Aggregated Sparse Attention for Steering Angle Prediction

Sen He, Dmitry Kangin, Yang Mi and Nicolas Pugeault
Department of Computer Sciences,University of Exeter, Exeter, EX4 4QF
Email: {sh752, D.Kangin, ym310, N.Pugeault}@exeter.ac.uk

***Abstract*—In this paper, we apply the attention mechanism to autonomous driving for steering angle prediction. We propose the first model, applying the recently introduced sparse attention mechanism to visual domain, as well as the aggregated extension for this model. We show the improvement of the proposed method, comparing to no attention as well as to different types of attention.**

I. Introduction

Consider a human driving a car on a countryside road. The driver's brain is subjected to a continuous flow of large quantities of visual information, interpreting it in real time to provide fast, precise and reliable control of the vehicle. An essential mechanism that allows such an efficient and fast processing of information is *visual attention*, which has been extensively studied by psychologists. Early computational models of attention, inspired by the seminal work of Itti & Koch [1], focused on the top-down mechanism that elicit eye movements when subjects perform a visual search of objects on images. The aim of such models is to estimate from an image a so-called *saliency map*: an estimate of how likely are the subject's eyes to look at image locations given the patterns it contains (see, eg, [2], [3]). Saliency models can either be engineered based on properties of images, or learnt from eye tracking records of human subjects. In both case, the quality of saliency models is estimated by comparing their prediction with actual eye fixations on dataset of images. Although such approaches can predict fairly well the eye fixations of human subjects when asked to perform a visual search task, their predictiveness is much worse when the subjects are performing an *active* task, such as playing video games or driving [4], [5]. More recently, several groups have proposed to learn attention not by mimicking the gaze of human subjects, but by optimising a system's performance at a specific task [6], [7]. In contrast to saliency, that is purely bottom-up, such models are explicitly task dependent.

This article proposes a novel attention mechanism for convolutional neural networks that is based on learning a task-specific sparse attention mechanism. In particular, we focus on the challenging task of predicting steering angle from visual input only [5]. We demonstrate that such a sparse attention focusing leads to better performance. Moreover, we provide experimental evidence that such an attention model is very sensitive to initial conditions and demonstrate that an ensemble of sparse attentional models can significantly improve not only the robustness of the learning process, but also overall performance.

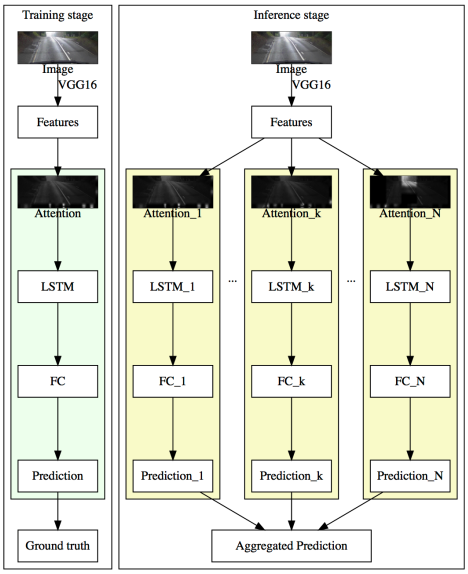


Fig. 1: Architecture of the proposed aggregated attention model.

The rest of this article is organised as follows: Section II reviews the use of attention mechanism in computer vision as well as steering angle prediction; Section III provides the detailed methodology used in this work; Experimental results, together with comparison are presented in Section IV; and conclusions are drawn in Section V.

II. Related Work

A broad range of attention models have been proposed over the years in the literature. This article is concerned in particular with the problem of *task-dependent attention*, where the focusing of attention is optimised to *improve a system's performance at a given task*. This is in contrast to saliency models which are designed to mimic human subjects' gaze patterns irrespective of the tasks demands. Existing models of task-dependent attention for neural networks can be classified in two groups: soft attention and hard attention.

