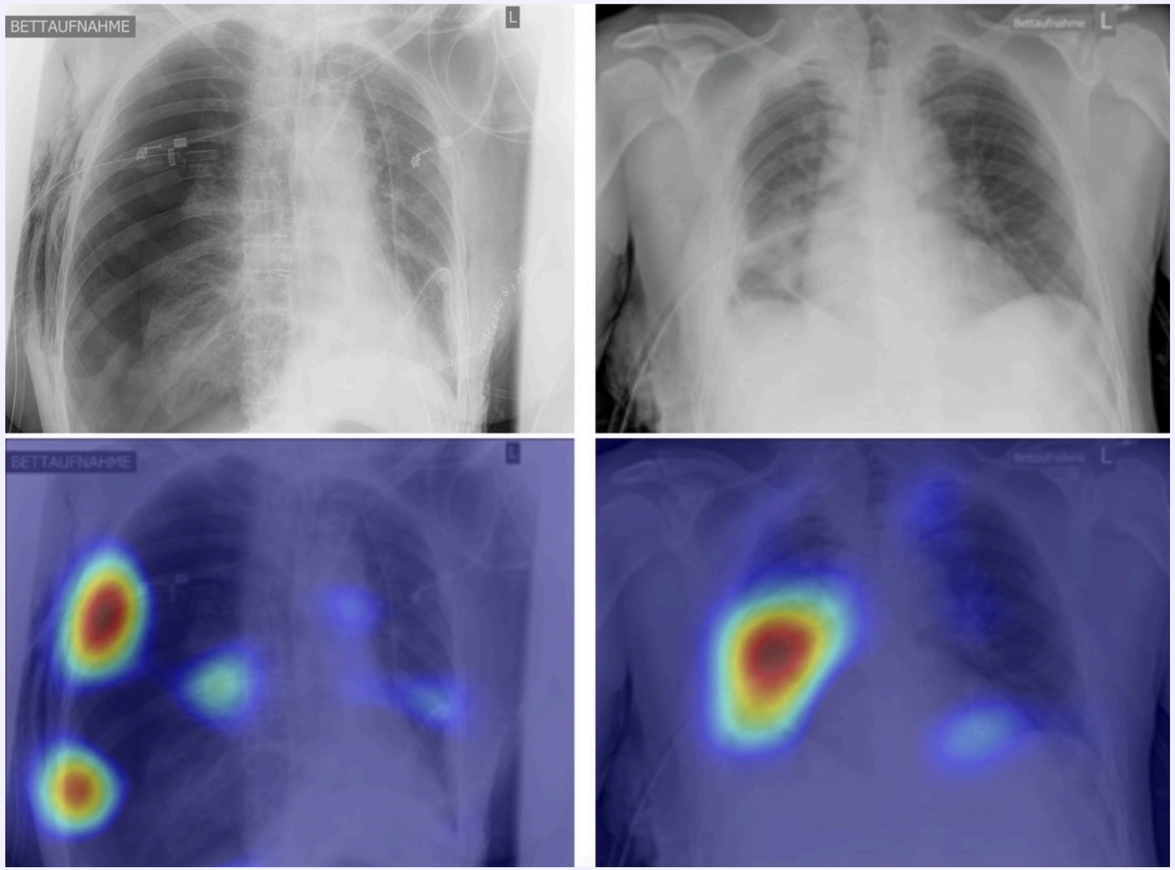
Robustness:
A Generalisation Indicator

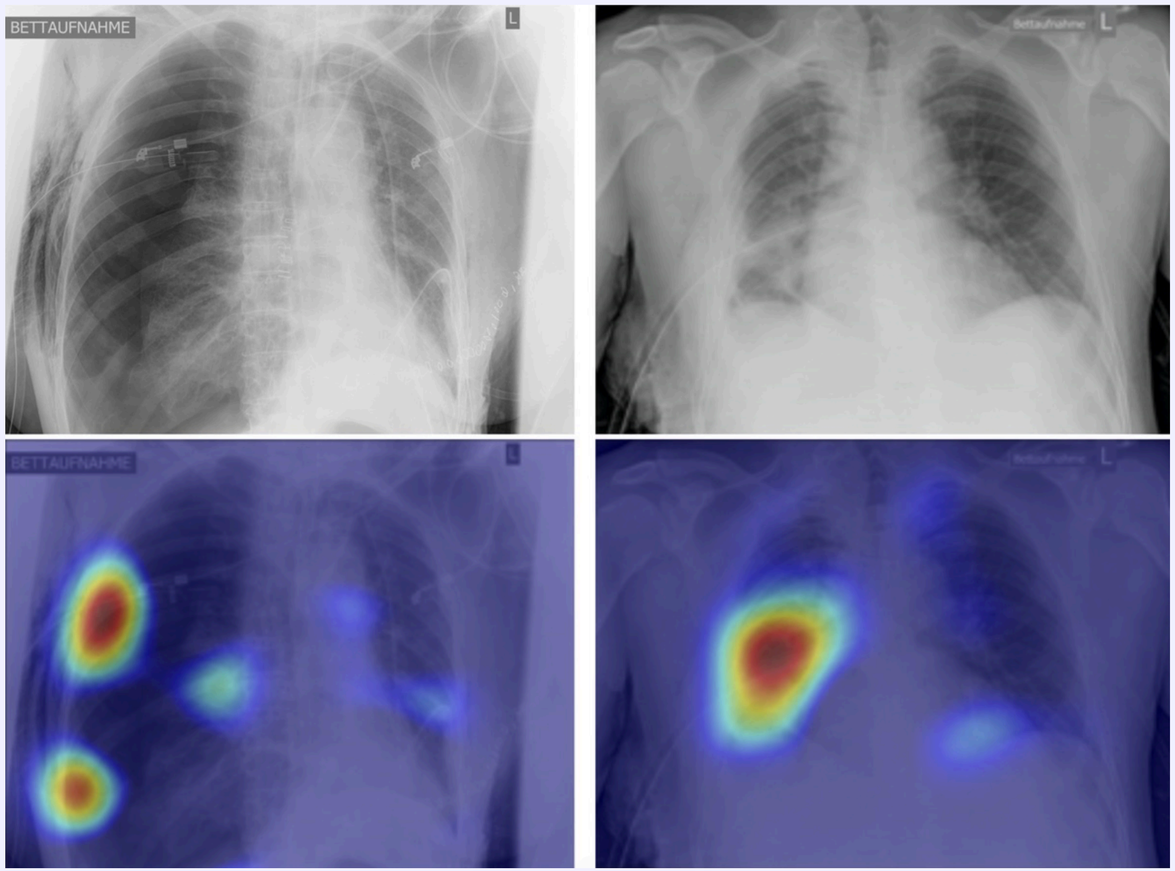
Antonia Marcu

University of Southampton

Generalisation

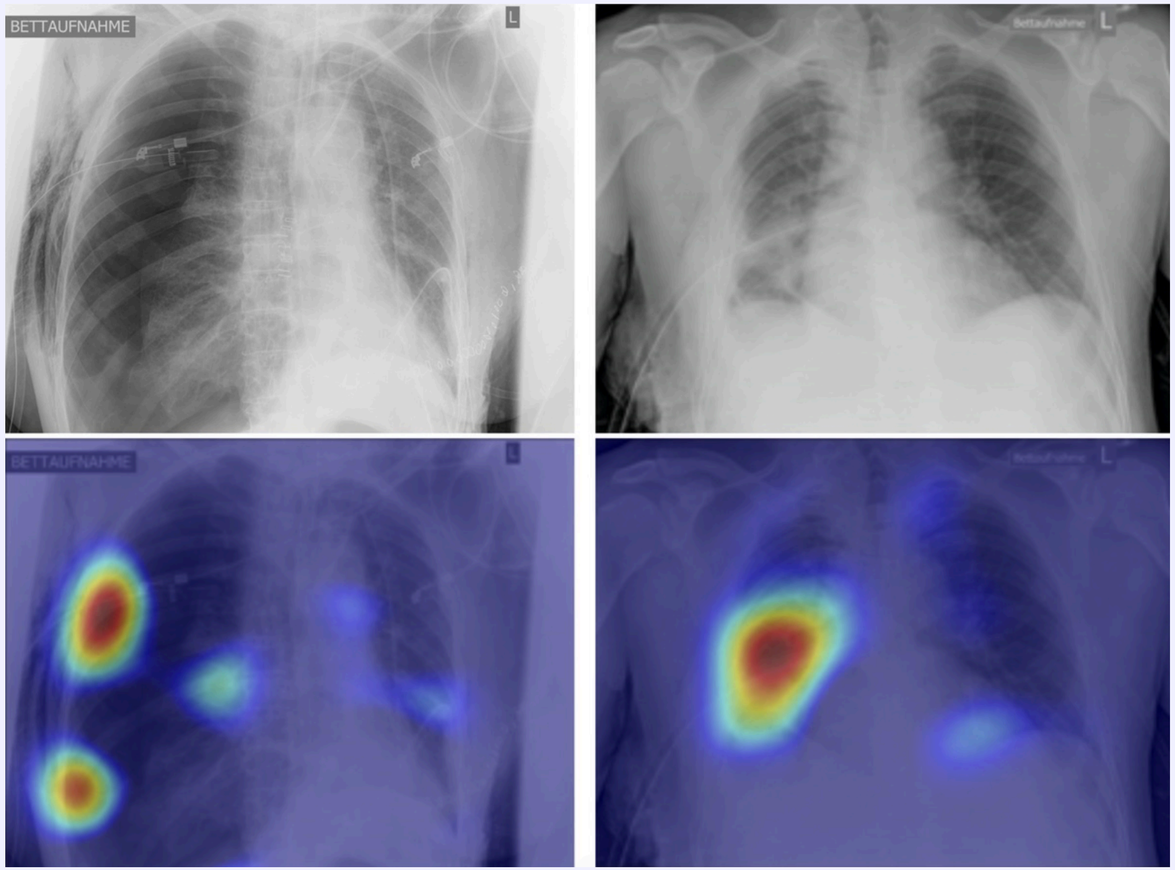


Generalisation



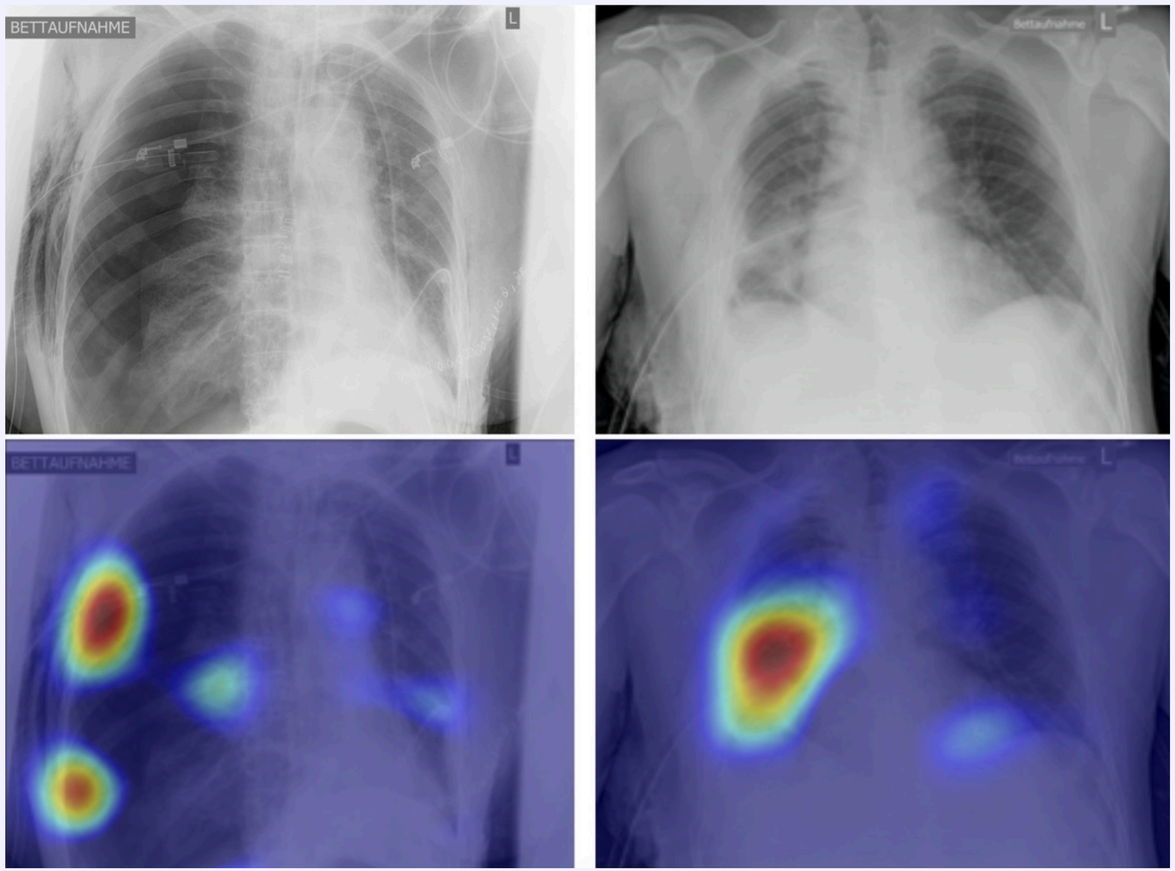
- Societal implications

Generalisation



- Societal implications
- Theoretical challenges

Generalisation



- Societal implications
- Theoretical challenges

Goal: A theoretical framework to detect and prevent such issues

Where is Generalisation Right Now? Theory vs Practice

Where is Generalisation Right Now?

