JKU

JOHANNES KEPLER
UNIVERSITY LINZ

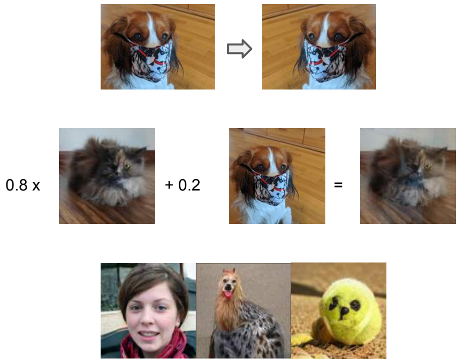
INSTITUTE OF
COMPUTATIONAL
PERCEPTION

LIT
AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Hamid Eghbal-zadeh

CP Lectures, Nov 24, 2020



JOHANNES KEPLER UNIVERSITY LINZ
LIT AI LAB
Institute of Computational Perception

Some notes

- There will be specific slides for taking questions

Some notes

- There will be specific slides for taking questions 
- A recording will be made available (Ask Alessandro!)

Some notes

- There will be specific slides for taking questions 
- A recording will be made available (Ask Alessandro!)
- Slides will be made available

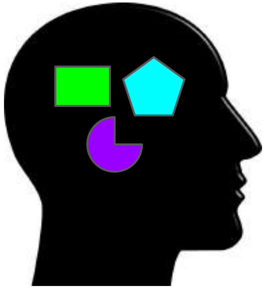
Overview

- Introduction
- An analysis framework: Adversarial Robustness in Data Augmentation
  - Performance Analysis
  - Stress Analysis
  - Influence Analysis
- Analysis results for 3 popular augmentation methods

Introduction

On the Inductive Biases in Data Augmentation and Adversarial Robustness

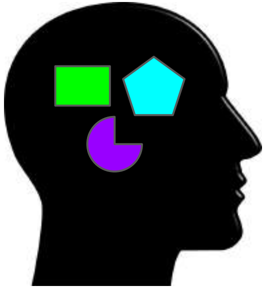
Inductive Bias: example 1



LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

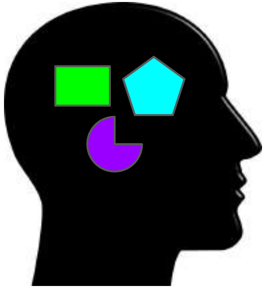
Inductive Bias: example 1

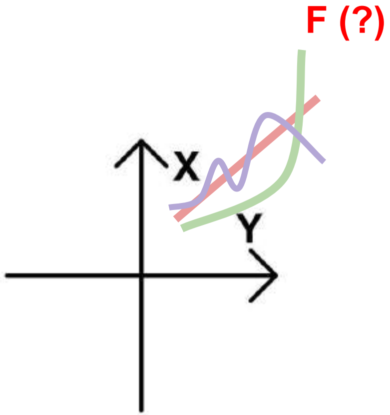


LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

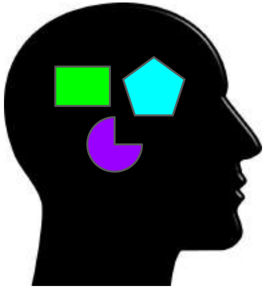
Inductive Bias: example 2

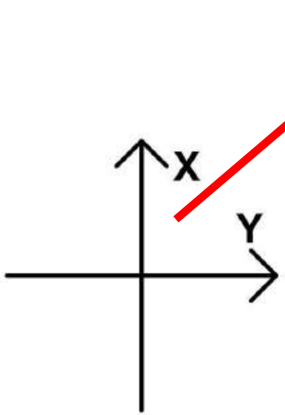




On the Inductive Biases in Data Augmentation and Adversarial Robustness

Inductive Bias: example 2





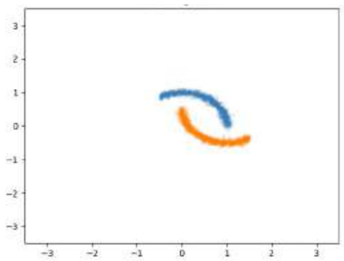
On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation:

LIT
AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

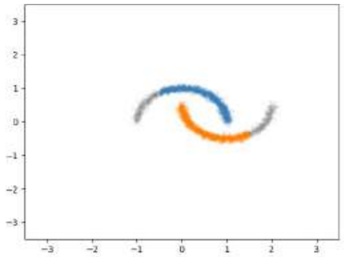
Data Augmentation:



LIT AI LAB

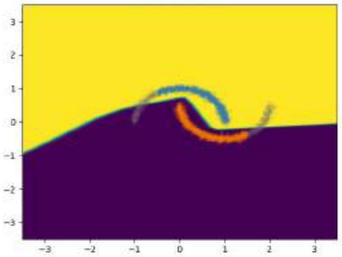
On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation:



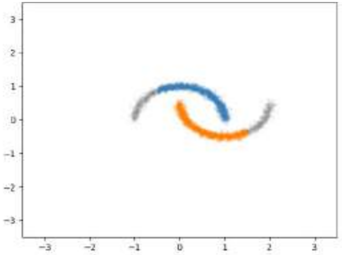
On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation:



On the Inductive Biases in Data Augmentation and Adversarial Robustness

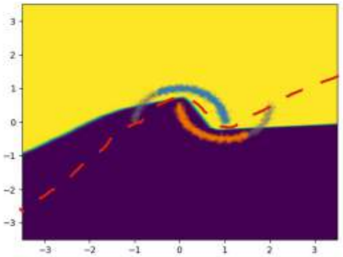
Data Augmentation:



LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

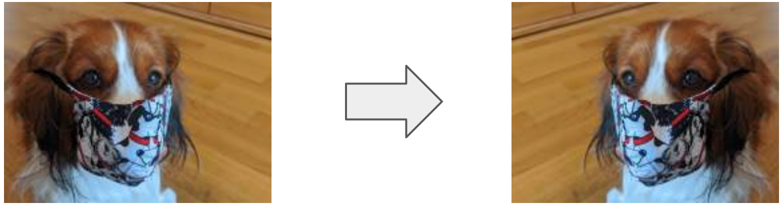
Data Augmentation:



LIT
AI LAB

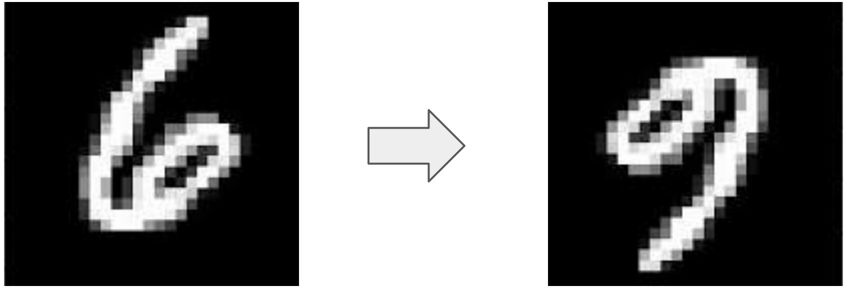
On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation: 1) Domain expert



On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation: 1) Domain expert



LIT
AI LAB

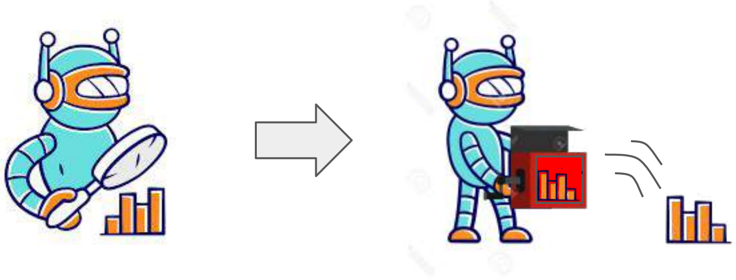
On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation: 2) Combining existing data

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Data Augmentation: 3) Generative models



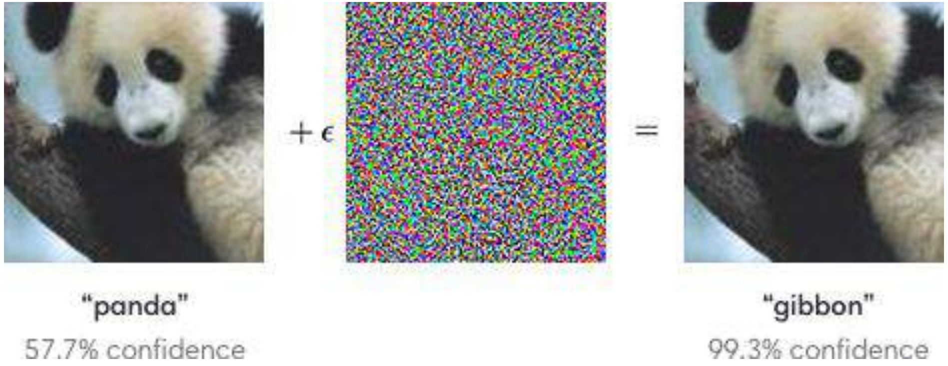
LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Adversarial examples:

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Adversarial examples:



"Explaining and Harnessing Adversarial Examples",
Ian J. Goodfellow et al, 2014.

