Robustness of Deep Convolutional Neural Networks for Image Recognition

Matej Uličný, Jens Lundström, and Stefan Byttner

Intelligent Systems Department, Halmstad University,
Box 823, S 301 18 Halmstad, Sweden
mtj.ulicny@gmail.com, jens.lundstrom@hh.se, stefan.byttner@hh.se
http://islab.hh.se

Abstract. Recent research has found deep neural networks to be vulnerable, by means of prediction error, to images corrupted by small amounts of non-random noise. These images, known as adversarial examples are created by exploiting the input to output mapping of the network. For the MNIST database, we observe in this paper how well the known regularization/robustness methods improve generalization performance of deep neural networks when classifying adversarial examples and examples perturbed with random noise. We conduct a comparison of these methods with our proposed robustness method, an ensemble of models trained on adversarial examples, able to clearly reduce prediction error. Apart from robustness experiments, human classification accuracy for adversarial examples and examples perturbed with random noise is measured. Obtained human classification accuracy is compared to the accuracy of deep neural networks measured in the same experimental settings. The results indicate, human performance does not suffer from neural network adversarial noise.

Keywords: adversarial examples, deep neural network, noise robustness

1 Introduction

In visual recognition problems, deep neural networks (DNN's) represent the state-of-the-art models outperforming all the other machine learning algorithms. The use of neural networks for visual recognition has application in many fields, from web applications to industrial products such as safeguards in automobile industry. Despite their outstanding performance, they have pitfalls in their understanding of problem they are trained to solve. Szegedy et al. have discovered robustness flaws in many machine learning methods [11]. Despite the fact that the most of machine learning methods exhibit these flaws, this article specializes exclusively on deep neural networks. Deep neural networks have problems to correctly classify images altered by non-random noise, imperceptibly different from images that have been classified correctly. Such flaw is unacceptable if neural networks are used for safety protocols or for verification programs. In this work we define robustness as the ability to correctly classify similar inputs.

Without a robust solution, attacker is able to create examples that can perturb the network.

We approach the problem by testing several robustness methods and by comparing their results. The article provides investigation of robustness not only to adversarial noise, but also to random noise. All experiments are performed on the MNIST [8] data-set. Different approaches, such as various configurations of dropout or pre-processing the input are examined. Robustness experiments are finalized with a study of adversarial training and robustness of various types of committees.

Several articles report human visual recognition accuracy on certain data-sets [1], [12]. We provide an estimate of human perception ability on noisy MNIST images and we compare it to the accuracy of deep neural networks.

The paper is organized as follows, Section 2 presents an outline of a related work in the field. Perturbations and robustness methods we use in experiments are briefly described in Section 3. Method description is followed by experimental setup (Section 4) and results of the experiments (Section 5). The paper is finalized by the main conclusions and by a discussion of their aspects (Section 6).

2 Related Work

Recent discoveries by Szegedy et al. [11] opened a whole new branch for research of DNNs. Instead of describing improvement in DNNs' generalization performance, they focused on discovering neural networks' weaknesses. Firstly, Szegedy et al. showed that it is the entire space of activations rather than individual units that contains semantic information. The rest of their work is oriented in finding the DNN's blind spots. The most important findings made by Szegedy et al. in [11] are:

1. For all the networks studied, for every tested image, the authors were able to generate an adversarial example, which was for humans visually almost indistinguishable from original image, that was misclassified by the original network.
2. Cross model generalization: a large number of adversarial examples are misclassified by networks trained with different hyper-parameters (number of layers, initial weights, etc.).
3. Cross training-set generalization: a large number of examples are misclassified by networks trained on a disjoint training set.

These discoveries pose question related to how the universal approximators can be so vulnerable against such subtle changes. This discovery undermines smoothness property of neural networks claiming that inputs close to each other are supposed to be assigned to the same class. Their experimental results suggest that using adversarial examples in training process may improve generalization performance.

Moreover, Nguyen et al. [9] were experimenting with creating visually meaningless images, which were classified by a neural network as one of the image

categories with high confidence reaching 99.99%. The authors named these examples "fooling images". Nguyen et al. made a hypothesis that these fooling examples are caused by the discriminative nature of classifier, permitting the algorithm to find an example that is far away from discriminative boundary as well as from all the data that have been seen before.

