Improving Reliability of Attention Branch Network by Introducing Uncertainty

Takuya Tsukahara*, Tsubasa Hirakawa*, Takayoshi Yamashita*, and Hironobu Fujiyoshi*
*Chubu University
Kasugai, Aichi, Japan
Email: {tsukahara, hirakawa}@mprg.cs.chubu.ac.jp, {takayoshi, fujiyoshi}@isc.chubu.ac.jp

***Abstract*—Convolutional neural networks (CNNs) are being used in various fields related to image recognition and are achieving high recognition accuracy. However, most existing CNNs do not consider uncertainty in their predictions; that is, they do not account for the difficulty of prediction, and the extent to which their predictions are reliable is unclear. This problem is considered to be the cause of erroneous decisions when we use CNNs in practice. By considering the uncertainty of the prediction result, it is thought that recognition accuracy would improve, and erroneous decisions would be suppressed. We propose a Bayesian attention branch network (Bayesian ABN) that incorporates uncertainty into an attention branch network (ABN). The method incorporates a Bayesian neural network (Bayesian NN) into the ABN to account for uncertainty in the prediction result. Also, it outputs prediction results from two branches and chooses the one having the lower uncertainty. In evaluations using standard object recognition datasets, we confirmed that the proposed method improves the accuracy and reliability of CNNs.**

I. Introduction

Machine learning is a technology that analyzes patterns of known data, extracts patterns and regularities from the data, and makes it possible to predict unknown data from patterns and regularities. Furthermore, convolutional neural networks (CNNs) [1] are machine learning methods that are mainly used for image recognition. Since a method using a CNN [2] won a general object recognition contest [3] held in 2012, CNNs have been applied to various fields, especially image recognition. In particular, they have achieved high recognition accuracy in computer vision, face, place, and object recognition, moving image analysis, and autonomous driving applications. Familiar technologies using CNNs include car driving support technology and image classification technology.

However, most existing CNNs do not consider uncertainty in their predictions; that is, they do not account for the difficulty of prediction, and the extent to which their predictions are reliable is unclear. This problem is considered to be the cause of erroneous decisions in practical use of CNNs. In particular, car driving support technology and image classification technology have been affected by misjudgments caused by CNNs. For instance, an electric car made by Tesla caused a fatal accident allegedly due to a faulty driver assistance system that misjudged a reflection on the white side of a trailer. Apparently, the driver assistance system and the driver did not notice the trailer and could not brake in time to prevent the accident.

If the uncertainty of the prediction result could be taken into account, it is thought that recognition accuracy would improve and erroneous judgments would be suppressed. In this paper, we propose a Bayesian attention branch network (Bayesian ABN) that introduces uncertainty in the attention branch network (ABN) [4]. ABN is a CNN that improves recognition accuracy and provides visual explanation by introducing a branch of attention mechanism. The proposed method considers the uncertainty of the prediction results by incorporating a Bayesian neural network (Bayesian NN) [5] in the ABN. In addition, the proposed method outputs prediction results from two branches and chooses the one having the lower uncertainty. We verified the effect of incorporating uncertainty in the prediction result and the change in recognition accuracy by conducting experiments using general object-recognition datasets.

A. Contribution

The contributions reported here are summarized as follows.

- In a recognition system using machine learning, the recognition result and its corresponding score are output. However, since the reliability of the score is not taken into account, it may cause misjudgments. In the proposed method, erroneous judgments are suppressed by considering the reliability of the score.
- In the recognition system using CNNs, it is difficult to find the cause of an erroneous judgment because the basis of that judgment is unclear. In the proposed method, an attention map can be used for visual explanation, so that an index can be obtained that is useful for investigating the cause of an erroneous judgment when it occurs.

II. Related work

Bayesian NN and Monte Carlo dropout (MC dropout) [6] are uncertainty estimation methods for neural networks (NNs). Moreover, as mentioned above, ABN is a CNN that improves recognition accuracy and provides visual explanation by introducing a branch of attention mechanism.

