UNIVERSAL ATTACKS ON EQUIVARIANT NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial attacks on neural networks perturb the input at test time in order to fool trained and deployed neural network models. Most attacks such as gradient-based Fast Gradient Sign Method (FGSM) by Goodfellow et al. (2015) and DeepFool by Moosavi-Dezfooli et al. (2016) are input-dependent, small, pixel-wise perturbations, and they give different attack directions for different inputs. On the other hand, universal adversarial attacks are input-agnostic and the same attack works for most inputs. Translation or rotation-equivariant neural network models provide one approach to prevent universal attacks based on simple geometric transformations. In this paper, we observe an interesting spectral property shared by all of the above input-dependent, pixel-wise adversarial attacks on translation and rotation-equivariant networks. We exploit this property to get a single universal attack direction that fools the model on most inputs. Moreover, we show how to compute this universal attack direction using principal components of the existing input-dependent attacks on a very small sample of test inputs. We complement our empirical results by a theoretical justification, using matrix concentration inequalities and spectral perturbation bounds. We also empirically observe that the top few principal adversarial attack directions are nearly orthogonal to the top few principal invariant directions.

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 try to maximize the prediction error of a given model by finding a small pixel-wise perturbation 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 denotes the input, y denotes the class label, θ denotes the model parameters, and $J(\theta, x, y)$ denotes the loss function used to train the model.

In subsequent work on FGSM, its iterative variant called Projected Gradient Descent (PGD) was introduced by Kurakin et al. (2017); also see Madry et al. (2018). 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 $[-\epsilon, +\epsilon]$. DeepFool by Moosavi-Dezfooli et al. (2016) gives a method to compute the minimal ℓ_2-norm perturbation. Broadly, all the above-mentioned adversarial attacks give directions to perturb any given test input in a way that is model-dependent and input-dependent. Tramer et al. (2017) also mention model-agnostic perturbations using the direction of the difference between the intra-class means, in an attempt to understand how adversarial attacks transfer across different models.

Universal adversarial attacks are input-agnostic so that the same perturbation or input transformation fools the trained model on most or nearly all test inputs. Recent work by Moosavi-Dezfooli et al. (2016; 2017) on universal adversarial attacks looks at the curvature of the decision boundary. Their universal attacks or perturbations are more sophisticated than the simple and fast, gradient-based adversarial perturbations, and require significantly more computation. Moosavi-Dezfooli et al. (2017) give a theoretical analysis of existence of universal adversarial attacks using directions in which the decision boundary has positive curvature.

Geometric transformations such as rotations, translations are simple input transformations that are model-agnostic and input-agnostic. Geometric transformations of input at test time are also possible natural attacks; see Gilmer et al. (2018). One way to counter such attacks is to use neural network models that are translation and rotation-equivariant by construction. Standard Convolutional Neural Networks (StdCNNs) 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). Among these, GCNNs are based on steerable filters, have a solid theoretical justification Kondor & Trivedi (2018), and achieve nearly state of the art results on MNIST-rot[1] and CIFAR10 as reported in Esteves et al. (2018). However, both StdCNNs and GCNNs do require rotation augmentation at training time to be robust to rotations.

The above discussion raises an important question: are there any interesting properties shared by translation and rotation-equivariant neural network models that can be exploited to come up with simple universal attacks? Can one find a universal attack direction using only a small fraction of test inputs but that fools the trained model on most test inputs? We answer both of these questions affirmatively in this paper.

2 Summary of our results

In this paper, we mostly study StdCNNs and GCNNs, which are translation-equivariant and rotation-equivariant, respectively. Some of our experiments also include RotEqNets. Our first observation is that even though the adversarial attack directions based on gradients, FGSM and DeepFool for StdCNN and GCNN models are input-dependent, overall they have only a small number of dominant principal components. For example, the top 5 principal components of the gradient directions on all test inputs contain more than 10% of the total spectrum for StdCNNs, GCNNs as well as RotEqNet models trained on MNIST. Our second observation is that a small pixel-wise perturbation in the direction of the top principal component alone can be used as a universal attack to fool these models on more than 80-90% test inputs. Our third observation is that the top principal component can be well-approximated using only 1% sample of FGSM or DeepFool attack directions on the test data, and this approximation can still give a smal pixel-wise perturbation to fool the model on more than 80% of test inputs. We give a theoretical justification of this phenomenon using matrix concentration inequalities and spectral perturbation bounds. An interesting aspect of our empirical study is that these observations hold across multiple input-dependent attacks (gradient, FGSM, DeepFool) and the choice of equivariant neural network models (StdCNNs, GCNNs).

