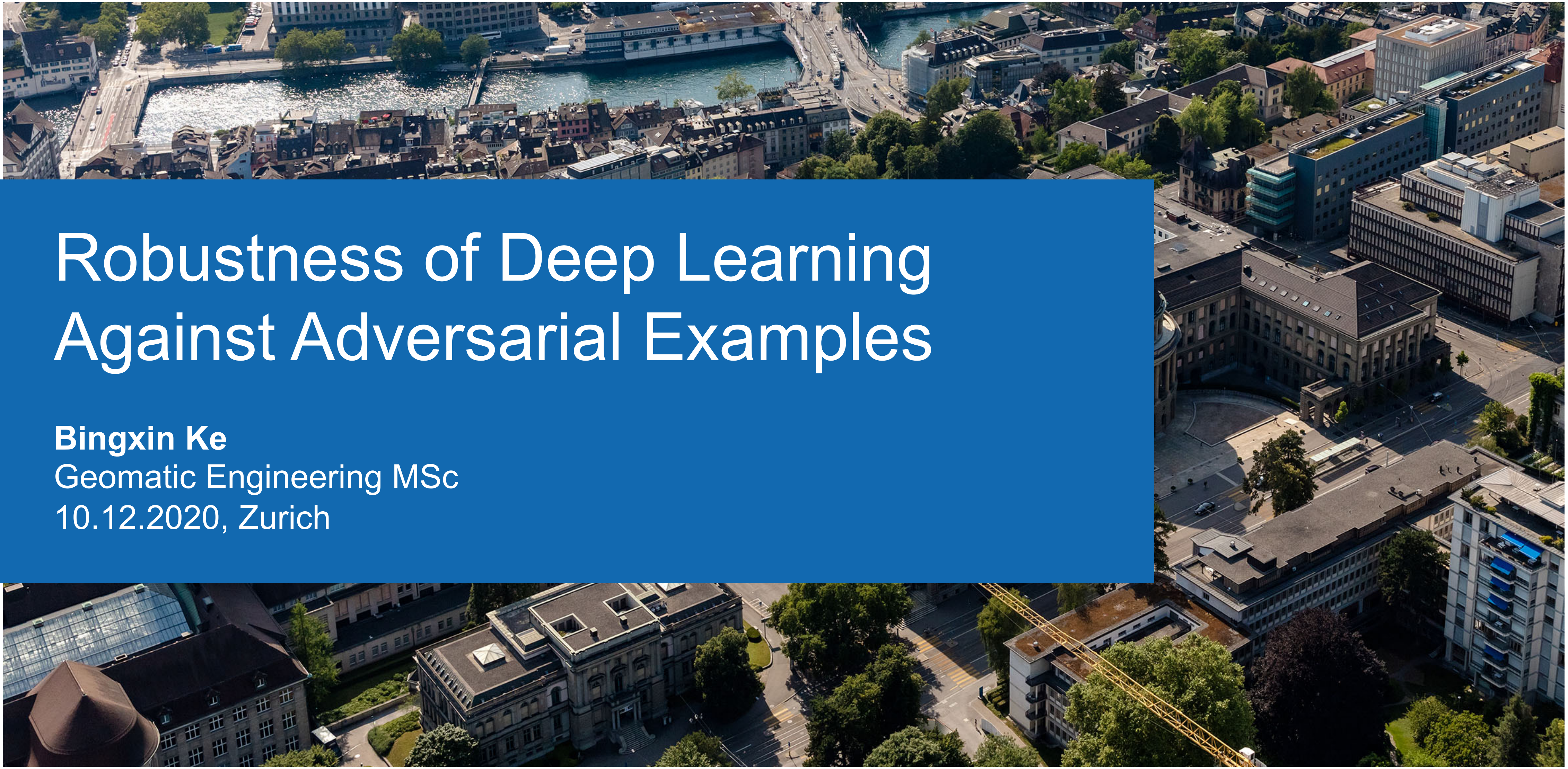
ETH zürich

D-BAUG
ETH Zurich



Robustness of Deep Learning Against Adversarial Examples

Bingxin Ke
Geomatic Engineering MSc
10.12.2020, Zurich

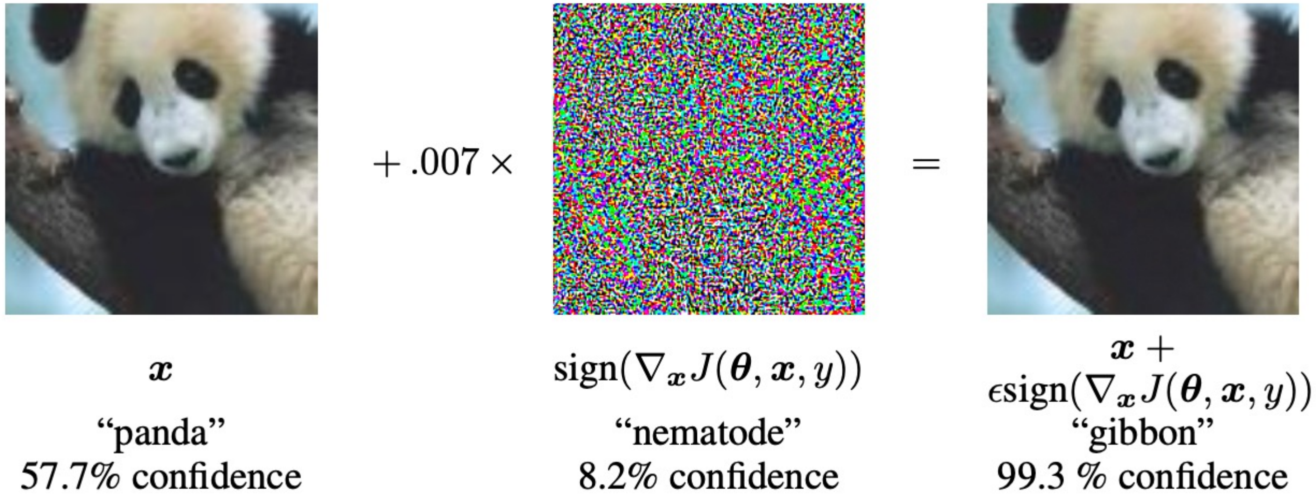
Agenda

1. What are adversarial examples?
2. How do adversarial examples work?
3. How to generate adversarial examples?
4. How to improve robustness?

What are adversarial examples?

What is adversarial example

- Adversarial examples are inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence.[1]



Misclassified
$y' \neq y$

Labeled data
(x, y)

Perturbation
η

Input data
$x' = x + \eta$

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014

What is adversarial example

[7] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1625-1634).

How do adversarial examples work?

How do adversarial examples work

- Speculative explanations:
  - extreme nonlinearity of deep neural networks
  - insufficient model averaging
  - insufficient regularization

- Primary cause of neural networks' vulnerability to adversarial perturbation is their **linear nature.**[1]
- Linear behavior in **high-dimensional space** is sufficient to cause adversarial examples.[1]

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014

How do adversarial examples work

Linear Explanation[1]

- Precision of 8-bit digital image: $\epsilon = 1/255$
- For a well-separated classifier c
  - when $x' = x + \eta, \ \| \eta \|_\infty < \epsilon$
  - expect $\boldsymbol{c(x) = c(x')}$
- Dot product: input and weight vector ω:

Activation growth

$$\boldsymbol{\omega^T x' = \omega^T x + \omega^T \eta}$$

- Maximize adversarial perturbation:
  - ➢ assign $\boldsymbol{\eta = sign(\omega)} \Longrightarrow \boldsymbol{\omega^T \eta = \epsilon \cdot m \cdot n}$
  - ➢ m – average magnitude of elements of ω
  - ➢ n – dimensionality of ω

Simple linear model can have adversarial examples if its input has **sufficient dimensionality**[1].

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014

How to generate adversarial examples?

