IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2023, accepted 29 March 2023, date of publication 5 April 2023, date of current version 10 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264855

RESEARCH ARTICLE

Pay Attention via Quantization: Enhancing Explainability of Neural Networks via Quantized Activation

YUMA TASHIRO[1] AND HIROMITSU AWANO[2], (Member, IEEE)
[1]Department of Information Systems Engineering, Graduate School of Information Science and Technology, Osaka University, Suita 565-0871, Japan
[2]Department of Communications and Computer Engineering, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Corresponding author: Yuma Tashiro (y-tashir@ist.osaka-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 21H03409.

ABSTRACT Modern deep learning algorithms comprise highly complex artificial neural networks, making it extremely difficult for humans to track their inference processes. As the social implementation of deep learning progresses, the human and economic losses caused by inference errors are becoming increasingly problematic, making it necessary to develop methods to explain the basis for the decisions of deep learning algorithms. Although an attention mechanism-based method to visualize the regions that contribute to steering angle prediction in an automated driving task has been proposed, its explanatory capability is low. In this paper, we focus on the fact that the importance of each bit in the activation value of a network is biased (i.e., the sign and exponent bits are weighted more heavily than the mantissa bits), which has been overlooked in previous studies. Specifically, this paper quantizes network activations, encouraging important information to be aggregated to the sign bit. Further, we introduce an attention mechanism restricted to the sign bit to improve the explanatory power. Our numerical experiment using the Udacity dataset revealed that the proposed method achieves a 1.14× higher area under curve (AUC) in terms of the deletion metric.

INDEX TERMS Self-driving, explainable AI, attention, quantized neural network.

I. INTRODUCTION

Deep neural networks (DNNs) have demonstrated overwhelmingly effective performance in tasks such as image recognition, in which they often outperform humans, and their application to speech synthesis and automatic translation has been expanding. To address growing shortages of workers arising from factors such as an aging society and declining birth rates, there has been growing interest in the integration of DNNs into robots and automobiles to further improve the efficiency of factory operations and logistics [1]. As new network structures are invented on a daily basis, DNNs continue to be the subject of intense research and have become far more complex than when they were first developed. For example, AlexNet, which was proposed in 2012 and sparked the DNN boom, comprises approximately 6 million parameters; by contrast, the famous VGG-19 network, which was published in 2014, incorporates approximately 144 million parameters and was optimized using millions of images [2], corresponding to an increase in DNN complexity by a factor of 24 in only two years. The decision-making processes of such a complex DNNs are no longer humanly traceable; this problem is known as “black box AI” and is a significant factor preventing DNNs from being applied to mission-critical applications [3], [4], [5]. To address the black box problem, explainable artificial intelligence (XAI) efforts attempt to make models interpretable by clarifying their inference processes, enabling human understanding of the rationales for actions. By increasing interpretability, the causes of erroneous judgments can be investigated and model judgments can be made more reliable.

One approach to improving the interpretability of DNNs is to visualize the regions of input images that are important in making predictions. For example, Ribeiro et al. proposed a method called “LIME” that provides an interpretable and faithful description of classifier predictions [6] by improving

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Flores.

model interpretability using simple, interpretable classifiers with linear kernels such as support vector machines (SVMs). For a given input image, random perturbations are applied to generate samples that are used to train a local classifier such as an SVM that can easily interpret the inference process and extract regions that contribute to inference. However, this approach is not suitable for real-time applications because it requires iterative computation. Subsequently, in 2017 Selvaraju et al. proposed a method called ''Grad-CAM'' that exploits the information of gradients flowing into the last convolutional layer of a convolutional neural network (CNN) to evaluate the importance of each input pixel for the final class label prediction without the time-consuming iterative training required by a local classifier [7]. Although Grad-CAM is computationally inexpensive, its inverse convolution-based approach suffers from low spatial resolution in its attention maps. To address this, a method for allowing the network to explain the basis for the model's decisions by modifying the network structure so that it outputs not only predictions but also the part that the network focused on making the predictions has been proposed [8]. This mechanism is called the ''attention mechanism'' and is widely used in neural networks, particularly in those focused on natural language processing. An approach based on the attention mechanism can directly extract regions of an image that have a significant impact on the output of the network [9], [10].

In 2017, Kim et al., reported the first successful application of the attention mechanism in self-driving systems [11]. Their model predicts steering angle commands from input raw images in an end-to-end manner and generates an attention heat map that visualizes where and what the model sees. Recent studies have shown that the attention mechanism is effective not only at improving the explainability of DNNs that perform a single task, such as following the road, but also in causal inference for selecting driving behaviors such as stopping at a traffic light or making a left-right turn [12]. However, an attention heatmap can contain unimportant regions, making it necessary to apply post-processing to sort out regions that do not contribute to the network prediction. Accordingly, improvements at the network structure level continue to be made, including an approach that improves explainability by introducing a gazing mechanism at information bottlenecks [13].

In this paper, we focus on the fact that the importance of each bit in the activation value of a network is biased (i.e., the sign and exponent bits are weighted more heavily than the mantissa bits), which has been overlooked in previous studies, and propose a method to further improve the explainability of the network. Specifically, by limiting the addition of attention to specific bits, an attention heatmap that does not require post-processing can be generated. As the bits representing the most significant bit (MSB) side are more informative than other bits, the importance of the bits that make up a floating-point value differ. For example, a change in the sign bit will reverse the direction of steering. Whereas conventional methods assign attachments equally to all bits, the proposed method limits attachment to important bits such as sign bits, making it possible to improve explanatory capability.

