Producing Decisions and Explanations: A Joint Approach Towards Explainable CNNs

MASTER'S DEGREE IN ELECTRICAL AND COMPUTERS ENGINEERING

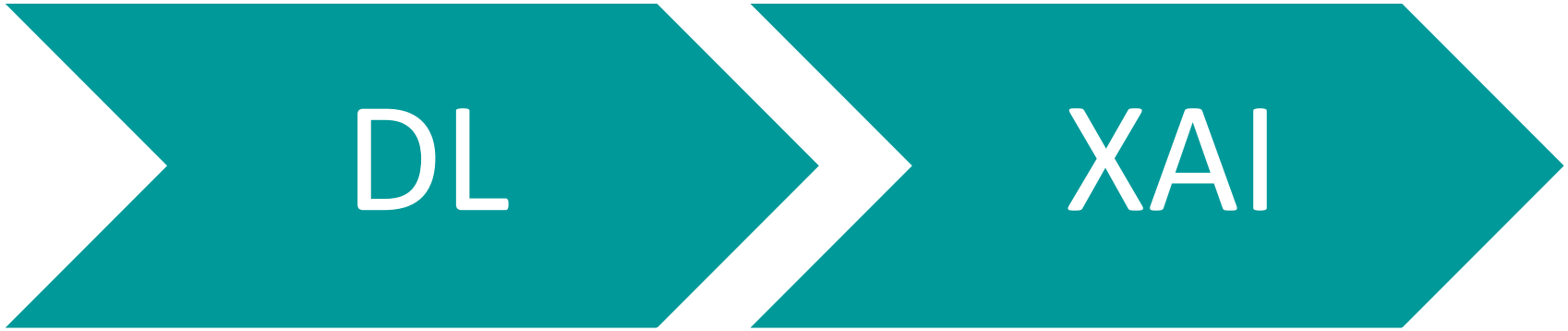
ISABEL RIO-TORTO DE OLIVEIRA

27 SEPTEMBER 2019

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

SUPERVISORS:
LUÍS TEIXEIRA
KELWIN FERNANDES

Context and Motivation



- Outstanding performance
- Great amount of labelled data
- Automatic feature extraction
- High complexity
- **Black-boxes**

- Transparency, auditability, trust, …
- High-stakes regulated areas (medicine, finance, …)
- Understand models' decisions
- Generate explanations

Related Work

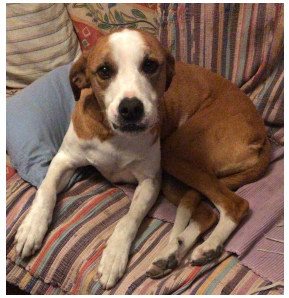
- Growing interest in this field
- Still lacks unified formal definition, taxonomy and metrics
- Existing literature focuses on post-model methods – saliency mapping
  - Gradient- or perturbation-based methods
  - Known pitfalls: independent on the data and the model parameters
  - Sensitivity Analysis, Deep-Lift, Layer-Wise Relevance Propagation, …