Goodfellow et al. [3] were trying to discover the reason why adversarial examples exist. They claim, existence of adversarial examples stem from models being *too linear*. Authors believe, adversarial perturbations are dependent on model's weights, which are similar for different models learned to perform the same task. They observed, a generalization of adversarial noise across different natural examples is caused by the fact that adversarial perturbations are not dependent on a specific point in space but on direction of the perturbation. Further in the work by Goodfellow et al., experiments comparing resistance of models with different capacity against adversarial and fooling examples have been performed. In their paper it was shown that models, which are simple to optimize yield easily to adversarial and fooling examples, thus they have no capacity to resist these perturbations. Adversarial training is presented by Goodfellow et al. as a possible tool for further regularization (than solely use methods such as dropout).

Gu & Rigazio [4] used various pre-processing methods to diminish adversarial perturbations. They have tested several denoising procedures including: injection of additional Gaussian noise with subsequent Gaussian blurring, more sophisticated methods using autoencoder trained on adversarial examples, and a standard denoising autoencoder. Gu & Rigazio believe DNN's sensitivity is affected by training procedure and the objective function rather than by network topology. As a possible solution to achieve local generalization in the input space, they propose a deep contractive neural network.

The contribution of this work, in relation to the mentioned studies is in investigating the network sensitivity to adversarial noise affected by dropout regularization applied to various combinations of network layers. Apart from adversarial noise examination, robustness methods are inspected when dealing with random noise. Moreover, this research contains a study of adversarial noise resistance of committees and combinations of robustness methods. Human object recognition accuracy has been reported [1], [12] and compared to accuracy of DNNs examined on natural images. An open question is how the accuracy changes on noisy images. To answer this question, this paper presents a comparison of human visual recognition ability of images distorted by different types of noise, with the performance of the state of the art deep convolutional neural network.

3 Perturbations and Robustness Methods

This section introduce the different types of perturbations used in the experiments, along with well-known robustness methods and the proposed method.

3.1 Perturbations

To examine neural network robustness, various perturbations are designed to corrupt an image. A perturbed image $\tilde{X}$ is composed of original image X where each element is in range $(0, 1)$ and additive noise R, expressed as $\tilde{X} = X + R$. In our framework, two noise types are used as an additive noise. Random noise taken from Gaussian distribution and non-random adversarial noise obtained by gradient sign method [3]:

$$R = \epsilon \operatorname{sign}(\nabla_X L(\boldsymbol{\theta}, X, \mathbf{y})). \tag{1}$$

The adversarial noise is created by a model with parameters $\boldsymbol{\theta}$ for input data X with the corresponding target vector $\mathbf{y}$ via a sign function for the gradient of the loss function $L(\boldsymbol{\theta}, X, \mathbf{y})$.

3.2 Distortion measure

To make a simple comparison of the amount of noise that is injected to images, a measure is designed, the average distortion per pixel *dist* for the whole data set, independent of type of the noise:

$$dist = \frac{1}{n}\frac{1}{c}\frac{1}{h}\frac{1}{w}\sum_{k=1}^{n}\sum_{l=1}^{c}\sum_{i=1}^{h}\sum_{j=1}^{w}\left|\tilde{X}_{klij} - X_{klij}\right|. \tag{2}$$

Average noise is calculated over all color channels c, image height h and image width w for every picture in the set containing n elements.

This distortion measure facilitates a comparison of perturbed images at similar noise quantification levels. Images perturbed with Gaussian-generated noise and with gradient sign noise, which are the main concern of the article, are visually compared in Fig. 1.

3.3 Robustness methods

To face problems caused by adversarial noise, several methods for increasing robustness of DNN's have been proposed. The following section start with description of well-known robustness methods and end by characterizing the proposed method for dealing with adversarial noise.

Dropout Dropout is a regularization method presented by Hinton et al. [5], which prevent networks from overfitting by dropping out random nodes along with their connections in each training iteration. This method can be applied to one or more layers. For each layer, a different probability to skip a node may be used. In each training iteration, a different thinned network is trained. Output from all the thinned networks can be easily approximated by using the whole network with activations scaled by the probability of the node is used in the training phase. The method comes with a price of longer training time (two to three times as reported by Srivastava et al. [10]).

Fig. 1. Visualization of MNIST images affected by random and gradient sign noise at different distortion levels.

