Fuzzy Detectors Against Adversarial Attacks

Yi Li, Plamen Angelov, Neeraj Suri
School of Computing and Communications, Lancaster University
Lancaster, UK
{y.li154, p.angelov, neeraj.suri}@lancaster.ac.uk

***Abstract*—Deep learning-based methods have proved useful for adversarial attack detection. However, conventional detection algorithms exploit crisp set theory for classification boundary. Therefore, representing vague concepts is not available. Motivated by the recent success in fuzzy systems, we propose a fuzzy rule-based neural network to improve adversarial attack detection accuracy. The pre-trained ImageNet model is exploited to extract feature maps from clean and attacked images. Subsequently, the fuzzification network is used to obtain feature maps to produce fuzzy sets of difference degrees between clean and attacked images. The fuzzy rules control the intelligence that determines the detection boundaries. In the defuzzification layer, the fuzzy prediction from the intelligence is mapped back into the crisp model predictions for images. The loss between the prediction and label controls the rules to train the fuzzy detector. We show that the fuzzy rule-based network learns rich feature information than binary outputs and offers to obtain an overall performance gain. Our experiments, conducted over a wide range of images, show that the proposed method consistently performs better than conventional crisp set training in adversarial attack detection with various fuzzy system-based neural networks. The source code of the proposed method is available at https://github.com/Yukino-3/Fuzzy.**

***Index Terms*—Deep learning, adversarial attack detection, classification boundary, fuzzy rule, fuzzy prediction**

I. Introduction

Adversarial attack detection, aiming to defend applications by detecting attacks using the difference between adversarial and clean image samples, is an important security topic useful in many real-world applications such as autonomous driving systems, object detection, medical image processing, and robotics [1]. Recently, a variety of deep learning approaches have been proposed [2], [3], for adversarial attack detection mainly divided into empirical statistics-based detection [4], image pre-processing and reconstruction-based detection [5], and detection networks [6]. In this paper, we focus on image pre-processing and reconstruction-based detection by using deep learning techniques.

A neural network typically predicts a *crisp* result, namely, a value 1 when the sample is attacked and 0 when it is clean. The loss between the crisp prediction and crisp label is then exploited to train the model. While some recent studies have explored the fuzzy classifier in adversarial attack detection [7], [8], crisp set-based detection methods [5], [9], [10] are the common choices as the crisp sets can easily be estimated from the input data. The crisp set-based detection methods directly determine whether an image is clean or attacked. However, calculating the loss is non-differentiable and hinders training through normal back-propagation.

In recent studies, the fuzzy system is shown to offer several advantages in handling crisp set-based problems. It allows representation imprecision of objects, relations, and knowledge, and aims at different levels of latent representation [11], [12]. It constitutes a unified framework for representing and processing both numerical and symbolic information, as well as structural information (constituted mainly by spatial relations in image processing) [13]. Hence this theory can potentially handle tasks at several levels, from a low level (e.g., binary classification) to a high level (e.g., model-based structural recognition and scene interpretation). It provides a flexible framework for information fusion as well as powerful tool support for reasoning and decision-making [14]. In this paper, we show how the use of fuzzy detectors offers significant benefits in adversarial attack detection. Specifically, we propose a fuzzification process with fuzzy rules of difference degree between clean and attacked images.

The paper is organized as follows. Section II provides a literature review of adversarial attack detection techniques. Our proposed approach is introduced in Section III, with the experimental settings and results described in Section IV. Section V presents the conclusions and future work.