In *soft attention*, the visual input is processed by a pre-trained convolutional neural network, and the output of the top convolutional layer is encoded by a feature tensor. Soft attention consists in weighing the feature tensor with an attention matrix that encodes the relative importance of all locations in the feature tensor. This weighted tensor is fed to another network (ie, a fully connected or a recurrent neural network) to optimise the desired task. The attention matrix is therefore learnt from the task and normalised using a softmax function. Li et al [6] used soft attention to develop a multilevel attention model for video captioning. Their model uses two attention layers. The first layer models *region-level attention*, which encodes the importance of each region in a frame. The second attention layer models *frame-level attention*, which encodes the importance of each frame in a short video. Sharma et al [7] proposed a soft attention model for action recognition. In their model, the output of a pre-trained deep convolutional neural network is fed to a *long short-term memory (LSTM)* network to output the action as well as the attention matrix. Xu et al [8] used soft attention for image captioning. Their model also uses a LSTM network that generates an attention matrix at each time step and generating sentences to describe the input image.

One limitation of soft attention is that it only reweighs the convolutional features and therefore everything is always attended to, although not with the same relative importance (hence *soft* attention). In contrast, *hard attention* models only process part of the input, which is assumed to be the most important region. Because hard attention is not differentiable, it is more challenging to optimise. Mnih et al [9] proposes to learn hard attention from reinforcement learning. There are two crucial component in their network: The first one is a glimpse sensor, which can be used to extract a retina-like representation centred at a given location in the input; The second component is a glimpse network, it is used to process the retina-like representation extracted from the glimpse sensor, and the processed information is then fed into a recurrent neural network (RNN) which estimates the attention focus for the glimpse sensor at the next iteration.

This article is especially concerned with active tasks, and in particular the problem of estimating steering from vision. Pugeault & Bowden [5] developed a pre-attentive model using gist [10] and random forests, while the deep network models are CNN or CNN+ LSTM based, Bojarski et al [11] used a convolutional neural network to map images to steering command. Du et al [12] explore two different models for steering angle prediction. The first is a 3D convolutional model with residual connections and LSTM cell. The second one uses transfer learning to fine-tune a pre-trained CNN and predict steering angles for individual images.

Much less work has been done on the application of attention mechanism in autonomous driving. In this article: i) we propose a new sparse attention model, based on the sparsemax function [13], yields better performance; ii) we demonstrate that bagging multiple sparse attention models can provide a significant performance improvement over single models; iii) we show that the proposed architecture performs better than the state-of-the-art soft attention model, CNN, CNN+LSTM for steering angle prediction.

III. Regression of steering angles with attention

This section describes the proposed sparse attention model: First, we present the overall architecture in Figure 1; we then describe the LSTM model and sparse attention formulation for steering regression; and finally, the proposed model aggregation approach.

A. Feature Extraction

The deep convolutional neural networks have achieved great success in computer vision due to its ability to learn hierarchical features. In our model, we extract the feature for each frame in a driving video using the convolutional part of VGG16 [14], which was trained for image recognition. After feature extraction, each frame was represented by a tensor of shape $M \times N \times K$ determined by the input size. We refer to feature tensor as a feature cube with $M \times N$ locations, and each location was represented by a feature vector of K elements:

$$X = [X_1, X_2, \cdots, X_{M\times N}] \tag{1}$$

The feature extraction part is fixed during the experiment and not fine-tuned.

B. LSTM

In order to take into account the previous context to predict steering angle, the recurrent neural network was used. Recurrent neural network (RNN) can process the time sequence by remembering the needed information and forgetting the redundant. Long Short Term Memory (LSTM) networks [15] are a kind of gated RNN, which can avoid the gradient vanishing or exploding problems encountered by standard RNNs.

C. Sparse Attention

A fundamental limitation of soft attention is that all image regions are in effect attended to at all times: Their importance is merely reweighed by the attention model. This is contrary to the very intent of attention learning.

