ROBUSTNESS AND EQUIVARIANCE OF NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks models are known to be vulnerable to geometric transformations as well as small pixel-wise perturbations of input. Convolutional Neural Networks (CNNs) are translation-equivariant but can be easily fooled using rotations and small pixel-wise perturbations. Moreover, CNNs require sufficient translations in their training data to achieve translation-invariance. Recent work by Cohen & Welling (2016), Worrall et al. (2016), Kondor & Trivedi (2018), Cohen & Welling (2017), Marcos et al. (2017), and Esteves et al. (2018) has gone beyond translations, and constructed rotation-equivariant or more general group-equivariant neural network models. In this paper, we do an extensive empirical study of various rotation-equivariant neural network models to understand how effectively they learn rotations. This includes Group-equivariant Convolutional Networks (GCNNs) by Cohen & Welling (2016), Harmonic Networks (H-Nets) by Worrall et al. (2016), Polar Transformer Networks (PTN) by Esteves et al. (2018) and Rotation equivariant vector field networks by Marcos et al. (2017). We empirically compare the ability of these networks to learn rotations efficiently in terms of their number of parameters, sample complexity, rotation augmentation used in training. We compare them against each other as well as Standard CNNs. We observe that as these rotation-equivariant neural networks learn rotations, they instead become more vulnerable to small pixel-wise adversarial attacks, e.g., Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), in comparison with Standard CNNs. In other words, robustness to geometric transformations in these models comes at the cost of robustness to small pixel-wise perturbations.

1 INTRODUCTION

Neural network-based models achieve state of the art results on several speech and visual recognition tasks but these models are known to be vulnerable to various adversarial attacks. Szegedy et al. (2013) show that small, pixel-wise changes that are almost imperceptible to the human eye can make neural networks models grossly misclassify. They find a small perturbation so as to maximizes the prediction error of a given model using box-constrained L-BFGS. Goodfellow et al. (2015) propose the Fast Gradient Sign Method (FGSM) as a faster approach to find such an adversarial perturbation given by $x' = x + \epsilon \text{ sign}\left(\nabla_x J(\theta, x, y)\right)$, where x is the input, y represents the targets, θ represents the model parameters, and $J(\theta, x, y)$ is the cost used to train the network.

Subsequent work has introduced multi-step variants of FGSM, notably, an iterative method by Kurakin et al. (2017) and Projected Gradient Descent (PGD) by Madry et al. (2018). On visual tasks, the adversarial perturbation must come from a set of images that are perceptually similar to a given image. Goodfellow et al. (2015) and Madry et al. (2018) study adversarial perturbations from the ℓ_∞-ball around the input x, namely, each pixel value is perturbed by a quantity within $[-\epsilon, +\epsilon]$. Broadly, all the above-mentioned adversarial attacks are model-dependent. Tramer et al. (2017) also mention model-agnostic perturbations using the direction of the difference between the intra-class means.

There is a large class of spatial transformations including translations, rotations, scaling that preserve perceptual similarity. Convolutional Neural Networks (CNNs) are translation-equivariant by construction. However, Engstrom et al. (2017) show that simple adversarial attacks using rotations and translations can fool CNNs, even when they are adversarially trained to make them robust to

ℓ_p-bounded adversaries. They observe that ℓ_p-bounded and spatial adversarial perturbations have additive or super-additive effect on the performance drop, suggesting that these two types of attacks have no bearing on each other. Engstrom et al. (2017) also show that CNNs achieve translation invariance only if the training data (or augmentation) contains some amount of translated inputs, however, their accuracy against the worst-case translations is significantly worse than the average-case.

CNNs are translation-equivariant but not equivariant with respect to other spatial symmetries such as rotations, reflections etc. Variants of CNNs to achieve rotation-equivariance and other symmetries have received much attention recently, notably, Harmonic Networks (H-Nets) by Worrall et al. (2016), cyclic slicing and pooling by Dieleman et al. (2016), Tranformation-Invariant Pooling (TI-Pooling) by Laptev et al. (2016), Group-equivariant Convolutional Neural Networks (GCNNs) by Cohen & Welling (2016), Steerable CNNs by Cohen & Welling (2017), Deep Rotation Equivariant Networks (DREN) by Li et al. (2017), Rotation Equivariant Vector Field Networks (RotEqNet) by Marcos et al. (2017), Polar Transformer Networks (PTN) by Esteves et al. (2018).