3 Principal components of attack and invariant directions

We take a StdCNN or GCNN network as given in Table 2 and train it with 50,000 MNIST training data augmented with random rotations in the range $[-180°, +180°]$. We then use the gradient (similarly FGSM, DeepFool) directions obtained for this trained model on the 10,000 MNIST test inputs (which are also augmented with random rotations in the range $[-180°, +180°]$). We form a matrix whose rows are unit vectors along these 10,000 gradient (similarly FGSM, DeepFool) directions and obtain is Singular Value Decomposition (SVD). For gradients, we observe, as in Figure (1), that the top 5 or less than 1% in number of the squared singular values of this matrix add up to more than 10% of the overall sum of squared singular values. This is irrespective of whether the trained model is StdCNN, GCNN or RotEqNet. Figures (2a), (3a), and (4a) show how the sigular values drop when we consider gradient, FGSM, DeepFool directions for StdCNNs, GCNNs, RotEqNets, respectively. These indicate that the drop in singular values is a common phenomenon for the different attacks and different translation and rotation-equivariant neural network models we consider.

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

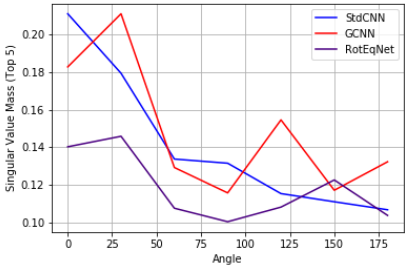


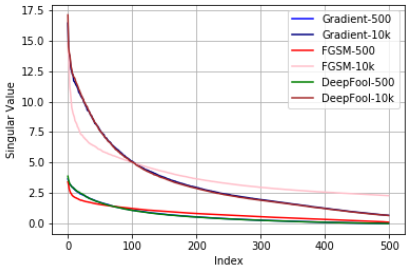
Figure 1: Fraction of the total spectrum contained in top 5 squared singular values of gradient directions for 10,000 MNIST test inputs, when train and test augmented with random rotations in $[-180°, 180°]$ for StdCNN, GCNN and RotEqNet, respectively.

Table 1: Subspace angles between Top 5 SVD vectors from gradients of test points for each network and invariant directions (small $2°$ rotations) respectively.

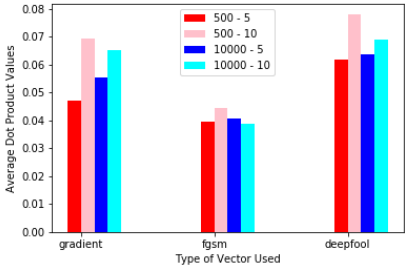
Model	1	2	3	4	5
StdCNN	89.82170934	88.06580815	84.76831292	81.22444162	78.85267548
GCNN	89.06822425	88.27986252	87.45134508	84.26710498	68.77651938
RotEqNet	89.75561818	88.88988678	85.60567329	82.28177397	68.96832307

We next consider the principal components of invariant directions. For this, the MNIST test data is augmented with random rotations in $[-180°, +180°]$, and we look at the difference vectors between such images and their small $2°$ rotations. For StdCNN, GCNN or RotEqNet trained with rotation augmentation, we expect the difference between a small rotation of an image and the image itself to be along the invariant subspace tangent to the level sets of the loss function. We take the principal components of these difference or *invariant* directions, and observe that the top 5 singular vectors of the adversarial directions are nearly orthogonal to the top 5 singular vectors of the invariant directions, i.e., their average dot product is less than 0.07. Figures (2b), (3b), (4b) indicate that this property is shared by StdCNNs, GCNNs and RotEqNets, and also holds when we look at a smaller sample of only 500 test points instead of all 10,000 test points. Table 1 shows deeper analysis using principal angles between subspaces that the two 5-dimensional SVD subspaces of adversarial directions and invariant directions, respectively, have nearly $90°$ principal angles.

