

# Adversarial Machine Learning

A comprehensive overview

---

Abhinav Venkataraman

Samsung SDS Research America

# Table of contents

1. Introduction
2. Adversarial Image Generation
3. Titleformats
4. Elements
5. Conclusion

# Introduction

---

# What's an Adversarial Example?

Machine learning models that misclassify examples that are slightly different (sometimes even imperceptible from human eye) from correctly classified examples drawn from the data distribution.

# Problem Definition

## Regular Neural Network Training

Train a model on a dataset such that you take the gradient of loss function w.r.t model parameters. In this way, you maximize on the score of the correct class.

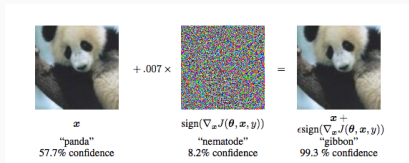
## Adversarial Learning

Generate an image by doing the following.

- Wiggle the pixel of an image in the direction of the loss function w.r.t to a class different from the target class. This perturbs the image by a tiny bit but the score of the target class is reduced.

Run the model on the generated image and see the classification result.

# Adversarial Example



- Digital images often use only 8 bits per pixel so they discard all information below  $\frac{1}{255}$  of the dynamic range.
- The classifier does not respond differently to an input  $x$  than to an adversarial input ( $\tilde{x} = x + \eta$ ) if every element of the perturbation is smaller than the precision of the features ( $\|\eta\|_\infty = \epsilon$ ).
- But then this perturbation causes the activation to grow by  $\epsilon mn$  times where  $m$  and  $n$  are dimensions of weight matrix.

# Adversarial Image Generation

---

# Neural Networks are linear too!

## Fooling CNNs

- Deep learning models are meant to express complex non-linear functions.



# Neural Networks are linear too!

## Fooling CNNs

- Deep learning models are meant to express complex non-linear functions.
- **how are these linear perturbations very effective??**

# Fast Gradient Sign Method

## Model Definition

Let  $\theta$  be model parameters,  $x$  be the input to the model and let  $y$  be the target associated with  $x$  then the loss for the model would be defined by  $L(\theta, x, y)$  .

We can obtain an adversarial example by having a perturbation in the following manner.

$$\eta = \epsilon \text{sign}(\nabla_x L(\theta, x, y))$$

The above gradient is computed through back propagation.

# Titleformats

---

**metropolis** supports 4 different titleformats:

- Regular
- SMALLCAPS
- ALLSMALLCAPS
- ALLCAPS

They can either be set at once for every title type or individually.

This frame uses the `smallcaps` titleformat.

## Potential Problems

Be aware, that not every font supports small caps. If for example you typeset your presentation with pdfTeX and the Computer Modern Sans Serif font, every text in smallcaps will be typeset with the Computer Modern Serif font instead.

This frame uses the `allsmallcaps` titleformat.

### Potential problems

As this titleformat also uses smallcaps you face the same problems as with the `smallcaps` titleformat. Additionally this format can cause some other problems. Please refer to the documentation if you consider using it.

As a rule of thumb: Just use it for plaintext-only titles.

This frame uses the `allcaps` titleformat.

## Potential Problems

This titleformat is not as problematic as the `allsmallcaps` format, but basically suffers from the same deficiencies. So please have a look at the documentation if you want to use it.

# Elements

---



The theme provides sensible defaults to  
`\emph{emphasize}` text, `\alert{accent}` parts  
or show `\textbf{bold}` results.

becomes

The theme provides sensible defaults to *emphasize* text, **accent** parts or  
show **bold** results.

# Font feature test

- Regular
- *Italic*
- SMALLCAPS
- **Bold**
- **Bold Italic**
- **Bold SmallCaps**
- Monospace
- *Monospace Italic*
- Monospace Bold
- *Monospace Bold Italic*

## Items

- Milk
- Eggs
- Potatos

## Enumerations

1. First,
2. Second and
3. Last.

## Descriptions

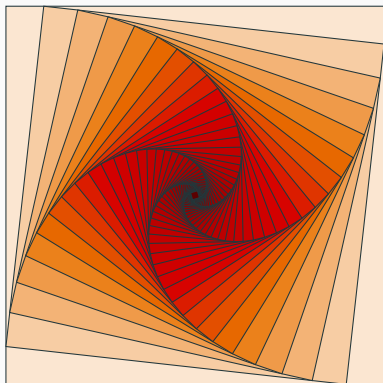
**PowerPoint** Meeh.  
**Beamer** Yeeeha.