Low-pass Filter Low-pass filtering is an input pre-processing method, which uses blurring/denoising convolution to clean noise, see Gu & Rigazio [4]. To boost the blurring effect on adversarial noise, before applying the convolutional filter, regions created by adversarial noise are destroyed with additive Gaussian noise. This process aims to move input image from unrecognizable regions to the form in which it can be correctly classified.

Denoising Autoencoder Denoising autoencoder (DAE) is a generative neural network that is used to reconstruct corrupted inputs. It was used by Gu & Rigazio [4] as a pre-processing method with purpose to clean input image of adversarial noise. This is facilitated by its symmetric bottleneck network topology.

Adversarial Training A one way to increase robustness of a neural network is to train the network on its own adversarial examples. Szegedy et al. [11] tested a procedure, in which a neural network was trained on set regularly updated by a pool of newly created adversarial examples. Goodfellow et al. [3] used

another approach, training a neural network on an ordinary training set using an adversarial objective function.

Adversarial Committee It has been shown [11] that other models are less affected by adversarial examples than the model, which the examples are designed to perturb. Models trained on adversarial examples show good performance when classifying adversarial examples of another model. Based on these observations, we propose a committee of models trained on adversarial examples. A standard model, trained on natural training set is consecutively trained on its own adversarial examples. Adversarial examples used to train the model are created in two stages. In the first stage, gradient sign noise examples are created. The second stage involves addition of gradient scaled by a constant to the gradient sign noise examples. For many natural images, a magnitude of the gradient is too small compared to the range of input parameters. Gradient magnitude of an adversarial example greatly exceeds gradient magnitude of the natural image, hence we use the gradient sign noise images to produce the gradient instead of natural images. The training image

$$\tilde{X} = X + R + c\nabla_X L(\boldsymbol{\theta}, X + R, \mathbf{y})) \tag{3}$$

is produced as a linear combination of a gradient sign noise example $X + R$ (Eq. 1) and a gradient of the loss function for the gradient sign image scaled by a constant c. After a fixed amount of training iterations a snapshot of the model is saved and used to produce new examples that update training set pool. All of these snapshots including the model trained purely on natural training set are combined into a committee. In deployment stage, the committee outputs an average prediction of all committee members. An advantage of this method is that it is difficult to generate gradient sign noise for the committee, since its members are trained to recognize images corrupted with noise derived from other committee members.

4 Experimental Setup

The robustness is observed on MNIST data-set, for which the baseline is obtained from a modification of the original LeNet network [7], see Fig. 2. The network is initialised by Xavier algorithm [2], trained by stochastic gradient descent while using momentum and L2 regularization.

4.1 Robustness Experiments

We measure robustness in form of generalization error obtained for 100 test sets, all originated from MNIST test set, perturbed to have different distortion levels. Results from different test sets facilitate graphical visualization of how the error changes with increased distortion.

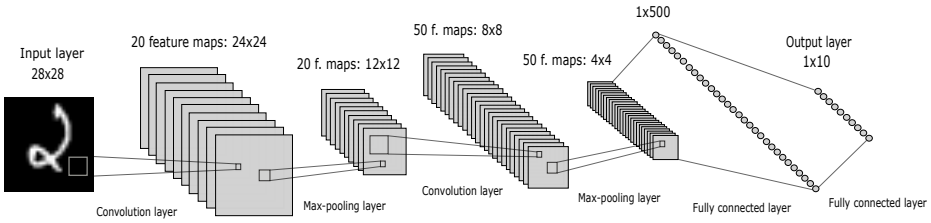

Fig. 2. Architecture of LeNet network, composed of two convolution layers, two max-pooling layers, and two fully connected layers.

Table 1. Layer-specific dropout rate (rate at which nodes are being randomly dropped out) of models used in experiments.

Abbreviation	Input	Conv1	Conv2	Full	Output	Train iterations
Inp	0.2	0	0	0	0	35000
All	0.2	0.2	0.2	0.5	0.5	40000

At first, a robustness of models with dropout applied to various layers is measured. Details (dropout rates) of models created by performing dropout on different layers of LeNet are contained in Table 1.

Further experimentation determines effect of low-pass filter and denoising autoencoder pre-processing on models' robustness. Low-pass filter is used to remove high frequencies representing the noise. For this purpose, three Gaussian convolution kernels of sizes 3x3, 5x5 and 7x7 were designed. Each kernel was filled with Gaussian function defined in range (-3, 3). Function's variance was chosen for each filter separately, under restriction that classification error on clean set cannot cross 1%. Variances of 0.6, 1.3 and 7 were chosen for filters of size 7, 5 and 3 respectively. The effect of filtering applied to adversarial examples is enhanced by injecting Gaussian noise to images before the filter is applied.

