

Art of singular vectors and universal adversarial perturbations

paper by Valentin Khrulkov and Ivan Oseledets

Group 23: Artyom Gadetsky, Darya Voronkova, Anastasia Fadeeva, Andrei
Atanov

Introduction

Adversarial attacks

- Negligible perturbations in input leads to misclassification
- Usually individual attack for an image
- What about universal perturbations?

$$\begin{array}{c} \text{A photograph of a panda's face.} \\ x \\ \text{"panda"} \\ 57.7\% \text{ confidence} \end{array} + .007 \times \begin{array}{c} \text{A small image of a colorful, noisy pattern.} \\ \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"nematode"} \\ 8.2\% \text{ confidence} \end{array} = \begin{array}{c} \text{The same photograph of a panda, but with a faint, multi-colored grid overlay.} \\ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \\ \text{"gibbon"} \\ 99.3 \% \text{ confidence} \end{array}$$

$$\frac{|\{x \in \mathcal{D} : \arg \max p(x) \neq \arg \max p(x + \varepsilon)\}|}{|\mathcal{D}|} \rightarrow \max_{\varepsilon}$$

Method

- Let's find small perturbation which cause the largest difference in some layer:

$$f_i(x + \varepsilon) - f_i(x) \approx J_i(x)\varepsilon$$

$$\|f_i(x + \varepsilon) - f_i(x)\|_q \approx \|J_i(x)\varepsilon\|_q$$

- Find best perturbation via the following problem:

$$\sum_{x_j \in X} \|J_i(x_j)\varepsilon\|_q^q \rightarrow \max \quad \|\varepsilon\|_p = 1$$

- This problem is equivalent to the finding the (p, q) singular vector:

$$\|J_i(X_b)\varepsilon\|_q \rightarrow \max \quad \|\varepsilon\|_p = 1$$

$$J_i(X_b) = \begin{bmatrix} J_i(x_1) \\ J_i(x_2) \\ \vdots \\ J_i(x_b) \end{bmatrix}$$

Method

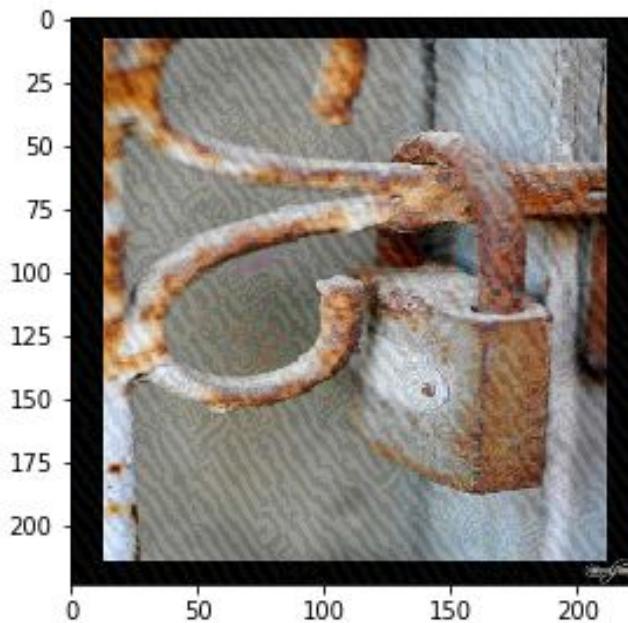
- How to deal with intractable Jacobi matrix?
- We only need matvec operation.

