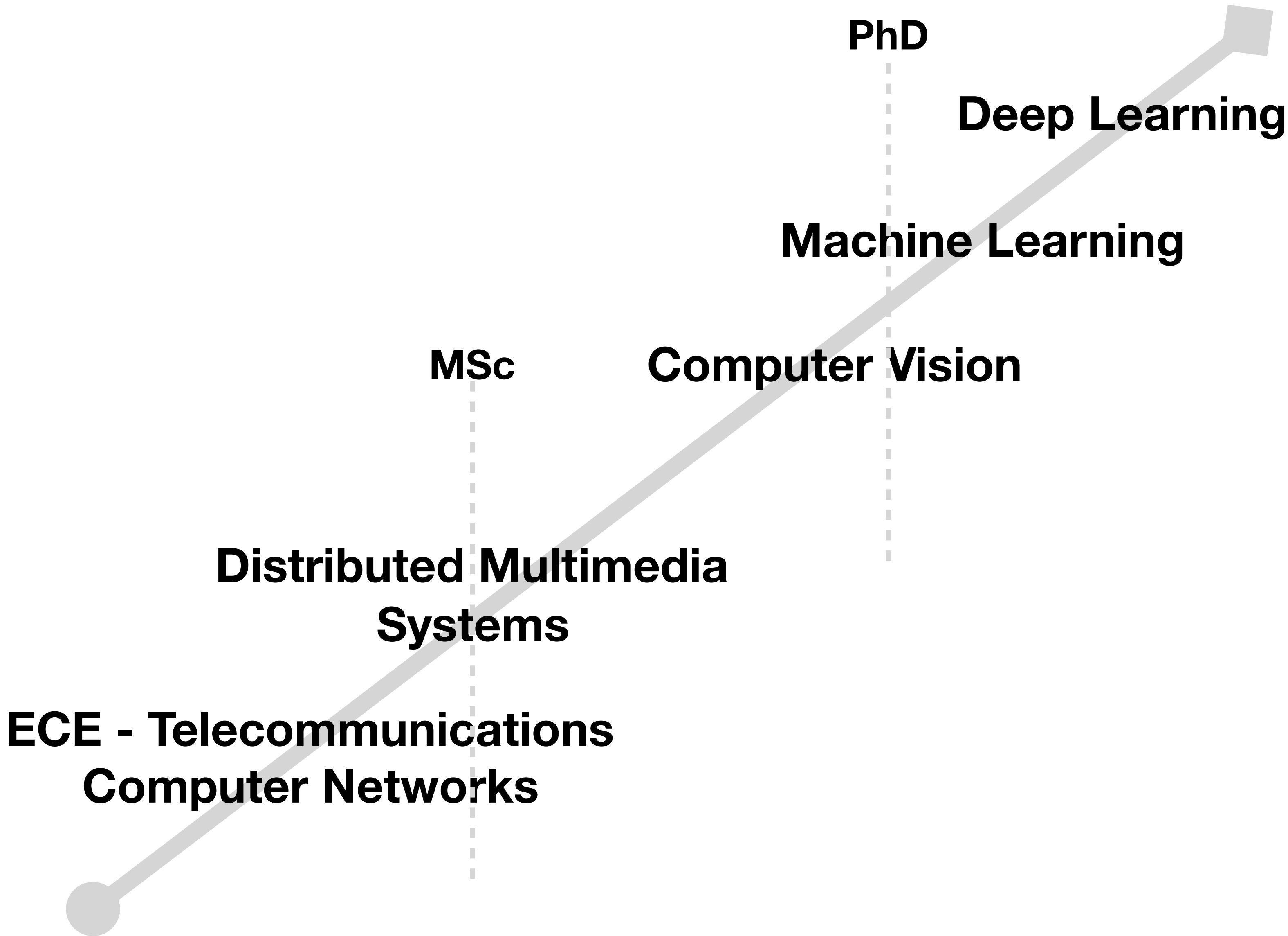
Producing Joint Decisions and Explanations with CNNs

Luis F. Teixeira
Universidade do Porto / FEUP, INESC TEC

University of Twente - July 1, 2019


PhD
Deep Learning
Machine Learning
MSc
Computer Vision
Distributed Multimedia Systems
ECE - Telecommunications
Computer Networks

PhD
"Contributions for the automatic description of multimodal scenes”

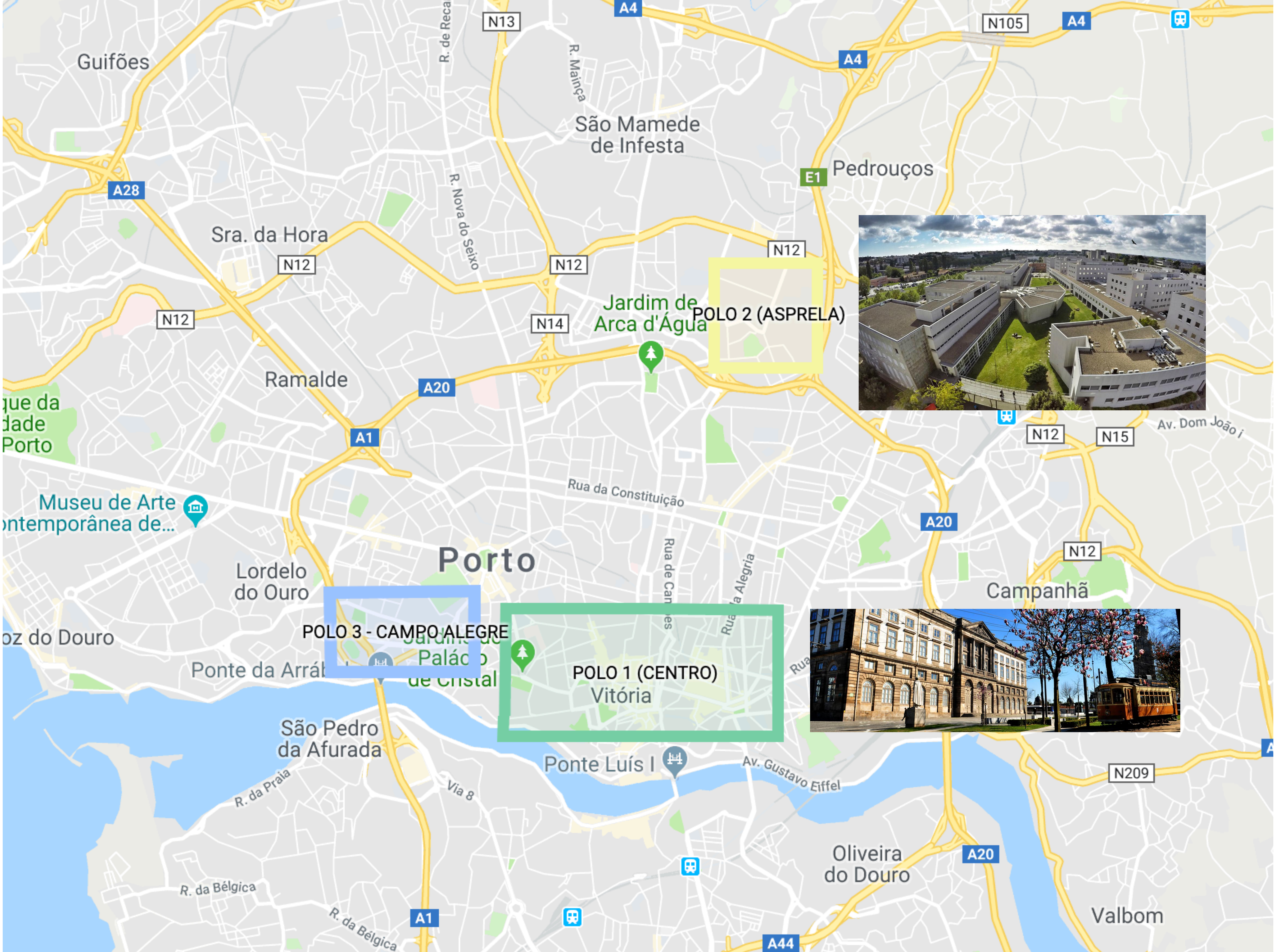
Post-Doc
Senior Scientist
"Assisted Living Solutions”

Assistant Professor
Informatics Engineering Department
Graphics Interaction and Games Group (GIG)
https://dei.fe.up.pt/gig

GIG main research areas:
- AR/VR interchangeability
- 360 multimedia
- Serious games
- Procedural 3D modeling

Senior Researcher
Visual Computing and Machine Intelligence Group (VCMI)
https://vcmi.inesctec.pt/ (to be updated)

VCMI main areas of research:
- Medical Imaging
- Biometrics
- Machine/Deep Learning


POLO 2 (ASPRELA)
POLO 3 - CAMPO ALEGRE
POLO 1 (CENTRO)
Porto
Guifões
São Mamede de Infesta
Pedrouços
Sra. da Hora
Jardim de Arca d'Água
Ramalde
Museu de Arte Contemporânea de...
Lordelo do Ouro
Campanhã
Vitória
Ponte da Arrábida
São Pedro da Afurada
Ponte Luís I
Rua da Constituição
Av. Gustavo Eiffel
Oliveira do Douro
Valbom
R. da Bélgica
R. da Praia
Via 8
R. Nova do Seixo
R. Mainça
Av. Dom João I
A4
N13
N105
A28
N12
N14
E1
A20
A1
N15
N209
A44

Interpretability
+
Deep Learning


Input
x
ML/DL Model
Prediction
$\hat{y}$

Most of the cases we don't really know what is happening…

… (or care about)

https://xkcd.com/1838/

Interpretability

Interpretation is the process of giving **explanations**

To Humans

adapted from Kim and Doshi-Velez

Explanations

- Explanations are a small (less complex) “model” that focuses on a small portion of the data

- Desirable properties of explanations:
  - **Completeness** -> susceptible of being applied in other cases where the audience can verify the validity of that explanation
  - **Correctness** -> generate trust (i.e., be accurate)
  - **Compactness** -> succinct

Wilson Silva and Kelwin Fernandes and Maria J. Cardoso and Jaime S. Cardoso, ”Towards complementary explanations using Deep Neural Networks, Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2018

Why?