We verified the effectiveness of the proposed method using Udacity's automatic driving dataset [14]. We first investigated the prediction accuracy of steering angle and confirmed that the proposed method introduced no degradation in accuracy. We then investigated whether the saliencies of the described attention maps correlated with their contributions to prediction accuracy using the deletion metric [15], a method of masking pixels in an input image in order of increasing attention and comparing the increases in prediction error. If the masking results in a large increase in prediction error, it can be concluded that attention is correctly assigned to the important regions that contribute to the prediction. Numerical experiments revealed that the proposed method maintains the same steering angle prediction accuracy as the existing method [11] while obtaining 1.14 times more explainability. We further investigated the explainability of the results produced by the proposed method under varied feature quantization and confirmed that the increase in prediction error arising from application of the deletion metric was maximized when the features were quantized with 10 bits. It was found that the bit size did not have a significant effect on explainability but could reduce prediction accuracy; accordingly, the optimal bit width for quantization was determined to be approximately 10 bits.

The contributions of this paper are summarized as follows.

- To the best of our knowledge, this paper firstly focuses on the previously overlooked benefits of discretization on explanatory power.
- We demonstrate that introducing a sign-limited attentional mechanism improves explanatory power.
- Numerical experiments on an automated driving task demonstrates a 1.14× improvements in AUC compared to existing methods.

This paper is an extended version of our previously published paper [16]. The enhancements are summarized as follows.

- While [16] required two dedicated convolutional feature extractors, this paper proposes to share a single convolutional feature extractor to reduce the computational complexity.
- While [16] has used floating point representations for network activations, this paper employs discretized activations to improve the explainability.
- Comparative experiments with Grad-CAM [7] and LIME [6] were conducted and demonstrated that the proposed method is superior to these sophisticated methods.

The remainder of this paper is organized as follows. In Sec. II, we provide research background and context, following which we provide the details of the proposed method in Sec. III. In Sec. IV, we discuss our experimental results. Finally, in Sec. V, we provide concluding remarks.

II. PRELIMINARY

A. AUTOMATIC DRIVING SYSTEMS BASED ON IMITATION LEARNING

Self-driving applications have made significant progress in the past few years; in particular, there is growing interest in self-driving cars based on imitation learning using DNNs. In this approach, human driving behavior is collected as training data and the DNN is trained to imitate it. Specifically, the system collects information from various sensors, including cameras mounted in front of the vehicle, and control information relevant to factors such as steering and throttling. The DNN is then trained to generate control commands for steering, throttle opening, etc., from sensor information such as camera images. Automated driving by imitation learning has a long history, with the earliest attempts dating to the Autonomous Land Vehicle In a Neural Network (ALVINN) published in 1989 [17]. ALVINN imitates human driving operation by having a neural network learn the correspondence between a forward camera image and the driver's steering wheel and pedal operations. More recently, the introduction of DNNs has made it possible to automate driving in more realistic situations via imitation learning. An example of this approach is the CNN-based approach proposed by Bojarski et al. [1], who developed a system capable of predicting steering maneuvers using three cameras that capture images of the front, left, and right sides of the car, respectively, as inputs. In a test on a public road, the system successfully drove 10 miles without human intervention. Although automatic driving systems based on DNN and imitation learning have performed well in practical situations, they often demonstrate a "black box problem" in which it is difficult to analyze the inference process and the basis through which the network carries it out.

B. EXPLAINABLE AI

The black box problem is a significant barrier to social implementation of DNNs, and various methods have been proposed to make the DNN inference process interpretable. A pioneer approach in this field was a method called Local Interpretable Model-agnostic Explanation (LIME), which allows neural networks to have explanatory properties. LIME allows interpretation of prediction results based on the idea that DNN models can be easily approximated in the neighborhood of a particular input even if they are very complex in a global context. For a DNN-based classifier and input samples f and $\boldsymbol{X}$, respectively, LIME adds random noise to $\boldsymbol{X}$ to generate samples in their neighborhood and the corresponding neural network output $f(\boldsymbol{X}+\epsilon)$. f is then approximated locally in the neighborhood of $\boldsymbol{X}$ using a sparse linear model, g. Because g is a linear function, i.e., $g = \boldsymbol{w} \cdot \boldsymbol{X}$, a weight vector w can be used to identify features of $\boldsymbol{X}$ that have a significant impact on discrimination.

Although LIME is versatile and can be applied to a variety of models, it is not suitable for applications that require real-time performance because it requires re-training of the classifier for each input. Thus, Grad-CAM, a more computationally efficient method, has been proposed [7]. Grad-CAM visualizes important pixels by weighting gradients against predictions based on the idea that pixels with large gradients have a large impact on predictions. The approach has been extended to Guided Grad-CAM, which visualizes the color maps obtained by Grad-CAM in combination with existing explanatory methods such as Guided Backpropagation and Deconvolution. Grad-CAM and Guided Grad-CAM are computationally less expensive than other XAI methods because they use gradients, which are essential for neural network training, making them suitable for implementation on resource-constrained equipment. However, they rely on a deconvolution layer, which results in an explained attention map with a low spatial resolution.

