Relevance Regularization of Convolutional Neural Network for Interpretable Classification

Chae Hwa Yoo, Nayoung Kim, and Je-Won Kang
Department of Electronic and Electrical Engineering, Ewha W. University, Seoul, Korea.
yulove3274@ewhain.net, 12skdud21@ewhain.net, and jewonk@ewha.ac.kr

Abstract

Conventional end-to-end learning algorithm considers only the final prediction output and ignores layer-wise relational reasoning during the training. In this paper, we propose to use a forward and backward interacted-activation (FBI) loss function that regularizes training a CNN so that the prediction model can provide interpretable results for classification. From our best knowledge, the proposed algorithm is the first work to use a regularization function without any prior knowledge or pre-defined terms to allow for a CNN to be more explainable. It is demonstrated with quantitative and qualitative analysis that the proposed technique can be used for efficiently train a CNN with more interpretability, applied to a well-known classification problem.

1. Introduction

Convolutional Neural Network (CNN) gains much attention as it has shown substantially improved prediction accuracy and apparently intelligent behaviors to solve many challenging problems [10]. Despite the excellent performance, however, the CNNs are not readily employed in industries such as medical services and autonomous driving because the prediction models provide little information why they make certain decisions. The large number of model parameters and nonlinear activation functions may maximize the prediction accuracy but make the prediction model hardly interpretable [13]. To improve more interpretability to the CNN, various researches are actively carried out in this field, namely explainable artificial intelligent (AI) [1, 12, 4].

Conventional end-to-end learning through a forward and a backward propagation considers only the final prediction output and ignores layer-wise relational reasoning during the training. In this paper, we propose a novel regularization technique using a new training loss function called a forward and backward interacted-activation (FBI) loss function to improve the interpretability of a CNN. The FBI is defined as a sum of layer-wise differences between neuron activations to the forward and the backward directions. The activation maps are propagated from the previous layers during a training, and, thus an intermediate run between a conventional forward and a back propagation is presented to reflect the loss in the proposed algorithm. It is demonstrated with quantitative and qualitative results that the proposed technique can be used for efficiently train a CNN with more interpretability, applied to a well-known classification problem [9].

2. Related work

As many domains of industries are interested in applying AI systems and having benefits from them, interpretable deep learning models become quite important topics in the literatures. The studies can be categorized to intrinsic interpretation methods [12, 11, 3, 7, 4, 14] and post-hoc interpretation methods [2, 6, 15, 5].

Intrinsic interpretable methods focus on modifying an internal structure of a complex black-box model for more interpretability [12]. Palm *et al.* [11] develop relation networks (RN) as additive sub-architectures to the original networks. The sub-networks are designed explicitly for computing relational reasoning. Goudet *et al.* [7] propose a causal network model to infer how a model can learn joint distributions of input data generatively. In [4, 3, 14], the authors introduce semantic templates from the human descriptions and use them as training constraints, so that learned kernels cannot be much deviated from human expectation. However, they need pre-defined semantic descriptions, requiring extra human labors.

Post-hoc interpretable methods are used for reverse engineering processes or visualization purposes to verify the reasoning of a system and give detailed analysis to human experts. In [6] Partial Differential Plot that is to visualize a partial relationship between one or more input variables and their impacts to the prediction results is presented. In [2], Layer-wise Relevance Propagation (RLP) is proposed

to visualize contributions of local pixels in an image to classification performance. Zhang *et al.* [15] and Fisher *et al.* [5] try to record how the model performance can vary with changing inputs or internal components.

3. Proposed Algorithm

3.1. Proposed Loss Function

We propose to use a FBI loss function that regularizes a training of a CNN so that the prediction model can yield explainable results for a classification. From our best knowledge, the proposed algorithm is the first work to use a regularization function without any prior knowledge or predefined terms to allow for a CNN to be more explainable. We compute relational layer-wise costs recursively in a way that the current activation is distributed from the previous layer back and forth. The distribution is determined depending on the activations and filter weights, motivated by an important insight from the work of Bach *et al.* [2].