Theory

- Goal: bound generalisation performance

Where is Generalisation Right Now?

Theory

- Goal: bound generalisation performance
- Uninformative bounds

Where is Generalisation Right Now?

Theory

- Goal: bound generalisation performance
- Uninformative bounds
- Making significant progress but still far from capturing generalisation

Where is Generalisation Right Now?

Theory

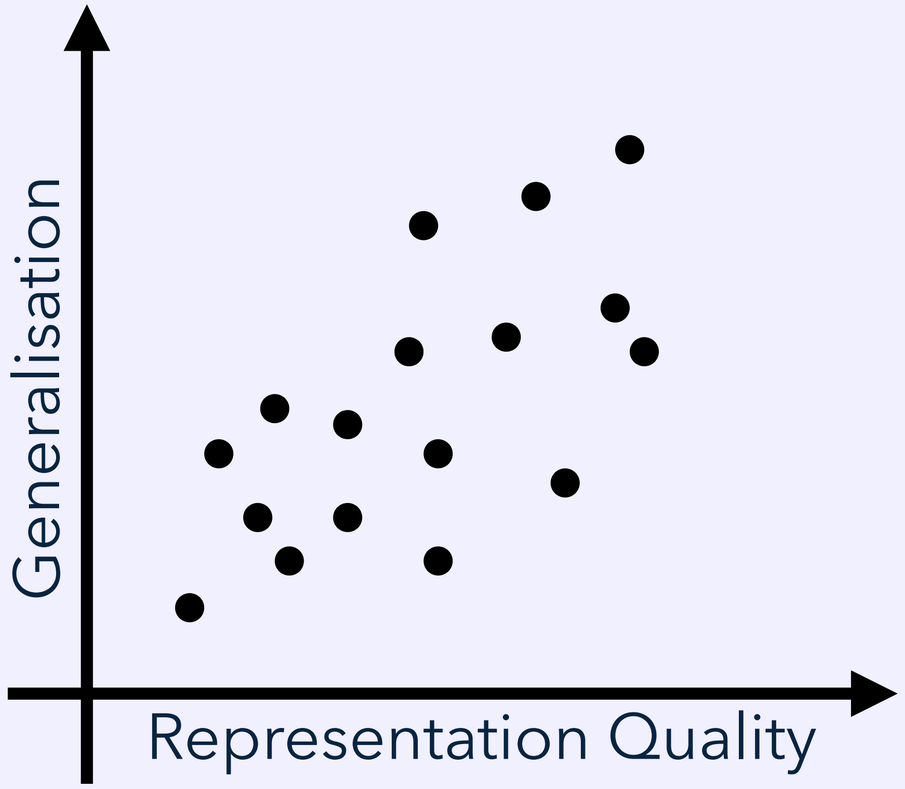
- Goal: bound generalisation performance
- Uninformative bounds
- Making significant progress but still far from capturing generalisation

My belief: we are missing the right intuitions

Where is Generalisation Right Now?

Practice

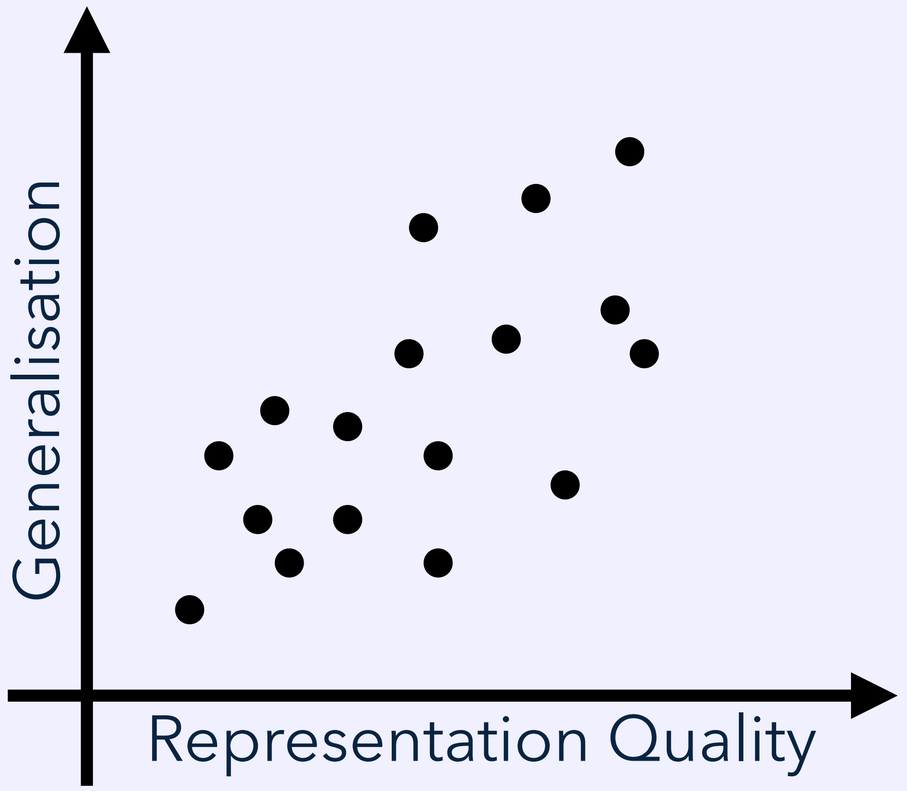
- Goal: predict performance of a learned model



Where is Generalisation Right Now?

Practice

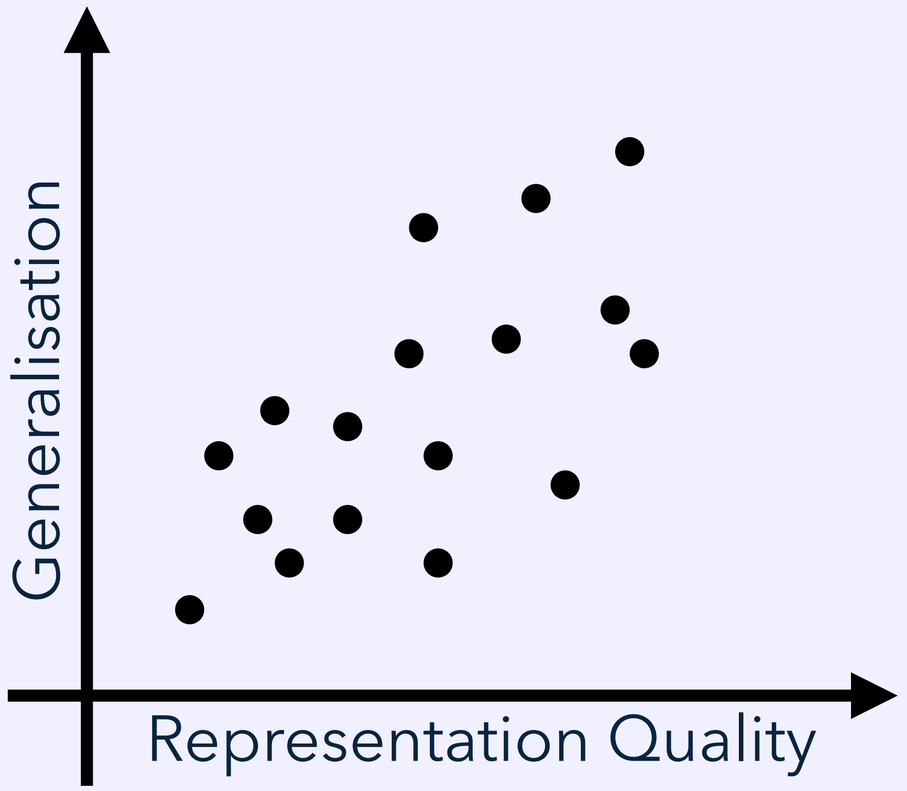
- Goal: predict performance of a learned model
- Poorly founded



Where is Generalisation Right Now?

Practice

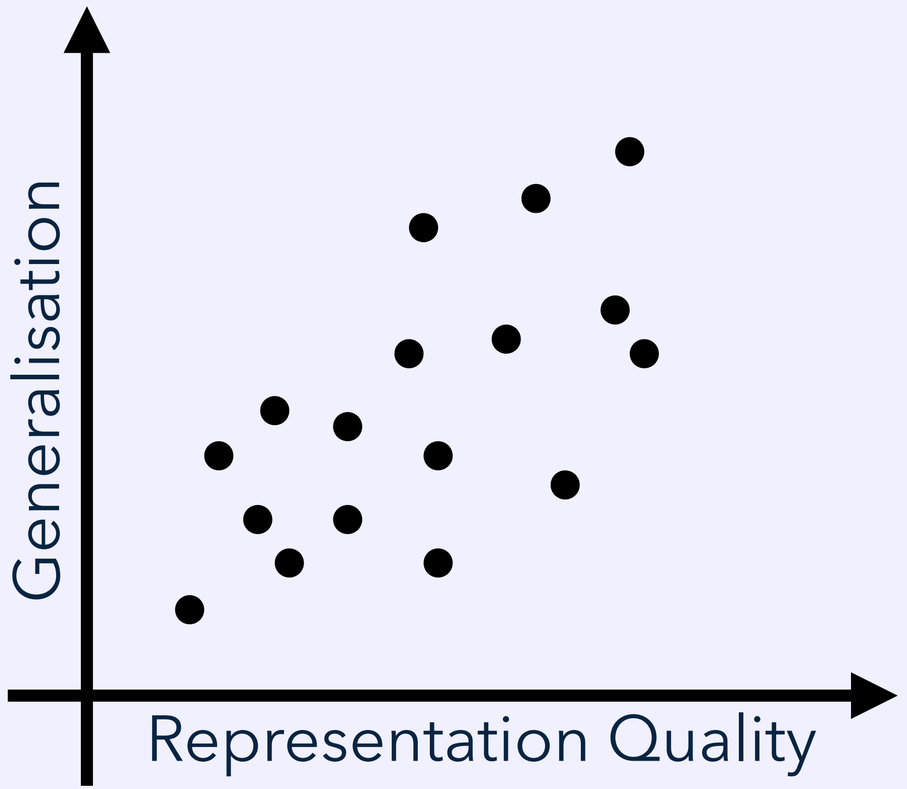
- Goal: predict performance of a learned model
- Poorly founded
- Difficult to create fair and extensive evaluation



Where is Generalisation Right Now?