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

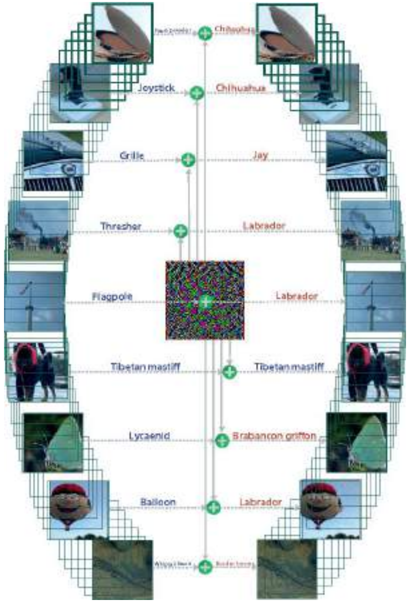
Adversarial examples:

"Robust physical-world attacks on deep learning visual classification.", Eykholt, Kevin, et al. CVPR, 2018.

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Adversarial examples:



"Universal adversarial perturbations.",
Moosavi-Dezfooli, Seyed-Mohsen, et al. CVPR, 2017.

LIT
AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \to \mathbb{R}^k$

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \to \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \to \mathbb{R}_+$

$\ell(h_\theta(x), y)$

LIT
AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \rightarrow \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \to \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \to \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

To find an adversarial example for h_θ

$$\underset{\hat{x}}{\text{maximize}} \, \ell(h_\theta(\hat{x}), y)$$

LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \rightarrow \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

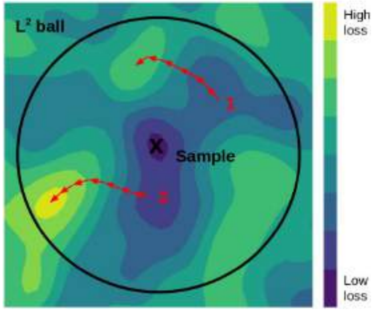
$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

To find an adversarial example for h_θ

$$\underset{\hat{x}}{\text{maximize}} \, \ell(h_\theta(\hat{x}), y)$$

And we constrain the changes to only small perturbations

$$\underset{\delta \in \Delta}{\text{maximize}} \, \ell(h_\theta(x + \delta), y)$$



On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \rightarrow \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

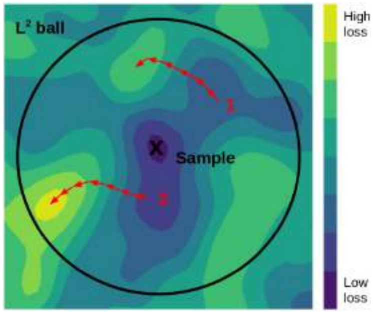
To find an adversarial example for h_θ

$$\underset{\hat{x}}{\text{maximize}} \ell(h_\theta(\hat{x}), y)$$

And we constrain the changes to only small perturbations

$$\underset{\delta \in \Delta}{\text{maximize}} \ell(h_\theta(x + \delta), y)$$

$\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$



LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \to \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \to \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

To find an adversarial example for h_θ

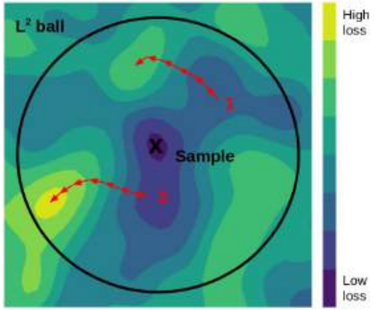
$$\underset{\hat{x}}{\text{maximize}} \ell(h_\theta(\hat{x}), y)$$

And we constrain the changes to only small perturbations

$$\underset{\delta \in \Delta}{\text{maximize}} \ell(h_\theta(x + \delta), y)$$

$$\Delta = \{\delta : ||\delta||_p \leq \epsilon\}$$

$$\Delta = \{\delta : ||\delta||_\infty \leq \epsilon\}$$



LIT AI LAB

On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \to \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \to \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

To find an adversarial example for h_θ

$$\underset{\hat{x}}{\text{maximize}} \ell(h_\theta(\hat{x}), y)$$

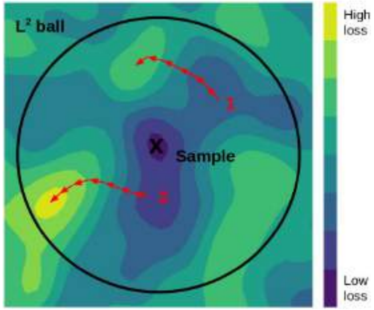
And we constrain the changes to only small perturbations

$$\underset{\delta \in \Delta}{\text{maximize}} \ell(h_\theta(x + \delta), y)$$

$\Delta = \{\delta : ||\delta||_p \leq \epsilon\}$

$\Delta = \{\delta : ||\delta||_\infty \leq \epsilon\}$

$\Delta = \{\delta : ||\delta||_2 \leq \epsilon\}$



On the Inductive Biases in Data Augmentation and Adversarial Robustness

Assume a Neural Net

$h_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$

And loss function ℓ

$\ell : \mathbb{R}^k \times \mathbb{Z}_+ \rightarrow \mathbb{R}_+$

$\ell(h_\theta(x), y)$

ℓ is used to train the NN

$$\underset{\theta}{\text{minimize}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

To find an adversarial example for h_θ

$$\underset{\hat{x}}{\text{maximize}} \, \ell(h_\theta(\hat{x}), y)$$

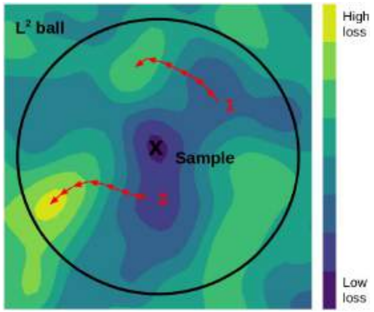
And we constrain the changes to only small perturbations

$$\underset{\delta \in \Delta}{\text{maximize}} \, \ell(h_\theta(x + \delta), y)$$

$$\Delta = \{\delta : ||\delta||_p \leq \epsilon\}$$

$$\Delta = \{\delta : ||\delta||_\infty \leq \epsilon\}$$

$$\Delta = \{\delta : ||\delta||_2 \leq \epsilon\}$$



LIT AI LAB

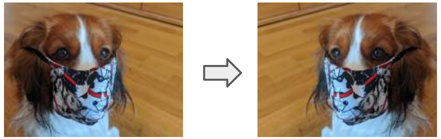
Introduction

Data augmentation is one of the standard techniques in deep learning, and been shown to greatly improve the generalisation abilities of models. The 3 popular examples are:

Introduction

Data augmentation is one of the standard techniques in deep learning, and been shown to greatly improve the generalisation abilities of models. The 3 popular examples are:

1. Traditional (classical) data augmentation: The idea is to incorporate domain expert knowledge into the model (e.g, if data is images of dogs, horizontal flipping helps).