II. Adversarial Attack Detection

Various techniques for adversarial attack detection have been developed in the deep learning community. Over the past several years, increasing research efforts have been devoted to improving the detection efficiency of neural networks. For example, in [1], a patch segmenter is designed to generate patch masks that provide pixel-level localization of adversarial patches. Then, these adversarial patches are removed to guarantee data security. Moreover, Soares et al. propose a similarity-based deep neural network (sim-DNN) to calculate the degree of similarity between training samples and their prototypes adversarial attacks [15]. By minimizing the similarity score, the concept changes are detected from the attacked data when comparing their similarities against the set of prototypes. In [10], binary classification datasets are constructed separately to train the binary classifier and then divides by the binary classification detector. The adversarial samples are constructed in two parts based on relevance features and model activation features for attack detection. Different from these techniques, Qi et al. utilize two deep learning models in the training stage. Some adversarial attack samples are generated toward the local DL model [9]. Subsequently, the target model is attacked to produce perturbed samples. In

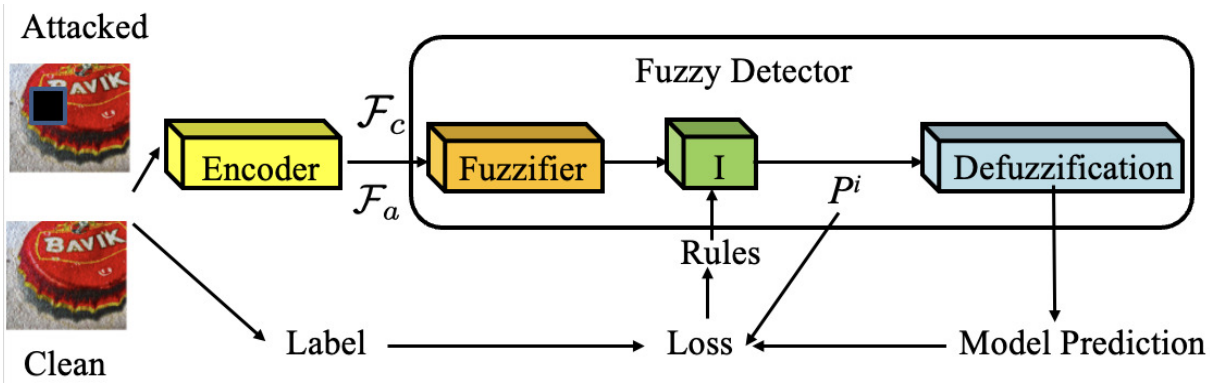

Fig. 1. The architecture of the proposed fuzzy detector.

the adversarial training, the misclassification probability of all training samples is estimated by the local model to detect and delete perturbed samples from the dataset.

III. Proposed Approach for Fuzzy Detectors

Fig. 1 illustrates the architectural approach of our proposed fuzzy detector. Conceptually, the feature maps from the attacked and clean images $\mathcal{F}_a$ and $\mathcal{F}_c$ constitute the inputs of the proposed fuzzy detector. The loss between the feature maps is converted into the fuzzy set for the intelligence (I).

The input of the encoder is either clean or attacked images with hard labels, i.e., clean or attacked. The attacked images are constructed with random error rates from 0.01 to 0.04 as in contemporary works. The ImageNet pre-trained model obtains these images and extracts their feature maps. The feature maps of clean and attacked images are presented as $\mathcal{F}_c$ and $\mathcal{F}_a$, respectively. The proposed fuzzy system-based detector then aims to map from the feature space to the label space. To achieve that, we propose a fuzzy detector with its constituent blocks progressively detailed in the following sections.

A. Fuzzifier

As aforementioned, the loss between $\mathcal{F}_c$ and $\mathcal{F}_a$ is required to be calculated. Subsequently, the fuzzifier converts the loss into a fuzzy set to describe the difference degrees between feature maps at the pixel level. The degree of differences in the fuzzy set quantifies the difference levels across the feature maps of clean and attacked images. The membership function is illustrated in Fig. 2.