Low-pass filter pre-processing is compared with pre-processing of denoising autoencoder (DAE). Experimental DAE with structure 784-1000-500-250-30-250-500-1000-784 is trained by Nesterov's accelerated gradient. Denoising effect is achieved by applying dropout on the input layer. Robustness to adversarial noise is facilitated by stacking fully trained denoising autoencoder to the bottom layer of LeNet. Denoising autoencoder cleans data from noise and feed clean images to LeNet's input layer. Three stacked networks have been created: denoising autoencoder stacked to LeNet and denoising autoencoder stacked to LeNet with dropout on input layer and to LeNet with dropout on input, conv1, conv2, full and output layer (Table 1).

Adversarial training in the proposed constellation is realized by training the model with its own adversarial examples. Networks trained on adversarial examples originate from LeNet model trained for 10000 iterations, regularized by weight decay. This model is further consecutively trained for 5 times, each time for 5000 iterations on current set of adversarial examples joined with training sets of previous models. After every phase of adversarial training a new model

is created. All of these 6 models are combined to a committee. To point out their robustness, a comparison with basic committee composed of 6 models is demonstrated.

4.2 Human Noise Processing Experiments

In our experiments, the performance of human ability to recognise digits is obtained for four different sets of images that are presented to test subjects, three of them originating from MNIST test set and one from USPS [6] data set. One of the sets is composed of 15 natural images, the other two sets consists of natural images corrupted either by Gaussian noise or by gradient sign noise. Each noise type set contains 6 subsets per 15 images with graduating noise levels. Noise levels are designed to have similar distortions for pairs of Gaussian and gradient sign noise subsets. Gradient sign noise levels are defined by $\epsilon = 0.07, 0.14, 0.21, 0.28, 0.34, 0.4$ matched with Gaussian noise levels created using $\sigma = 0.09, 0.18, 0.27, 0.36, 0.45, 0.54$. The last set of 10 images are taken from USPS dataset and resized to match dimensions of MNIST pictures. Digits on these images are written in a way that may be difficult for humans to decipher, thus they are noted in this work as the "human adversarial noise".

Each mentioned noise level subset is randomly picked from the whole set of 10000 images, except for USPS images, of which 5 images per each digit are available and exactly one per digit is randomly chosen for the test. Subsets are shuffled and presented to test subject via a Python script, which captures subject's decisions and stores them for future analysis.

The data essential to carry out the experiment was gathered from 57 subjects. Every subject has classified a set of 205 images described in previous paragraphs. The accuracy is compared for each noise level separately. Each participant contributes with an average accuracy of his choices for each noise type and level separately. A collection of these accuracies achieved by all tested subjects is used to compute mean and standard deviation, which are compared with values obtained from tests of 20 LeNet models in the same experimental setup.

5 Results

5.1 Dropout

A two LeNet models (see Table 1) were trained with dropout regularization and tested for resistance to gradient sign adversarial noise. Regularizing network with dropout made the network more robust to adversarial distortion up to large levels of distortion (> 0.15) where all models perform poorly. The most robust solution was obtained by applying dropout on every layer. Error on adversarial examples with $\epsilon = 0.2$ was reduced from basic model error of 78.88% to 42.97% (see Fig. 3).

The same models regularized with dropout were tested for robustness by gradually increasing levels of Gaussian noise. As Fig. 4 depicts, the most resistant