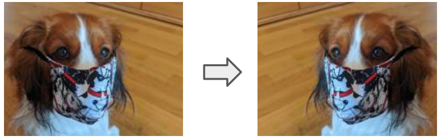


LIT
AI LAB

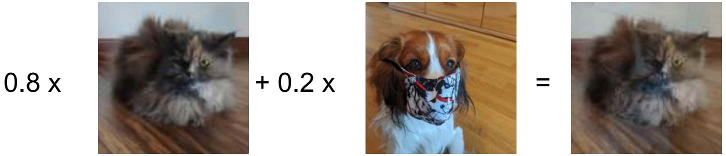
Introduction

Data augmentation is one of the standard techniques in deep learning, and been shown to greatly improve the generalisation abilities of models. The 3 popular examples are:

1. Traditional (classical) data augmentation: The idea is to incorporate domain expert knowledge into the model (e.g, if data is images of dogs, horizontal flipping helps).



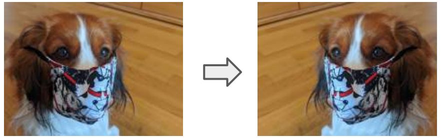
2. Mixup: Linearly combining data and their labels.



Introduction

Data augmentation is one of the standard techniques in deep learning, and been shown to greatly improve the generalisation abilities of models. The 3 popular examples are:

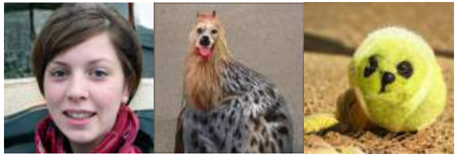
1. Traditional (classical) data augmentation: The idea is to incorporate domain expert knowledge into the model (e.g, if data is images of dogs, horizontal flipping helps).



2. Mixup: Linearly combining data and their labels.

0.8 x  + 0.2 x  = 

3. Generative models (GANs): Conditioning a generative model on labels.



Introduction

- In today's talk, we detail an analysis framework for systematically evaluating data augmentation methods with respect to **risk under attack** and **classification risk**.

LIT AI LAB

Introduction

- In today’s talk, we detail an analysis framework for systematically evaluating data augmentation methods with respect to **risk under attack** and **classification risk**.

- Given this framework, we analyze three popular data augmentation methods (Classic, mixup, GAN-augmentation)

Introduction

- In today's talk, we detail an analysis framework for systematically evaluating data augmentation methods with respect to **risk under attack** and **classification risk**.

- Given this framework, we analyze three popular data augmentation methods (Classic, mixup, GAN-augmentation)

- We provide a formal formulation for data augmentation based on random functions.
    - This allows us to express combinations of data augmentations as composition of functions

Introduction

- In today's talk, we detail an analysis framework for systematically evaluating data augmentation methods with respect to **risk under attack** and **classification risk**.

- Given this framework, we analyze three popular data augmentation methods (Classic, mixup, GAN-augmentation)

- We provide a formal formulation for data augmentation based on random functions.
    - This allows us to express combinations of data augmentations as composition of functions

- We provide a new measure known as **prediction-change stress**, and show that this property is related to the adversarial vulnerability of models.

Introduction

- In today's talk, we detail an analysis framework for systematically evaluating data augmentation methods with respect to **risk under attack** and **classification risk**.

- Given this framework, we analyze three popular data augmentation methods (Classic, mixup, GAN-augmentation)

- We provide a formal formulation for data augmentation based on random functions.
  - This allows us to express combinations of data augmentations as composition of function

- We provide a new measure known as **prediction-change stress**, and show that this property is related to the adversarial vulnerability of models.

- We use **Influence functions** to examine **how much influence** models have from real and augmented data

LIT
AI LAB

Formal Definition of Data Augmentation

Data Augmentation - Formal definition

A random function $A : (\mathcal{X} \times \mathcal{Y})^s \to \{X \times Y : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d\}^r$ is an ***Augmentation***, if it maps a sample $S = ((\mathbf{x}_1, l(\mathbf{x}_1)), \ldots, (\mathbf{x}_s, l(\mathbf{x}_s))) \in (\mathcal{X} \times \mathcal{Y})^s$, with measure P_X on $\mathcal{X}$, and labeling function $l : \mathcal{X} \to \mathcal{Y}$, to some vector $A(S) = (X_1 \times Y_1, \ldots, X_r \times Y_r)$ of independent random vectors $X_1 \times Y_1, \ldots, X_r \times Y_r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ with measure $P_{X_I \times Y_I}$ on $\mathcal{X} \times \mathcal{Y}$ and marginal measure P_{X_I} dominating P_X.

LIT AI LAB

Data Augmentation - Formal definition

By this definition, an augmented sample $\tilde{S} = ((\tilde{\mathbf{x}}_1, \tilde{y}_1), \ldots, (\tilde{\mathbf{x}}_s, \tilde{y}_s))$ can be obtained from a sample $S \in (\mathcal{X} \times \mathcal{Y})^s$ by observing the random variable $A(S)$.

The assumption P_{X_I} dominating P_X ensures data augmentations take the original sample into account, i.e. if $P_X(D) > 0$ then also $P_{X_I}(D) > 0$ for any measurable D.

LIT AI LAB

Data Augmentation - Formal definition (Cont'd)

Lemma:

If $A(S)$ and $B(S)$ are augmentations then $A(S) \circ B(S)$ is also an augmentation.

Data Augmentation - Formal definition (Cont'd)

Lemma:

If $A(S)$ and $B(S)$ are augmentations then $A(S) \circ B(S)$ is also an augmentation.

Therefore:

- Classical data augmentation (flipping, rotating, etc) is an augmentation

Data Augmentation - Formal definition (Cont'd)

Lemma:

If A(S) and B(S) are augmentations then $A(S) \circ B(S)$ is also an augmentation.

Therefore:

- Classical data augmentation (flipping, rotating, etc) is an augmentation
- Conditional generative models (GANs) are an augmentation.

Data Augmentation - Formal definition (Cont'd)

Lemma:

If $A(S)$ and $B(S)$ are augmentations then $A(S) \circ B(S)$ is also an augmentation.

Therefore:

- Classical data augmentation (flipping, rotating, etc) is an augmentation
- Conditional generative models (GANs) are an augmentation.
- Sampling from vicinity distributions (Mixup) is an augmentation.

LIT
AI LAB

Data Augmentation - Formal definition (Cont'd)

Lemma:

If A(S) and B(S) are augmentations then $A(S) \circ B(S)$ is also an augmentation.

Therefore:

- Classical data augmentation (flipping, rotating, etc) is an augmentation
- Conditional generative models (GANs) are an augmentation.
- Sampling from vicinity distributions (Mixup) is an augmentation.

LIT AI LAB

Adversarial Robustness in Data Augmentation

Adversarial Robustness in Data Augmentation

The proposed analysis framework for the augmentation functions as defined before is structured into three parts:

Adversarial Robustness in Data Augmentation

The proposed analysis framework for the augmentation functions as defined before is structured into three parts:

1. **Performance analysis**: where we look at the effect of data augmentation on classification performance and adversarial robustness.

Adversarial Robustness in Data Augmentation

The proposed analysis framework for the augmentation functions as defined before is structured into three parts:

1. **Performance analysis**: where we look at the effect of data augmentation on classification performance and adversarial robustness.

2. **Stress analysis**: where we analyse how the predictions of a model under adversarial attacks, is affected by the augmentation.

Adversarial Robustness in Data Augmentation

The proposed analysis framework for the augmentation functions as defined before is structured into three parts:

1. **Performance analysis**: where we look at the effect of data augmentation on classification performance and adversarial robustness.

2. **Stress analysis**: where we analyse how the predictions of a model under adversarial attacks, is affected by the augmentation.

3. **Influence analysis**: where we look at how much a model relies on augmented training samples when predicting on the real test examples and their adversarial counterparts.

Performance Analysis

Performance Analysis

We analyse the models w.r.t **usefulness** and **adversarial robustness**:

- We apply each data augmentation method with a probability changing from 0 to 1.
  We train a model with a specific augmentation probability fixed, and then evaluate it.

Performance Analysis

We analyse the models w.r.t **usefulness** and **adversarial robustness**:

- We apply each data augmentation method with a probability changing from 0 to 1.
  We train a model with a specific augmentation probability fixed, and then evaluate it.