Practice

- Goal: predict performance of a learned model
- Poorly founded
- Difficult to create fair and extensive evaluation
- Can help build intuitions



Estimating Generalisation

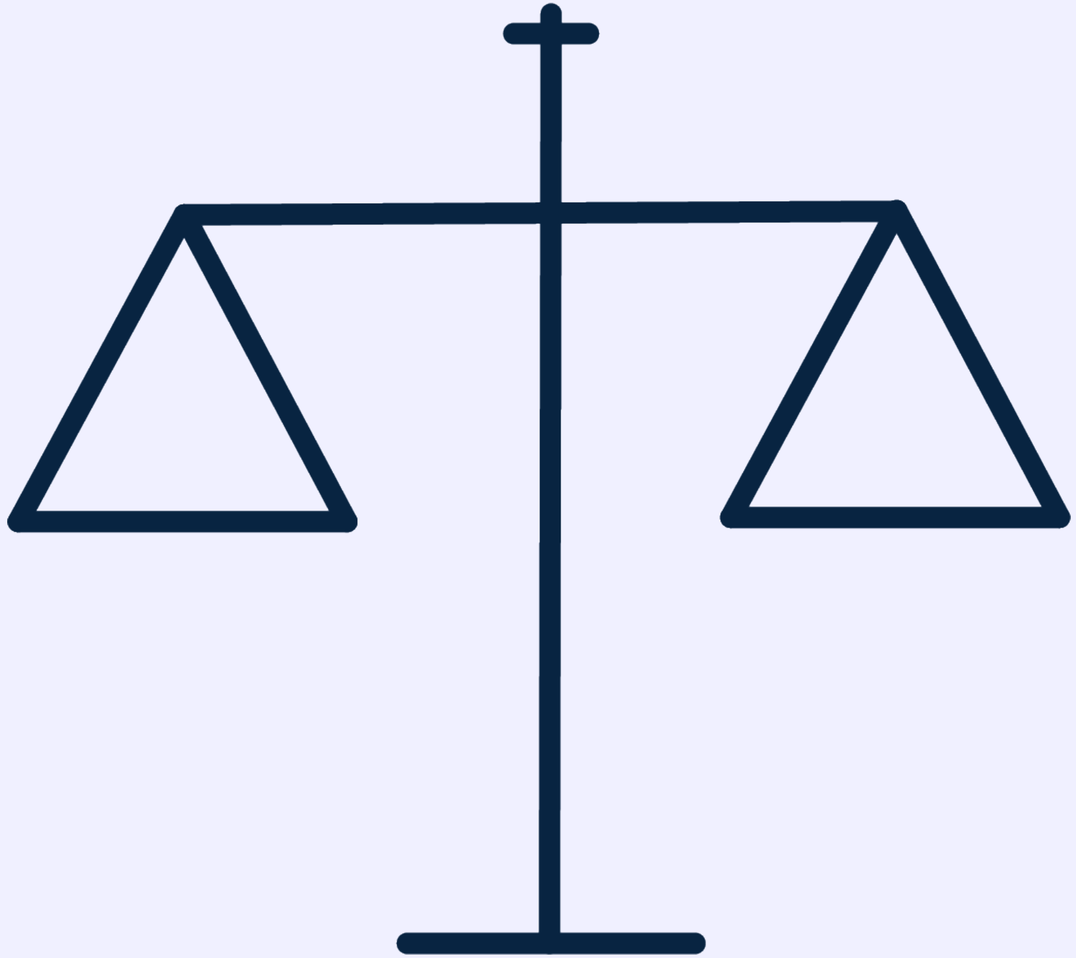
Model Instance Quality

Estimating Generalisation

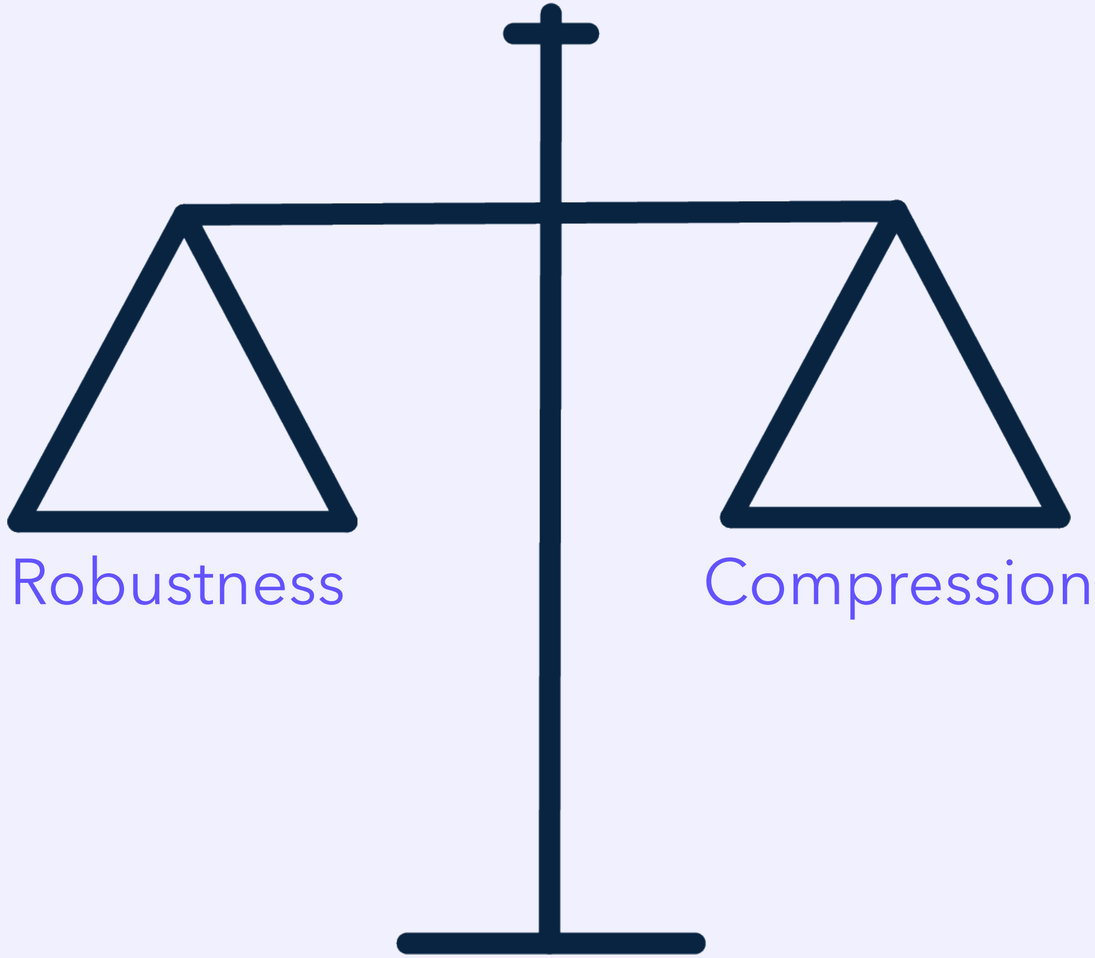
Model Instance Quality

My Perspective

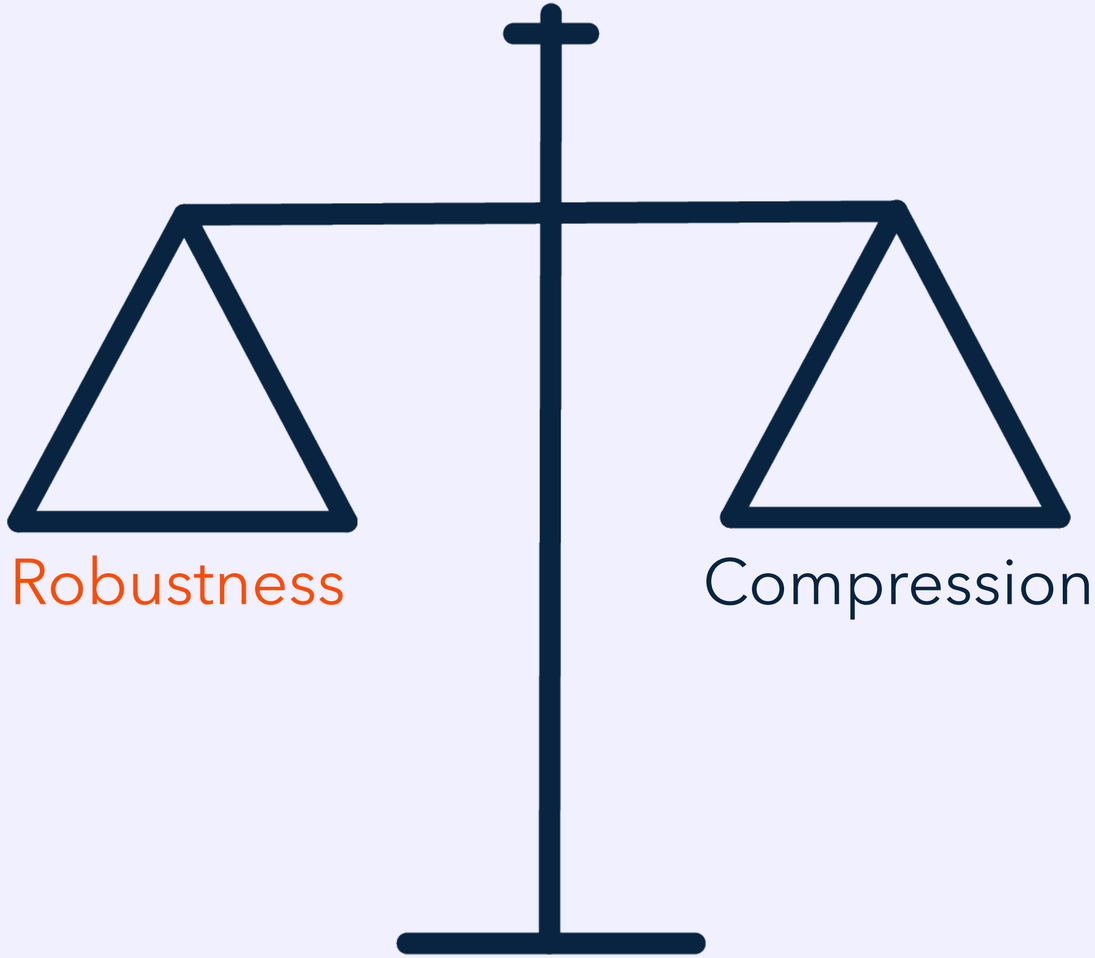
My Perspective



My Perspective



My Perspective



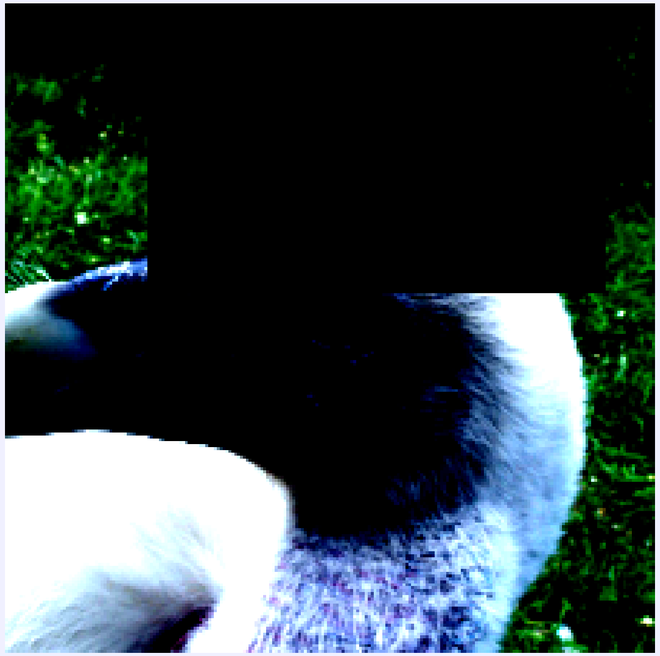
How Has Robustness Been Measured?

Data Distortions

Modifications to Image Pixels

Data Distortions

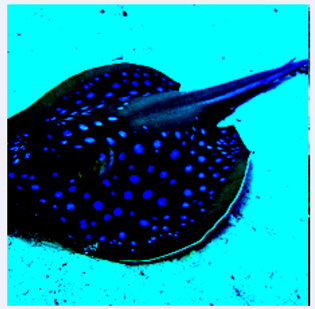
Modifications to Image Pixels



Data Distortions

Mixed Sample Data Augmentation (MSDA)

Source Image 1

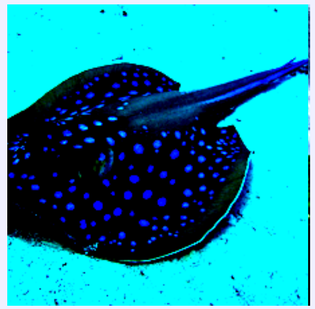


Source Image 2

Data Distortions

Mixed Sample Data Augmentation (MSDA)

