A SELF-EXPLANATORY METHOD FOR THE BLACK PROBLEM ON DISCRIMINATION PART OF CNN

Anonymous authors
Paper under double-blind review

ABSTRACT

Convolution neural networks (CNNs) have surpassed human's abilities in some specific tasks. However, they are considered difficult to understand and explain. Recently, the black box problem of CNN, especially concerning the discrimination part, has been studied by different scientific communities. Many methods were proposed for extracting an interpretable model from the discrimination part, which can explain the prediction of the part. However, it is hard for the interpretable models to approximate the discrimination part because of the tradeoff problem between interpretability performance and generalization performance of the discrimination part. We suppose the tradeoff problem is mainly attributed to the fact that the sufficient and necessary condition for the consistent convergence of the both performances is hard to be guaranteed by tradition learning algorithm. So the tradeoff problem could be solved by shrinking the distance between the interpretable model and the discrimination part. This paper firstly introduces a Markov random field model (MRF), namely Deep Cognitive Learning Model(DCLM), which explains the causal relationship between the features(the weight matrixes in the first layer of the discrimination part of CNN can capture) and the output result of the discrimination part of CNN. A greedy algorithm is proposed for initiatively extracting the DCLM from the discrimination part by solving a MAX-SAT problem. And then, a game process between two MAP inferences is implemented for shrinking an interpretation distance which can evaluate how close the discrimination part is to the DCLM. Finally, the proposed self-explanatory approach is evaluated by some contrastive experiments with certain baseline methods on some standard image processing benchmarks. These experiments indicate that the proposed method can improve the interpretability performance of the discrimination part without largely reducing its generalization performance, the generalization performance of the DCLM also can be improved and the discrimination part can be explained by the DCLM in real time during the training process after the interpretation distance converges.

1 INTRODUCTION

Convolution neural network(CNN) has surpassed human abilities in some specific tasks such as computer game and computer vision etc. However, they are considered difficult to understand and explain(Brandon, 2017), which leads to many problems in aspects of fairness, privacy leaking, reliability and robustness. Even wrong decision could possibly be made due to spurious correlations in training data, such as recognizing an object in an image by the properties of the background or lighting (Riccardo et al., 2018). Explanation technology is of immense help for companies to creat safer, more trustable products, and to better manage any possible liability of them (Riccardo et al., 2018). Recently, the unexplainable problem of CNN, especially concerning the discrimination part, has been studied by different scientific communities. Many methods were proposed for extracting an interpretable model from the discrimination part, which can explain causal relationship between inputs and outputs of the discrimination part. However, because of data bias and noisy data in training data set the sufficient and necessary condition for the consistent convergence of an interpretability performance and the generalization performance of the discrimination part is difficult to be guaranteed in traditional machine learning methods. The interpretability performance is a distance between the discrimination part and its optimal interpretable model. So when a discrimination part has a well

generalization performance, it usually stay away from its optimal interpretable model, which leads to an incorrect interpretable model. We think that the tradeoff problem could be solved by the following procedure. Firstly, extracting an interpretable model which expresses causal relationships in the discrimination part in training process, and then, iteratively shrinking a distance between the model and the part, and amending them. It is because that the method not only can make the discrimination part approximate the interpretable model which can improve the interpretability performance of the part but also can make the interpretable model close to the part which makes it correctly express the causal relationship in the part.

Main contributions of this paper can be summarized as follows:

- An interpretable model, Deep Cognitive Learning Model(DCLM), is proposed to express the causal relationship in the discrimination part, and a greedy method is given for initiatively extracting the DCLM from the discrimination part by solving its MAX-SAT problem.
- A new game method is proposed to improve the interpretability performance of the discrimination part without largely reducing its generalization performance by iteratively shrinking a distance between the two models and amending the two models. At the same time, the generalization performance of the DCLM can also be improved.
- Interpretable distance is proposed to evaluate a distance between the discrimination part and the interpretable model on the unexplainable problem.

2 Related work

There are usually two types of methods for the unexplainable problem of the discrimination part, such as post-hoc method and ante-hoc method (Holzinger et al., 2019). However, because ante-hoc method is a transparent modeling method(Arrietaa et al., 2020), it can not obtain an explanation about the discrimination part. So, the post-hoc method will be reviewed.

Early post-hoc method can obtain global explanations for a neural network by extracting an interpretable model. Some references(Craven & Shavlik, 1999; Krishnan et al., 1999; Boz, 2002; Johansson & Niklasson, 2009) proposed a few methods that can find a decision tree for explaining a neural network by maximizing the gain ratio and an estimation of the current model fidelity. Other references (Craven & Shavlik, 1994; Johansson & Niklasson, 2003; Augasta & Kathirvalavakumar, 2012; Sebastian et al., 2015; Zilke et al., 2016) proposed rule extraction methods for searching optimal interpretable rules from a neural network.

