without further training.

We first used Kullback-Leibler divergence (KL divergence, D_{KL}) to quantify how well the predicted attention maps matched human attention maps and compared the results between our model, Dreyeve model and a baseline model. The baseline model always predicts the averaged human attention map of training videos. The default version of our model demonstrated significantly smaller testing error than both the baseline model and the Dreyeve model in terms of KL divergence (Table 2). Our model reduced the average testing error by 15% with respect to the baseline model and by 35% with respect to the Dreyeve model.

As we discussed above, the average testing error should not be the only criterion for evaluating driving related models. Human drivers need to look at many objects while driving, e.g. the cars in front, the traffic lights, the pedestrians that may enter the roadway, etc. Therefore, it is important to evaluate whether a human attention prediction model can successfully attend to these objects. We used a state-of-the-art object detection algorithm ([22]) to detect the objects in the videos of our dataset. A confidence threshold of 0.45 was used for detecting the objects. The three categories that had the most detected instances in our videos were car, traffic light, and person. Looking at these three types of objects is often crucial for driving. For each of these three categories, we calculated the proportion of all the detected objects of that category that were actually attended to by humans versus the models. An object was defined as being attended if more than 20% of the area of its bounding box was an attended region. The attended regions were defined as the regions where the probability density assigned by the attention map was at least 5 times the probability density of a uniform distribution.

The attended proportions of these three object categories were calculated for human attention, the baseline model, the Dreyeve model, and the default version our model (Figure 2). Human attention included cars, traffic lights and people at levels of 33%, 24% and 35%, respectively. The baseline model attended to traffic lights and pedestrians substantially less than humans. The baseline model attended to cars because the car immediately ahead usually appeared at the center of the frame. The Dreyeve model showed a similar proportion as human attention

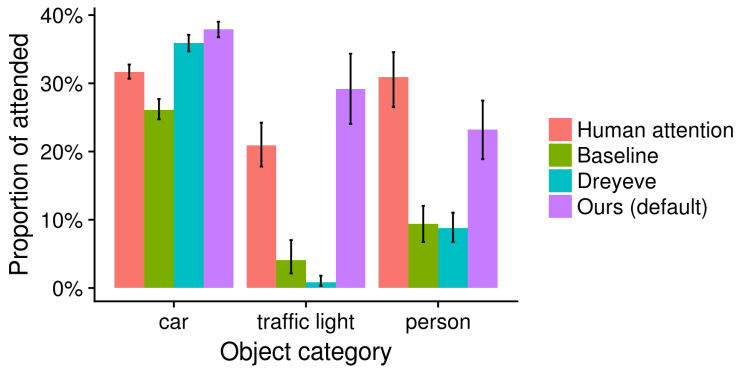

Figure 2: Proportions of attended objects of different categories for human attention and different models

for cars, but for traffic lights and people it was comparable or worse than the baseline model. This statistical result matched our empirical observation that the attention maps predicted by the Dreyeve model were biased toward the center of the road, or the car immediately ahead, but largely ignored other important objects. In contrast, the default version of our model was significantly better than both the baseline model and the Dreyeve model at predicting attending to traffic lights and people, and showed a performance comparable to human attention.

Importantly, the default version our model did not simply select objects like an object detection algorithm. Our model selectively attended to the pedestrians that were also attended to by humans. Let us refer to the pedestrian that were actually attended to by humans as the important pedestrians and the rest of them as non-important pedestrians. Among all the pedestrians detected by the object detection algorithm, the proportion of important pedestrians was 33%. If our model was simply detecting pedestrians at a certain level and could not distinguish between important pedestrians and non-important pedestrians, the proportion of important pedestrians among the pedestrians attended to by our model would also be 33%. However, the actual proportion of important pedestrians that our model attended to was 56% with a bootstrapped 95% confidence interval of $[47\%, 65\%]$. Thus, the default version of our model predicts which of the pedestrians are the ones most relevant to human drivers.

Some concrete examples are shown in Figure 3 to demonstrate that our model could pay more attention to the people that were actually crucial for the particular driving situation than to other non-important pedestrians. In the example in row 1 of Figure 3, the pedestrian highlighted by the red rectangle was about to cross the road. The pedestrian was on a phone and not looking at the driver. So it was crucial to be aware of that pedestrian and stop the car even if the driver had the green light. Our model attended to that pedestrian, but did not attend to the other pedestrian more to the right who was larger in the video frame but not crucial for the driver's decision. In the example in row 2, the driver had a yellow light and some pedestrians (highlighted by the red rectangle) were about to enter the roadway. Our model attended to those pedestrians, who informed the driver whether to proceed through the yellow light; our model did not attend to the other (less important) pedestrian. In the example in row 3, our model successfully attended to a cyclist who was very close to the car.

