Deep Neural Networks Do Not Recognize Negative Images

Hossein Hosseini and Radha Poovendran
Network Security Lab (NSL)
Department of Electrical Engineering, University of Washington, Seattle, WA
Email: {hosseinh, rp3}@uw.edu

***Abstract*—Deep Neural Networks (DNNs) have achieved remarkable performance on a variety of pattern-recognition tasks, particularly visual classification problems, where new algorithms reported to achieve or even surpass the human performance. In this paper, we test the state-of-the-art DNNs with negative images and show that the accuracy drops to the level of random classification. This leads us to the conjecture that the DNNs, which are merely trained on raw data, do not recognize the semantics of the objects, but rather memorize the inputs. We suggest that negative images can be thought as "semantic adversarial examples", which we define as transformed inputs that semantically represent the same objects, but the model does not classify them correctly.**

(a) (b)

Fig. 1: Examples of the original image (left) and the negative image (right). a) An example from MNIST dataset, b) An example from GTSRB dataset. DNNs fail to recognize the negative images and classify them randomly into other classes.

I. Introduction

Deep Neural Networks (DNNs) have transformed the machine learning field and are now widely used in many applications [1]. One of the fields that has benefited the most from deep learning is computer vision, where DNNs have achieved state-of-the-art results on variety of problems [2]–[4]. Several recent works in this domain have reported that new algorithms can reach or even surpass the human performance on tasks such as image classification problems [5]–[7].

Humans have an overall sense of the objects and can recognize them in various forms such as different scales, orientations, colors or brightness. Several recent works have hypothesized that DNNs also develop an understanding about the objects based on the training data, as such that they are even able to generate new images [8], [9].

To assess the behavior of the image classification models, we evaluate the performance of DNNs on *negative images* of the training data. A negative is referred to an image with reversed brightness, i.e., the lightest parts appear the darkest and the darkest parts appear lightest. These complemented images are often easily recognizable by humans. We show when testing on negative images, the accuracy of DNNs drops to the level of the random classification, i.e., the network maps the inputs randomly to one of the output classes. This shows that the DNNs, which are merely trained on raw data, do not learn the structures or semantics of the objects and cannot generalize the concepts.

Moreover, the fragility of learning models to transformed inputs has a security implication. An adversary, who has no access to the learning system, can generate transformed inputs which are semantically representing the same objects, yet the system does not correctly classify them. We call such transformed inputs as *semantic adversarial examples*.

II. Negative Images

New approaches in computer vision try to understand and imitate the human vision system [10], [11]. However, it has been shown that image classification algorithms, although capable of achieving high accuracies on regular data, show certain differences with the human perception of the objects [12], [13]. In this paper, we intend to assess whether DNNs can recognize the transformed samples of the training data. In particular, we examine them with *negative images*. [1]

Formally, we define the negative image as the image complement. Let X be an image and $X_i \in [0, 1]$ be the i-th feature (pixel). The negative image is defined as X^*, where $X_i^* = 1 - X_i$. Image complementing is a simple transformation, which is easy to compute and, although it may introduce a large perturbation to the original image, it typically does not impact the human perception of the object. Figure 1 shows examples of original and negative images from two image datasets.

III. Experimental Results

We demonstrate the results on DNNs trained on two image datasets, MNIST [14] and German Traffic Signs Recognition Benchmark (GTSRB) [15]. MNIST is a dataset of handwritten digits with $70,000$ gray-scale images of size 28×28, split into three groups of $50,000$ training samples, $10,000$ validation samples, and $10,000$ test samples. The task involves the

This work was supported by ONR grants N00014-14-1-0029 and N00014-16-1-2710, ARO grant W911NF-16-1-0485 and NSF grant CNS-1446866.

[1] We borrowed the term "negative images" from photography.

TABLE I: Accuracies of different classifiers trained on MNIST and GTSRB datasets. We train the classifier on regular training data and evaluate it on test images and negative images of the training data.

	Classification Accuracy on MNIST		Classification Accuracy on GTSRB	
DNN Classifier	Test Images	Negative Images of Training Data	Test Images	Negative Images of Training Data
DNN	99.20%	6.68%	95.63%	4.39%
Robust DNN ($\epsilon = 0.05$)	99.34%	6.72%	98.08%	7.07%
Robust DNN ($\epsilon = 0.1$)	98.30%	3.89%	97.39%	8.44%
Robust DNN ($\epsilon = 0.2$)	98.77%	6.24%	94.55%	11.12%
Robust DNN ($\epsilon = 0.5$)	98.74%	16.69%	94.80%	13.54%

classification of images of digits into one of 10 classes, corresponding to digits 0 to 9. GTSRB is a dataset of color images of 43 traffic signs, with 39,209 training samples, and 12,630 test samples. We randomize the ordering of the training data and use the first 5,000 samples as the validation set. Also, since images vary in size, we resize all images to the same size as MNIST inputs (28×28 gray-scale images) and pass them through a histogram equalization filter.