Recently, some feature relevance methods have become progressively more popular. Montavon et al.(Montavon et al., 2017) proposed a decomposition method from a network classification decision into contributions of its input elements based on deep Taylor decomposition. Shrikumar et al.(Shrikumar et al., 2016) proposed DeepLIFT which can compute importance scores in a multilayer neural network by explaining the difference of the output from some reference output in terms of differences of the inputs from their reference inputs.

Some other works make complex black box model simpler. Che et al.(Che et al., 2017) proposed a simple distillation method called Interpretable Mimic Learning for extracting an interpretable simple model by gradient boosting trees. Thiagarajan et al.(Thiagarajan et al., 2016) build a Treeview representation of the complex model by hierarchical partitioning of the feature space. In addition, some references (Hinton et al., 2015; Bucila et al., 2006; Frosst & Hinton, 2017; Traore et al., 2019) proposed the distillation method of knowledge from an ensemble of models into a single model. Wu et al.(M. Wu, 2018) proposed a tree regularization method via knowledge distillation to represent the output feature space of a RNN based on a Multilayered perception. However, these methods can only solve the unexplainable problem of trained neural network or trained deep neural networks with explicit input characteristics. Wan et al.(Wan et al., 2020) constructed a decision tree using the last fully connection layer of the discrimination part of a CNN based on a prior structure.

In the paper, our goal is to improve the interpretability performance of multi full connection layers of CNN without hurting its generalization performance by initiatively extracting its logic relationships with no prior structure and finally obtain its explanation by these logic relationships.

3 DEEP COGNITIVE LEARNING MODEL

Convolution layers and pooling layers of a CNN constitute a feature extractor. The outputs from the feature extractor are the inputs to the fully connected layers of the CNN, namely feature maps, $\tau_1, \tau_2, ..., \tau_k$ where k is the number of feature maps. All these feature maps form a feature set Γ. The objective of the fully connected layers is to use the feature maps to classify the input image x of the CNN into a label. So, the fully connected layers compose a discrimination part of the CNN.

We suppose that the discrimination part should better be explained by the logic relationships between the features(the weight matrixes in the first layer of the part can capture) and its output vector. This is because that the relationships is the inherent relation of such CNN structure which is neither related to the input data nor related to the extent of the features' abstract. In order to express the relationships, deep cognitive learning model (DCLM)is proposed. The DCLM is shown in Fig.1.

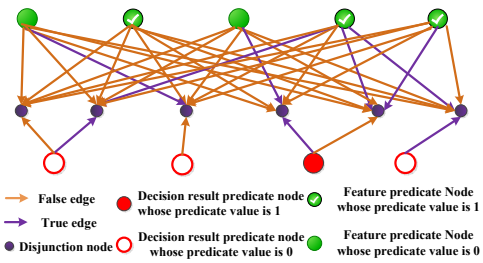

Figure 1: Deep Cognitive Learning Model

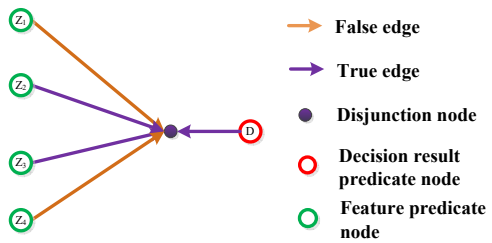

Figure 2: A disjunction relation in DCLM

The DCLM consists of three layers of nodes:feature predicate layer, disjunction layer and decision result predicate layer. The top layer is the feature predicate layer which consisits of feature predicate nodes. Every feature predicate node has a feature predicate function $Z(\Gamma)$ which expresses whether a neuron in the first fully connected layer of discrimination part of the CNN can capture a feature. The feature predicate function $Z(\Gamma)$ is defined as follows.

$$Z(\Gamma) = \begin{cases} 1, & \frac{\tau_i * w_i}{\|\tau_i\|\|w_i\|} > 0 \text{ and } \tau_i \in \Gamma, \quad (1) \\ null, & \frac{\tau_i * w_i}{\|\tau_i\|\|w_i\|} = 0 \text{ and } \tau_i \in \Gamma, \quad (1') \\ 0, & otherwise. \quad (1'') \end{cases}$$

where $i \in \{1, 2, ..., k\}$, τ_i is the ith feature map in Γ, w_i is its corresponding weight vector on the first fully connection layer, namely feature-capture map and "$*$" is a convolution.