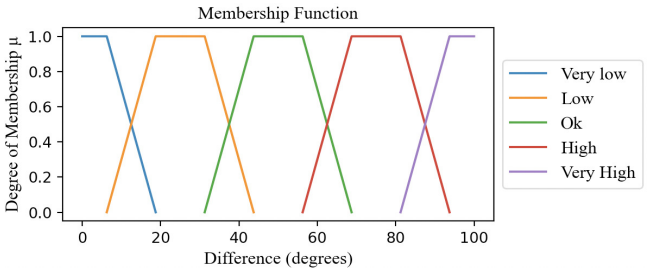

Fig. 2. The membership function.

B. Fuzzy Rules

Intelligence is controlled by rules that determine the detection boundaries. The rules of the proposed fuzzifier follow commonly used fuzzy-rule-based classifiers [11], [12] as:

$$R^i: \text{IF}\,(x_1 \textit{ is around } x_1^i) \quad \text{AND} \quad (x_2 \textit{ is around } x_2^i) \quad \text{AND} \cdots \quad \cdots \quad \text{AND}\,(x_n \textit{ is around } x_n^i) \quad \text{THEN}\,(P^i) \tag{1}$$

where $x = [x_1; x_2; \vdots; x_n]^T$ is the pixels of feature maps. In the intelligence layer, $(x_j$ is around $x_j^i)$ indicates the jth fuzzy set of the ith fuzzy rule R^i. To achieve that, we consider the Eucledian Distance d between x_j and x_j^i with a hyperparameter $_j$. When the distance d is smaller than $_j$, the fuzzy prediction with the ith fuzzy rule is P^i that predicts how much the model trusts the image. The hyperparameter $_j$ is further updated to improve the boundary accuracy in the training stage.

C. Defuzzification

A centroid defuzzification method is exploited to convert the fuzzy prediction set into the model prediction [16]. Particularly, the center of gravity of the fuzzy set is calculated along the difference degree as:

$$P = \frac{\sum_i (P_i)\, P_i}{\sum_i (P_i)} \tag{2}$$

where P is the model prediction, i.e., 0 or 1 for clean or attacked image, respectively.

D. Training

The training loss is calculated as follows. Firstly, we calculate the fuzzy loss $\mathcal{L}_F$ between the label and fuzzy prediction. Secondly, the overall loss $\mathcal{L}$ is calculated by the loss between the label and model prediction $\mathcal{L}_M$ with hyperparameters $_1$ and $_2$ as:

$$\mathcal{L} = \begin{cases} _1 \cdot \mathcal{L}_F; & \text{if } \mathcal{L}_M \neq 0 \\ \mathcal{L}_F = _2; & \text{otherwise} \end{cases} \tag{3}$$

Both $_1$ and $_2$ are empirically set between 1 and 10 over the different experiments. The fuzzy rules are refined by $\mathcal{L}$

TABLE I

ATTACK DETECTION RATIO ON THE CIFAR-10 AND IMAGENET-R DATASETS. EACH RESULT IS THE AVERAGE OF 10,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Detection Ratio (%)															
	CIFAR-10								ImageNet-R (%)							
Method	Clean		FGSM		PGD		SSAH		Clean		FGSM		PGD		SSAH	
FCB [7]	75.5	1.8	49.8	1.5	47.1	1.0	43.6	1.8	75.2	1.9	48.5	1.8	46.8	1.9	43.9	1.4
SAC [1]	78.7	1.5	60.1	2.3	59.7	2.2	56.8	2.4	78.0	1.9	58.9	2.0	57.5	2.1	52.9	1.8
sim-DNN [15]	81.8	1.0	70.5	1.4	60.0	1.6	49.4	0.7	80.9	1.5	71.0	1.7	66.2	1.3	61.4	1.1
DTBA [9]	85.7	1.3	78.3	1.2	75.6	1.0	71.7	0.8	85.2	1.0	78.0	1.1	72.4	1.1	68.9	0.8
ESMAF [10]	87.4	1.0	80.5	1.6	76.9	1.9	75.4	0.7	87.4	1.2	79.7	1.3	75.6	0.9	71.8	0.6
F-Res	89.1	0.5	86.3	1.0	83.5	0.7	82.1	0.9	89.0	0.5	86.0	1.1	82.8	1.2	80.2	0.8
F-PF	89.9	0.8	86.5	0.7	82.3	1.2	80.8	0.8	89.9	0.8	**87.0**	**1.3**	81.6	1.0	80.8	0.9
F-YL	**90.2**	**0.4**	**87.2**	**0.5**	**84.8**	**0.9**	**84.1**	**0.4**	**89.9**	**0.7**	**87.0**	**1.2**	**83.4**	**0.7**	**82.9**	**0.5**

