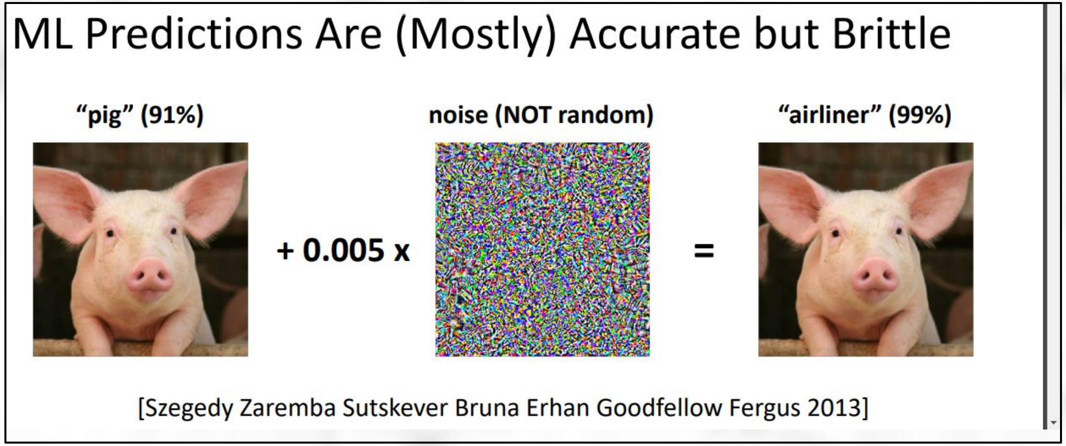
Universal attacks on equivariant networks

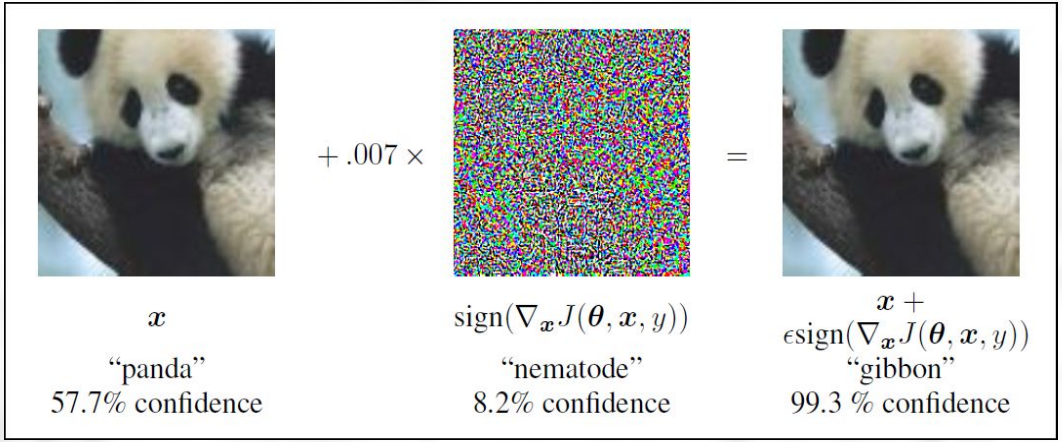
Anton Smerdov, Nurislam Tursynbek, Yuriy Biktairov

Adversarial attacks



- Tiny (imperceptible to human) perturbation of input can easily fool neural network.
- Recent successful attacks:

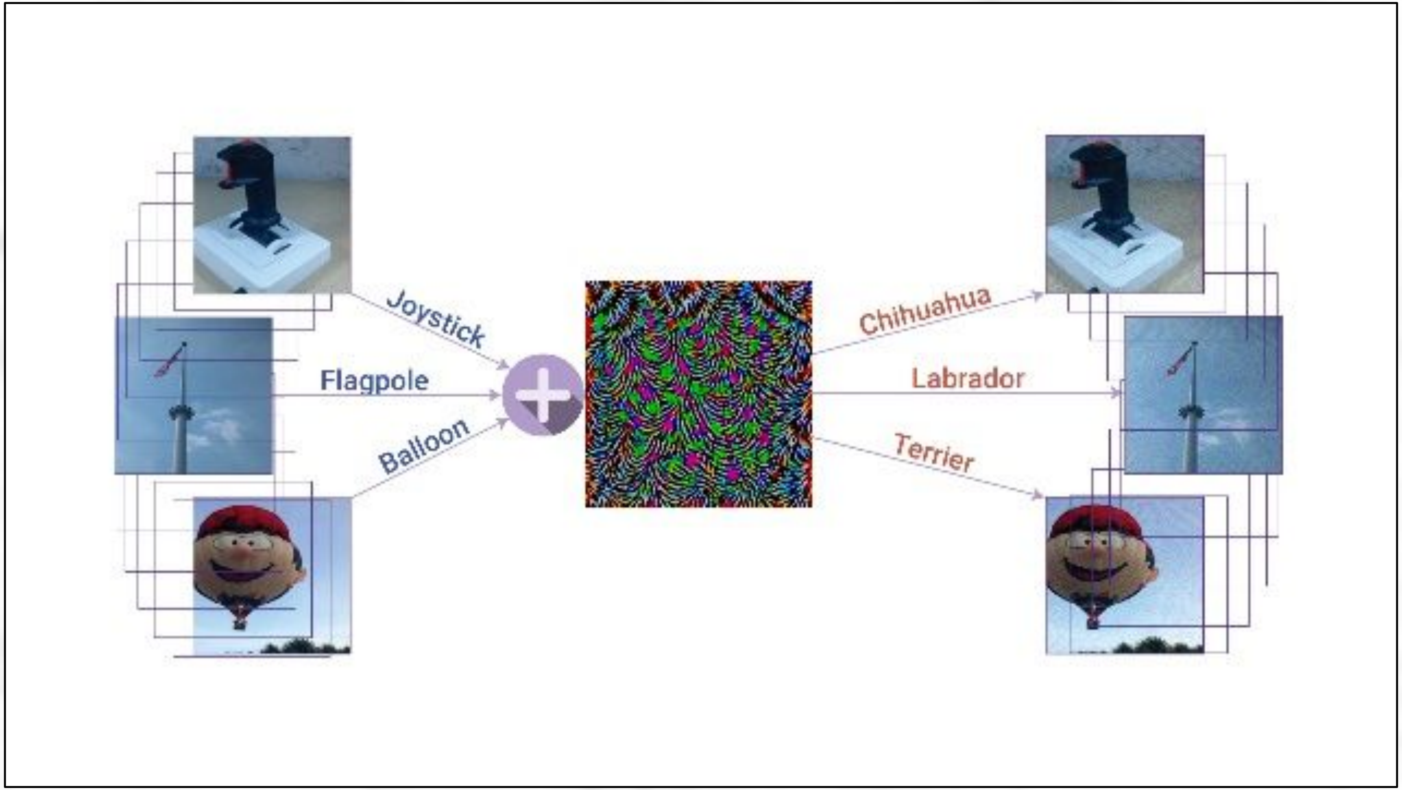
1) FGSM (Fast Gradient Sign Method);
2) PGD (Projected Gradient Descent);
3) DeepFool.



Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples (2014) //arXiv preprint arXiv:1412.6572