The bottom layer is the decision result predicate layer which includes all decision result predicate nodes. Every decision result predicate node has a decision result predicate function which expresses whether the output y of an output neuron of the discrimination part is greater than 0. It is defined as following as

$$D(y) = \begin{cases} 1, & y > 0, \quad (2) \\ 0, & otherwise. \quad (2') \end{cases}$$

All feature predicate nodes and every decision result predicate node are connected to one or more disjunction nodes in the middle layer, namely as disjunction layer, with true or false edges. Every disjunction node represents a disjunction relation between all feature predicate nodes and a decision result predicate node. A disjunction relation is expressed by a disjunctive normal form. If a predicate node is connected to the disjunction node by a false edge, its predicate function follows after a non-operator in the disjunctive normal form. Potential function of a disjunctive normal form can be obtained by using the Lukasiewicz method(Giles, 1975).

$$\phi_c(y) = \min(1, T(\Gamma, y)) \quad (3)$$

where $T(\Gamma, y) = \sum_{j=1}^{N} \{a_j[1 - Z_j(\Gamma)] + (1 - a_j)Z_j(\Gamma)\} + D(y)$ and N is the number of the feature predicate nodes. If $a_j = 1$, there is a false edge between predicate node and a disjunction node.If $a_j = 0$, there is a true edge.

In Fig.2 suppose a DCLM includes a disjunction relation $\forall\Gamma\neg Z_1(\Gamma) \vee Z_2(\Gamma) \vee Z_3(\Gamma) \vee \neg Z_4(\Gamma) \vee D(y)$ which is equivalent to $\forall\Gamma[Z_1(\Gamma) \wedge \neg Z_2(\Gamma) \wedge \neg Z_3(\Gamma) \wedge Z_4(\Gamma) \rightarrow D(y)]$, where $Z_j()$ is a feature predicate function and y is an output of an output neuron of CNN. Its potential function is $\phi_c(y) = \min(1, \sum_{i=1}^{4}\{a_j[1 - Z_j(\Gamma)] + (1 - a_j)Z_j(\Gamma)\} + D(y))$ where $a_1 = a_4 = 1$ and $a_2 = a_3 = 0$.

The conditional probability distribution that a ground DCLM is true is

$$p(y_{dclm}, \Gamma) = \frac{1}{\Xi}\exp(\frac{\sum_{i=1}^{G}\lambda_i\phi_{ci}(y_{dclm,i})}{\sum_{i=1}^{G}\lambda_i}) \tag{4}$$

where G is the number of all ground formulas, $\Xi = \sum_{\Gamma\in\mathbb{F}}\exp(\frac{\sum_{i=1}^{G}\lambda_i\phi_{ci}(y_{dclm,i})}{\sum_{i=1}^{G}\lambda_i})$ is a partition function, $y_{dclm} = (y_{dclm,1}, y_{dclm,2}, ..., y_{dclm,G})$, $y_{dclm,i}$ is an output value of an output neuron of the CNN and λ_i is a weighted value of the ith ground formula.

By maximizing its likelihood function, the optimal y_{dclm}, a_i and λ_i in the DCLM can be obtained.

$$C(\Gamma) = arg\max_{y_{dclm},a_i,\lambda_i}[\log p(y_{dclm}, \Gamma)] = arg\max_{y_{dclm},a_i,\lambda_i}(\frac{\sum_i\lambda_i\phi_{ci}(y_{dclm,i})}{\sum_i\lambda_i}) \tag{5}$$

For extracting a DCLM, a Maximum A Posterior(MAP) algorithm on the Maximum Satisfiability Problem (MAX-SAT) was designed, which is shown in the appendix.

4 Evaluation of interpretability performance

We consider that if the shape of function curve of the discrimination part of a CNN is similar as that of its optimal interpretable model, it can be very easily understood by humans and has well interpretability performance. Therefore, the interpretable performance of the discrimination part can be measured by the shape similarity between it and the optimal interpretable model. We posit that given the same input data set, the similarity may be measured by variance of differences between outputs of the both models. It can tentatively be named interpretation distance. It is easily proved that the smaller the interpretation distance is, the more similar the shape of the discrimination part is to the shape of the optimal interpretable model and the better the interpretability performance of the discrimination part would be.

Definition 1 If X is a compact metric space and ν is a Borel measure in X, such as Lebesgue measure or marginal measures, in $\mathcal{L}_\nu^2(X)$, a square integrable function space on X, the interpretation distance, $\phi_d(P^*, f)$, between a discrimination part $f(x)$ and its optimal interpretable model $P^*(x)$ is

$$\phi_d(P^*, f) = \int_Z (f(x) - P^*(x) - \mu^{P^*}(f))^2 d\nu \tag{6}$$

where

$$\mu^{P^*}(f) = \int_Z (f(x) - P^*(x))d\nu \tag{7}$$

In the definition, the optimal interpretable model can have many forms, such as DCLM, other probabilistic graphical model, decision tree, decision rules and so on.