- **Safety** -> can help expose safety issues
- **Mismatched objectives and multi-objective trade-offs** -> what you optimise is not what you meant to optimise
- **Debugging** -> understand why the system doesn’t work, and fix it
- **Sensitive domain** -> decisions in medicine, criminal justice, etc
- **Legal/Ethics** -> legally required to provide an explanation (e.g. GDPR) and/or we don't want to discriminate against particular groups
- **…**

adapted from Kim and Doshi-Velez

How?

- Ideal case — supervised ML approach
  - A dataset containing $\{features_{k,i}, question_k, answer_k\}$

- (Almost) never the case —> proxy models or approaches are needed

How?

- **Pre-model**
  - Exploratory data analysis
  - Visualisation for data exploration

- **In-model**
  - Build inherently interpretable models (e.g. rule-based - decision trees, rule list, rule sets -, case-based)
  - Regularisation (e.g. sparsity, monotonicity)

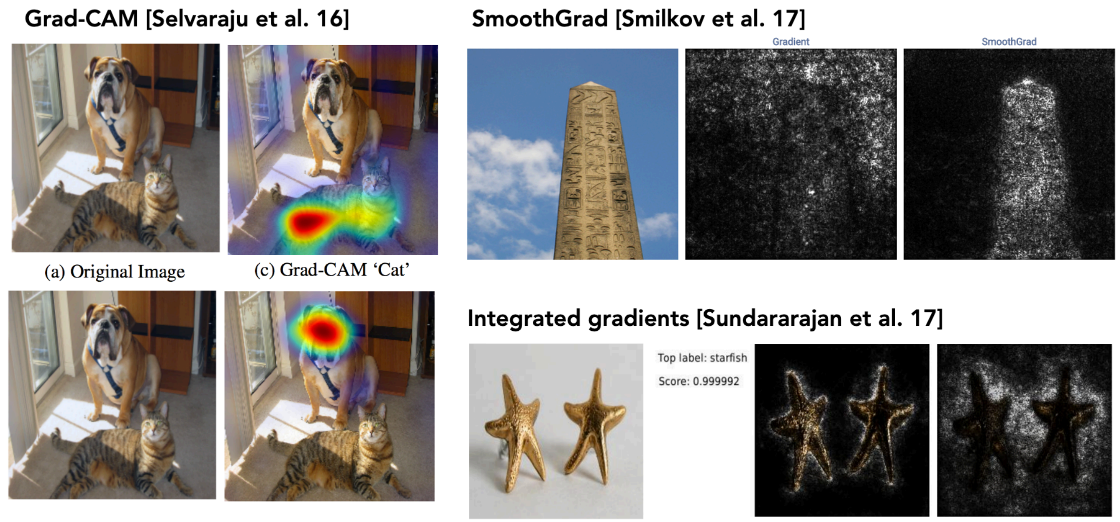
How?

- **Post-model**
  - White box
    - Saliency maps
    - Investigation on hidden activations
  - Black box
    - Sensitivity analysis
    - Mimic models

White Box - Saliency Maps

What are the features in the input space that influenced the most the classification?

$$\frac{\partial y}{\partial x_i}$$



backpropagate gradient of the output to the pixels in order to understand which pixels need to change the least to affect the class score the most

White Box - Hidden Layers

Gradient ascent (class model visualisation) — update the input image that maximizes the score of a certain class + some regularisation

Deconvolution — use deconvolution blocks to go from an activation map to a reconstructed image only with the most relevant parts

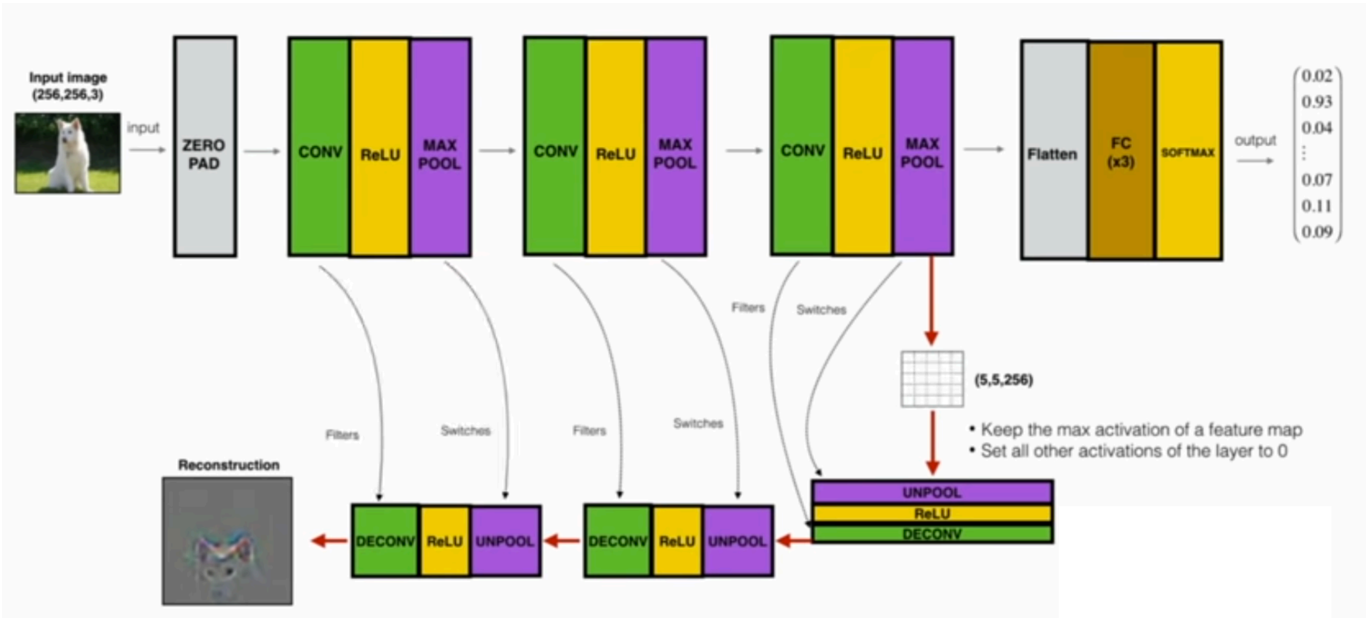


image from Stanford CS230, Fei-Fei Li and Justin Johnson

Black Box - Sensitivity Analysis

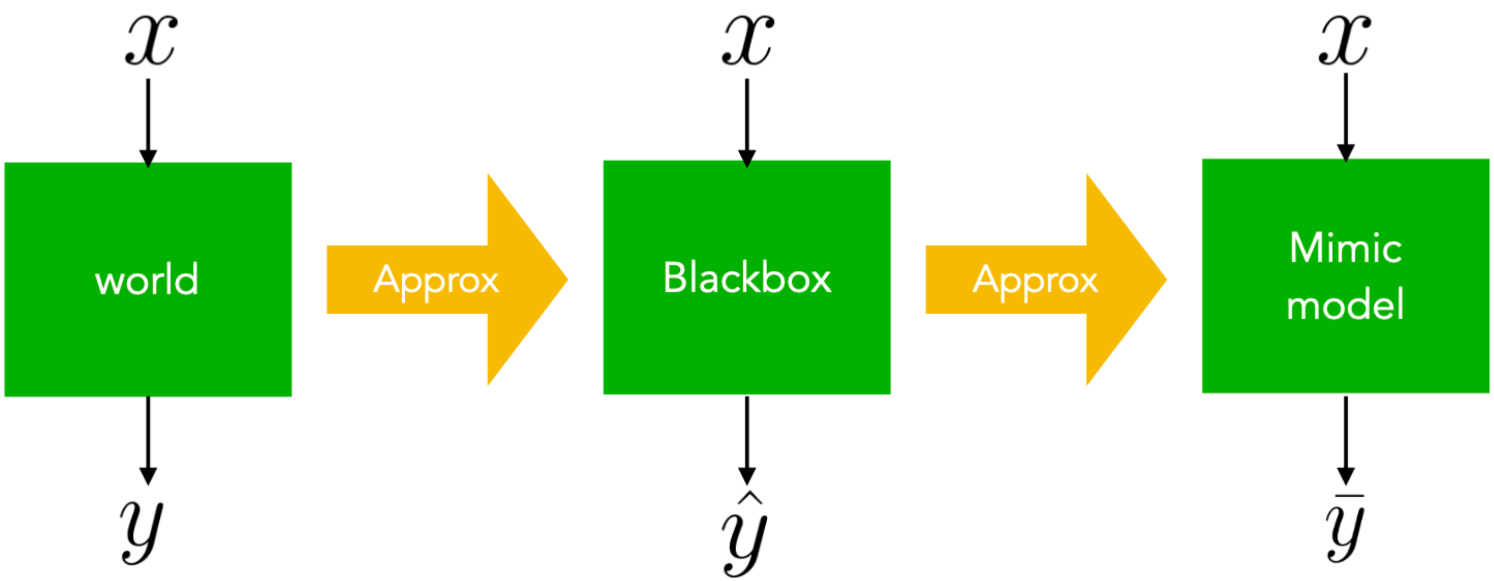
What would happen to output $\hat{y}$, if we perturb the input x?