TABLE II

ATTACK DETECTION RATIO ON THE COCO AND ILSVRC DATASETS. EACH RESULT IS THE AVERAGE OF 10,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Detection Ratio (%)															
	COCO								ILSVRC (%)							
Method	Clean		FGSM		PGD		SSAH		Clean		FGSM		PGD		SSAH	
FCB [7]	72.7	1.6	47.2	1.6	46.5	1.9	40.6	2.0	76.8	1.5	51.2	2.2	50.4	1.5	48.5	2.5
SAC [1]	78.0	2.0	49.6	2.1	48.4	2.0	45.9	1.9	80.2	2.2	63.6	1.6	63.0	2.0	59.4	1.9
sim-DNN [15]	79.3	2.0	66.8	2.0	64.3	2.2	60.0	1.8	82.5	1.3	75.2	2.1	72.5	1.6	69.6	1.5
DTBA [9]	83.1	0.9	75.3	1.1	71.0	1.4	68.7	1.0	86.8	0.7	81.1	1.4	80.5	0.8	76.0	0.9
ESMAF [10]	85.0	1.1	76.5	1.3	74.2	1.6	70.8	1.0	87.9	1.0	82.6	2.0	77.0	2.1	75.2	0.9
F-Res	88.4	0.6	84.0	1.4	80.7	1.5	78.6	1.2	89.7	0.8	86.9	1.2	84.1	0.6	82.8	1.1
F-PF	**89.8**	**0.3**	85.8	1.2	82.0	1.0	80.1	1.1	90.3	0.4	87.3	1.0	82.4	1.1	81.2	0.6
F-YL	**89.8**	**0.3**	**86.8**	**0.9**	**84.0**	**1.1**	**84.1**	**0.8**	**91.0**	**0.3**	**87.9**	**0.7**	**85.9**	**0.9**	**84.4**	**0.8**

and control the intelligence that makes more accurate fuzzy predictions. The pseudo-code of the proposed fuzzy rule-based attack detection method is summarized in Algorithm 1.

Algorithm 1: Fuzzy rule-based detector.

Input: Feature maps $\mathcal{F}_c$ and $\mathcal{F}_a$, Label X, learning rate , epoch $E_{\max}$,
Output: Model prediction P
Data: Training set $\mathcal{D}$

Initialize R^i, ;
Initialize hyperparameters $_1$, $_2$, $_j$ and ;
for $E = 1; 2; \ldots; E_{\max}$ **do**
  $\mathcal{L}_F = H(\mathcal{F}_c; \mathcal{F}_a)$ // Calculate the cross entropy loss;
  if $d(x_j; x_j^i) <$ $_j$ **then**
    $P_i =$ $(\mathcal{F}_a)$;
  end
  $P \leftarrow P_i$; // Defuzzification;
  if $X = P$ **then**
    $\mathcal{L} = \mathcal{L}_F =$ $_2$;
  else
    $\mathcal{L} =$ $_1 \cdot \mathcal{L}_F$
  end
  ; $\leftarrow \mathcal{L}_F$;
  $_1$; $_2 \leftarrow \mathcal{L}$ // Update hyperparameters ;
end

IV. EXPERIMENTAL RESULTS

A. Datasets

We perform experiments on several public datasets, including ImageNet-R [17], Canadian Institute For Advanced Research-10 (CIFAR-10) [18], Common Object and Concept (COCO) [19], and ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [20]. We randomly select 50,000, 10,000, and 10,000 images from each dataset for the training, validation, and test stages.