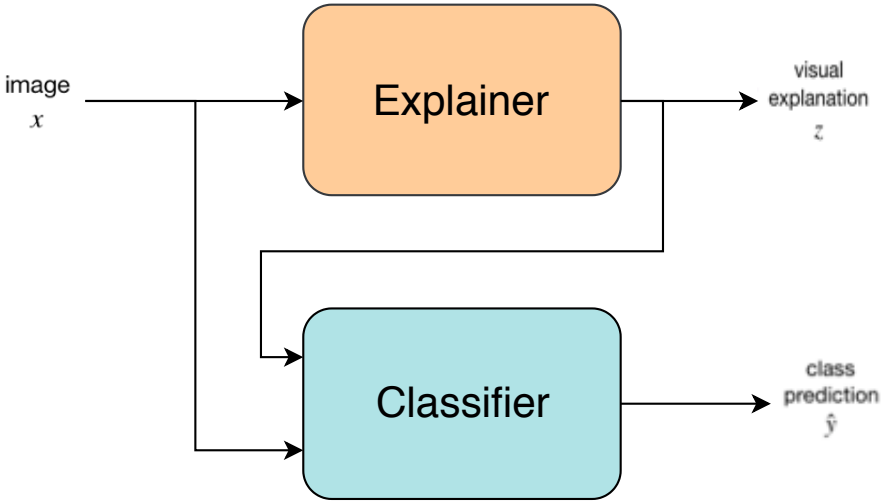
Research Questions

Would it be possible to…

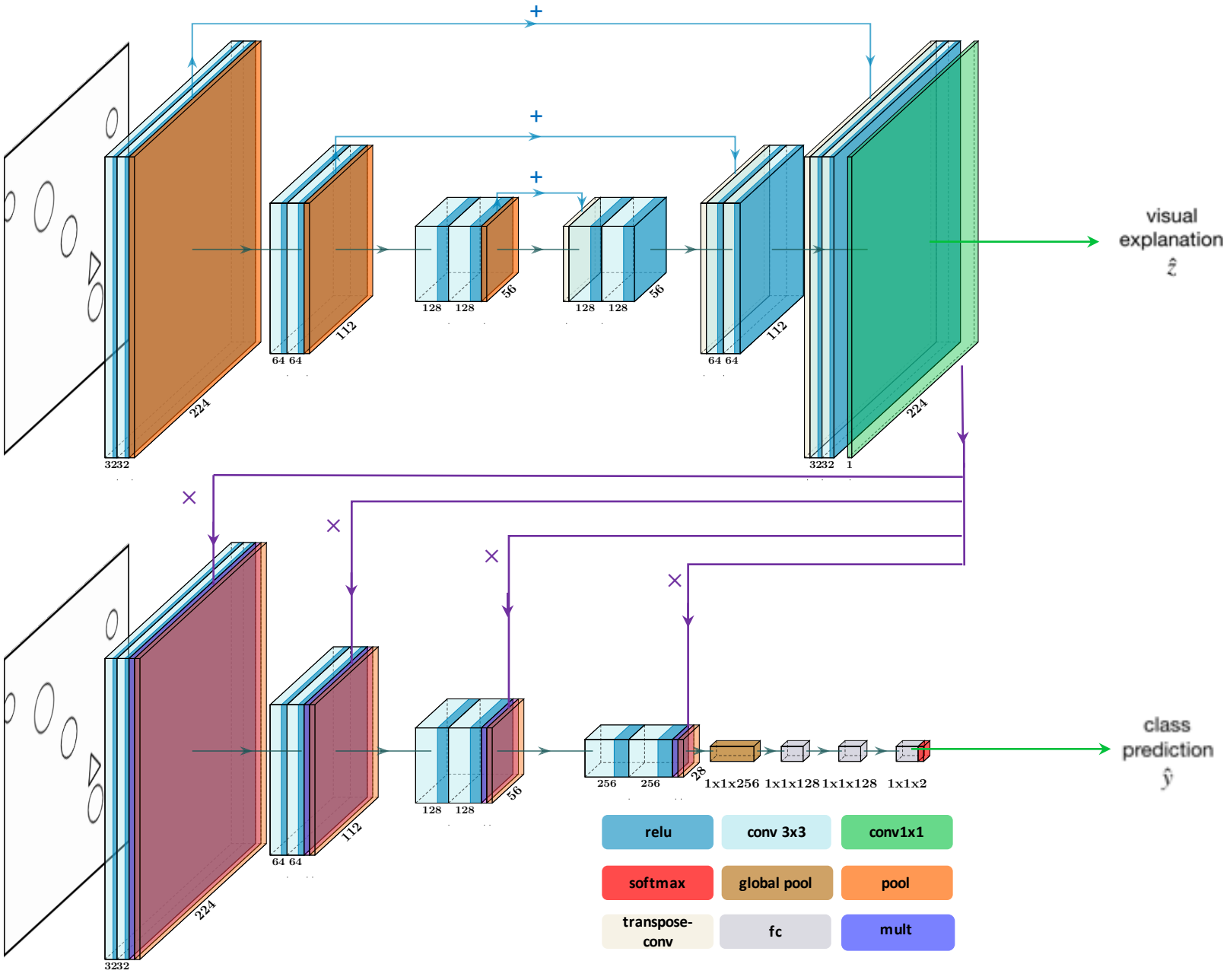
- generate visual explanations without supervision?
- design an in-model explainability method?
- apply this method to different CNN architectures?
- validate this architecture on a real medical application?