$$x \to x + \epsilon$$

Occlusion sensitivity — occlude some part (sliding window) of the image and check how that affected the output

Black Box - Mimic Models

- Train a black box on x and y : $f(x) = \hat{y}$
- Train an interpretable model on x and $\hat{y}$: $f(x) = \bar{y}$



adapted from Kim and Doshi-Velez

Evaluation

Machine Learning

Function-based	Application-based
a variety of synthetic and standard benchmarks e.g, UCI datasets, imagenet	Backing up claims e.g., performance on a cool medical dataset, winning Go games

adapted from Kim and Doshi-Velez

Evaluation

Interpretable Machine Learning

	Function-based	Application-based
Quantitative	How sparse are the features?	How much did we improve patient outcomes?
	Does it look reasonable?	Do scientists find the explanations useful?

adapted from Kim and Doshi-Velez

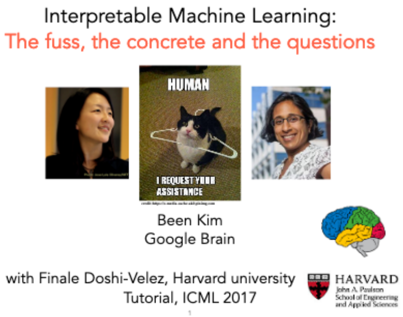
Evaluation

Interpretable Machine Learning

Function-based	Application-based
How sparse are the features?	How much did we improve patient outcomes?
Does it look reasonable?	Do scientists find the explanations useful?

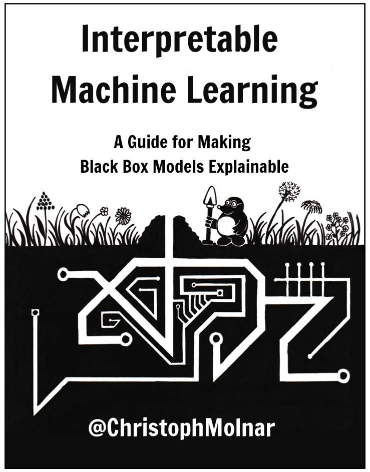
Qualitative

adapted from Kim and Doshi-Velez



Been Kim and Finale Doshi-Velez, “Interpretable Machine Learning: The fuss, the concrete and the questions”, ICML Tutorial, 2017

https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf



Interpretable Machine Learning

A Guide for Making Black Box Models Explainable”, Christoph Molnar, 2019

https://christophm.github.io/interpretable-ml-book/

Towards a Joint Approach to Produce Decisions and Explanations Using CNNs

Isabel Rio-Torto, Kelwin Fernandes, Luis F. Teixeira

IbPRIA 2019

(to be presented tomorrow - shortlisted for best paper)

Background

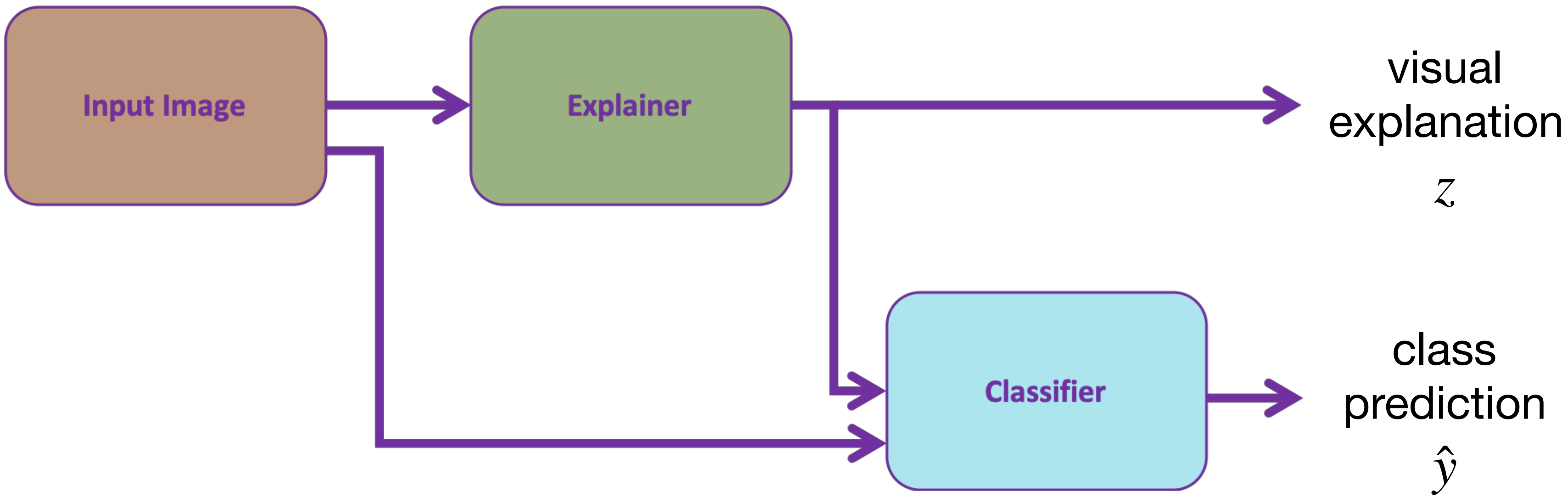
- Interpretability still lacks a unified formal definition and metrics
- Definition used (L.H. Gilpin *et al.* : “Explaining explanations: An overview of interpretability of machine learning):
  - explainability > interpretability

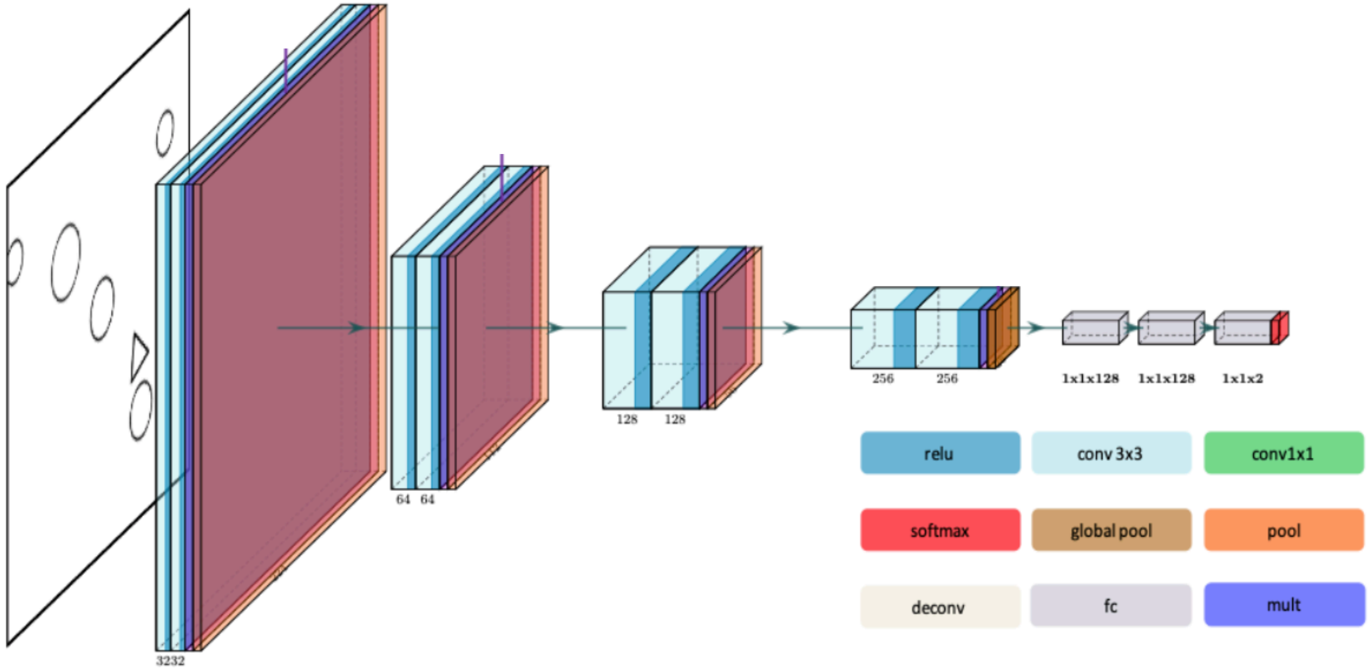
Explainable Model

- **Explainable model** is one that can **summarise the reasons** for its behaviour or the causes of its decisions