To go one step further, when human drivers need to look at a person during driving, do they look at different parts of the body with equal probabilities or do they look at a particular part selectively? If there was a particular pattern, would our model be able to learn the same pattern? In order to test this, we intended to calculate the average attention distribution within the

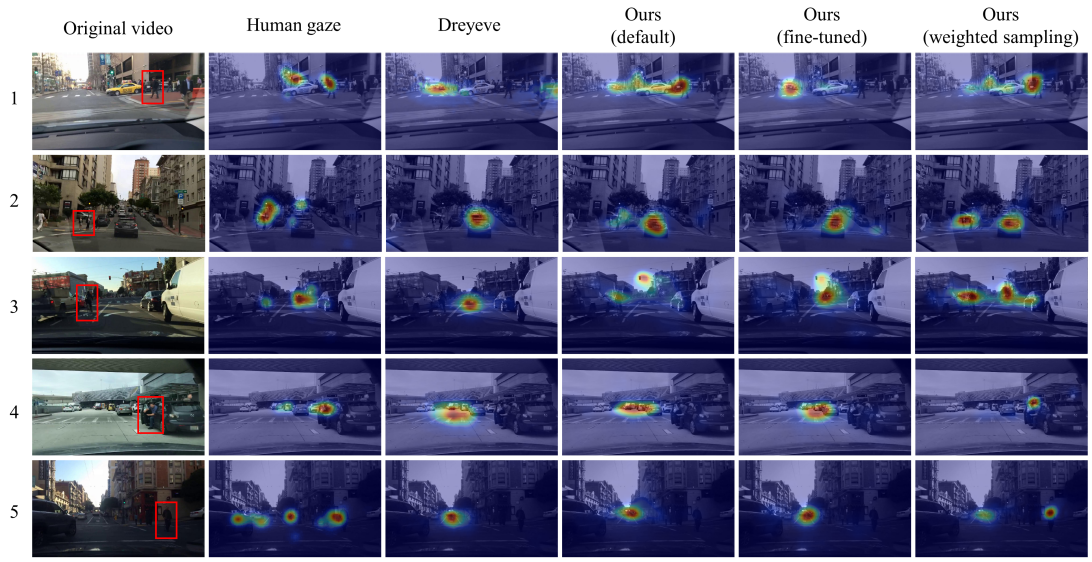

Figure 3: Examples of the videos in our dataset, ground-truth human attention maps and the prediction of different models. The red rectangles in the original video column highlight the pedestrians that pose a potential hazard. A successful driver attention prediction should cover the highlighted pedestrians but not the other pedestrians. Row 1: a pedestrian was on a phone and about to cross the road without noticing the car. Row 2: some pedestrians were about to enter the roadway when the driver had a yellow light. Row 3: a cyclist very close to the car. Row 4: a driver suddenly exited from a parked car. Row 5: A pedestrian was crossing the road when the driver was making a right turn. Faces are blurred in this display but were not blurred during training and testing.

bounding box of a person. To do this, we had to control any biases in the spatial distribution of gaze and in the distribution of pedestrians across the visual field. For example, observers might tend to gaze toward the center of the visual field and pedestrians are not uniformly distributed over space, and the coupling of these two biases could generate an artificial pattern in the average attention distribution within the bounding box of a pedestrian. We therefore modeled the attention probability at a certain pixel in a certain frame using the following formulas in order to account for the potential gaze and pedestrian biases:

$$p_i(\vec{x}) = q_i(\vec{x}) p_c(\vec{x} - \vec{x}_{ci})$$

where $p_i(\vec{x})$ is the attention probability at location $\vec{x}$ in the ith frame, $\vec{x}_{ci}$ is the location of the road center in the ith frame, $p_c(\vec{x})$ is the average attention probability at location $\vec{x}$ relative to the road center when no object is present at that location and $q_i(\vec{x})$ is an attention ratio between $p_i(\vec{x})$ and $p_c(\vec{x} - \vec{x}_{ci})$ quantifying the contribution of the specific visual input of the ith frame at location $\vec{x}$ after accounting for the road-center bias.