Proposed Architecture

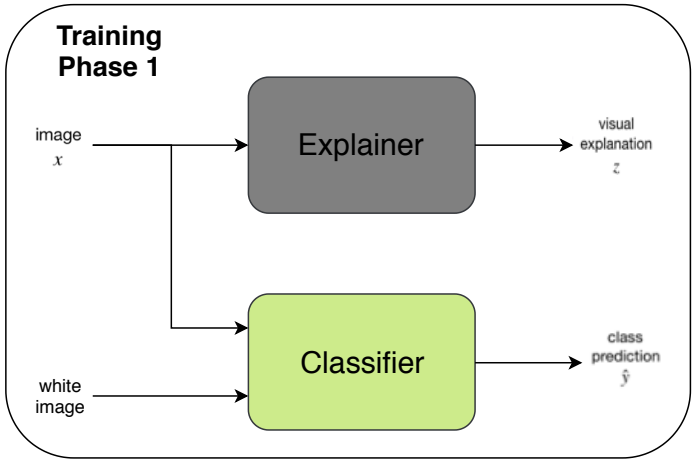


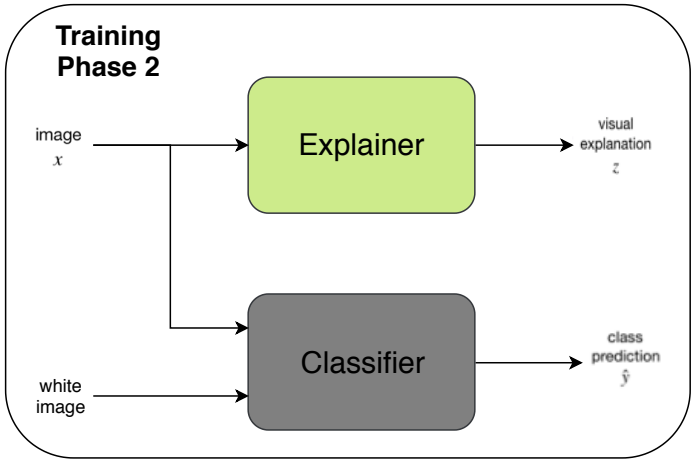


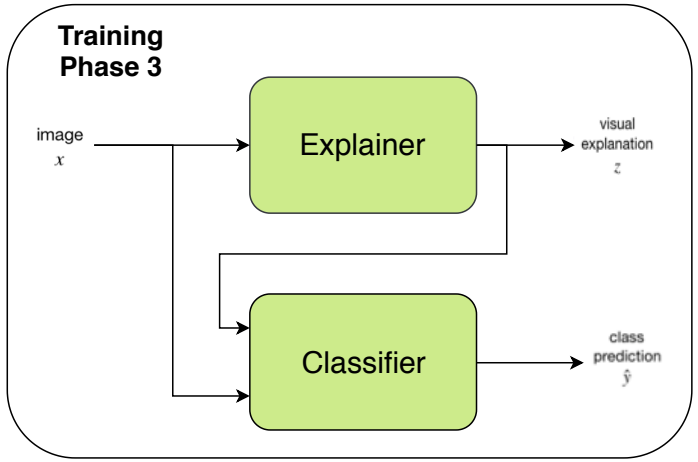
DOG

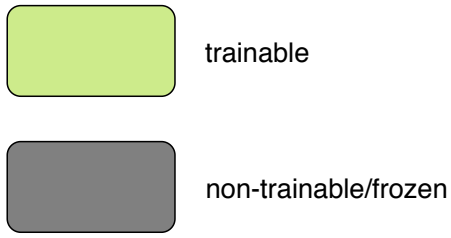

visual explanation
$\hat{z}$
class prediction
$\hat{y}$
3232
224
64 64
112
128 128
56
256 256
28
1x1x256
1x1x128
1x1x128
1x1x2
relu
conv 3x3
conv1x1
softmax
global pool
pool
transpose-conv
fc
mult

Training Process









Loss Function

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

$$\mathcal{L}_{class} = -\sum_{i=1}^{N} y_i^{\top} \times log(\hat{y}_i)$$

Categorical cross entropy

y_i class labels for instance i
$\hat{y}_i$ predictions for instance i
N number of instances

Unsupervised Explainer Loss

$$\mathcal{L}_{expl\_unsup} = \beta \times \sum_{i=1}^{N} \mathcal{L}_{sparsity}(\hat{z}_i) + (1-\beta) \times \sum_{i=1}^{N} \mathcal{L}_{contiguity}(\hat{z}_i)$$

$$\mathcal{L}_{sparsity}(\hat{z}) = \frac{1}{m \times n} \sum_{i,j} |\hat{z}_{i,j}|$$

$l1$penalised norm

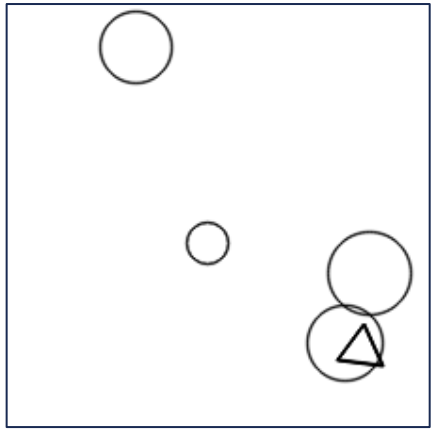
$$\mathcal{L}_{contiguity}(\hat{z}) = \frac{1}{m \times n} \sum_{i,j} |\hat{z}_{i+1,j} - \hat{z}_{i,j}| + |\hat{z}_{i,j+1} - \hat{z}_{i,j}|$$

Total variation

$\hat{z}$ generated feature map
m, n dimensions of $\hat{z}$
N number of instances

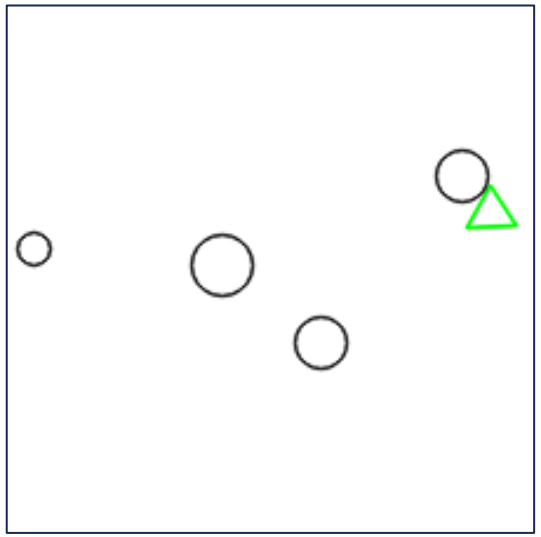
Weakly Supervised Explainer Loss

$$\mathcal{L}_{expl\_weakly} = \left| \frac{\sum_{i,j}(1 - z_{i,j}) \times \hat{z}_{i,j}}{\sum_{i,j}(1 - z_{i,j})} \right|$$

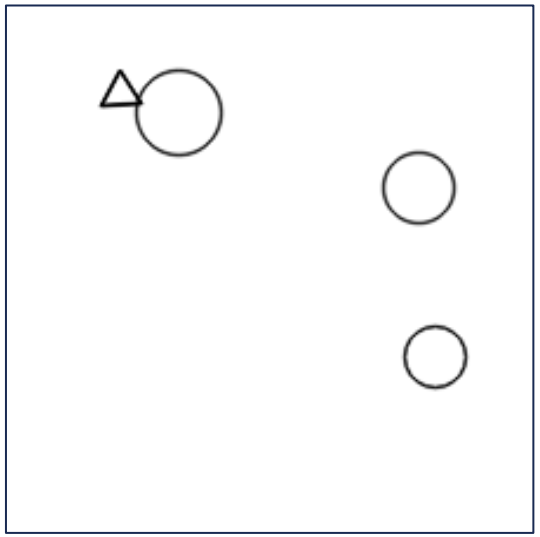


$\hat{z}$ generated feature map
z target feature map (mask)

