MTDeep: Boosting the Security of Deep Neural Nets Against Adversarial Attacks with Moving Target Defense

Sailik Sengupta[1(✉)], Tathagata Chakraborti[2], and Subbarao Kambhampati[1]

[1] Arizona State University, Tempe, AZ, USA
{sailiks,rao}@asu.edu
[2] IBM Research, Cambridge, MA, USA
tchakra2@ibm.com

Abstract. Present attack methods can make state-of-the-art classification systems based on deep neural networks mis-classify every adversarially modified test example. The design of general defense strategies against a wide range of such attacks still remains a challenging problem. In this paper, we draw inspiration from the fields of cybersecurity and multi-agent systems and propose to leverage the concept of *Moving Target Defense (MTD)* in designing a meta-defense for 'boosting' the robustness of an ensemble of deep neural networks (DNNs) for visual classification tasks against such adversarial attacks. To classify an input image at test time, a constituent network is randomly selected based on a mixed policy. To obtain this policy, we formulate the interaction between a Defender (who hosts the classification networks) and their (Legitimate and Malicious) users as a *Bayesian Stackelberg Game (BSG)*. We empirically show that our approach *MTDeep*, reduces misclassification on perturbed images for various datasets such as MNIST, FashionMNIST, and ImageNet while maintaining high classification accuracy on legitimate test images. We then demonstrate that our framework, being the first meta-defense technique, can be used *in conjunction* with any existing defense mechanism to provide more resilience against adversarial attacks that can be afforded by these defense mechanisms alone. Lastly, to quantify the increase in robustness of an ensemble-based classification system when we use MTDeep, we analyze the properties of a set of DNNs and introduce the concept of differential immunity that formalizes the notion of attack transferability.

1 Introduction

State-of-the-art systems for image classification based on Deep Neural Networks (DNNs) are used in many important tasks such as recognizing handwritten digits on cheques [10], object classification for automated surveillance [9] and autonomous vehicles [6]. Adversarial attacks to make these classification systems misclassify inputs can lead to dire consequences. For example, in [15], road signs saying 'stop' are misclassified, which can make an autonomous vehicle behave

T. Alpcan et al. (Eds.): GameSec 2019, LNCS 11836, pp. 479–491, 2019.
https://doi.org/10.1007/978-3-030-32430-8_28

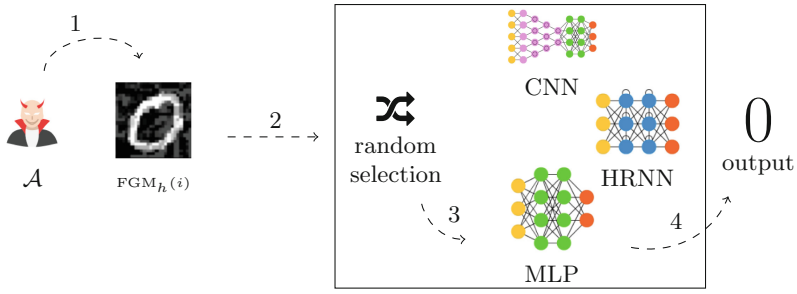


Fig. 1. At attack perturbation crafted for HRNN (FGM_h) is rendered ineffective when MTDeep picks the MLP (at random) for classification at test time.

dangerously. If $\hat{D}(i)$ denotes the class of an image i output by a Deep Neural Network $\hat{D}$, an adversarial perturbation ϵ when added to the image i tries to ensure that $\hat{D}(i) \neq \hat{D}(i+\epsilon)$. In addition, attackers try to minimize some norm of ϵ, which ensures that the changed image $i+\epsilon$ and the original image i are indistinguishable to humans. The effectiveness of an attack method is measured by the accuracy of a classifier on the perturbed images generated by it.

Defenses against adversarial examples are designed to be effective against a certain class of attacks by either training the classifier with perturbed images generated by these attacks or making it hard for these attacks to modify some property of the neural network. Some recent works construct defenses that enforcing classification of images that are ϵ distance away from an image in the training set to the same class. Unfortunately, this has the side effect of bringing down the classification accuracy [21].