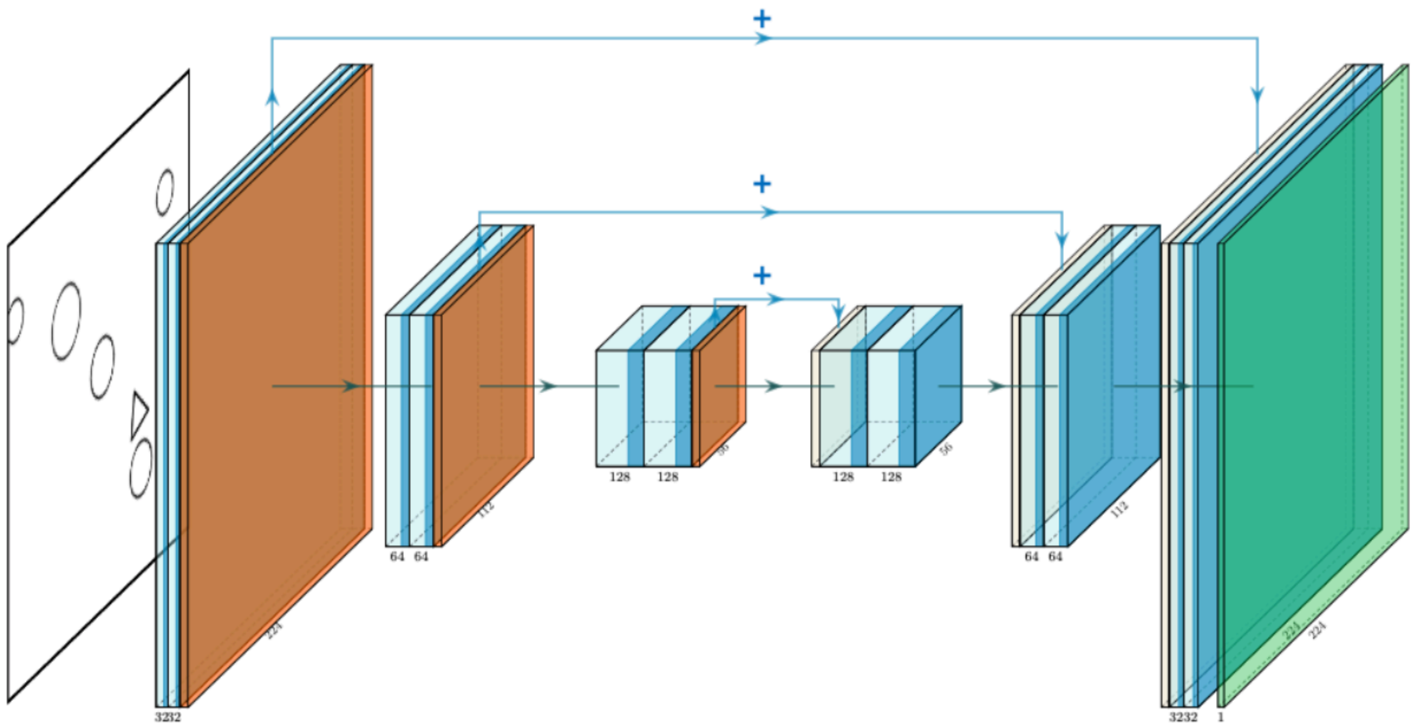
- A good explanation should be able to balance the **interpretability-completeness trade-off**, because the more accurate an explanation, the less interpretable it may be to humans

Proposed Architecture

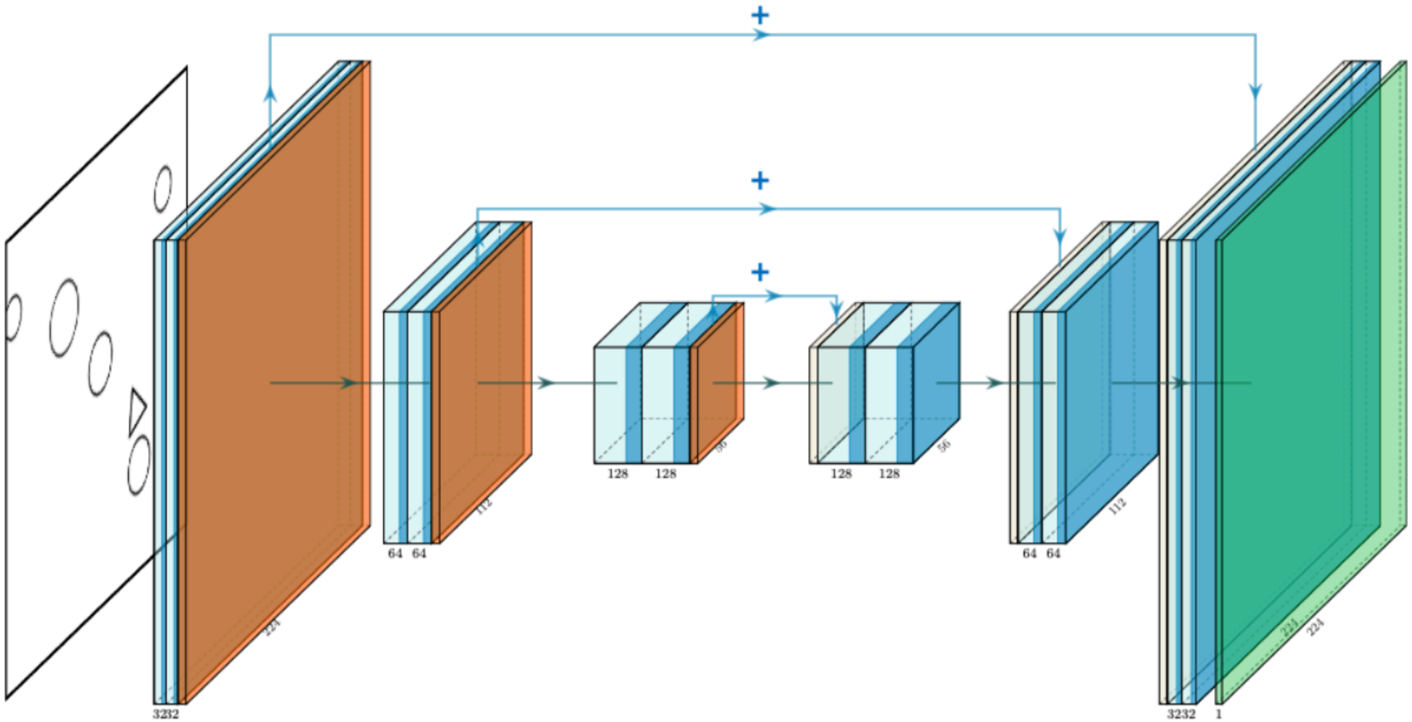



3232
64 64
128 128
256 256
1x1x128 1x1x128 1x1x2
relu
conv 3x3
conv1x1
softmax
global pool
pool
deconv
fc
mult