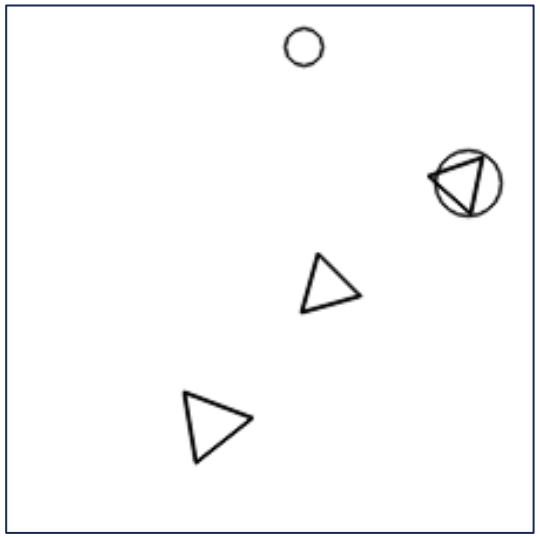
Synthetic Datasets



Simple dataset with colour cues



Simple dataset without colour cues

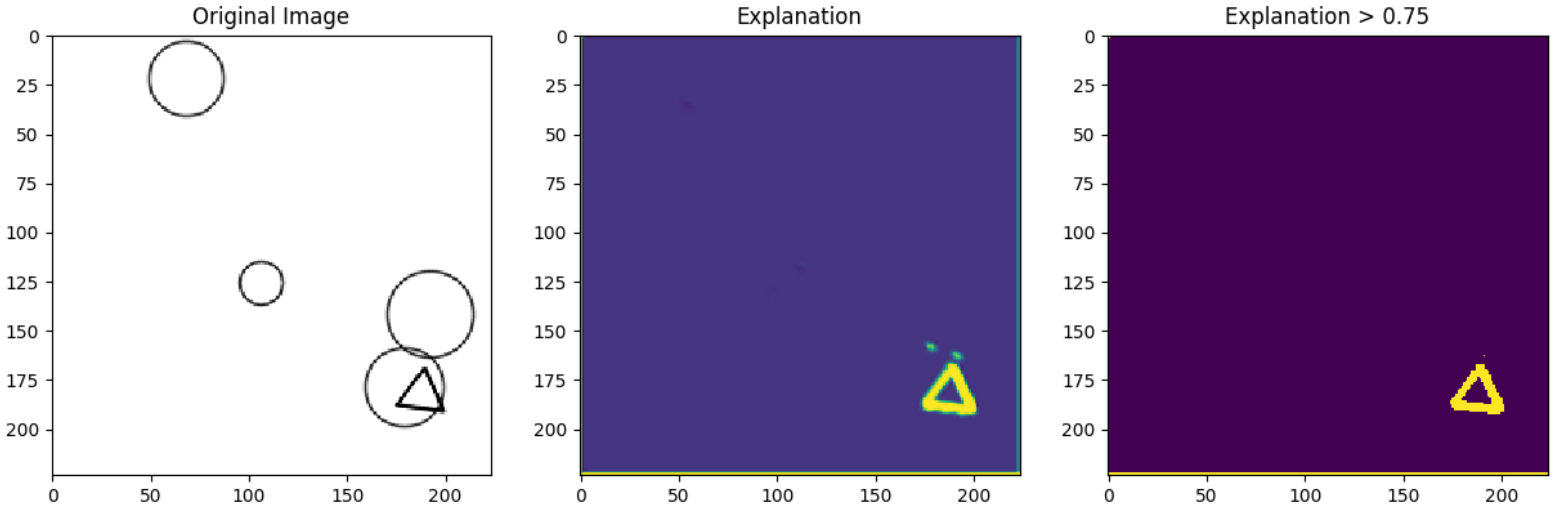


Dataset with multiple targets and without colour cues

Binary classification problem: exists/does not exist (at least) one triangle

Results on Synthetic Datasets (1/4)

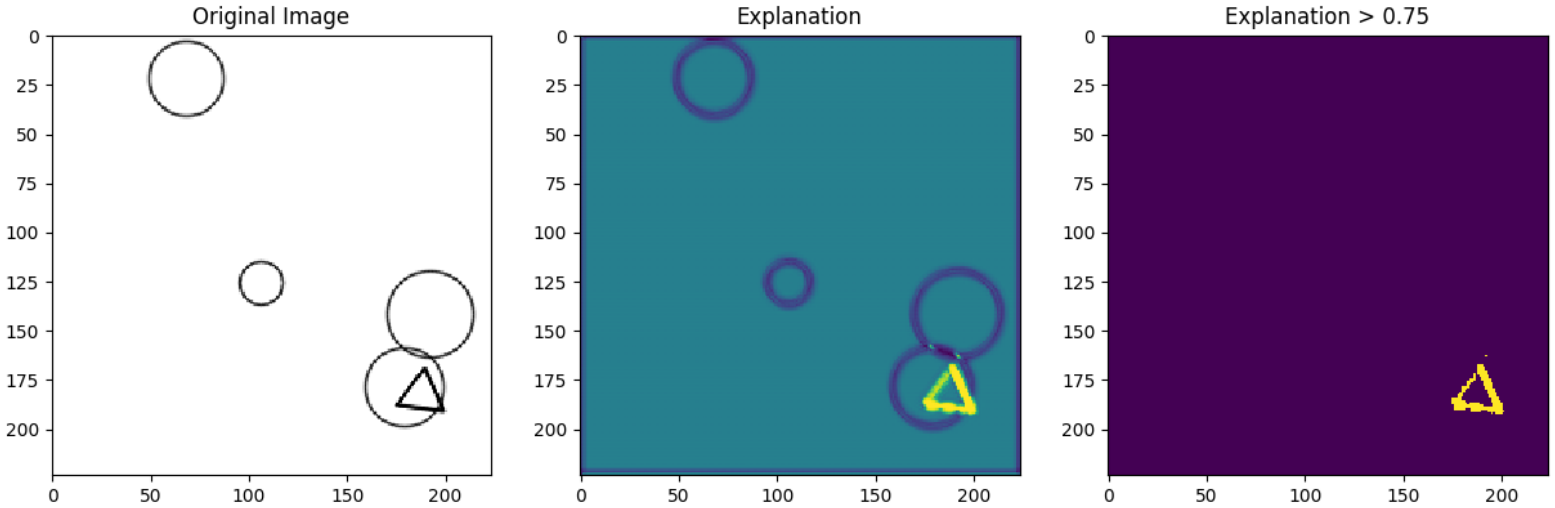
Unsupervised Approach



Explanation obtained with $\alpha = 0.99$ and $\beta = 0.1$

Results on Synthetic Datasets (2/4)