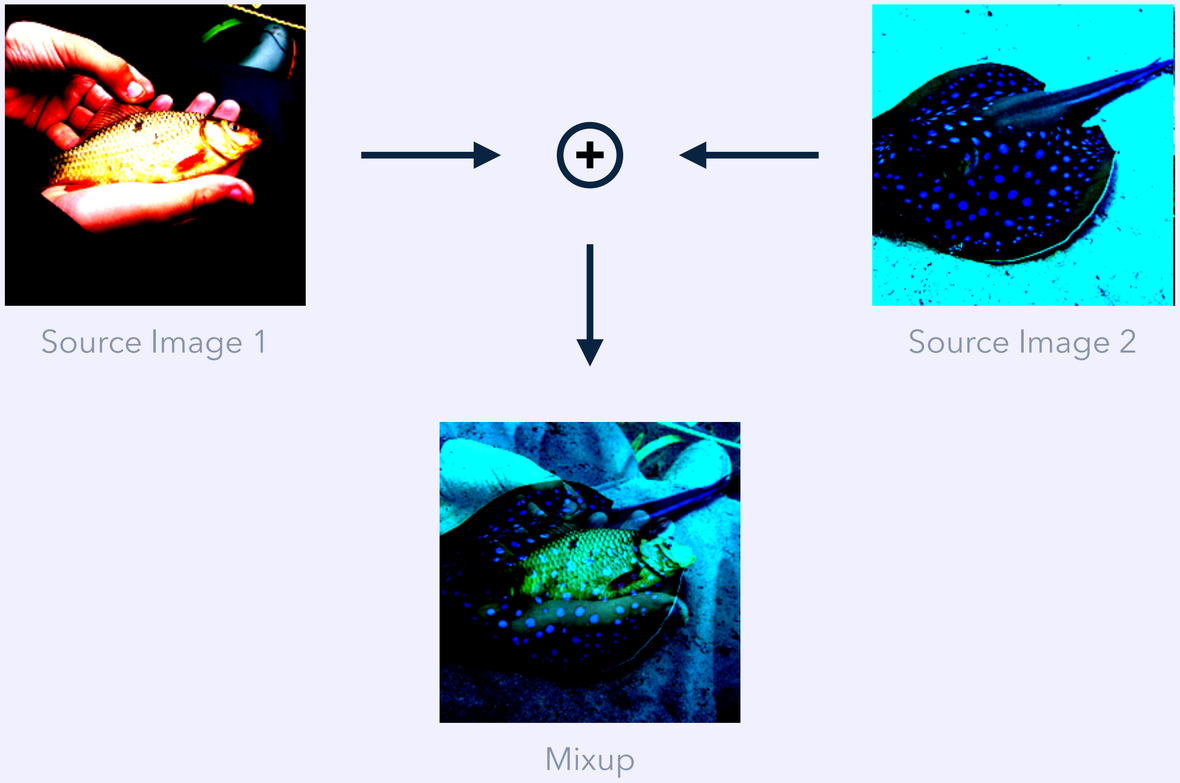
Source Image 1



Source Image 2

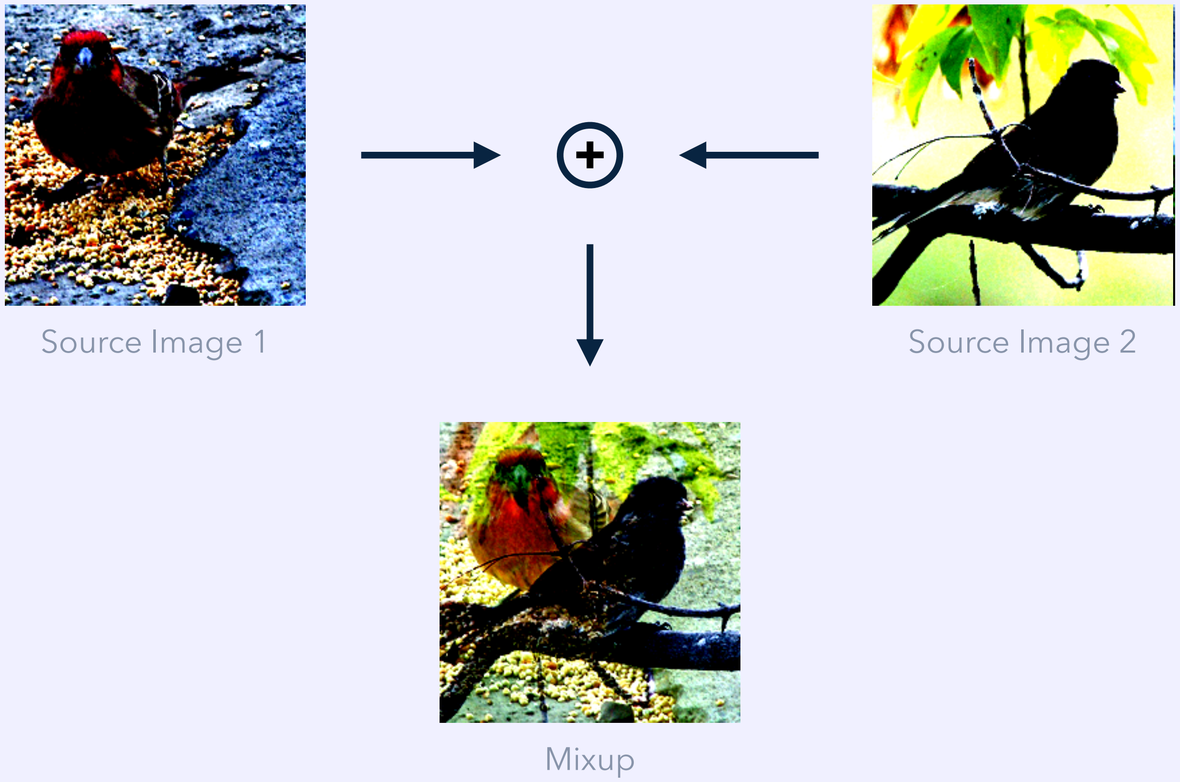
Data Distortions

Mixed Sample Data Augmentation (MSDA)



Data Distortions

Mixed Sample Data Augmentation (MSDA)



Data Distortions

Mixed Sample Data Augmentation (MSDA)

Source Image 1

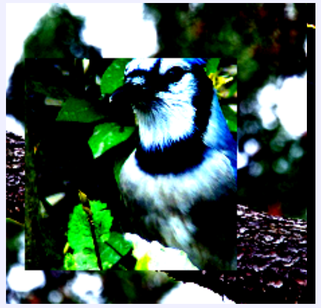
Source Image 2

Data Distortions

Mixed Sample Data Augmentation (MSDA)

Source Image 1

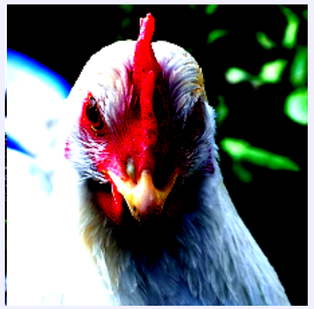
Source Image 2



CutMix

Data Distortions

Mixed Sample Data Augmentation (MSDA)



Source Image 1

Source Image 2

CutMix

Data Distortions

Mixed Sample Data Augmentation (MSDA)

Source Image 1

Source Image 2

Data Distortions

Mixed Sample Data Augmentation (MSDA)

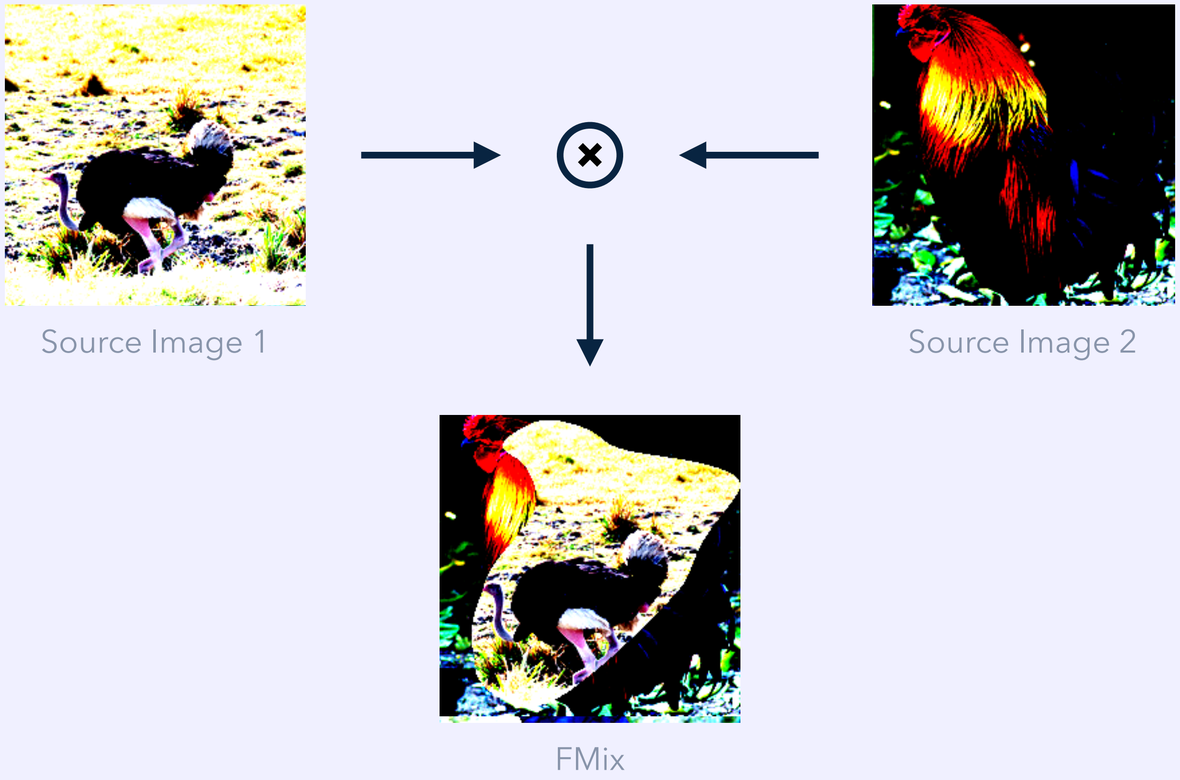
Source Image 1

Source Image 2

FMix

Data Distortions

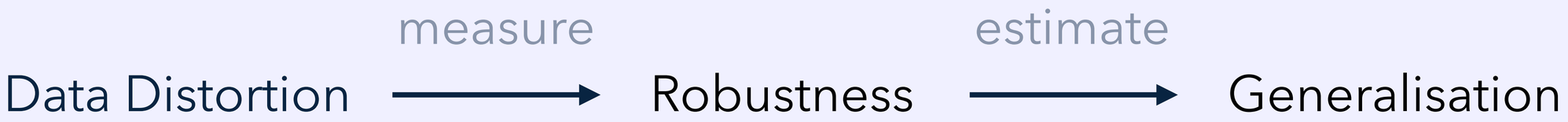
Mixed Sample Data Augmentation (MSDA)



Putting Things Together

Putting Things Together

Putting Things Together



Putting Things Together

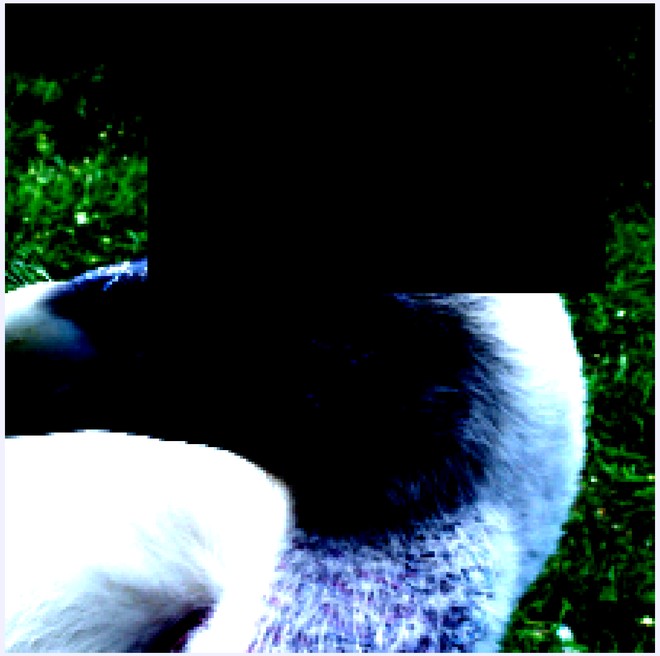
Pitfalls of Measuring Robustness Through Distortion

Confounding Factors

Introducing Artefacts

Confounding Factors

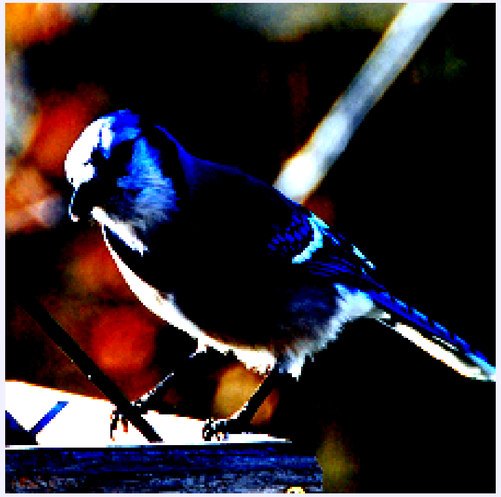
Introducing Artefacts



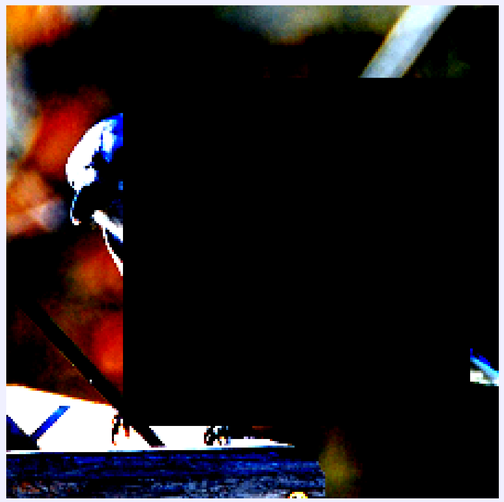
Are distorted images consistently associated with a particular class?

