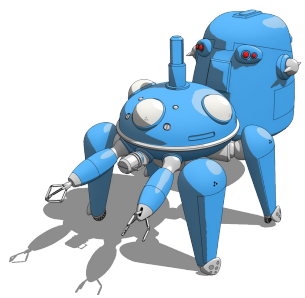
On the Machine Illusion

Empirical Study on Adversarial Samples

Zhitao Gong

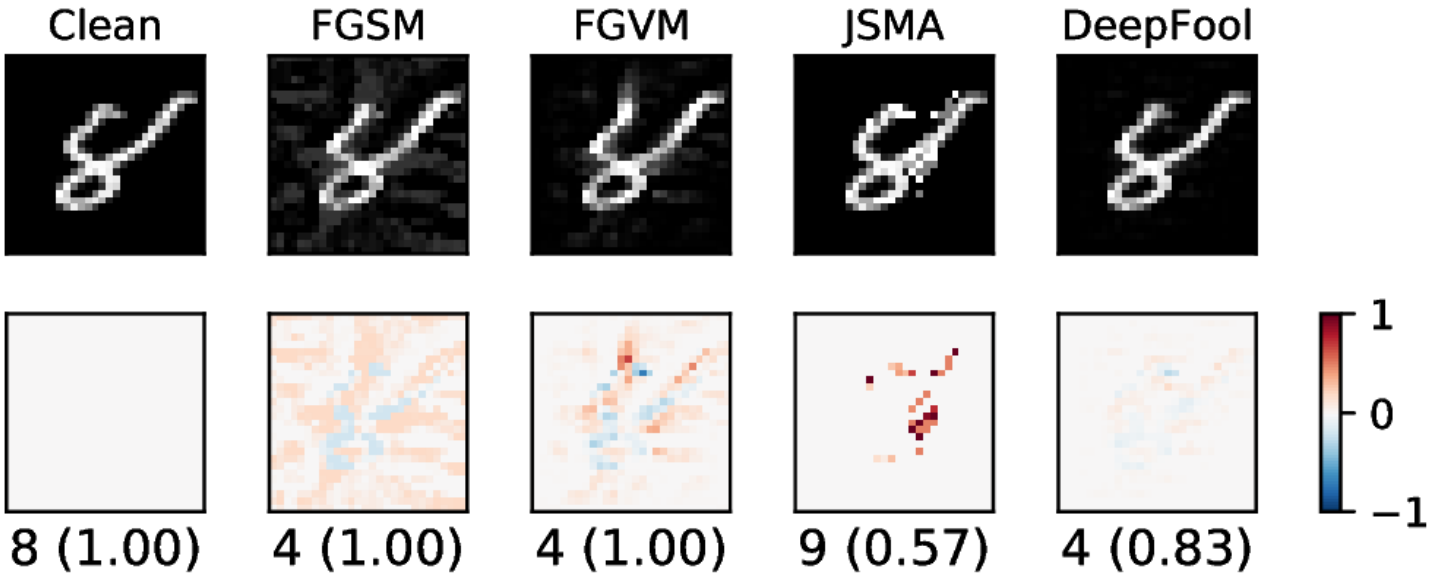
Auburn University

March 26, 2019



Outline

Adversarial Samples I



1. Noises are very subtle, visually indistinguishable.
2. Trick machines into wrong predictions with high confidence.

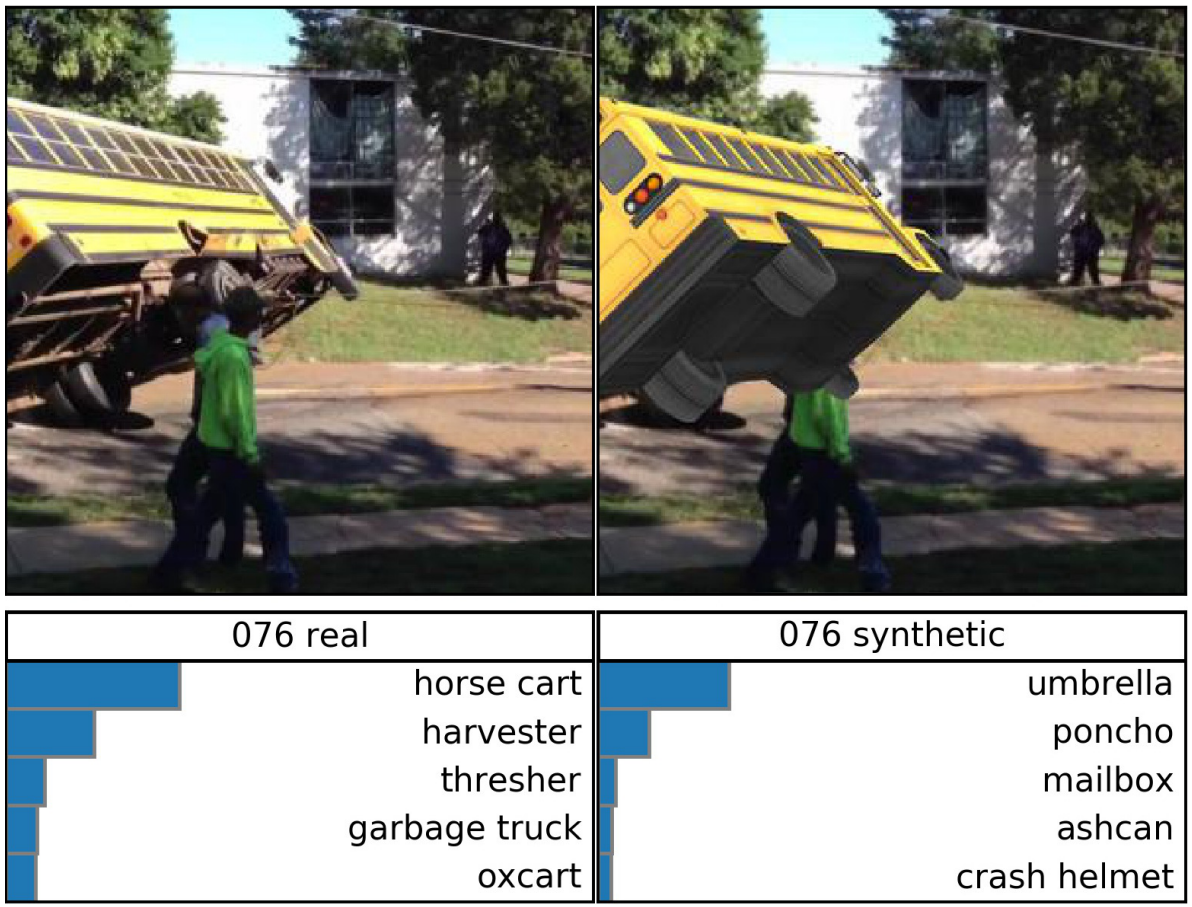
Adversarial Samples II

Clean Text	Label	WMD (n/L)	Adversarial Text
Quick summary of the book : **[. . .]** The book was n't bad , but was soooooooo cliché < br / > < br / > Now about the movie **[. . .]** (IMDB)	0→1	0.0317 (0.0050)	Quick summary of the book : **[. . .]** The book was n't bad , but was soooooooo TahitiNut < br / > < br / > Now about the movie **[. . .]**
zulchzulu < SM > TO OFFER SPECIAL DIVIDEND Southmark Corp said it will issue its shareholders a special dividend right **[. . .]** (REUTERS-2)	1→0	0.0817 (0.0125)	zulchzulu < SM > TO OFFER OFFERS SHARES Southmark Corp said it will issue its shareh olders a special dividend right **[. . .]**
U . K . MONEY MARKET GIVEN FURTHER 68 MLN STG HELP The Bank of England said it provided the market with a further **[. . .]** (REUTERS-5)	3→2	0.0556 (0.0077)	U . K . MONEY MARKET GIVEN FURTHER 68 ARL STG HELP The Bank of England said it provided the market with a further **[. . .]**

The highlighted words are changed. n/L is the number of words changed divided by the total number of words. (credit: [Gon+18])