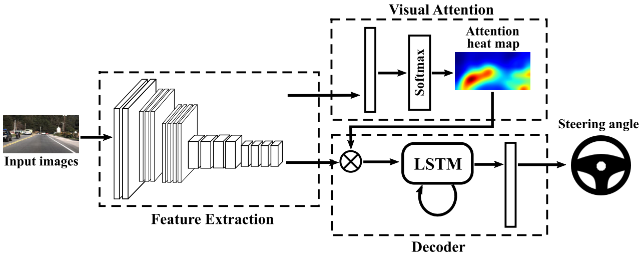


FIGURE 1. Example of Neural Network that predicts steering wheel steering angle via the application of an attention mechanism.

Attention is a method for representing the underlying "reasoning" underlying model decision-making through the introduction of a mechanism to indicate the components of the input data that the model focuses on when making inferences. The attention mechanism was initially introduced to machine translation models in the field of natural language processing as a mechanism for learning the origins within a source text of specific words in a translation generated from it. Recently, it has been applied to CNNs, which are primarily used for image recognition, making it possible to visualize as a heat map the regions in an input image that are focused on during the inference process.

The application of XAI technologies to automated driving has also been expanding. For example, Kim et al. introduced an attention mechanism into an automated driving system comprising a CNN and a long short-term memory (LSTM) to improve model interpretability by displaying image regions that the model focuses on when predicting the steering angle [11]. Fig. 1 shows the model proposed by Kim et al. First, a feature map is extracted through a CNN in the same manner as a general image-input model. This feature map is then passed through a convolutional layer to generate an attention map with the same spatial resolution as the feature map but with a single channel. A spatial Softmax, defined as follows, is then used to generate the attention map:

$$y_{i,j} = \frac{e^{x_{i,j}}}{\Sigma_{i,j} e^{x_{i,j}}}, \tag{1}$$

where $x_{i,j}$ and $y_{i,j}$ are the pixel values of attention map and attention weight at pixel (i, j), respectively.

As the entire attention map is normalized so that the sum of all attention maps is equal to one, the network parameters are naturally trained so that the attention is concentrated on regions that are useful for inference. Therefore, by observing the attention map during inference, it is possible to visually track which regions of the input image have been paid attention to. However, as the naive attention mechanism occasionally assigns high attention to regions that do not contribute to the inference result, Kim et al.'s method performs post-processing to ensure that only regions of interest that are likely to actually contribute significantly to the prediction are retained.

III. PROPOSED METHOD

The structure of the proposed network is shown in Fig. 2. First, a feature map is extracted from the input image using a CNN with five convolutional layers. The extracted feature maps are then weighted by attention weights estimated by the attention mechanism. Unlike the conventional attention mechanism, in which all bits are equally weighted, the proposed network applies attention only to the signs of the bits. The output of the attentional mechanism, which is designed to have a flat shape, is input to the LSTM, which predicts the steering angle. The following sections describe the flow of each of these processes.

A. PREPROCESSING

Given an image captured by a camera at the front of a vehicle, the proposed model predicts continuous steering angle values in an end-to-end fashion. First, an exponential smoothing method is applied to the measured steering angle to reduce measurement noise and improve learning stability as follows:

$$\begin{pmatrix} \theta_t \\ \hat{v}_t \end{pmatrix} = \alpha_s \begin{pmatrix} \theta_t \\ v_t \end{pmatrix} + (1-\alpha_s) \begin{pmatrix} \theta_{t-1} \\ \hat{v}_{t-1} \end{pmatrix}, \qquad (2)$$

where $\hat{\theta}_t$ and $\hat{u}_t$ are the smoothed steering angle and vehicle speed time series data, respectively. α_s is a parameter that adjusts the degree of smoothing; as it approaches zero, the smoothing effect increases. Because the magnitude of the steering angle depends on the structure of the vehicle in terms of, e.g., the wheelbase, this method predicts the inverse turning radius u_t instead of the steering angle [1], [11]. The relationship between the steering angle and the inverse turning radius u_t is approximated by the following equation:

$$u_t = \frac{\hat{\theta}_t}{d_w K_s \left(1 + K_{slip} \hat{v_t}^2\right)}, \qquad (3)$$

where d_w is the distance between the front and rear tires, K_s is the ratio of steering wheel rotation to wheel rotation, and K_{slip} is the relative motion between the wheels and the road surface. The images used as input data are resized to 80 × 160 × 3 to reduce computational cost, and the color space is converted from RGB to HSV.

TABLE 1. Structure of feature extraction encoder.

layer name	output size	filter size, # of channel, stride
conv1	40×80	5×5, 24, stride 2
conv2	20×40	3×3, 36, stride 2
conv3	10×20	3×3, 48
conv4	10×20	3×3, 64
conv5	10×20	3×3, 64

B. NETWORK STRUCTURE

The proposed network comprises three parts: a feature extractor, an attention mechanism, and a steering angle predictor.

1) FEATURE EXTRACTOR

Feature maps are extracted based on the conventional method using a network comprising five convolutional layers [11]. The structure of the feature extraction encoder of the proposed network is shown in Table 1, in which the first column contains the names of the layers, the second column shows the output size, and the third column shows the filter window size, number of channels and stride size. A series of convolution operations on an input image at time t yields a tensor X_t with height H, width W, and channel C. The (i, j) elements of X_t are referred to as $x_{t,i,j} = (x_{t,i,j,1}, x_{t,i,j,2}, \ldots, x_{t,i,j,C})$.