B. Attacks and Performance Measure

The adversarial samples from these datasets are constructed using the Fast Gradient Sign Method (FGSM) [21], Projected Gradient Descent (PGD) [22], and Semantic similarity attack on high-frequency components (SSAH) [23] attacks. We select these attacks because they are robust to novel adversarial attack detection and defense techniques [15], [23]. The error rate is randomly set from 0.01 to 0.04 for training and test data similar to as in contemporary works [5], [15].

In the experiment, the detection rate (DR) [15] is used as the performance measure.

$$DR(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{4}$$

where TP and TN are true positive and true negative results, and FP and FN are false positive and false negative results. Moreover, we evaluate the true positive on clean image samples as:

$$DR(\%) = \frac{TP}{TP + TN} \times 100 \tag{5}$$

TABLE III

ATTACK DETECTION RATIO ON THE CIFAR-10 AND IMAGENET-R DATASETS WITH UNSEEN ATTACKS. EACH RESULT IS THE AVERAGE OF 10,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Detection Ratio (%)															
	CIFAR-10								ImageNet-R (%)							
Method	Clean		FGSM		PGD		SSAH		Clean		FGSM		PGD		SSAH	
FCB [7]	73.6	1.9	46.3	1.4	45.2	1.3	40.7	2.0	72.0	2.1	44.9	1.9	44.5	1.9	41.4	1.7
SAC [1]	74.8	1.6	57.7	2.2	56.0	2.5	53.5	2.6	74.9	2.1	56.0	2.3	53.9	2.4	50.2	2.0
sim-DNN [15]	79.5	1.2	67.9	1.6	56.9	1.8	47.5	1.0	77.6	1.8	68.8	1.6	65.1	1.5	60.1	1.2
DTBA [9]	84.2	1.6	77.0	1.4	74.0	1.1	70.2	1.1	83.9	1.4	76.4	1.3	70.9	1.2	67.3	1.2
ESMAF [10]	84.8	1.3	77.9	1.5	74.3	2.2	73.8	0.9	84.7	1.5	77.1	1.6	72.9	1.2	70.0	1.0
F-Res	88.8	0.5	85.9	0.9	82.8	0.9	81.6	1.1	88.5	0.7	85.7	1.5	82.7	1.1	79.4	0.9
F-PF	**89.8**	**0.6**	86.3	0.6	82.0	1.2	80.4	0.9	**89.5**	**0.9**	**86.8**	**1.1**	81.4	1.1	80.7	1.0
F-YL	89.4	0.6	**86.4**	**0.7**	**83.6**	**1.0**	**82.8**	**0.7**	89.0	0.4	85.9	1.5	**82.0**	**0.8**	**82.6**	**0.5**

TABLE IV

ATTACK DETECTION RATIO ON THE COCO AND ILSVRC DATASETS WITH UNSEEN ATTACKS. EACH RESULT IS THE AVERAGE OF 10,000 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULTS. *Italic* SHOWS THE PROPOSED METHODS.