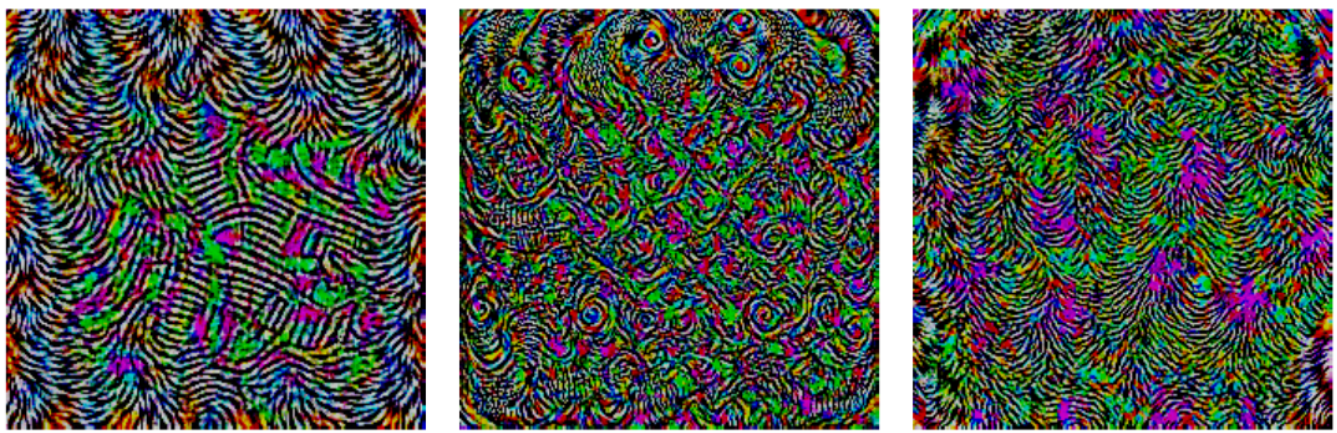
Adversarial Samples III

Objects in weird poses are also tricky! (credit: [Alc+18])



Adversarial Patterns for Machines

(a) CaffeNet (b) VGG-F (c) VGG-16

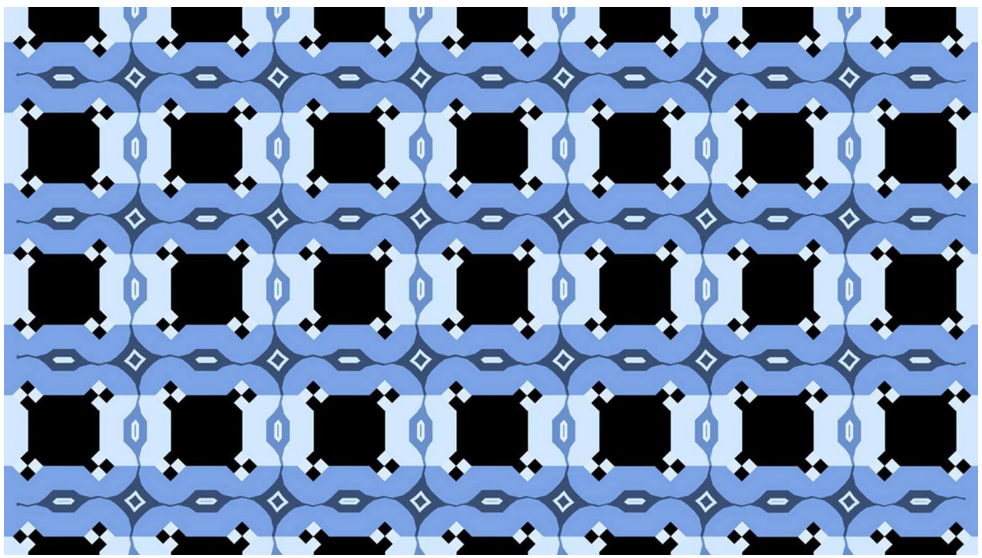


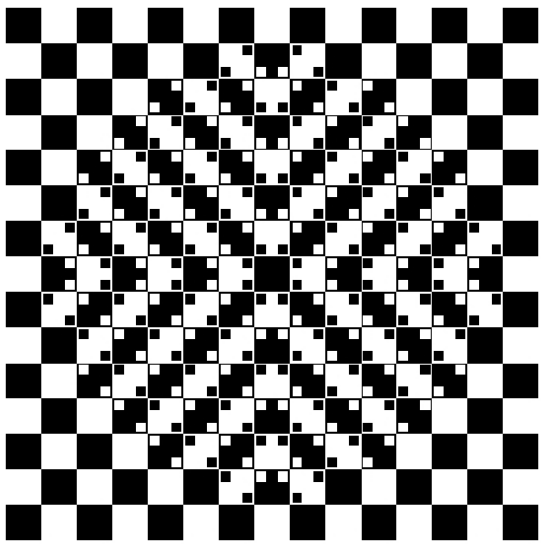
(d) VGG-19 (e) GoogLeNet (f) ResNet-152

Figure: Adversarial patterns for different neural nets [Moo+16].

Adversarial Patterns for Humans

Illusion possibly caused by the fringed edges [KPB04]. More examples http://www.psy.ritsumei.ac.jp/~akitaoka





Motivation

This phenomenon is interesting both in practice and in theory.

1. It undermines the models' reliability.
2. Hard to ignore due to it being transferable and universal.
3. It provides new insights into neural networks:
   - Local generalization does not seem to hold.
   - Data distribution: they appear in dense regions.
   - Trade-off between robustness and generalization.
   - ...

Outline

Neural Networks

It is a connectionist model.

1. Any state can be described as an N-dimensional vector of numeric activation values over neural units in a network.
2. Memory is created by modifying the strength of the connections between neural units.

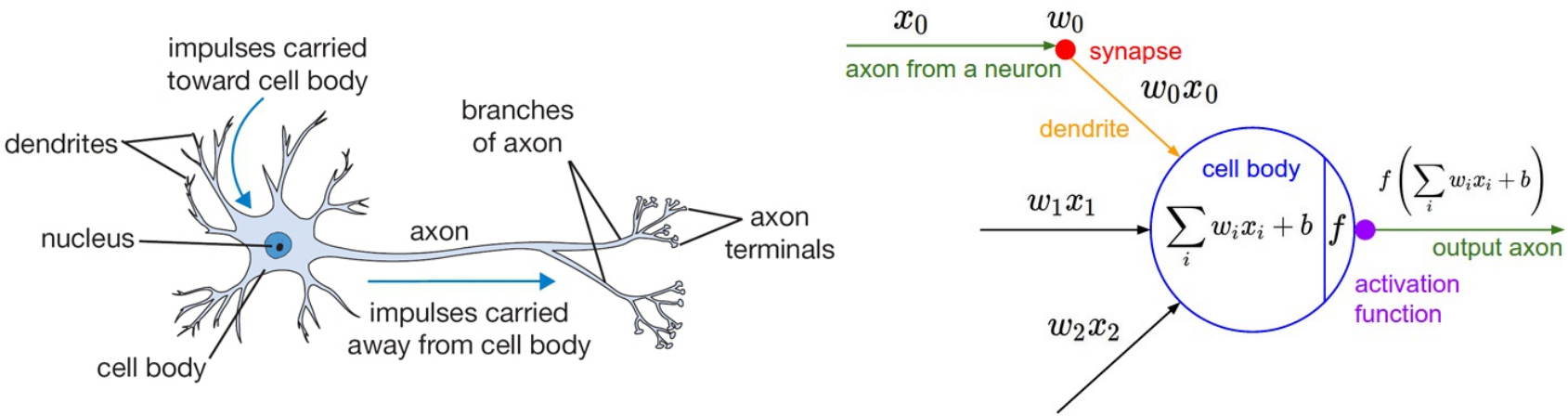


Figure: Biological neuron versus neuron model (credit: cs231n)

Case Study: Multi-Layer Perceptron (MLP)

MLP is one of the most simple feedfoward architectures.

1. Each neuron outputs to the neurons in the next layer.
2. Neurons in the same layer have no connections.

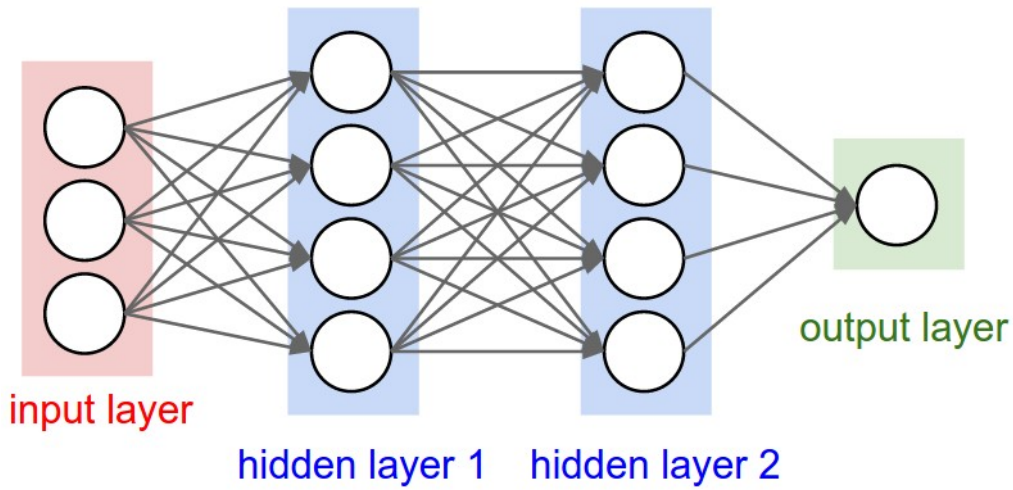


Figure: Multi-layer perceptron (credit: cs231n)

Case Study: Convolutional Neural Network (CNN)

CNN is inspired by eye structure, widely used in computer vision.