For our study, we choose GCNNs as they achieve close to the current state of the art results on MNIST-rot[1] and CIFAR10 data sets as reported in Esteves et al. (2018). GCNNs provide good representative networks to understand the effect of ℓ_p-bounded and spatial transformation adversaries on symmetric networks. GCNNs use G-convolutions, they have more weight-sharing than regular convolution layers, and they are easy to implement with minimal computational overhead for discrete groups of symmetry generated by translations, reflections, and rotations. We do show similar qualitative trends on Harmonic Networks (H-Nets), Polar Transformer Networks(PTN) and Rotation Equivariant Vector Field Networks (RotEqNet).

2 Robustness of rotation-equivariant networks

We study the robustness of GCNNs to adversarial attacks based on rotations as well as pixel-wise perturbations on MNIST data set and compare it with Standard CNNs(StdCNNs). To the best of our knowledge this is the first study of a rotation-equivariant network towards pixel-wise perturbations. There are other types of networks like CapsNetSabour et al. (2017) which do show natural robustness to AffNist dataset[2] though not trained on it. And further derived networks based on EM Routing Hinton et al. (2018) which along with better spatial robustness also seem to be robust to FGSM attacks. We have also checked the robustness of CapsNet to small rotations.

2.1 Robustness to rotations

We study the robustness of equivariant networks to attacks based on rotations on MNIST and compare it with StdCNNs. The main takeaways of our empirical results are (a) Rotation equivariant networks are robust to small degrees of rotations away from the ones present in the training data, (b) applying data augmentation increases their robustness, (c) Rotation equivariant networks do achieve state of the art results with smaller sample size for training.

We first trained all the networks on MNIST with no augmentation and tested against inputs augmented with varying range of rotations. In Figure (1)(left) we observe the inherent robustness of all the equivariant networks to small(below 40°) rotations. Their accuracy is always greater than StdCNNs with accuracy well above 98% for rotations upto 20°. We also did the same experiment for CapsNet(see Figure 12) and found their values to be sandwiched between GCNNs and HNets. Figure (1)(right) shows the performance of these networks to the entire range of rotations. Here we observe PTN and RotEqNet to be more robust than the other networks and their accuracy remains safely above 60%.

2.1.1 Robustness with training augmentation

We trained the equivariant networks with input augmented with varying range of rotations. We observe that their accuracy for the respective range of rotation augmentation remains above 98%,

[1] http://www.iro.umontreal.ca/ lisa/twiki/bin/view.cgi/Public/MnistVariations

[2] https://www.cs.toronto.edu/ tijmen/affNIST/

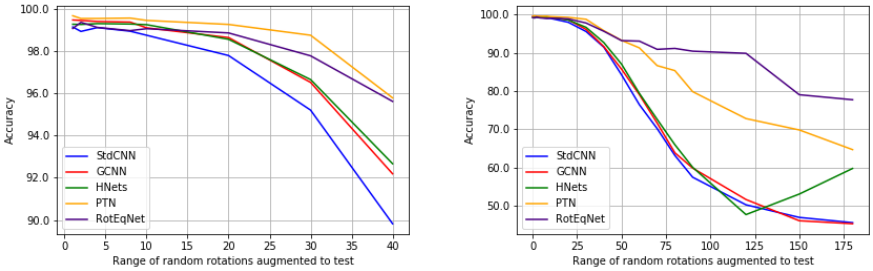

Figure 1: On MNIST, networks trained with no augmentation, test augmented with random rotations in $[-x^\circ, x^\circ]$ range. (left) Rotations(small) upto 40° (right) Rotations from 0° to 180°

even when the test data is augmented with the same range. However, in Figure 2 we only show results for StdCNNs and GCNNs.

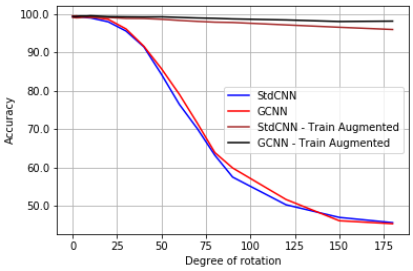

Figure 2: Networks trained with and without augmentation to MNIST, random rotation augmentations in $[-x^\circ, x^\circ]$ range.