5 Game between a DCLM and the discrimination part of a CNN

As discussed above,when the shapes of the discrimination part of a CNN and its optimal interpretable model are enough similar, the discrimination part has well interpretability performance. However, its generalization performance will tend to decrease. This is mainly attributed to the fact that the sufficient and necessary condition for the consistent convergence of the two performances, $\phi_d(P^*, f^*) = 0$(f^* is the optimal predication model), is difficult to be guaranteed. For proving the conclusion, we focus on a neuron of a CNN. From the foot we may judge of Hercules. Because every input channel $f(x)$ of the neuron can be seen as a kernel function $K(x, w)$(w is a weight vector including a bias of the neuron), it may span a kernel Hilbert space $\mathcal{H}_K = \{f(x) \in \mathcal{L}_\nu^2(X) \mid f(x) = K(x, w) = \sum_{k=1}\phi_k(w)\phi_k(x)\}$. $\mathcal{H}_K$ is a linear function set on

Algorithm 1 Game between DCLM and the discrimination part of a CNN(Its time complexity(TC) is $O(N+M)$ where N is TC of training CNN,M is TC of construction of Logic Net.)

Input: data X, target Y_t
initialize Logic Net LN, CNN CN
for $i = 1$ **to** $batch - size$ **do**
  $FMs, f_{nn} = CN(X[i])$.
  $y_{dclm} = LN(FMs, Y_t[i])$.
  $CN = Updata_CNN_Gradient(y_{dclm}, f_{nn}, y_t[i])$
  $FMs, FCMs, f_{nn} = CN(X[i])$.
  $LN = ConstructionLogicNet(FMs, FCMs, f_{nn})$
end for

$\mathcal{L}_\nu^2(X)$. However, because of data bias and noisy data in training data set, its optimal interpretable model which obey to human's cognition usually is a nonlinear function in $\mathcal{L}_\nu^2(X)$ in the majority of engineering applications. Because continuous linear functional set is nowhere dense in $\mathcal{L}_\nu^2(X)$. So the sufficient and necessary condition is hard to be satisfied and the tradeoff problem always exists. This is also the case with the discrimination part of a CNN. However its optimal interpretable model P^* and its optimal discrimination part are unknown. Therefore, for the tradeoff problem, in training process, extracting an interpretable model from the discrimination part and then iteratively reducing the interpretation distance between the two models may be a feasible solution. A detailed discussion about the problem can be found in the appendix.

To avoid to reduce the generalization performance, the maximum probability $p(w \mid X, y_t)$ should be guaranteed, where X is a training sample, w is parameter set of the CNN and y_t is target vector of X.

$$p(w \mid X, y_t) = \frac{p(w \mid X)p(y_t \mid w, X)}{p(y_t \mid X)} \propto p(w)p(y_t \mid w, X) \tag{8}$$

where $p(y_t \mid w, X) = \int p(y_t \mid f, w, X) \int p(f \mid y_{dclm}, w, X)p(y_{dclm} \mid w, X)dy_{dclm}df$.

When the DCLM is known, y^*_{dclm} is its optimal solution and $p(y^*_{dclm} \mid w, X) = 1$. Then

$$\int p(f \mid y_{dclm}, w, X)p(y_{dclm} \mid w, X)dy_{dclm} = p(f \mid y^*_{dclm}, w, X) \tag{9}$$

Similarly, known the input X and w, f_{nn} is the optimal solution of the CNN.

$$p(y_t \mid w, X) = p(y_t \mid f_{nn}, w, X)p(f_{nn} \mid y^*_{dclm}, w, X) \tag{10}$$

If w and X are given and the loss function $\phi_r(y_t, f_{nn}) = -\frac{1}{2}\sum_l \mid y_t - f_{nn} \mid^2$, the conditional probability distribution function

$$p(y_t \mid f_{nn}, w, X) = \frac{\exp(\phi_r(y_t, f_{nn}))}{\Xi_1} \tag{11}$$

Meanwhile,

$$p(f_{nn} \mid y^*_{dclm}, w, X) = \frac{\exp(-\phi_d(y^*_{dclm}, f_{nn}))}{\Xi_2} \tag{12}$$

where Ξ_1 and Ξ_2 are partition functions. Then by maximizing a likelihood function of $p(w \mid X, y_t)$ the optimal w can be obtained. In particular, assuming that w follows Gaussian distribution, we get:

$$C_w(X, y_t) = arg\max_w[-\frac{\alpha}{2} \parallel w \parallel^2 + \phi_r(y_t, f_{nn}) - \phi_d(y^*_{dclm}, f_{nn})] \tag{13}$$

where α is a meta-parameter determined by the variance of the selected Gaussian distributions. Turn it into a minimization problem:

$$C_w(X, y_t) = arg\min_w[\frac{\alpha}{2} \parallel w \parallel^2 - \phi_r(y_t, f_{nn}) + \phi_d(y^*_{dclm}, f_{nn})] \tag{14}$$

The iterative optimization algorithm is shown as follows:

6 EXPERIMENTAL VERIFICATION

We designed some experiments to verify the effectiveness of the proposed method. The first experiment verified whether the self-explanatory method could improve the interpretability performance of the CNN without sacrificing its generalization performance. The second experiment verified whether the proposed method can tend towards stability and convergence in the game process.