How to generate adversarial examples
Fast Gradient Sign Method (FGSM)[1]

- Linearize L around θ:

$$\eta = \epsilon \cdot sign\left(\nabla_x L(\theta, x, y)\right)$$

  - ϵ – feature precision
  - $L(\theta, x, y)$ - cost function of a neural network
    - θ – parameters of the model
    - x – input data
    - y – targets (labels)
- backpropagation → gradient

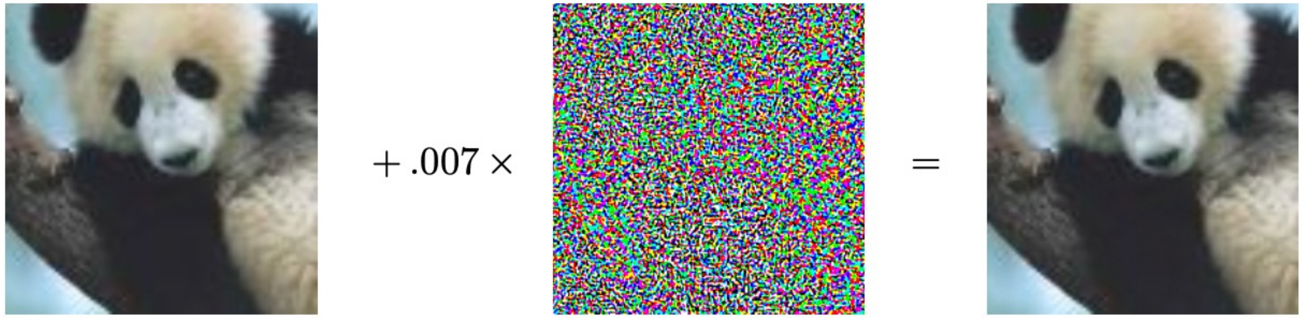
- Pros: computationally simple and efficient
- Cons: low attack success rate

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014
[6] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

How to generate adversarial examples
Fast Gradient Sign Method (FGSM)[1]

GoogLeNet[5] classifier
on ImageNet

8-bit image
$\epsilon = 0.07 (\approx 2/255)$



Experiment Result of FGSM[1]

Classifier	Data set	ϵ	Error Rate	Average confidence
shallow softmax	MNIST	.25	99.9%	79.3%
maxout	MNIST	.25	89.4%	97.6%
convolutional maxout	CIFAR-10	.1	87.15%	96.6%

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, andA. Rabinovich. Going deeper with convolutions. InProceedings of the IEEE conference oncomputer vision and pattern recognition, pages 1–9, 2015

How to generate adversarial examples
Projected Gradient Descent (PGD)[6]

- FGSM is considered a one-step method.[2]
- Projected gradient descent (PGD) also known as iterative FGSM (I-FGSM) is a more powerful **multi-step** variant