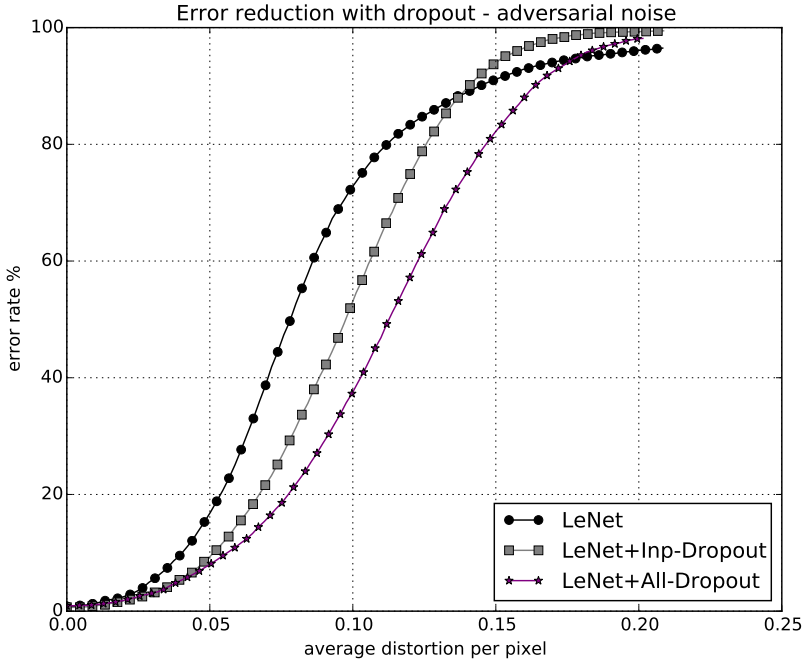

Fig. 3. Generalization error of models regularized with dropout on input layer (Inp-Dropout) and on all the layers (All-Dropout) compared to standard model when classifying adversarial examples.

model to random noise was a model regularized with dropout only in the input layer. Applying dropout to other layers seem to have negative effect on robustness to random noise.

5.2 Pre-processing Methods

In the experiments for pre-processing methods, images are processed either by low-pass filter or by denoising autoencoder. Pre-processed pictures were further fed into two different models to be classified, to an ordinary LeNet model and to a LeNet regularized with dropout. The denoising autoencoder has been found better at preparing adversarial examples for classification than low-pass filter. The highest accuracy has been achieved by pre-processing adversarial examples by a denoising autoencoder and subsequent classification by LeNet regularized by dropout on all the layers (see Fig. 5).

By examining the resistance of these methods to random noise, different results were obtained. Low-pass filter prepared randomly distorted images for classification better than denoising autoencoder. When dealing with Gaussian

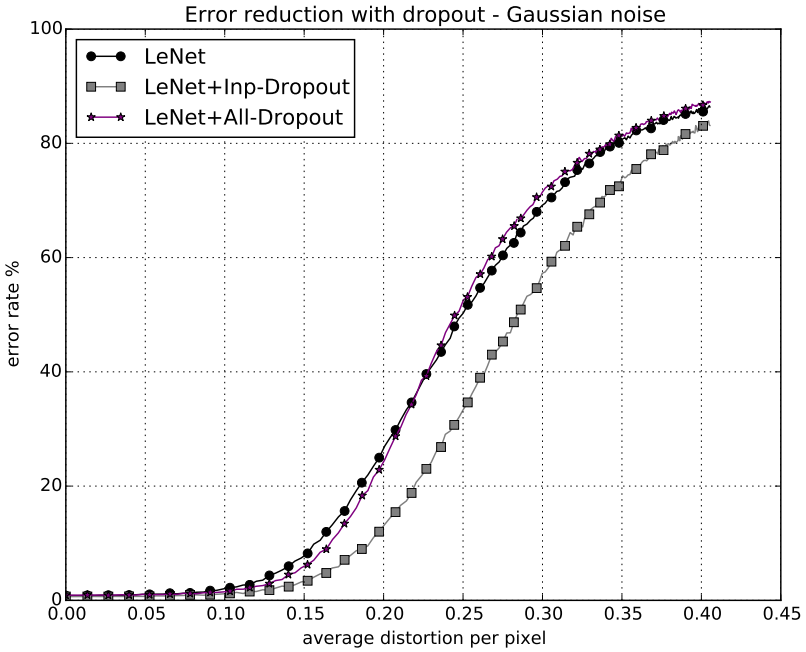

Fig. 4. Generalization error of models regularized with dropout on input layer (Inp-Dropout) and on all the layers (All-Dropout) compared to standard model when classifying images corrupted with Gaussian noise.

noise, encouraging results were achieved by low-pass filtering of images, later classified by LeNet regularized with dropout on the input layer (see Fig. 6).