In the experiments, the structure of the CNN3(includes 3 convolution layers,3 MaxPooling layers, 3 fully connect layers(FCLs) and 1 output layer),CNN5(includes 5 convolution layers,5 MaxPooling layers, 3 FCLs and 1 output layer) and CNN8(includes 8 convolution layers,8 MaxPooling layers, 3 FCLs and 1 output layer) were used. The final structures obtained after the training on these CNNs by the proposed method are named as CNN3-DCLM, CNN5-DCLM and CNN8-DCLM respectively. All experiments used Mnist(Lecun et al., 1998), FashionMnist(Zalando, 2017) and Emnist(Cohen et al., 2017) benchmark data sets. All algorithms were implemented in Python using the Pytorch library(Paszke et al., 2019). All experiments ran on a server with Intel Xeon 4110(2.1GHz) Silver Processor, 20GB RAM and Nvidia Telsa T4.

Experiment 1: Performance verification of the proposed method on CNN. The experiment also compared soft decision tree(SDT) (Frosst & Hinton, 2017) based on trained CNN3,CNN5 and CNN8,namely "CNN3-SDT", "CNN5-SDT" and "CNN8-SDT" respectively. These baseline methods also include SDT (Frosst & Hinton, 2017) based on trained CNN3-DCLM,CNN5-DCLM and CNN8-DCLM, namely "CNN3*-SDT", "CNN5*-SDT" and "CNN8*-SDT" respectively. These indicators of all algorithms on all testing data sets were shown in Table 1. Some values are "—", which indicates that these results do not exist.

It is observed in Table 1 that the accuracy of all CNN-DCLMs are higher than the two interpretable models, such as SDT and DCLM, on all benchmark data sets and are around 1.4 percentage points lower than those of CNN. But it is worth noticing that on the most of data sets the interpretation distances of all CNN-DCLMs are around 5 percentage points lower than interpretation distances between the majority SDTs and their CNNs except CNN3-SDT on Emnist data set. These might prove that that the self-explanatory method can improve the interpretability performance of the discrimination part of a CNN without largely reducing its generalization performance.

On Mnist data set the accuracies of the DCLMs are only 0.7 percentage points lower than those of the SDTs on CNN5-SDT and CNN8-SDT except CNN3-SDT. Though for Emnist data set and FashionMnist data set,we can find that the accuracies of the DCLMs are 2 percentage points lower than those of the SDTs on CNN3-SDT, we also can find that on the two data sets the accuracies of the DCLMs are around 4.3 percentage points higher than those of the SDTs on CNN5-SDT and CNN8-SDT. It is mainly because that the feature extract parts of CNN5 and CNN8 output more abstract feature maps than CNN3 for Emnist data set and FashionMnist data set. The abstract feature maps can be easily expressed without distortion by feature predicates of the DCLM, so it does not impede the generation of the DCLM.

Experiment 2: Convergence test of the proposed method We designed the experiments to demonstrate convergence of the proposed method. CNN3, CNN5 and CNN8 were used for comparing with CNN3-DCLM, CNN5-DCLM, and CNN8-DCLM respectively. Every training works out 25 epoches. All results were measured at every epoch and shown as the four figures, Fig.3, Fig.4, Fig.6 and Fig.5. Every figure includes nine subplots. The three subplots on the left column were shown for the experiments on Mnist data set. These subplots on the middle column were for FashionMnist data set and these subplots on the right column were for Emnist data set.

In Fig.3, the accuracies of the DCLMs and the CNN-DCLMs of every epoch in the game process were shown. From these figures, it is obvious that accuracies of the DCLMs and these CNN-DCLMs steadily increase in the early stage. In the next stage, their accuracies tend to be stable. This reflects that the game method did not affect the improvement of the generalization performances of these DCLMs and these CNNs.

Fig.4 shows the interpretation distances between the DCLMs and the CNNs and CNN-DCLMs. As seen from these subplots, the interpretation distances of CNNs not involved in the game are greater than those of CNN-DCLMs at the most of the epochs, especially by the end of the training. The results indicate that the game method can effectively improve the interpretability performance of CNN-DCLMs.

Table 1: Classification accuracies and interpretation distances for three game methods on DCLM-CNN.

DATA SET: MNIST			
GAME METHOD	ACCURACY	ACCURACY OF SDT OR DCLM	INTERPRETATION DISTANCE
CNN3	0.988±0.012	—	—
CNN3-SDT	—	0.970±0.017	0.005±0.0039
CNN3*-SDT	—	0.953±0.014	0.008±0.0043
CNN3-DCLM	0.983±0.016	0.970±0.018	0.002±0.0023
CNN5	0.988±0.013	—	—
CNN5-SDT	—	0.968±0.017	0.005±0.0046
CNN5*-SDT	—	0.959±0.020	0.006±0.0032
CNN5-DCLM	0.983±0.015	0.962±0.022	0.002±0.0024
CNN8	0.991±0.012	—	—
CNN8-SDT	—	0.978±0.012	0.003±0.0019
CNN8*-SDT	—	0.974±0.014	0.006±0.0039
CNN8-DCLM	0.987±0.013	0.970±0.023	0.002±0.0022