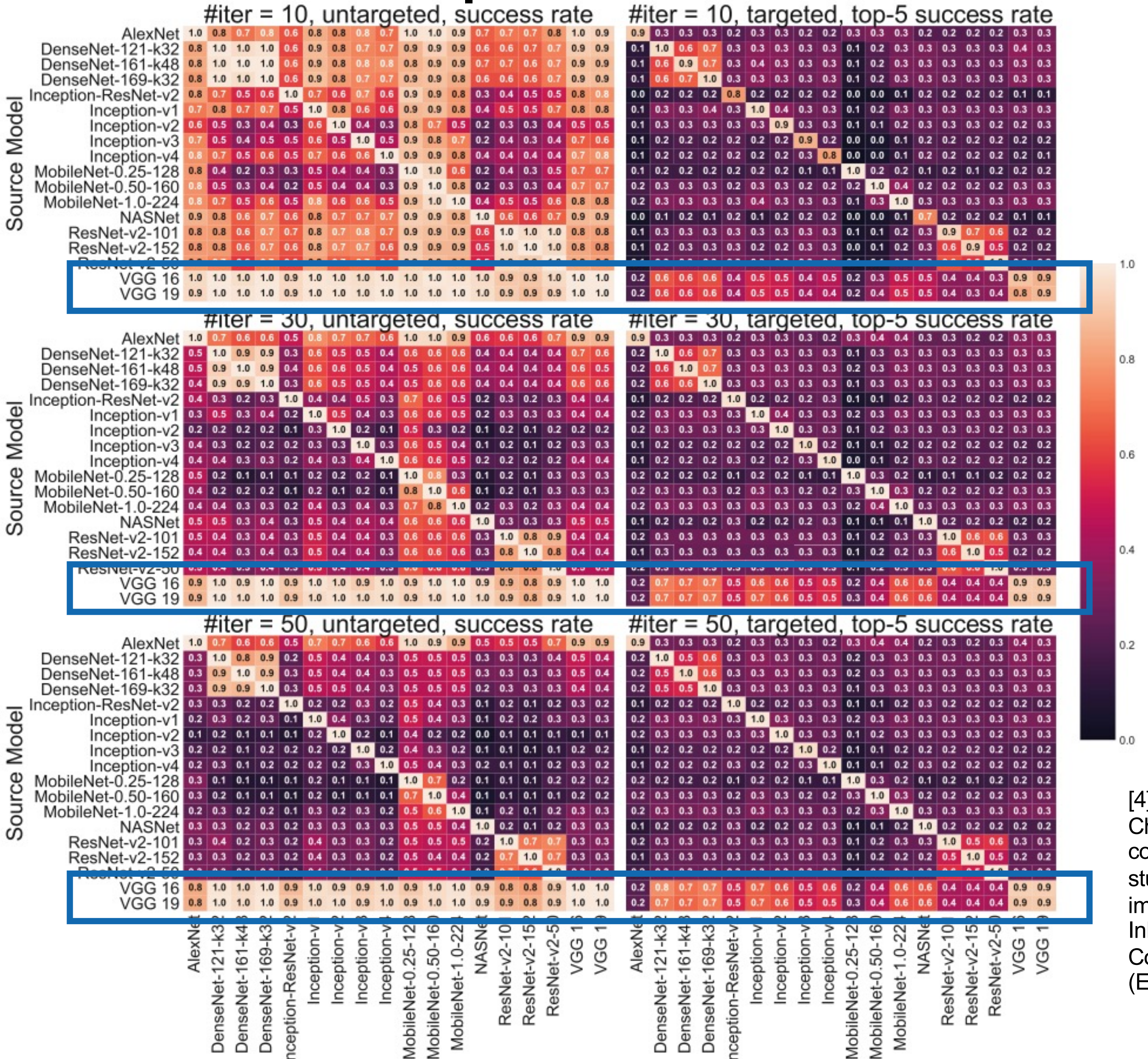
$$x^0 = x, \qquad x^{t+1} = Clip_{X,\epsilon}\left(x^t + \alpha\, sign\left(\nabla_x L(\theta, x^t, y)\right)\right)$$

- ϵ – feature precision
- $L(\theta, x, y)$ - cost function of a neural network
  - θ – parameters of the model
  - x^t – last adversarial example
  - x^{t+1} – next adversarial example
  - y – targets (labels)

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning modelsresistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.
[6] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

How to generate adversarial examples
Transfer attack

Transferability of I-FGSM attack over 18 ImageNet models [4]



[4] Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. InProceedings of the European Conference on Computer Vision (ECCV), pages 631–648

How to improve robustness?

How to improve robustness

A unified view of attacks and defenses (Saddle point problem)

- Saddle point problem[2] aims to summarize the attack-defense problem:

$$\min_{\boldsymbol{\theta}} \rho\,(\theta), \quad \text{where} \quad \rho(\theta) = E_{(x,y)\sim\mathcal{D}}\left[\max_{\boldsymbol{\delta\in\mathcal{S}}} L\,(\theta, x+\delta, y)\right]$$

- Inner maximization - to find an adversarial
- Outer minimization – to minimize adversarial