A. Bayesian Neural Network

Bayesian NN [5] is a probabilistic model in which the dependence of multiple random variables is represented by a graph structure, and the relationships between variables are represented by conditional probabilities. The difference

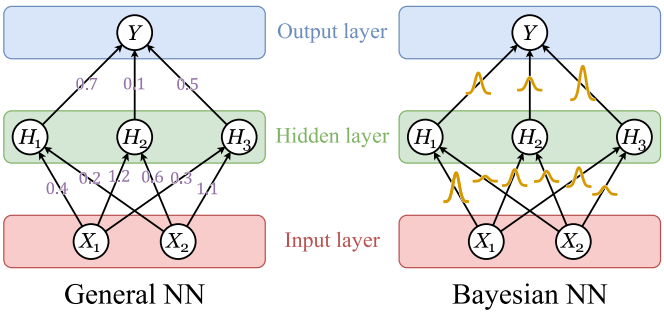


Fig. 1: Difference in structure between general NN and Bayesian NN [5]

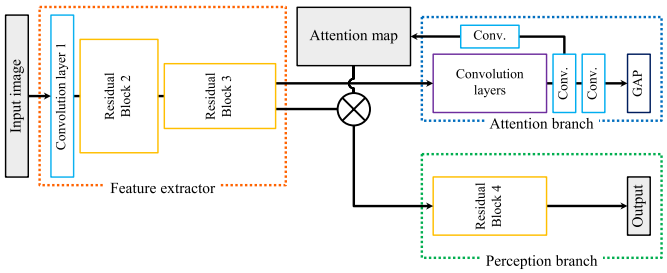


Fig. 2: Network structure of ABN [4]

in structure between a general NN and a Bayesian NN is shown in Fig. 1. General NN weights are point estimates, because they are unique. Therefore, it is impossible to estimate uncertainty with a general NN. Bayesian NNs are different from general NNs in that they represent the weight of a network model by using a probability distribution, and this enables them to estimate uncertainties along with prediction results.

B. Monte Carlo dropout

Applying a Bayesian NN to a CNN is computationally expensive and difficult to optimize. The MC dropout method [6] has been proposed as a way to solve this problem. MC dropout approximates the probability distribution of weights by expressing the weights of the network model in terms of a Bernoulli distribution. The approximate variation distribution $q_\theta^*(\boldsymbol{w}_i)$ of the weights of the unit $j = 1, \ldots, K_{i-1}$ and the layer $i = 1, \ldots, L$ are defined as follows.

$$\boldsymbol{z}_{i,j} \sim \text{Bernoulli}(p_i) \tag{1}$$

$$q_\theta^*(\boldsymbol{w}_i) = \boldsymbol{m}_i \cdot \text{diag}\left(\left[\boldsymbol{z}_{i,j}\right]_{j=1}^{K_i}\right) \tag{2}$$

At this time, an approximate model of the Gaussian process is obtained from the distribution of the random variable vector $\boldsymbol{z}_{i,j}$ and the variation parameter $\boldsymbol{m}_i$ obtained by the Bernoulli distribution. This definition is equivalent to setting the unit of the network model to 0 randomly by dropout [7] with a dropout probability p_i. The prediction distribution is obtained from the average of the sampling using dropout, and the variance and entropy of the prediction distribution provide an index that indicates uncertainty. Bayesian SegNet [8] is a method of applying MC dropout to semantic segmentation tasks, and variational RNN [9] is a method of applying MC dropout to recurrent neural network (RNN).

C. Attention Branch Network

ABN [4] is a CNN that improves recognition accuracy and provides visual explanation by introducing a branch of attention mechanism. The network structure of ABN based on a residual network (ResNet) [10] is shown in Fig. 2. An ABN is constructed from a network used in general object recognition. The network used for general object recognition is divided into a feature extractor and perception branch, and the attention branch is placed after the feature extractor. The attention branch consists of convolutional layers and global average pooling (GAP) [11], and it enables an attention map representing the visual explanation from the feature map of the feature extractor. An ABN improves recognition accuracy by visualizing the attention area in the attention branch, reflecting it in the feature map of the feature extractor, and estimating the prediction result in the perception branch.

III. Proposed Method

We propose Bayesian ABN, which introduces uncertainty to the ABN. Bayesian ABN uses a MC dropout to improve the accuracy and reliability of the prediction results of the ABN by considering their uncertainty. The proposed method is described below.

A. Construction of Bayesian Attention Branch Network

The network structure of a Bayesian ABN based on ResNet is shown in the Fig. 3. The Bayesian ABN is based on networks used in general object recognition like the ABN. In addition, the Bayesian ABN uses a MC dropout to estimate the uncertainty of its prediction results. Therefore, if the network is based on one that does not use dropout, dropout has to be added to it. For instance, ResNet and ResNeXt [12] are networks that do not use dropout. In Bayesian ABN based on ResNet and ResNeXt, dropout is added to the last residual block of the feature extractor and the residual block of the perception branch. In particular, dropout is added after the last convolution layer in each residual module that makes up the residual block.