DATA SET: EMNIST			
GAME METHOD	ACCURACY	ACCURACY OF SDT OR DCLM	INTERPRETATION DISTANCE
CNN3	0.982±0.014	—	—
CNN3-SDT	—	0.949±0.021	0.007±0.0035
CNN3*-SDT	—	0.863±0.032	0.022±0.0088
CNN3-DCLM	0.974±0.013	0.911±0.041	0.010±0.0043
CNN5	0.985±0.016	—	—
CNN5-SDT	—	0.870±0.034	0.022±0.0094
CNN5*-SDT	—	0.942±0.026	0.082±0.0224
CNN5-DCLM	0.973±0.012	0.942±0.027	0.006±0.0038
CNN8	0.989±0.012	—	—
CNN8-SDT	—	0.872±0.031	0.025±0.0092
CNN8*-SDT	—	0.855±0.039	0.037±0.0092
CNN8-DCLM	0.974±0.011	0.923±0.034	0.007±0.0042

DATA SET: FASHIONMNIST			
GAME METHOD	ACCURACY	ACCURACY OF SDT OR DCLM	INTERPRETATION DISTANCE
CNN3	0.905±0.029	—	—
CNN3-SDT	—	0.842±0.041	0.083±0.0355
CNN3*-SDT	—	0.744±0.049	0.058±0.0150
CNN3-DCLM	0.880±0.034	0.825±0.032	0.016±0.0028
CNN5	0.900±0.027	—	—
CNN5-SDT	—	0.792±0.041	0.064±0.0119
CNN5*-SDT	—	0.715±0.053	0.078±0.0205
CNN5-DCLM	0.883±0.032	0.819±0.036	0.016±0.0028
CNN8	0.905±0.027	—	—
CNN8-SDT	—	0.785±0.043	0.122±0.0397
CNN8*-SDT	—	0.791±0.043	0.053±0.0129
CNN8-DCLM	0.876±0.033	0.811±0.032	0.017±0.0037

From Fig.5 it is evident that all DCLMs of CNN3, CNN5 and CNN8 have been found to have stable entropies at the end of the game. On Mnist data set, the entropies finally converge between 120 and 145. On FashionMnist data set, the entropies finally converge between 125 and 136. On Emnist data set, the entropies finally converge between 100 and 120. The results indicate that the game algorithm can ensure that the DCLM converges to a stable state.

In Fig.6,the accuracies of the CNNs and the CNN-DCLMs of every epoch in the game process were shown. From these subplots, it can be seen that the accuracies of the CNNs and those of the CNN-DCLMs steadily increase in the early stage and accuracies of CNN-DCLMs are lower than