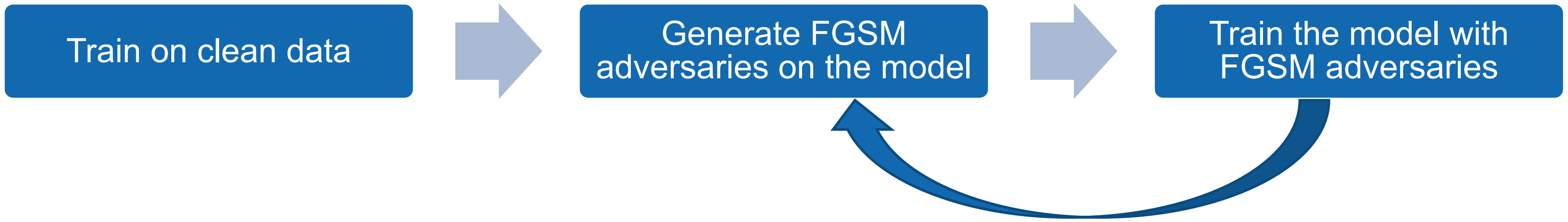
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning modelsresistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.

How to improve robustness
Adversarial Training

- Widely used
- Training on a mixture of adversarial and clean example can regularize the model[5].
- Continually train the model on updating adversarial examples, which resist the current version of the model.

  e.g. using FGSM adversarial examples[1]:

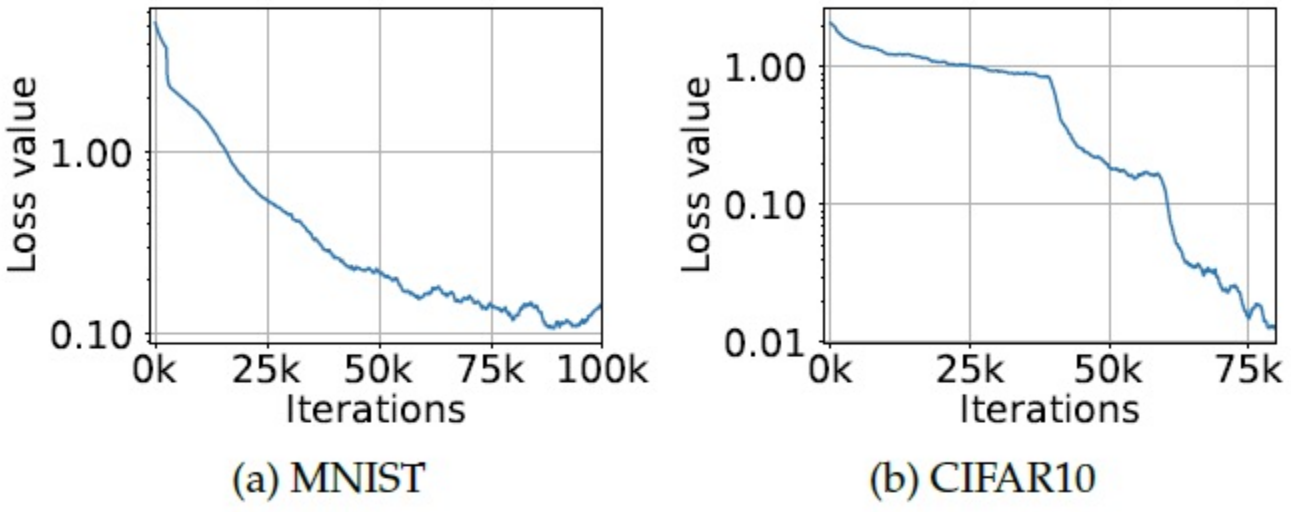
$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1-\alpha) J\big(\theta, x + \epsilon\, sign\big(\nabla_x J(\theta, x, y)\big)\big)$$



[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, andA. Rabinovich. Going deeper with convolutions. InProceedings of the IEEE conference oncomputer vision and pattern recognition, pages 1–9, 2015

How to improve robustness
Adversarial Training using PGD attacking[2]

- Input: PGD perturbed adversarial data + clean data
- Use stochastic gradient descent to minimize the loss function.