The intensity maps of top singular vectors for the gradient, FGSM, DeepFool directions for StdCNNs, GCNNs and RotEqNets also have interesting structure as seen in Figures (8)-(16). As we have already observed, these are nearly orthogonal to the invariant directions which lead to steer-

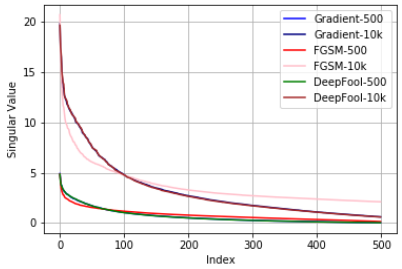


(a) Singular values of attack directions over a sample of 500 and 10,000 test points

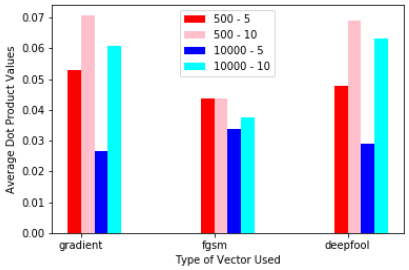


(b) Avg. dot product of top 5, top 10 singular vectors of adversarial and invariant directions, respectively, for a sample of 500 and 10,000 test points

Figure 2: Principal components of adversarial and invariant directions for StdCNN

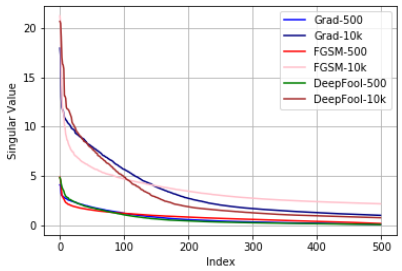


(a) Singular values of attack directions over a sample of 500 and 10,000 test points

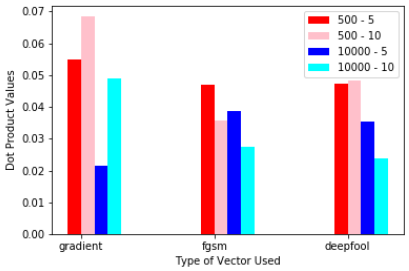


(b) Avg. dot product of top 5, top 10 singular vectors of adversarial and invariant directions, respectively, for a sample of 500 and 10,000 test points

Figure 3: Principal components of adversarial and invariant directions for GCNN



(a) Singular values of attack directions over a sample of 500 and 10,000 test points



(b) Avg. dot product of top 5, top 10 singular vectors of adversarial and invariant directions, respectively, for a sample of 500 and 10,000 test points

Figure 4: Principal components of adversarial and invariant directions for RotEqNet

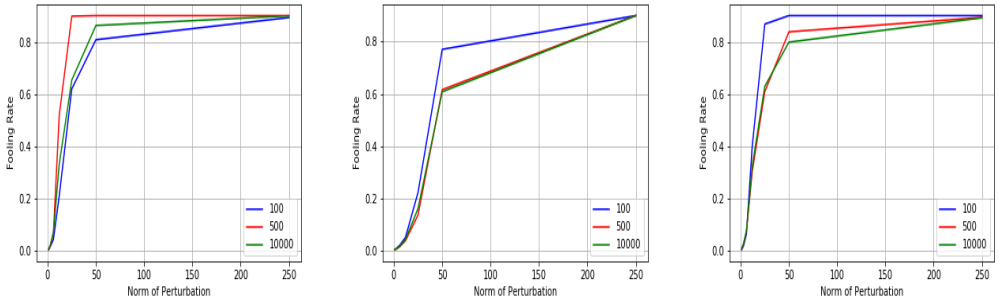


Figure 5: StdCNN: fooling rate vs. norm of perturbation along top singular vector of attack directions on 100/500/10000 sample, (left) gradients (center) FGSM (right) DeepFool