B. Learning algorithm

The learning method of Bayesian ABN is the same as that of ABN. The loss function used during training is calculated from the predictions obtained from the attention branch and the predictions obtained from the perception branch. Denoting the error of the attention branch as $L_{\text{attention}}$ and the error of the perception branch as $L_{\text{perception}}$, the loss function L is expressed as

$$L = L_{\text{attention}} + L_{\text{perception}} \tag{3}$$

We employ softmax cross-entropy for both loss functions. According to this loss function, learning is performed using the same learning algorithm as using in a general CNN.

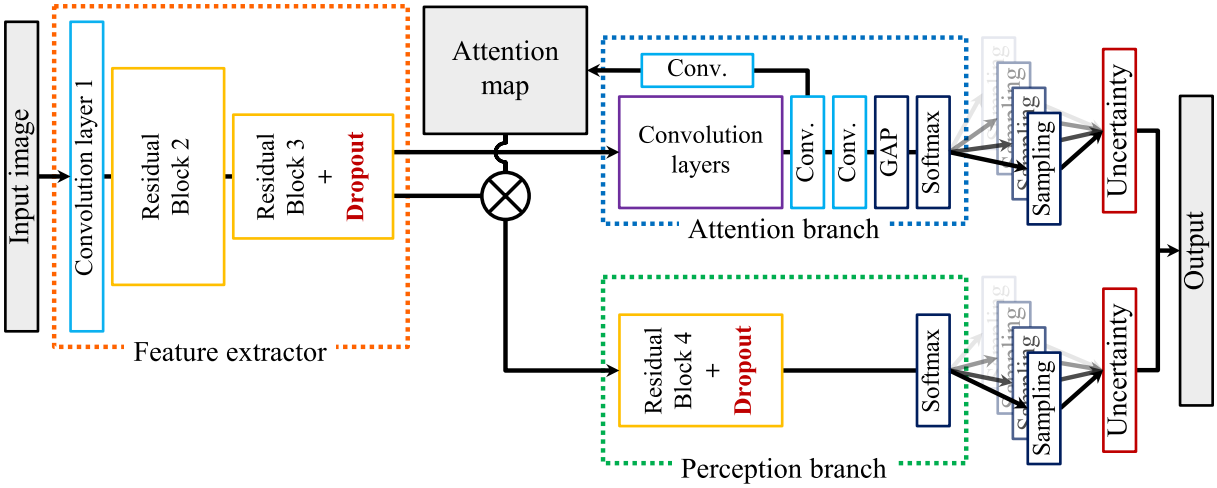


Fig. 3: Network structure of Bayesian ABN

C. Uncertainty estimation

Bayesian ABN estimates the uncertainty of prediction results by using MC dropout during inference. The output result is sampled using dropout, and the predicted distribution $\boldsymbol{p}_{\text{branch}}$ is obtained from the average of the samples. The uncertainty $H\left(\boldsymbol{p}_{\text{branch}}\right)$ is estimated using the entropy of the predicted distribution P_c for each class $c = 1, \ldots, C$:

$$H\left(\boldsymbol{p}_{\text{branch}}\right) = -\sum_{c=1}^{C} P_c \ln P_c \tag{4}$$

$H\left(\boldsymbol{p}_{\text{branch}}\right)$ captures the uncertainty of each prediction distribution in the attention branch and the perception branch.

D. Estimating prediction results using uncertainty

The Bayesian ABN can estimate the uncertainty $H\left(\boldsymbol{p}_{\text{branch}}\right)$ for each predicted distribution $\boldsymbol{p}_{\text{branch}}$ in the attention branch and the perception branch. However, it is not possible to improve the recognition accuracy only by estimating the uncertainty. Therefore, Bayesian ABN recognition accuracy and reliability are improved by using the prediction distribution $\boldsymbol{p}$ of the branch with the lowest uncertainty. If the prediction distribution of the attention branch is $\boldsymbol{p}_{\text{attention}}$ and the prediction distribution of the perception branch is $\boldsymbol{p}_{\text{prediction}}$, the final prediction result is expressed as

$$\boldsymbol{p} = \begin{cases} \boldsymbol{p}_{\text{attention}} & H\left(\boldsymbol{p}_{\text{attention}}\right) < H\left(\boldsymbol{p}_{\text{perception}}\right) \\ \boldsymbol{p}_{\text{perception}} & H\left(\boldsymbol{p}_{\text{attention}}\right) \geqq H\left(\boldsymbol{p}_{\text{perception}}\right) \end{cases} \tag{5}$$

The final recognition result y is expressed as

$$y = \underset{c}{\operatorname{argmax}} \left\{ P_c \in \boldsymbol{p} \right\} \tag{6}$$

IV. Experiment

We performed experiments on general object recognition using networks for general object recognition, an ABN, and Bayesian ABN. The recognition accuracy was evaluated in relation to that of general object recognition, and the effect of introducing uncertainty in the prediction result and the change in recognition accuracy were analyzed.