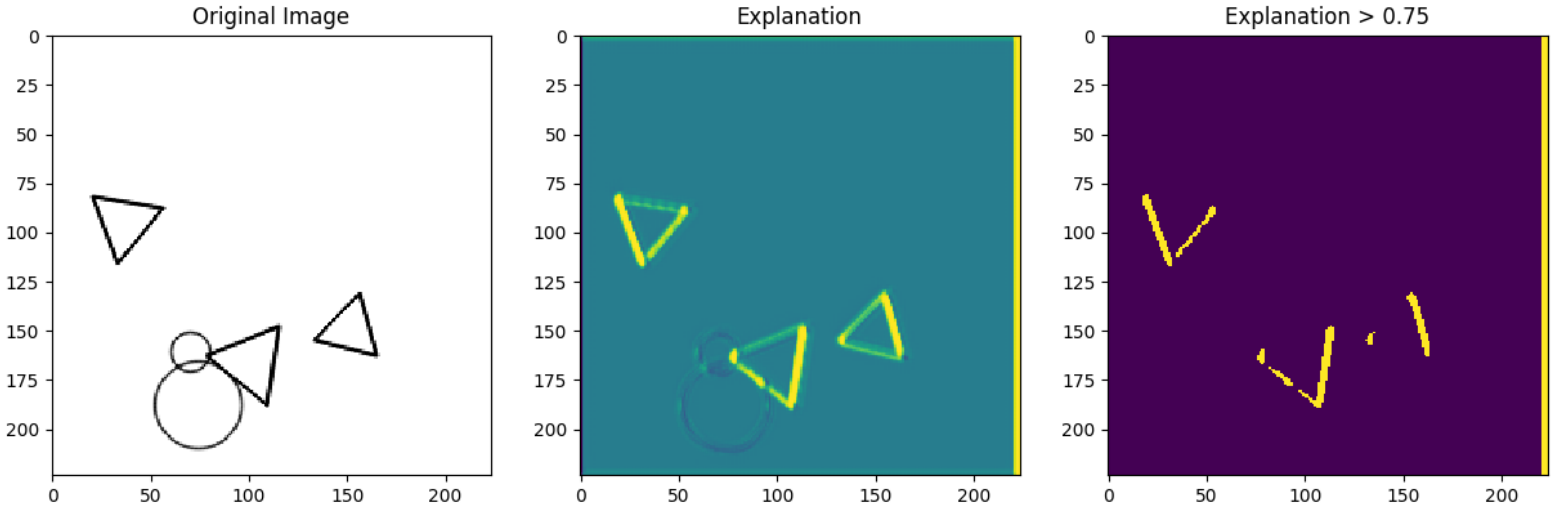
Weakly Supervised Approach



Explanation obtained with $\alpha = 0.99$

Results on Synthetic Datasets (3/4)