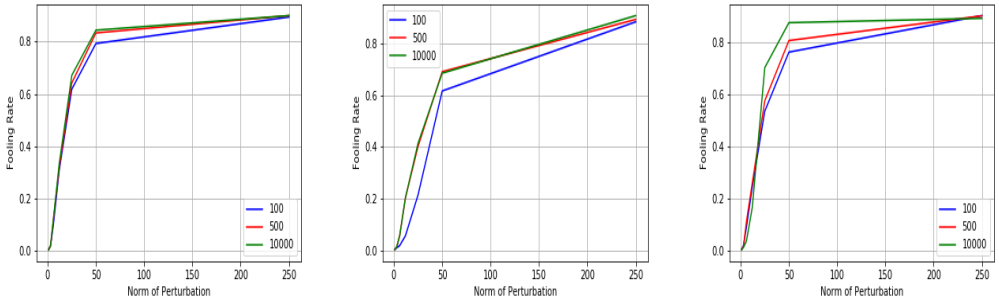


Figure 6: GCNN: fooling rate vs. norm of perturbation along top singular vector of attack directions on 100/500/10000 samples, (left) gradients (center) FGSM (right) DeepFool

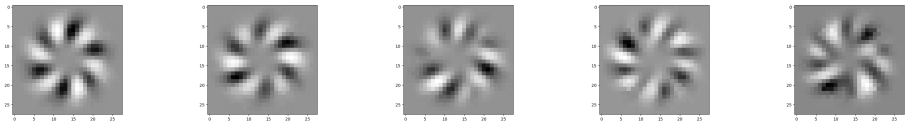


Figure 7: Top 5 SVD vectors from image difference

able filters similar to Figure (7). We believe this underlying structure is useful and of independent interest.

4 Universal adversarial attack using a small sample

We now plot the fooling rate or the fraction of test inputs that the trained model missclassifies, when we perturb all of them using the top singular vector of gradient direction scaled by ϵ that denotes the norm of perturbation. Figures (5) and (6) plot these fooling rate to show that we can fool 80-90% of test inputs using the same small-norm universal adversarial attack. Moreover, if we pick a small sample of 100 random test points (about 1% of test data) and take their gradients, their top singular vector also gives a good universal adversarial attack direction with a comparable fooling rate, namely, simultaneously fooling the model on roughly 80% of test points using the same small-norm perturbation.

