Disentangling Adversarial Robustness and Generalization

David Stutz, Matthias Hein and Bernt Schiele

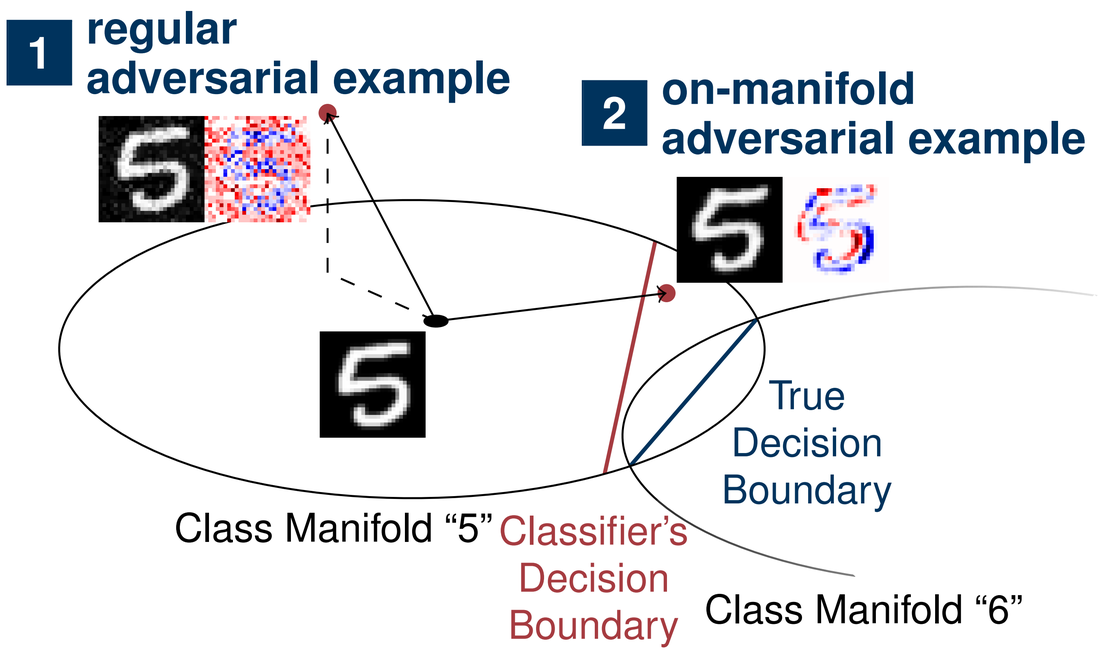
max planck institut informatik

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Problem

Relationship between **adversarial robustness** and **generalization**: **are accurate *and* robust models possible?**

Contributions



3 On-manifold robustness *is* generalization.

4 Robustness and generalization *not* contradicting.

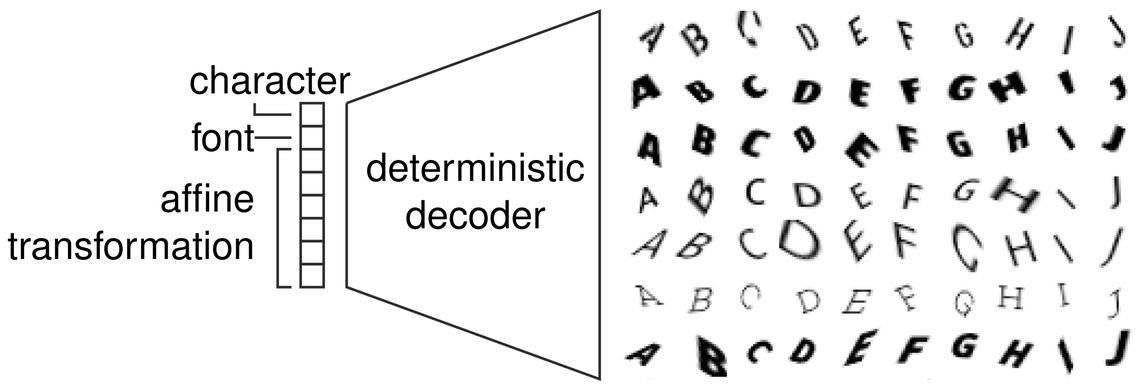
- **Robustness has higher sample complexity.**