We calculated $p_c(\vec{x})$ using the following formula:

$$p_c(\vec{x}) = \frac{\sum_i p_i(\vec{x} + \vec{x}_{ci}) \delta_{i,\vec{x}+\vec{x}_{ci}}}{\sum_i \delta_{i,\vec{x}+\vec{x}_{ci}}}$$

where $\delta_{i,\vec{x}+\vec{x}_{ci}} = 1$ if the pixel $\vec{x} + \vec{x}_{ci}$ of the ith frame is not in the bounding box of any detected object, and $\delta_{i,\vec{x}+\vec{x}_{ci}} = 0$ otherwise.

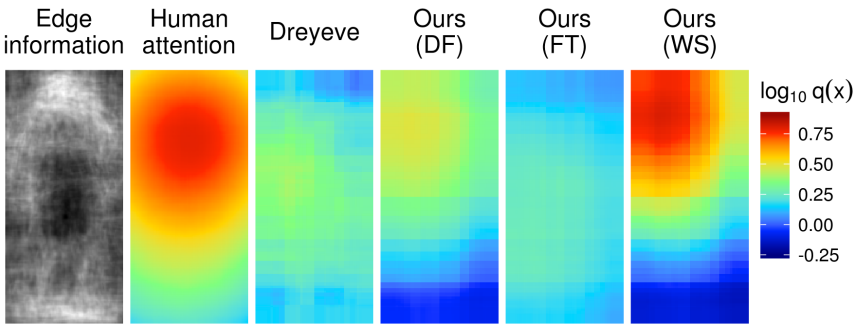

Figure 4: Average edge information and mean common logarithm attention ratio distributions over a human body for huamn attention, the state-of-the-art model (Dreyeve), the defaut version of our model (DF), the fine-tuned version of our model (FT) adn the weighted sampling version of our model (WS)

With this setup, $q_i(\vec{x})$, which we will refer to as the attention ratio in the following discussion, is a measure of how much the specific visual content at a particular location of a particular frame changes the attention probability with respect to the attention probability that location would have merely according to the road-center gaze bias. We then transformed the human attention map of each frame into its common logarithm attention ratio map ($\log_{10} q_i(\vec{x})$), took out the patches that corresponded to the bounding boxes of attended people, stretched them to a common size of 100 px $\times$ 214 px and averaged them. That gave an average common logarithm attention ratio distribution within the bounding box of a person (Figure 4). This distribution showed a single significant peak at the upper part of the body of a person. We also averaged the edge information of the camera image patches that corresponded to the bounding boxes of attended people in the same manner and the averaged edge information revealed a rough outline of a person (Figure 4). By comparing the averaged edge information with the averaged common logarithm gaze ratio distribution, we concluded that the peak in the distribution corresponded to people's faces. The peak value was 0.79, which means the human attention probability at that part of a person bounding box was on average 6.2 ($10^{0.79}$) times the human attention probability at that location when the person was not present. This result suggested that human observers looked at a person's face much more than any other part of the body. This is reasonable, as human drivers would need to look at the faces of pedestrians to know their intentions [24].

In order to test whether the models could learn this high-level cue, we calculated the same common logarithm attention ratio distribution within a person based on the attention maps predicted by Dreyeve model and the default version of our model (Figure 4). The Dreyeve model seemed to slightly overweight the center of a body, whereas our model demonstrated a preference to faces similar to human gaze.

5.2 Experiment 2: Test our model with fine-tuned Alexnet

In the section above we showed that the default version of our model surpassed the baseline model and the state-of-the-art Dreyeve model in terms of testing loss and the ability to attend to important objects, especially pedestrians. However, the proportion of people attended to by our model was still significantly lower than the proportion of people attended to by human observers. Besides, even though our model also demonstrated a preference to faces, its preference was not as sharply tuned as the preference shown by human attention.

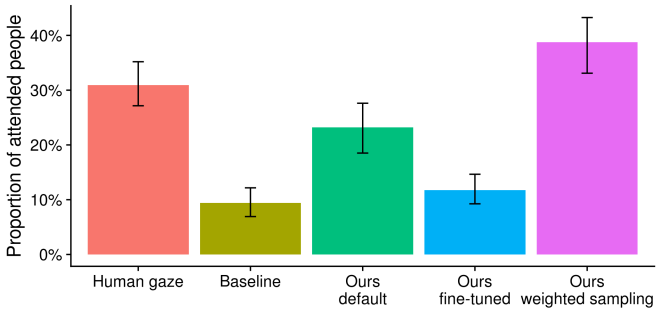

Figure 5: Proportions of attended pedestrians for human attention, the baseline model and the different versions of our model

The visual feature extraction module of our model is ImageNet pre-trained AlexNet for general object classification. Since the weights of this module were fixed in our training, it is reasonable to assume that replacing this module with a network trained for driving specific task might make our model predict human driver's attention better.