Increase in Incorrect Predictions

Increase in Incorrect Predictions

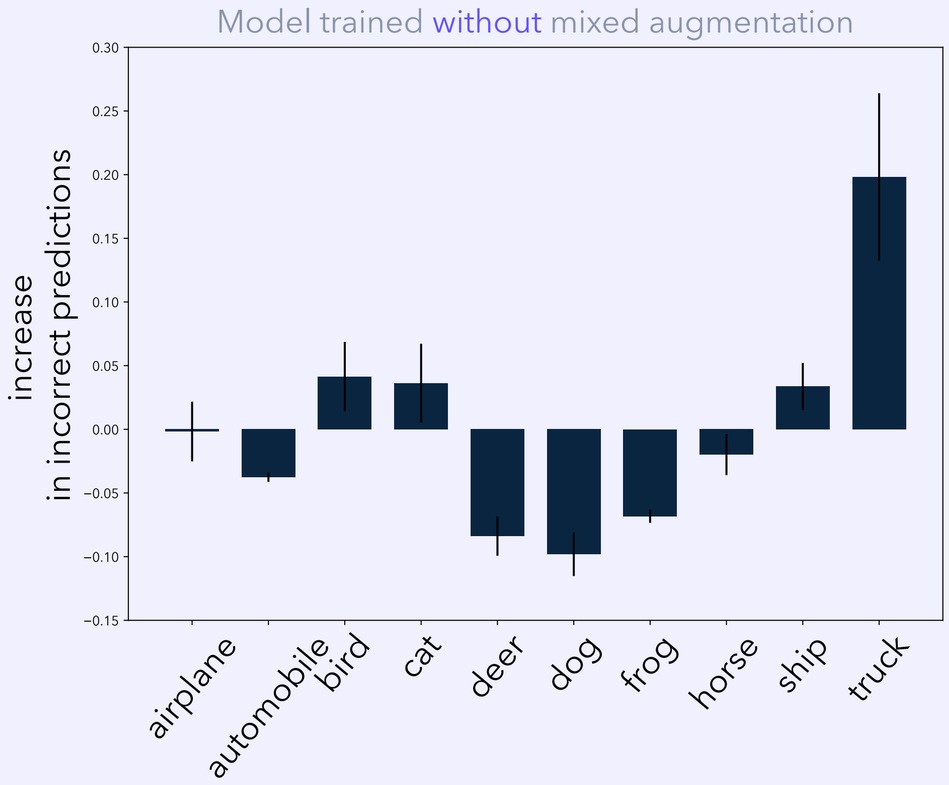


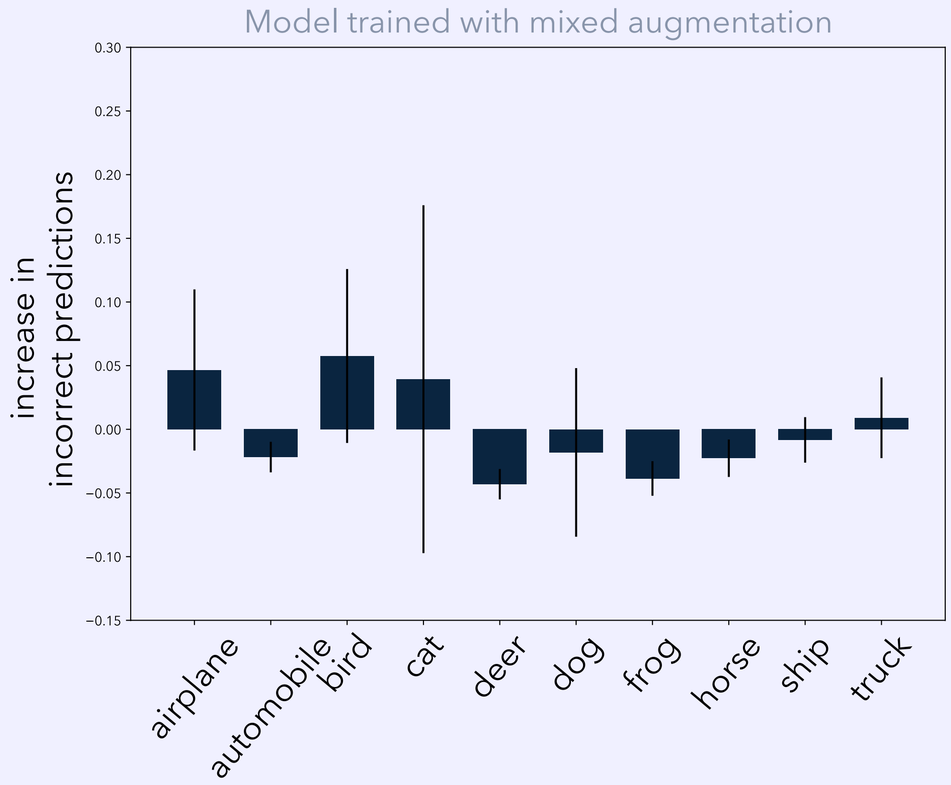
Increase in Incorrect Predictions



CIFAR-10 Example

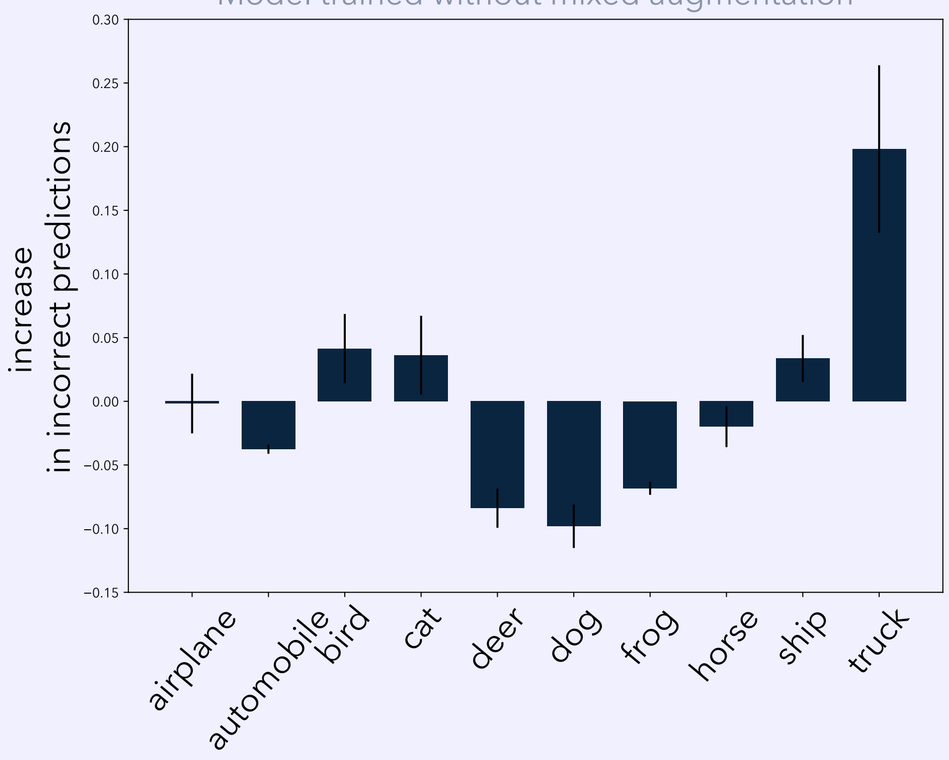
CIFAR-10 Example



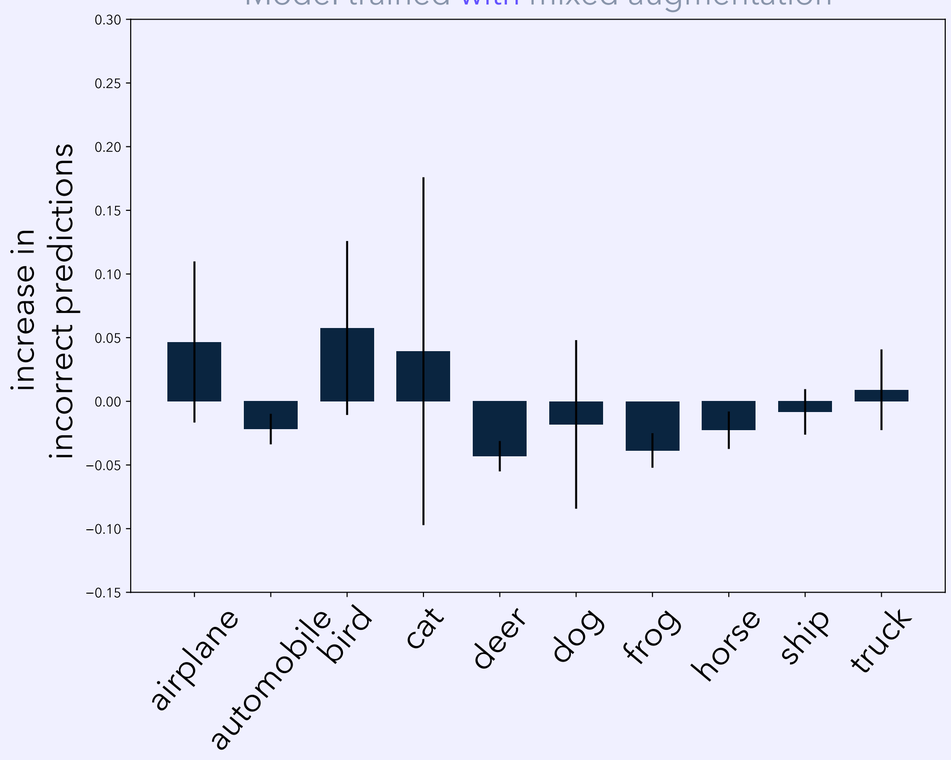


CIFAR-10 Example

Model trained without mixed augmentation



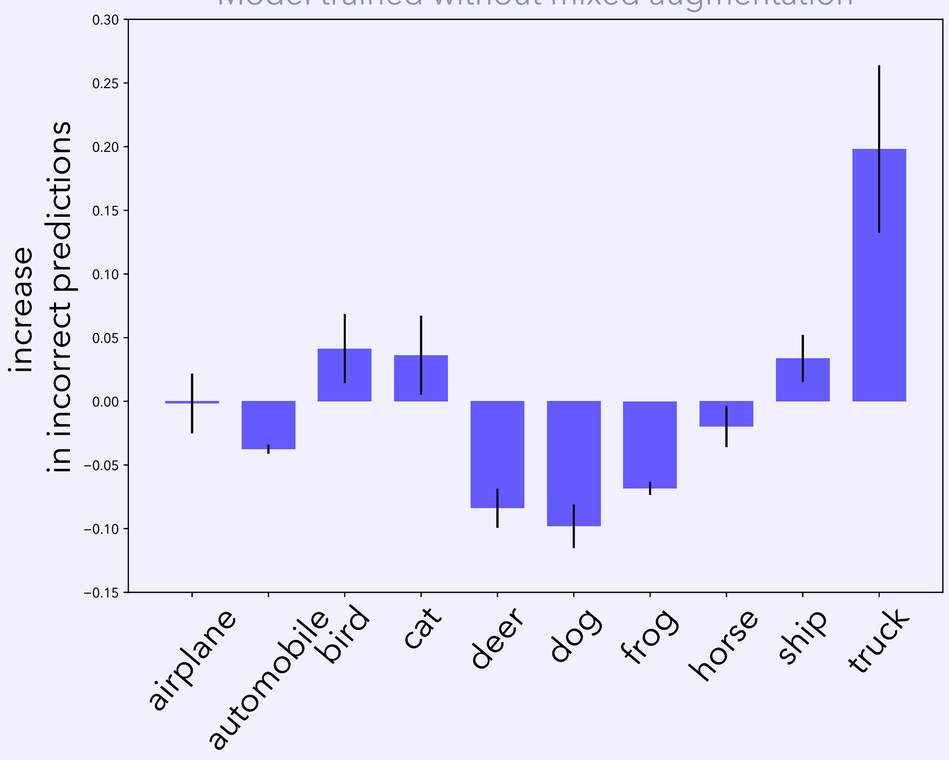
Model trained with mixed augmentation



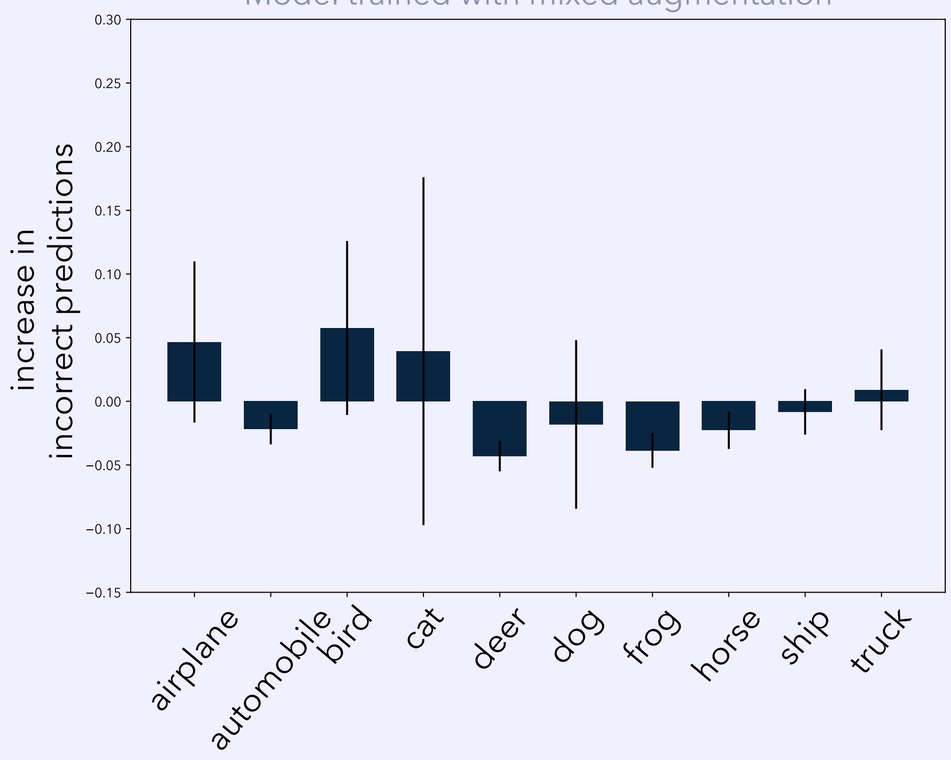
CIFAR-10 Example

Data Interference

Model trained without mixed augmentation



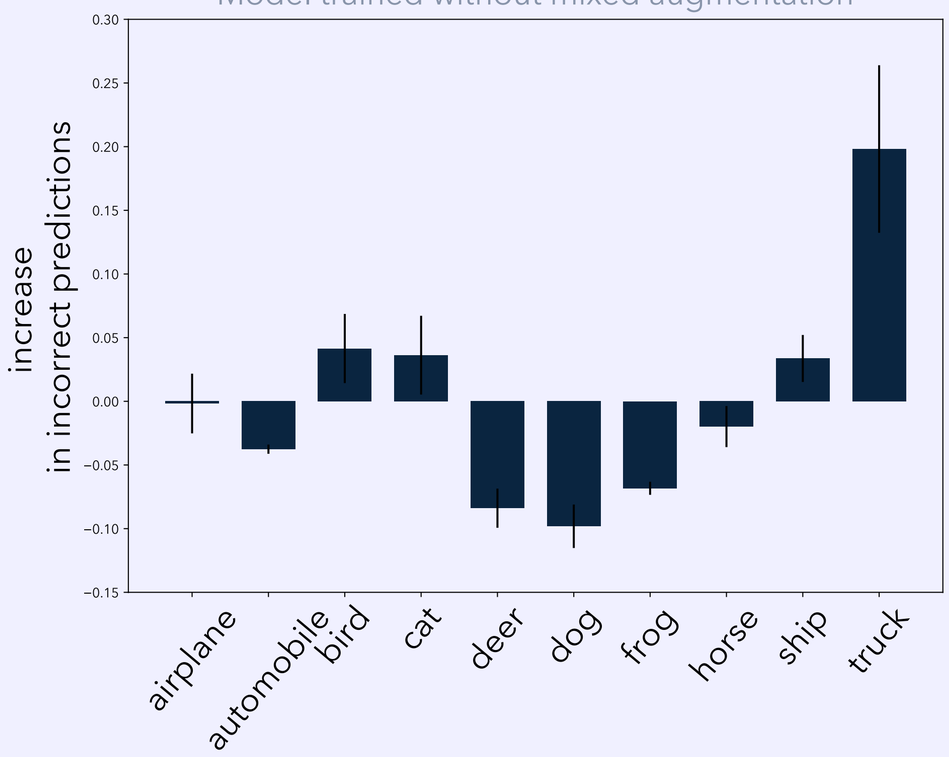
Model trained with mixed augmentation



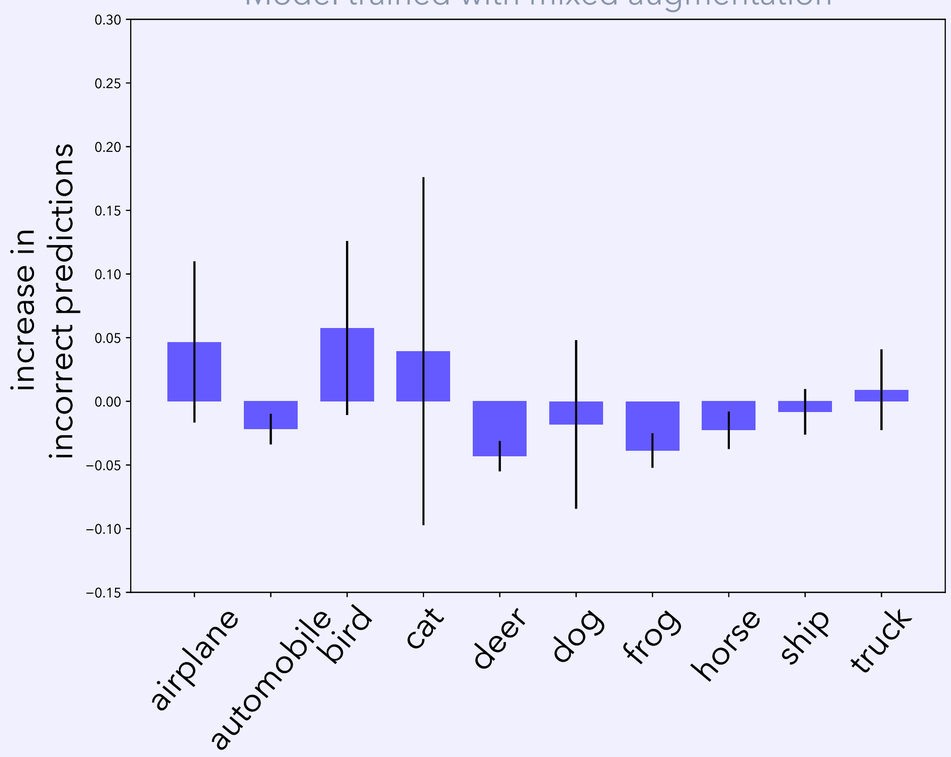
CIFAR-10 Example

Data Interference

Model trained without mixed augmentation



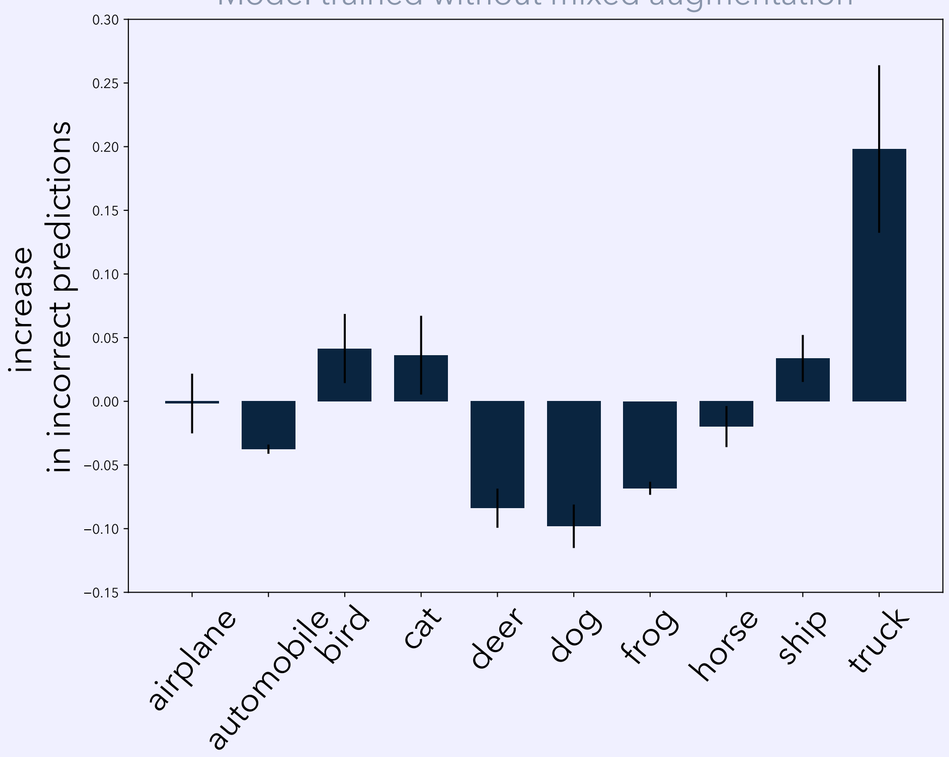
Model trained with mixed augmentation



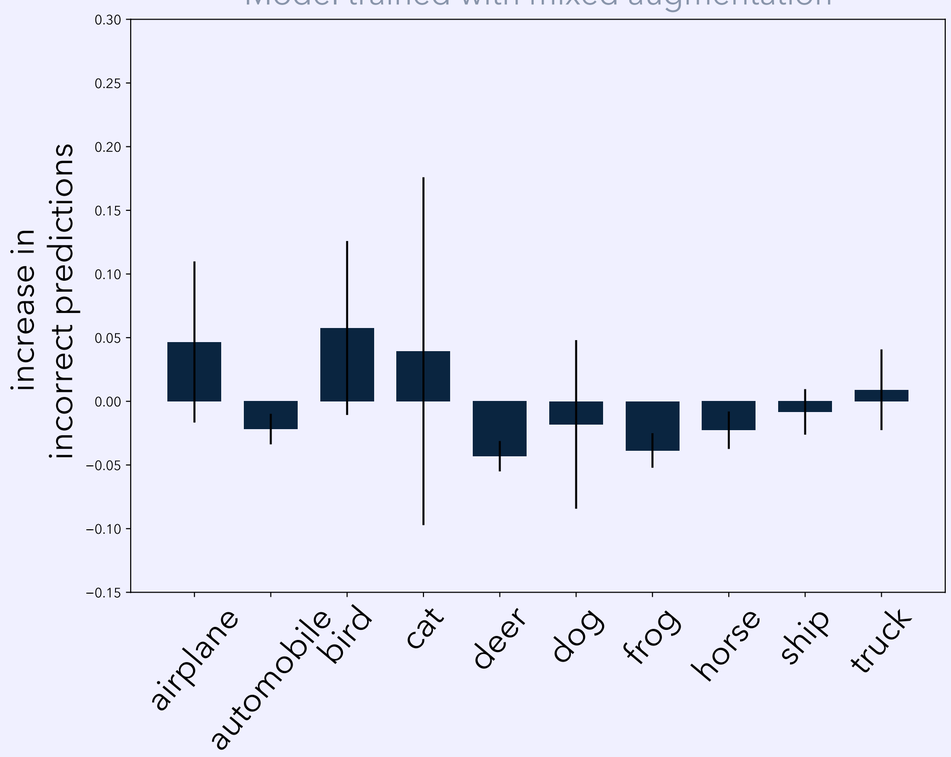
CIFAR-10 Example

Data Interference

Model trained without mixed augmentation