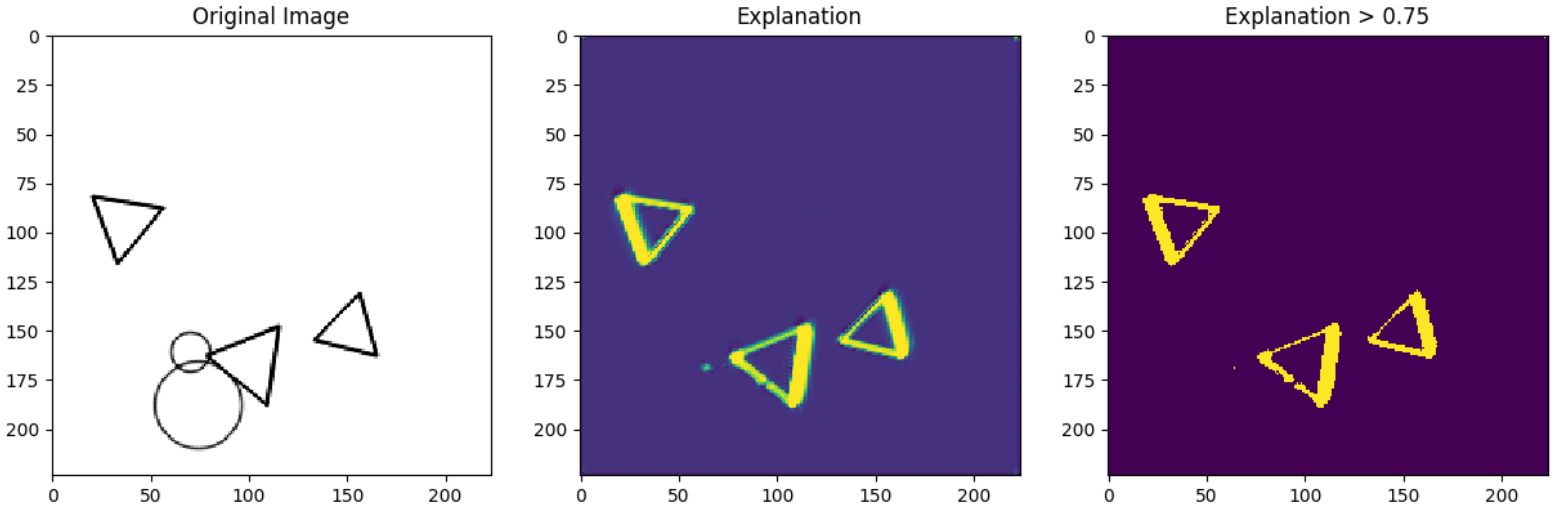
Unsupervised Approach



Explanation obtained with $\alpha = 0.99$ and $\beta = 0.25$

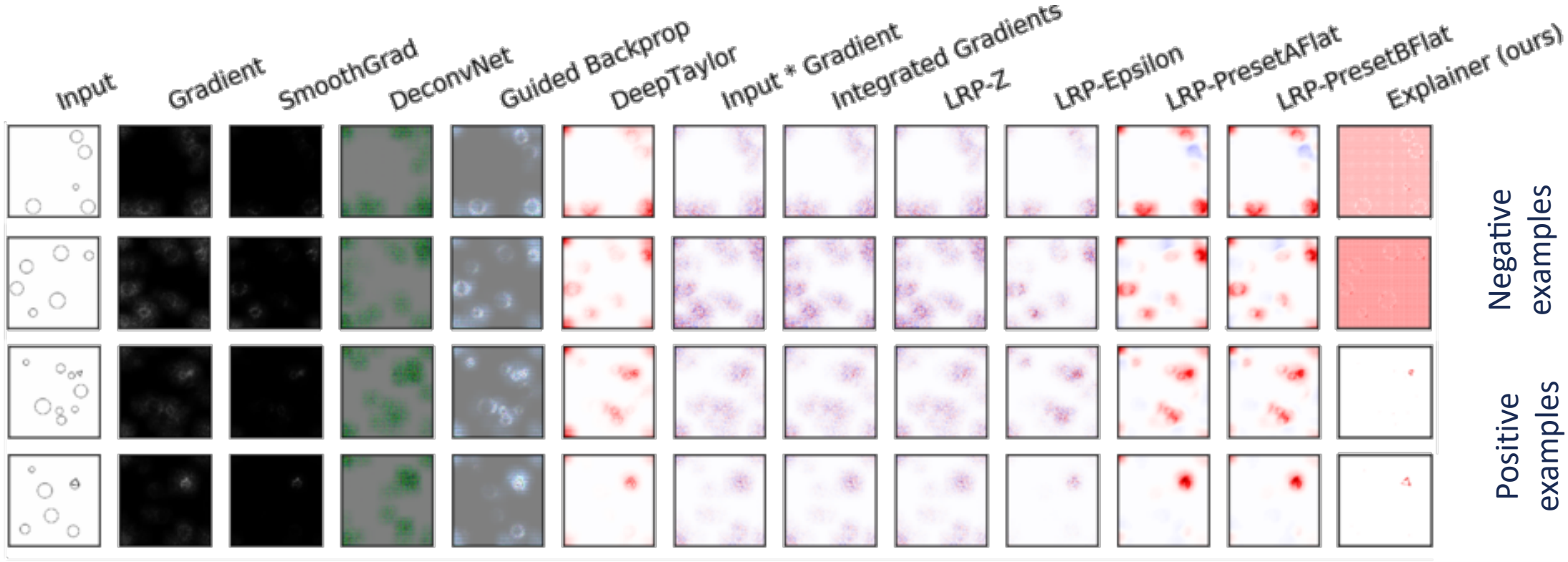
Results on Synthetic Datasets (4/4)

Weakly Supervised Approach



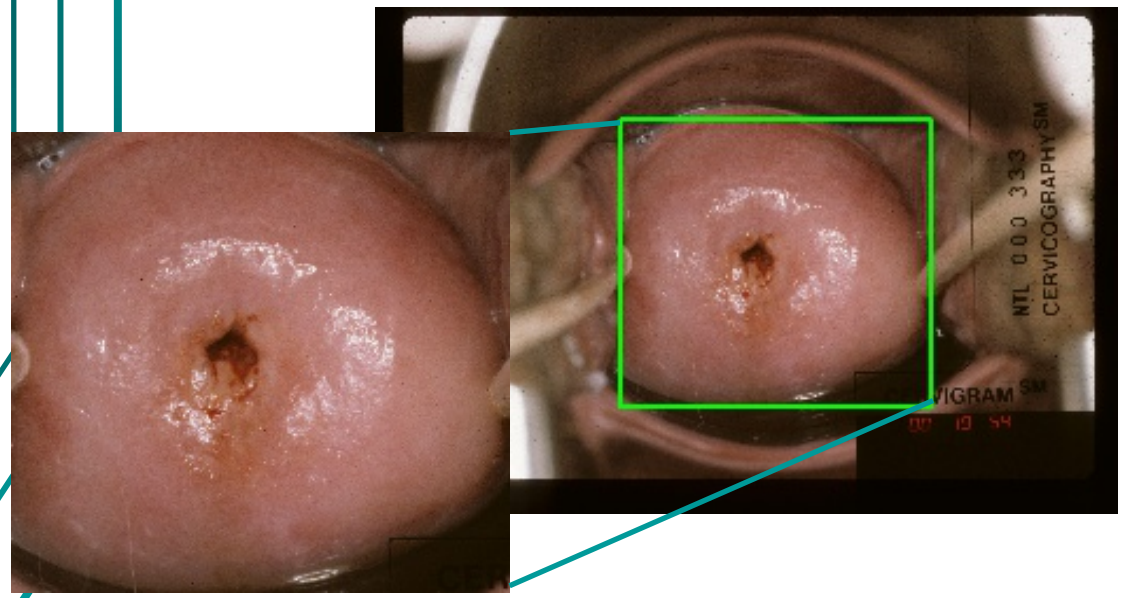
Explanation obtained with $\alpha = 0.99$