We chose to use AlexNet fine-tuned by Xu et el. [27]. They finetuned the AlexNet jointly with their whole model on the Berkeley DeepDrive Video dataset to predict the car ego-motion. We replaced our visual feature extraction module with the fine-tuned AlexNet and tested our model's performance while keeping all other factors the same as Experiment 1. Counter-intuitively, the result showed that using the fine-tuned AlexNet did not increase the rate at which our model attended to pedestrians and instead dropped it to the baseline model level by a decrease of 49% (Figure 5). Moreover, when using the fine-tuned AlexNet, our model was no longer able to learn the preference to faces (Figure 4). It is important to note that this performance decrement would not have been noticed if one had only looked at the average testing loss because this fine-tuned version of our model showed an average testing loss only barely larger than the default version of our model (Table 2).

ImageNet pre-trained AlexNet was able to support our model in learning to selectively attend to pedestrians and faces. But, after being fine-tuned on a large-scale driving video dataset to predict accurate car ego-motion, the same network could no longer support our model in attending to pedestrians or faces, even though it was supposed to have learned how to react to pedestrians for the sake of car egomotion prediction. We think this is because in real life we only need to avoid pedestrians (e.g., swerve, brake) rarely and this likelihood is honestly reflected in the dataset that the AlexNet was fine-tuned on. Since the training loss was prediction error averaged across the frames equally, the model was forced to compromise its reaction to pedestrians in order to perform better in most of the frames, presumably where the driver only needed to follow the lane or the car in front. It has been largely held like a golden standard in the field of machine learning that the dataset should truthfully mimic the statistics of the real world. But the question is, could the learning algorithms weight the learning examples unequally and appropriately, just like human beings do? This issue becomes especially problematic in the context of autonomous driving since by nature a small proportion of moments are much more important than the rest of the moments. Regardless of the small probability with which these moments occur, the price of not being able to make the right decision at those particular moments is not small, because once they happen any mistake may lead to a fatal cost. For this sake, we proposed a weighted sampling training paradigm that we describe in the following section.

5.3 Experiment 3: test our model with weighted sampling

In order to tackle the problem that more important learning examples are less common in the dataset, we proposed to sample the learning examples with unequal probabilities, which appropriately reflect their differing levels of importance. But, then the problem became how the appropriate sampling weights could be determined. It seems impossible that any state-of-the-art model can evaluate by itself the importance of driving situations since this evaluation would involve so much high-level and social information. Manually labelling the importance of all the frames of driving datasets also seems to be too costly given the huge scale of the datasets. Alternatively, we propose that human drivers/observers' gaze data, which can be collected naturally and at a relatively high speed, could be used to determine the importance of the frames of driving datasets.

In real driving situations, most of the time the driving is smooth and the driver only needs to look at the center of the road or the car in front. These moments are less dangerous and therefore less important compared to moments when the driver has to make an eye movement to look at a darting out pedestrian, a changing traffic light, an incursion of a nearby cyclist, etc. When we averaged the attention maps of one video, we could get a distribution concentrated at the center of the road. Therefore, the more different an attention map is from the average attention map of that video, the more important the corresponding training frame is. We used the KL divergence to measure the difference between the attention map of a particular frame and the averaged attention map of that video. This KL divergence determined the sampling weight of this video frame during training. The histogram of this KL divergence of all the training video frames is shown in Figure 6. As we expected, the histogram was strongly skewed to the left side. Our goal was to boost up the proportion of the frames of high KL divergence values by weighted sampling. The sampling weight was determined as a function of KL divergence (D_{KL}) illustrated in Figure 6 in panel B. The middle part of this function ($D_{KL} \in [1, 3]$) was set to be proportional to the inverse of the histogram so that after weighted sampling the histogram of KL divergence would become flat on this range. The left part of the function ($D_{KL} < 1$) was set to a low constant value so that those frames would be sampled occasionally but not completely excluded. The right part of the function was set to a saturated constant value instead of monotonically increasing values in order to avoid over-fitting the model to this small proportion of data. The histogram of KL divergence after weighted sampling is shown in Figure 6 in panel A. In our experiment we needed to sample the training frames in continuous sequences of 6 frames. For a particular sequence, its sampling weight was equal to the sum of the sampling weights of its member frames. These sequences were sampled at probabilities proportional to the sequence sampling weights.