Model trained with mixed augmentation



Data Interference Index

- Model associates distortions with a particular class

- This happens consistently across runs

Data Interference Index

$$\mathbb{E}_r \left[\frac{\mathrm{inc}_r(c_{\max})}{\sum_c \mathrm{inc}_r(c)} \, \mathbb{E}_{r'} \left[\mathrm{inc}_{r'}(c_{\max}) \right] \right]$$

- $\mathrm{inc}_r(c)$: increase of class c for run r
- $c_{\max}$: class with maximum average increase across runs

Interference Occurs in Benchmark Data Sets

- Models: Basic, FMix, MixUp, CutMix

Interference Occurs in Benchmark Data Sets

- Models: Basic, FMix, MixUp, CutMix

DI Index for identifying shape bias

	basic	MixUp	FMix	CutMix
CIFAR-10	$2.82_{\pm 0.44}$	$2.40_{\pm 0.59}$	$0.59_{\pm 0.12}$	$0.31_{\pm 0.10}$
CIFAR-100	$0.99_{\pm 0.27}$	$0.88_{\pm 0.24}$	$0.18_{\pm 0.10}$	$0.09_{\pm 0.04}$
Tiny ImageNet	$1.28_{\pm 1.13}$	$0.57_{\pm 0.11}$	$0.67_{\pm 0.10}$	$0.25_{\pm 0.11}$
ImageNet	0.82	1.49	0.58	–