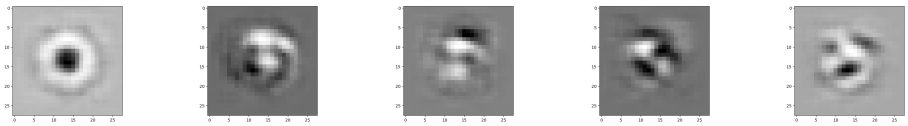


Figure 8: Top 5 SVD vectors from Gradients in StdCNN with rotation augmentations

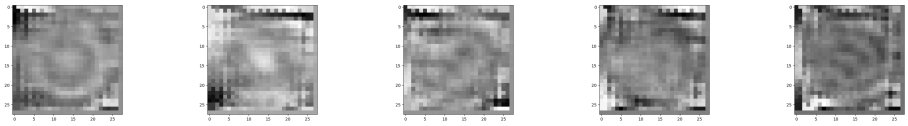


Figure 9: Top 5 SVD vectors from FGSM in StdCNN with rotation augmentations

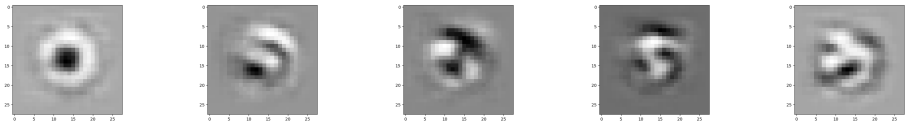


Figure 10: Top 5 SVD vectors from DeepFool in StdCNN with rotation augmentations

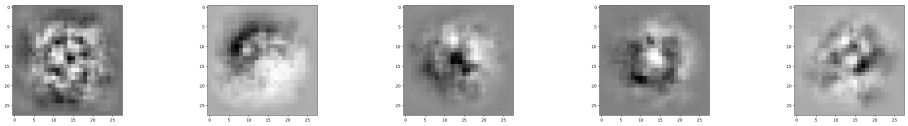


Figure 11: Top 5 SVD vectors from Gradients in GCNN with rotation augmentations

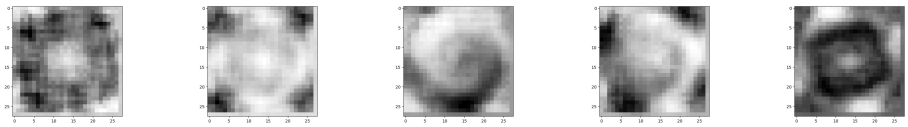


Figure 12: Top 5 SVD vectors from FGSM in GCNN with rotation augmentations

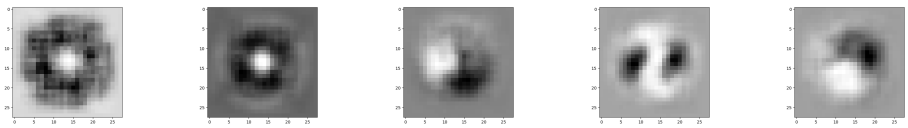


Figure 13: Top 5 SVD vectors from gradients in DeepFool on GCNN

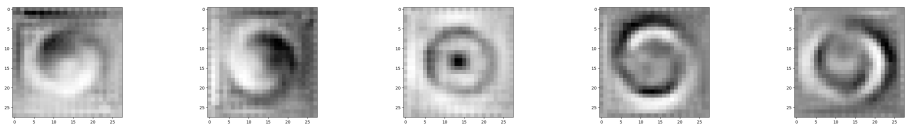


Figure 14: Top 5 SVD vectors from Gradients in RotEqNet with rotation augmentations

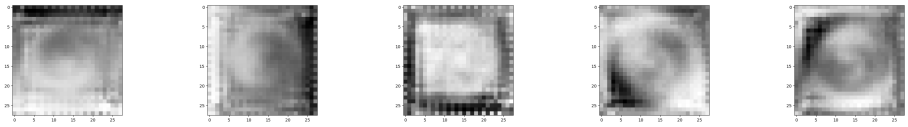


Figure 15: Top 5 SVD vectors from FGSM in RotEqNet with rotation augmentations

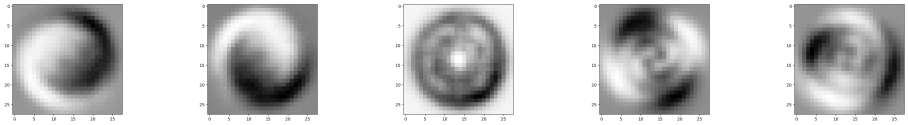


Figure 16: Top 5 SVD vectors from DeepFool in RotEqNet with rotation augmentations

5 ANALYSIS OF UNIVERSAL ADVERSARIAL PERTURBATIONS

In this section, we attempt to provide a theoretical justification for the existence of universal adversarial perturbations. Assume that our data is in d dimensions $X \subseteq \mathbb{R}^d$ with an underlying distribution given by probability density $\mu(\mathbf{x})$ at each point $\mathbf{x} \in X$. Let $\mathbf{a_x}$ be an adversarial perturbation direction for each $\mathbf{x}$, respectively. Note that $\mathbf{a_x}$ could be obtained by any of the several available methods such as the Fast Gradient Sign Method (FGSM), the Kurakin attack, DeepFool or by finding the nearest point on the decision boundary etc. Two important questions are: (a) Why does there exist a *universal* vector or direction that works as an adversarial perturbation for many or most data points simultaneously? (b) Can we compute such a *universal* adversarial perturbation by looking only at a small subsample of the data?

Given any method that obtains adversarial perturbations for individual data points, if the matrix of these adversarial perturbation directions taken over all data points satisfies a certain property, then we show that the top singular vectors of this matrix are good candidates for *universal* adversarial perturbations that make many points to be misclassified simultaneously.

