Semantic Adversarial Attacks: Parametric Attacks that fool Deep Classifiers

Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde
Iowa State University

ICCV 2019 Seoul, Korea

Introduction

Types of Attacks.

- Imperceptible adversarial attacks → Pixel space but unrealized in real life.
- Perceptible semantic attacks → Latent parameter space and **realizable**.

We propose a method to generate semantic adversarial examples for deep classifiers by perturbing meaningful attributes of the given data.

Problem

- A parametric framework for generating adversarial examples for deep classifiers by modifying *semantic* attributes using attribute conditioned generative models.
- Theoretical upper bound on the robust classification error for subspace constrained semantic attacks.
- Experimental evidence of theoretical bounds for strength of various parametric attacks.

Semantic Attacks

Semantic Adversarial attack: Transforming an input image, $\mathbf{x}$ via a parametric model to produce a new example $\tilde{\mathbf{x}} = G(\mathbf{x}, \mathbf{a})$ such that $f(\tilde{\mathbf{x}}) \neq f(\mathbf{x})$. Therefore, the generation of semantic attacks can be decomposed into two steps:

- Train a parametric model, $G(\mathbf{x}, \mathbf{a})$ that decouples semantic attributes from underlying data.
- Search over the attribute space of the trained generator to find an adversarial example for $f(\mathbf{x})$.

Algorithm 1 Adversarial Parameter Optimization

Require: $\mathbf{x}_0$:Input image, $\mathbf{a}_0$:Initial attribute vector, $E(.)$: Attribute encoder, $G(.,.)$:Pre-trained parametric transformation model, $f(.)$: Target classifier, $\mathbf{y}$: Original label

$h_0 \leftarrow f(\mathbf{x})$, $l_{\text{adv}} \leftarrow \infty$, $i \leftarrow 0$
success $= 0$
while do $l_{\text{adv}} \neq 0$ and $i \leq$ MaxIter
  $\bar{\mathbf{a}} \leftarrow E(\mathbf{a})$
  $h_i \leftarrow f(G(\mathbf{x}_i, \bar{\mathbf{a}}_i))$
  $l_{\text{adv}} \leftarrow L_{\text{adv}}(\mathbf{y}, h_i)$
  $\mathbf{a}_{i+1} \leftarrow \text{BackProp}\{\mathbf{a}_i, \nabla l_{\text{adv}}(f(G(\mathbf{x}_i, E(\mathbf{a}_i))))\}$
  $\tilde{\mathbf{x}} \leftarrow G(\mathbf{x}, E(\mathbf{a}_{i+1}))$
  if $f(\mathbf{x}) \neq f(\tilde{\mathbf{x}})$ **then**
    return success, $\tilde{\mathbf{x}}$
  end if
  $i \leftarrow i + 1$
end while

Architecture

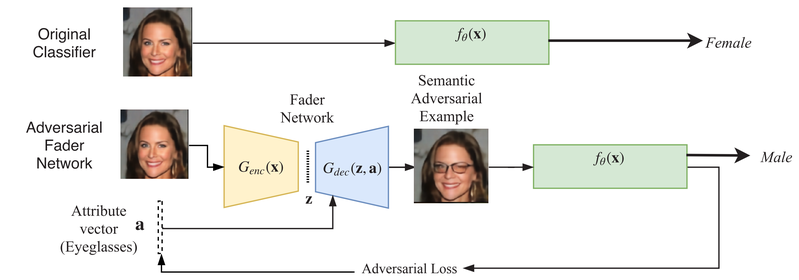

Figure 1: *Our architecture. The Fader Network is a pretrained generative model that decouples attributes such as glasses from the underlying identity (represented by* $\mathbf{a}$*).*

Modified Carlini-Wagner Adversarial Loss: $\max\left(0, \max_{t \neq i}(f(\tilde{\mathbf{x}})_t) - f(\tilde{\mathbf{x}})_i\right)$ s.t. $\tilde{\mathbf{x}} = G(\mathbf{x}, \mathbf{a})$

Qualitative Results

Figure 2: *Semantic adversarial examples generated with multiple attribute models.*