Paper, Code and Data:
davidstutz.de/cvpr2019

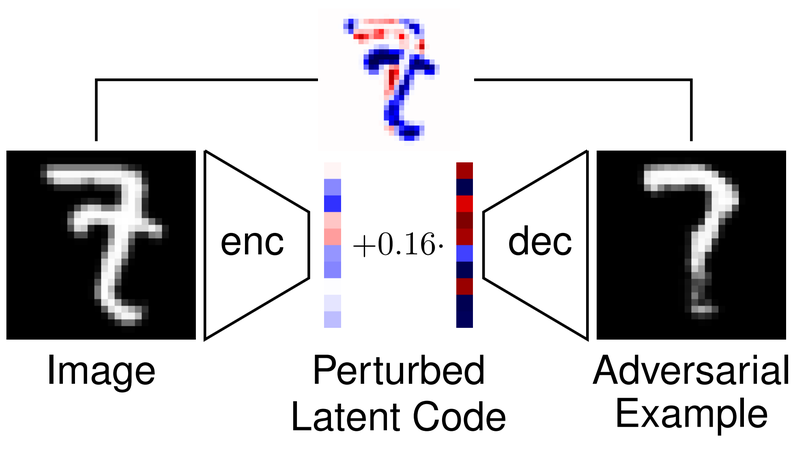
Related Work

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

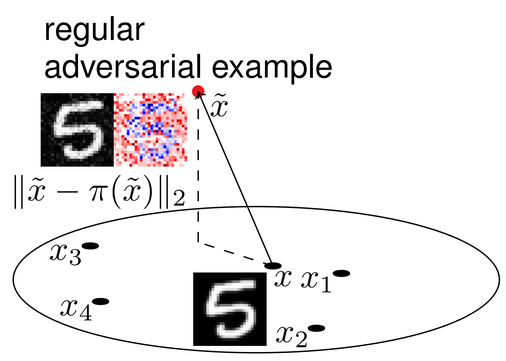
FONTS (Synthetic)