We define a set of forward activation maps $\mathbf{a}_i^+ = [a_{i,1}^+, a_{i,2}^+, ..., a_{i,C_i}^+]$ of i-th layer during forward propagation as

$$\mathbf{a}_i^+ = f(\mathbf{a}_{i-1}^+, \mathbf{w}_i), \tag{1}$$

where C_i is a size of the channels in the i-th layer, and $\mathbf{w}_i = [w_{i,1}, w_{i,2}, ...w_{i,C_i}]$ is a convolutional kernel in each channel. f presents a functional layer, *e.g.* a convolution layer or a fully connected layer, depending on the current layer. It is noted that $\mathbf{a}_i^+$ is the same activation map obtained from the conventional forward propagation in a deep neural network.

We also define a set of backward activation maps $\mathbf{a}_i^- = [a_{i,1}^-, a_{i,2}^-, ..., a_{i,C_{i-1}}^-]$ of the i-th layer. $\mathbf{a}_i^-$ is calculated with

$$\mathbf{a}_i^- = f(\mathbf{a}_{i-1}^+, \sum_{j=1}^{C_i} w_{i,j} \frac{a_{i+1,j}^-}{a_{i,j}^+}), \tag{2}$$

where the kernels are re-normalized with the ratios of $\mathbf{a}_i^+$ and $\mathbf{a}_{i+1}^-$. As shown in Fig. 1, $\mathbf{a}_i^+$ and $\mathbf{a}_{i+1}^-$ are the inward activation maps to the i-th layer.

Then, the FBI loss function is defined as

$$L_{FBI} = \frac{1}{K} \sum_{i=1}^{K} L_1(\mathbf{a}_{i-1}^+, \mathbf{a}_i^-), \tag{3}$$

where K is the number of layers in a network. The index i starts from the first layer, so $\mathbf{a}_0^+$ is an input image of a network. $L_1(.)$ is an l_1 loss function, computing an absolute difference of the two input terms.

Then, the overall loss function is formulated as follows.

$$L = (1 - \lambda) L_{TSK} + \lambda L_{FBI}, \tag{4}$$

where L_{TSK} is a task-specific loss, *e.g.* a cross-entropy loss between an actual label and a predicted label in an image classification problem. λ controls a weight of a FBI loss, set to 0.001 in our experiments. Finally, we train a set of network parameters $\mathcal{Y}$ to minimize the loss in (4), *i.e.*,

$$\mathcal{Y}^* = \arg\min_{\mathcal{Y} \in \mathbf{Y}} L. \tag{5}$$

3.2. Network Training

The proposed algorithm trains a network using three steps: (1) a forward pass, (2) a relevant computing pass, and (3) a backward pass. The forward pass is the same as the conventional training scheme, where an input training sample goes through the network to the end. A task-specific loss is computed after a forward pass. The relevant computing pass is conducted between a forward and a backward pass to collect a FBI loss in each layer, as shown in Fig. 1. While starting from the neurons of the final layer to the first layer, the pass collects each loss in the corresponding layer. After all, the backward pass is performed once all the loss terms are summed up.

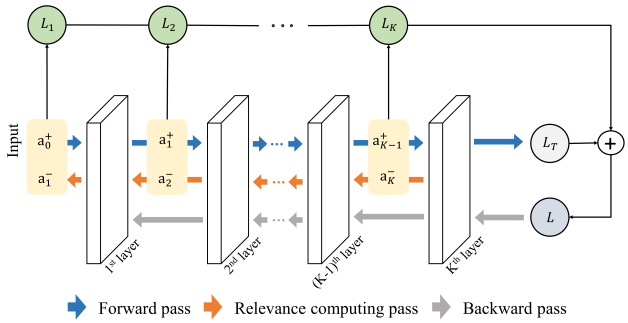

Figure 1. The proposed training scheme.

Since the terms regarding the relevance computing pass are not stabilized enough in an early stage of learning, we need to train the network using only the task-specific loss as a conventional training to some extents. Once the training is saturated with enough epoch numbers and validation errors, the FBI loss begins to be considered in the training. In our experiments, we set the moment when the validation accuracy becomes above 70%.

4. Experimental Results

4.1. Experimental Configurations

We evaluate the interpretability of the proposed algorithm quantitatively and qualitatively. For this, we apply the proposed algorithm for an image classification problem using MNIST [9]. We use an AlexNet [8] model with minor modification in kernel and channel sizes.