A. Datasets

The experiment used the CIFAR-10 dataset, CIFAR-100 dataset [13], and ImageNet-1K dataset [14]. The input image size of the CIFAR-10 and CIFAR-100 datasets is 32×32 pixels, and that of the ImageNet-1K dataset is 224×224 pixels. The number of categories for each dataset is as follows: The CIFAR-10 dataset consist of 10 classes, the CIFAR-100 dataset consist of 100 classes, and the ImageNet-1K dataset consist of $1,000$ classes. During training, we applied the standard data augmentation. For the CIFAR-10 and CIFAR-100 datasets, the images are first zero-padded with 4 pixels for each side then randomly cropped to again produce 32×32 pixels images, and the images are then horizontally mirrored at random. For the ImageNet-1K dataset, the images are resized to 256×256 pixels then randomly cropped to again produce 224×224 pixels images, and the images are then horizontally mirrored at random. The numbers of training and evaluation images of each dataset are as follows: The CIFAR-10 and CIFAR-100 datasets use $50,000$ images for training and $10,000$ images for evaluation, and the ImageNet-1K dataset uses $1,281,167$ images for training and $50,000$ images for evaluation.

B. Experimental conditions

The ABN and Bayesian ABN of the experiments were based on networks used for general object recognition. ResNet, wide residual network (WRN) [15], dense convolutional network (DenseNet) [16], and ResNeXt were chosen as the networks for general object recognition. For the CIFAR-10 and CIFAR-100 datasets, ResNet had 110 layers, and WRN had 28 layers; their widening factor was 10. DenseNet had 190 layers; the growth rate was 40, and the compression parameter was 0.5. ResNeXt had 29 layers; the branch number was 16, and the width parameter was 4. For the ImageNet-1K dataset, ResNet

TABLE I: Recognition accuracy of each method [%]

Methods			CIFAR-10 dataset		CIFAR-100 dataset		ImageNet-1K dataset	
Base	ABN	Bayesian ABN	Top-1 accuracy	Top-5 accuracy	Top-1 accuracy	Top-5 accuracy	Top-1 accuracy	Top-5 accuracy
ResNet			93.57	–	75.86	–	77.81	–
	✓		94.25	99.77	76.09	92.80	79.35	94.55
		✓	**94.28**	**99.78**	**78.97**	**94.58**	**80.31**	**95.01**
WRN			95.83	–	79.50	–	76.61	–
	✓		96.04	99.89	82.01	95.53	76.93	92.97
		✓	**96.06**	**99.90**	**82.04**	**95.75**	**77.75**	**93.20**
DenseNet			94.08	–	75.85	–	77.80	–
	✓		94.48	99.79	76.51	93.57	75.85	92.87
		✓	**94.75**	**99.83**	**79.47**	**94.87**	**78.61**	**94.16**
ResNeXt			96.42	–	81.68	–	77.60	–
	✓		96.93	99.91	82.05	86.73	78.48	94.10
		✓	**96.97**	**99.93**	**83.11**	**96.94**	**79.39**	**94.62**

had 152 layers, and WRN had 34 layers; their widening factor was 2.

Regarding the experimental conditions used during training, the dropout probability of WRN and ABN and Bayesian ABN based on WRN was set to 0.3. We optimize the networks by stochastic gradient descent (SGD) with momentum. The total numbers of iterations to update the networks is as follows: The number of training epochs was 600 on the CIFAR-10 dataset and CIFAR-100 dataset and 200 on the ImageNet-1K dataset. The initial learning rate was set to 0.1, and is divided by 10 at $1/2$ and $5/6$th of the total number of training epochs. As for the experimental conditions at the time of the evaluation, the dropout probability of Bayesian ABN was 0.3, and the number of samplings for MC dropout was 50.

C. Experimental result

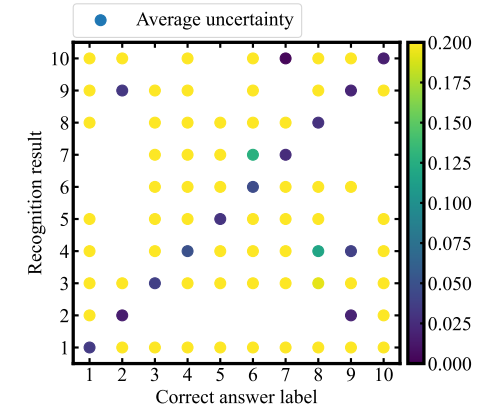
We evaluate the recognition accuracy and the effectiveness of uncertainty in the proposed method by experiments on general object recognition.