2) ATTENTION MECHANISM

As explained earlier, to improve the explainability, an attention map is applied only to the signs. To this end, the tensor $\boldsymbol{X}_t$ is split into two elements: a sign $x^{sig}_{t,i,j,c}$ and a magnitude $x^{mag}_{t,i,j,c}$. Defining the weight of the attention at time t as $\boldsymbol{\alpha}_t = \{\alpha_{t,1,1}, \alpha_{t,1,2}, \cdots, \alpha_{t,W,H}\}$, the sign of $\boldsymbol{X}_t$ can be computed as follows:

$$x^{sig}_{t,i,j,c} = \begin{cases} +1 \; if \; x_{t,i,j,c} \cdot \alpha_{t,i,j} \geq 0, \\ -1 \; otherwise. \end{cases} \qquad (4)$$

where the attention weight satisfies $\Sigma_{i,j} \alpha_{t,i,j} = 1$. The absolute value of X_t is extracted using the "Softplus" function as follows:

$$x^{\text{amp}}_{t,i,j,c} = \log(1 + e^{x_{t,i,j,c}}). \qquad (5)$$

where $x^{\text{amp}}_{t,i,j,c}$ is a floating-point value in which the significant bits have been distributed to the sign and exponent bits. To isolate the important bits into sign bits alone, the activation is discretized by the discretization function $\mathcal{Q}$:

$$x^{\text{quant}}_{t,i,j,c} = \mathcal{Q}_{a,b,n}\left(x^{\text{amp}}_{t,i,j,c}\right). \qquad (6)$$

The proposed model utilizes Quantization-Aware Training (QAT) to avoid significant degradation of inference accuracy in discretization. QAT was originally used to quantize the parameters and activations of neural networks to reduce model size. Approximating floating-point parameters with small numbers of bits can speed up calculations and reduce memory usage. However, simply quantizing the numerical values as they results in a deterioration of the accuracy of the model after quantization. Therefore, QAT was devised

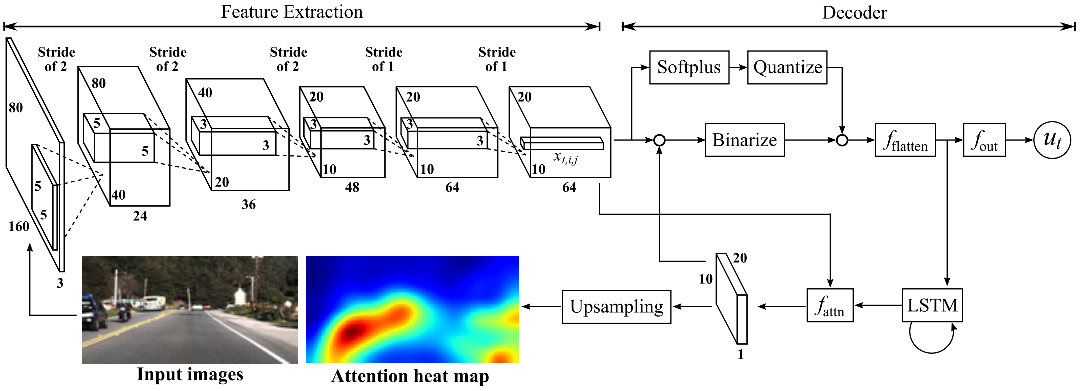


FIGURE 2. Structure of proposed network.

to learn to take the effects of discretization into account and maintain accuracy when quantized.

In QAT, the quantization function $\mathcal{Q}(\cdot)$ is expressed as follows:

$$\mathcal{Q}_{a,b,n}(x) = \left\lfloor \frac{\min\left(\max(x,a),b\right)-a}{\frac{b-x}{n-1}} \right\rceil \frac{b-x}{n-1} + a, \qquad (7)$$

where r is the number before quantization, $[a; b]$ is the range of quantization, $\lfloor \cdot \rceil$ is rounding to the nearest integer, and n is the quantization scale, e.g., $n = 2^8 = 256$ for 8-bit discretization. The derivative of the $\lfloor \cdot \rceil$ function is zero in most places, making it impossible to directly apply the general gradient descent algorithm during training. To solve this problem, a technique called Straight Through Estimator, in which the gradient of the $\lfloor \cdot \rceil$ function is given as follows [18]:

$$\frac{\partial}{\partial x}\lfloor x \rceil = \begin{cases} 1 \ \textit{if}\ a \leq x \leq b, \\ 0 \ \textit{otherwise}. \end{cases} \qquad (8)$$

is utilized. Finally, the outcomes of the first and second paths are multiplied together to reconstitute the signed features:

$$\hat{\boldsymbol{x}}_{t,i,j} = \left(x^{\text{sig}}_{t,i,j,1} \cdot x^{\text{quant}}_{t,i,j,1}, \cdots, x^{\text{sig}}_{t,i,j,C} \cdot x^{\text{quant}}_{t,i,j,C} \right). \qquad (9)$$

The resultant tensor $\hat{X}_t$, whose (i, j)-element is given by $\hat{\boldsymbol{x}}_{t,i,j}$ is flattened to fit into the LSTM module as follows:

$$\boldsymbol{y}_t = f_{\text{flatten}}(\hat{X}_t),$$

where $\boldsymbol{y}_t$ has $W \times H \times C$ elements.