We use the same DNN structure for both datasets, which is described as follows. The network has two convolutional layers with 32 and 64 filters and with a kernel size of 3×3, followed by a max-pooling layer. It is then followed by two fully-connected layers with 200 rectified linear units. The classification is made by a softmax layer.

We present the results for a DNN and the robust DNNs trained to be resilient against adversarial perturbations [16]. A robust DNN with parameter ϵ is a DNN trained on both the regular and adversarily modified images, where $\epsilon \in [0, 1]$ is the maximum allowed change to each feature value. Note that the input features (image pixel values) are continuous and bounded between 0 and 1.

In experiments, we train the DNN classifiers on regular training images and evaluate them on test images and also negative images of the training data. Table I provides the accuracies of different DNNs trained on the two datasets MNIST and GTSRB. DNNs typically yield very high accuracy ($\sim 100\%$) on training images and seem to generalize very well to test images, since the test accuracies are also very high. However, the networks fail to correctly classify the negative images of the training data, meaning that they cannot recognize the negative images. Note that, for MNIST and GTSRB datasets, a classifier that maps the inputs to random classes would yield 10% and $\sim 2.33\%$ accuracy, respectively, since the datasets have 10 and 43 number of classes.

IV. Discussion

A. *Fragility of DNNs to Transformed Inputs*

It has been shown that the effective capacity of neural networks is sufficient for memorizing the entire training dataset [17]. As a result, DNN classifiers generally correctly classify the training samples with very high confidence. Besides, the network loss function vary smoothly around the input samples, i.e., a *randomly* perturbed sample is likely to be classified into the same class as the regular sample [18]. Also, since the test samples are typically collected from the same distribution as the training samples, the test data points occur mostly in vicinity of the training points. Therefore, with the availability of large datasets, it is likely that the network can associate each test sample with one or several training samples from the same class and thus achieve high test accuracy.

However, since a transformed sample may be far from the original sample, the network cannot correctly classify it. While, for a particular transformation, we can train the DNN also on the transformed data to get high accuracy on them, relying on large and diverse datasets, which cover all aspects of possible novelties in the test data, seems to pose a fundamental problem to machine learning systems. It causes the models to require a lot of data in order to understand every feature, which clearly does not scale for real-world applications.

While many computer vision problems are data rich, for some critical applications, e.g., training driveless cars, gathering diverse training data is costly. Several works have proposed ideas, such as one shot learning [19]–[21], to learn with small datasets. Humans, in contrast, are able to capture the semantics of the new objects by incorporating their understanding of the concepts such as shape and color. Likewise, learning models need to import such general concepts in form of "accessories" from other models. In essence, learning to reason can compensate for the lack of diversity in the training data and help the model to "semantically generalize". In fact, considering the human vision system, semantic generalization is possibly a more important aspect of the artificial intelligence.

B. *Security Implication*

In machine learning classifiers, while only a tiny fraction of the input space is used for training, all possible inputs are allowed during the test phase. This causes the classifier to show unexpected behaviors when it is queried with invalid inputs. Several recent works have pointed out such behaviors of the image classifiers, e.g., against fooling [13] and adversarial [12] images. Fooling images are images which are completely unrecognizable to humans, but learning models are *fooled* into classifying them as valid objects with high confidence. Also, adversarial examples are inputs which are maliciously modified to deceive the classifier model, while a human observer can correctly recognize them.

In contrast, we showed the fragility of the learning models against transformed inputs which semantically represent the same object, but the learning model does not classify them correctly. Therefore, an adversary, who does not have any access to the learning system, can easily deceive it by applying transformations to the input images, which do not affect the human perception. We call such inputs as *semantic adversarial examples*.

We presented the image complementing as one such transformation. Note that for generating regular adversarial examples, the perturbation must be small to remain unnoticeable by a human observer [22]. In contrast, complementing the image may introduce a large perturbation to the image, yet a human observer can correctly recognize it.

V. Conclusion

In this paper, we showed that, despite the impressive performance of DNNs on regular data, their accuracy on negative images is at the level of random classification. This observation indicates that the DNNs that are simply trained on raw data cannot recognize the semantics of the objects and possibly only memorize the inputs. The inability of recognizing the transformed inputs shows the shortcoming of current training methods, which is that learning models fail to semantically generalize. We also introduced the concept of semantic adversarial examples, as transformed inputs which appear the same to a human observer, yet the learning model does not classify them correctly.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[5] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1058–1066, 2013.

[6] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[7] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *AISTATS*, vol. 2, p. 5, 2015.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[9] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1538–1546, 2015.

[10] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[11] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1847–1871, 2013.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[13] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

[14] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.

[15] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[18] P. Tabacof and E. Valle, "Exploring the space of adversarial images," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 426–433, IEEE, 2016.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[20] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," in *Advances in neural information processing systems*, pp. 935–943, 2013.

[21] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065*, 2016.

[22] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pp. 372–387, IEEE, 2016.