In this paper, we take a different view and design a meta-defense that can function both as (1) a first line of defense against new attacks and (2) a second line of defense when used in conjunction with any existing defense mechanism to boost the security gains the latter can provide. We consider a game theoretic perspective and investigate the use of Moving Target Defense (MTD) [25], in which we randomly select a network from an ensemble of networks when classifying an input image (i.e. *strategic randomization at test time*), for boosting the robustness against adversarial attacks (see Fig. 1). Our contributions are–

- MTDeep – an MTD-based framework for an ensemble of DNNs.
- A Bayesian Stackelberg Game formulation with two players – MTDeep and the users. The Stackelberg Equilibrium of this game gives us the optimal randomization strategy for the ensemble that maximizes the classification accuracy on regular as well as adversarially modified inputs.
- We show empirically that MTDeep can be used as (1) a standalone defense mechanism to increase the accuracy on adversarial samples by $\approx$ 24% for MNIST, $\approx$ 22% for Fashion MNIST and $\approx$ 21% for ImageNET data-sets against a variety of well-known attacks and (2) in conjunction with existing defense mechanisms like Ensemble Adversarial Training, MTDeep increases

the robustness of a classification system (by $\approx 50\%$ for MNIST). We also show that black-box attacks (see related work) on a distilled network are ineffective (in comparison to white-box attacks) against the MTDeep system.
– We define the concept of *differential immunity*, which is (1) the first attempt at defining a robustness measure for an ensemble against attacks and (2) a quantitative metric to capture the notion of attack transferability.

Although prior research has shown that effectiveness of attacks can sometimes transfer across networks [19], we show that there is still enough residual disagreement among networks that can be leveraged to design an add-on defense-in-depth mechanism by using MTD. In fact, recent work has demonstrated that it is possible to train models with limited adversarial attack transferability [2], making our meta-level defense approach particularly attractive.

2 Related Work

In this section, we first discuss existing work on crafting adversarial inputs against DNNs (at test-time) and defenses developed against them. Then, we briefly discuss some work in Moving Target Defense that inspires this defense.[1]

Recent literature has shown multiple ways of crafting adversarial samples for a DNN [11,13,15,19] using the gradient information or by examining the geometric space around an input. These attacks require complete knowledge about the classification network. On the other hand, attacks that craft gradient-based perturbations on distilled networks [14,19] or use zeroth-order optimization [5] can cripple DNNs even when the attacker has no knowledge about the actual classification network (and are thus called black-box attacks).

Defense techniques against the two types of attacks described above commonly involve generating adversarial perturbed training images using one (or all) of the attack methods described and then using the generated images along with the correct labels to fine tune the parameters of the DNN. Ensemble adversarial training [20] and stability training [24] are improvements on this defense technique. Unlike us, the former does not use the ensemble at classification time. We do not discuss other defenses further because *our proposed framework can be used in conjunction with any of these to improve their security guarantees*. Our approach is well supported by findings in previous research works that show introduction of randomized switching makes it harder for any attacker to reverse engineer a classification system [22], which is necessary for constructing effective white-box attacks. Note that ensemble based defenses [1,8] can be viewed as simply adding an extra pooling layer whose weights are equal to the importance given to the votes of the constituent networks. Thus, all attacks on a DNNs are trivially effective against such voting-based ensembles. To this extent, researchers have also shown that an ensemble of vulnerable DNNs cannot result in a classifier robust to attacks [7]. In contrast, MTDeep builds in an implicit mechanism based on *randomization at prediction time*, making it difficult for an

[1] A detailed overview can be found at https://arxiv.org/abs/1705.07213.

Table 1. The actions of the players and the utilities of the two user types–$\mathcal{L}$ and $\mathcal{A}$ for the MNIST dataset.