2.2 Sample Complexity of Networks

To understand the sample complexity of the networks, we perform two experiments. In the first we train the networks with varying sample sizes of MNIST training set and test them on the entire MNIST test set. And in the second experiment we do the same as the first with the inputs in train and test augmented with $[-180^\circ, 180^\circ]$ range of rotations. From Figure 3, we can see that rotation equivariant networks achieve their best performance safely using $10k$ - $30k$ training samples. This confirms that rotation equivariant networks exploit symmetry and reduce training sample size.

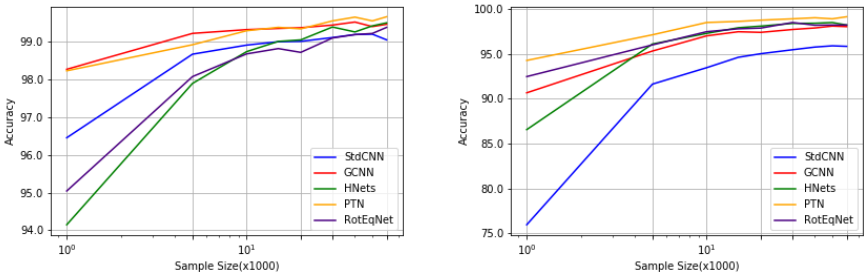

Figure 3: Networks trained with varying training sample size on X-axis. (left) Only MNIST, (right) MNIST train and test augmented with random rotations in $[-180^\circ, 180^\circ]$ range.

2.3 ROBUSTNESS TO PIXEL-WISE PERTURBATIONS

From Figure 2 we observe that with augmentation the networks become robust to rotations. With networks made robust to rotations with respective augmentation we now study the vulnerability of StdCNNs and GCNNs to pixel-wise perturbations.

We first motivate the comparison of the two networks by considering pixel-wise attack on architectures as give in Table 1. We observe that as GCNNs become more robust to rotations they are more vulnerable to pixel-wise attacks, more so than StdCNNs. For a fairer comparison we make them roughly equivalent on parameter count by doubling the number of filters in StdCNNs and reducing the number of filters in GCNNs. Even with this change the relative behaviour of GCNNs to StdCNNs remains the same. In parallel we also show what happens when the networks are adversarially trained and tested with/without adversarial perturbations.
We first give complete details for the networks attacked with FGSM and follow it up with PGD. For a finer analysis of their vulnerability we plot their accuracy drop as a function of ϵ budget used in the attack for each network made robust to a specific range of rotations.

FGSM attack

We augment both the train and test with random rotations from the same fixed range to handle spatial perturbations, and see that GCNNs not only outperform StdCNNs but also maintain accuracy above 98% even for rotations in the larger range of 180°.
With this setup, we check the vulnerability of the networks to FGSM perturbed inputs, with $\epsilon = 0.3$. Figure 4 clearly demonstrates that as GCNNs become more robust to larger rotations they become more vulnerable to FGSM attack. However, for small rotations GCNNs are definitely more robust to FGSM.

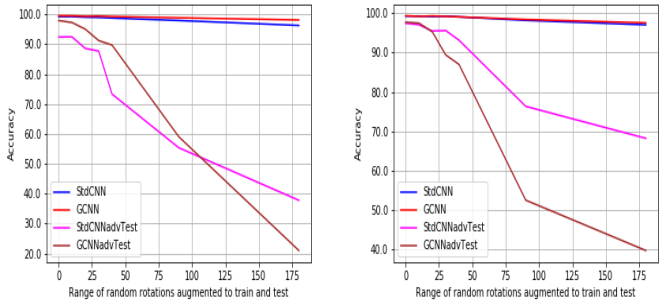

Figure 4: On MNIST, without FGSM training and test FGSM perturbed (left) Networks as given in Table1, (right) Networks as given in Table1 but changed to make them parameter equivalent.

Next, we check their performance with FGSM adversarial training and unperturbed test. From Figure 5(left), as expected, GCNNs outperform StdCNNs in every count. However, when their parameter counts are roughly the same, we see in Figure 5(right) that StdCNNs perform as well as GCNNs.