$$\boxed{\nabla \langle v_1, f_i(x) \rangle(x)} = (v_1^T J_i(x))^T = J_i^T v_1$$

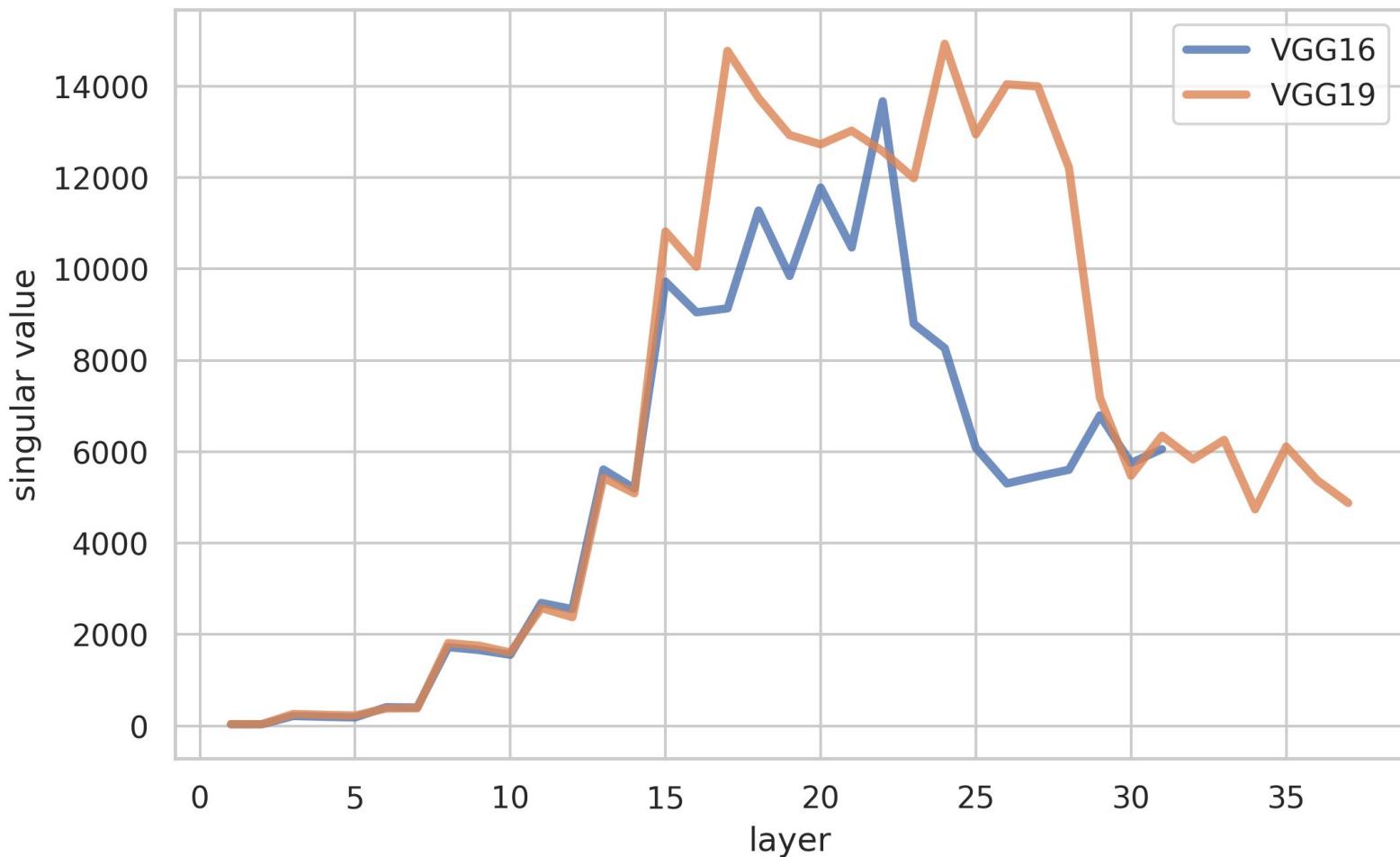
$$\boxed{\nabla \langle J_i(x)v_1, v_2 \rangle} = J_i v_2$$

Automatic differentiation

Example of the attack



Singular values for different layers



Fooling rates

50.000 pictures in test, pretrained architectures from pytorch and inf norm of perturbation is 10

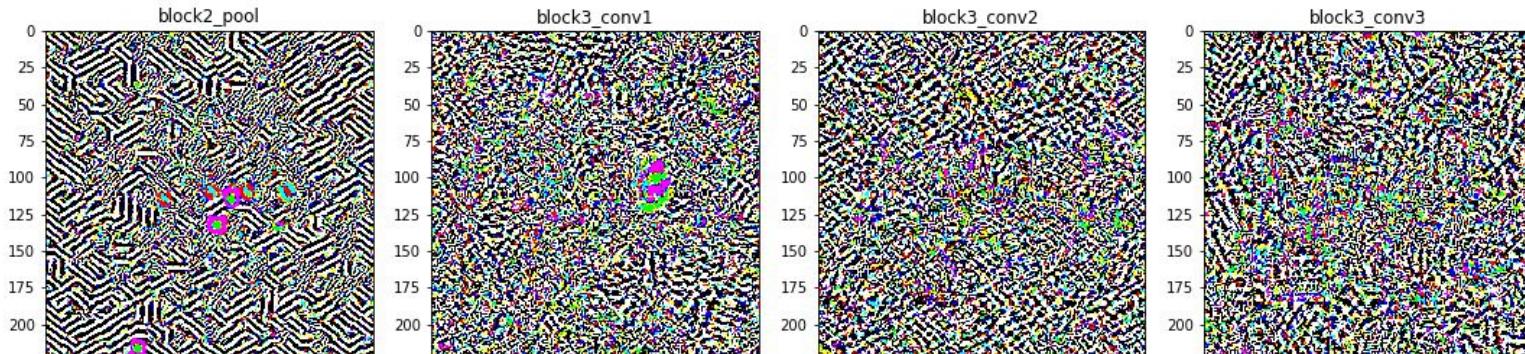
VGG16	block2_pool	block3_conv1	block3_conv2	block3_conv3
singular values	1567.24	2446.83	5056.81	8585.74
fooling rate	55.99	43.3	46.8	44.31

VGG19	block2_pool	block3_conv1	block3_conv2	block3_conv3
singular values	1630.61	2415.33	5044.3	10517.26
fooling rate	55.95	44.39	47.25	45.69

