Universal Adversarial Perturbations

Against Semantic Image Segmentation

Samim Zahoor[1]

[1]Department of Information Technology
National Institute of Technology Srinagar

B.Tech Seminar, 2017

Outline

Outline

What are Adversarial Perturbations?

- Imperceptible changes to the **inputs** to **deep network classifiers** that cause them to mis-predict labels.

$$\widetilde{x} = x + \eta \tag{1}$$

What are Adversarial Perturbations?

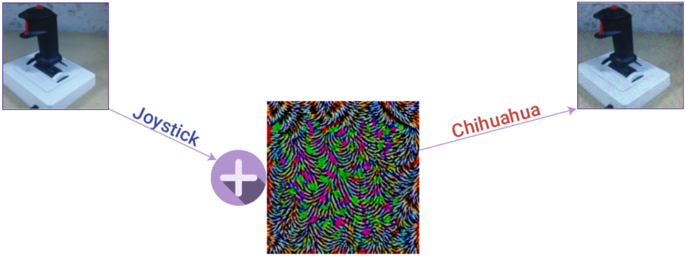


Figure: Adversarial Perturbation

Outline

Universal Adversarial Perturbations

- A single perturbation (η) that causes vast majority of the images to be mis-classified.
- Image agnostic perturbation.

What are Adversarial Perturbations?

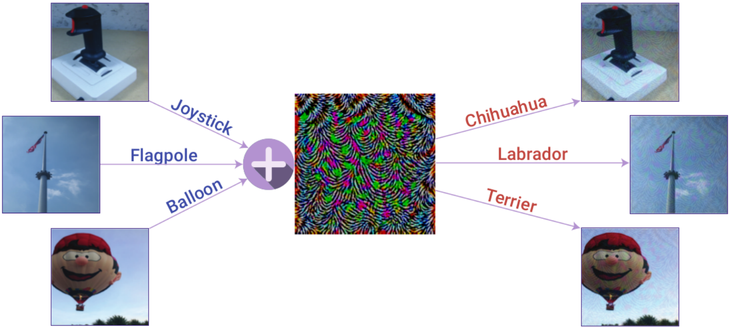


Figure: Universal Adversarial Perturbation

Outline

Generating Adversarial Perturbations

- Fast Gradient Sign Method (FGSM) by Goodfellow et al based on their hypothesis that neural networks are too linear to resist linear adversarial perturbations.

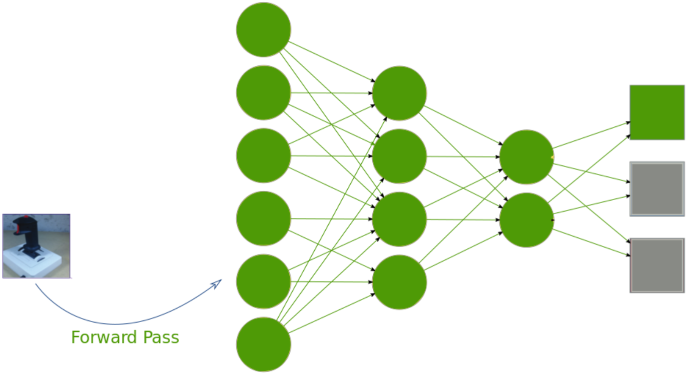
Generating Adversarial Perturbations

- Fast Gradient Sign Method (FGSM) by Goodfellow et al based on their hypothesis that neural networks are too linear to resist linear adversarial perturbations.
- Defined by

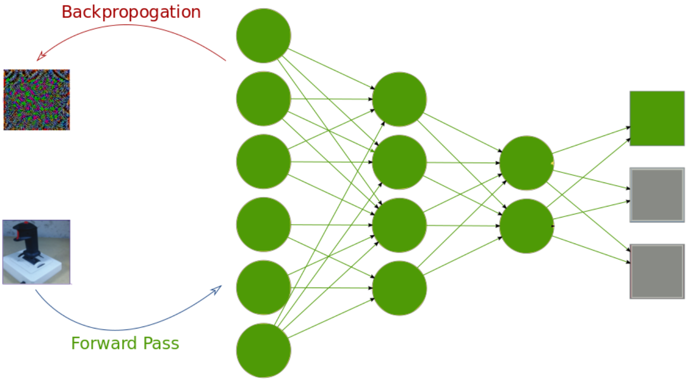
$$\eta = \epsilon sign(\nabla_x J(\boldsymbol{\Theta}, \boldsymbol{x}, y)). \tag{2}$$

J is the cost function, x is the input, y the label and ϵ is set to 0.25 for the max norm constraint.

Generating Adversarial Perturbations



Generating Adversarial Perturbations



Generating Universal Adversarial Perturbation

- A Slight Modification of the FGSM proposed by Dezfooli et al. produces image agnostic perturbations.
- They called it **DeepFool**.

Deepfool

- For the first image, DeepFool identifies a standard image-dependent perturbation. For subsequent images, it is checked whether adding the previous adversarial perturbation already fools the classifier; if yes the algorithm continues with the next image, otherwise it updates the perturbation using DeepFool such that also the current image becomes adversarial. The algorithm stops once the perturbation is adversarial on a large fraction of the train set.

Adversarial Perturbations for Semantic Image Segmentation

- Semantic image segmentation denotes a dense prediction task that addresses the what is where in an image? question by assigning a class label to each pixel of the image.

Adversarial Perturbations for Semantic Image Segmentation

- Semantic image segmentation denotes a dense prediction task that addresses the what is where in an image? question by assigning a class label to each pixel of the image.

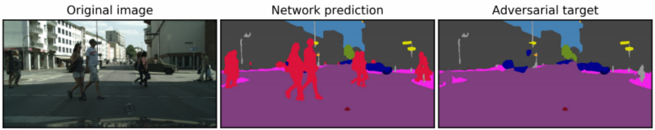


Figure: Dynamic target segmentation for hiding pedestrians.

- Recently, deep learning based approaches have become the dominant and best performing class of methods for this task and thus susceptible to adversarial examples.

Generating Universal Adversarial Perturbations

- In the context of semantic image segmentation, the universal adversarial perturbations Ξ are generated on a set of m training examples.

Generating Universal Adversarial Perturbations

- In the context of semantic image segmentation, the universal adversarial perturbations $\boldsymbol{\Xi}$ are generated on a set of m training examples.
- $\boldsymbol{\Xi}$ is defined by

$$\boldsymbol{\Xi}^{(0)} = 0, \tag{3}$$

$$\boldsymbol{\Xi}^{(n)} = Clip_{\varepsilon}\{\boldsymbol{\Xi}^{(n)} - \alpha sign(\nabla(\boldsymbol{\Xi}))\}. \tag{4}$$

With$\nabla(\boldsymbol{\Xi})$ being the loss gradient averaged over the entire training data for the semantic image segmentation loss function.

Key Points and Insights

- Adversarial examples Generalize well across different architectures and data.
- Adversarial examples can be explained as a property of high-dimensional dot products. They are a result of models being too linear, rather than too nonlinear.
- Linear models lack the capacity to resist adversarial perturbation.
- Outlook
  - Future work to address how machine learning can become more robust against (adversarial) perturbations. (Adversarial Training)
  - How adversarial attacks can be detected.