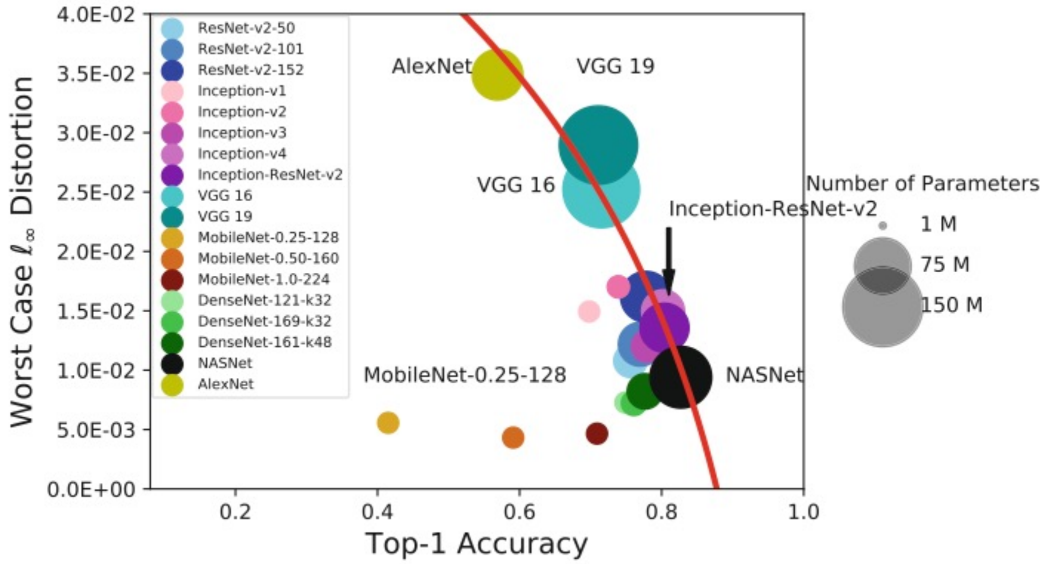


Cross-entropy loss on adversarial examples during training[2]

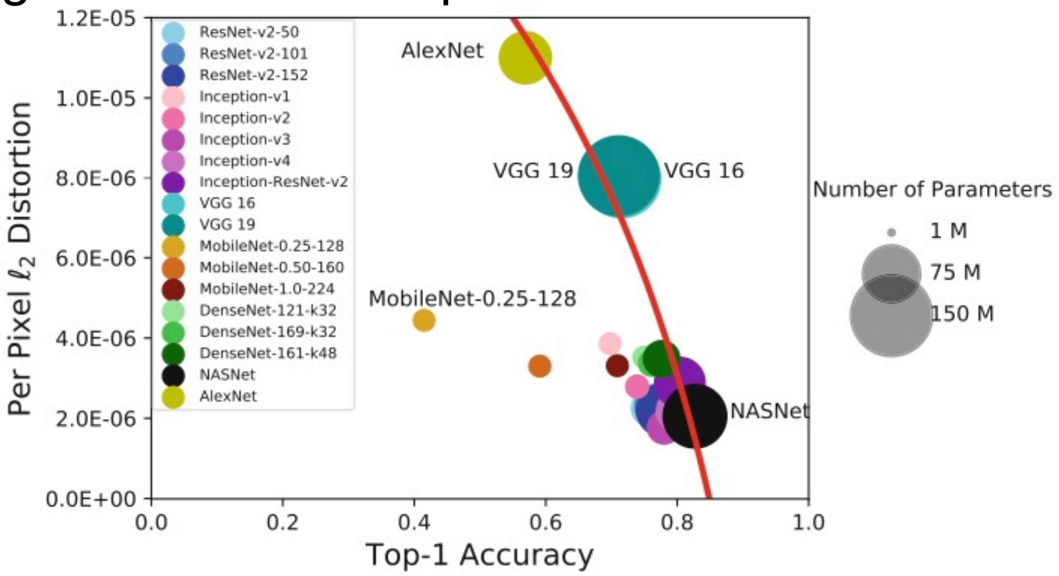
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning modelsresistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.

