Disentangling Adversarial Robustness and Generalization
David Stutz, Matthias Hein and Bernt Schiele

1 Regular Adversarial Examples Leave Manifold

Regular Adversarial Examples Leave Manifold

Approximate Manifold using Nearest Neighbors

2 On-Manifold Adversarial Examples

3 On-Manifold Robustness is Generalization

4 Robustness Independent of Generalization

FONTS (Synthetic)  EMNIST

Character font affine deterministic decoder

Image Perturbed Latent Code Adversarial Example

Normal Training Adversarial Training
Adversarial Training with On-True-Manifold Adversarial Examples
Adversarial Training with On-Learned-Manifold Adversarial Examples
Adversarial Training with Adversarial Transformations

Contributions

1. Regular adversarial example
2. On-manifold adversarial example

3. On-manifold robustness is generalization.
4. Regular robustness and generalization not contradicting.
   Robustness has higher sample complexity.

Problem

Investigating the relationship between adversarial robustness and generalization – are accurate and robust models possible?

Related Work

▶ [4, 2]: trade-off between robustness and generalization;
▶ [3, 1]: off- or on-manifold adversarial examples.

Paper, Code and Data:
davidstutz.de/cvpr2019


Font Options: FONTS (Synthetic), EMNIST

Contributions

1. Regular adversarial example
2. On-manifold adversarial example

3. On-manifold robustness is generalization.
4. Regular robustness and generalization not contradicting.
   Robustness has higher sample complexity.

Problem

Investigating the relationship between adversarial robustness and generalization – are accurate and robust models possible?

Related Work

▶ [4, 2]: trade-off between robustness and generalization;
▶ [3, 1]: off- or on-manifold adversarial examples.

Paper, Code and Data:
davidstutz.de/cvpr2019