Comparison with State-of-the-Art Methods



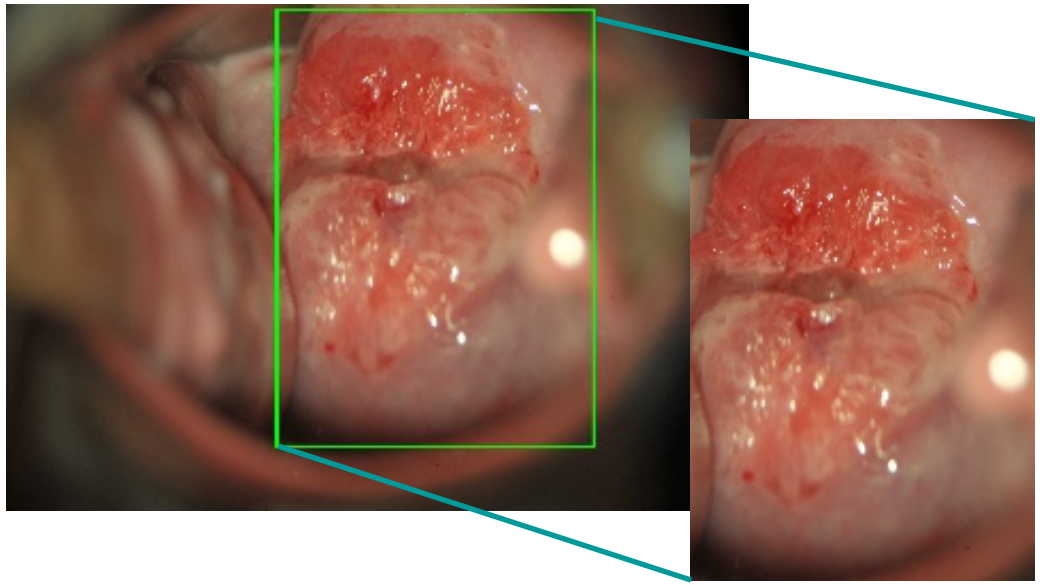
Comparison between our explanation method and methods implemented in the iNNvestigate toolbox (Alber *et al.:* iNNvestigate neural networks!)

Cervical Cancer Dataset



Healthy cervix (negative instance):

- Pinkish colour
- Planar surface

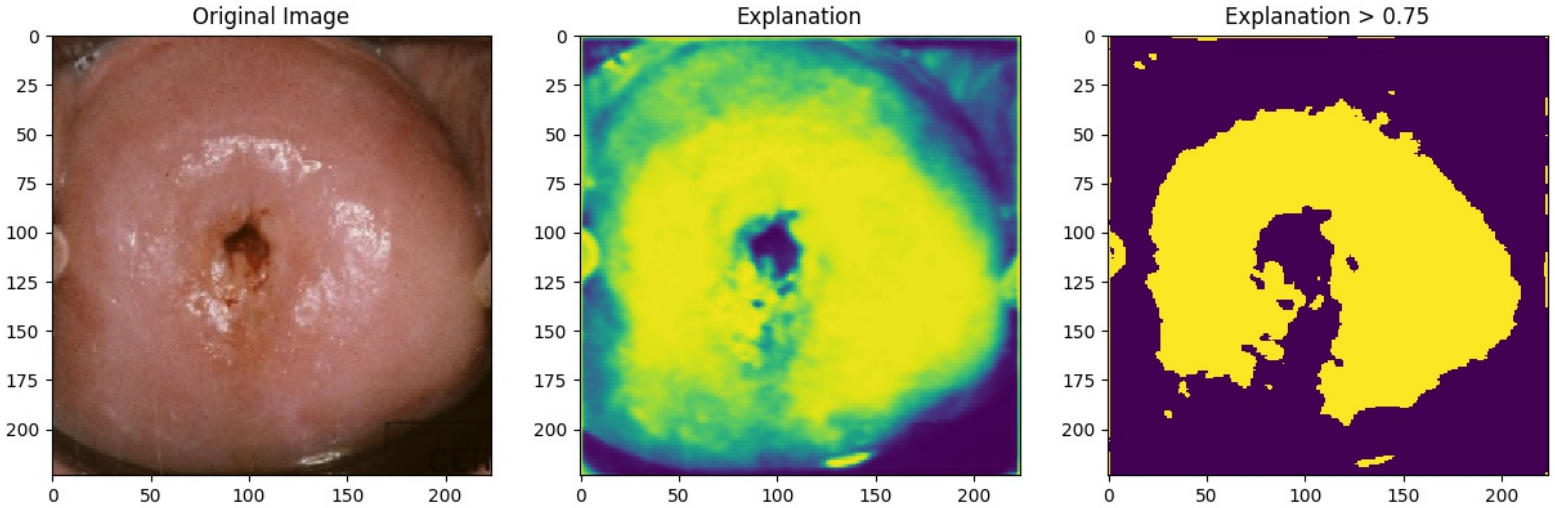


Cancerous cervix (positive instance):

- Whitish lesions
- Irregular contours
- Morphology changes

Results on the Cervical Cancer Dataset (1/5)