5.3 Adversarial Training and Committees

As has been shown before [3], a standard committee does not provide desired robustness. To demonstrate the resistance of a committee of models trained on adversarial examples we propose, robustness tests comparing generalization error of this committee with the error of a standard committee and a single model are conducted (see Fig. 7). A committee of 6 models trained on adversarial examples gave an error of 36.5% on MNIST adversarial examples with $\epsilon = 0.37$ compared to error of a single model 96.1% or to the error of a basic averaging committee with the same amount of members, 97.2%. Table 2 compares results with other robustness methods.

5.4 Human Recognition Experiments

The performance is evaluated by calculating the accuracy as a ratio of correctly classified images to the total number of tested images. The results of human vi-

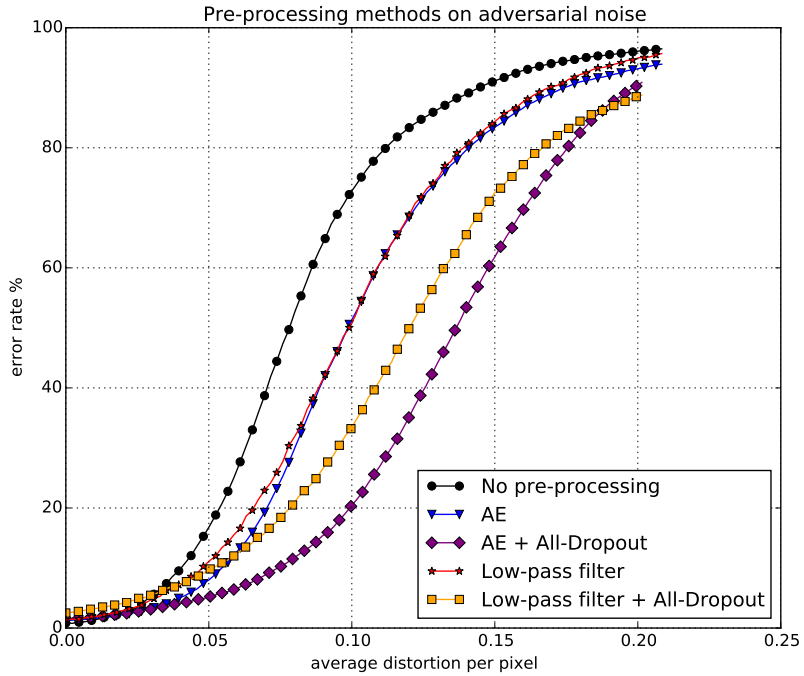

Fig. 5. Generalization error of LeNet when classifying pre-processed adversarial examples. Images are pre-processed either by low-pass filter or by denoising autoencoder.

Table 2. Generalization error (%) of robustness methods on MNIST test set perturbed at three different distortion levels.

Method	dist=0	dist=0.1	dist=0.2
No method	0.83	73.6	96.1
Dropout	0.91	35.7	98.1
Low-pass filter	1.38	52.0	95.0
DAE	1.27	52.4	93.2
Low-pass filter + dropout	2.51	34.8	88.5
DAE + dropout	1.56	21.4	90.3
Standard committee	0.86	70.9	97.2
Adversarial training after 5 stages	0.87	44.5	92.1
Committee of models trained on adv. examples	0.65	9.8	36.5

sion experiments are illustrated as curves composed of mean values for gradually increasing levels of each noise type separately, see Fig. 8. Curves are encapsulated by their 95% confidence regions. The experiment indicates that humans classify images (corrupted by Gaussian noise) with similar accuracy as the deep neural networks. Experiment also suggests, humans find adversarial and random noise