Finally, we check the performance of these networks with both train and test data FGSM perturbed. Here also we also we see in Figure 6(left) that GCNNs perform better than StdCNNs. Once again, when their parameter counts are roughly the same, StdCNNs perform as well as GCNNs as shown in Figure 6(right).

PGD attack

Similar to the FGSM attack studied above, we now check and compare the vulnerability of these networks to the stronger PGD attack(with $\epsilon = 0.3$) once they are made robust to rotation with augmentation.
In the plots Figure 7(left) and 7(center) we compare the performance of the individual networks towards FGSM and PGD. And, as expected, we observe that in both the networks PGD is a stronger

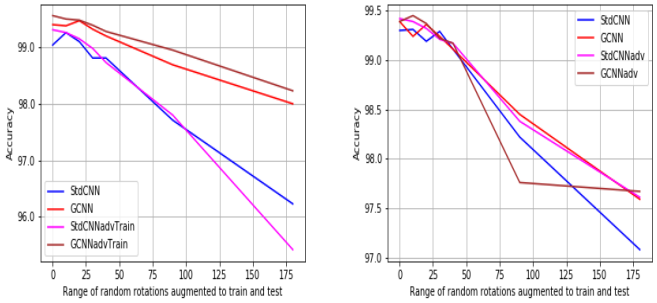

Figure 5: On MNIST, with FGSM adversarial training only (left) Networks as given in Table1, (right) Networks as given in Table1 but changed to make them parameter equivalent.

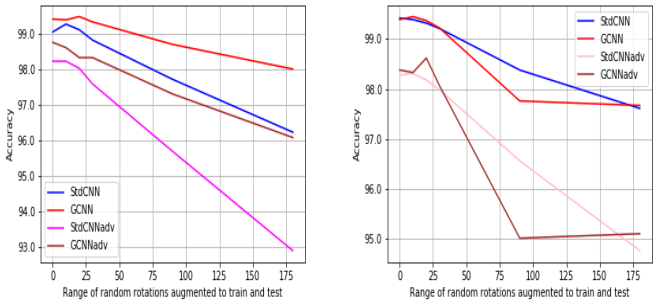

Figure 6: On MNIST, with both train and test FGSM perturbed (left) Networks as given in Table1, (right) Networks as given in Table1 but changed to make them parameter equivalent.

attack than FGSM. From Figure 7(right) we see that as in the case of FGSM attack, GCNNs are more vulnerable to PGD attack than StdCNNs as they become more robust to larger rotations. For smaller rotations GCNNs are still more robust to PGD than StdCNNs.
This trend is similar when GCNNs and StdCNNs are made roughly parameter count equivalent. These observations are given in Figure 8.
In Figure 9, we also have checked the vulnerability of RotEqNet to pixel-wise perturbations like FGSM and PGD. There is no clear trend as in GCNNs. RotEqNet seem to be severely affected by FGSM even at no rotation augmentation. Their accuracy changes roughly from 40% to 60% as we augment the network from $0°$ to $180°$.

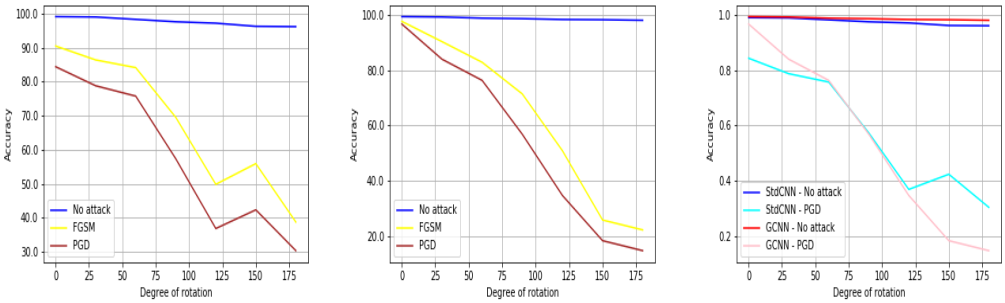

Figure 7: For Networks give in table 1, comparison between Test with No attack, FGSM and PGD attack, $\epsilon = 0.3$ (left) StdCNNs, (center) GCNNs, (right) StdCNNs vs GCNNs - PGD attack.