	Detection Ratio (%)															
	COCO								ILSVRC (%)							
Method	Clean		FGSM		PGD		SSAH		Clean		FGSM		PGD		SSAH	
FCB [7]	69.9	1.7	45.8	1.7	44.8	2.2	38.9	2.4	74.0	1.7	48.4	2.0	47.8	1.8	45.9	2.8
SAC [1]	72.8	2.1	48.2	2.1	45.9	2.2	44.5	2.1	78.7	2.3	60.8	1.4	61.7	2.1	57.6	2.2
sim-DNN [15]	78.8	2.4	63.6	1.9	62.5	2.0	57.8	2.1	82.0	1.6	74.3	2.4	70.3	1.9	66.8	1.8
DTBA [9]	82.4	1.1	74.6	1.6	70.1	1.8	67.6	1.5	86.3	0.5	79.7	1.8	79.0	1.2	74.7	1.5
ESMAF [10]	82.2	1.1	72.9	1.9	72.0	2.0	67.1	1.8	84.3	1.9	80.2	2.2	73.8	2.5	72.9	1.6
F-Res	88.3	0.5	83.3	1.4	80.1	1.9	77.9	1.4	89.5	1.0	86.0	1.3	83.3	0.8	81.7	1.5
F-PF	**89.7**	**0.4**	85.2	1.1	81.6	1.4	79.8	1.2	89.9	0.6	**87.0**	**0.9**	82.1	1.4	80.8	0.8
F-YL	89.0	0.6	**85.6**	**1.2**	**82.9**	**1.2**	**82.8**	**1.0**	**90.0**	**0.5**	86.8	1.0	**85.0**	**0.8**	**83.1**	**1.3**

C. Model Configuration

The pre-trained EfficientNetV2-XL [24] on the ILSVRC dataset is exploited to extract features. We select this model because it achieves the state-of-the-art benchmark on the ILSVRC challenge. Moreover, we apply the proposed fuzzy detector on different backbones, e.g., Res2Net-v1b-101 [25], YOLOX-L [26], and PRB-FPN6-2PY [27]. Different from the pre-trained encoder, these models are initialized and re-trained with fuzzy logic.

At training stage, the proposed model is trained using the M-SGD optimizer with a learning rate set empirically to 0.0008 according to a grid search. The batch size is set to 32. We train the networks for 200 epochs. All experiments are run on the High End Computing (HEC) Cluster with Tesla V100 GPUs.

D. Results

1) Comparison with Same Attacks: In the first experiment, we compare the proposed method to state-of-the-art adversarial attack detection methods [1], [7], [9], [10], [15] with the same attack between the training and test stage. The proposed fuzzy detector-based Res2Net-v1b-101, YOLOX-L, and PRB-FPN6-2PY are simplified as F-Res, F-YL, and F-PF, respectively.

From Tables I & II, it can be observed that: (1) In all the evaluated models, the proposed methods with different backbones offer the best effectiveness. Different from crisp set-based decision-making pipelines, the proposed fuzzy detectors convert the loss between feature maps into fuzzy sets and provide difference scores ('high', 'ok', and 'low'). Therefore, the proposed method exploits more feature information than binary decisions. The fuzzy rules are trained with difference scores to help the detector make more accurate decisions. (2) The proposed F-YL model offers the best attack detection performance on all datasets. The reason is likely due to the combined implicit knowledge and explicit knowledge in the YOLOX decoder [26]. (3) Compared to the improvement in FGSM, PGD, and SSAH attacks, the improvement of detection accuracy tends to fall drastically when evaluating the true positives on clean image samples. For example, compared to ESMAF model, the proposed F-YL obtain 7.9% improvement on PGD attacked CIFAR-10 dataset, while it is only 2.8% on the true positive evaluation.

2) Comparison with Different Attacks: In this experiment, we compare the proposed method to state-of-the-art adversarial attack detection methods [1], [7], [9], [10], [15] with different attacks between the training and test stage. To achieve that, we randomly select an unseen attack to construct the test data in Tables III & IV.

From Tables III & IV, it can be observed that in all the evaluated models, the proposed fuzzy rule-based methods with different backbones offer the best effectiveness. The goal of the adversarial training provided by the DTBA and DSMAF is to increase the model's robustness, however, they lack generalisation to unseen domains, i.e., datasets and attacks. The proposed fuzzy detector maintains its performance stable even when adversarial or clean images from unknown datasets are presented to the detection model due to its inner fuzzy rules and detection mechanism that was projected for such scenarios.

The visualizations are shown in Fig. 3 as related to the reconstructions after detecting attacks of three randomly se-