4.2. Visualization of Neuron Activations

We show kernel activations of the first three convolutional layers to understand classification decisions using the

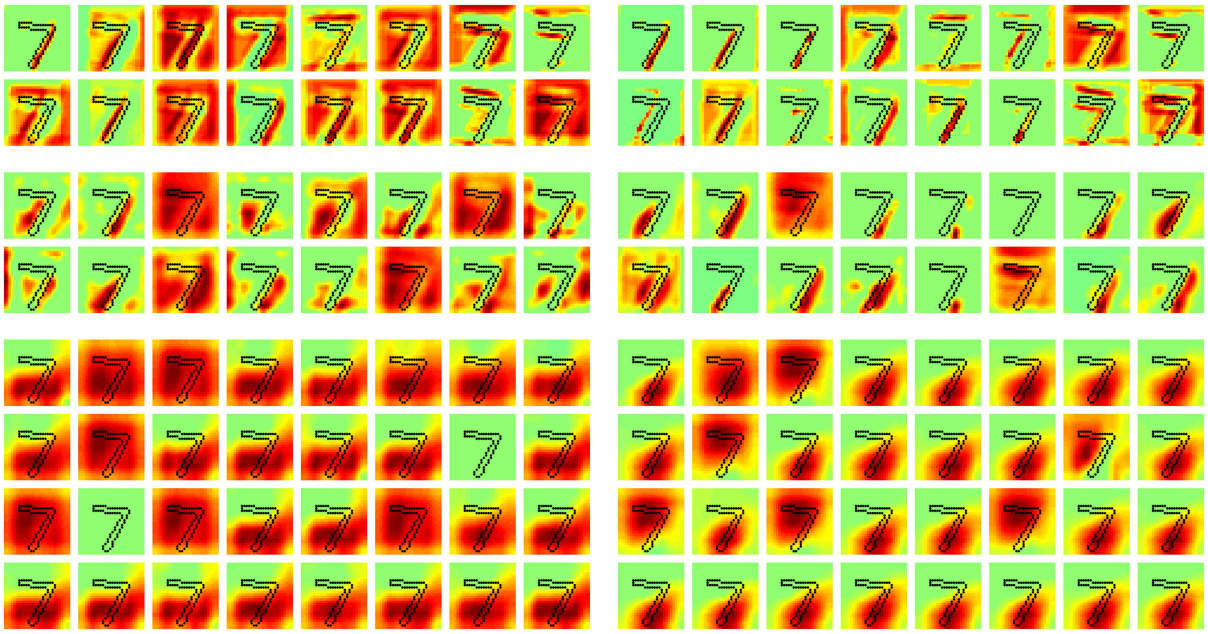

Figure 2. Activations of filters in the three convolution layers, using the "conventional training"**(left)** and the "proposed training"**(right)**. From the top to the bottom, filter activations of the first, second, and third convolutional layers of the tested network.

pixel-wise decomposition. For comparisons, we train the network using a task-specific loss with a back-propagation algorithm, denoted by "conventional training" and a FBI loss with the three-step training algorithm, denoted by "proposed training." Fig. 2 show the results of the conventional training (left) and the proposed training (right). Note that the heatmaps of both training are normalized within same range for better comparison. From the top to the bottom, those figures present the activations in the three convolutional layers, where a number of the corresponding filters as 16, 16, and 32, respectively.

As shown in the heat maps, the red color implies higher activations, where the filters actually see. In contrast, the green implies lower activations close to 0. It is clearly seen that the proposed algorithm facilitates the CNN to learn a region or a feature which it actually expects. For instance, the number 7 has distinguished features near edges and contours from the other letters. However, the CNN trained with the conventional algorithm often shows strong activations in the background, as depicted in the first row of the filter responses. On the contrary, the proposed algorithm tries to avoid any confusion to predict the character by focusing on a region of interest. We also observe in the last rows (*i.e.*, the filter activations of the last convolution layers) that there are more zero-out filters in the proposed algorithm. It implies that the proposed algorithm achieves more energy compaction, and, thus gives more confidence to specific neurons.

conventional / **relevance**					
layer	conv1	conv2	conv3	conv4	conv5
max	1.377 / **2.576**	1.270 / **1.205**	0.941 / **1.647**	2.355 / **3.883**	14.741 / **16.524**
skew-ness	0.603 / **2.576**	1.484 / **2.086**	0.806 / **1.484**	0.707 / **1.314**	0.651 / **1.012**

Table 1. Maximum and skewness of activations from all convolutional layers in the Alexnet, using "conventional training" and the "proposed training".