Legitimate User ($\mathcal{L}$)	
MTDeep	Classification Image
MLP	99.1
CNN	98.3
HRNN	98.7

Adversarial User ($\mathcal{A}$)								
FGM_m	FGM_c	FGM_h	DF_m	DF_c	DF_h	PGD_m	PGD_c	PGD_h
3.1	20.39	38.93	1.54	89.8	93.83	0.00	49.00	61.00
55.06	10.28	71.39	98.87	0.87	98.55	78.00	0.00	90.0
25.12	27.24	11.43	95.38	83.17	3.66	23.00	51.00	0.00

adversary to fool the classification system. There has been some previous work that leverage randomization at test time [4] but cannot be used out-of-the-box for DNNs. The authors try to prevent misclassification rate under attack and end-up affecting the classification accuracy on non-adversarial test inputs.

Universal perturbations [12], based on the DeepFool attack [13], needs to generate only one "universal" perturbation per network. Authors show that adversarial training is ineffective against this attacks. On the contrary, we show that MTDeep can prove to be an effective defense against these attacks because such attacks are network specific and thus, often have low transferability.

Moving Target Defense (MTD) is a paradigm used in software security that tries to reduce the success rate of an attack by pro-actively switching between multiple software configurations [25]. Devising effective switching strategies for MTD systems requires reasoning about attacks in a multi-agent game theoretic fashion in order to provide formal guarantees about the security of such systems [18]. Thus, we model the interaction between an image classification system (an ensemble) and its users, both legitimate and adversarial, as a Bayesian Stackelberg Game, providing provable guarantees on the expected performance on both legitimate and adversarial inputs.

3 MTDeep: MTD for Deep Neural Networks

In our system, the defender has multiple system classifiers for a given task. The attacker has a set of attacks that it can use to cripple the constituent classifiers. Given an input to the system, the defender selects, *at random*, one of the configurations to run the input and returns the output generated by that system. Since the attacker does not know which system is specifically selected, its attacks are less effective than before (Fig. 1). As stated earlier, randomization in selecting a configuration for classification of each input is paramount. Unfortunately, an MTD framework for classification systems, that leverages randomization, might end up reducing the accuracy of the overall system in classifying non-perturbed images. Thus, in order to retain good classification accuracy and guarantee high security, we model the interaction between MTDeep and its users as a Bayesian Stackelberg Game and show that the equilibrium results in the optimal selection strategy. We now discuss our game-theoretic formulation.

Players and Action Sets. The configuration space for the defender, i.e. MTDeep, comprise of various DNNs that are trained on the particular image classification task. The second player in this game is the user of the classification system. The second player has two player types – Legitimate User ($\mathcal{L}$) and

the Adversary ($\mathcal{A}$). $\mathcal{L}$ has one action – to input non-perturbed images to the MTDeep system. The adversary $\mathcal{A}$ has various attack actions and uses one of these to perturbs an input image. In our threat model, we consider a strong adversary who knows the different constituent architectures in our MTDeep system. This means they can easily generate powerful white-box attacks.

Utilities. Existing works that design defense methods against adversarial attacks for DNNs model the problem as a zero-sum game where the attacker tries to maximize the defender's loss function by coming up with perturbed images that the network misclassifies, whereas the defender tries to reduce the loss on these adversarially perturbed examples [11]. Fine tuning the classifier to have high accuracy on adversarially perturbed inputs often has the side effect of reducing the classification accuracy on non-perturbed inputs from the test set [21]. In this paper, we move away from the zero-sum game assumption and try to ensure that the defender minimizes the loss functions for both types of inputs images– images from the initial test-set and the adversarially perturbed ones. Thus, we want MTDeep to be effective for $\mathcal{L}$ (proportional to minimizing the loss on the original test set) and, at the same time, increase the accuracy of classification for the perturbed images (proportional to minimizing the loss against adversarial inputs at test-time), making this a multi-objective optimization problem. The utilities for each player in this game are as follows.

- The Legitimate User ($\mathcal{L}$) and the defender both get a reward value equal to the % accuracy of the DNN system.
- The Adversary ($\mathcal{A}$) and the defender play a constant($= 100$) sum game, where the former's reward value for an attack against a network is given by the *fooling rate* and the defender's reward is the accuracy on perturbed inputs.