1. Each neuron receives inputs from a pool of neurons in previous layer, just like the convolution operation.
2. Neurons in the same layer have no connections

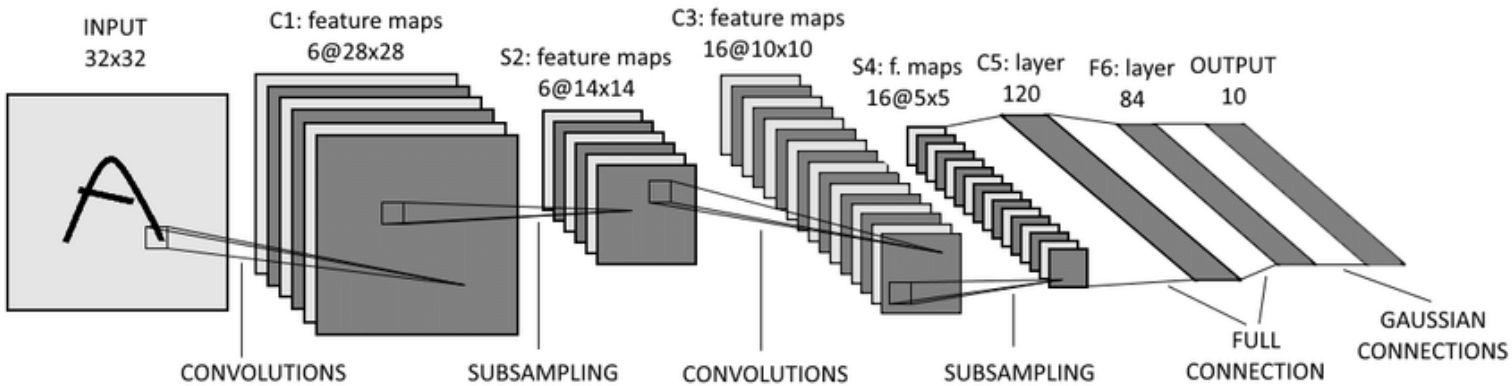


Figure: LetNet-5 [LeC+98]

Case Study: Recurrent Neural Network (RNN)

Some neurons get part of input from its output.

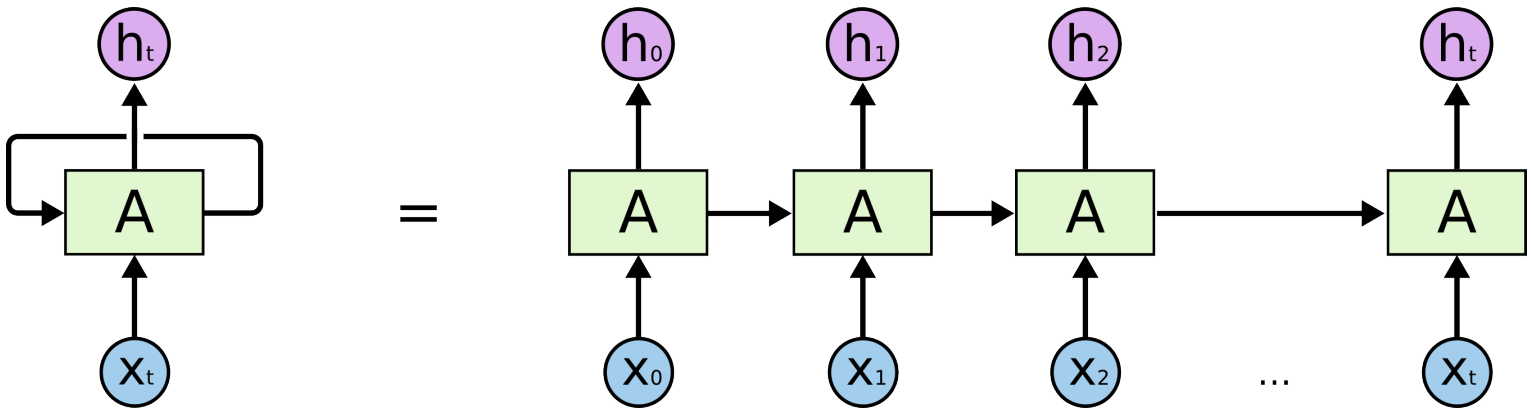


Figure: Dynamic unrolling of recurrent cells. (credit: colah's blog)

Case Study: Recurrent Neural Network (RNN)

Some neurons get part of input from its output.

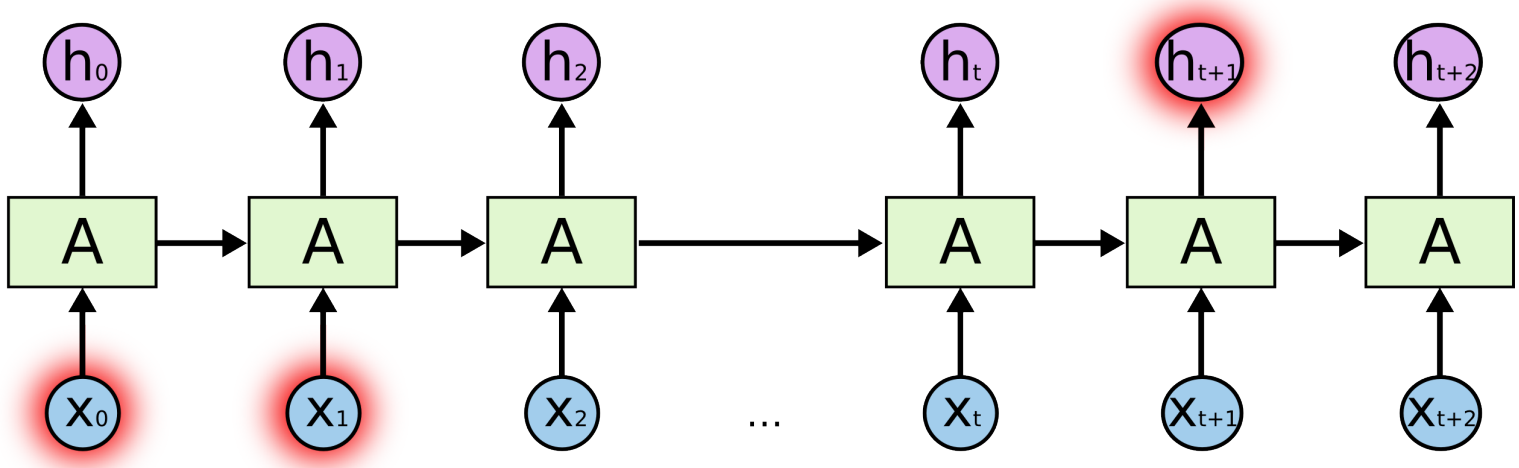


Figure: The double-edged sword: long term dependencies between outputs and inputs. (credit: colah's blog)

Generate Adversarial Images

Intuitions behind the adversarial methods

1. Move the data points
   - towards the decision boundary [MFF15; Moo+16],
   - in the direction where loss increases for the clean samples [GSS14; KGB16a], or decreases for the adversarial samples [Sze+13], or
   - where the probability of the correct label increases or the probability of the target label increases [Pap+15; CW16].
2. Map between clean and adversarial data points [ZDS17; BF17; Xia+18].