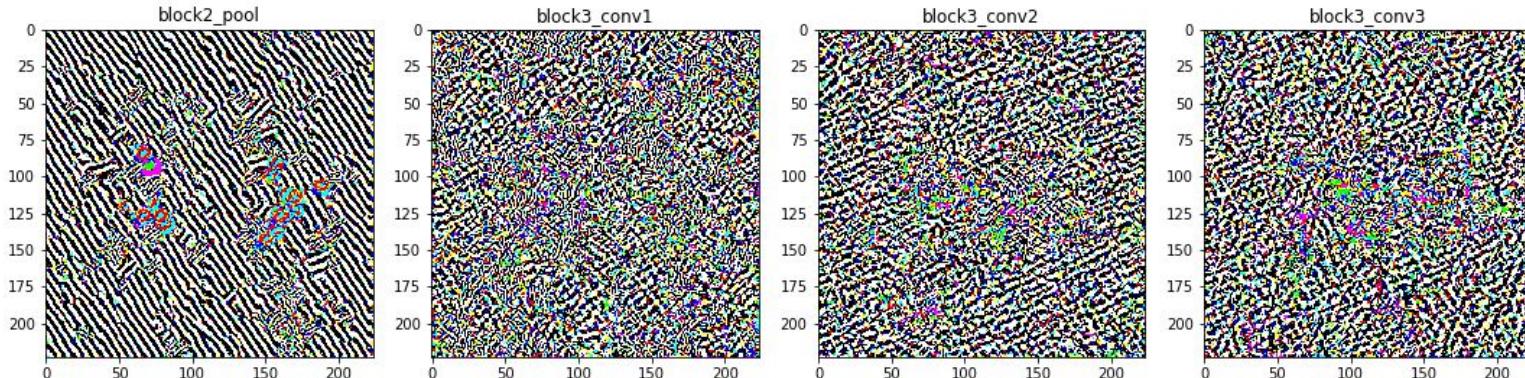
ResNet50	conv1	bottleneck_1	bottleneck_2	bottleneck_3
singular values	61.11	43.4	117.34	669.24
fooling rate	47.33	34.76	33.67	29.44