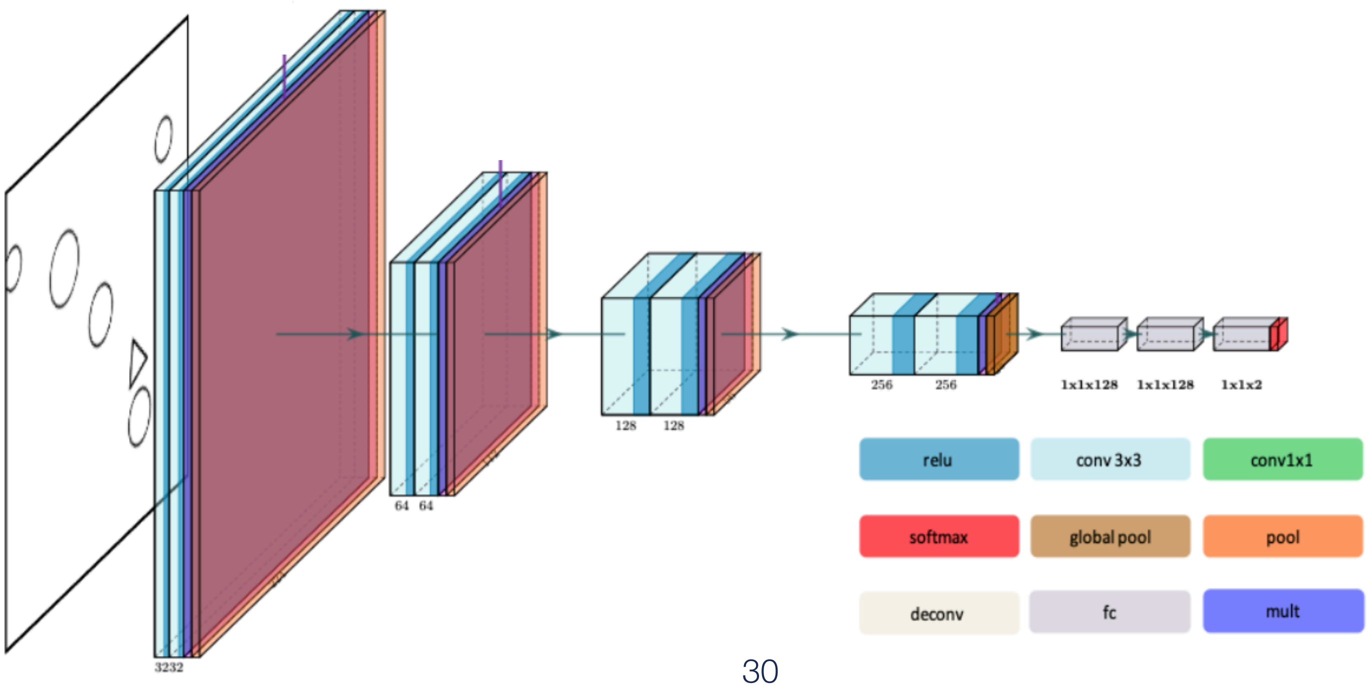
Classifier



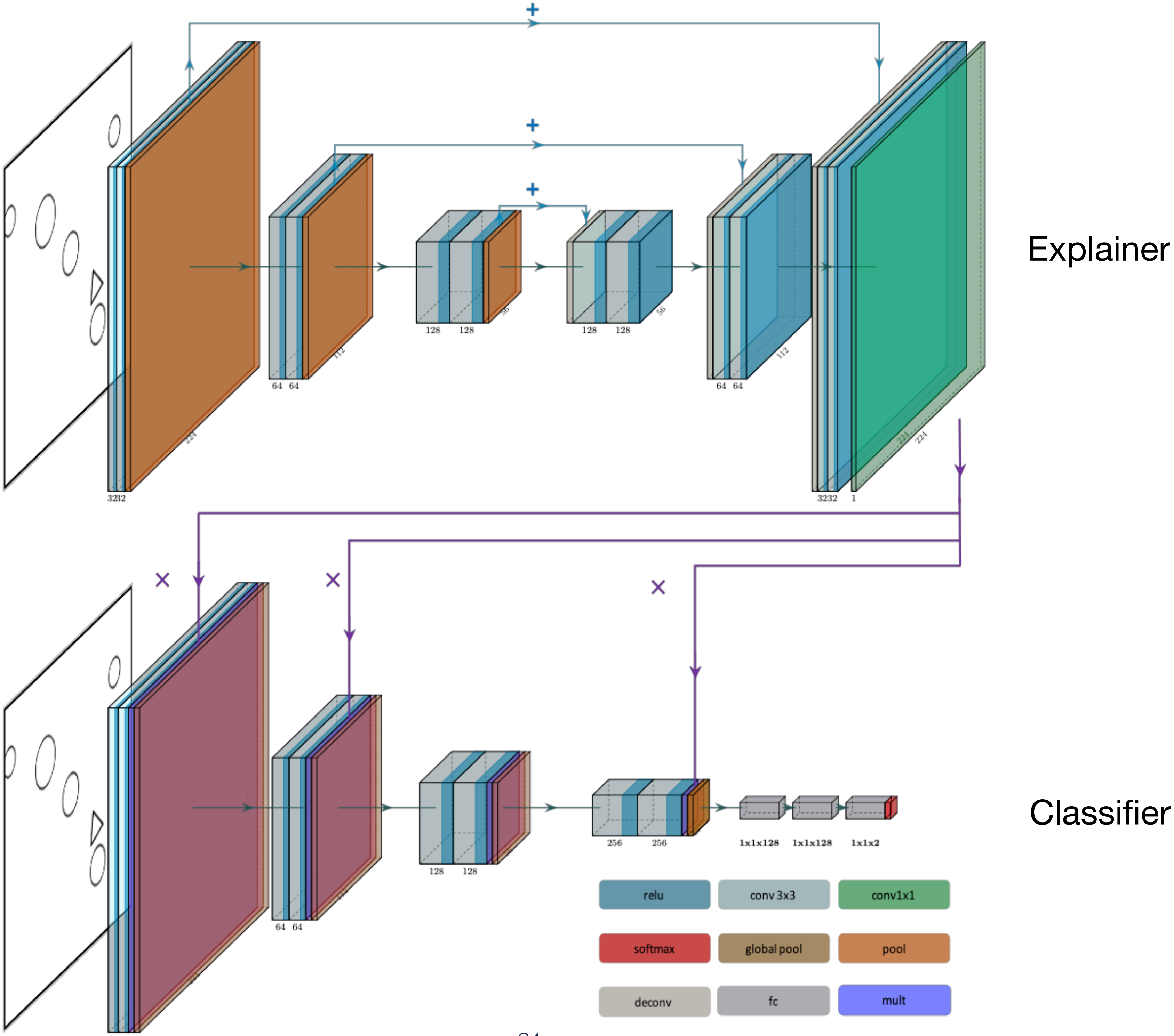
Explainer



Explainer

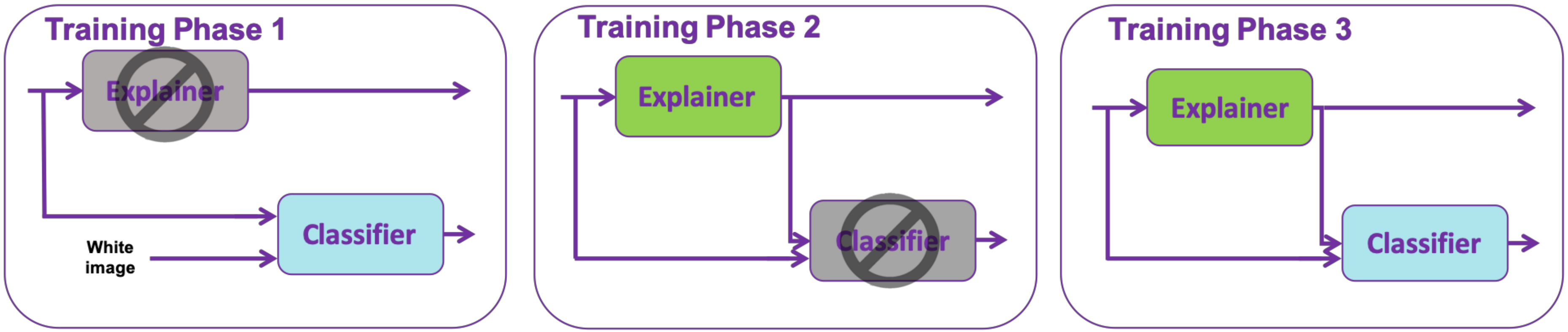


Classifier



Explainer
Classifier
+
+
+
×
×
×
3232
64 64
128 128
128 128
64 64
3232
1
224
112
56
3232
64 64
128 128
256 256
1x1x128
1x1x128
1x1x2
relu
conv 3x3
conv1x1
softmax
global pool
pool
deconv
fc
mult

Training Process



Loss

$$\mathscr{L} = \alpha \mathscr{L}_{class} + (1 - \alpha)\mathscr{L}_{expl}$$

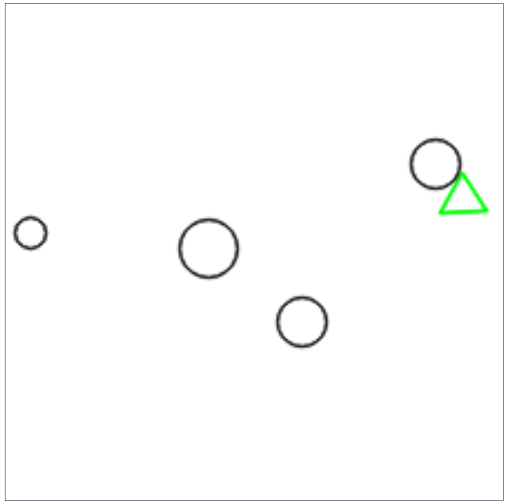
$$\mathscr{L}_{class} = -\sum_{c} y_{o,c} log(p_{o,c})$$

categorical cross entropy

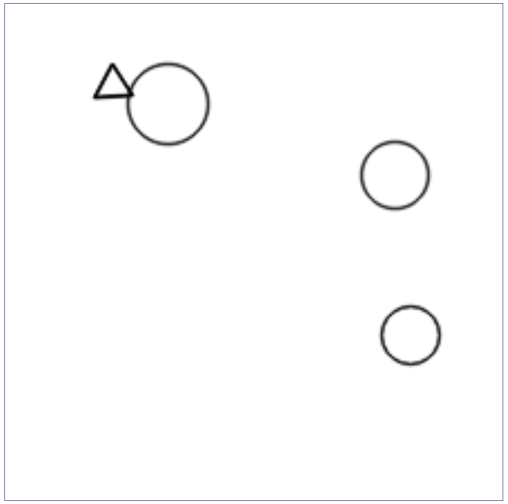
$$\mathscr{L}_{expl} = \lambda \frac{1}{m \times n} \sum_{i,j} |z_{i,j}|$$