Intuition

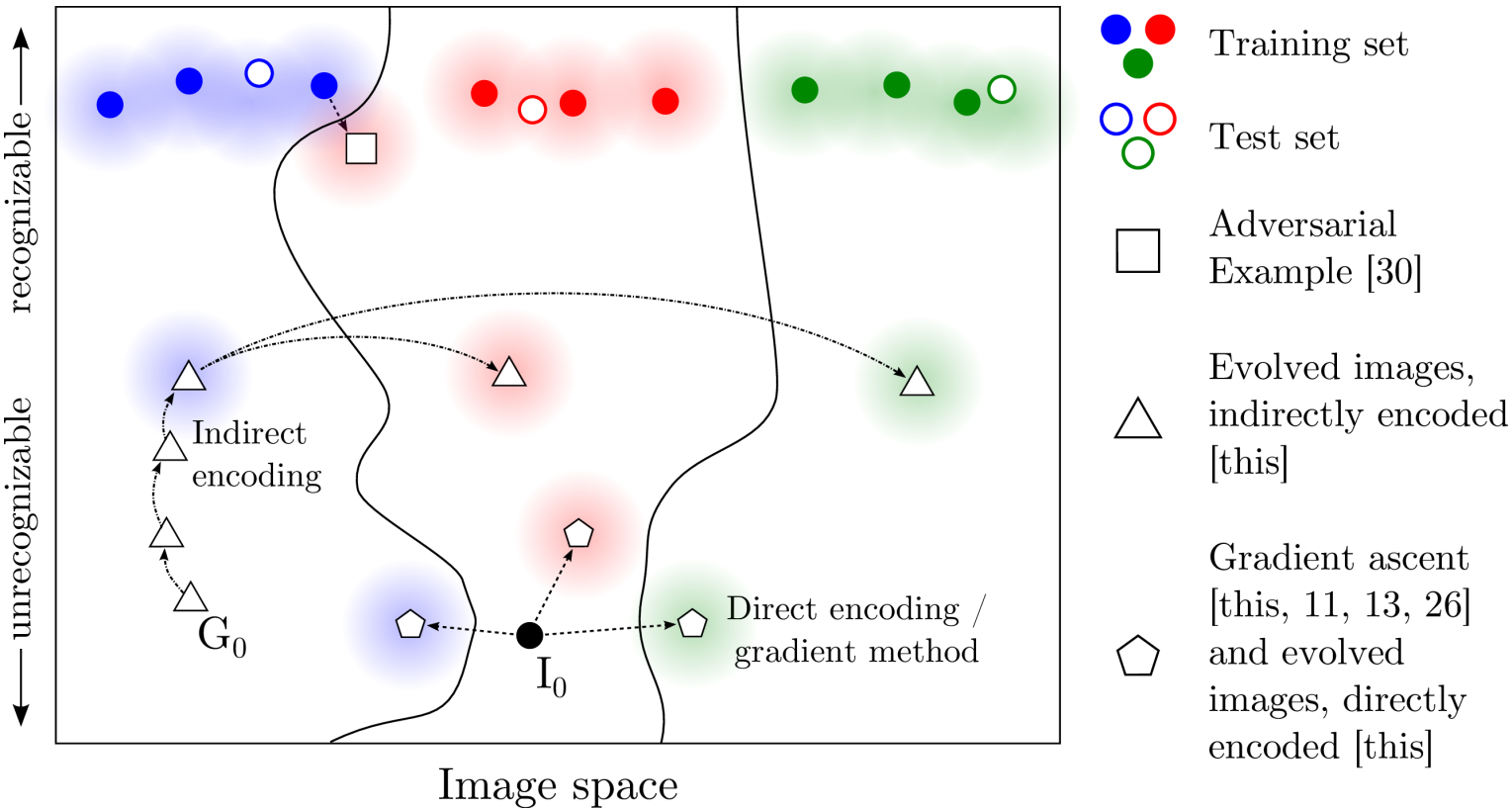


Figure: Data space hypothesis [NYC14]

Outline

Gist

We investigate binary classifier as a defense method

- Recognizes adversarial samples of the same distribution.
- Does not generalize to arbitrary adversarial samples.

Related Work

Adversarial training Augment training data with adversarial samples [GSS14; Mad+17].

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{x \in \mathcal{X}} \left[\max_{\delta \in [-\epsilon, \epsilon]^N} L(x + \delta; f_\theta) \right]$$

Preprocess Transform input images, e.g., denoising [Xie+18; Lia+17b], compression [Pra+18], quilting [Guo+17].

Detecting classifier [Met+17], density ratio estimation [Gon17].

Binary Classifier as A Defense

We propose to use a binary classifier to separate adversarial samples from clean ones [GWK17] based on the following observations:

1. The adversarial noise follows a specific direction [GSS14].
2. The neural nets are sensitive to individual pixel values [Sze+13].

Code: https://github.com/gongzhitaao/adversarial-classifier

Adversarial Examples

Dataset	X	$\widetilde{X}$
MNIST	0.9914	0.0213
CIFAR-10	0.8279	0.1500
SVHN	0.9378	0.2453

Table: The target model accuracy.

Classifier Efficiency and Robustness

Dataset	X	$\widetilde{X}_f$	$\widetilde{X}_g$	$\{\widetilde{X}_f\}_g$
MNIST	1.00	1.00	0.00	1.00
CIFAR-10	0.99	1.00	0.01	1.00
SVHN	1.00	1.00	0.00	1.00

Table: The classifier g's accuracy. f is the target model. And $\widetilde{X}_f$ denotes adversarial examples targeting model f.

1. X and $\widetilde{X}_f$ columns shows the classifier g is effective.
2. $\widetilde{X}_g$ and $\{\widetilde{X}_f\}_g$ columns shows the classifier g is robust.

Problem with Classifier Defense

Limitation: different hyperparameters, different adversarial algorithms may elude the binary classifier or adversarial training.

ϵ	X	$\widetilde{X}$
0.3	0.9996	1.0000
0.1	0.9996	1.0000
0.03	0.9996	0.9997
0.01	0.9996	0.0030

Table: The binary classifier, trained with FGSM adversarials with $\epsilon = 0.03$, is unable to recognize the adversarials with $\epsilon = 0.01$ (more subtle noise).

Problem with Adversarial Training

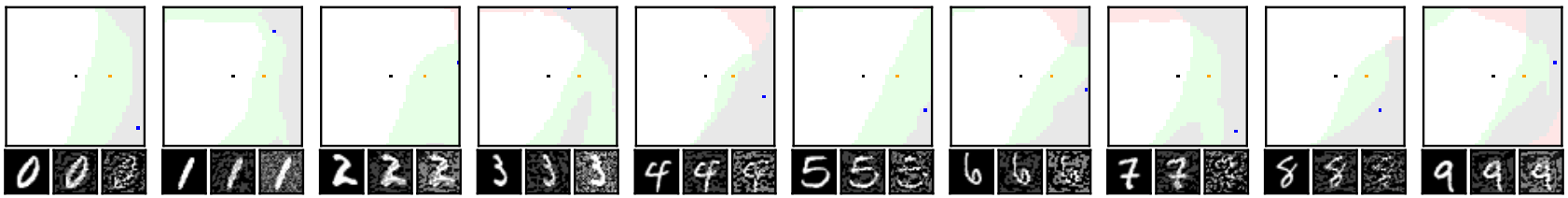


Figure: Adversarial training [Hua+15; KGB16b] is not sufficient. In the church window plot [WG16], each pixel (i, j) is a data point $\widetilde{x}$ such that $\widetilde{x} = x + \mathbf{h}\epsilon_j + \mathbf{v}\epsilon_i$, where $\mathbf{h}$ is the FGSM direction and $\mathbf{v}$ is a random orthogonal direction. The ϵ ranges from $[-0.5, 0.5]$. (credit: [GWK17])

1. ☐ (☐) always correct (incorrectly).
2. ☐ correct with adversarial training.
3. ☐ correct without adversarial training.