In this work, we propose to mitigate this limitation by implementing a *sparse attention mechanism* based on [13], but extending it to visual inputs. The output of the attention transformation is defined as

$$X_{weighted} = X \cdot A, \tag{2}$$

where the elements of sparse attention matrix A sums to 1, and it is determined by the feature cube of the current input frame and the model hidden state in the last time step and normalised by the sparsemax function:

$$A = \text{sparsemax}(\tanh(W_f X + W_h H + b)) \tag{3}$$

where the W_f is weight for the current input frame's feature, W_h is the weight for the model hidden state, H is the hidden state of the model, both of the weights are learned during training to form the attention matrix, the sparsemax function which is defined by [13] in Algorithm 1.

Algorithm 1 sparsemax

```
Input : $\boldsymbol{z}$
Sort $z_{(1)} \geq \cdots \geq z_{(M \times N)}$
Find $k(z) := \max \{ k \in [M \times N] | 1 + kz_{(k)} > \sum_{j \leq k} z_{(j)} \}$
Define $\tau(z) = \frac{(\sum_{j \leq k_{(z)}} z_{(j)})-1}{k_{(z)}}$
Output : $\boldsymbol{p}$ s.t. $p_i = [z_i - \tau(z)]_+$
```

One can see that the sparsemax function is not continuous. More importantly, compared to the softmax function, it has the ability to inhibit the unimportant but enhance the significant elements of the input [13]. The final prediction is generated by a two layer fully connected layers (FCN):

$$S(t) = W_{fcn2}(W_{fcn1}H_t + b_{fcn1}) + b_{fcn2} \tag{4}$$

where, $S(t)$ is the predicted steering angle, W_{fcn1} and W_{fcn2} are the weights of each fully connected layer, b_{fcn1} and b_{fcn2} are the bias of each layer, and H_t is the output of LSTM.

D. Model Aggregation

Due to the non-continuity of the sparsemax function, we suggest that the result of training a sparse attention model is highly dependent on (random) initialisation. This means that the resulting attention models after training, although converging to similar performance levels, correspond to very different local minima depending on the random initialisation. In other words, the same task can afford multiple attention models of similar quality. If those models all capture different aspects of the task, a combination of those models could lead to better performance. Therefore, we propose to train a collection of N (we choose $N = 3$ in the experiment) randomised attention models. Because model variance can be ensured from the random initialisation, we can train them all using the same dataset (experiments confirmed that training each model on a separate bootstrap samples did not alter the results significantly). At inference time we propose to combine these attention models and average their predictions, similarly to model bagging.

IV. Experimental Results

The advantages of the proposed method are shown using DIPLECS dataset [5], containing indoor and outdoor scenarios, and *Comma.ai* dataset [16]. The proposed method is compared to soft attention and aggregated soft attention (ASA). Also the method is compared to the gist-based approach [5], method with no attention and no LSTM (CNN), as well as LSTM without attention (CNN+LSTM). To measure the prediction quality, we use the mean absolute error.

A. Dataset description

The indoor part of the *DIPLECS* dataset [5] is collected using a radio controlled car (see Figure 2, left). There are two tracks, P-shaped and O-shaped, eight recordings for each of them from different starting point. For each of the tracks, three recordings are used for training, one for validation, and the rest for testing.

The outdoor part of the *DIPLECS* dataset [5] contains real world driving scenarios (see Figure 2, middle), totalling about 47 minutes of driving, or $84,690$ frames. This dataset has been divided into eight subsequences of the same length, with junctions removed as causing ambiguity which cannot be resolved using vision-based information. Six subsequences were used to train the model, another two were used for validation and testing. In order to factor out focusing attention on the steering wheel and the mirror, these regions were cropped.

Comma.ai dataset consists of 10 day- and night-time highway driving video clips of variable size, in total 7.5 hours (see Figure 2, right). We extract 8 sequences from the dataset, each of them contains 4000 frames, and use 6 of them to train and the rest for validating and testing.

B. Training parameters

Our models are trained using Tensorflow [17] with the L^1 norm loss function, we set the learning rate as 10^{-4} and use Adam [18] optimisation method to train the model, all weights to be trained in the model are initialised using Xavier [19] initialisation method.