The result showed that our model with weighted sampling predicted attention to pedestrians at a significantly higher chance than our model trained under the regular paradigm, and even slightly higher than human observers (Figure 5). Again, our model with weighted sampling selectively attended to the important pedestrians. While the baseline proportion of important pedestrians was 33%, the proportion of important pedestrians among the pedestrians this weighted sampling version of our model attended to was 48% with a bootstrapped 95% confidence interval of $[42\%, 55\%]$. Additionally, compared to the default version of our model, the weighted sampling version showed an even stronger face preference and matched the face preference of human attention at a quantitative level (Figure 4).

In the three examples discussed above, our model with weighted sampling assigned even more attention to the important people than our model trained under the regular paradigm. Our model with weighted sampling demonstrated some sophisticated behaviours. In the example in row 4 in Figure 3, our model with weighted sampling was the only one that successfully predicted attending to the person exiting from the parked car. Notice that our model with

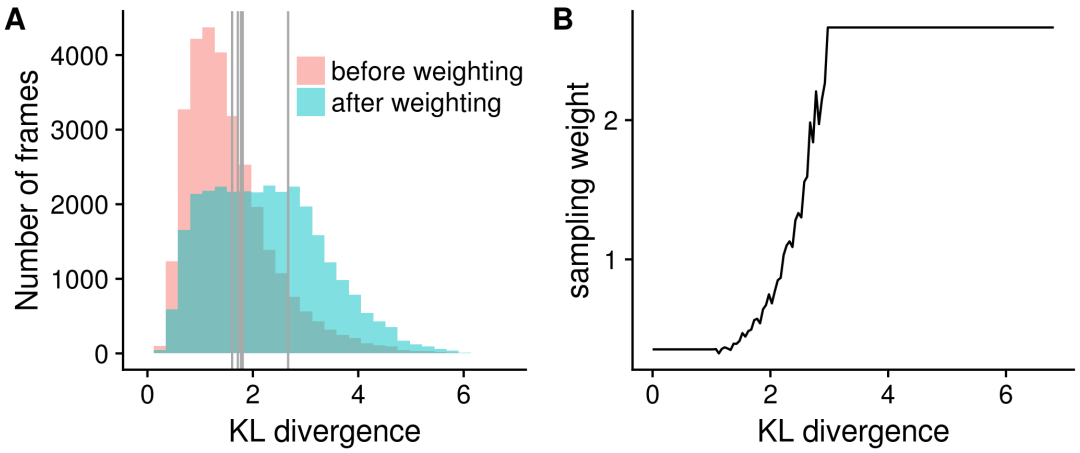

Figure 6: Histogram of KL divergence between individual human attention maps and video average attention maps and sampling weight as a function of the KL divergence value. Panel A: histograms of KL divergence between individual human attention maps of training video frames and their corresponding video average attention maps before and after weighted sampling. The gray vertical lines are showing the KL divergence values of the five example frames illustrated in Figure 3. Panel B: Normalized sampling weight as a function of the KL divergence value. A value of 1 for the normalized sampling weight means the video frame will be on average sampled 1 time when the training algorithm goes through as many training frames as the whole training set.

weighted sampling did not attend to the passenger who exited from the other side of the car. In the example in row 5, the driver was making a right turn and needed to yield to the crossing pedestrian. That pedestrian appeared in a quite peripheral area in that video frame. Our model with weighted sampling, was the only one that successfully overcame the central bias and was able attend to that pedestrian.

Despite its superior ability to attend to important pedestrians, this weighted sampling version of our model showed a higher mean testing error than the other two versions of our model (Table 2). This again suggested that a simple mean testing error that weights all the testing examples equally should be interpreted with caution in the context of driving related learning. Even though we only experimented with human driver attention prediction, we believe that the usefulness of human gaze data and the effectiveness of our weighted sampling paradigm can be generalized to other driving-related machine learning problems.

6 Conclusion

In this work, we have collected human eye movement data and generated multi-focal human attention maps for large-scale crowd-sourced driving videos, which will be made publicly available. We propose a novel FCN-ConvLSTM model that predicts human driver attention merely from monocular dash camera videos. Our model has surpassed the state-of-the-art performance and demonstrated sophisticated behaviors akin to human attention.