Outline

Text Embedding Layer

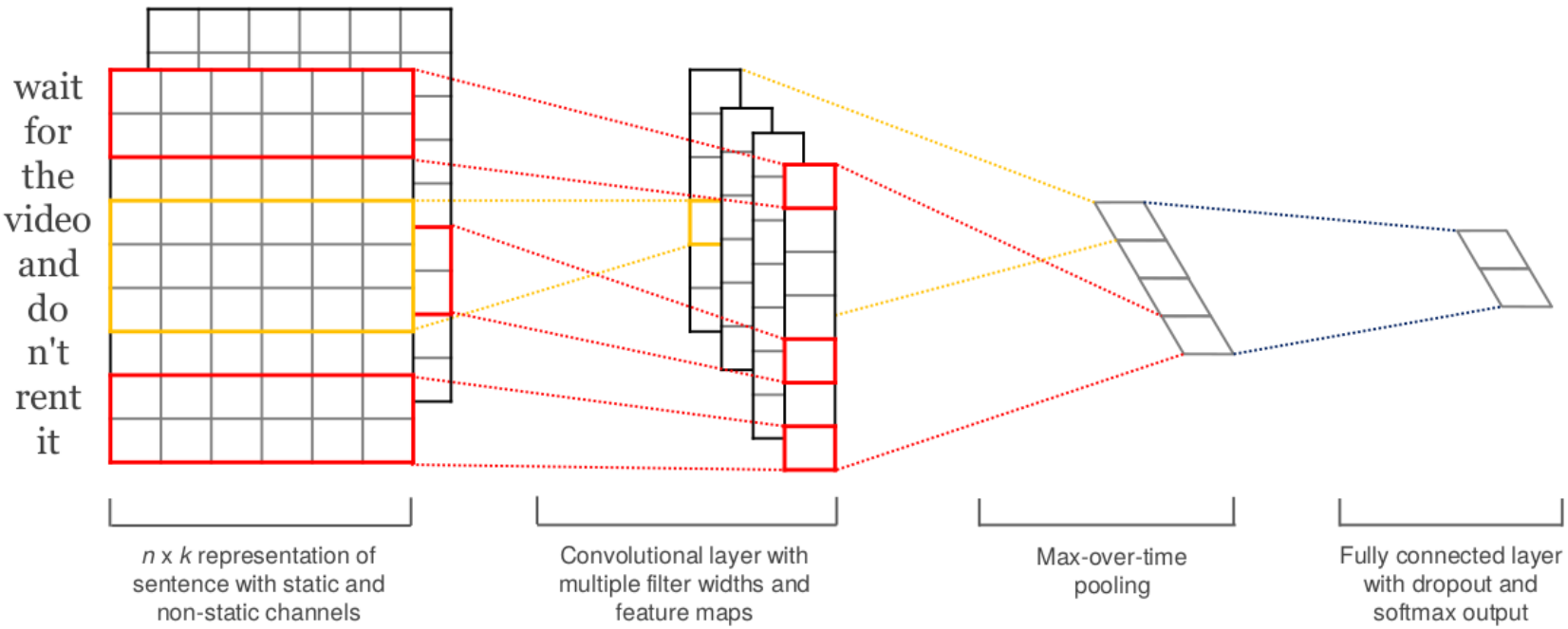


Figure: Architecture for sentence classification with CNN [Kim14]

Text Embedding Example

"wait for the video" $\xrightarrow{\text{tokenize}}$ ["wait", "for", "the", "video"] $\xrightarrow{\text{indexer}}$ [2, 20, 34, 8] $\xrightarrow{\text{embedding}}$ $\mathbb{R}^{4\times D}$, where D is the embedding size.

- Each sentence with be converted to $\mathbb{R}^{L\times D}$ before being fed into the convolution layer, where L is the sentence length.
- We usually truncate/pad sentences to the same length so that we could do *batch training*.
- Embedding may also be on the character-level.

Problem Overview

Difficulties we face:

1. The text space is discrete. Moving the data points in small steps following a certain direction does not work, directly.
2. Text quality is hard to measure. *Much to learn, you still have* (the Yoda-style) v.s. *You still have much to learn* (the mundane-style)

General directions:

1. Three basic operations are available, *replacement*, *insertion*, and *deletion*.
2. They may work at character, word or sentence level.

Methods

In text space This class of methods need to solve two problems:

1. what to change, e.g., random, ∇L [Lia+17a], manually picking [SM17].
2. change to what, e.g., random, synonyms [SM17] or nearest neighbors in embedding space, or forged facts [JL17; Lia+17a].

In latent space GAN [Goo+14] is used to map from a latent space (e.g., Gaussian noise) to sentences [ZDS17].

Adversarial Text Framework

We propose another method in the embedding space.

GENERATE-ADVERSARIAL-TEXTS(f, x)

1 **for** $i = 1$ **to** $x.length$
2 $\quad z_i = \text{EMBEDDING}(x_i)$
3 $z' = \text{ADV}(f, z)$
4 **for** $i = 1$ **to** $z'.length$
5 $\quad x'_i = \text{NEAREST-EMBEDDING}(z'_i)$
6 $\quad s_i = \text{REVERSE-EMBEDDING}(x'_i)$
7 **return** s

Assumptions:

1. The text embedding space preserve the semantic relations.
2. Important features get more noise.

Result: https://github.com/gongzhitaao/adversarial-text

Results On Word-Level

Method	Dataset	Accuracy				
		ϵ	0.40	0.35	0.30	0.25
FGSM	IMDB		0.1334	0.1990	0.4074	0.6770
	Reuters-2		0.6495	0.7928	0.9110	0.9680
	Reuters-5		0.5880	0.7162	0.7949	0.8462
		ϵ	15	30	50	100
FGVM	IMDB		0.8538	0.8354	0.8207	0.7964
	Reuters-2		0.7990	0.7538	0.7156	0.6523
	Reuters-5		0.7983	0.6872	0.6085	0.5111
		ϵ	20	30	40	50
DeepFool	IMDB		0.8298	0.7225	0.6678	0.6416
	Reuters-2		0.6766	0.5236	0.4910	0.4715
	Reuters-5		0.4034	0.2222	0.1641	0.1402

Table: Word-level CNN accuracy under different parameter settings. ϵ is the noise scaling factor.

Case Study: DeepFool I

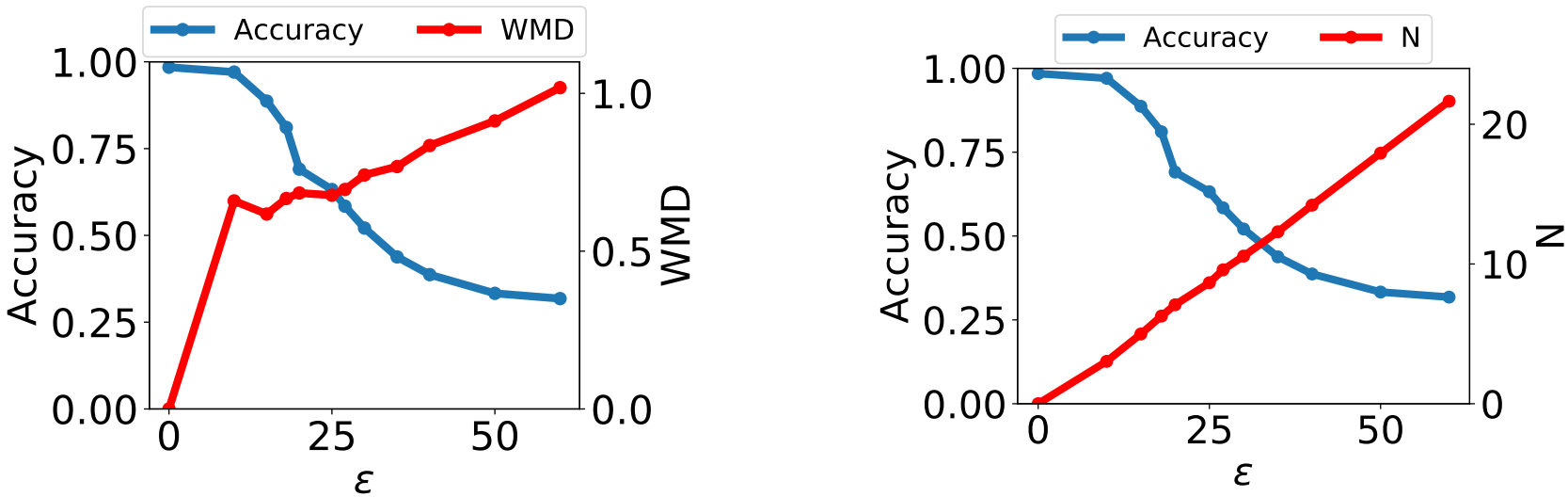


Figure: Word-level model's accuracy with varying DeepFool overshoot value. The WMD and N (number of words changed) empirically show the quality of the adversarial texts. (credit: [Gon+18])

Case Study: DeepFool II

0.0556 1 (0.77%)	U . K . MONEY MARKET GIVEN FURTHER 68 ~~MLN~~ ARL STG HELP The Bank of England said it provided the market with a further 68 mln stg assistance this afternoon , **[...]**
0.1762 8 (2.93%)	GERMAN BANKERS ' REMARKS ~~REVIVE~~ AXIOM TALK OF ~~RATE~~ FROM CUT **[...]** discussion , ; currency dealers said . **[...]** required cuts in interest ~~rates~~ higher . Separately , West Berlin state central bank president Dieter ~~Hiss~~ Thein told journalists **[...]** forecast ~~on~~ out interest rates , however . **[...]** It allocated 6 . 1 billion marks in new liquidity , much less than the ~~14 .~~ 9 ... 9 billion leaving the market as a prior pact expired . **[...]** Koehler said in a speech in **[...]** regardless of whether central ~~banks~~ banking intervened or ~~exchange~~ forex rates fell . " **[...]**
0.4406 12 (6.14%)	ITALY ~~DEFICIT~~ NOT DUE TO ~~LIBERALIZATION~~ SOUNDSCAPE - ~~MINISTER~~ GOVERNOR Italy ' s Foreign Trade Minister ~~Mario~~ Nintendo ~~<unk>~~ monk,commenting on speculation in the Italian press , said a sharp balance of payments ~~deficit~~ spending in May **[...]** - bearing deposits ~~on~~ down foreign securities purchases . " The deficit can be better attributed to premature and delayed foreign ~~trade~~ five payments and receipts (leads and lags) rather than capital outflow to portfolio investment ~~<unk>~~ Hippo_Tron ~~<unk>~~ Libera_me said in a statement . **[...]** " non - banking ~~capital~~ investment ~~outflows~~ cashflows <unk> " In practice , it seems that there has been a constant flow of capital to foreign securities or investments outside our borders ~~<unk>~~ cat_girl25 said the newspaper . **[...]**

Figure: Adversarial texts sample from Reuters-5 dataset. Original is the original token, replaced is the adversarial token. **[...]** denotes omitted tokens due to space constraint. $\epsilon = 50$ in DeepFool.