Fooling rates

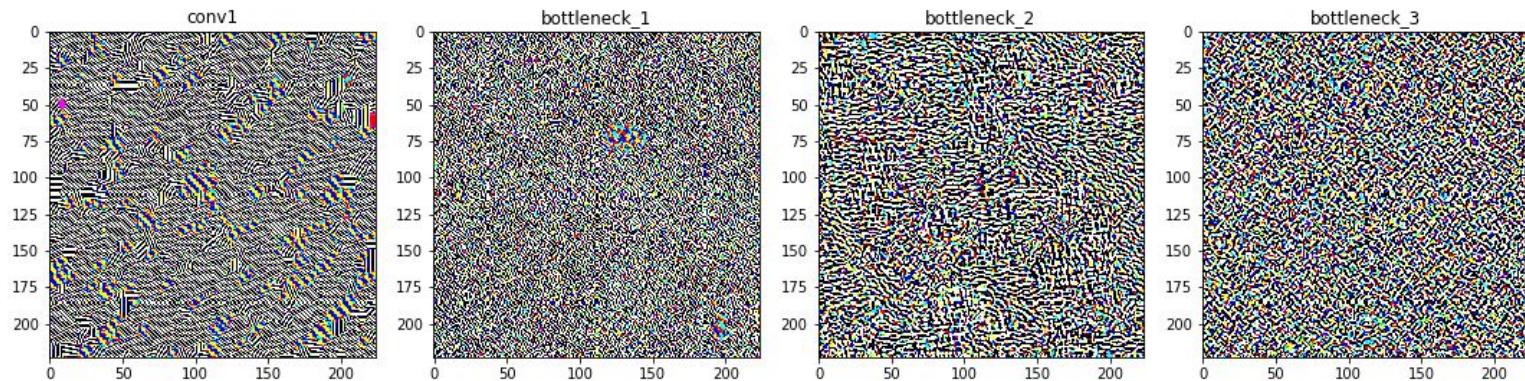
VGG16



VGG19

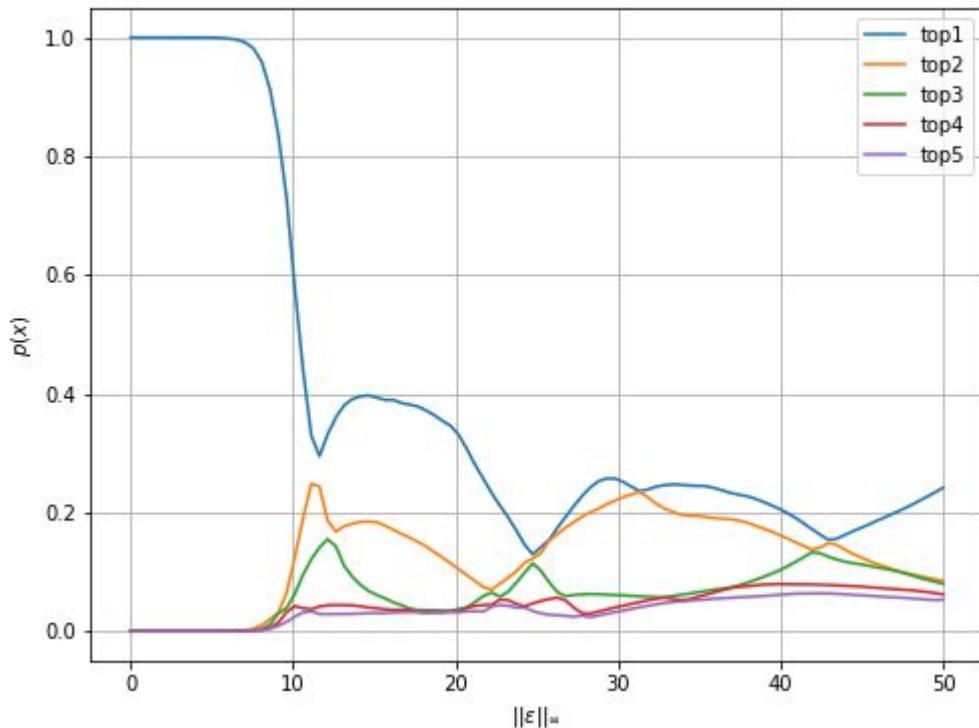


ResNet



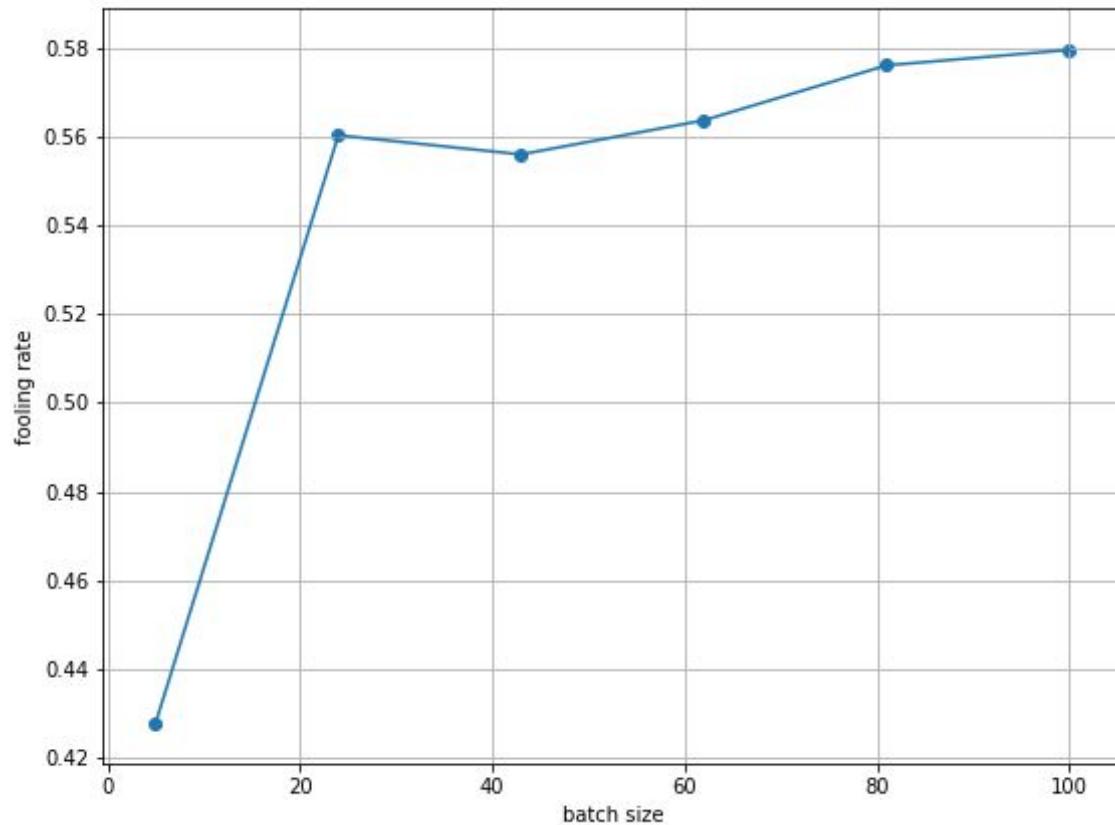
Fooling rates

Generalization	VGG16	VGG19	ResNet50
VGG16	55.99	58.04	62.15
VGG19	57.36	55.95	56.37
ResNet50	37.3	36.65	47.33



Fooling rate

Dependence of the fooling rate on the batch size block2_pool layer in VGG-19 was used.