DI Index for robustness to occlusion

	basic	MixUp	FMix	CutMix
CIFAR-10	$1.25_{\pm 0.17}$	$0.47_{\pm 0.11}$	$0.11_{\pm 0.04}$	$2.20_{\pm 0.81}$
CIFAR-100	$1.24_{\pm 0.35}$	$0.34_{\pm 0.09}$	$0.12_{\pm 0.10}$	$1.06_{\pm 0.32}$
FashionMNIST	$0.21_{\pm 0.08}$	$0.38_{\pm 0.06}$	$0.16_{\pm 0.05}$	$0.12_{\pm 0.05}$
Tiny ImageNet	$0.52_{\pm 0.17}$	$0.39_{\pm 0.03}$	$0.14_{\pm 0.04}$	$3.46_{\pm 2.45}$
ImageNet	0.50	1.50	0.50	–

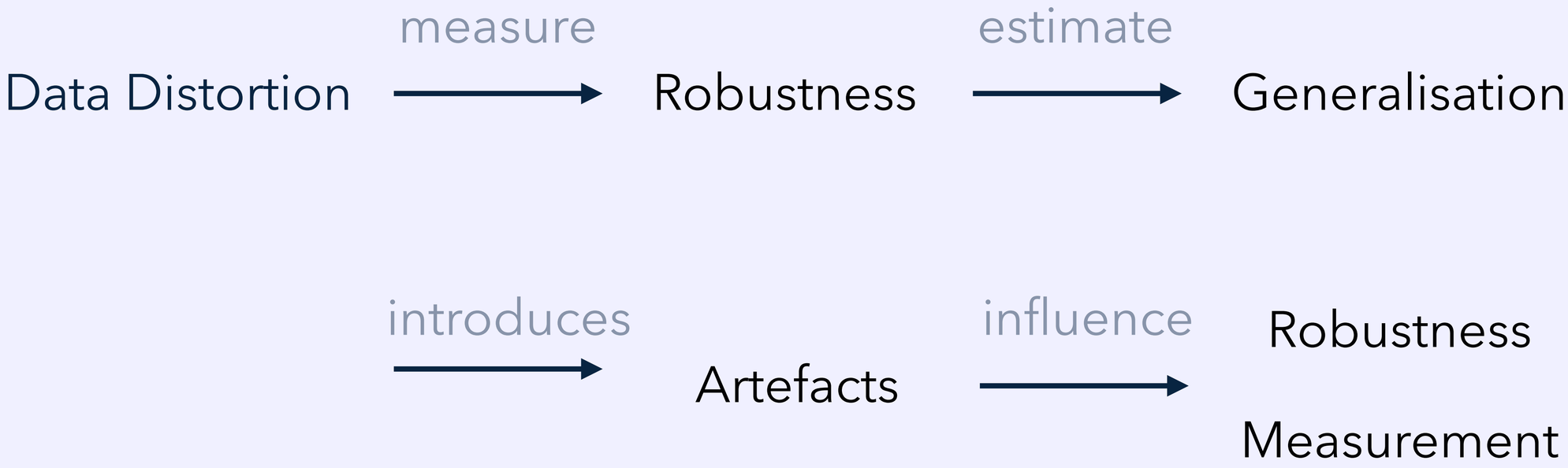
Interference Occurs in Benchmark Data Sets

- This affects robustness measurements

Putting Things Together

Putting Things Together

Putting Things Together



Are there fairer alternatives?

iOcclusion

An Alternative Measurement

Prior Approach: CutOcclusion

Measure: accuracy on distorted data

Prior Approach: CutOcclusion

Measure: accuracy on distorted data

Sensitive to:

- Shape of the occluder

Prior Approach: CutOccclusion

Measure: accuracy on distorted data

Sensitive to:

- Shape of the occluder
- Local information inside the occluding patch

Prior Approach: CutOccclusion

Measure: accuracy on distorted data

Sensitive to:

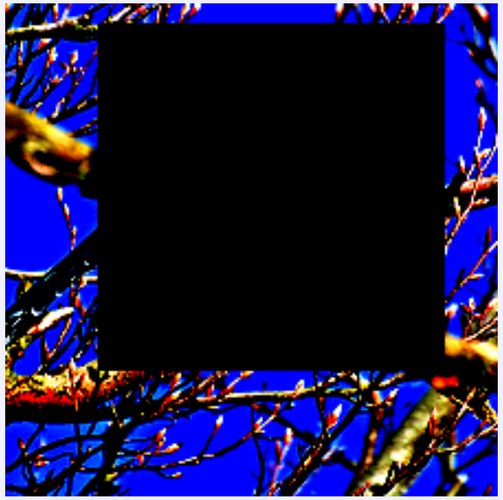
- Shape of the occluder
- Local information inside the occluding patch
- Overall model performance

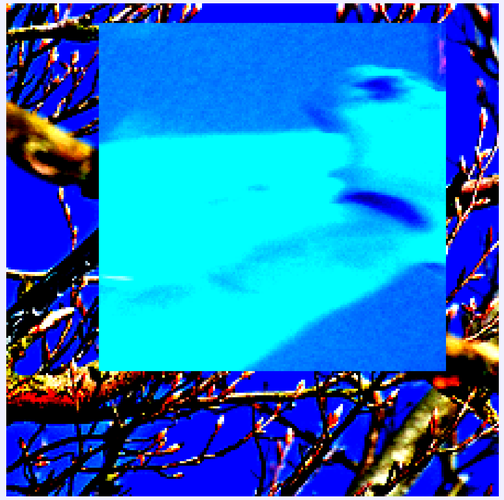
iOcclusion

$$\left| \frac{\mathcal{A}(\mathcal{D}^{p}_{\text{train}}) - \mathcal{A}(\mathcal{D}^{p}_{\text{test}})}{\mathcal{A}(\mathcal{D}_{\text{train}}) - \mathcal{A}(\mathcal{D}_{\text{test}})} \right|$$

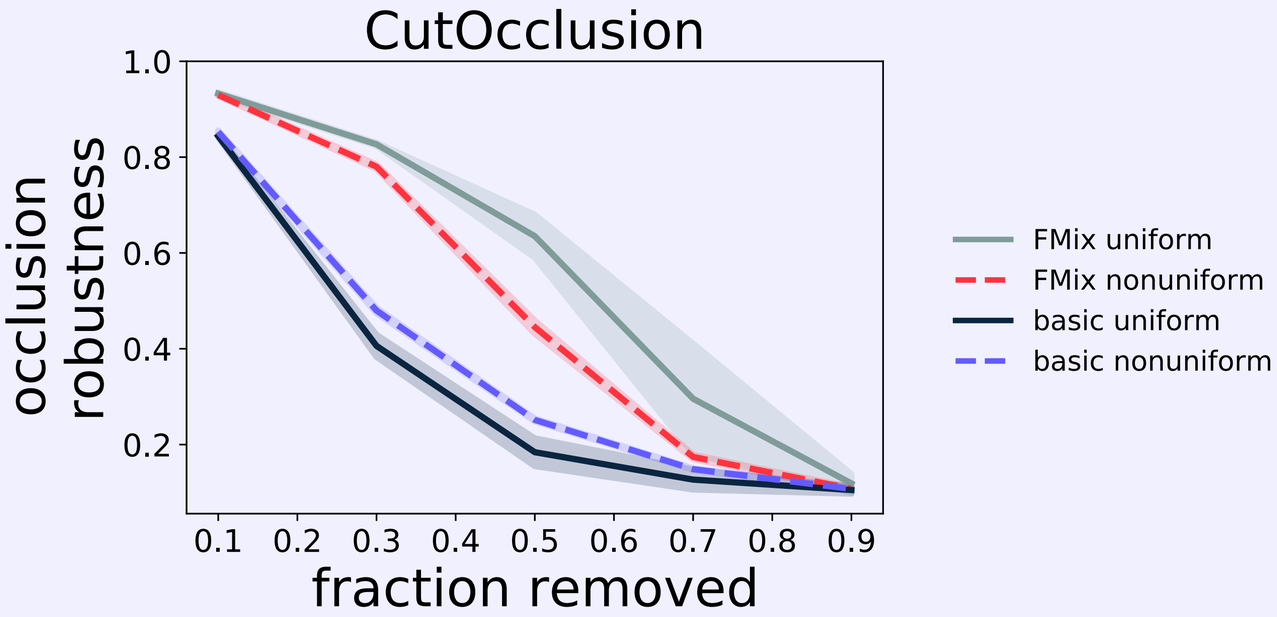
- $\mathcal{A}(\mathcal{D})$: accuracy on data set $\mathcal{D}$
- p: fraction of pixels distorted

Illustrative Example: Patch Information Invariance

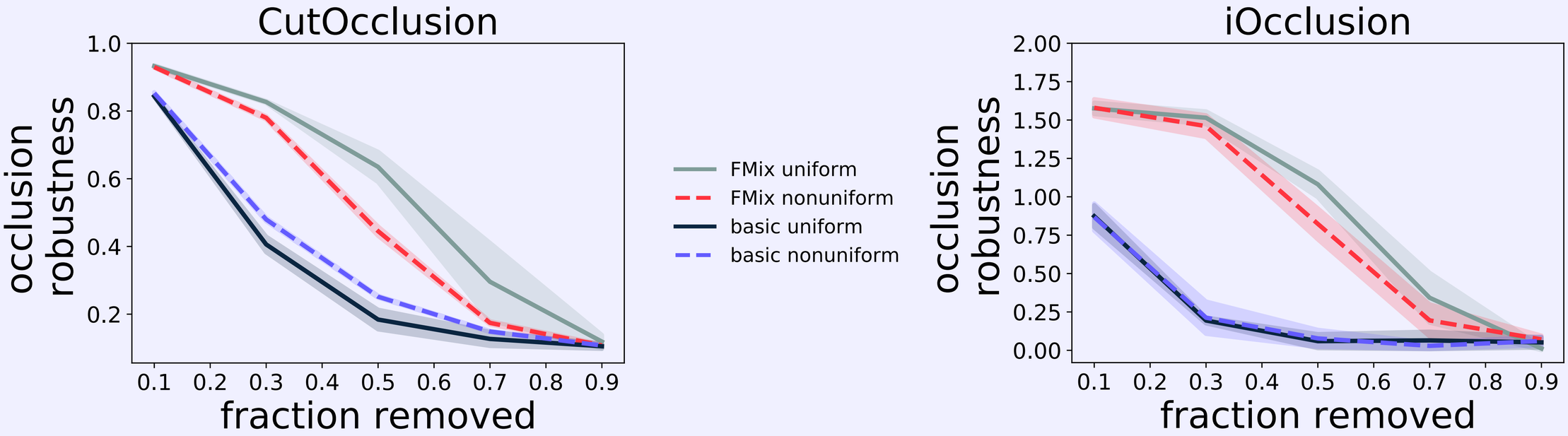




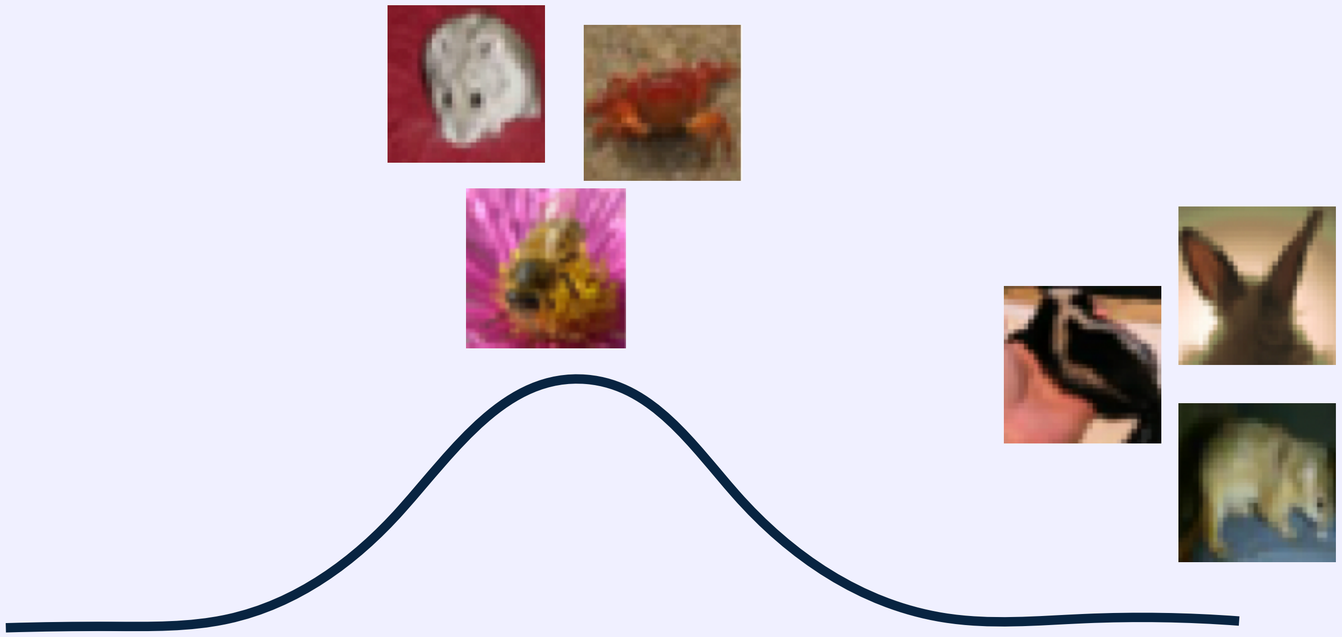
Illustrative Example: Patch Information Invariance