More results: https://gongzhitaao.org/adversarial-text

Transferability

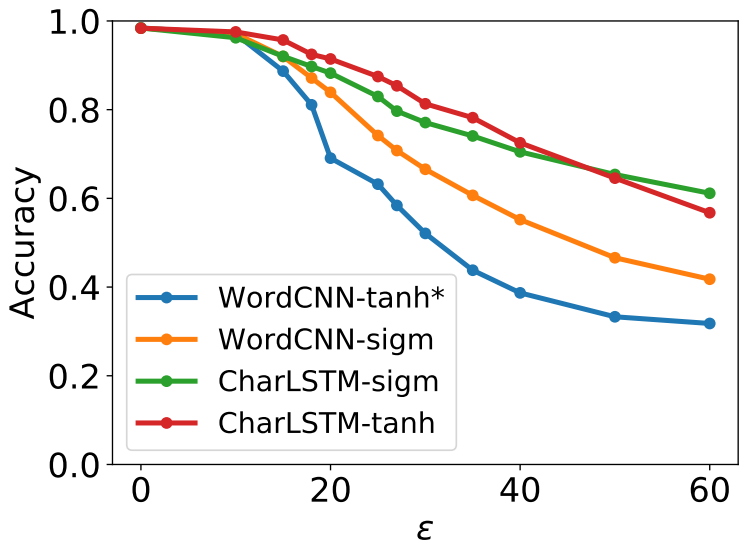


Figure: Transferability of adversarial texts generated via our framework on word-level.

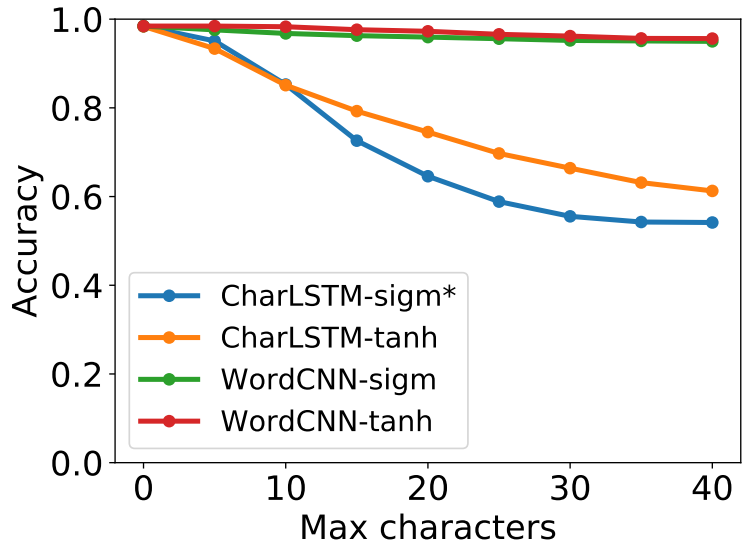
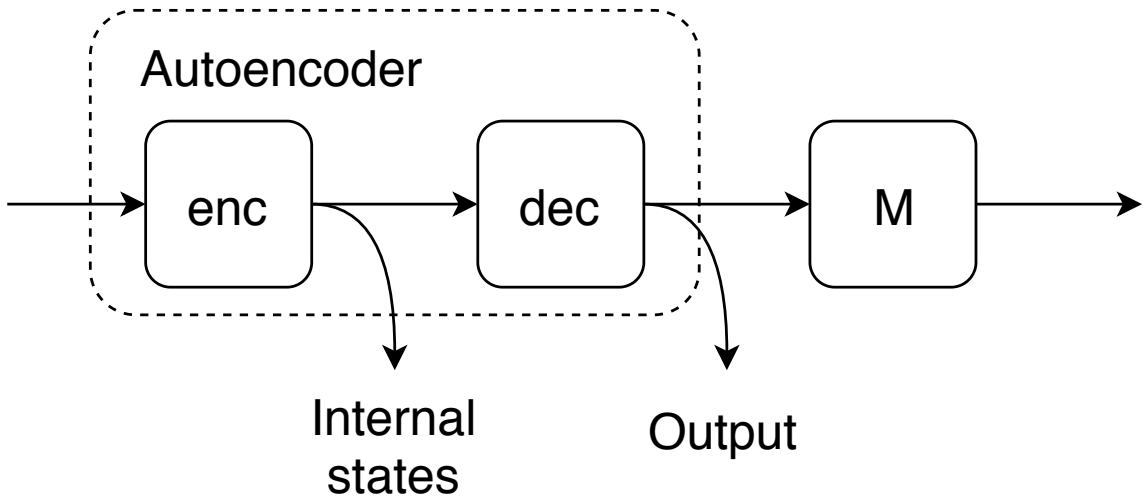


Figure: Transferability of adversarial texts generated via Hotflip on character-level.

* denotes the target model. (credit: [Gon+18])

Future Work

- Evaluation metrics.
- Nearest neighbor search is non-differentiable.



Outline

Overview

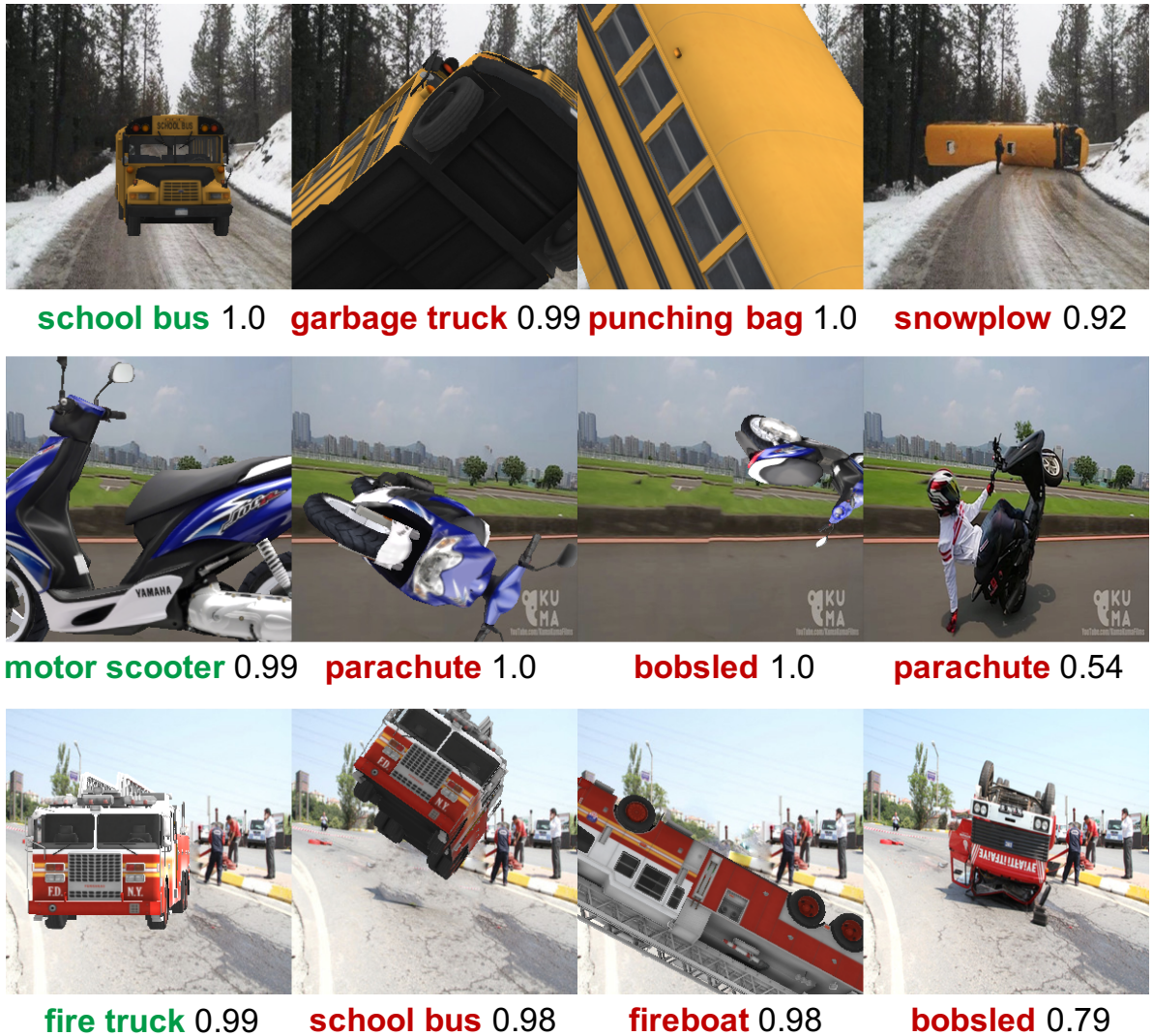


Figure: Objects in weird poses. (credit: [Alc+18])