C. Indoor Dataset Results

In the indoor dataset, the remote control steering angle signal was normalised to $[-1, 1]$ (-1 corresponds to the leftmost and 1 to the rightmost angle). In Figures 4, we can see that the steering regression improves with the addition of any attention mechanism. Also, sparse attention performs better than soft attention, and the proposed aggregated sparse attention model performs best among those models. Importantly, we note that the models with attention provide a steering control that is not only more accurate but also smoother, which may be favourable for control applications. This particularly true for the proposed model. Figure 5 shows the attention maps for soft and sparse attention respectively. Note that the attended regions for most models appear to be focused on the road markings, some of them on the boundary marking or on the central one and some of them on both. We note also that the attention map for sparse attention is sparser than soft attention, which was the purpose of using the $\mathrm{sparsemax}$ function.

D. Outdoor Dataset Results

In practice, a driver's actions are not instantaneous: due to reaction time, the driver's actions at any instant t are based on the visual input received some time before. According to the studies in [20], [21], a driver's reaction time can vary from a few hundred milliseconds to several seconds. Before this section, we were predicting just the steering angle for the current frame ($s(t) = f(i(t))$). Here we use the current frame to predict the steering angle for different time delays ($s(t + d) = f(i(t))$. We choose the delays corresponding to 0s, 0.25s, 0.5s, 0.75s, 1s, and compare the results for different attention models.

One can see from Figure 7 that the proposed aggregated sparse attention model achieves the best performance among

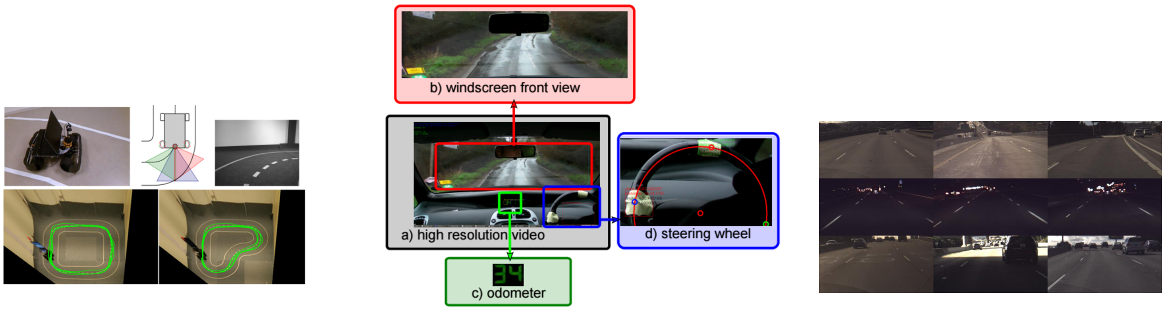
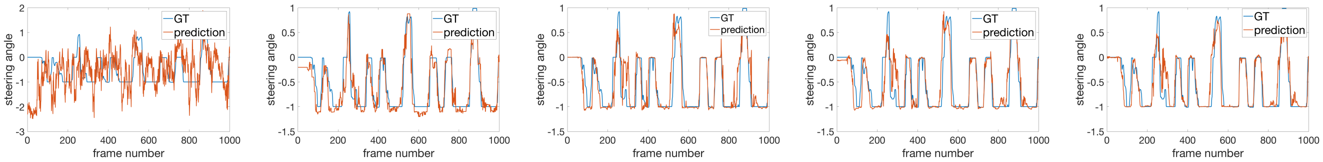


Fig. 2: The datasets used in this article: left, the *DIPLECS* indoor dataset (figure reproduced from [5]); middle, the *DIPLECS* outdoor dataset (figure reproduced from [5]; right, the *Comma.ai* dataset [16].



(a) CNN (b) CNN+LSTM (c) Soft attention (d) ASA (e) Proposed method

Fig. 3: Steering angle prediction of different method in one recordings of the indoor dataset.