Theorem 1. *Let $X \subseteq \mathbb{R}^d$ be any given input data with the underlying probability density $\mu(\mathbf{x})$ at $\mathbf{x} \in X$. Let $\mathbf{a_x}$ denote an adversarial perturbation vector at point $\mathbf{x} \in X$. Let $0 \leq \lambda \leq 1$ be the top eigenvalue of the matrix $M \in \mathbb{R}^{d \times d}$ defined as*

$$M = \mathsf{E}\left[\frac{\mathbf{a_x}}{\|\mathbf{a_x}\|_2}\frac{\mathbf{a_x}^T}{\|\mathbf{a_x}\|_2}\right],$$

and let $\mathbf{v}$ be its corresponding eigenvector. Then

$$\mathsf{Pr}\left(\{\mathbf{x} \ : \ |\langle \mathbf{a_x}, \mathbf{v}\rangle| \geq \delta \ \|\mathbf{a_x}\|_2\}\right) \geq \frac{\lambda - \delta^2}{1 - \delta^2}.$$

In particular, plugging in $\delta = \sqrt{\lambda/2}$ we get

$$\mathsf{Pr}\left(\left\{\mathbf{x} \ : \ |\langle \mathbf{a_x}, \mathbf{v}\rangle| \geq \sqrt{\frac{\lambda}{2}} \ \|\mathbf{a_x}\|_2\right\}\right) \geq \frac{\lambda}{2}.$$

Proof. Define $S = \{\mathbf{x} \ : \ |\langle \mathbf{a}_{\mathbf{x}}, \mathbf{v}\rangle| \geq \delta \ \|\mathbf{a}_{\mathbf{x}}\|_2\} \subseteq X$. Since λ is the top eigenvalue of M with $\mathbf{v}$ as its corresponding eigenvector,

$$\begin{aligned}
\lambda &= \int_X \left\langle \frac{\mathbf{a}_{\mathbf{x}}}{\|\mathbf{a}_{\mathbf{x}}\|_2}, \mathbf{v} \right\rangle^2 \mu(\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathbf{x}\in S} \left\langle \frac{\mathbf{a}_{\mathbf{x}}}{\|\mathbf{a}_{\mathbf{x}}\|_2}, \mathbf{v} \right\rangle^2 \mu(\mathbf{x}) d\mathbf{x} + \int_{\mathbf{x}\notin S} \left\langle \frac{\mathbf{a}_{\mathbf{x}}}{\|\mathbf{a}_{\mathbf{x}}\|_2}, \mathbf{v} \right\rangle^2 \mu(\mathbf{x}) d\mathbf{x} \\
&\leq \int_{\mathbf{x}\in S} \mu(\mathbf{x}) d\mathbf{x} + \delta^2 \int_{\mathbf{x}\notin S} \mu(\mathbf{x}) d\mathbf{x} \qquad \text{because } \|\mathbf{v}\|_2 = 1 \\
&= \mathsf{Pr}\,(S) + \delta^2 \left(1 - \mathsf{Pr}\,(S)\right) \\
&= (1-\delta^2)\, \mathsf{Pr}\,(S) + \delta^2.
\end{aligned}$$

This implies that $\mathsf{Pr}\,(S) \geq (\lambda - \delta^2)/(1-\delta^2)$. □

Observe that $\mathrm{tr}\,(M) = 1$. Theorem 1 implies that as the top eigenvalue λ dominates the spectrum, its eigenvector $\mathbf{v}$ get more aligned with the adversarial perturbations directions $\mathbf{a}_{\mathbf{x}}/\|\mathbf{a}_{\mathbf{x}}\|_2$ for most points $\mathbf{x} \in X$. This means that $\mathbf{v}$ is a potential candidate for a *universal* adversarial perturbation. More generally, Theorem 1 works for any of the top eigenvalues of M and their corresponding eigenvectors, and thus, gives a subspace spanned by multiple orthogonal directions all of which are potential candidates for being *universal* adversarial perturbations that fool a given classifier on many input examples.