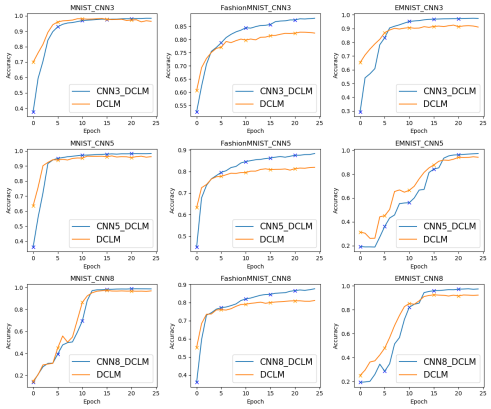

Figure 3: Acc on $DCLM$ and CNN_DCLM

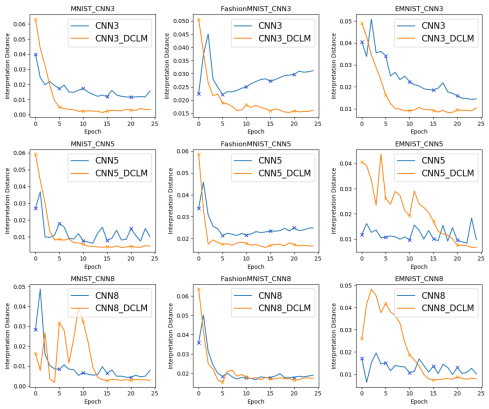

Figure 4: Interpretation distance of $DCLM$

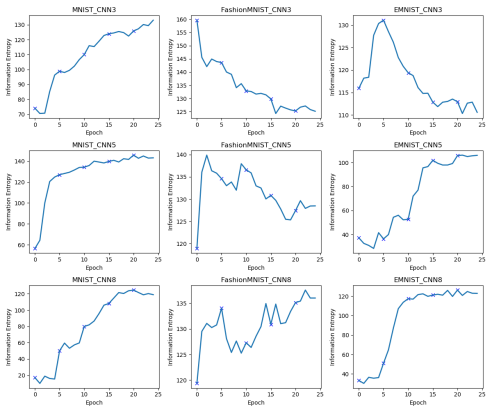

Figure 5: Information Entrop of $DCLM$

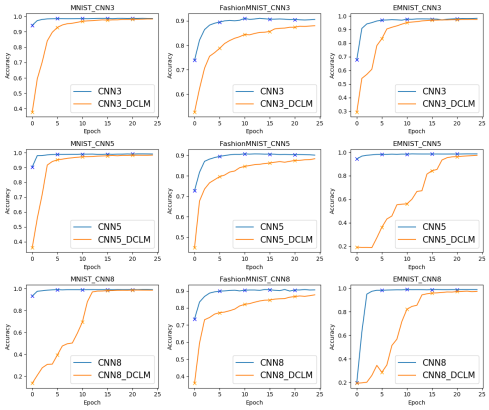

Figure 6: Acc on CNN and CNN_DCLM

those of CNNs. But in the last stage, their accuracies tend to be stable and consistent. The main reason is that in the early stage, a tradeoff problem between the generalization performance and interpretability performance of CNN-DCLMs must reduce its generalization performance of CNN-DCLMs to increase its interpretability performance. But the proposed game method can effectively reduce the gap between the two performances. This reflects that such method is effective for the tradeoff problem.

From these subplots in Fig.3and Fig.4, we also can see that on every epoch a DCLM could be obtained. And after the 15th epoch, their interpretation distances tend to converge. The phenomenon indicates that the game process can decrease the distances and the training process of the discrimination part of the CNNs can be explained by their DCLMs in real time after the fifteenth epoch.

7 Conclusion

The performance of the proposed method was demonstrated by experiments on benchmark data sets. The proposed method showed prominent advantages over traditional learning algorithm on CNN for improving the generalization performance and the interpretability performance of the discrimination part of the CNN.

In practical engineering, the proposed method may provide a new learning paradigm. The method can not only predict an accurate result for new input data but also provide a reasonable causal interpretation for the prediction of the discrimination part of a CNN. We suppose that it can solve the black box problem in the discrimination part. We believe that the proposed method provide a way how human understands the discrimination part.

REFERENCES

A.B. Arrietaa, N. Diaz-Rodriguezb, J.D. Sera, A. Bennetotb, S. Tabikg, A. Barbadoh, S. Garciag, S. Gil-Lopeza, D. Molinag, R. Benjaminsh, R. Chatilaf, and F. Herrerag. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

M. Gethsiyal Augasta and T. Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural Process. Lett*, 35(2):131–150, 2012.

O. Boz. Extracting decision trees from trained neural networks. pp. 456C461, the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.

J. Brandon. An ai god will emerge by 2042 and write its own bible. will you worship it?, 2017. https://venturebeat.com.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006.

Z.P. Che, S. Purushotham, R. Khemani, and Y. Liu. Interpretable deep models for icu outcome prediction. *Amia Annu Symp Proc*, 2016:371–380, 2017.

G. Cohen, S. Afshar, J. Tapson, and A.v. Schaik. EMNIST: an extension of MNIST to handwritten letters. *CoRR*, abs/1702.05373, 2017. URL http://arxiv.org/abs/1702.05373.

Mark W. Craven and Jude W. Shavlik. Using sampling and queries to extract rules from trained neural networks. *Machine Learning Proceedings*, pp. 37–45, 1994.

Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8:24–30, 1999.

N. Frosst and G. Hinton. Distilling a neural network into a soft decision tree, 2017. arXiv preprint arXiv:1711.09784 (2017).

R. Giles. Lukasiewicz logic and fuzzy set theory. *International Journal of Man-Machine Studies*, 8 (3):313–327, 1975.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *Computer Science*, 2015.

A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Mller. Causability and explainability of artificial intelligence in medicine. *Causability and explainability of artificial intelligence in medicine*, 9(4):e1312, 2019.

U. Johansson and L. Niklasson. Rule extraction from trained neural networks using genetic programming. *Int.conf.neural Inform.processing*, 2003.

U. Johansson and L. Niklasson. Evolving decision trees using oracle guides. In *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on*, 2009.

R. Krishnan, G. Sivakumar, and P. Bhattacharya. Extracting decision trees from trained neural networks. *Pattern Recognition*, 32(12):1999–2009, 1999.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

S. Parbhoo M. Zazzi V. Roth andF. Doshi-Velez M. Wu, M. C. Hughes. Beyond sparsity: Tree regularization of deep models for interpretability. AAAI, 2018.

Grgoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus Robert Mller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Algorithm 2 Construction Logic Net(Its time complexity is $O(scn)$ where s is number of iterations of MAX-SAT,c is number of sample categories and n is number of samples.)