penalised ℓ_1 norm

Synthetic Datasets

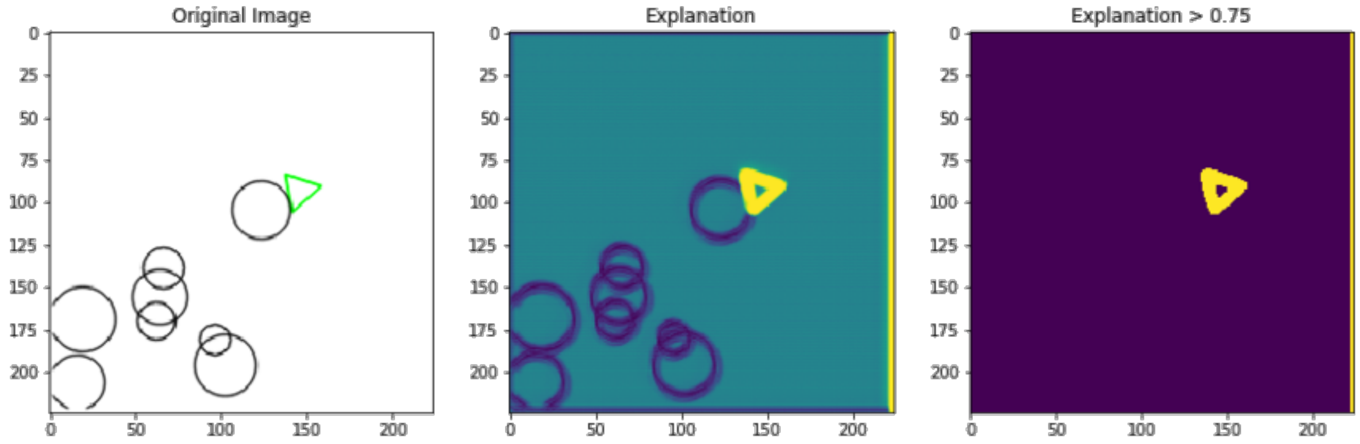


Simple dataset with colour cues
Binary classification problem: exists/does not exist a triangle



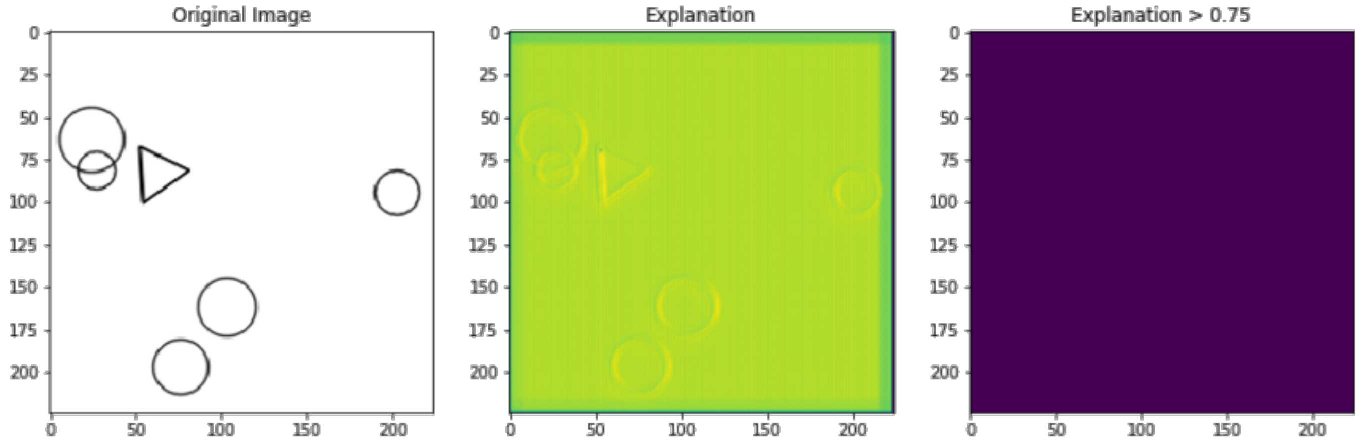
Simple dataset without colour cues
Binary classification problem: exists/does not exist a triangle