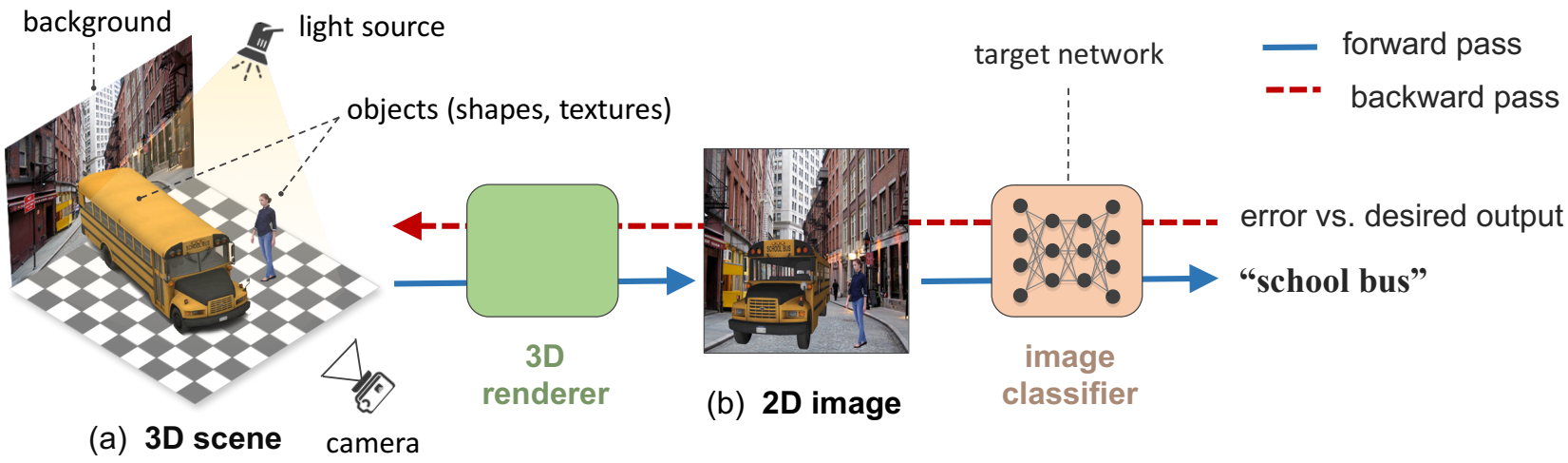
Gist

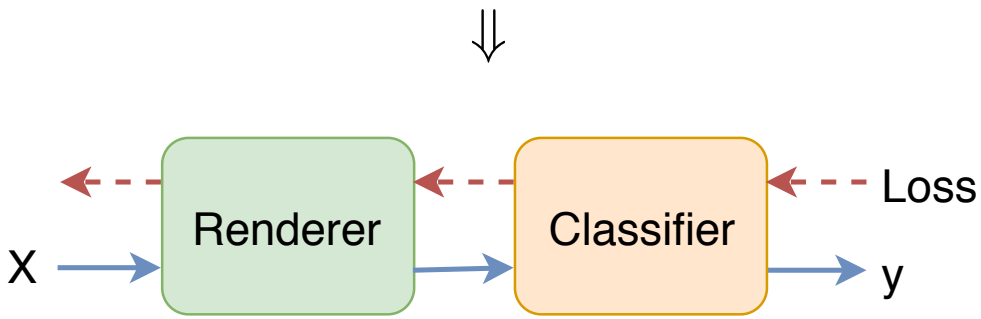
A *descriptive* study on the adversarial pose properties:

1. Effectiveness. Only 3% are correctly recognized.
2. Imperceptible. Small rotation (10.30° in yaw) results in an adversarial sample.
3. Good transferability. 99% against Inception-v3 transfer to AlexNet and ResNet-50, 75% transfer to YOLO-v3.
4. Adversarial training is not helpful.

Intuition: https://gongzhitaao.org/strike-with-a-pose

Framework





X pose parameters, 6D, $(x, y, z, \theta_x, \theta_y, \theta_z)$

y prediction, a probability distribution over all labels.

Methods

Random search Randomly sample the 6D space.

Gradient descent

$$X_{k+1} = X_k + \nabla_{X_k} L(y_k, \widetilde{y})$$

- Differentiable renderer, neural renderer [KUH18]
- Non-differentiable renderer, ModernGL [Dom19]

Random Search

The distributions of each pose parameters for high-confidence ($p \geq 0.7$) correct/wrong classifications. (credit: [Alc+18])

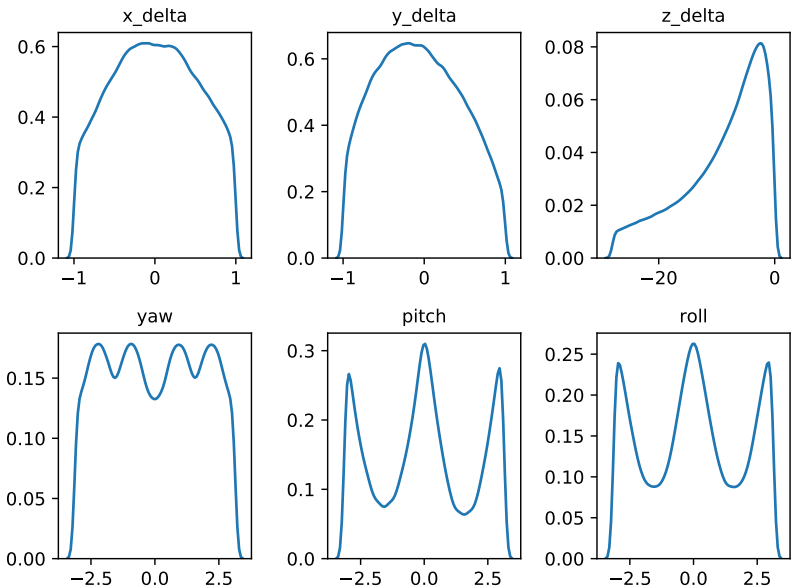


Figure: Correct

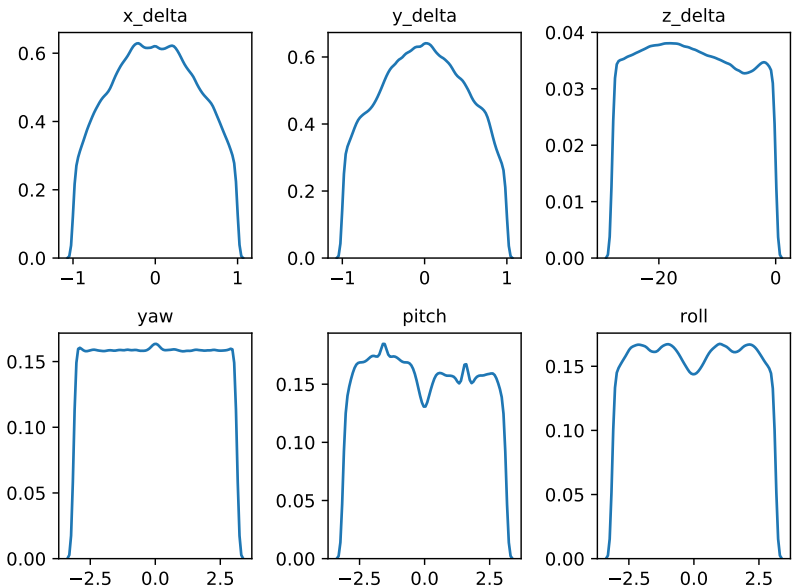


Figure: Wrong

Random Search

The distributions of each pose parameters for high-confidence ($p \geq 0.7$) correct/wrong classifications. (credit: [Alc+18])