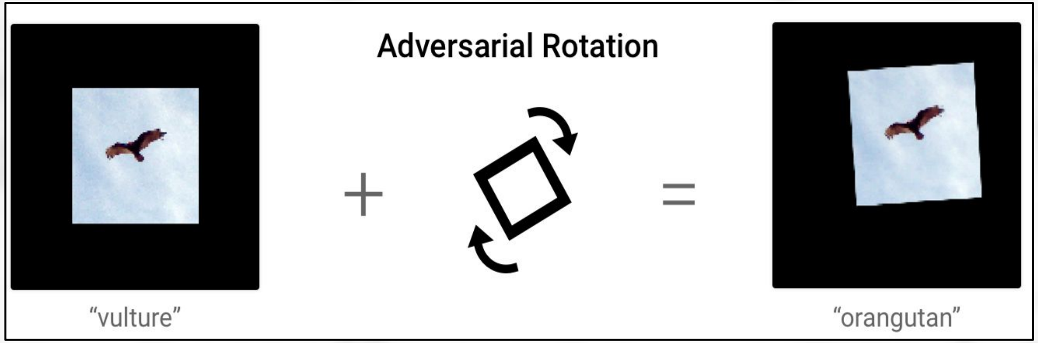

THEY SAY THE BEST WEAPON
IS ONE YOU NEVER HAVE TO FIRE.
I RESPECTFULLY DISAGREE.
I PREFER THE WEAPON YOU ONLY HAVE TO FIRE
ONCE.

Universal Adversarial attacks

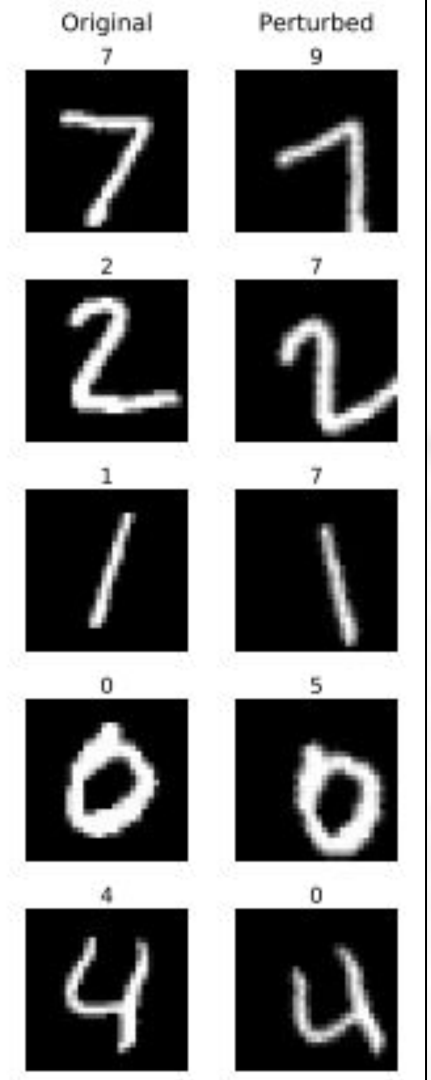


Moosavi-Dezfooli S. M. et al. Universal adversarial perturbations //arXiv preprint. – 2017