Input: feature maps X, feature-capture maps F,y_{nn},batch size m.
Initialize L_a ,$y_{dclm} = y_{nn}$,$c = 0$.
repeat
  for $i = 1$ **to** sizeof(X) **do**
    for $j = first(i)$ **to** $final(i)$ **do**
      $IP[i][j]$ = Cosine similarity($X[i]$,$F[j]$).
      $lamda[i][j] =| X[i] || F[j] |$
    end for
  end for
  for $i = 1$ **to** sizeof(y_m) **do**
    $NewL_a[i] = maxsat(IP[][j], lamda[][j], y_{dclm}[i])$.
    $L_a[i] = NewL_a[i] \cap L_a[i]$.
  end for
  $c++$.
until $c == m$ is $true$

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

G. Riccardo, M. Anna, R. Salvatore, T. Franco, G. Fosca, and P. Dino. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2018.

B. Sebastian, B. Alexander, M. Grgoire, K. Frederick, M. Klaus-Robert, S. Wojciech, and S.O. Deniz. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 10(7):e0130140–, 2015.

A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences, 2016. arXiv preprint arXiv:1605.01713 (2016).

J.J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy. Treeview: Peeking into deep neural networks via feature-space partitioning, 2016. arXiv preprint arXiv:1611.07429(2016).

R. Traore, H. Caselles-Dupre, T. Lesort, T. Sun, G. Cai, N. D. Rodriguez, and D. Filliat. Discorl:continual reinforcement learning via policy distillation, 2019. arXiv preprint arXiv:1907.05855(2019).

A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, H. Jin, S. Petryk, S. A. Bargal, and J. E. Gonzalez. Nbdt: Neural-backed decision trees, 2020.

S.E. Zalando. Fashion mnist data set, 2017. https://github.com/zalandoresearch/fashion-mnist.

J.R. Zilke, E.L. Mencia, and F. Janssen. Deepred-rule extraction from deep neural networks. In *International Conference on Discovery Science, Springer*, pp. 457–473, 2016.

A Appendix: A Maximum A Posterior(MAP) algorithm for extracting a DCLM

A Maximum A Posterior(MAP) algorithm on the Maximum Satisfiability Problem (MAX-SAT) for extracting an DCLM is as follows.

B APPENDIX: SUFFICIENT AND NECESSARY CONDITION FOR CONSISTENT CONVERGENCE OF THE GENERALIZATION PERFORMANCE AND THE INTERPRETABILITY PERFORMANCE

If the minimization of loss function of CNN can guarantee the minimum of its interpretability performance, learning algorithm of the CNN can improve its interpretability performance. If not, the tradeoff problem between the generalization performance and the interpretability performance will exit. For proving the existence of the problem, we focus on a neuron of CNN. From the foot we may judge of Hercules. If an input channel $f(x)$ of the neuron is seen as a kernel function $K(x, w)$(w is weight vector including a bias of the neuron), it will span a kernel Hilbert space $\mathcal{H}_K = \{f(x) \in \mathcal{L}_\nu^2(X) \mid f(x) = K(x, w) = \sum_{k=1}^{\infty} a_k \phi_k(x)\}$ for the neuron. $\mathcal{H}_K$ can be regarded as a linear function set on $\mathcal{L}_\nu^2(X)$. It is a solution space of the neuron. The necessary and sufficient conditions for consistent convergence between the generalization performance and the interpretability performance are discussed below in $\mathcal{L}_\nu^2(X)$ based on the following lemmas.

Lemma 1. Continuous linear functional set on a separable Hilbert space X is nowhere dense in a square integrable function space $\mathcal{L}_\nu^2(X)$.

Lemma 2. Continuous nonlinear functional set of the separable Hilbert space X is everywhere dense in $\mathcal{L}_\nu^2(X)$.

When the optimal input channel $f^*(x)$ approximate a linear functional in $\mathcal{H}_K$ while the optimal interpretable model $P^*(x)$ don't approximate any linear functional or do not exist in $\mathcal{H}_K$, traditional training process cannot guarantee $f(x)$ approximates $P^*(x)$ according Lemma 1. From Lemma 2, the approximate will cannot converge until $P^*(x)$ approximate $f^*(x)$. Here, if we define the approximation as the similarity between the shapes of function curve of $f^*(x)$ and $P^*(x)$, the sufficient and necessary condition for the consistent convergence of the two performances is $\phi_d(P^*, f^*) = 0$. For the discrimination part of the CNN, the sufficient and necessary condition is still true. However because of data bias and noisy data in training data set, the condition is difficult to to be ensured in the majority of engineering applications. The tradeoff problem always exists between the two performances of the discrimination part.

According to the above conclusion, in order to completely solve the tradeoff problem, the $\phi_d(P^*, f^*)$(Here $f^*(x)$ is the optimal discrimination part) should be reduced. However P^* and $f^*(x)$ are unknown. Therefore, in training process, extracting a interpretable model $P(x)$ from a discrimination part $f(x)$ and then iteratively reducing $\phi_d(P, f)$ may be a convenient method.