1) Evaluation of recognition accuracy: We evaluated the recognition accuracies of networks used for general object recognition, ABN, and Bayesian ABN. TABLE I shows the recognition accuracies of the methods on the CIFAR-10 dataset, CIFAR-100 dataset, and ImageNet-1K dataset in terms of the top-1 and top-5 accuracies. The accuracy of the base network (i.e., ResNet, WRN, DenseNet, and ResNeXt) is cited from [10], [15], [16], [12], respectively. In each case, Bayesian ABN achieved the highest recognition accuracy. Although the results of the proposed method on CIFAR-10 dataset are not increased so much, in case of the results on CIFAR-100 and ImageNet-K datasets, more difficult datasets, our method can improve those accuracies. This is because more information is needed to calculate the uncertainty and more accurate measurements can be obtained.

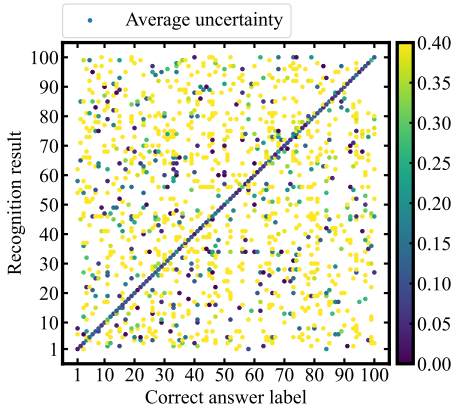
2) Visualization of uncertainty: The experiment used the CIFAR-10, CIFAR-100, and ImageNet-1K datasets to visualize the uncertainty of the results of a ResNet-based Bayesian ABN. Fig. 4 shows a visualization of the uncertainty results for each data. The x axis represents the correct answer label, the y axis represents the recognition result, and the color of the point represents the average uncertainty. The uncertainty of the data in the case of an incorrect recognition result tends to be high, indicating that the uncertainty was correctly estimated. From this trend, uncertainty can be used to determine whether the prediction is accidental or inevitable. Focusing on the labels 2 and 9 in Fig. 4 (a), it can be confirmed that the recognition is wrong although the uncertainty is low. Label 2 represents the car class and label 9 represents the ship class. It can be easily confused because of its similar shape.

3) Assessing the effectiveness of incorporating uncertainty: We evaluated the recognition accuracy of an ABN and a Bayesian ABN based on a network for general object recognition on the CIFAR-100 dataset and ImageNet-1K dataset. We compared their recognition accuracy over the different thresholds based on class score or uncertainty. As the reliability measures, we use uncertainty and a class score provided by the Bayesian ABN and a class score obtained from the ABN. The class score is the maximum value of the class likelihood that is the prediction result. $H\left(\boldsymbol{p}_{\mathrm{branch}}\right)$

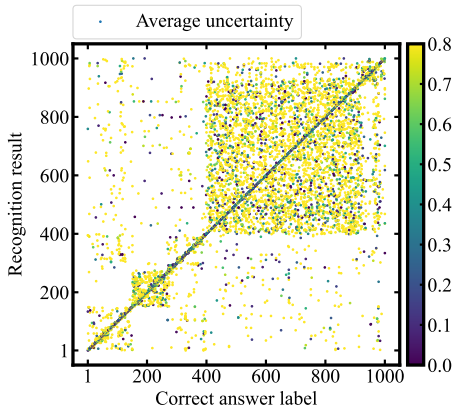
Fig. 5 shows the results for the CIFAR-100 dataset, and Fig. 6 shows those for the ImageNet-1K dataset. The horizontal axis shows the threshold of reliabilities, i.e., the uncertainty or class score. The vertical axis shows the ratio of evaluation samples being correctly classified with lower uncertainty or higher class score than the threshold. Note that, in order to deal with both the uncertainty and class score in a graph, we



(a) CIFAR-10 dataset



(b) CIFAR-100 dataset



(c) ImageNet-1K dataset

Fig. 4: Visualization of uncertainty on (a) CIFAR-10, (b) CIFAR-100, and (c) ImageNet-1K datasets. each colorized dot shows the averaged uncertainty of samples.

scale the uncertainty values $H(\boldsymbol{p})$ to $H'(\boldsymbol{p})$ by

$$H'(\boldsymbol{p}) = 1 - \frac{H(\boldsymbol{p})}{H_{\max}} \tag{7}$$

where $H_{\max}$ is the maximum value of uncertainty in the evaluation dataset. We normalize the class score into [0, 1] by using the maximum value of class scores obtained from the evaluation samples. Therefore, higher reliability (e.g., 1) indicates more certain classification results, and lower reliability (e.g., 0) indicates more uncertain results.