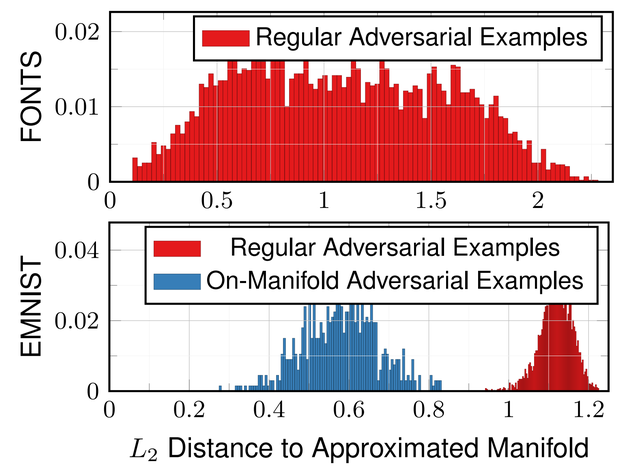
EMNIST



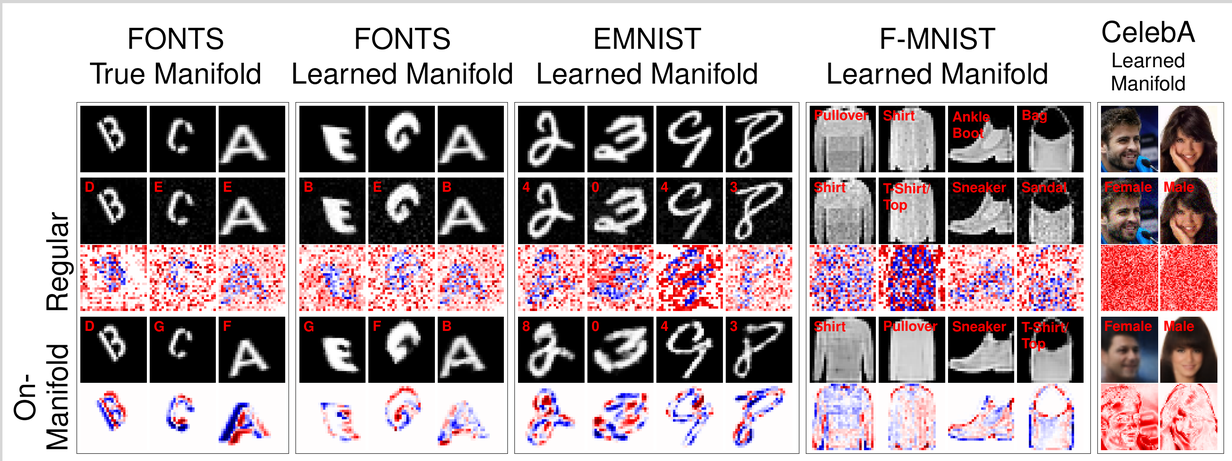
1 Adversarial Examples Leave Manifold



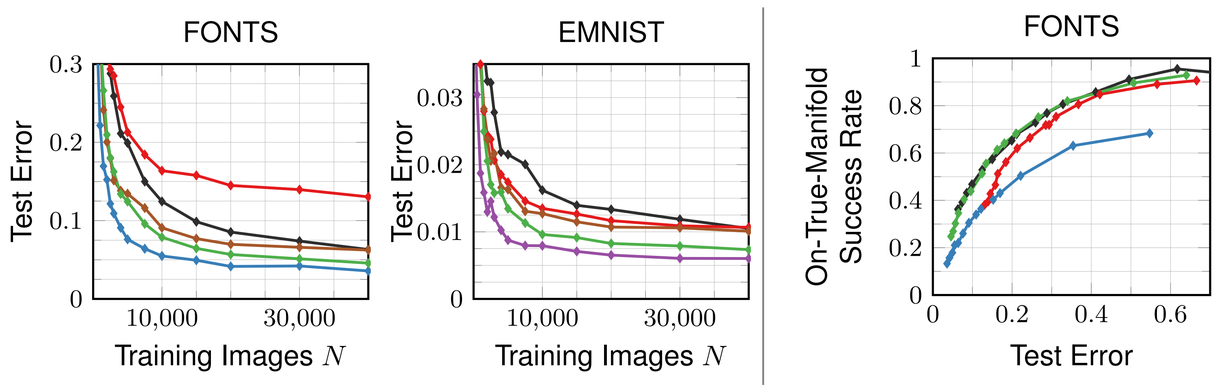
Approximate Manifold using Nearest Neighbors



2 On-Manifold Adversarial Examples

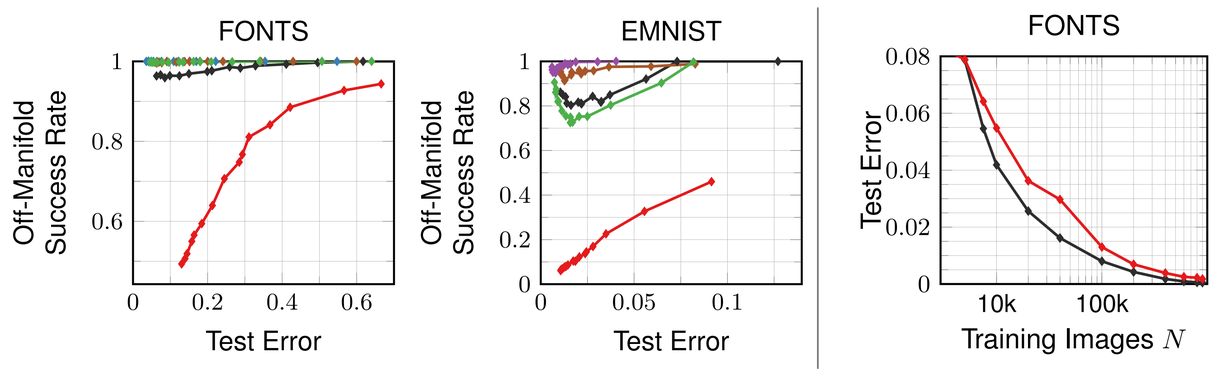


3 On-Manifold Robustness *is* Generalization



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

4 Robustness Independent of Generalization



[1] Justin Gilmer et al. "Adversarial Spheres". In: *arXiv.org* abs/1801.02774 (2018).
[2] Dong Su et al. "Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models". In: *arXiv.org* abs/1808.01688 (2018).
[3] Thomas Tanay and Lewis Griffin. "A boundary tilting persepective on the phenomenon of adversarial examples". In: *arXiv.org* abs/1608.07690 (2016).
[4] Dimitris Tsipras et al. "Robustness May Be at Odds with Accuracy". In: *arXiv.org* abs/1805.12152 (2018).