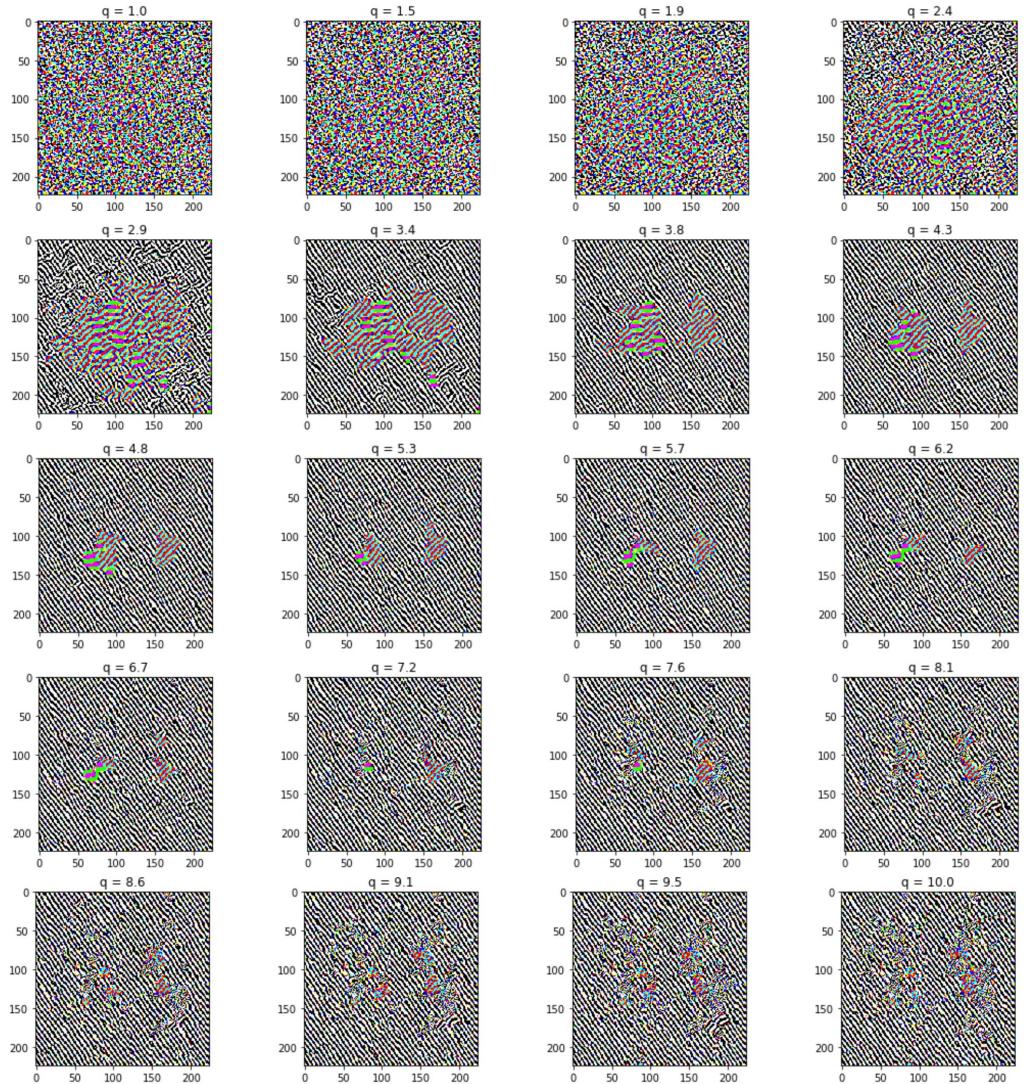


Dependence of the fooling rate on the value of q

Adversarial perturbations
constructed for various values of q.