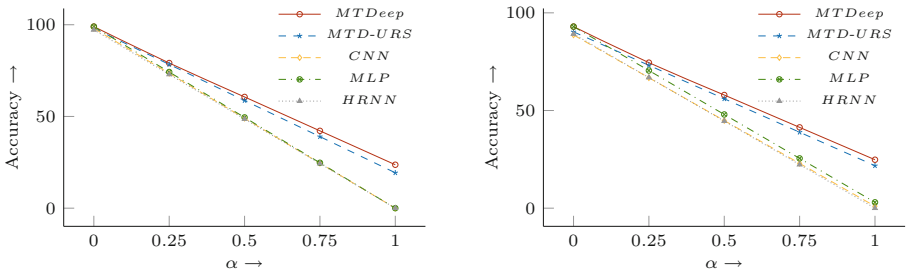
We also consider a parameter α that defines the probability of the player types $\mathcal{A}$ and $\mathcal{L}$. It lets the defender weigh the importance of catering to legitimate test samples *vs.* correctly classifying adversarial samples. The game-matrix for the MNIST classification task is shown in Table 1.[2]

MTDeep's Switching Strategy. Note that the defender D has to play first, i.e. deploy a classification system that either a legitimate user $\mathcal{L}$ can use or an adversary $\mathcal{A}$ can attack. This imparts a leader-follower paradigm to the formulated Bayesian Game. The defender leads by playing first and then the attacker follows by choosing an attack action having inferred the leader's (mixed) strategy. Satisfying the multi-objective criterion, mentioned above, is now equivalent to finding the Stackelberg Equilibrium of this game. We find this equilibrium by using the mixed integer quadratic program (MIQP) formulation in [16].

4 Experimental Results

We first compare the effectiveness of MTDeep as a standalone defense mechanism for classifying MNIST, Fashion-MNIST and ImageNet datasets. We then

[2] More details and examples of games (for the Fashion-MNIST and Imagenet classification tasks) can be found at https://arxiv.org/abs/1705.07213.



(a) MNIST (b) Fashion-MNIST

Fig. 2. Accuracy of MTDeep with non-adversarially trained networks compared to accuracy of individual constituent networks and uniform random strategy.

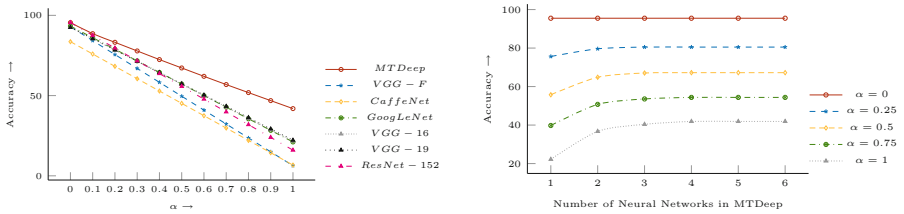
show that MTDeep piggybacked onto an existing defense mechanism can help boost the classification accuracy against adversarial attacks by almost 50%. We then analyze the effect of black-box attacks created on a distilled network and introduce the notion of differential immunity for ensembles. We discuss how that this metric can capture the informal notion of transferability of attacks and be used to measure the effectiveness of MTDeep.

4.1 MTDeep as a Standalone Defense Technique

We compare the effectiveness of MTDeep with two baselines– the individual networks in the ensemble and the a randomized ensemble that uses Uniform Random Strategy (MTD-URS) to pick one of the constituent networks with equal probability. In contrast, MTDeep uses the Stackelberg equilibrium strategy of the defender to pick a constituent DNN.

MNIST and Fashion-MNIST. For each data-set, we trained three classification networks that were built using either Convolution layers (CNN), Multi-layer Perceptrons (MLP) or Hierarchical Recurrent layers (HRNN). The size of the train and test sets were 50000 and 10000 respectively.

We considered three attack methods for the attacker – the Fast Gradient Based (FGM) attack (with $\epsilon = 0.3$), the DeepFool (DF) attack (with three classes being considered at each step when searching for an attack perturbation), and the Projected Gradient Descent (PGD) attack (with $\epsilon : 0.3$, $\epsilon - iter : 0.05$). An adversarial example generated using the PGD algorithm on the loss information of the CNN is termed as PGD_c in Table 1 (similarly $PGD_{h/m}$). We then find the classification accuracy of each network on these adversarial examples to compute the utility values. Note that an adversarial example developed using information about one network may not be as effective for the other networks. We find that this is especially true for attacks like DF that exploit information about a particular network's classification boundary. On the other hand, attacks that exploit the gradient signals of a particular network are more effective against



(a) Accuracy of classifiers under attack. (b) Participation of constituent networks.