Results on Synthetic Datasets



Explanation obtained without any regularisation

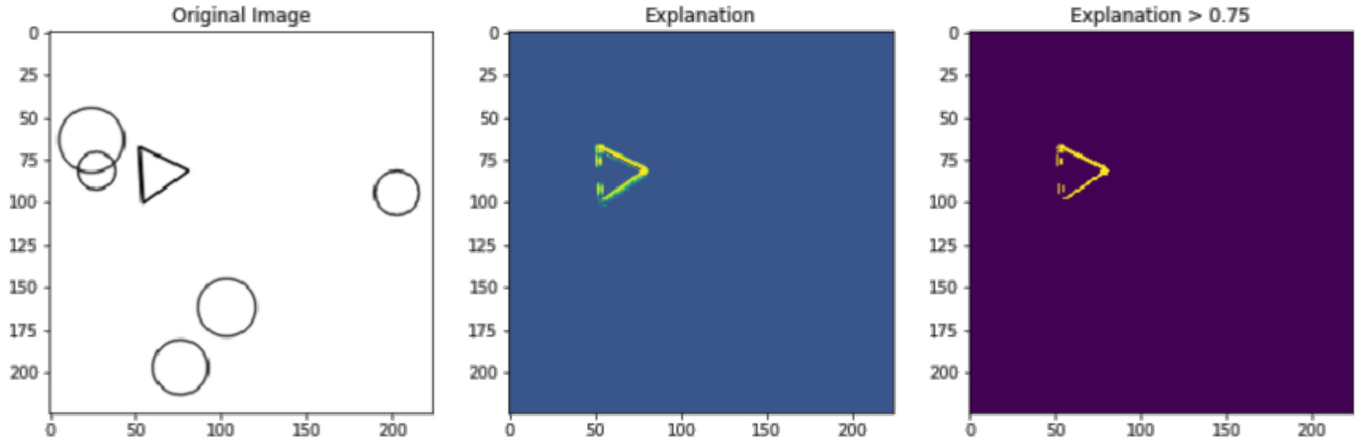
Results on Synthetic Datasets



Explanation obtained without any regularisation

0
255

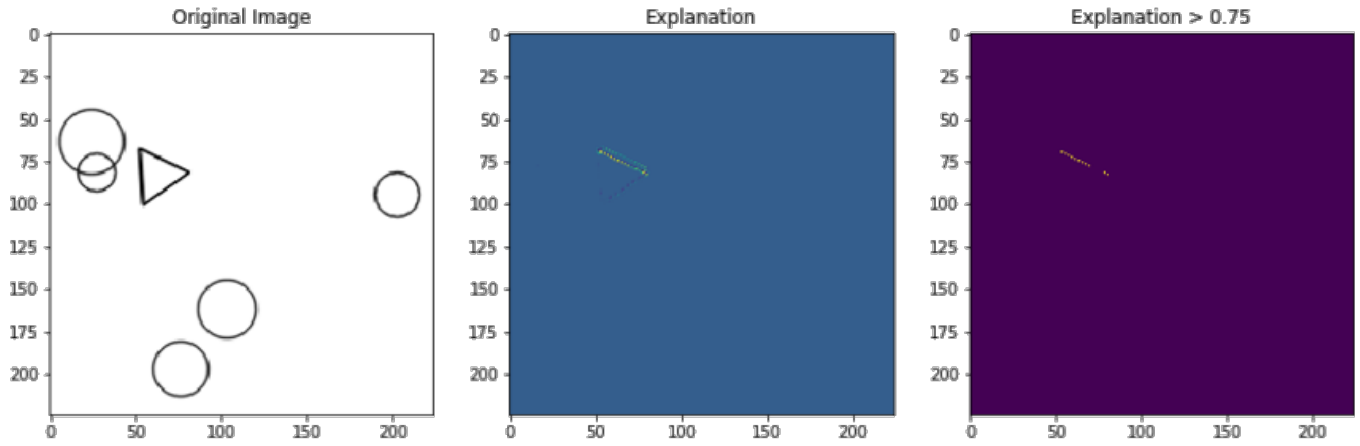
Results on Synthetic Datasets



Explanation obtained with ℓ_1 penalty $\lambda = 10^{-6}$

0
255

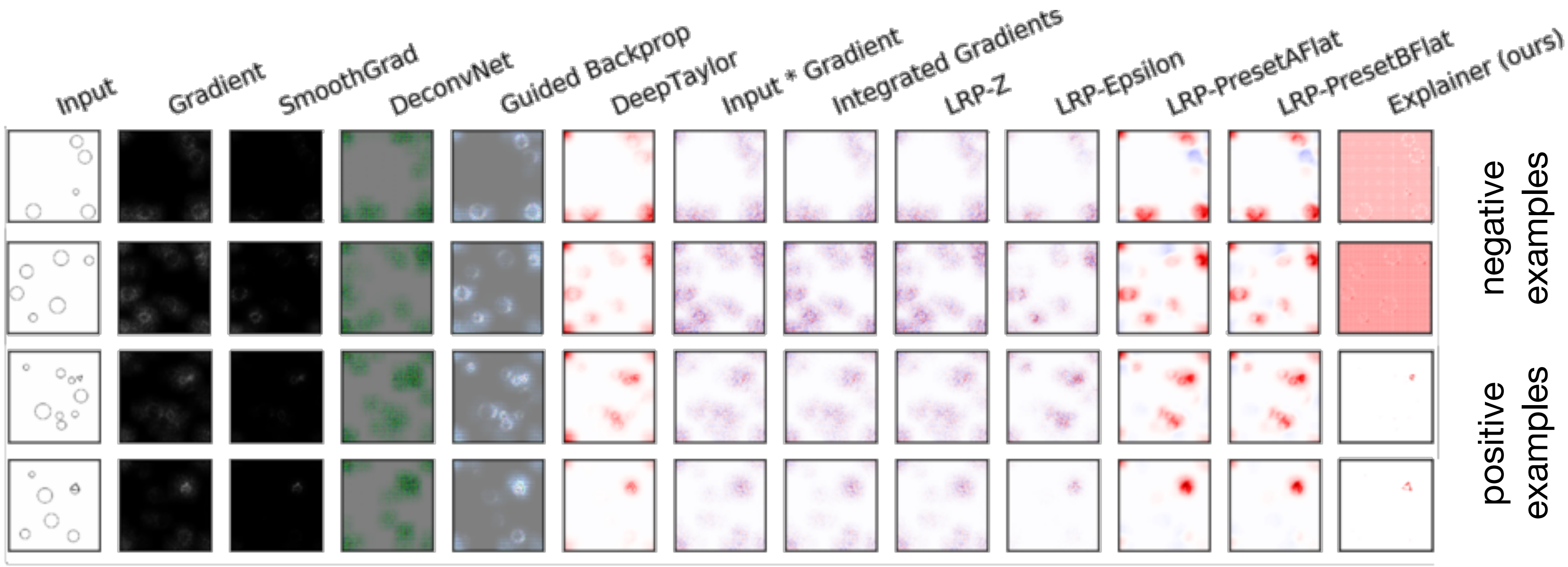
Results on Synthetic Datasets



Explanation obtained with ℓ_1 penalty $\lambda = 10^{-4}$

0
255

Comparison with State-of-the-Art Methods



Comparison between our explanation method and methods implemented in the iNNvestigate toolbox (Alber *et al.*: iNNvestigate neural networks!)

Real Datasets

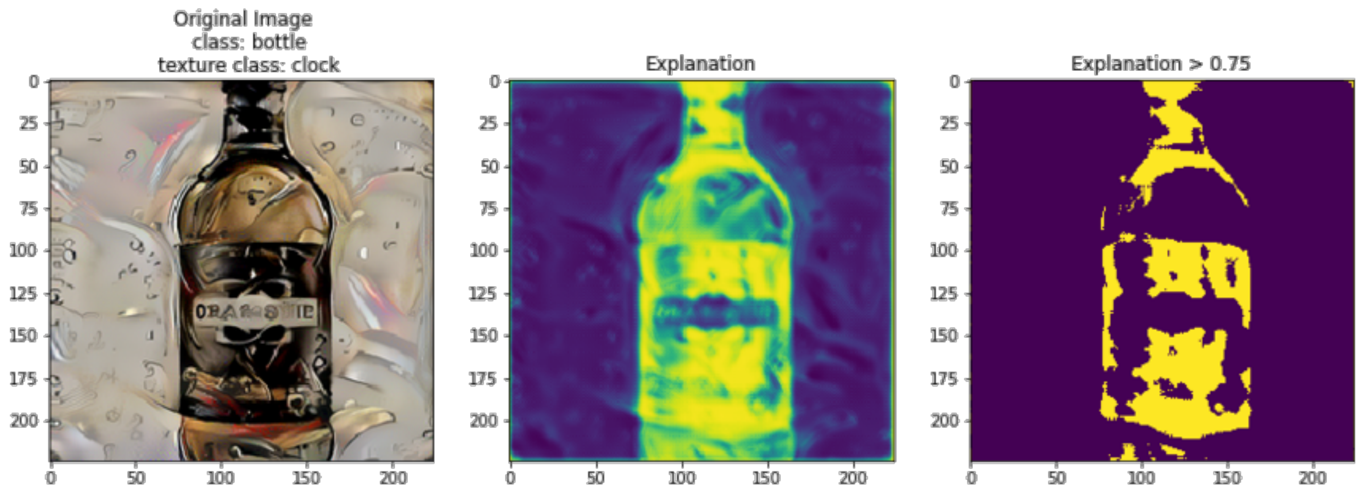
Cue conflict dataset introduced in Geirhos *et al.*: ”ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”

Original class: truck
Texture class: elephant

Original class: bicycle
Texture class: truck

Original class: bottle
Texture class: clock

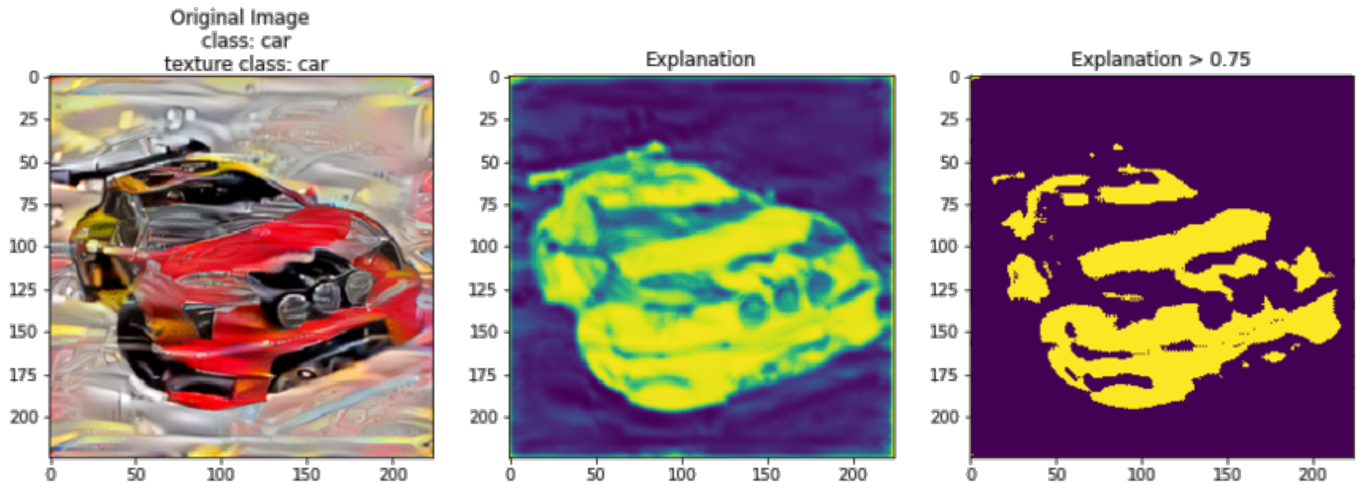
Results on Cue Conflict Dataset



Example explanation obtained from the cue conflict dataset without regularisation

0
255

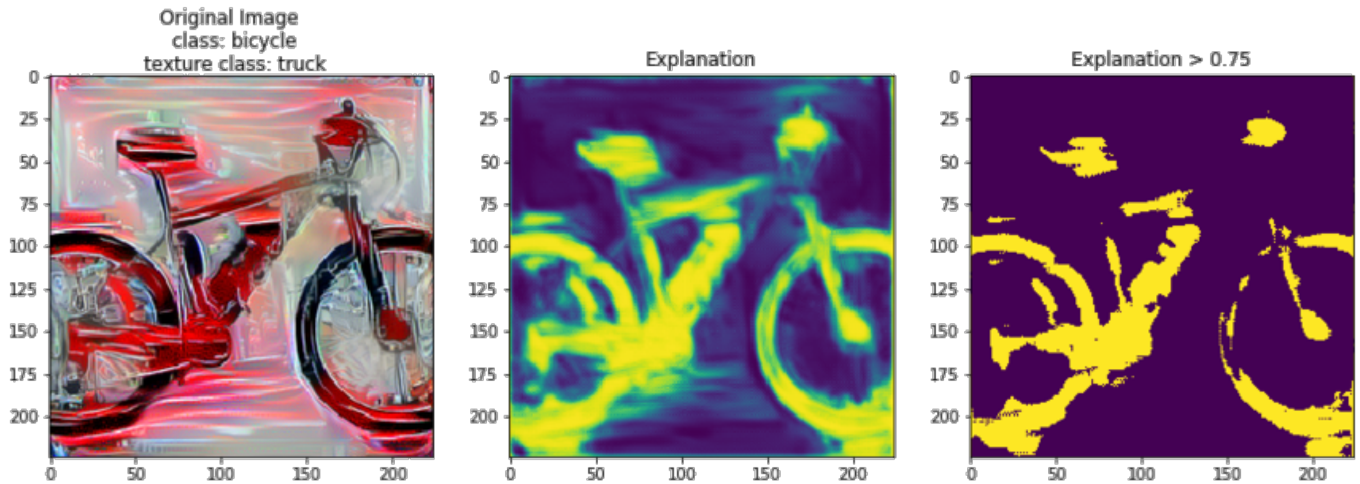
Results on Cue Conflict Dataset



Example explanation obtained from the cue conflict dataset without regularisation

0
255

Results on Cue Conflict Dataset



Example explanation obtained from the cue conflict dataset without regularisation

