Adversarial Defense based on Structure-to-Signal Autoencoders

German Research Center for Artificial Intelligence

Presenter **Joachim Folz***, Sebastian Palacio*, Jörn Hees, Andreas Dengel

joachim.folz@dfki.de * equal contribution

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Findings

Our method makes gradients (almost) useless for adversarial attacks.

- Are you just shattering gradients? **Nope, still there.**
- Is accuracy on clean samples decreased? **Nope.**
- Do I need to change my model? **Nope.**
- Is this magic? **Nope**, just autoencoders replacing **semantics** with **structure**.

Bottom line:

- ★ **Attack agnostic** defense.
- ★ **Post-hoc** implementation.
- ★ **No compromise** to accuracy.
- ★ **Architectural** and **representational bottlenecks** are effective defenses against all tested adversarial attacks.


Joachim Folz
joachim.folz@dfki.de

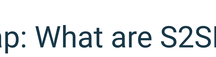

Sebastian Palacio
sebastian.palacio@dfki.de


Jörn Hees
joern.hees@dfki.de

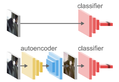

Andreas Dengel
andreas.dengel@dfki.de

Experimental Setup

1. Recap: What are S2SNets

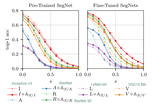


We trained a large SegNet autoencoder on 92 million images from the YFCC100m dataset. It has excellent generalization capabilities.



To analyze the behavior of deep neural networks, we place this autoencoder in front of several classifiers.

Accuracy is down initially, but was recovered by fine-tuning the decoder to suit the classifier. This is the **S2SNet training method**.



We also noticed that S2SNets are more resilient to random noise.

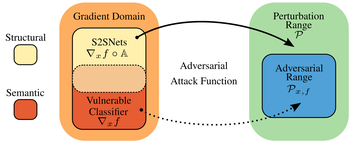
This is the starting point of our paper: to investigate properties and behavior of these combined networks with adversarial noise.

Fine-tuned autoencoders are known as S2SNets, short for structure-to-signal networks.

2. Combine fine-tuned autoencoders with state-of-the-art classifiers

Formally, view gradient-based adversarial attacks as function mapping from gradient domain to perturbation range.

We propose that there are two partially overlapping subspaces of structurally and semantically relevant gradients.



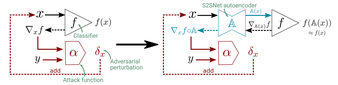
Thus, moving gradients into the structural subspace is an effective defense against gradient-based attacks.

3. White-box Attack defended ResNet 50 & Inception v3, analyze behavior:

- **FGSM** — Fast Gradient Sign Method
- **BIM** — Basic Iterative Method = Iterative FGSM
- **Carlini-Wagner** — CW; iterative optimization-based method

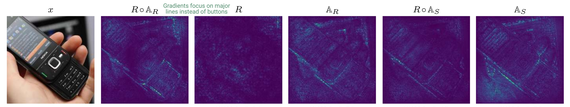
For undefended networks gradients are intercepted at the input layer and perturbations are added to the input. These attacks are usually very effective.

With our defense gradients are intercepted at the input layer of the S2SNet autoencoder. These attacks are far less effective. They target the reconstruction, not the classification.



Results

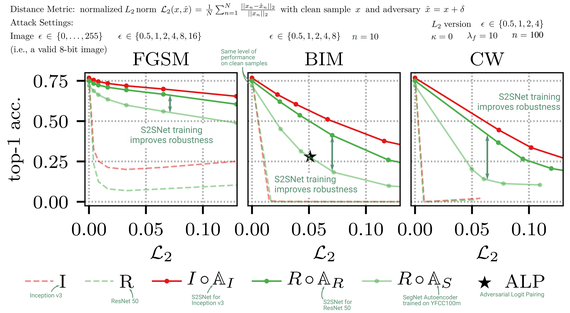
1. Structural nature of gradients



We compared gradients of classifiers, autoencoders, S2SNets, and combined networks using structural similarity (SSIM). Gradients of autoencoders are more similar among themselves than compared to any classifier on its own.

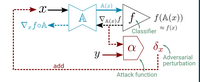
2. S2SNets defend against well-known adversarial attacks

We performed a large-scale evaluation of our defense method on the full ILSVRC 2012 validation set.

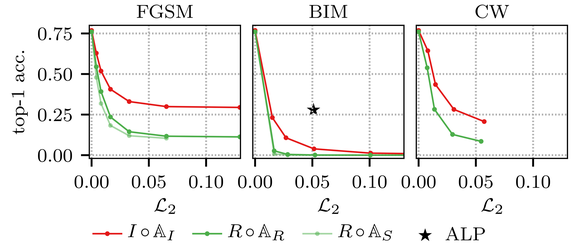


The autoencoder on its own adds an **architectural bottleneck**. Fine-tuning into S2SNet adds a **representational bottleneck**, thus greatly improving robustness to tested attacks.

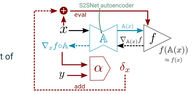
3. Counterfactual: Attack S2SNet with classifier gradients.



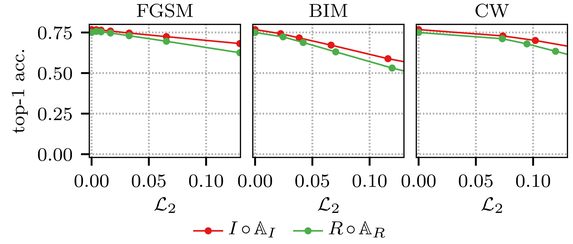
Gradients are intercepted at the classifier and the resulting perturbation added to the input of the autoencoder. These perturbations are much more potent, as they are close to expected signals and thus preserved by the autoencoder.



4. Counterfactual: Attack classifier with S2SNet gradients.



Gradients are intercepted at the autoencoder and the resulting perturbation added to the input of the classifier. These perturbations are not understood by the classifier and thus much less potent. This is further evidence that previously the autoencoder and thus the structure was attacked.



Future Work:

- Investigate whether interpretability techniques provide further insights on how adversarial attacks interact with S2SNets.
- Develop specific (lighter weight; production-ready) S2SNet autoencoder architecture.
- What about non-gradient-based attacks?

This work was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab (NVAIL) program. We also thank all members of the Deep Learning Competence Center at the DFKI for their comments and support.

NVIDIA. Federal Ministry of Education and Research