- We train a Resnet50 on the training data, and we report:
  - Normal test error (usefulness)

Performance Analysis

We analyse the models w.r.t **usefulness** and **adversarial robustness**:

- We apply each data augmentation method with a probability changing from 0 to 1. We train a model with a specific augmentation probability fixed, and then evaluate it.

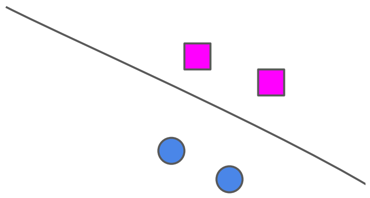
- We train a Resnet50 on the training data, and we report:
  - Normal test error (usefulness)
  - Risk under attack (error under adversarial attack) for 4 cases of PGD attack (robustness)
    - with epsilon=0.25 and 0.5
    - 10 and 100 iterations

LIT AI LAB

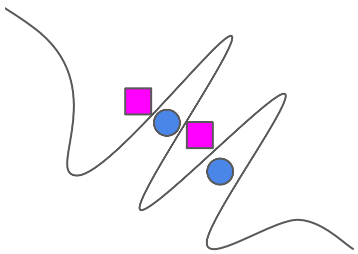
Stress Analysis

Stress Analysis

Boundary 1:

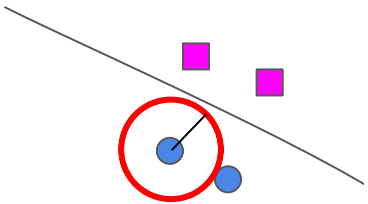


Boundary 2:

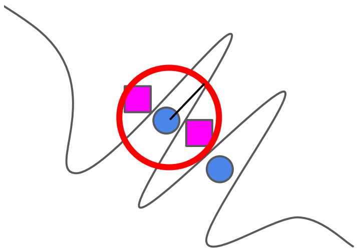


Stress Analysis

Boundary 1:



Boundary 2:

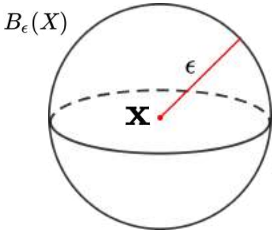


Stress Analysis

On a sample $\tilde{S} := ((\mathbf{x}'_1, l(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_s, l(\mathbf{x}'_s)))$, where all points are from the surface of ∂B_ϵ
We introduce prediction-change stress as follows

$$\widehat{\text{stress}}_{\text{pc}}(f, \tilde{S}, \epsilon) := \frac{1}{s} \sum_{i=1}^{s} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{f(\mathbf{x}'_i) \neq f(\mathbf{y}_{ij})}$$

Where ∂B_ϵ is the surface of a ball $B_\epsilon(X)$ around $\mathbf{X}$ with radius $\epsilon > 0$.



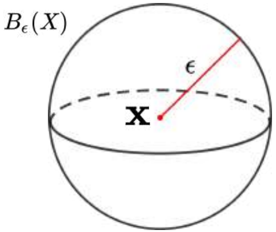
Stress Analysis

On a sample $\tilde{S} := ((\mathbf{x}'_1, l(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_s, l(\mathbf{x}'_s)))$, where all points are from the surface of ∂B_ϵ
We introduce prediction-change stress as follows

$$\widehat{\text{stress}}_{\text{pc}}(f, \tilde{S}, \epsilon) := \frac{1}{s} \sum_{i=1}^{s} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{f(\mathbf{x}'_i) \neq f(\mathbf{y}_{ij})}$$

Where ∂B_ϵ is the surface of a ball $B_\epsilon(X)$ around $\mathbf{X}$ with radius $\epsilon > 0$.

In other words, for a given input $\mathbf{X}$ and its predicted label $f(\mathbf{x})$, stress relates to the **probability** that a random neighbor from the ϵ-sphere of $\mathbf{X}$ will be **assigned a different label** by the model.



LIT
AI LAB

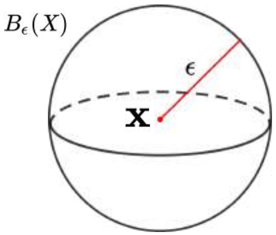
Stress Analysis

On a sample $\tilde{S} := ((\mathbf{x}'_1, l(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_s, l(\mathbf{x}'_s)))$, where all points are from the surface of ∂B_ϵ
We introduce prediction-change stress as follows

$$\widehat{\text{stress}}_{\text{pc}}(f, \tilde{S}, \epsilon) := \frac{1}{s} \sum_{i=1}^{s} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{f(\mathbf{x}'_i) \neq f(\mathbf{y}_{ij})}$$

Where ∂B_ϵ is the surface of a ball $B_\epsilon(X)$ around $\mathbf{X}$ with radius $\epsilon > 0$.

In other words, for a given input $\mathbf{X}$ and its predicted label $f(\mathbf{x})$, stress relates to the **probability** that a random neighbor from the ϵ-sphere of $\mathbf{X}$ will be **assigned a different label** by the model.



LIT
AI LAB

Influence Analysis

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

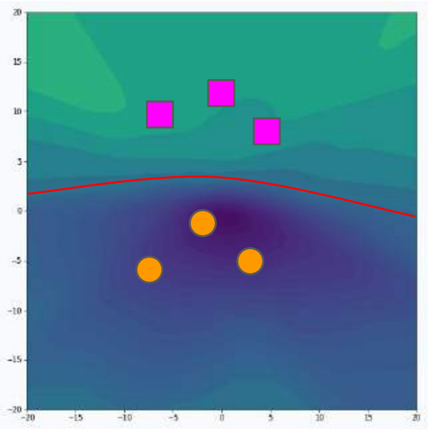
- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

LIT
AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

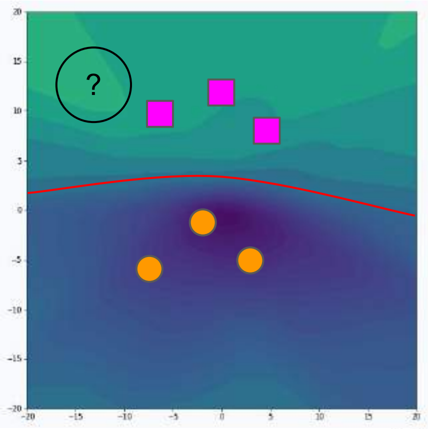


LIT
AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

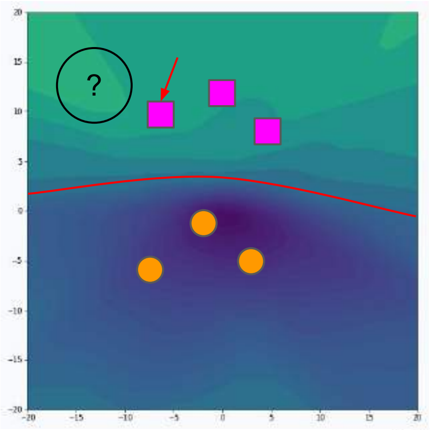
- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.



Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

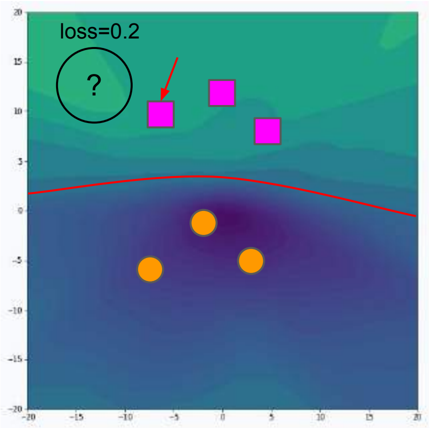


LIT AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

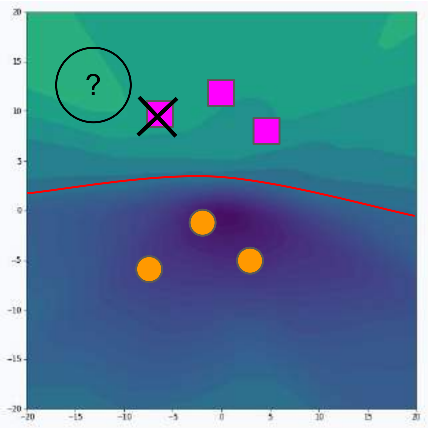


LIT AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

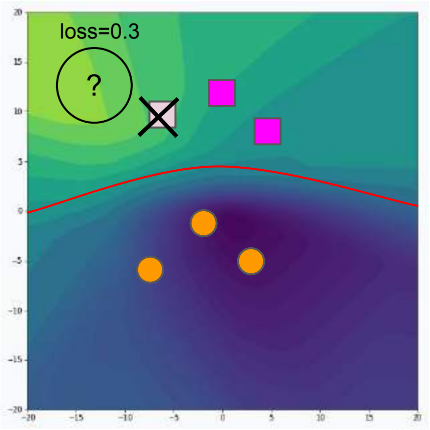


LIT
AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.