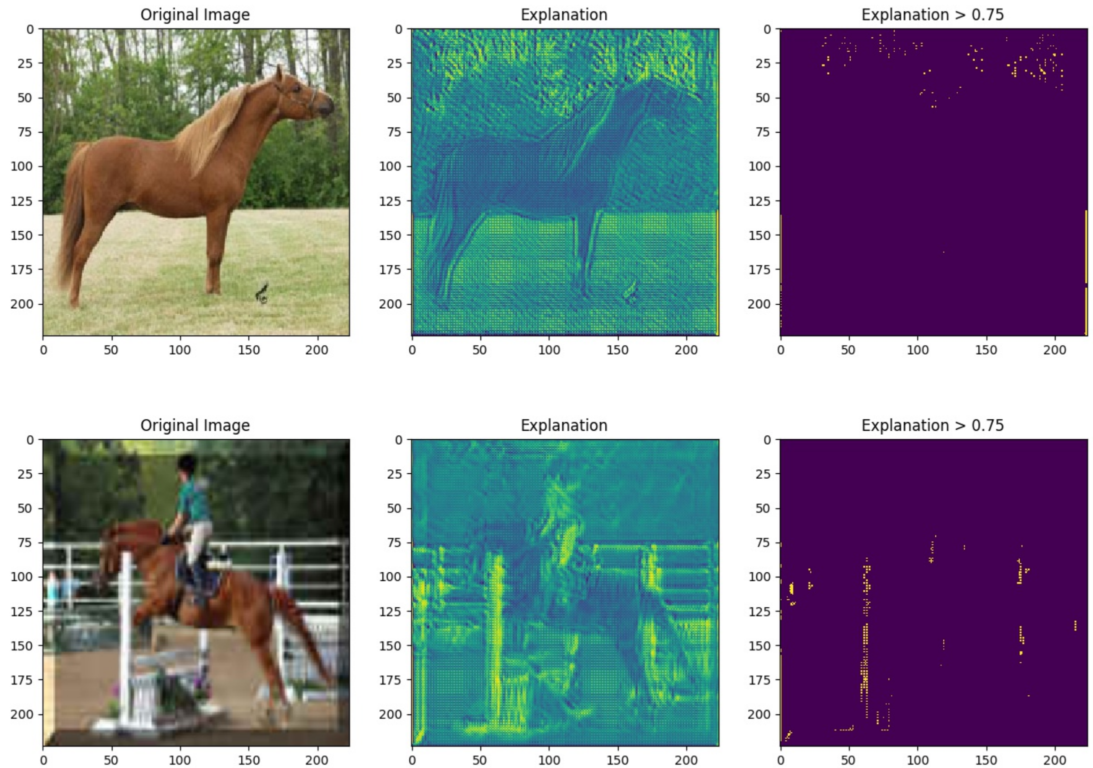
0
255

Other Preliminary Experiments

- Using a different (and larger) classifier
  - ResNet-50 pre-trained w/ ImageNet
- Horses vs. Zebras

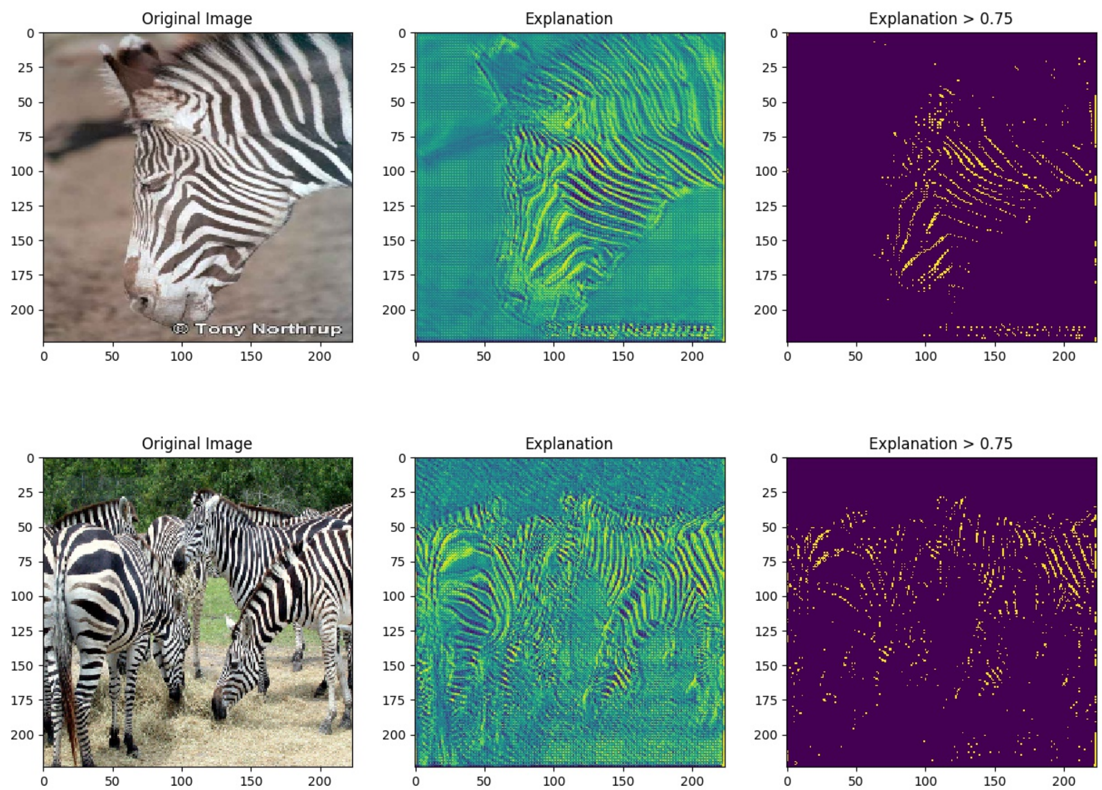
Other Preliminary Experiments

- During the training process (before plateauing):



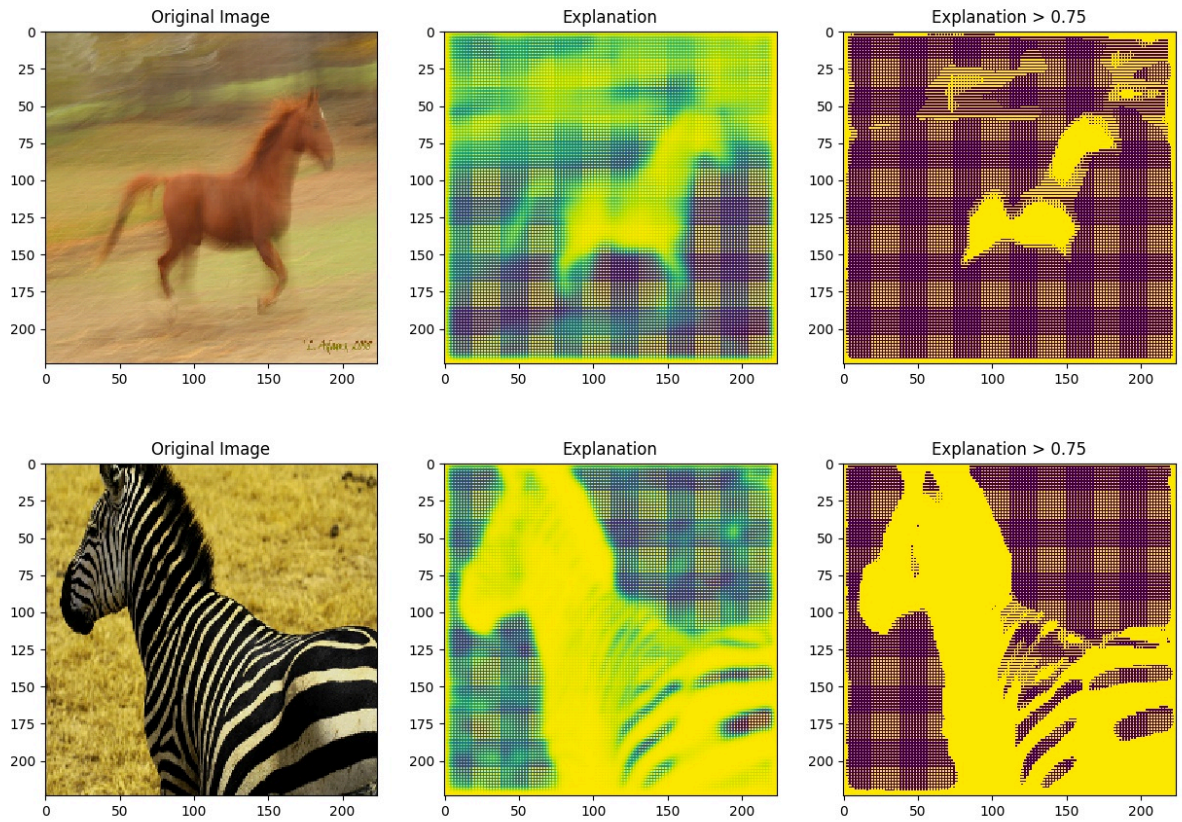
Other Preliminary Experiments

- During the training process (before plateauing):



Other Preliminary Experiments

- After the training process (no regularisation -> some artefacts):



Conclusion and Future Work

- Joint approach to produce decisions and explanations using CNNs
- Shows potential especially when compared to existing methods
- Future work includes:
    - Experimenting with other explainer losses, e.g. using Total Variation
    - Weak (and Semi) supervision of the explainer
    - Other modalities

Producing Joint Decisions and Explanations with CNNs

Luis F. Teixeira
Universidade do Porto / FEUP, INESC TEC

University of Twente - July 1, 2019