Fig. 3. Results on the ImageNET classification task.

the other networks, i.e. have high transferability. We observe this trend for both the MNIST and the Fashion-MNIST data-set.

In Fig. 2, we plot the accuracy of a particular classification system as α varies from 0 to 1. When $\alpha = 0$ and the defender ignores the possibility of playing against an adversary, the optimal strategy for MTDeep boils down to a pure strategy for selecting the most accurate classifier. In contrast, MTD-URS has lower classification accuracy than MTDeep because it chooses the two less-accurate classifiers with probability 0.33. Given that classification accuracy of the constituent networks are relatively high, the drop in accuracy is small.

When $\alpha = 1$ and the defender only receives adversarial examples as inputs, strong attacks like PGD can fool individual networks 100% of the time for MNIST and 97% for Fashion-MNIST. In contrast, randomized selection of networks at test-time perform much better because an adversarial perturbation developed based on information from one network fails to fool other networks that may be selected at classification time. MTDeep achieves a classification accuracy of 24% for MNIST and 25% for Fashion-MNIST while MTD-URS has a classification accuracy of $\approx 20\%$ for both the data sets. The difference in classification accuracy stems form the fact that MTD-URS picks more vulnerable networks with equal probability.

ImageNET. We use six different networks which have excelled on ILSVRC-2012s validation set [17] to construct the ensemble for MTDeep[3]. Generating attacks like FGM, DF and PGD for ImageNET are time and resource intensive. Thus, we assume the adversary uses Universal Perturbations (UP) developed for each network [12], which are built on top of DF and only one attack mask is generated for each constituent network (as opposed to each test image).

Defense mechanisms like adversarial training are ineffective against this type of attack [12]. Furthermore, no other defenses have been shown to be effective against this attack. In such cases, MTDeep is a particularly attractive approach because it increases the robustness of the classification system even when all other defense mechanisms are ineffective.

In Fig. 3(a), we plot the expected accuracy for the MTDeep along with the expected accuracy of each of the constituent networks when the probability of

[3] More details can be found at https://arxiv.org/abs/1705.07213.

Legitimate User ($\mathcal{L}$)	
MTDeep	Classification Image
MLP_{eat}	97.99
CNN_{eat}	98.97
HRNN_{eat}	97.22

Adversarial User ($\mathcal{A}$)

FGM_m	FGM_c	FGM_h	DF_m	DF_c	DF_h	PGD_m	PGD_c	PGD_h
95.06	75.32	70.1	1.5	96.97	95.73	0.00	88.00	69.00
61.44	96.55	68.58	98.36	0.79	96.09	72.00	20.00	81.00
81.24	84.79	93.1	96.85	95.9	4.41	82.00	71.00	10.00

Fig. 4. The utilities for the players when the adversary uses the aforementioned attacks against the classifiers fine-tuned using Ensemble Adversarial Training (EAT) with FGM attacks.

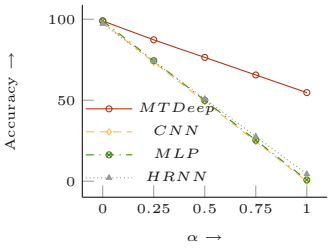


Fig. 5. Accuracy gains of MTDeep when constituent classifiers are adversarially trained.

the adversary type α varies. Given there are six constituent networks in the ensemble, to avoid clutter, we don't plot MTD-URS for brevity but observe that it always has $\approx 4\%$ less accuracy than MTDeep, which is a large accuracy difference in accuracy in the context of ImageNET. At $\alpha = 0$, MTDeep uses the most accurate network (ResNet-152) to maximize classification accuracy. As adversarial inputs become more ubiquitous and α becomes 1, the accuracy on perturbed inputs drops for all the constituent networks of the ensemble. Thus, to stay protected, MTDeep switches to a mixed policy that utilizes more networks. At $\alpha = 1$, the accuracy of MTDeep is 42% compared to 20% for the best of the single DNN architectures. The optimal strategy in this case is $\boldsymbol{x} = \langle 0.0, 0.171, 0.241, 0.0, 0.401, 0.187 \rangle$ which discards two of the six constituent networks. Note that the 22% accuracy bump for modified images comes despite (i) high misclassification rates of the constituent networks against Universal Perturbations, and (ii) lack of proven defense mechanisms against such attacks.