Now let's consider the second question of finding a good approximation to our candidate *universal* adversarial perturbation or the top eigenvector $\mathbf{v}$ of M using only a small subsample of X. Theorem 2 shows that only a small sample of size independent of $|X|$ but depending only on the dimension d of the data and the spectral properties of the matrix M suffices.

Theorem 2. *Let $X \subseteq \mathbb{R}^d$ be any given input data with the underlying probability density $\mu(\mathbf{x})$ at $\mathbf{x} \in X$. Let $\mathbf{a}_{\mathbf{x}}$ denote an adversarial perturbation vector at point $\mathbf{x} \in X$. Let $0 \leq \lambda \leq 1$ be the top eigenvalue of the matrix $M \in \mathbb{R}^{d\times d}$ defined as*

$$M = \mathsf{E}\left[\frac{\mathbf{a}_{\mathbf{x}}}{\|\mathbf{a}_{\mathbf{x}}\|_2} \frac{\mathbf{a}_{\mathbf{x}}^T}{\|\mathbf{a}_{\mathbf{x}}\|_2}\right],$$

and let $\mathbf{v}$ be its corresponding eigenvector. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be a i.i.d. sample of m points from X drawn from the distribution defined by probability density μ, and let $\tilde{\lambda}$ be the top eigenvalue of the matrix $\tilde{M}$,

$$\tilde{M} = \frac{1}{m}\sum_{i=1}^{m} \frac{\mathbf{a}_{\mathbf{x}_i}}{\|\mathbf{a}_{\mathbf{x}_i}\|_2} \frac{\mathbf{a}_{\mathbf{x}_i}^T}{\|\mathbf{a}_{\mathbf{x}_i}\|_2},$$

and $\tilde{\mathbf{v}}$ be the top eigenvector of $\tilde{M}$.

Also suppose that there is a gap of at least $\gamma\lambda$ between the top eigenvalue λ and the second eigenvalue of M. Then for any $0 \leq \epsilon < \gamma$ and $m = O(\epsilon^{-2} d \log d)$, we get $\|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \epsilon/(\gamma - \epsilon)$, with a constant probability. This probability can be boosted to $1-\delta$ by having an additional $\log(1/\delta)$ in the $O(\cdot)$.

Proof. Using matrix Bernstein inequality from Tropp (2015), we get that for $m = O(\epsilon^{-2} d \log d)$, we have that the spectral or the operator norm $\left\|M - \tilde{M}\right\|_2 \leq \epsilon\lambda$, with a constant probability. Now we can use results about spectral perturbation bounds from Bhatia (1997). By Weyl's theorem, this implies $\left|\lambda - \tilde{\lambda}\right| \leq \epsilon\lambda$. If there is gap of at least $\gamma\lambda$ between the first and the second eigenvalue of M with $\gamma > \epsilon$, then by Davis-Kahan theorem we get that $\|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \epsilon/(\gamma - \epsilon)$, with a constant probability. □

6 Conclusion

We show how to use a small sample of input-dependent adversarial attack directions on test inputs to find a single universal attack directions that fools translation and rotation-equivariant models such

as CNNs and GCNNs on a large fraction of test inputs. Our main observation is a spectral property shared by different attacks directions such as gradients, FGSM, DeepFool on these models. We give a theoretical justification for how this spectral property helps in using the top singular vector as a universal attack direction computed using only a small sample of test inputs.

REFERENCES

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997. ISBN 0387948465.

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.

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

Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Robustness of rotation-equivariant networks to adversarial perturbations. *CoRR*, abs/1807.06732, 2018. URL `http://arxiv.org/abs/1807.06732`.

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

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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Analysis of universal adversarial perturbations. *CoRR*, abs/1705.09554, 2017. URL `http://arxiv.org/abs/1705.09554`.

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

Table 2: Architectures used for experiments

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.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL http://dx.doi.org/10.1561/2200000048.

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 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. The collection of gradients and other attack vectors and the testing of the universal perturbations were on the same network.

Data sets MNIST[2] 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 the operations are as per CNNs. Refer to Table 2 for details. RotEqNet architecture is are given in Marcos et al. (2017).

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