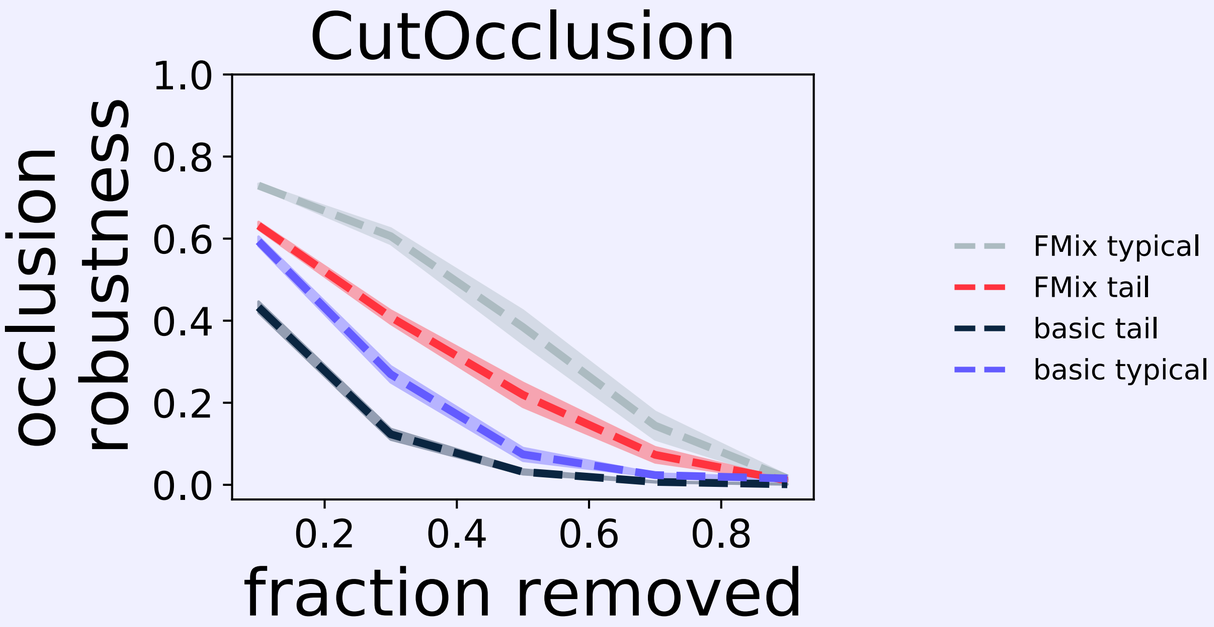
Illustrative Example: Patch Information Invariance



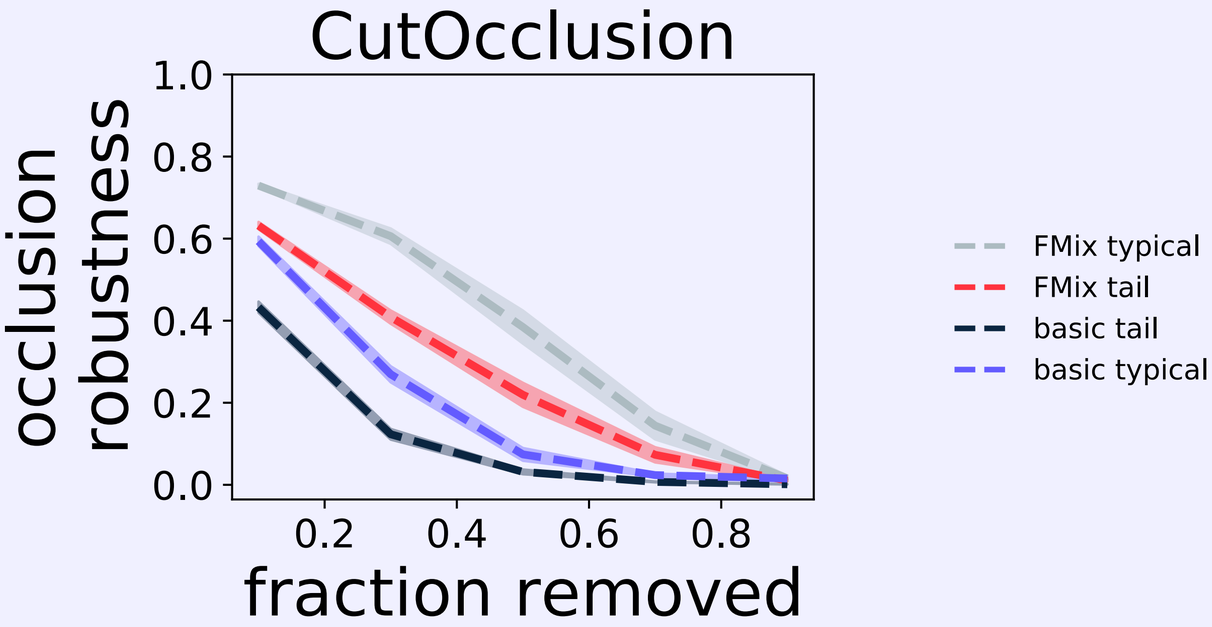
Illustrative Example: Performance Invariance

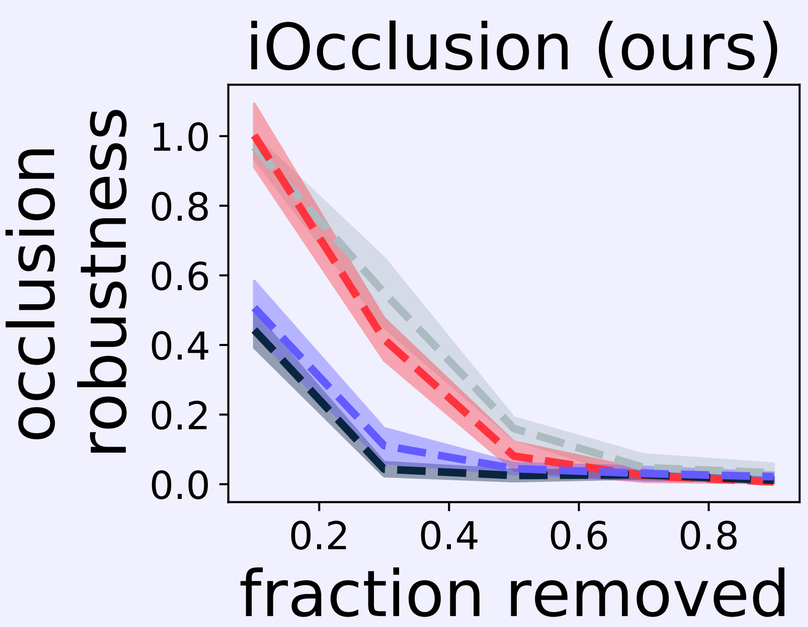


Illustrative Example: Performance Invariance



Illustrative Example: Performance Invariance





Limitations

Good for uncontrolled settings but …

- Relies on implicit assumption

Limitations

Good for uncontrolled settings but …

- Relies on implicit assumption
- Robustness to occlusion only

That still leaves us with the need for a way of measuring robustness more generally

Empirically Predicting Generalisation

- Predicting Generalisation in Deep Learning Challenge

Empirically Predicting Generalisation

- Predicting Generalisation in Deep Learning Challenge

- Why would model that is more robust to this specific distortion necessarily be better than another?

Empirically Predicting Generalisation

	Model 1	Model 2
Test accuracy	$95.51_{\pm 0.10}$	$93.39_{\pm 0.52}$
MixUp accuracy	$87.18_{\pm 0.19}$	$87.53_{\pm 0.62}$
Compression	$4.25_{\pm 0.03}$	$4.27_{\pm 0.02}$

- Models with overlapping compression and MixUp accuracy can have different generalisation performance

Empirically Predicting Generalisation

- Similar argument holds for any particular distortion

- Can we think of all the possible distortions and evaluate against them?

So how can we better measure and enforce robustness?

Decision Decompositionality

Vision for Future Work

Occlusion experiments

+

Simplicity bias

Decision is over-reliant

on a very small

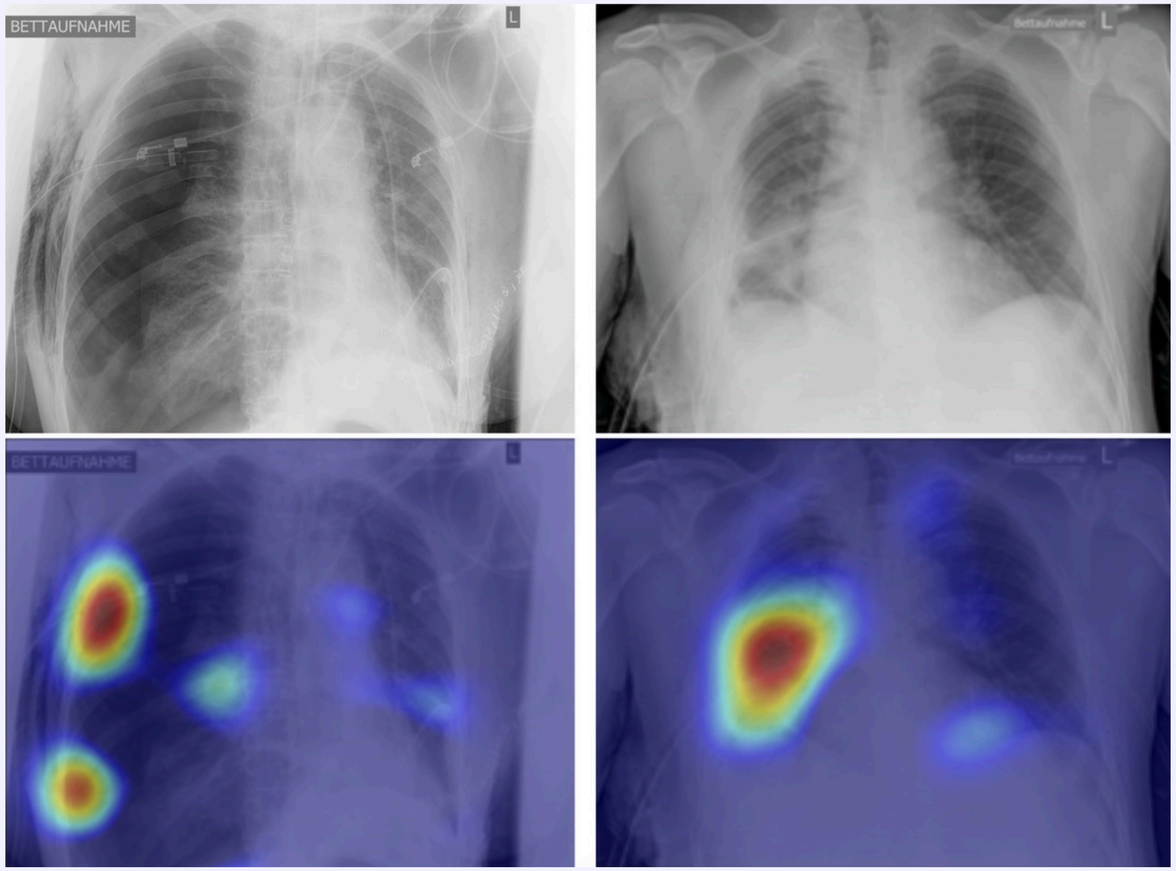
number of features

Decision Decompositionality

Vision for Future Work

Can we enforce a more “distributed”/“decoupled” decision?

Going Back...



Take-aways

- Measuring robustness in an unbiased way is difficult
- Doing so can help us better understand generalisation
- Decision “decompositionality” could be a way forward