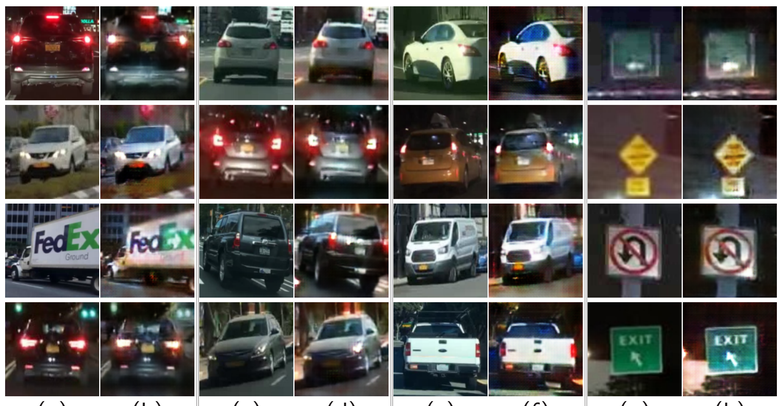

Figure 3: *Semantic adversarial examples produced by Attribute GAN trained on Time of Day labels from BDD dataset.*

Analysis: Dimensionality of Attack Space

Data model: $\tilde{\mathbf{x}} \sim$ Mixture of Gaussians with parameters $(\theta^\star, \sigma)$
Classifier: Well trained linear classifier, i.e. $\langle \hat{\boldsymbol{w}} \rangle \theta^\star \leq \epsilon$
Attack: Subspace attack with dimensionality, k; $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta$; $rank\mathbf{U} = k$.

Theorem (Robust classification error for subspace attacks)

Let $\hat{\boldsymbol{w}}$ be such that $\langle \hat{\boldsymbol{w}}, \theta^\star \rangle \geq k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon$. Then, the linear classifier $f_{\hat{\boldsymbol{w}}}$ has a $\mathcal{S}_\epsilon$-robust classification error upper bounded as:

$$\beta \leq \exp\left(-\frac{\left(\langle \hat{\boldsymbol{w}}, \theta^\star \rangle - k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon\right)^2}{2\sigma^2}\right) \quad (1)$$

Empirical Analysis

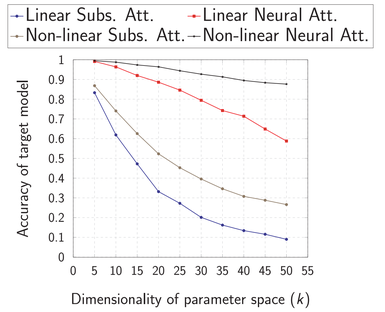

Figure 4: *Effect of dimensionality of the parametric attack space.*

Dataset: Mixture of 10 Gaussians Classifier: 2-layer binary classifier

- Subspace attack: $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta$
- Neural Attack: $\tilde{\mathbf{x}} = \mathbf{U}\text{diag}(\delta)\mathbf{U}^T$
- Subspace attacks ≈ Pixel space additive attacks
- Neural attacks ≈ Semantic attacks

Quantitative Results

Table 1: *Performance of the our approach when compared to random sampling of the attribute space.*

Attack Type	No. of Attributes	Accuracy of target model (%)	Random Sampling (%)
Single Att. Attack	1	52.0	89.00
	1	35.0	96.00
	1	14.0	90.00
Multi Att. Attack	3	3.00	89.00
	3	1.00	81.00
	3	3.00	87.00
Cascaded Multi Att. Attack	3	18.00	55.6
	3	20.00	93.00
Multi Att. AttGAN Attack	5	70.40	32.80
	5	39.40	40.40

Table 2: *Comparison of adversarial attacks with other attack strategies.*

Attack ($\epsilon = 1.74$)	Accuracy(%)
Single Att. Semantic Attack	**14.01**
Multi Att. Semantic Attack	**1.00**
FGSM	91.6
PGD	26.2
CW-l_2	0.00
Spatial Transformation Attacks	41.00

Acknowledgements

This work was supported in part by NSF, DARPA, AFOSR, ISU, a GPU gift grant from NVIDIA corporation, and faculty fellowships from the Black and Veatch Foundation.

For more related work: DICE Lab — Self-aware Complex Systems Laboratory