3) PREDICTION OF INVERSE TURNING RADIUS AND ATTENTION

After passing through the attention mechanism, the activation is forwarded to the LSTM module, which predicts the inverse turning radius u_t and the attention $\boldsymbol{\alpha}_t$ conditioned on the previous hidden state of the LSTM, $\boldsymbol{h}_{t-1}$. Formally, the prediction of $\boldsymbol{u}_t$ can be expressed as

$$\begin{aligned} \hat{u}_t &= f_{\text{out}}(\boldsymbol{h}_{t-1}, \hat{X}_t) \\ &= \left(Sigm(W_\alpha \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}_\alpha) \circ \hat{X}_t \right) \cdot W_\beta + \boldsymbol{b}_\beta, \end{aligned} \qquad (10)$$

where W_α and W_β are the trainable weight matrices, $\boldsymbol{b}_\alpha$ and $\boldsymbol{b}_\beta$ are trainable biases, $Sigm(\cdot)$ is the sigmoid function, and $\circ$ is the element-wise multiplication.

To generate the attention $\boldsymbol{\alpha}$, we first compute the additional hidden layer:

$$\begin{aligned} \boldsymbol{e}_t &= f_{\text{attn}}(X, \boldsymbol{h}_{t-1}) \\ &= \tanh\left(W_\gamma \cdot X_t + W_\delta \cdot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_\gamma, \end{aligned} \qquad (11)$$

where W_γ, W_δ are trainable weight matrices and $\boldsymbol{b}_\gamma$ is a trainable bias. The Softmax function is then applied to ensure that $\sum_i \alpha_{t,i} = 1$ as follows:

$$\hat{\alpha}_{t,i} = \frac{\exp(e_{t,i})}{\Sigma_i \exp(e_{t,i})}. \qquad (12)$$

Finally, $\hat{\boldsymbol{\alpha}}_t$ is reshaped into the two-dimensional attention matrix $\boldsymbol{\alpha}_t$.

To initialize of the cell and hidden states of the LSTM at time $t = 0$, we follow the method shown in [11]. Specifically, the states are initialized by the following equation using the hidden layer $f_{init,c}$ and $f_{init,h}$:

$$c_0 = f_{init,c}\left(\frac{1}{L} \sum_{i=1}^{L} x_{0,i} \right), \qquad (13)$$

$$h_0 = f_{init,h}\left(\frac{1}{L} \sum_{i=1}^{L} x_{0,i} \right), \qquad (14)$$

where c_0 and h_0 are the cell and hidden states, respectively, of the LSTM at $t = 0$.

The pseudo code of the network definition is shown in Fig. 3.

IV. EXPERIMENTAL RESULT ON SELF DRIVING TASK

A. EXPERIMENTAL SETUP

We trained and evaluated the proposed method using the Udacity dataset [14], an open-source automated driving project launched in 2016. The dataset contains video images captured by a front view camera mounted on the rear of a vehicle windshield and time-stamped sensor measurements such as vehicle speed and steering angle recorded at

```
# Convolution
x = Conv(in_channels=3, out_channels=24,
         kernel_size=5, stride=2, padding=1)(x)
x = Conv(in_channels=24, out_channels=36,
         kernel_size=5, stride=2, padding=1)(x)
x = Conv(in_channels=36, out_channels=48,
         kernel_size=3, stride=2, padding=1)(x)
x = Conv(in_channels=48, out_channels=64,
         kernel_size=3, stride=1, padding=1)(x)
x = Conv(in_channels=64, out_channels=64,
         kernel_size=3, stride=1, padding=1)(x)

# Attention
e = Tanh(Linear(64, 64)(x) + Linear(1024, 64)(hid))
alpha = Softmax(e)
x_sig = x[:32] * x_sig

x_sig = Binarize(x_sig)
x_amp = Quantize(Softplus(x[32:]))
x = x_sig * x_amp

x,hid = LSTM(x, hid)
```

FIGURE 3. Pseudo-code.

TABLE 2. Comparison of steering angle prediction accuracies.

Methods	MAE [°]
CNN+LSTM w/ Attention [11]	4.94
CNN+LSM w/ Sign only Attention (Proposed)	4.67
CNN+LSTM w/o Attention	4.77

20 frames/second over 3.6 hours of daytime driving on highways and city streets. The dataset contains 263,075 frames, of which 257,796 are used to train the model and the remaining 5,279 are used to evaluate the prediction accuracy of the steering angle.

The model was trained using an NVIDIA Geforce GTX2080Ti GPU and the code was written using the PyTorch framework. Xavier initialization was used for network weight initialization and an Adam optimizer with a learning rate of 10^{-4} was used for training. The dropout probability and size of the LSTM hidden state were set to 0.5 and 1,024, respectively.

B. PERFORMANCE ANALYSIS

1) STEERING ANGLE PREDICTION ACCURACY

We first compared the steering angle prediction accuracy of the proposed method with that of the model proposed in [11]. To examine the impact of the attention mechanism on the prediction performance, we also implemented a combined CNN-and-LSTM structure that was identical to ours except that it lacked an attention mechanism. Tab. 2 lists the mean absolute errors (MAEs) obtained by each method, demonstrating that the proposed method is able to achieve performance comparable to existing methods.