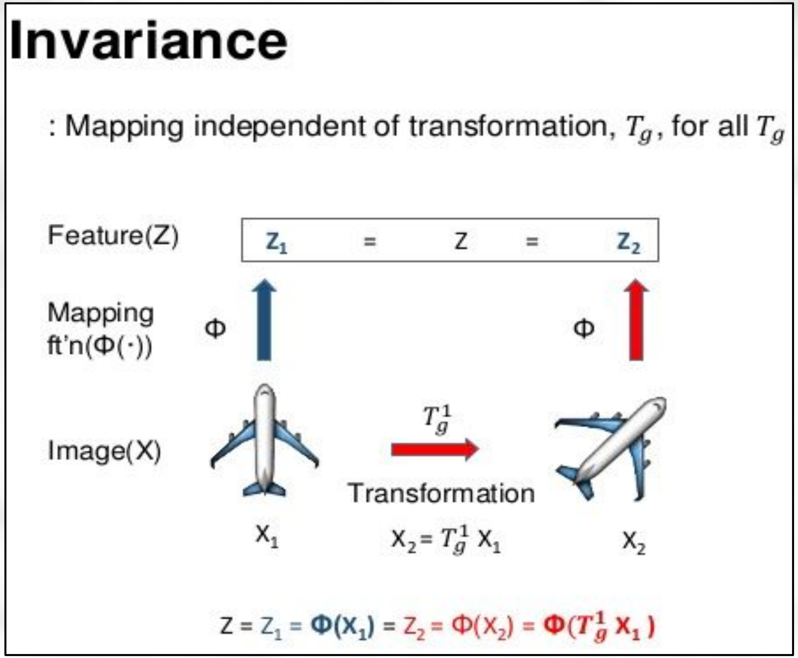
Image Transformation as an attack

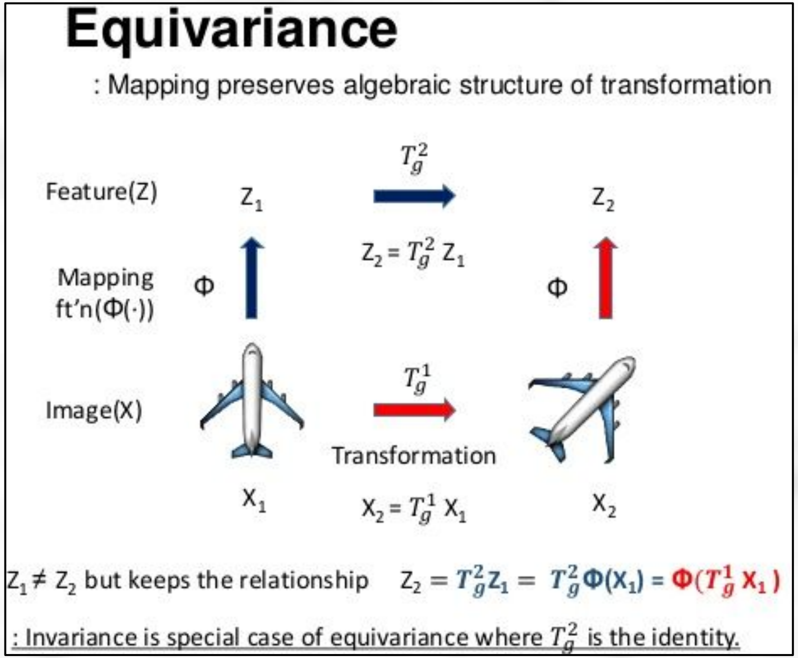


Engstrom L. et al. A rotation and a translation suffice: Fooling cnns with simple transformations //arXiv preprint arXiv:1712.02779. – 2017



Equivariant networks





- Some networks are equivariant to different types of geometric transformations.
- Examples of such networks are:

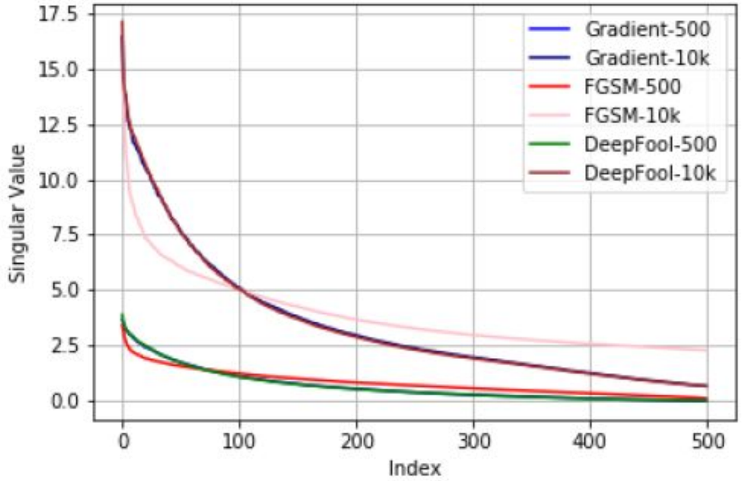
1. StdCNN (translation-equivariant);
2. GCNN (rotation-equivariant);
3. H-Net (rotation-equivariant).