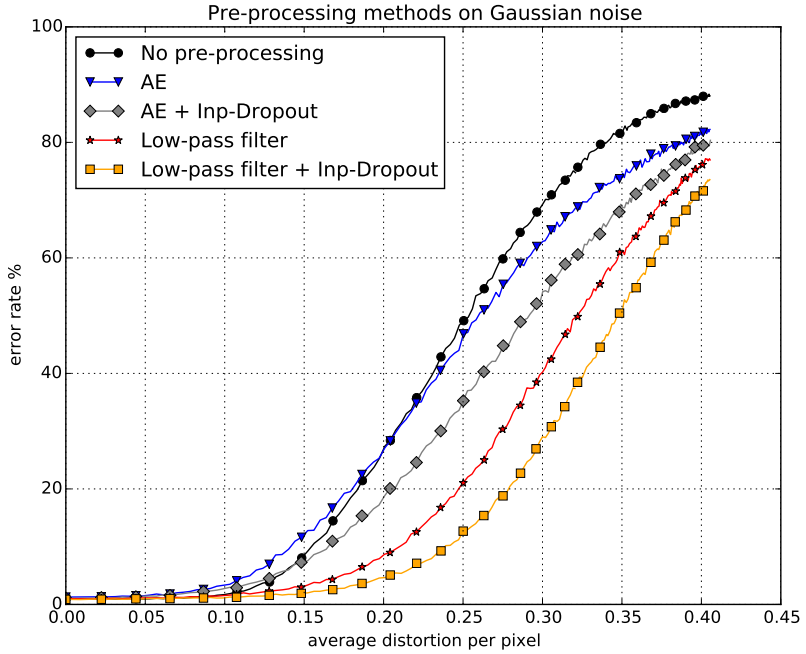

Fig. 6. Generalization error of LeNet when classifying preprocessed images corrupted by Gaussian noise. Images are preprocessed either by low-pass filter or by denoising autoencoder.

similarly problematic. In contrary, DNNs suffer a significant decrease in classification performance when classifying adversarial examples, compared to human performance on adversarial examples (see Fig. 8) or to DNNs' performance on random noise.

6 Discussion and Conclusion

For the MNIST dataset, we exploited fast and simplistic method (the gradient sign method) of creating adversarial examples. Due to the small number of output classes, one color channel and low image resolution, adversarial examples on MNIST are notably different from original images, however, far more harmful than randomly distorted images. Thus robustness experiments might be more accurate on datasets such as CIFAR100 or ImageNet that contain many image classes.

Dropout experiments in this paper suggest, regularization by dropout on every network layer is more effective on adversarial examples than using dropout just on input layer. Adversarial noise is affecting every layer, through which

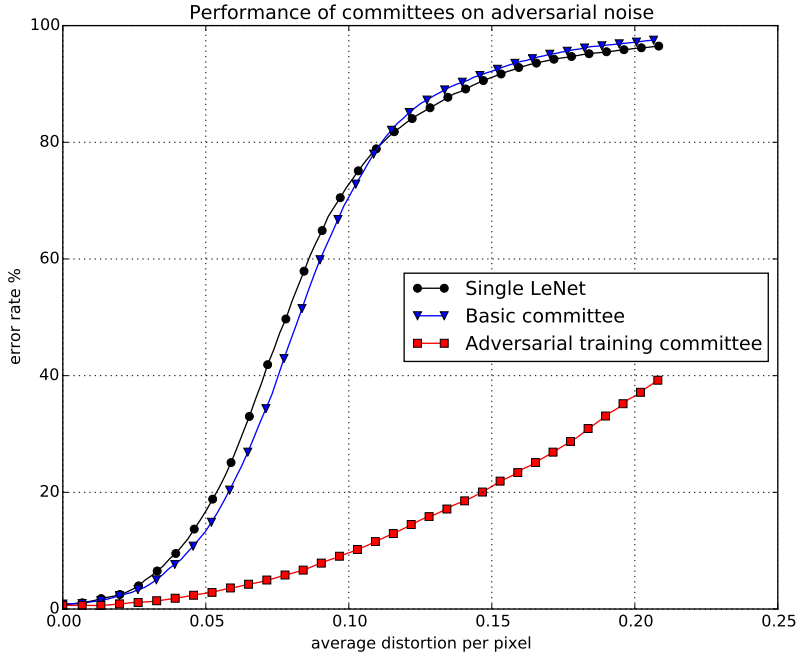

Fig. 7. Generalization error of committees on adversarial noise. Performance of committee trained on adversarial noise is compared to performance of a basic committee.

the gradient has been backpropagated, making adversarial noise more dependent on model's internal structure. Using dropout on model internal structure complicates creation of new adversarial examples. Random noise is more likely compensated by learning from incomplete inputs than by dropping out nodes in every network layer. A possible reason to this lies in the fact that, random noise is independent of model structure. During training, model is learning to recognize incomplete patterns, becoming more robust to single pixel random deviations.