When using the CIFAR-100 and ImageNet-1K datasets, comparing the ABN with the Bayesian ABN shows that the Bayesian ABN classifies more evaluation samples correctly. The results on ImageNet-1K datasets (Fig. 6) show that the uncertainties achieved higher accuracy than the class score. Consequently, these results show that the reliability of a model can be improved by using uncertainty.

4) Visualization of attention maps: We visualized the attention maps of ResNet-based ABN and Bayesian ABN on the CIFAR-10, CIFAR-100, and ImageNet-1K datasets. Fig. 7 shows the obtained attention maps for an input image. Because of the MC dropout of Bayesian ABN, the proposed method generates different attention maps for each sampling. Therefore, we visualize the mean and variance of the attention map. Here, we visualize the attention map as a heat map; red means more highlighted areas, where the attention map takes values between 0 and 1. For the variance of the attention map, white is high and black is low variances.

ABN and Bayesian ABN attention maps capture characteristic regions of objects in an image. Comparing the attention map of ABN and the mean attention map of Bayesian ABN, those seem to be almost the same. Meanwhile, the variance of attention maps obtained from Bayesian ABN represents the confidence of each pixel's attention map values, which would be helpful for further understanding of the decision-making of a network model.

V. Conclusion

We evaluated the recognition accuracy of the proposed method and evaluated the effectiveness of incorporating uncertainty by conducting experiments on general object recognition. The proposed method achieved the highest recognition accuracy among the methods evaluated in every case of the experiments. Incorporating uncertainty was proved effective because recognition accuracy improved when a threshold based on uncertainty was set. In the future, we aim to improve the recognition accuracy and reduce learning costs of CNNs by incorporating uncertainty into active learning.

References

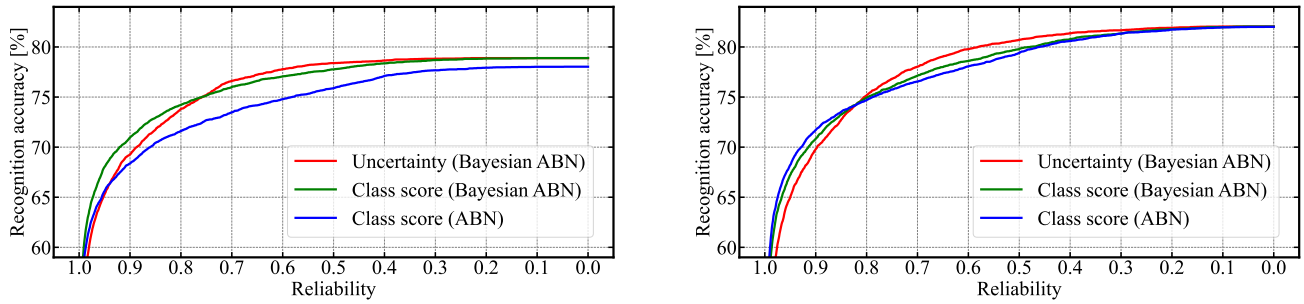
[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

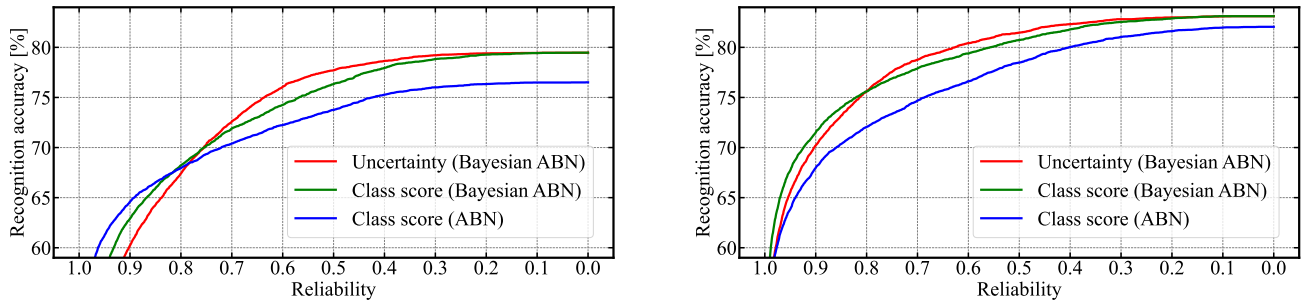
[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Computer Vision and Pattern Recognition*, 2019.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International Conference on Machine Learning*, 2015, pp. 1613–1622.