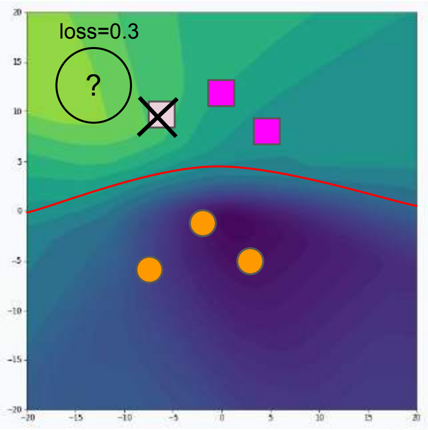
LIT AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

Influence => 0.3 - 0.2 = +0.1



Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

- We compare the influence of **real training** to **augmented training** for **normal** and **adversarial** test examples.
    - We show the distribution of influence values

LIT AI LAB

Influence Analysis

- We use influence functions *(Koh & Liang, 2017)* to analyse the importance of normal, and augmented training data in relation to adversarial vulnerability of the resulting classifiers.

- **Short Recap**: For a given test example, influence functions compute an **importance value** for **each training point** that shows **how much it contributed** to the **prediction of that test example**, by **estimating the change in the loss** on that test example that would result **if the training point were removed** from the training set.

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

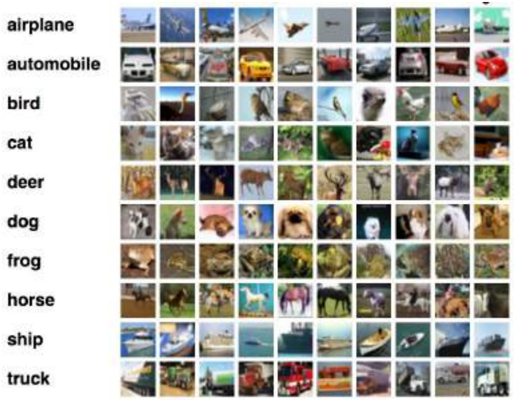
- We compare the influence of **real training** to **augmented training** for **normal** and **adversarial** test examples.
  - We show the distribution of influence values

LIT AI LAB

Experimental Setup

Experimental Setup

- We train a Resnet50 with SGD on CIFAR10, achieving normal acc of 94.92%.



LIT AI LAB

Experimental Setup

- We train a Resnet50 with SGD on CIFAR10, achieving normal acc of 94.92%.
- Augmentation were applied by probabilities of {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}

Experimental Setup

- We train a Resnet50 with SGD on CIFAR10, achieving normal acc of 94.92%.
- Augmentation were applied by probabilities of {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
- Each experiment repeated 3 times, mean and std are reported

Experimental Setup

- We train a Resnet50 with SGD on CIFAR10, achieving normal acc of 94.92%.
- Augmentation were applied by probabilities of {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
- Each experiment repeated 3 times, mean and std are reported
- Two GAN models (NS, WGP) were trained, and evaluated with FID (20.11, 18.30)

LIT
AI LAB

Experimental Setup

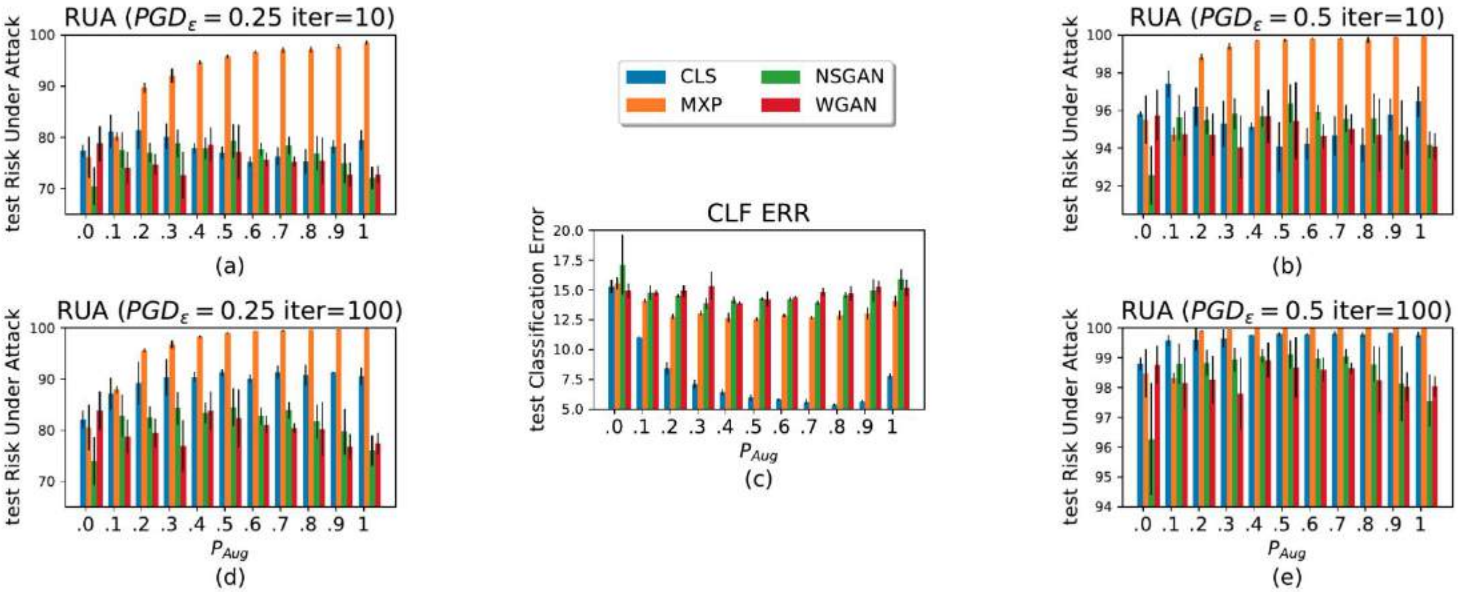
- We train a Resnet50 with SGD on CIFAR10, achieving normal acc of 94.92%.
- Augmentation were applied by probabilities of {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
- Each experiment repeated 3 times, mean and std are reported
- Two GAN models (NS, WGP) were trained, and evaluated with FID (20.11, 18.30)
  - Generators were conditioned on labels of the train set