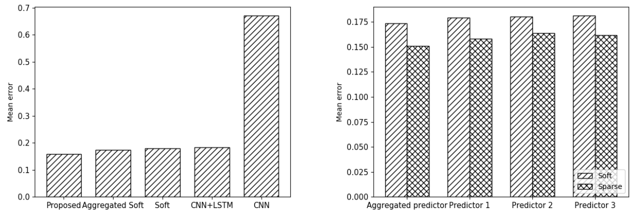


Fig. 4: The mean error of different methods in the DIPLECS indoor dataset (left),and the mean error of each single predictor and aggregated predictor for sparse and soft attention in DIPLECS indoor dataset (right).

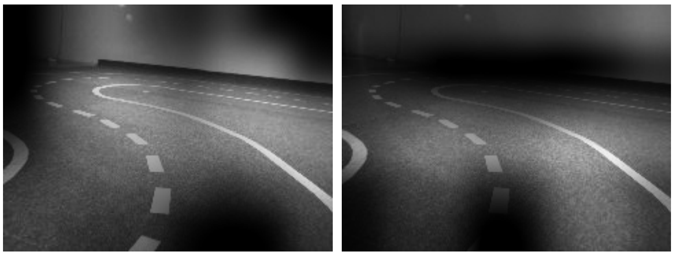
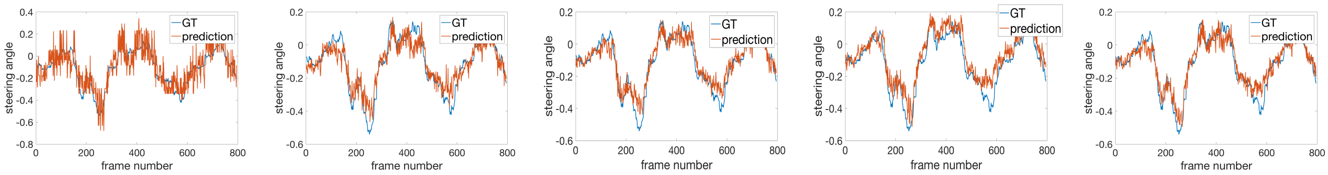


Fig. 5: The attention map for soft attention (left) and sparse attention (right) in DIPLECTS indoor dataset

all methods with 0.5s time delay, which is also the minimum mean error for all time delay among all methods. All models, which have an attention mechanism, perform much better than those without attention, and Figure 6 shows that the model with attention mechanism is more stable, with less perturbation for steering angle prediction, than the model without attention. Figures 8 and 9, show the area of the visual field where attention is focused, for selected frames. Each single attention map only focus on a few different areas (the bright parts). Initially, attended locations are mostly at the bottom of the screen, but when time delay increases, we start seeing locations higher in the image being attended—this is especially visible for prediction time delay of $0.25s$ and $0.5s$.

It is also remarkable that in Figures 10, even though every single predictor within the aggregated sparse attention is trained using the same dataset, the aggregated model performs better than any single sparse attention predictor. For the aggregated soft attention model, two varieties of the model were compared: each single model within the aggregation has been trained on the same training set (ASA) or on different random subsets (ASAR). After model aggregation, the aggregated soft attention model performed worse than the single soft attention model for time delays 0.25 s,0.5 s. We suggest this is due to the cross-correlation between the attention maps of each single soft attention predictor. In Figure 11, one can see that there is a high correlation between the attention maps; even the smallest correlation coefficient of the soft attention map pairs is larger than the largest correlation coefficient for the sparse attention map, which suggests those soft attention models tend to focus on more or less the same region. In this case, if a single predictor model over- or underestimates the steering angle at some time, the other correlated predictors would also have the same trend in steering angle estimation, and after model aggregation it would negatively impact the final error. This result also confirms our suggestion about diversity of individual sparse attention maps, made in section III-D.



(a) CNN+LSTM (b) Soft attention (c) ASA (d) ASAR (e) Proposed method

Fig. 6: Predicting the steering angle 0.5 seconds later, by different methods for a subsequence of the DIPLECS outdoor testing dataset, the blue curve is the ground truth and the red one is the predicted steering angle.