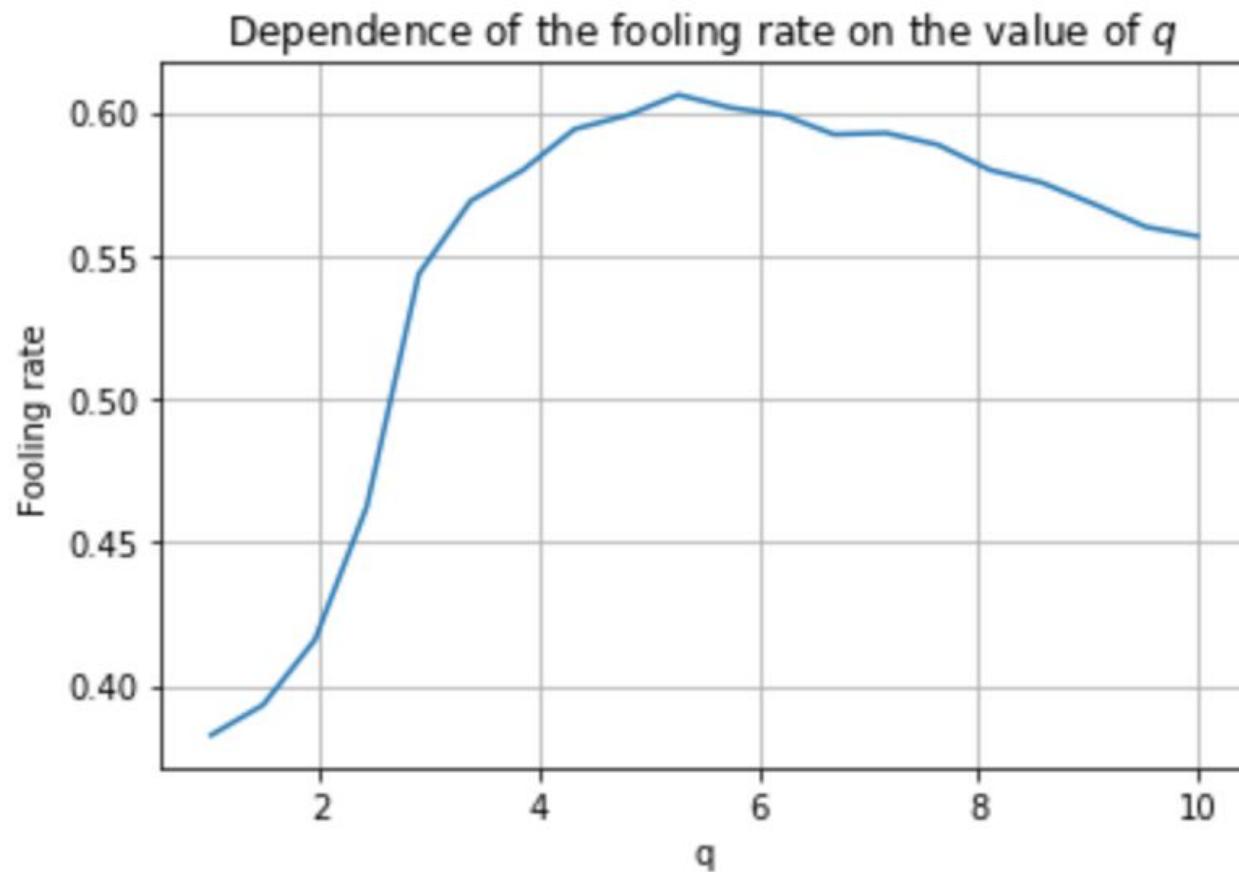
Presented images correspond to
values q increasing from 1.0 to
10.0.

block2_pool layer of VGG-19 was
used.



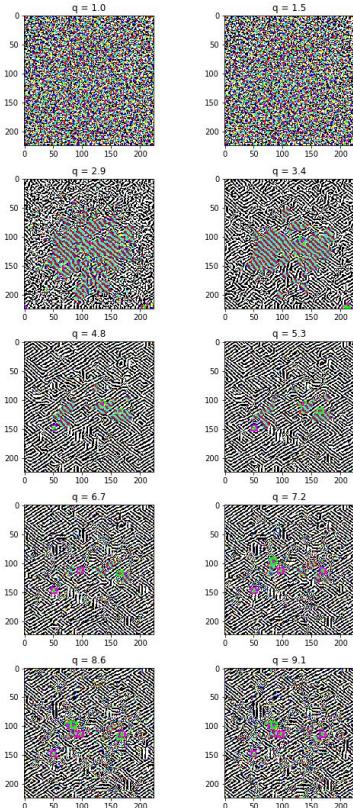
Dependence of the fooling rate on the value of q

Dependence of the fooling rate on the value of q for block2_pool layer of VGG-19.