Universal Attacks on Equivariant Networks [under review on ICLR'19]

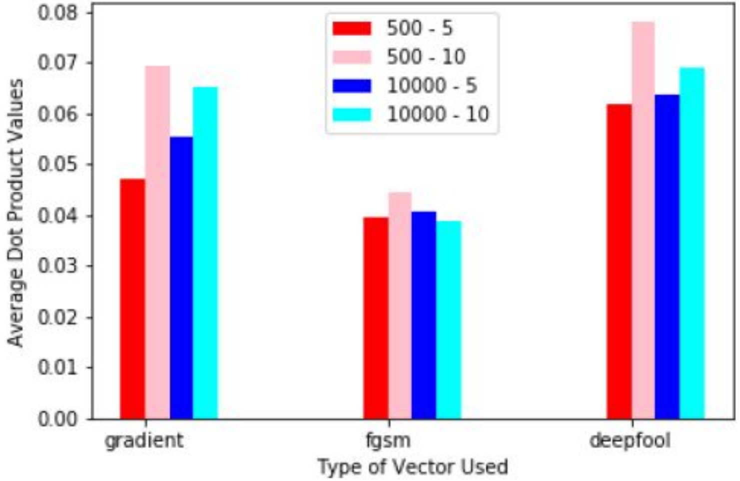
The experiment performed by authors of the article:

1. Trained equivariant networks (GCNN, RotEqNet) on different datasets (including MNIST).
2. Found principal components of adversarial directions:
   2.1. Calculated attack directions (FGSM, DeepFool) for some inputs.
   2.2. Formed a matrix using these directions and computed SVD of this matrix.
   2.3. Investigated the spectrum of this matrix.
   2.4. Observed that top-1 singular vector is a good universal attack.
3. Analyzed principal components of invariant directions in the same way.
4. Observed that top-5 singular vectors of adversarial directions and top-5 singular vectors of invariant directions are nearly orthogonal.

Universal Attacks on Equivariant Networks [under review on ICLR'19]



(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

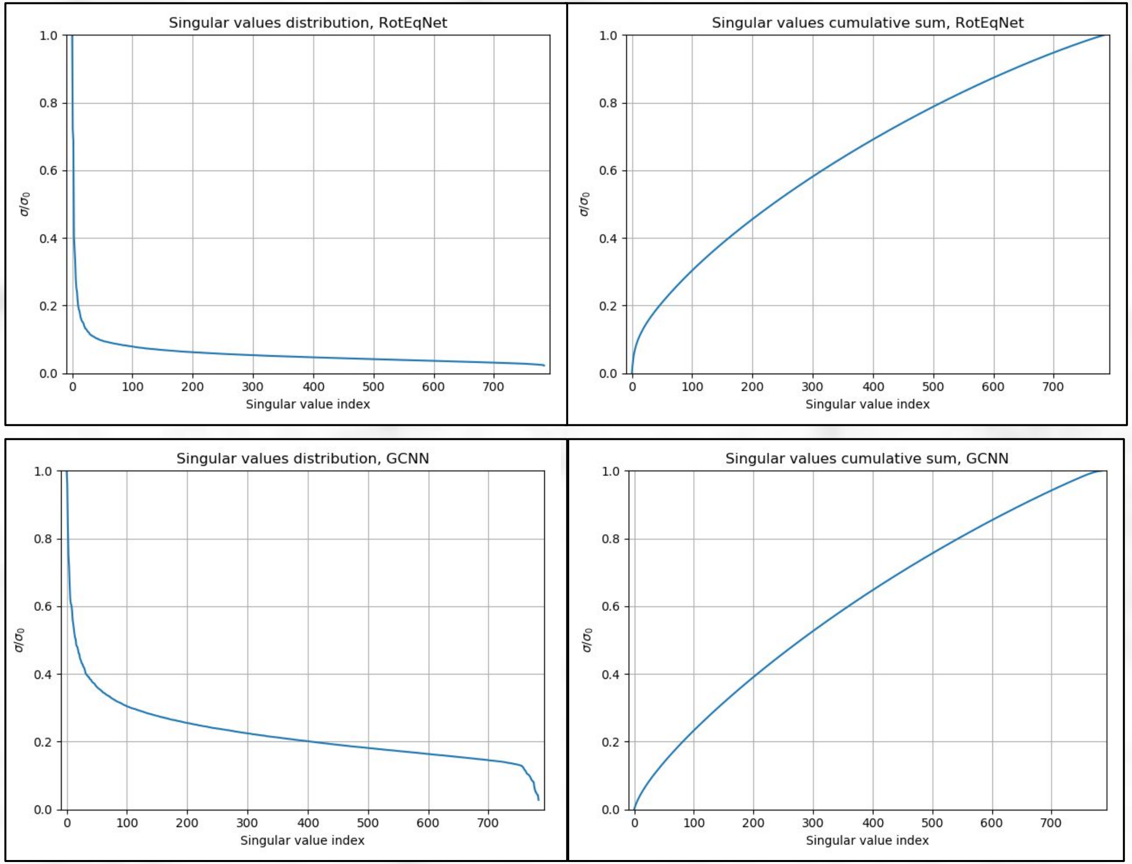
Figure 2: On MNIST, Principal components of adversarial and invariant directions for StdCNN