2) VISUALIZATION OF SALIENCY MAP

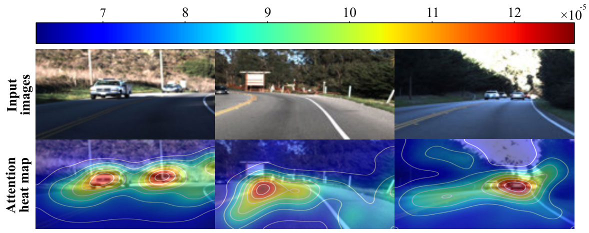


FIGURE 4. Examples of input images and corresponding attention maps.

Fig. 4 shows raw input images from the front camera and corresponding attention heat maps arranged from left to right. As attention is added to the output of the CNN-based feature extraction mechanism, the obtained attention has a resolution of one-eighth that of the input image. To match the resolution of the input images, up-sampling has been applied followed by Gaussian blurring to soften the edges of the image. It is seen from the figure that the proposed method assigns attention to the center line of the road, the white spring, and oncoming cars, all of which are considered to be important when driving.

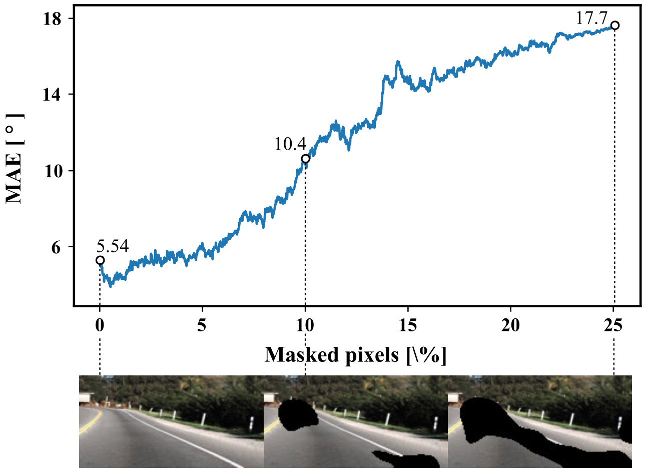


FIGURE 5. Procedures for evaluating explainability via the deletion metric, involving assignment of high-attrition masking regions and evaluation of the increase in inference error.

We further investigated whether the reported attention can effectively reflect the contribution of the deletion metric to the prediction of steering angle [15]. The deletion metric progressively masks pixels according to the ranking provided by the attention mechanism and measures the corresponding increase in the prediction error. For a good explanation, we should observe a monotonic increase in the prediction error. Examples of relationships between masked input images and prediction accuracies are shown in Fig. 5, which demonstrates that the prediction error for the original input image was 5.54° and increased to 17.7° when one quarter of the image was masked. We conducted the same procedure on 5,279 frames in the test dataset and calculated the resulting MAE. Fig. 6(a) shows the relationship between the percentage of pixels masked and the calculated MAE, with the black and red lines corresponding to the MAEs of the conventional attention-based method [11], Grad-CAM [7], LIME [6], and proposed methods, respectively. It is clearly seen that the proposed method produces a sharp increase in error; specifically, when a quarter of the input image is masked, the error produced by the proposed method increases to 8.35° whereas those of [6], [7], and [11] increase only to 6.33°, 6.63°, 7.43°, respectively, confirming the improved explainability of the proposed method. As shown in Fig. 6(b), we further calculated the areas under curve (AUCs) to quantitatively compare the reliability of the proposed attention mechanism. The AUC of the proposed method is 162 whereas those of the conventional attention-based method [11], Grad-CAM [7], and LIME [6] are 141, 140, 154, respectively. A high AUC indicates that the attention weights closely correlate with the significance of the input regions. The 1.14× increase in the AUC achieved by the proposed method relative to the conventional attention-based method [11] again confirms that the explanation of the proposed method is better than that of the conventional method. Some readers may find that the explanatory ability of the proposed method is only marginally improved compared to LIME. However, the computational cost of the proposed method is significantly lower because it can extract the region of interest in a single inference, whereas LIME requires multiple inferences with perturbations to the input.

C. QUANTIZED ATTENTION

The proposed method discretizes the feature map activations to ensure that sign bits have the greatest weight. The discretization accuracy should be selected carefully because it affects the prediction accuracy and explainability of the steering angle. Specifically, decreasing the discretization accuracy (approximating the activation with fewer bits) should degrade prediction accuracy while improving explanatory ability. On the other hand, increasing the discretization accuracy (approximating activations using more bits) improves prediction accuracy while decreasing the explanatory power owing to the relatively lower weights of the sign bits. Therefore, we investigated the relationship between bit width, prediction accuracy, and explanatory power when converting features to fixed-point representation. The AUCs produced by the deletion metric when the bit width is increased in steps of 2 bits from 2 to 24 bits are shown in Fig. 7(a). Similarly, the relationship between the steering angle prediction error and bit accuracy is shown in Fig. 7(b). From Fig. 7(a), it is seen that the AUC is highest when activations are quantized at 10 bits. From Fig. 7(b), it is seen that changing the bit width does not significantly change the prediction accuracy, which indicates that 16 bits—at which point the highest AUC is achieved—is the optimal value.