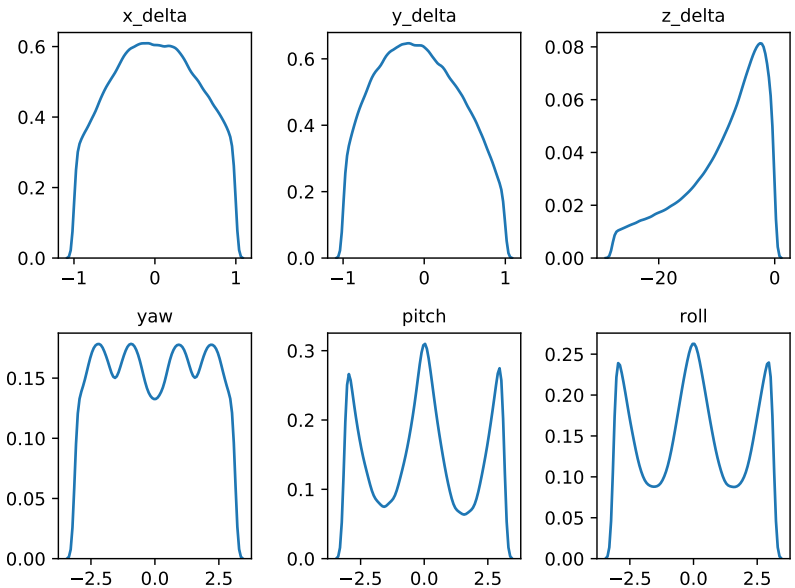


Figure: Correct

Parameter	Fail %	Δ_{min}
x_δ	42	2.0
y_δ	49	4.5
z_δ	81	5.4%
θ_y	69	10.31°
θ_p	83	8.02°
θ_r	81	9.17°

Methods Comparison

ZRS: z-focused random search
FD-G: finite difference approximated gradient
DR-G: differentiable renderer

	Hit Rate %	Target Probability
ZRS	78	0.29
FD-G	92	0.41
DR-G†	32	0.22

Problem with Adversarial Training (again)

PT: AlexNet trained with vanilla ImageNet
AT: training data augmented with adversarial samples

	Error	PT	AT
All	Train	99.67	6.7
	Test	99.81	89.2
$p \geq 0.7$	Train	87.8	1.9
	Test	48.2	33.3

Conclusion: adversarial training does not help models generalize to unseen adversarial samples.

Outline

Summary

1. Binary classifier as a defense is effective and limited.
2. Text adversarials are also not difficult to generate.
3. Objects in weird poses are also difficult for neural nets.

Future Work

Image credit [Kar16]

Machine detects

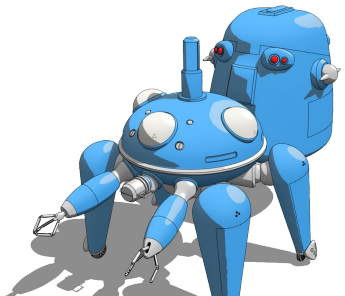
- objects
- faces
- figure components
- ...

Future Work

Image credit [Kar16]

Cannot understand

- mirror
- shadows
- jokes
- ...



Outline

[Alc+18] Michael A. Alcorn et al. "Strike (with) a Pose: Neural Networks Are Easily Fooled By Strange Poses of Familiar Objects". In: *CoRR* abs/1811.11553 (2018). arXiv: 1811.11553. URL: http://arxiv.org/abs/1811.11553.

[BF17] Shumeet Baluja and Ian Fischer. "Adversarial Transformation Networks: Learning To Generate Adversarial Examples". In: *CoRR* abs/1703.09387 (2017). URL: http://arxiv.org/abs/1703.09387.

[CW16] Nicholas Carlini and David Wagner. "Towards Evaluating the Robustness of Neural Networks". In: *CoRR* abs/1608.04644 (2016). URL: http://arxiv.org/abs/1608.04644.

[Dom19] Szabolcs Dombi. *ModernGL - ModernGL 5.4.1 documentation*. https://moderngl.readthedocs.io/en/stable/index.html. (Accessed on 11/14/2018). 2019.

[Gon+18] Zhitao Gong et al. "Adversarial Texts With Gradient Methods". In: *arXiv e-prints*, arXiv:1801.07175 (Jan. 2018), arXiv:1801.07175. arXiv: 1801.07175 [cs.CL].

[Gon17] Lovedeep Gondara. "Detecting Adversarial Samples Using Density Ratio Estimates". In: *arXiv preprint arXiv:1705.02224* (2017).

[Goo+14] I. J. Goodfellow et al. "Generative Adversarial Networks". In: *ArXiv e-prints* (June 2014). arXiv: 1406.2661 [stat.ML].

[GSS14] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ArXiv e-prints* (Dec. 2014). arXiv: 1412.6572 [stat.ML].

[Guo+17] C. Guo et al. "Countering Adversarial Images Using Input Transformations". In: *ArXiv e-prints* (Oct. 2017). arXiv: 1711.00117 [cs.CV].

[GWK17] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. "Adversarial and Clean Data Are Not Twins". In: *CoRR* abs/1704.04960 (2017).

[Hua+15] Ruitong Huang et al. "Learning With a Strong Adversary". In: *CoRR* abs/1511.03034 (2015). URL: http://arxiv.org/abs/1511.03034.

[JL17] Robin Jia and Percy Liang. "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *arXiv preprint arXiv:1707.07328* (2017).

[Kar16] Andrew Karparthy. "Connecting Images and Natural Language". Ph.D. dissertation. Stanford University, 2016. URL: https://cs.stanford.edu/people/karpathy/main.pdf.

[KGB16a] A. Kurakin, I. Goodfellow, and S. Bengio. "Adversarial Examples in the Physical world". In: *ArXiv e-prints* (July 2016). arXiv: 1607.02533 [cs.CV].

[KGB16b] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial Machine Learning At Scale". In: *CoRR* abs/1611.01236 (2016). URL: http://arxiv.org/abs/1611.01236.

[Kim14] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *CoRR* abs/1408.5882 (2014).

[KPB04] Akiyoshi Kitaoka, Baingio Pinna, and Gavin Brelstaff. "Contrast Polarities Determine the Direction of Café Wall Tilts". In: *Perception* 33.1 (2004), pp. 11–20.

[KUH18] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. "Neural 3D Mesh Renderer". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[LeC+98] Yann LeCun et al. "Gradient-Based Learning Applied To Document Recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Lia+17a] Bin Liang et al. "Deep Text Classification Can Be Fooled". In: *arXiv preprint arXiv:1704.08006* (2017).

[Lia+17b] Bin Liang et al. "Detecting Adversarial Examples in Deep Networks With Adaptive Noise Reduction". In: *arXiv preprint arXiv:1705.08378* (2017).

[Mad+17] Aleksander Madry et al. "Towards Deep Learning Models Resistant To Adversarial Attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[Met+17] Jan Hendrik Metzen et al. "On Detecting Adversarial Perturbations". In: *arXiv preprint arXiv:1702.04267* (2017).

[MFF15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a Simple and Accurate Method To Fool Deep Neural Networks". In: *CoRR* abs/1511.04599 (2015). arXiv: 1511.04599. URL: http://arxiv.org/abs/1511.04599.

[Moo+16] Seyed-Mohsen Moosavi-Dezfooli et al. "Universal Adversarial Perturbations". In: *arXiv preprint arXiv:1610.08401* (2016).

[NYC14] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images". In: *CoRR* abs/1412.1897 (2014). URL: http://arxiv.org/abs/1412.1897.

[Pap+15] Nicolas Papernot et al. "The Limitations of Deep Learning in Adversarial Settings". In: *CoRR* abs/1511.07528 (2015). URL: http://arxiv.org/abs/1511.07528.

[Pra+18] Aaditya Prakash et al. "Protecting JPEG Images Against Adversarial Attacks". In: *CoRR* abs/1803.00940 (2018). arXiv: 1803.00940. URL: http://arxiv.org/abs/1803.00940.

[SM17] Suranjana Samanta and Sameep Mehta. "Towards Crafting Text Adversarial Samples". In: *arXiv preprint arXiv:1707.02812* (2017).

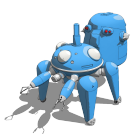
[Sze+13] Christian Szegedy et al. "Intriguing Properties of Neural Networks". In: *CoRR* abs/1312.6199 (2013). URL: http://arxiv.org/abs/1312.6199.

[WG16] D Warde-Farley and I Goodfellow. "Adversarial Perturbations of Deep Neural Networks". In: *Perturbation, Optimization and Statistics*. Ed. by Tamir Hazan, George Papandreou, and Daniel Tarlow. 2016.

[Xia+18] C. Xiao et al. "Generating Adversarial Examples With Adversarial Networks". In: *ArXiv e-prints* (Jan. 2018). arXiv: 1801.02610 [cs.CR].

[Xie+18] Cihang Xie et al. "Feature Denoising for Improving Adversarial Robustness". In: *CoRR* abs/1812.03411 (2018). arXiv: 1812.03411. URL: http://arxiv.org/abs/1812.03411.

[ZDS17] Z. Zhao, D. Dua, and S. Singh. "Generating Natural Adversarial Examples". In: *ArXiv e-prints* (Oct. 2017). arXiv: 1710.11342 [cs.LG].



by arrghman.deviantart.com @DeviantArt