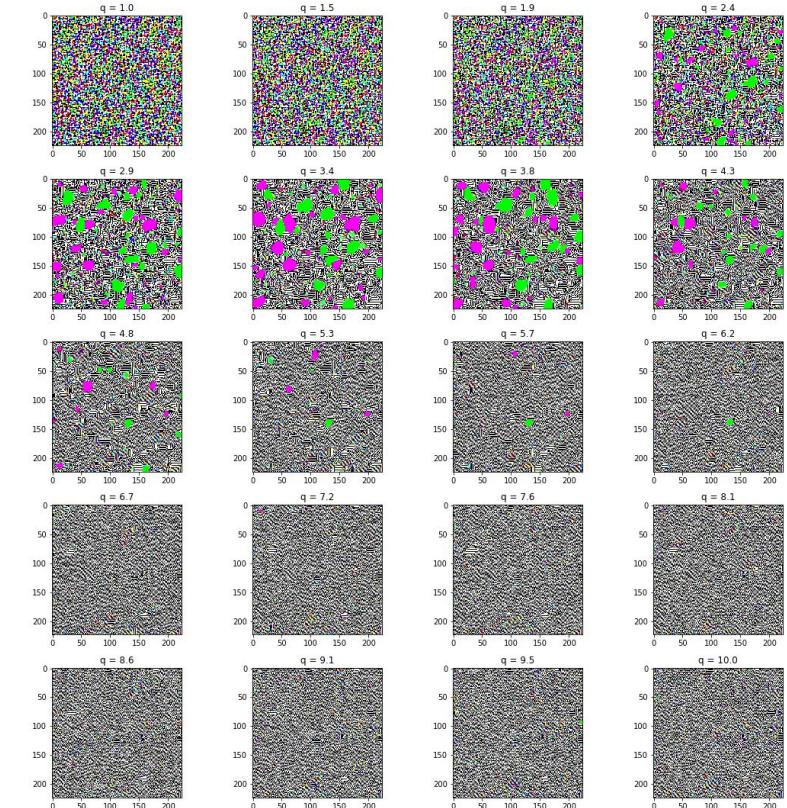


Dependence of the fooling rate on the value of q

Adversarial perturbations constructed for various values of q for VGG-16 and ResNet50.



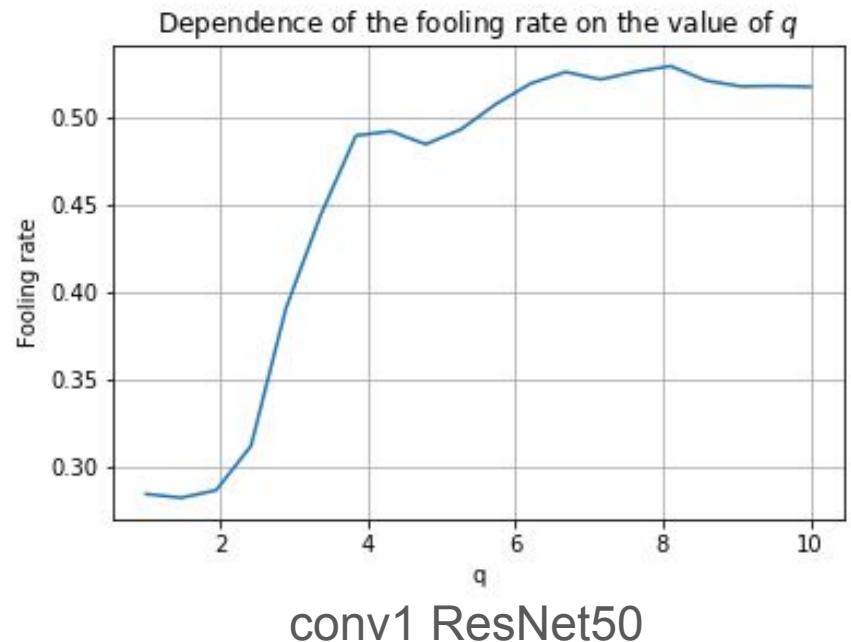
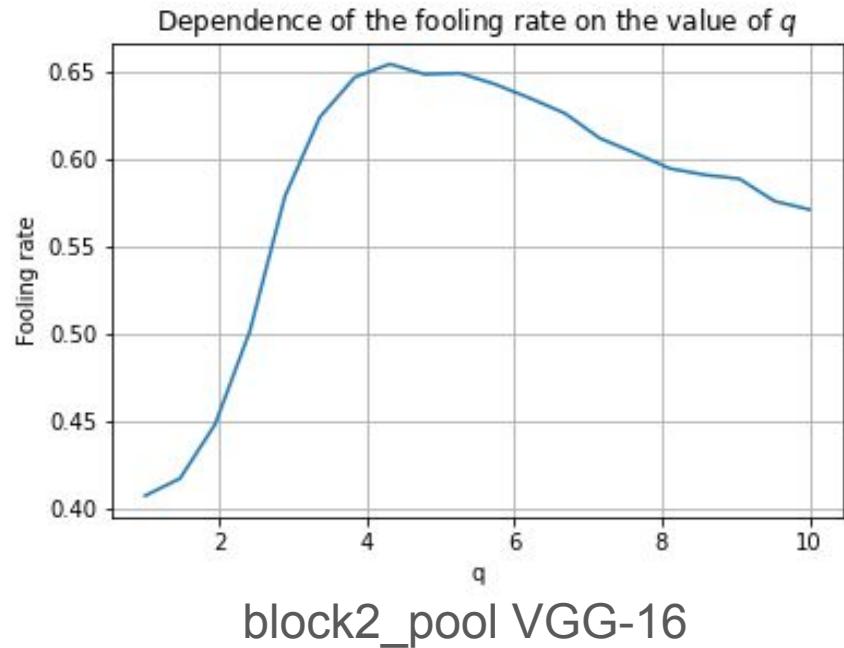
block2_pool VGG-16