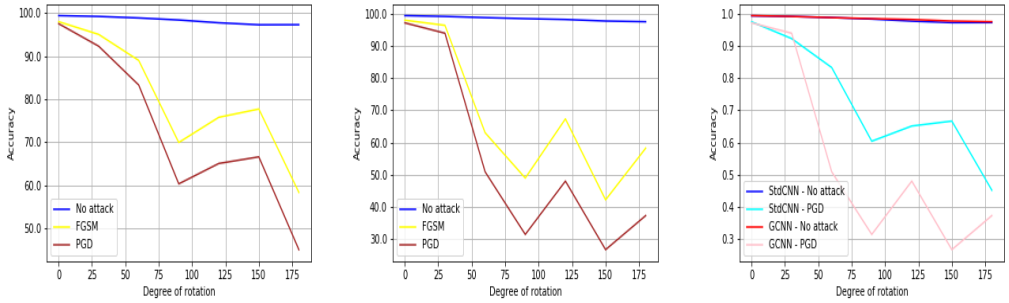

Figure 8: For Networks similar to table 1 but with double filters in each convolution layer for StdCNNs and half the filters for GCNNs, comparison between Test with No attack, FGSM and PGD attack, $\epsilon = 0.3$ (left) StdCNNs, (center) GCNNs, (right) StdCNNs vs GCNNs - PGD attack.

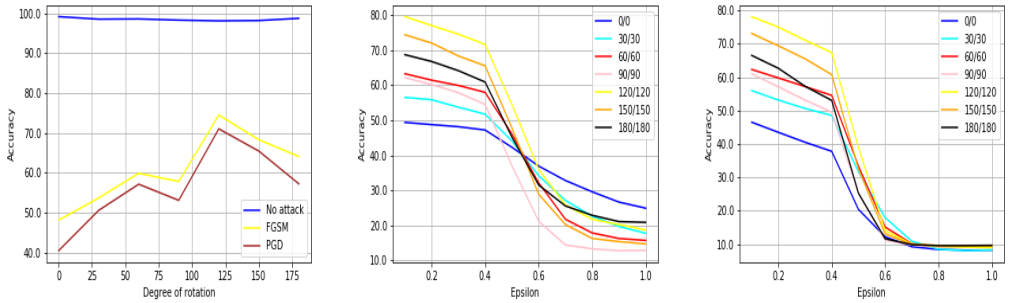

Figure 9: RotEqNet, MNIST with train and test augmented with $[-x^\circ, x^\circ]$ range, (left) Adversarial test with FGSM and PGD, $\epsilon = 0.3$, (center) Varying epsilon ball for adversarial test with FGSM, (right) Varying epsilon ball for adversarial test with PGD

Accuracy drop with varying ϵ

It's clear from the above experiments that as GCNNs become more robust to larger rotations, they become more vulnerable to pixel-wise attacks, in comparison to StdCNNs. We do a finer analysis of the attacks with varying ϵ values, ϵ being the maximum perturbation allowed for the attack. Plots in Figure 10 are for FGSM and PGD attack on StdCNNs with changing ϵ and plots in Figure 11 are for FGSM and PGD attack on GCNNs with changing ϵ. The labels a/a in the legend of the plots denote the $[-a^\circ, a^\circ]$ range of rotations augmented to train/test, respectively. We observe that even for ϵ as small as 0.1 the networks exhibit a behaviour similar to that seen from above experiments. *As GCNNs become robust to larger rotations they become more vulnerable to pixel-wise attacks even for smaller epsilon.*

3 Details of experiments

All experiments performed on neural network-based models were done using MNIST dataset with appropriate augmentations applied to the train/validation/test set.

Data sets MNIST[3] dataset consists of $70,000$ images of 28×28 size, divided into 10 classes. $55,000$ used for training, $5,000$ for validation and $10,000$ for testing.