How to improve robustness
Trade-off between robustness and accuracy

- Robustness is scarified when solely pursuing a higher classification performance.[4]



(a) Fitted Pareto frontier of ℓ_∞ distortion (I-FGSM attack) vs. top-1 accuracy: ℓ_∞ dist $= [2.9 \cdot \ln(1 - \text{acc}) + 6.2] \times 10^{-2}$



(b) Fitted Pareto frontier of ℓ_2 distortion (C&W attack) vs. top-1 accuracy: ℓ_2 dist $= [1.1 \cdot \ln(1 - \text{acc}) + 2.1] \times 10^{-5}$

Robustness vs. classification accuracy plots of I-FGSM attack[4]

- Theoretical decomposition of the prediction error for adversarial examples[3]:

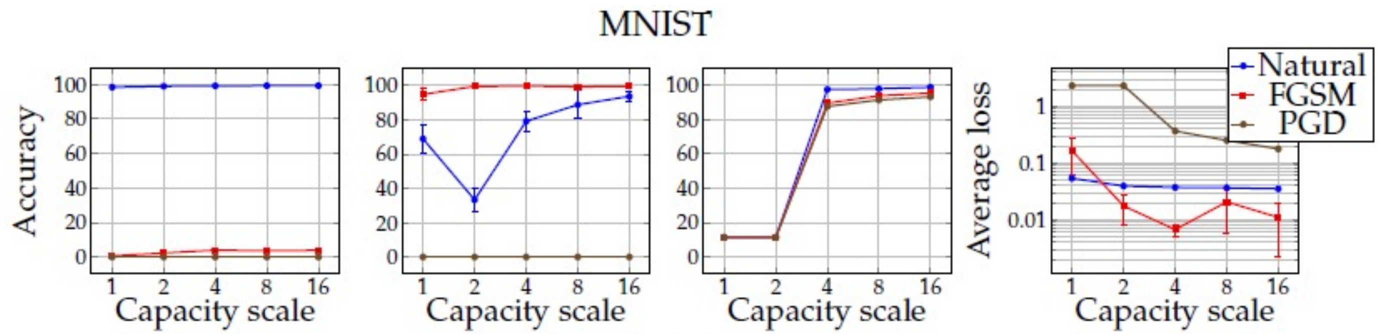
$$\mathcal{R}_{rob}(f) = \mathcal{R}_{nat}(f) + \mathcal{R}_{bdy}(f)$$

[3] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principledtrade-off between robustness and accuracy.arXiv preprint arXiv:1901.08573, 2019.

[4] Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648

How to improve robustness
Model Capacity

- Model capacity: Larger capacity →more information →better discrimination ability[8]
- Model capacity is crucial for the ability to successfully train against adversaries.
- For a similar network architecture, increasing network depth slightly improves robustness in l_∞ distortion metric[4]



CIFAR10

	(a) Standard training: Simple	Wide	(b) FGSM training: Simple	Wide	(c) PGD training: Simple	Wide	(d) Training Loss: Simple	Wide
Natural	92.7%	95.2%	87.4%	90.3%	79.4%	87.3%	0.00357	0.00371
FGSM	27.5%	32.7%	90.9%	95.1%	51.7%	56.1%	0.0115	0.00557
PGD	0.8%	3.5%	0.0%	0.0%	43.7%	45.8%	1.11	0.0218

(a) Standard training (b) FGSM training (c) PGD training (d) Training Loss

Fig. The effect of network capacity on the performance of the network[2]

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning modelsresistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.
[4] Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648
[8] Wang, H. Zhou, W. Xu, and X. Chen. Deep neural network capacity.arXiv preprintarXiv:1708.05029, 2017

How to improve robustness

Model Architecture

- Network architecture has a larger impact on robustness than model size.[4]

- Some networks naturally have better robustness against adversarial examples, e.g. RBF networks are resistant to adversarial examples.[1]
- There is no “best” network architecture, yet.

[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014
[4] Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648

Reference

- **[1] J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples.arXiv preprint arXiv:1412.6572, 2014**
- **[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks.arXiv preprint arXiv:1706.06083, 2017.**
- **[3] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy.arXiv preprint arXiv:1901.08573, 2019.**
- **[4] Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648**
- [5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, andA. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015
- [6] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
- [7] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1625-1634).
- [8] Wang, H. Zhou, W. Xu, and X. Chen. Deep neural network capacity.arXiv preprintarXiv:1708.05029, 2017

Thank you for listening

Take-home thinking

- Can these ideas be used for other models (not only deep network)?
- Can we still rely on existing ML applications?
- ...

Q&A Session