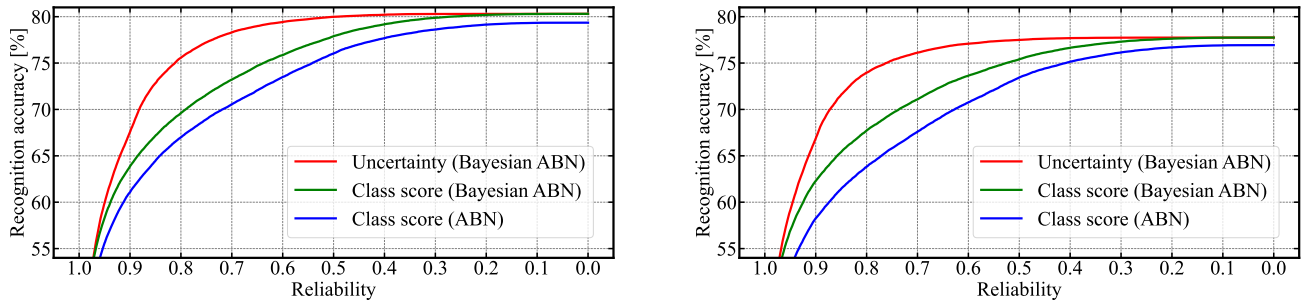


(a) ResNet (b) WRN

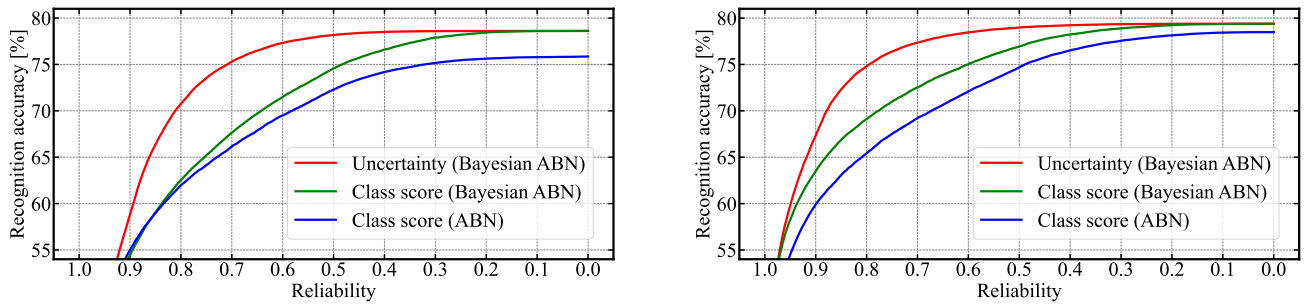


(c) DenseNet (d) ResNeXt

Fig. 5: Recognition accuracy over different reliability (uncertainty or class score) threshold on the CIFAR-100 dataset.



(a) ResNet (b) WRN



(c) DenseNet (d) ResNeXt

Fig. 6: Recognition accuracy over different reliability (uncertainty or class score) threshold on the ImageNet-1K dataset.

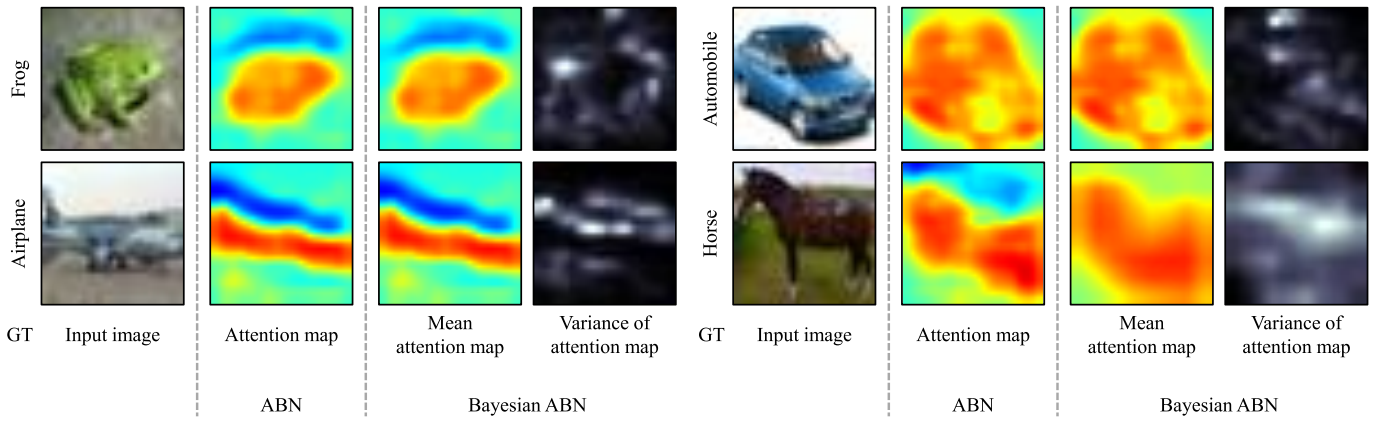
[6] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, vol. 48, 2016, pp. 1050–1059.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