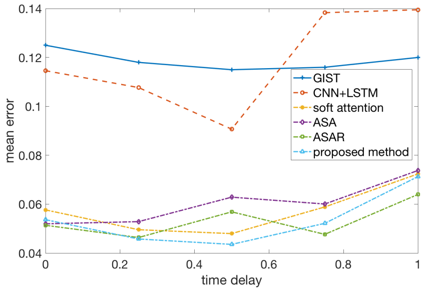


Fig. 7: The mean regression error of different attention models on the DIPLECS outdoor testing dataset, for different time delay.

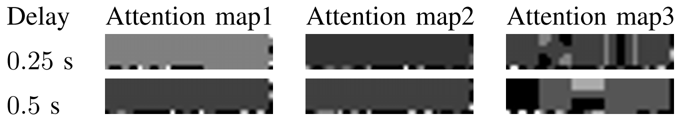


Fig. 8: Attention maps for $0.25s$ and $0.5s$ delays of each single sparse predictor

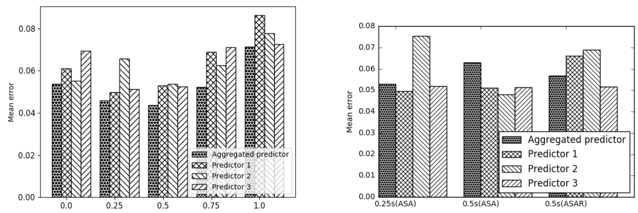
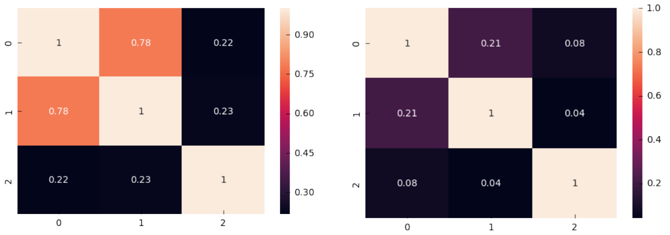


Fig. 10: The mean error of each single predictor as well as aggregated predictor for different time delay of sparse (left) and soft(right) attention in DIPLECS outdoor dataset.

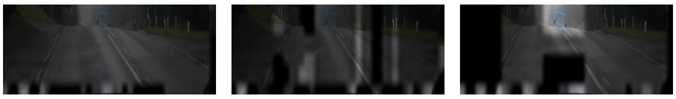


(a) Soft attention (b) Sparse attention

Fig. 11: The cross correlation between the attention map of each single attention model for aggregated sparse attention model and aggregated soft attention model with 0.5s time delay.

E. *Comma.ai* Dataset

The testing procedure for *Comma.ai* dataset is the same as for DIPLECS outdoor dataset. One can see from Figure 12, 13 that the proposed aggregated sparse attention model still achieves the best performance among all methods, but for different time delay (1s time delay), we suggest that this is due to the driving environment being a highway with a broad view far ahead of the car, and therefore possibly requiring less attention from the driver than the countryside road in the DIPLECS outdoor dataset. On this dataset as previously, all models with attention mechanism perform much better than models without attention. Also, Figure 14 (left) confirms, as was the case with the DIPLECS outdoor dataset, that even though every single predictor within the aggregated sparse attention is trained using the same dataset, the aggregated model performs significantly better than any single sparse attention predictor. The performance improvement from attention model aggregation is less evident when considering soft attention, as shown in Figure 14 (right).

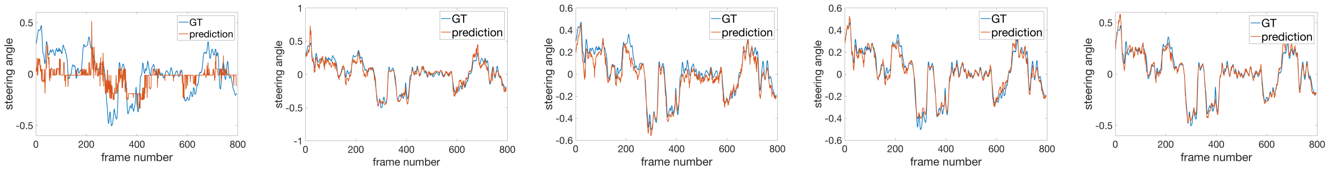


(a) 0s delay (b) 0.25s delay (c) 0.5s delay

Fig. 9: The input frame overlapped by the attention map for different time delay.