D. IMPROVED EXPLAINABILITY WITH SOFTMAX WITH TEMPERATURE

To promote attention to specific areas in an input image, a temperature parameter T can be introduced to the Spatial Softmax function as follows:

$$S(y_{i,j}) = \frac{e^{x_{i,j}/T}}{\sum_{i,j} e^{x_{i,j}/T}} \tag{15}$$

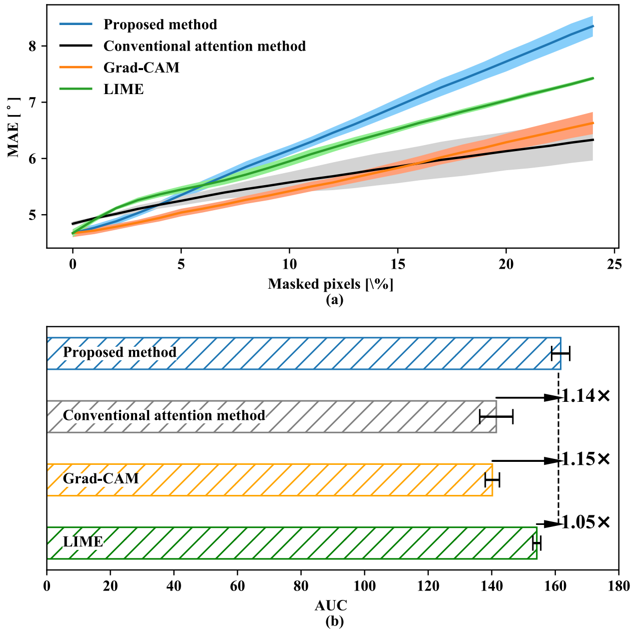


FIGURE 6. (a) Relationship between percentage of masking and increase in inference error and (b) AUCs of proposed and existing methods.

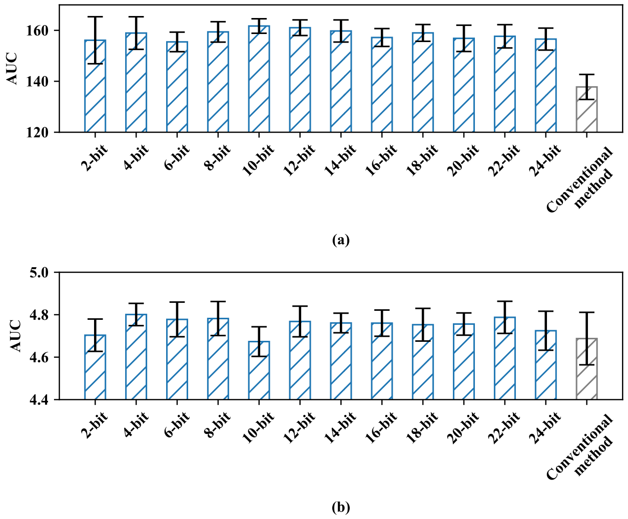


FIGURE 7. (a) Relationship between number of quantization bits (*n*) and AUC and (b) Relationship between mean absolute error (MAE) of steering angle prediction and number of quantization bits (*n*).

The temperature parameter T can be used to control the smoothness of the probability distribution, i.e., as T decreases a higher probability is assigned to the probability corresponding to the (i, j) with the largest $y_{i,j}$ ($S(y_i)$) and, conversely, as T increases $S(y_{i,j})$ is assigned a uniform probability regardless of whether $y_{i,j}$ is large or small. For example, at the limit of $T \rightarrow 0$, only $S(y_{i,j})$ corresponding to the i that maximizes $y_{i,j}$ will be equal to one. Conversely, at the limit of $T \rightarrow \infty$, $S(y_{i,j})$ takes equal values for all (i, j).

By using the property of Spatial Softmax with temperature, which assigns probabilities to specific inputs when the

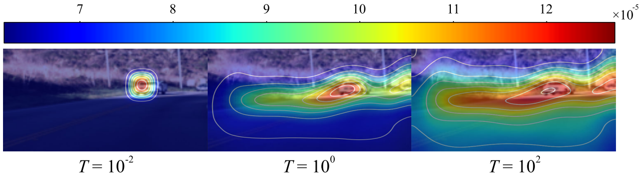
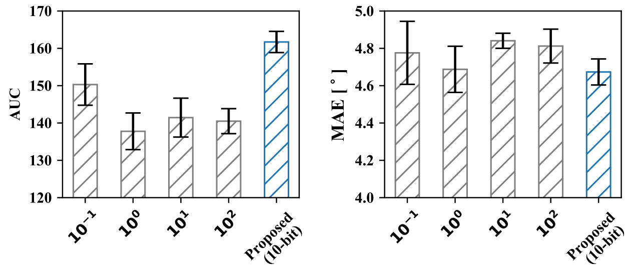


FIGURE 8. Relationship between temperature parameter (*T*) and attention.



(a) Temperature (*T*) vs. AUC

(b) Temperature (*T*) vs. inference error.