Moreover, a few observations while comparing input pre-processing methods deserved to be noted. Low-pass filtering prepares an image distorted by random noise for correct classification better than the denoising autoencoder. Gaussian low-pass filter is a simple and powerful tool to suppress random Gaussian noise by averaging. Our results suggest that, adversarial examples reconstructed by denoising autoencoder are easier for neural network to classify than adversarial examples blurred by a low-pass filter. A possible explanation may be that the gradient sign noise splits image into regions, which are moved by the noise in the same direction. Averaging filter that performs well on random noise has small

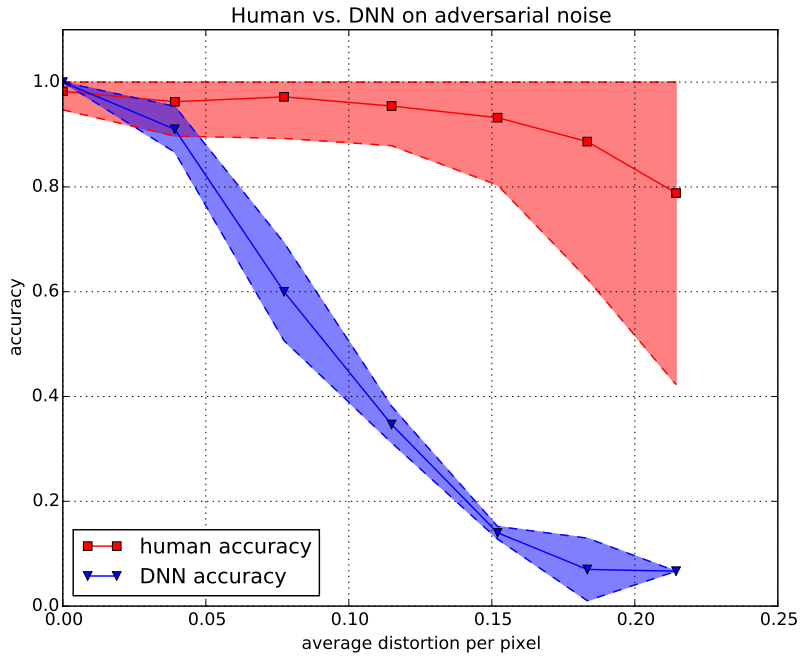

Fig. 8. Classification accuracy of humans compared to the accuracy of 20 LeNet models when classifying adversarial examples. Mean values are bounded by 95% confidence regions. Plot depicts, the performance of LeNet on adversarial noise is far inferior to performance of humans.

or almost no effect within these regions, whereas autoencoder is not limited to simple averaging.

Also, models trained on adversarial examples are consistently more robust to adversarial noise created by other models. Average prediction of these models forming a committee diminishes the chance to perturb the committee by exploiting the weakest model. We come to conclude, an ensemble of models trained on adversarial noise is more resistant to adversarial noise than any single model we have tested so far, for the MNIST dataset.

This paper also reported human accuracy when classifying images corrupted by random and adversarial noise. From obtained results, we come to an assumption, humans classify images perturbed by adversarial and by random noise with similar accuracy, unlike DNNs. DNNs' performance suffers greatly when classifying adversarial examples. Accuracy has been measured on an inconsistent test set. To avoid overfitting to a small set of images, every test subject was presented with different subset of images. Hence input set variance may have biased the results. For future work, there is an opportunity for a similar experiment: test-

ing each participant on the same image set, comparing the results to the results already obtained by this paper's experiments.

Acknowledgments We would like to express our gratitude to all participants of the human visual recognition experiment.

References

1. Cireşan, D., Meier, U., Masci, J., Schmidhuber, J.: Multi-column deep neural network for traffic sign classification. Neural Networks 32, 333–338 (2012)
2. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics (2010)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. CoRR abs/1412.6572 (2014), http://arxiv.org/abs/1412.6572
4. Gu, S., Rigazio, L.: Towards deep neural network architectures robust to adversarial examples. CoRR abs/1412.5068 (2014), http://arxiv.org/abs/1412.5068
5. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR abs/1207.0580 (2012), http://arxiv.org/abs/1207.0580
6. Hull, J.J.: A database for handwritten text recognition research. Pattern Analysis and Machine Intelligence, IEEE Transactions on 16(5), 550–554 (1994)
7. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (Nov 1998)
8. Lecun, Y., Cortes, C.: The MNIST database of handwritten digits http://yann.lecun.com/exdb/mnist/
9. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images (2015)
10. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15, 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
11. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. CoRR abs/1312.6199 (2013), http://arxiv.org/abs/1312.6199
12. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1701–1708 (June 2014)