Model Architectures For the MNIST based experiments we use the 7 layer architecture of GCNN similar to Cohen & Welling (2016). The StdCNN architecture is similar to the GCNN except that

[3]http://www.iro.umontreal.ca/ lisa/twiki/bin/view.cgi/Public/MnistVariations

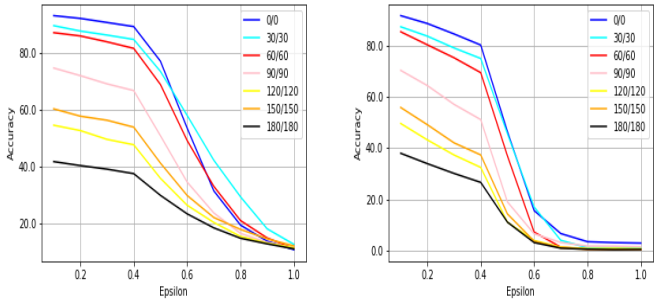

Figure 10: StdCNNs, MNIST with train and test augmented with $[-x^\circ, x^\circ]$ range and varying epsilon ball as perturbation budget on X-axis. (left) Adversarial test with FGSM, (right) Adversarial test with PGD.

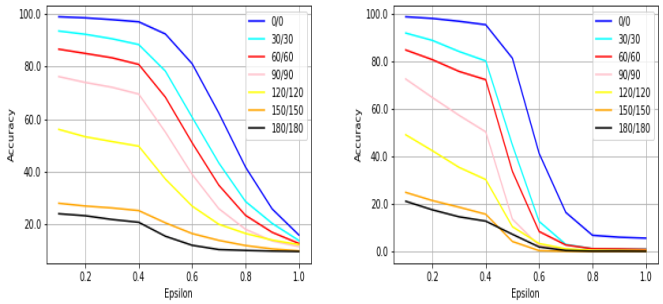

Figure 11: GCNNs, MNIST with train and test augmented with $[-x^\circ, x^\circ]$ range and varying epsilon ball for perturbation budget on X-axis. (left) Adversarial test with FGSM, (right) Adversarial test with PGD.

the operations are as per CNNs. Refer to Table 1 for details. RotEqNet architecture is as given in Marcos et al. (2017).

Table 1: Architectures used for experiments

Standard CNN	**GCNN**
Conv(10,3,3) + Relu	P4ConvZ2(10,3,3) + Relu
Conv(10,3,3) + Relu	P4ConvP4(10,3,3) + Relu
Max Pooling(2,2)	Group Spatial Max Pooling(2,2)
Conv(20,3,3) + Relu	P4ConvP4(20,3,3) + Relu
Conv(20,3,3) + Relu	P4ConvP4(20,3,3) + Relu
Max Pooling(2,2)	Group Spatial Max Pooling(2,2)
FC(50) + Relu	FC(50) + Relu
Dropout(0.5)	Dropout(0.5)
FC(10) + Softmax	FC(10) + Softmax

4 Conclusion

We observe that the robustness to geometric transformations in equivariant networks comes at the cost of their robustness to pixel-wise adversarial perturbations. We do an extensive comparative study of various equivariant network models ranging from StdCNNs to GCNNs, HNets, PTNs,

RotEqNets. We believe that good neural network models should be robust to both geometric transformations and pixel-wise adversarial perturbations, and understanding trade-offs similar to the ones in our paper is an important direction for future work.

REFERENCES

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *In Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. *In International Conference on Learning Representations*, 2017.

Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *In Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *In International Conference on Learning Representations*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *In International Conference on Learning Representations*, 2015.

Geoffrey Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. *In International Conference on Learning Representations*, 2018.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *In Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2017.

Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys. TI-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 289–297, 2016.

Junying Li, Zichen Yang, Haifeng Liu, and Deng Cai. Deep rotation equivariant network. *arXiv preprint arXiv:1705.08623*, 2017.

Aleksander Madry, Aleksandar A Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *In International Conference on Learning Representations*, 2018.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. *In International Conference on Computer Vision*, 2017.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *CoRR*, abs/1710.09829, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramer, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. *arXiv preprint arXiv:1612.04642*, 2016.

A APPENDIX

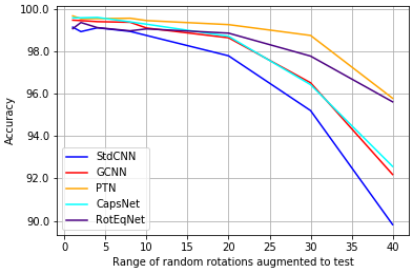

Figure 12: On MNIST, (left) networks trained with no augmentation, test augmented with random rotations in $[-x^\circ, x^\circ]$ range.