4.2 MTDeep as an Add-On Defense-in-depth Solution

We study the use of MTDeep on top of a state-of-the-art defense mechanism called Ensemble Adversarial Training (EAT) [20]. EAT is an improvement of adversarial training that uses adversarial examples generated on non-target networks to fine tune weights of the target network. Given MTDeep works with an ensemble, it renders itself naturally to this robustification method. Unfortunately, using EAT can only make the networks robust against attack images generated by the particular attack algorithm. We observe that the individual networks are still vulnerable to stronger (i.e. more computationally intensive) attacks. In Fig. 4, we show that the utility values obtained using the three constituent networks whose parameters are fine-tuned using EAT (which, in turn uses FGM). Note that although there is a boost in overall accuracy against adversarial examples generated using FGM, the other attacks (1) DF, which is generated in a very different manner compared to FGM, and (2) PDG, which

Table 2. Differential immunity of various ensembles and their accuracy ($\alpha = 1$).

Networks	Differential immunity (δ)	Accuracy of best constituent net	Accuracy of MTDeep	Gain
FashionMNIST	0.11	3%	24.8%	21.8%
MNIST	0.19	0%	23.68%	23.68%
ImageNET	0.34	22.2%	42.88%	20.68%
MNIST + EAT	0.78	4.41%	54.71%	50.3%

represents a stronger class of attacks, are both still able to cripple the individual constituent networks. Although we do not presently understand why EAT helps is reducing the transferability of the PDG and DF attacks, this phenomenon helps MTDeep, when used in conjunction to the EAT, obtain impressive accuracy gains. In Fig. 5, we see that when $\alpha = 1$ (only adversarially perturbed inputs at test time) the accuracy of the constituent networks are 0–4% while MTDeep achieves an accuracy of $\approx 55\%$ Thus, we see a gain of more than 50% when classifying only adversarially perturbed images.

4.3 Blackbox Attacks on MTDeep

MTDeep designs a strategy based on a set of known attacks. Once deployed, an attacker can train a substitute network via distillation, i.e. use MTDeep as an oracle to obtain labels for the (chosen-ciphertext like) training set for the substitute network. Given that the distilled network captures information relating to the randomization at test time, we wanted to see how effective such a distillation procedure is in generating an expected network that mimics MTDeep. More specifically, we want to know if adversarial samples generated on this distilled network [14] successfully transfer against the MTDeep ensemble.

For this purpose, we used the three networks designed for MNIST data and experimented with $\alpha = 1$. We notice that MTDeep has higher immunity to blackbox attacks and is able to classify attack inputs $\approx 32\%$ of the time compared to the $\approx 24\%$ accuracy against white-box attacks. Thus, there exists a white-box attack in the attacker's arsenal that strictly dominates the black-box attack[4]. Thus, the defender's optimal mixed strategy remains unaffected.

4.4 Differential Immunity

If an attack $u \in U$ could cripple all the networks $n \in N$, using MTDeep will provide no gains in robustness. In this section, we try to quantify the gains MTDeep can provide. Let $E : N \times U \rightarrow [0, 100]$ denote the fooling rate function

[4] Note that even if a blackbox attack proves to be a more effective attack against the ensemble (for a different dataset), this attack is not modeled by the defender in the original game. They may choose to include it in the formulated game.

Attacks	0	1	2	3
FGM_C	4788	3641	1449	118
FGM_H	389	2728	6667	212
FGM_M	1513	5790	2479	214
FGM_{BB}	2305	2569	2678	2444

Fig. 6. Agreement among constituent networks when classifying perturbed inputs for the MNIST data-set.