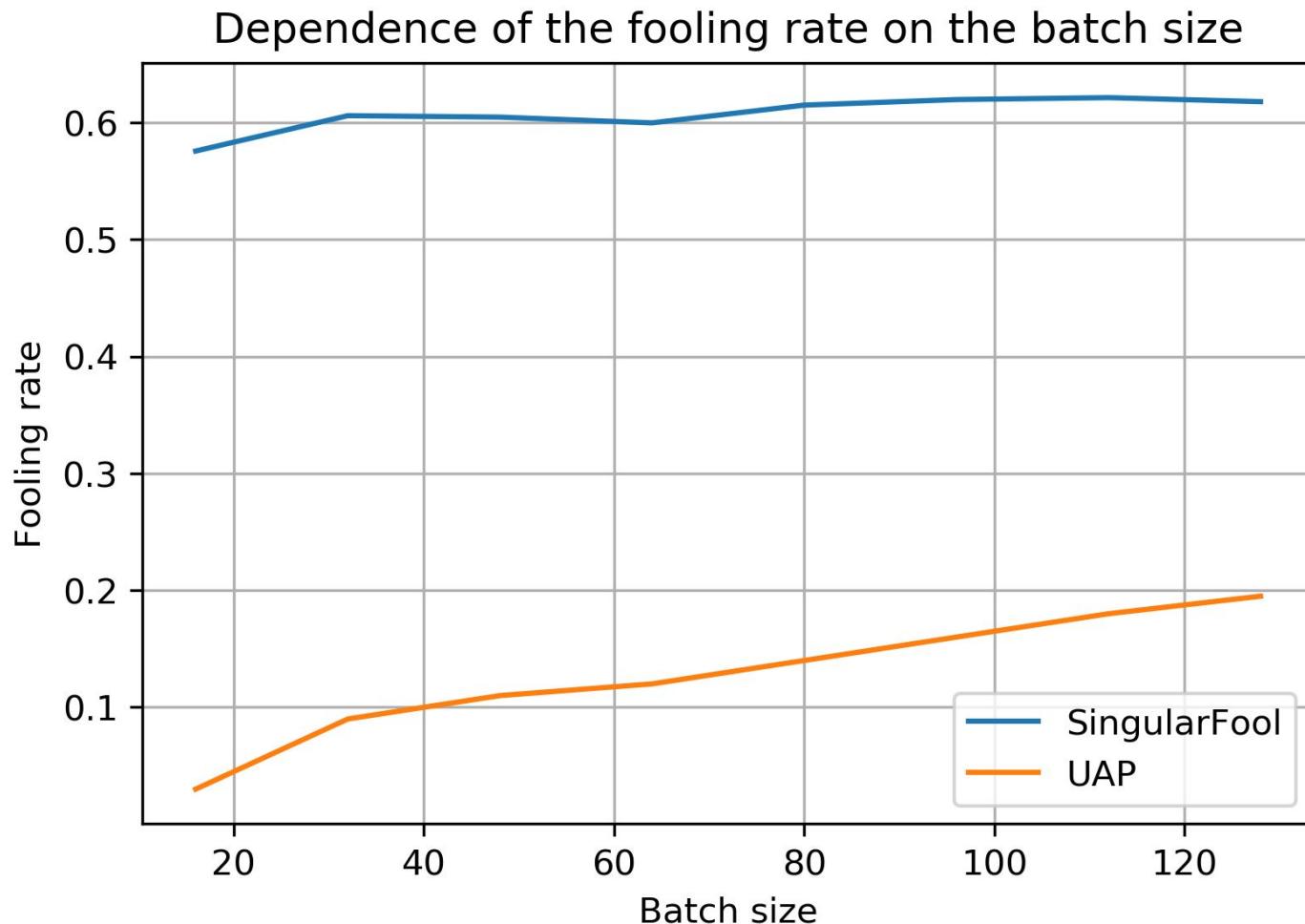
conv1 ResNet50

Dependence of the fooling rate on the value of q

Dependence of the fooling rate on the value of q for VGG-16 and ResNet50.



Dependence of the fooling rate on the batch size



S. Moosavi-Dezfooli*, A. Fawzi*, O. Fawzi, P. Frossard: [Universal adversarial perturbations](#), CVPR 2017

Conclusion

- Reproduced paper proposes a novel state of the art approach for universal adversarial attacks.
 - It is efficient in comparison with other approaches!
 - It needs only 64 images to produce adversarial attack for 60% fooling rate on all validation set of ImageNet (50k images)!

Follow us on Github



goo.gl/Qu36y9

[AndrewAtanov / nla-project](#)

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Unwatch 3 Star 0 Fork 0

NLA Project on reproducing "Art of singular vectors and universal adversarial perturbation"

12 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

agadetsky	Add files via upload	Latest commit c05583f 3 minutes ago
.gitignore	add experiment for singular values for differentet layers	3 days ago
BatchSizeFoolingRates.ipynb	Add files via upload	3 minutes ago
README.md	Update README.md	5 days ago
SingularValues.ipynb	Add files via upload	3 minutes ago
method.py	minor upd	6 minutes ago
singular_exp.py	minor upd	6 minutes ago

README.md

NLA Project on reproducing "Art of singular vectors and universal adversarial perturbation"