FIGURE 9. Relationships between (a) temperature parameter (*T*) and AUC and (b) temperature parameter (*T*) and inference accuracy in MAE.

temperature T is decreased, it is possible to force the concentration of attention on specific regions in an input image. Fig. 8 shows output attention maps obtained by applying Spatial Softmax with different temperatures to the same input image. The results clearly indicate that reducing the temperature causes more attention to be focused on a single point, whereas increasing the temperature increasingly disperses the attention.

It would seem plausible that reducing the temperature parameter would make the model pay more attention to the important areas involved in the prediction, thereby improving its explanatory power. Therefore, we examined the effect of temperature parameters on explanatory power using the deletion metric [15]. The relationship between temperature and AUC is shown in Fig.9(a); for comparison, the AUC of the proposed method is also shown. From the figure, it is seen that the AUC increases insignificantly as the temperature is decreased. To quantitatively investigate the relationship between temperature and AUC, we applied Welch's t-test, which is used to test the hypothesis that two populations have equal means. Unlike the Student's t-test, it can be used on two samples that might have unequal variances. Welch's t-test defines the statistic t using the following equation:

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}. \tag{16}$$

To test whether the AUCs obtained at $T = 0.1$ and $T = 1$ differ significantly, the p-value corresponding to the two curves was calculated using Welch's T-test as $p = 0.193$. At a significance level of 5%, this is considered not to be a significant difference. On the other hand, testing of the AUCs of the proposed method and that at $T = 0.1$ returned a p-value of 6.55×10^{-5}, which represents a significant difference. From the relationship between inference accuracy and temperature shown in Fig. 9(b), it is further seen that the inference accuracy changes insignificantly even if the temperature is varied. Furthermore, the inference accuracy of the proposed method, which is shown with the blue bar, differs to no significant degree from that of the conventional method. It is clear from these results that the proposed method is able to improve explanatory power without sacrificing inferential accuracy.

V. CONCLUSION

In this paper, we proposed a novel attention mechanism to enhance the explanatory power of deep learning algorithms. Exploiting the fact that bits are not equally important, i.e., MSBs should carry more information than other bits, we developed a method under which the attention mask is applied only to "sign" bits. By exploiting the STE technique, the proposed model can be trained via a standard back propagation algorithm in an end-to-end manner, resulting in no degradation in prediction accuracy under the proposed sign-only attention mechanism. We investigated whether the reported attention mechanism effectively reflects the significance of input in making predictions using the deletion metric and confirmed that the proposed method achieves an AUC 1.14× higher than that achieved by the conventional method [11], indicating that the proposed method can correctly assign attention to important image regions.

The proposed method can visualise the areas that the neural network focuses on during inference in a single inference, so it is expected to be suitable for applications such as automated driving and robot control, where real-time performance is required. Recent research has shown that safety can be improved by attaching eyes to vehicles so that automated vehicles and pedestrians can communicate through eye contact [19]. If the proposed method can help the automated driving algorithm to adaptively direct its eyes to the areas where attention is being paid, further safety improvements could be possible.

REFERENCES

[1] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Müller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, pp. 1–9, Apr. 2016.

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

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[4] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[5] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *Proc. Int. Joint Conf. Artif. Intell. Workshop Explainable AI (XAI)*, 2017, vol. 8, no. 1, pp. 8–13.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[9] 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 *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[10] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *CoRR*, vol. abs/1511.04119, pp. 1–11, Nov. 2015.

[11] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2942–2950.

[12] Y.-C. Liu, Y.-A. Hsieh, M.-H. Chen, C.-H.-H. Yang, J. Tegner, and Y.-C.-J. Tsai, "Interpretable self-attention temporal reasoning for driving behavior understanding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2338–2342.

[13] J. Kim and M. Bansal, "Attentional bottleneck: Towards an interpretable deep driving network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 322–323.

[14] Udacity. *Udacity Self-Driving Car Driving Data*. Accessed: Jun. 23, 2020. [Online]. Available: https://github.com/udacity/self-driving-car

[15] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–13.

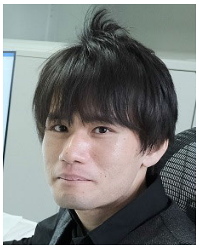
[16] Y. Tashiro and H. Awano, "Pay attention via binarization: Enhancing explainability of neural networks via binarization of activation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 3160–3164.

[17] D. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Neural Inf. Process. Syst.*, Dec. 1989, pp. 305–313.

[18] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized Neural Networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4114–4122.

[19] C.-M. Chang, K. Toda, X. Gui, S. H. Seo, and T. Igarashi, "Can eyes on a car reduce traffic accidents?" in *Proc. Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, 2022, pp. 349–359.

YUMA TASHIRO received the B.S. degree in computer science from Osaka University, in 2021, where he is currently pursuing the M.S. degree.



HIROMITSU AWANO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2010, 2012, and 2016, respectively. He was with Hitachi Ltd., Tokyo, Japan, in 2016; the VLSI Design and Education Center, The University of Tokyo, Japan, from 2017 to 2018; and the Graduate School of Information Science and Technology, Osaka University, Osaka, Japan, from 2019 to 2020. In 2020, he joined the Graduate School of Informatics, Kyoto University, where he is currently an Associate Professor. His research interests include CAD for VLSI design and hardware accelerator for machine learning. He was a Research Fellow of the Japan Society for the Promotion of Science and a member of IEICE and IPSJ.