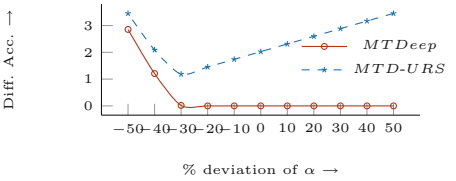


Fig. 7. Loss in % accuracy when correct α is different from assumed α.

where $E(n, u)$ is the fooling rate when an attack u is used against a network n. Now, the differential immunity δ of an ensemble N against a set of known attacks U can measured as follows,

$$\delta(U, N) = \min_u \frac{\max_n E(n, u) - \min_n E(n, u) + 1}{\max_n E(n, u) + 1}$$

If the maximum and minimum fooling rates of u on n differ by a wide margin, then the differential immunity of MTDeep is higher. The denominator ensures that an attack which has high impact (or fooling rate) reduces the differential immunity of a system compared to a low impact attack even when the numerator is the same. The $+1$ factor the numerator ensures that higher values of $\max_n E(n, u)$ reduce the δ when $\max_n E(n, u) = \min_n E(n, u)$. Note that $\delta \in [0, 1]$. As per this measure, the differential immunity of the various ensembles used in our experiments are highlighted in Table 2.

As per our expectation, we observe a general trend that the differential immunity of an ensemble is proportional to the accuracy gains obtained by MTdeep when compared to the most secure constituent network in the ensemble. Although we notice the lowest gain in case of ImageNET, note that this 20.68% gain in accuracy is substantially better than the $\approx 22\%$ (or t $\approx 24\%$) gain in accuracy for the Fashion-MNIST (or MNIST) dataset(s) with non-adversarially trained DNNs because the number of classes in ImageNET is 1000 compared to 10 for the latter two datasets.Lastly, existing measures of robustness are mostly designed for a single DNN [3,23] and thus, cannot capture the effect of attack transferability on robustness (of an ensemble). Thus, we propose *differential immunity* as one of the metrics for evaluating the robustness of ensembles that use any form of randomization at test time.

Disagreement Metrics. In Fig. 6, we highlight the number of perturbed test images (total 10000) on which 0–3 constituent DNN's classification output(s) agree with the correct class label. We conducted these experiments using the non-adversarially trained networks for MNIST classification and for brevity purposes, use only the FGM attack method. FGM_C is the strongest attack that can make all the $n \in N$ misclassify at least 70% of the images. As generating δ can be costly at times, which needs the fooling rates for each pair (u, n), one can

generate the agreement metrics on a small data set to provide upper bounds for δ. This provides an idea as to how using a MTDeep ensemble can increase the robustness against adversarial samples. In this case, $\delta_{MNIST} \leq 0.51$ because for the strongest attack, every network in the ensemble will misclassify (approx.) 49% of the time. Also, note that a majority based ensemble can will only be able to guarantee an accuracy of $\approx 14\%$ against the FSM_C attack because in all the other cases, only net 0 or net 1 is able to predict the correct class. In comparison, MTDeep can obtain an accuracy of 26.8% against FGM attacks.

4.5 Participation of Individual Networks

In Fig. 3(b), we explore the participation of individual networks in the mixed strategy equilibria for MTDeep used to classify ImageNET data. The results clearly show that while it is useful to have multiple networks providing differential immunity (as testified by the improvement of accuracy in adversarial conditions), the leveling-off of the objective function values with more DNNs in the mix does underline that there is much room for research in actively developing DNNs that can provide greater *differential immunity*. Note that no more than four (out of the six) networks participate in the equilibrium. An ensemble of networks with higher differential immunity equipped with MTD can thus provide significant gains in both security and accuracy.

4.6 Robustness Against Miscalibrated α

If the value of α, which is assumed up-front, turns out to be incorrect, the computed strategy ends up becoming sub-optimal. In Fig. 7, we plot the deviation of the chosen policy (based on the assumed α) from the optimal as the real α is varied $\pm 50\%$ from the one assumed. The BSG-framework remains quite robust (as opposed to a uniform random strategy) i.e. the accuracy is within 0–3% of the optimal accuracy. The robustness to α further highlights the usefulness of MTDeep as a meta-defense meant to work not only against adversarial attacks but also in the context of a deployed classifier that will have to deal with adversaries as well as legitimate users.