We have also demonstrated that fine-tuning AlexNet on a large scale driving dataset degraded its ability to register pedestrians, a counter-intuitive result that arises because typical training

paradigms treat all frames and all errors equally. As a solution, we have proposed a new weighted sampling training paradigm, which takes advantage of human gaze data to assign appropriate sampling weights to the learning frames. We demonstrated its effectiveness in the prediction of human driver attention and we believe it can also be generalized to other driving-related machine learning tasks. Using human gaze data as a rich form of annotation can fundamentally change the way in which deep learning can be applied to real world control problems, whereas ignoring the human factor may trap machine learning in an empty race to minimize ill-informed loss functions.

References

[1] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. "Dr (Eye) Ve: A dataset for attention-based tasks with applications to autonomous and assisted driving". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 54–60.

[2] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. "Recurrent mixture density network for spatiotemporal visual attention". In: *arXiv preprint arXiv:1603.08199* (2016).

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. "End to end learning for self-driving cars". In: *arXiv preprint arXiv:1604.07316* (2016).

[4] Martin Buehler, Karl Iagnemma, and Sanjiv Singh. *The DARPA urban challenge: autonomous vehicles in city traffic*. Vol. 56. springer, 2009.

[5] Patrick Cavanagh and George A Alvarez. "Tracking multiple targets with multifocal attention". In: *Trends in cognitive sciences* 9.7 (2005), pp. 349–354.

[6] Lu Chi and Yadong Mu. "Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues". In: *arXiv preprint arXiv:1708.03798* (2017).

[7] Frans W Cornelissen, Enno M Peters, and John Palmer. "The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox". In: *Behavior Research Methods* 34.4 (2002), pp. 613–617.

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. "Predicting human eye fixations via an LSTM-based saliency attentive model". In: *arXiv preprint arXiv:1611.09571* (2016).

[9] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. "Going Deeper: Autonomous Steering With Neural Memory Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 214–221.

[10] Tianxiang Gao and Vladimir Jojic. "Sample Importance in Training Deep Neural Networks". In: (2016).

[11] Angelos Katharopoulos and François Fleuret. "Biased Importance Sampling for Deep Neural Network Training". In: *arXiv preprint arXiv:1706.00043* (2017).

[12] Jinkyu Kim and John Canny. "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention". In: *arXiv preprint arXiv:1703.10631* (2017).

[13] Jinkyu Kim and John Canny. "Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention". In: *arXiv preprint arXiv:1703.10631* (2017).

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[15] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet". In: *arXiv preprint arXiv:1411.1045* (2014).

[16] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. "DeepGaze II: Reading fixations from deep features trained on object recognition". In: *arXiv preprint arXiv:1610.01563* (2016).

[17] Michael F Land and Mary Hayhoe. "In what ways do eye movements contribute to everyday activities?" In: *Vision research* 41.25 (2001), pp. 3559–3565.

[18] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. "Towards fully autonomous driving: Systems and algorithms". In: *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE. 2011, pp. 163–168.

[19] Yufan Liu, Songyang Zhang, Mai Xu, and Xuming He. "Predicting salient face in multiple-face videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4420–4428.

[20] Andrea Palazzi, Francesco Solera, Simone Calderara, Stefano Alletto, and Rita Cucchiara. "Learning where to attend like a human driver". In: *IEEE Intelligent Vehicles Symposium, Proceedings* (Nov. 2017), pp. 920–925. DOI: 10.1109/IVS.2017.7995833. URL: http://arxiv.org/abs/1611.08215.

[21] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. "Predicting the Driver's Focus of Attention: the DR (eye) VE Project". In: *arXiv preprint arXiv:1705.03854* (2017).

[22] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: (2016). ISSN: 0146-4833. DOI: 10.1142/9789812771728{_}0012. URL: http://arxiv.org/abs/1612.08242.

[23] Giacomo Rizzolatti, Lucia Riggio, Isabella Dascola, and Carlo Umiltá. "Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention". In: *Neuropsychologia* 25.1 (1987), pp. 31–40.

[24] Alison Smiley. en. 3rd ed. Lawyers and Judges Publishing, 2015, pp. 550–550. ISBN: 978-1-933264-88-2.

[25] Ashish Tawari and Byeongkeun Kang. "A computational framework for driver's visual attention using a fully convolutional architecture". In: *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE. 2017, pp. 887–894.

[26] Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher Baker, Robert Bittner, MN Clark, John Dolan, Dave Duggins, Tugrul Galatali, Chris Geyer, et al. "Autonomous driving in urban environments: Boss and the urban challenge". In: *Journal of Field Robotics* 25.8 (2008), pp. 425–466.

[27] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. "End-to-end learning of driving models from large-scale video datasets". In: *arXiv preprint arXiv:1612.01079* (2016).