V. Conclusion

Attention plays an essential role in human driving. This article experiments with existing neural network models for task-directed attention, and proposed improved models the task of steering a car autonomously. Our experiments show that: i) all attention models improve steering prediction significantly; ii) a sparse attention model yields better performance than classical soft attention; and iii) an aggregated ensemble based on randomised attention models can achieve significantly better performances than a single attention model, even when



(a) CNN+LSTM (b) Soft attention (c) ASA (d) ASAR (e) Proposed method

Fig. 12: Predicting the steering angle 1 seconds later, by different methods for a subsequence of the *Comma.ai* testing dataset, the blue curve is the ground truth and the red one is the predicted steering angle.

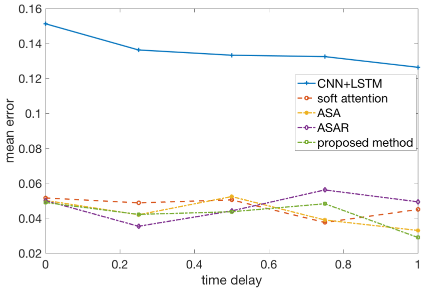


Fig. 13: The mean error of different methods for different time delay in *Comma.ai* dataset.

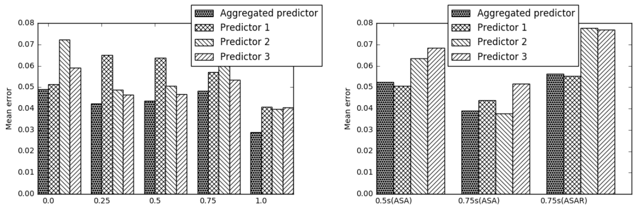


Fig. 14: The mean error of each single predictor as well as aggregated predictor for different time delay of sparse (left) and soft (right) attention in *Comma.ai* dataset.

trained on the same data. The method has been assessed in a variety of scenarios on three datasets, and achieves better performance than state-of-the-art. Additionally, as was done in previous published works the problem of steering angle prediction with a perception-action delay has been considered, demonstrating that the model achieves the best performance for 0.5s delay for countryside road and 1s delay for a highway.

VI. Acknowledgements

This work was supported by the EPSRC project DEVA EP/N035399/1.

References

[1] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[2] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, 2017.

[3] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 598–606, 2016.

[4] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? Modeling top-down visual attention in complex interactive environments," *IEEE Transactions on Systems, Man and Cybernetics: Systems*, vol. 44, no. 5, pp. 523–538, 2014.

[5] N. Pugeault and R. Bowden, "How much of driving is preattentive?," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5424–5438, 2015.

[6] X. Li, B. Zhao, and X. Lu, "Mam-rnn: Multi-level attention model based rnn for video captioning," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[7] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, pp. 2048–2057, 2015.

[9] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, pp. 2204–2212, 2014.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[11] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[12] S. Du, H. Guo, and A. Simpson, "Self-driving car steering angle prediction based on image recognition,"

[13] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *International Conference on Machine Learning*, pp. 1614–1623, 2016.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[16] E. Santana and G. Hotz, "Learning a driving simulator," *arXiv preprint arXiv:1608.01230*, 2016.

[17] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[20] M. Green, "How long does it take to stop? Methodological analysis of driver perception-brake times," *Transportation Human Factors*, vol. 2, no. 3, pp. 195–216, 2000.

[21] H. Summala, "Brake reaction times and driver behavior analysis," *Transportation Human Factors*, vol. 2, no. 3, pp. 217–226, 2000.