5 Conclusion

In this paper, we introduced MTDeep – a framework inspired by Moving Target Defense in cybersecurity – as 'security-as-a-service' to help boost the security of existing classification systems based on Deep Neural Networks (DNNs). We modeled the interaction between MTDeep and the users as a Bayesian Stackelberg Game, whose equilibrium gives the optimal solution to the multi-objective problem of reducing the misclassification rates on adversarially modified images while maintaining high classification accuracy on the non-perturbed images. We empirically showed the effectiveness of MTDeep against various classes of attacks for

the MNIST, the Fashion-MNIST and the ImageNet data-sets. Lastly, we demonstrated how using MTDeep in conjunction with existing defense mechanisms for DNNs result in more robust classifiers, thereby highlighting the importance of developing ensembles with higher differential immunity.

Acknowledgments. We thank the reviewers for their comments. This research is supported in part by NASA grant NNX17AD06G and ONR grants N00014161-2892, N00014-13-1-0176, N00014-13-1-0519, N00014-15-1-2027. The first author is also supported by an IBM Ph.D. Fellowship.

References

1. Abbasi, M., Gagné, C.: Robustness to adversarial examples through an ensemble of specialists. arXiv:1702.06856 (2017)
2. Adam, G.A., Smirnov, P., Goldenberg, A., Duvenaud, D., Haibe-Kains, B.: Stochastic combinatorial ensembles for defending against adversarial examples. arXiv:1808.06645 (2018)
3. Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., Criminisi, A.: Measuring neural net robustness with constraints. In: NIPS (2016)
4. Biggio, B., Fumera, G., Roli, F.: Adversarial Pattern Classification Using Multiple Classifiers and Randomisation. In: da Vitoria, Lobo N. (ed.) SSPR /SPR 2008. LNCS, vol. 5342, pp. 500–509. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89689-0_54
5. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. arXiv:1708.03999 (2017)
6. De La Escalera, A., Moreno, L.E., Salichs, M.A., Armingol, J.M.: Road traffic sign detection and classification. IEEE Trans. Ind. Electron. **44**(6), 848–859 (1997)
7. He, W., Wei, J., Chen, X., Carlini, N., Song, D.: Adversarial example defenses: ensembles of weak defenses are not strong. arXiv preprint arXiv:1706.04701 (2017)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015)
9. Javed, O., Shah, M.: Tracking and object classification for automated surveillance. In: ECCV (2006)
10. Jayadevan, R., Kolhe, S.R., Patil, P.M., Pal, U.: Automatic processing of handwritten bank cheque images: a survey. J. Doc. Anal. Recogn. **15**(4), 267–296 (2012). https://doi.org/10.1007/s10032-011-0170-8
11. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
12. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. arXiv:1610.08401 (2016)
13. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: CVPR (2016)
14. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: ACM CCS (2017)
15. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P) (2016)

16. Paruchuri, P., Pearce, J.P., Marecki, J., Tambe, M., Ordonez, F., Kraus, S.: Playing games for security: an efficient exact algorithm for solving Bayesian stackelberg games. In: AAMAS (2008)
17. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
18. Sengupta, S., et al.: A game theoretic approach to strategy generation for moving target defense in web applications. In: AAMAS (2017)
19. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv:1312.6199 (2013)
20. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P.: Ensemble adversarial training: attacks and defenses. arXiv:1705.07204 (2017)
21. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
22. Vorobeychik, Y., Li, B.: Optimal randomized classification in adversarial settings. In: AAMAS (2014)
23. Weng, T.W., et al.: Evaluating the robustness of neural networks: an extreme value theory approach. arXiv preprint arXiv:1801.10578 (2018)
24. Zheng, S., Song, Y., Leung, T., Goodfellow, I.: Improving the robustness of deep neural networks via stability training. In: CVPR (2016)
25. Zhuang, R., DeLoach, S.A., Ou, X.: Towards a theory of moving target defense. In: Proceedings of the First ACM Workshop on Moving Target Defense, pp. 31–40. ACM (2014)