Numerical experiment

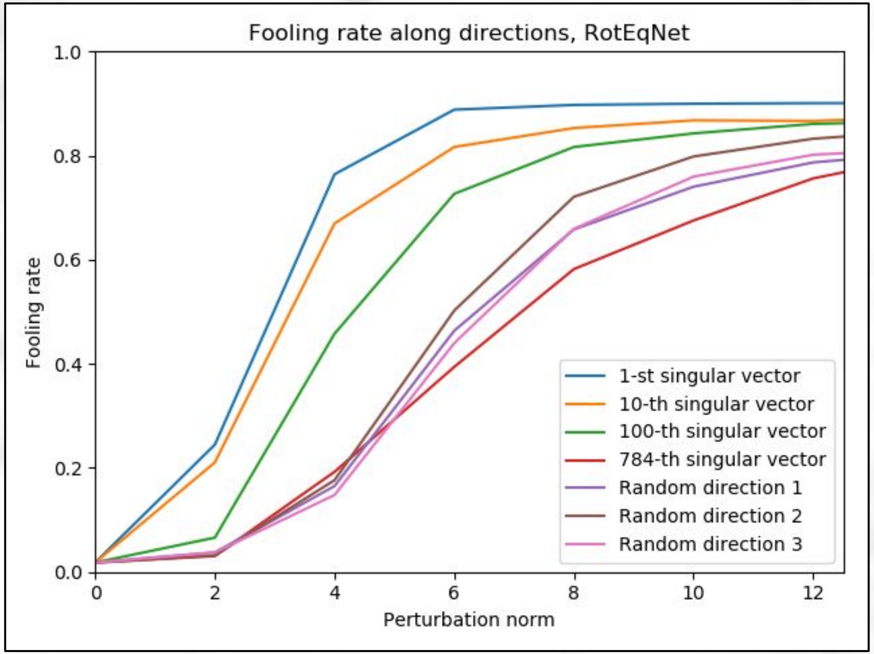
1. We have trained RotEqNet and GCNN on MNIST dataset with respective accuracies 98.2% and 99.3%.
2. Networks were attacked by FGSM on each input.
3. Attack directions were stacked into one matrix for each network.
4. Top singular vectors were used as an universal attack.