LIT AI LAB

Results

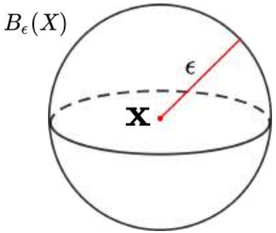
Results

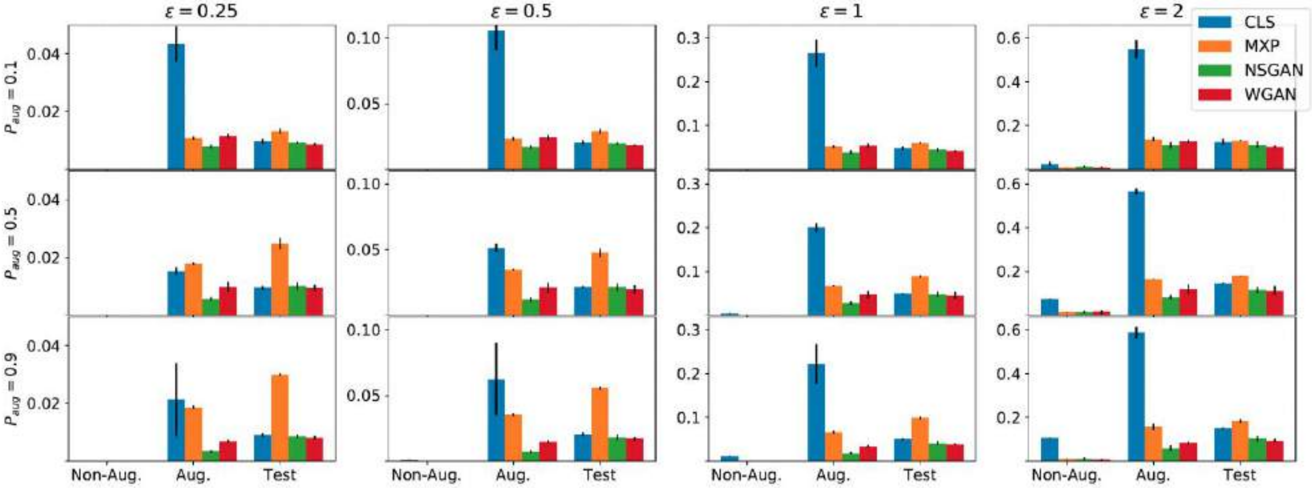
Classification and Adversarial Risk:



Results

Stress analysis

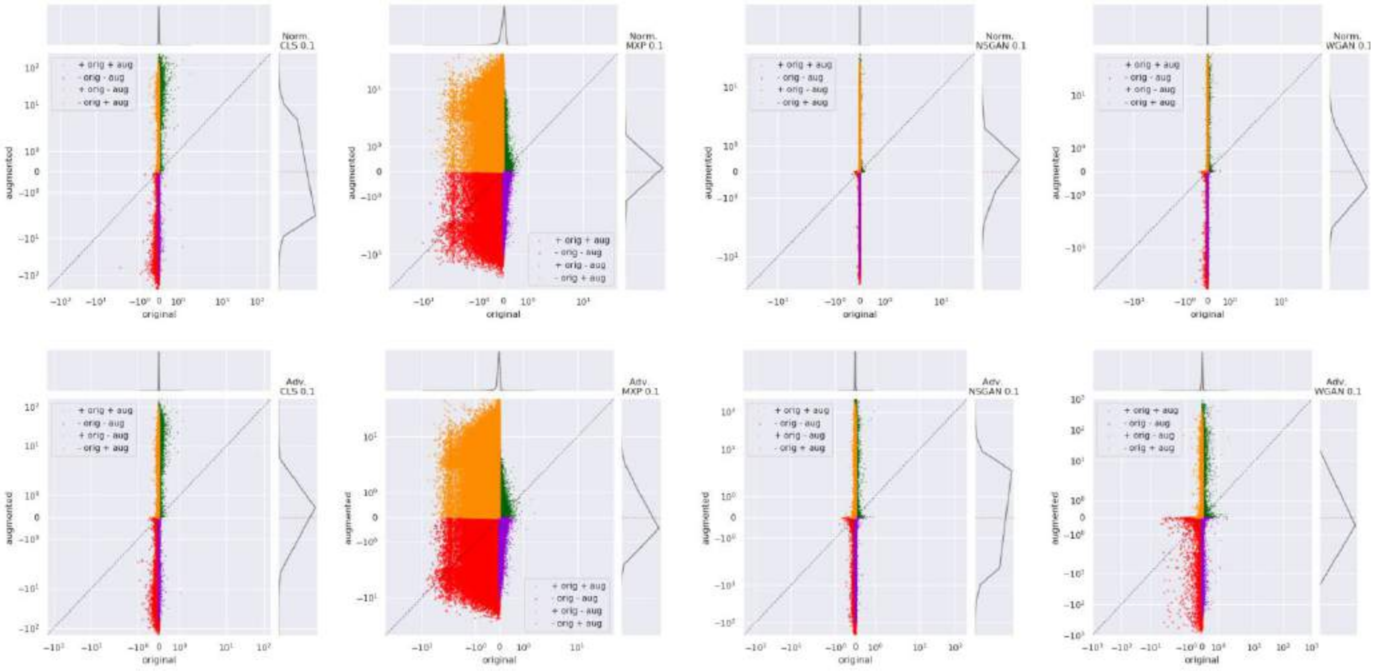




LIT AI LAB

Results

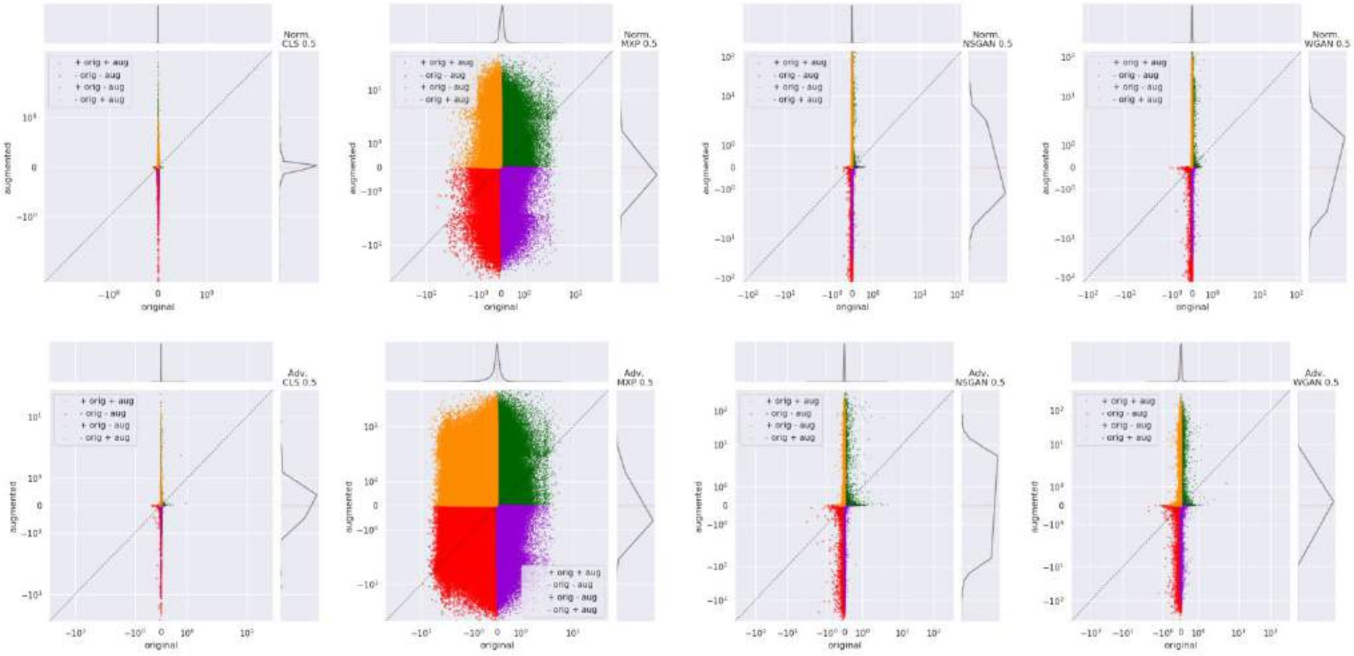
Influence analysis



LIT
AI LAB

Results

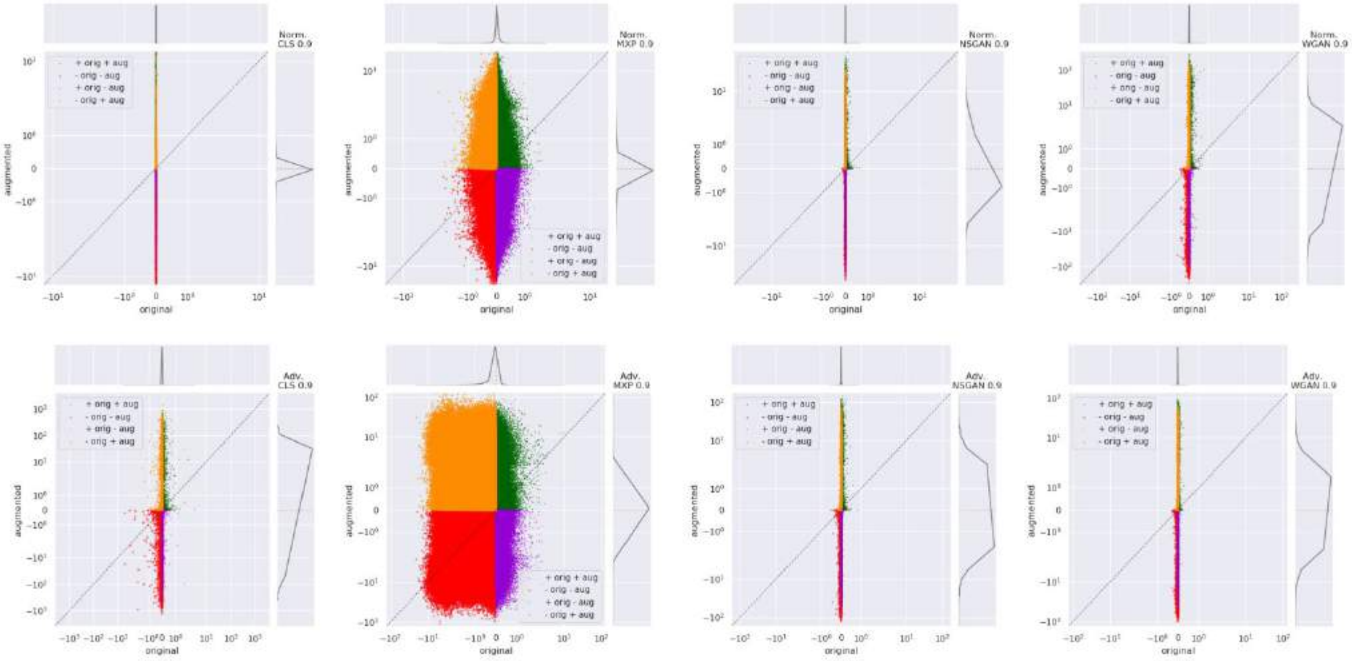
Influence analysis



LIT AI LAB

Results

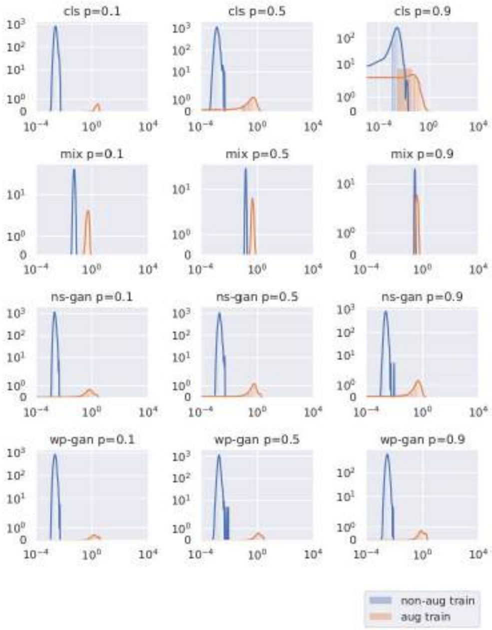
Influence analysis

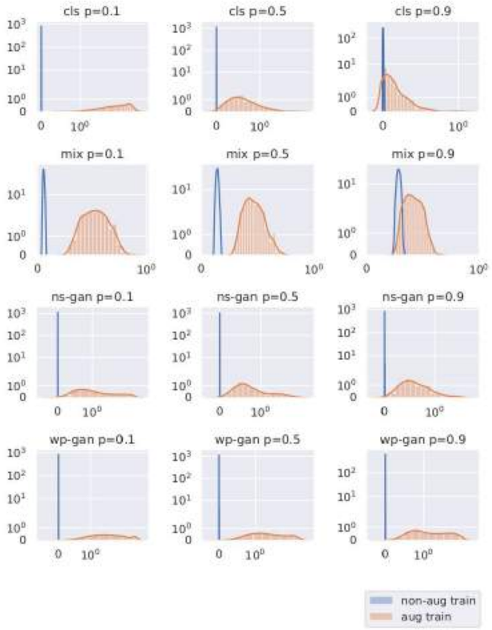


Results

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Gradient-norm



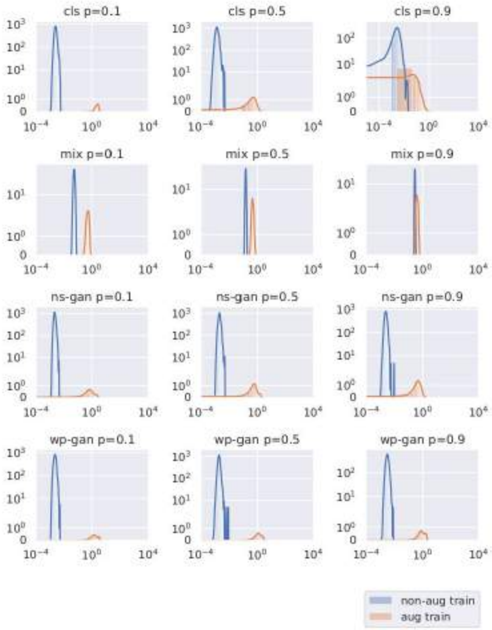


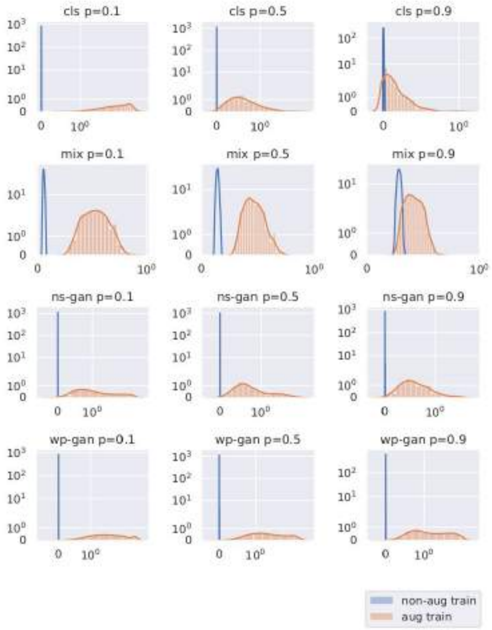
LIT AI LAB

Results

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top \cancel{H_{\hat{\theta}}^{-1}} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Gradient-norm