4.3. Quantitative Analysis of Neuron Activations

For a quantitative analysis, we measure the energy compaction of the activations of all convolutional layers of the Alexnet. In Fig. 3 blue and red curves indicate the conventional training algorithm and the proposed relevant training algorithm, respectively. In Fig. 3, the x-axis and the y-axis present an activation value and a density, respectively. The activation values are always positive since they are obtained after the rectification in the layer. As shown, the curves of the proposed algorithm are more skewed to 0 because the activations are more focused to specific regions. For quantitative comparison, we measure the maximum and the skewness, denoting the maximal value and the inverse of the variance in the energy-normalized curves. The values are shown in Table 1. The larger the skewness value is, the more skewed the distributions are. It is shown in Table 1 that the values are larger when using the proposed training. This is because that the trained model has higher

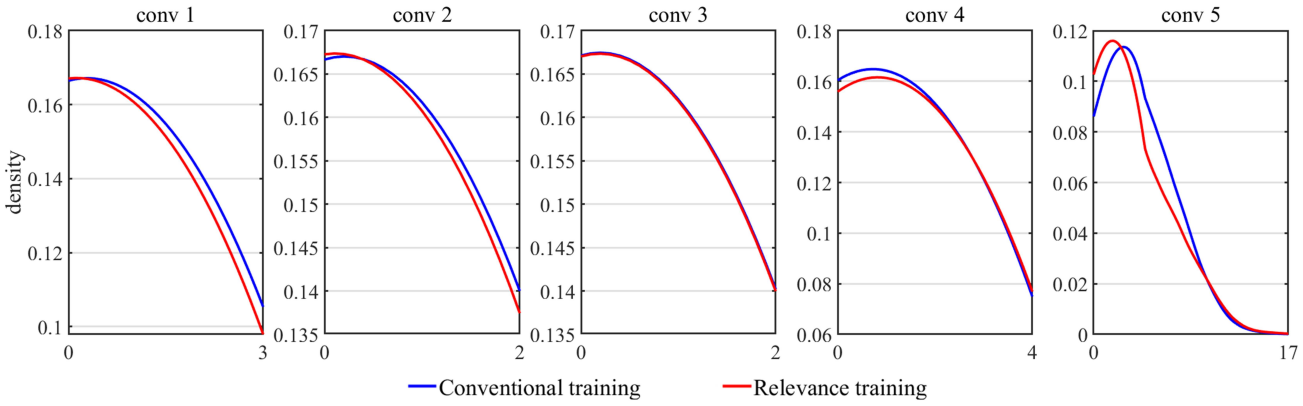

Figure 3. probability density estimation of layer-wise activations. Blue and red lines indicate the conventional algorithm and the proposed algorithm, respectively.

confidence (a larger maximum value) and lower confusion (a large skewness).

5. Conclusion

We proposed a new regularization method to enhance interpretability of a CNN model. For this, we introduced a FBI loss function and the relevance computing pass during the training. It was demonstrated with qualitative and quantitative analysis that the proposed algorithm provides more interpretability with the trained filters. In the future work, we will evaluate the proposed technique with more datasets and various network models.

Acknowledgment

This work has supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.NRF-2019R1C1C1010249)

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] Yinpeng Dong, Hang Su, Jun Zhu, and Bo Zhang. Improving interpretability of deep neural networks with semantic information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4306–4314, 2017.

[4] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*, 2018.

[5] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*, 2018.

[6] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*.

[7] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

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

[9] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[11] Rasmus Palm, Ulrich Paquet, and Ole Winther. Recurrent relational networks. In *Advances in Neural Information Processing Systems*, pages 3372–3382, 2018.

[12] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 19–36. Springer, 2018.

[13] Sanjoy Sarkar, Tillman Weyde, A Garcez, Gregory G Slabaugh, Simo Dragicevic, and Chris Percy. Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In *CEUR Workshop Proceedings*, volume 1773. CEUR Workshop Proceedings, 2016.

[14] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018.

[15] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.