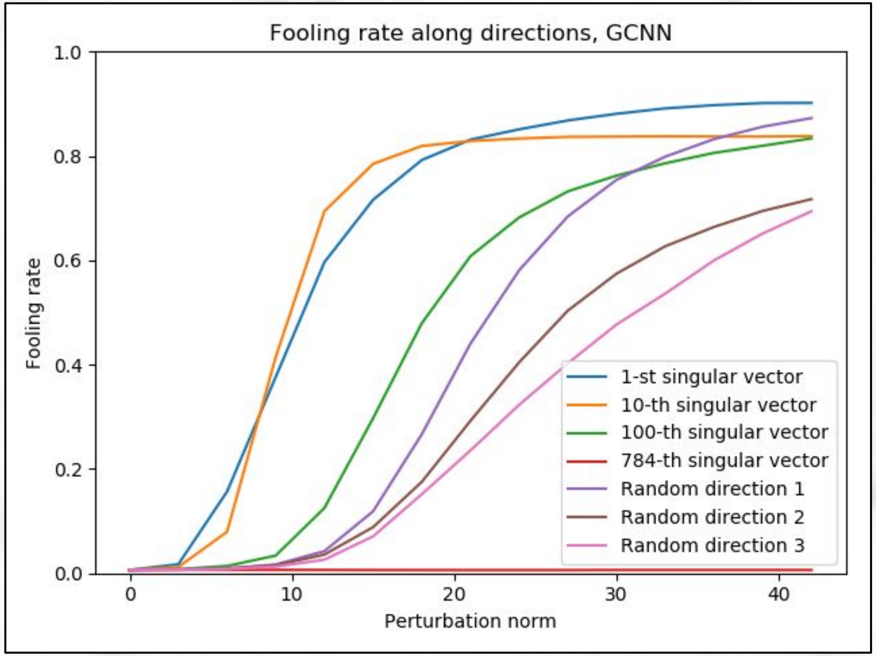
Singular values distribution

	RotEqNet	GCNN
Top-5 values	7.0%	2.4%
Top-10 values	9.9%	4.2%

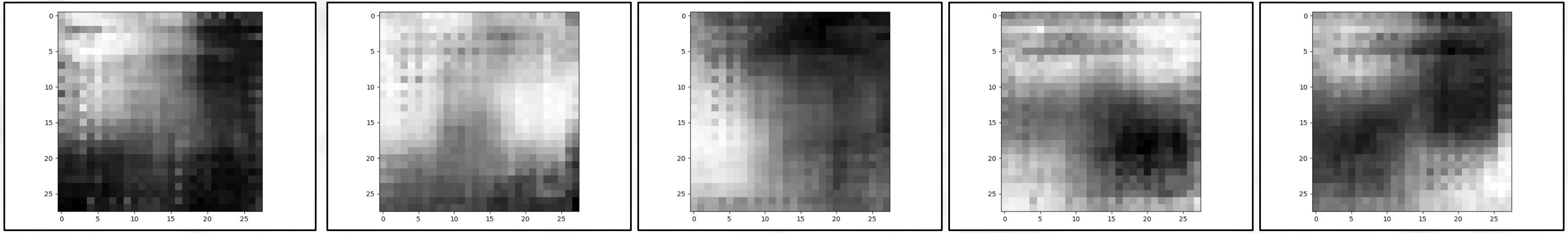


Fooling rates along attack directions

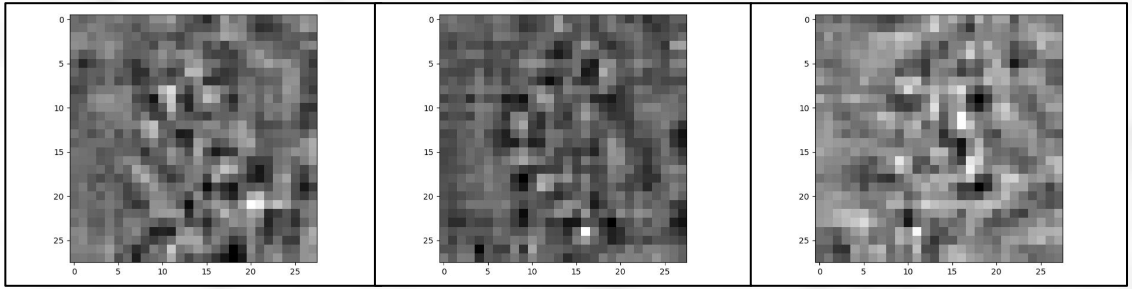




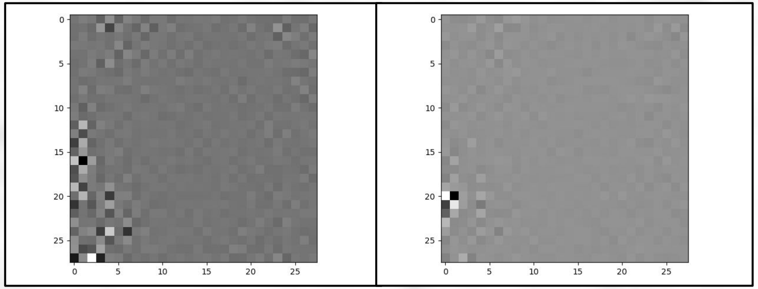
Attack directions, RotEqNet



Top 5 singular vectors

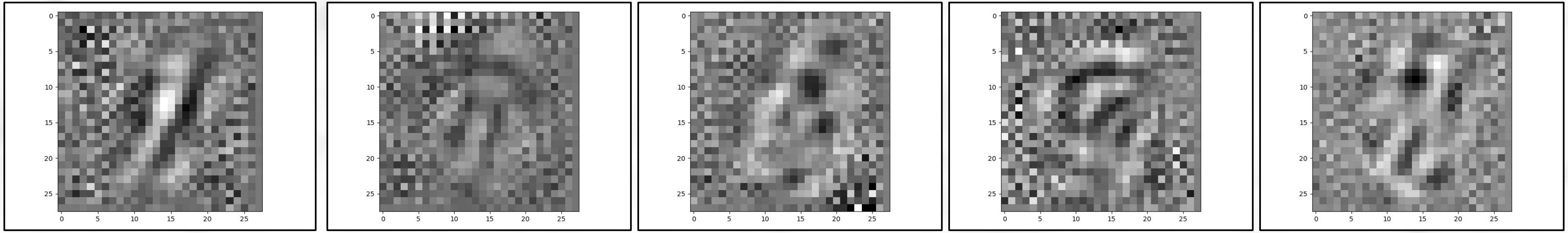


100-102 singular vectors

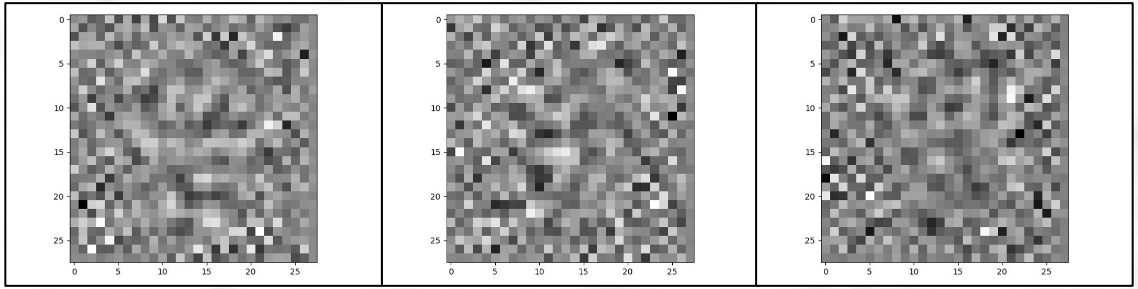


783-784 singular vectors

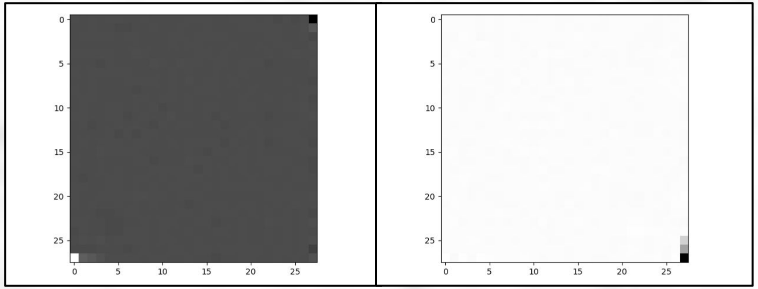
Attack directions, GCNN



Top 5 singular vectors



100-102 singular vectors



783-784 singular vectors

Summary

Our study has confirmed key conclusions of the analyzed article:

1. the significant part of spectrum of adversarial attacks for randomly selected inputs is indeed concentrated in first few singular values. Moreover, this statement holds for various network models;
2. the applicability of the principal component of these attacks as an universal attack has been confirmed;
3. moreover, the ability of this attack to fool specifically rot-equivariant networks has also been confirmed.

Thank you for your attention!