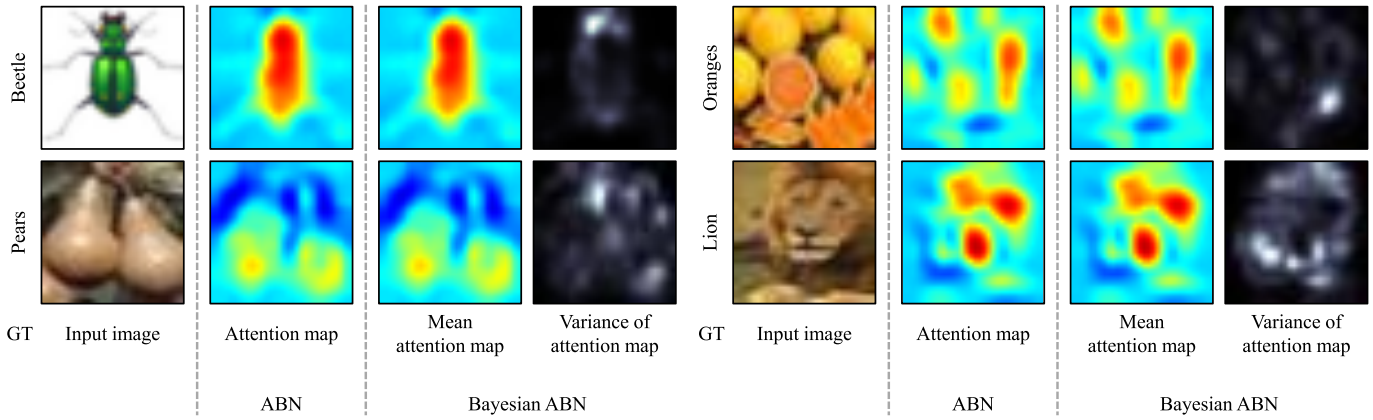
[8] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[9] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 1019–1027.

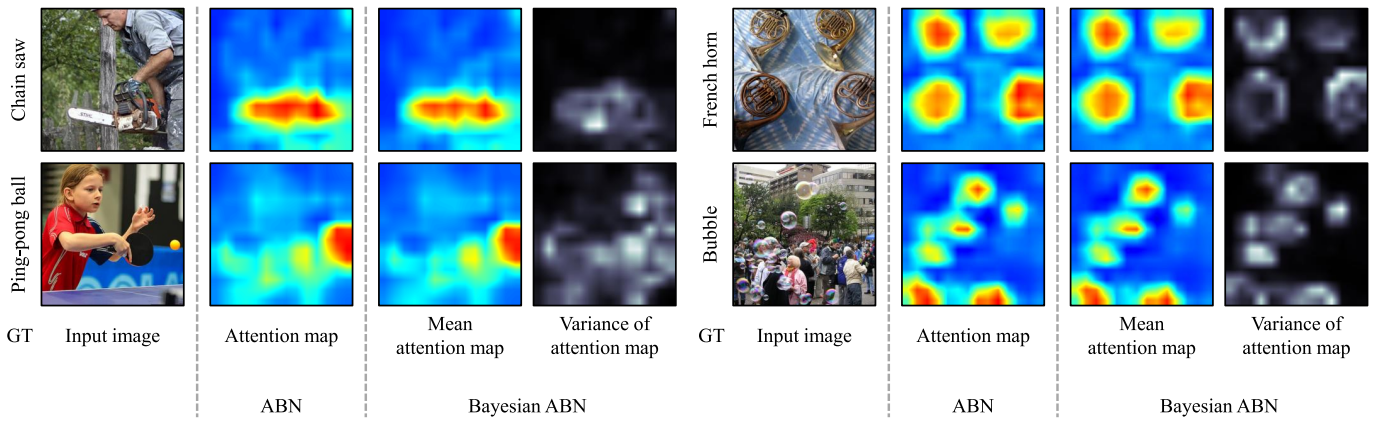
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image



(a) CIFAR-10 dataset



(b) CIFAR-100 dataset



(c) ImageNet-1K dataset

Fig. 7: Visualized attention maps

recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2014.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Computer Vision and Pattern Recognition*, 2017, pp. 5987–5995.

[13] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, University of Tront, 2009.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[15] S. Zagoruyko and N. Komodakis, "Wide residual networks," *British Machine Vision Conference*, pp. 87.1–87.12, 2016.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.