- This is important

- This is important
- Now this

- This is important
- Now this
- And now this

- This is really important
- Now this
- And now this



**Figure 1:** Rotated square from texample.net.



**Table 1:** Largest cities in the world (source: Wikipedia)

City	Population
Mexico City	20,116,842
Shanghai	19,210,000
Peking	15,796,450
Istanbul	14,160,467

Three different block environments are pre-defined and may be styled with an optional background color.

## Default

Block content.

## Alert

Block content.

## Example

Block content.

## Default

Block content.

## Alert

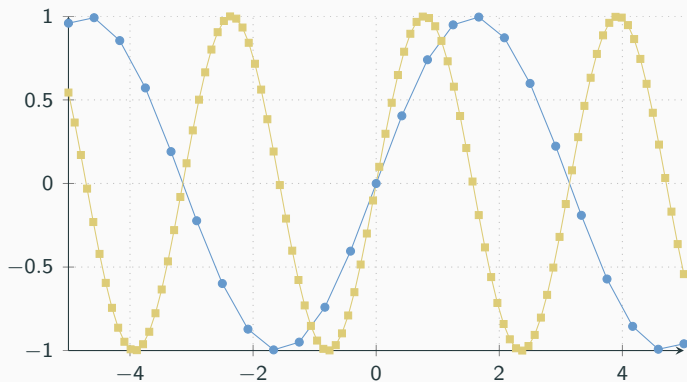
Block content.

## Example

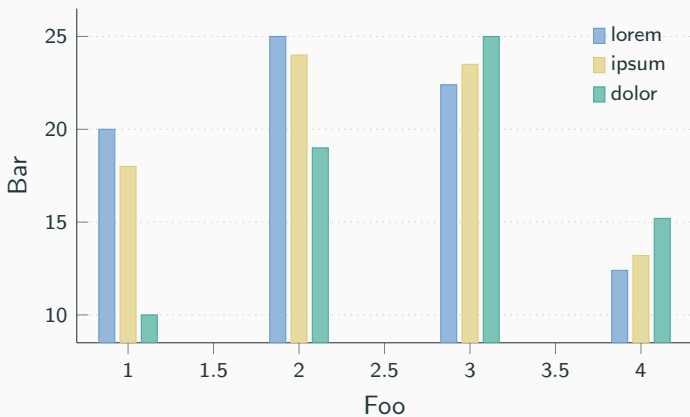
Block content.

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

# Line plots



# Bar charts



*Veni, Vidi, Vici*

**metropolis** defines a custom beamer template to add a text to the footer. It can be set via

```
\setbeamertemplate{frame footer}{My custom footer}
```

Some references to showcase `[allowframebreaks]` [4, 2, 5, 1, 3]



# Conclusion

---

Get the source of this theme and the demo presentation from

`github.com/matze/mtheme`

The theme *itself* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.



**Questions?**

# Backup slides

Sometimes, it is useful to add slides at the end of your presentation to refer to during audience questions.

The best way to do this is to include the `appendixnumberbeamer` package in your preamble and call `\appendix` before your backup slides.

**metropolis** will automatically turn off slide numbering and progress bars for slides in the appendix.



P. Erdős.

**A selection of problems and results in combinatorics.**

In *Recent trends in combinatorics (Matrahaza, 1995)*, pages 1–6.  
Cambridge Univ. Press, Cambridge, 1995.



R. Graham, D. Knuth, and O. Patashnik.

**Concrete mathematics.**

Addison-Wesley, Reading, MA, 1989.



G. D. Greenwade.

**The Comprehensive Tex Archive Network (CTAN).**

*TUGBoat*, 14(3):342–351, 1993.



D. Knuth.

**Two notes on notation.**

*Amer. Math. Monthly*, 99:403–422, 1992.



H. Simpson.

**Proof of the Riemann Hypothesis.**

preprint (2003), available at

<http://www.math.drofnats.edu/riemann.ps>, 2003.