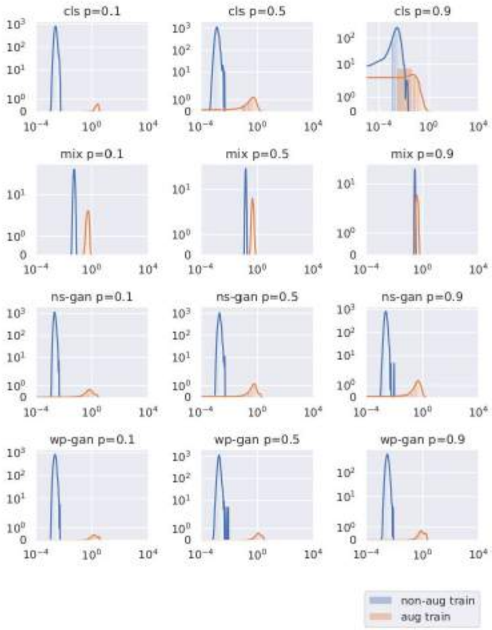


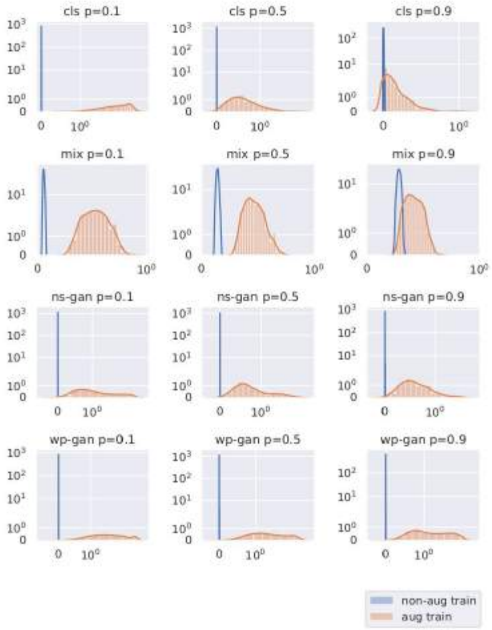
LIT AI LAB

Results

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Gradient-norm



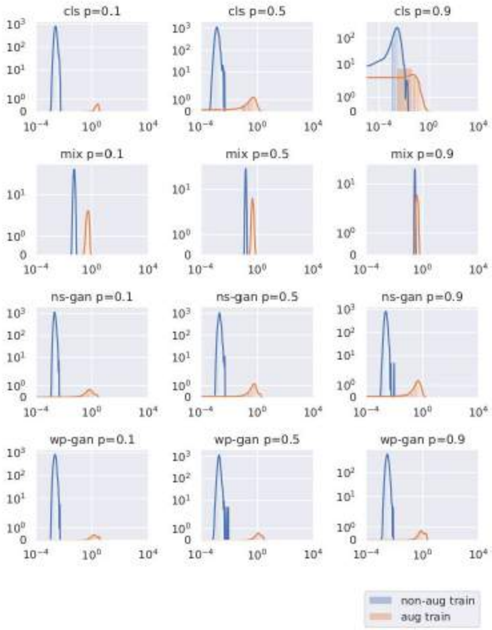


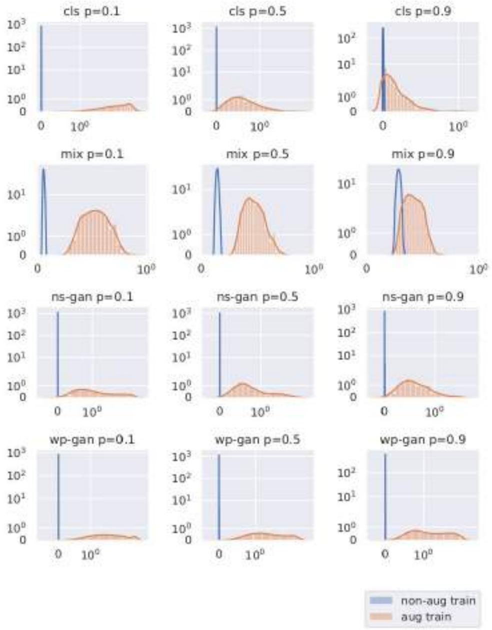
LIT
AI LAB

Results

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Gradient-norm



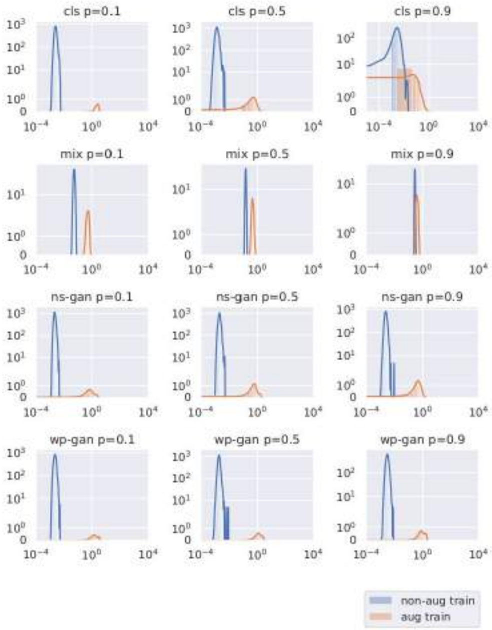


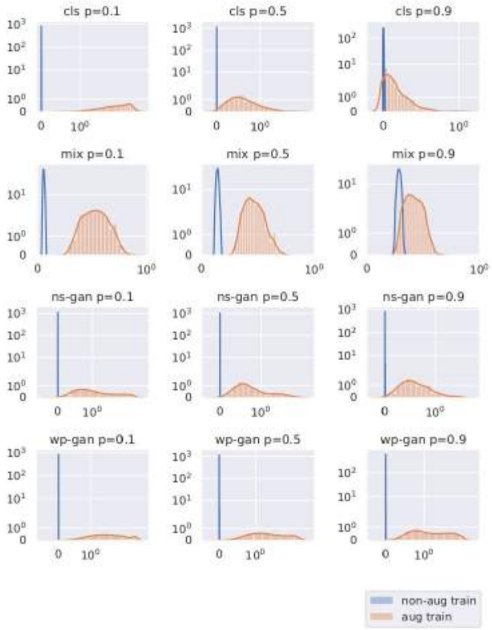
LIT AI LAB

Results

$$\hat{I}(\mathbf{x}, \mathbf{x}_{\text{test}}) := -\nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}_{\text{test}}), l(\mathbf{x}_{\text{test}}))^\top H_{\hat{\theta}}^{-1} \nabla_\theta L(f_{\hat{\theta}}(\mathbf{x}), l(\mathbf{x})),$$

Gradient-norm





LIT AI LAB

Summary

Summary

- In this talk, we provided a framework for evaluating the **classification risk** and **risk under attack** of data augmentation algorithms

Summary

- In this talk, we provided a framework for evaluating the **classification risk** and **risk under attack** of data augmentation algorithms
- We provided a theoretical definition for data augmentation that enables us to apply function composition on data augmentation

LIT AI LAB

Summary

- In this talk, we provided a framework for evaluating the **classification risk** and **risk under attack** of data augmentation algorithms
- We provided a theoretical definition for data augmentation that enables us to apply function composition on data augmentation
- We showed that the expert-introduced augmentations were the most robust and useful augmentation

Summary

- In this talk, we provided a framework for evaluating the **classification risk** and **risk under attack** of data augmentation algorithms
- We provided a theoretical definition for data augmentation that enables us to apply function composition on data augmentation
- We showed that the expert-introduced augmentations were the most robust and useful augmentation
- We analysed the decision boundary of models using the proposed prediction-change stress and showed that non-robust augmentations result in higher stress around test examples.

Summary

- In this talk, we provided a framework for evaluating the **classification risk** and **risk under attack** of data augmentation algorithms
- We provided a theoretical definition for data augmentation that enables us to apply function composition on data augmentation
- We showed that the expert-introduced augmentations were the most robust and useful augmentation
- We analysed the decision boundary of models using the proposed prediction-change stress and showed that non-robust augmentations result in higher stress around test examples.
- We analysed the influence of augmentation on models, and showed that models get more influenced by augmented data.

LIT AI LAB

Collaborators

Hamid Eghbal-zadeh

LIT AI LAB

Institute of Computational Perception

JKU JOHANNES KEPLER UNIVERSITY LINZ

Khaled Koutini

Institute of Computational Perception

JKU JOHANNES KEPLER UNIVERSITY LINZ

Verena Haunschmid

Institute of Computational Perception

JKU JOHANNES KEPLER UNIVERSITY LINZ

Paul Primus

Institute of Computational Perception

JKU JOHANNES KEPLER UNIVERSITY LINZ

Michal Lewandowski

scch software competence center hagenberg

Werner Zellinger

scch software competence center hagenberg

Bernhard Moser

scch software competence center hagenberg

Gerhard Widmer

LIT AI LAB

Institute of Computational Perception

JKU JOHANNES KEPLER UNIVERSITY LINZ

[1] On Data Augmentation and Adversarial Risk: An Empirical Analysis
Hamid Eghbal-zadeh, Khaled Koutini, Paul Primus, Verena Haunschmid, Michal Lewandowski,Werner Zellinger, Bernhard A. Moser, Gerhard Widmer
arXiv preprint arXiv:2007.02650., 2020.

[2] Adversarial Robustness in Data Augmentation
Hamid Eghbal-zadeh, Khaled Koutini, Paul Primus, Verena Haunschmid, Michal Lewandowski,Werner Zellinger, Gerhard Widmer
Towards Trustworthy ML: Rethinking Security and Privacy for ML, ICLR 2020 Workshop (talk), 2020.

Thank you!

https://www.jku.at/en/institute-of-computational-perception/news-media-events/cp-lectures/

https://eghbalz.github.io/

hamid.eghbal-zadeh@jku.at

Welcome to Q&A!