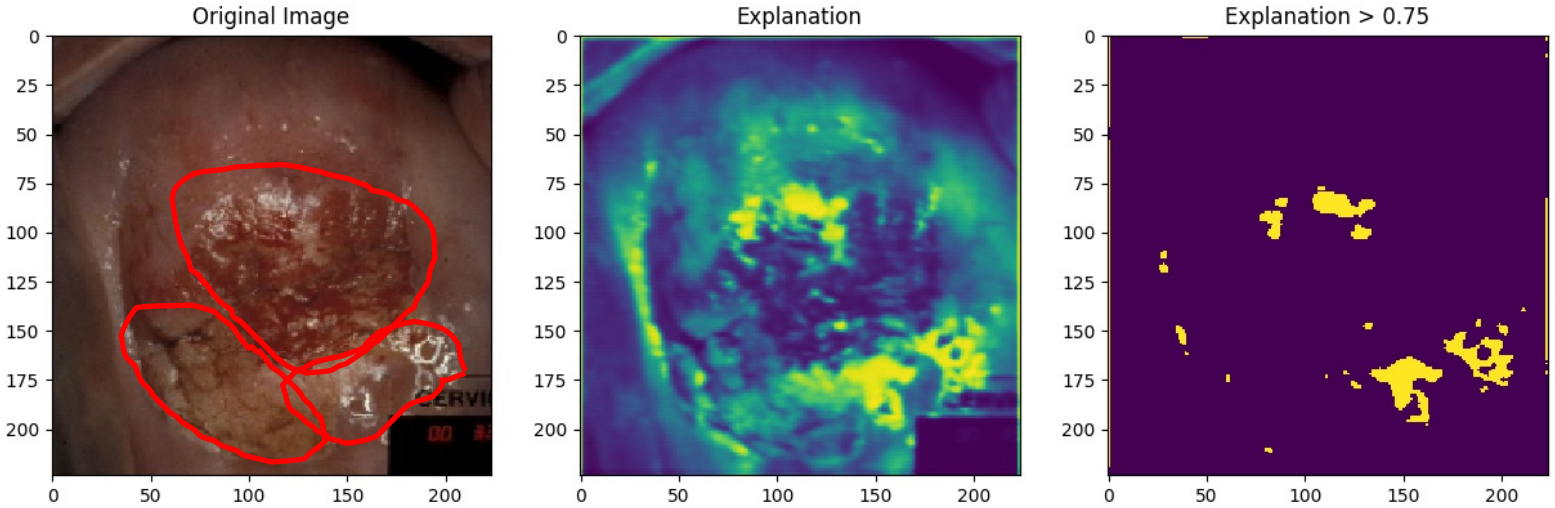
Unsupervised Approach



Explanation obtained with $\alpha = 0.99$ and $\beta = 0.9$
Normal case correctly classified (91.52%)

Results on the Cervical Cancer Dataset (2/5)

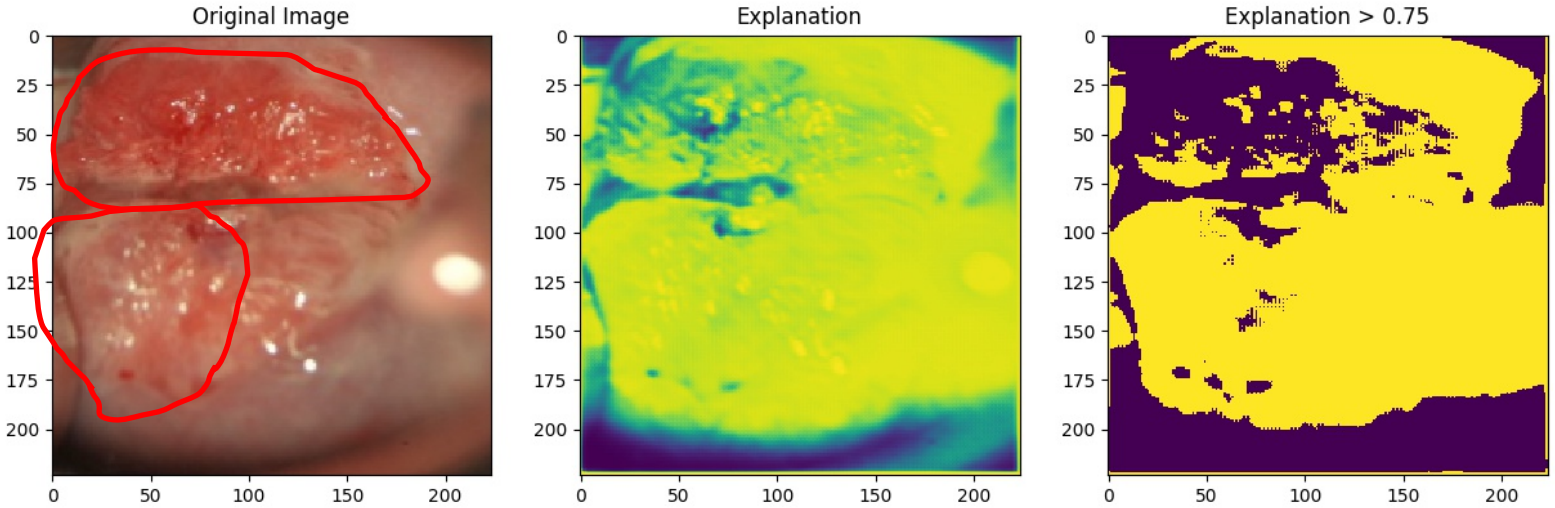
Unsupervised Approach



Explanation obtained with $\alpha = 0.99$ and $\beta = 0.9$
Cancer case misclassified (76.62%)

Results on the Cervical Cancer Dataset (3/5)

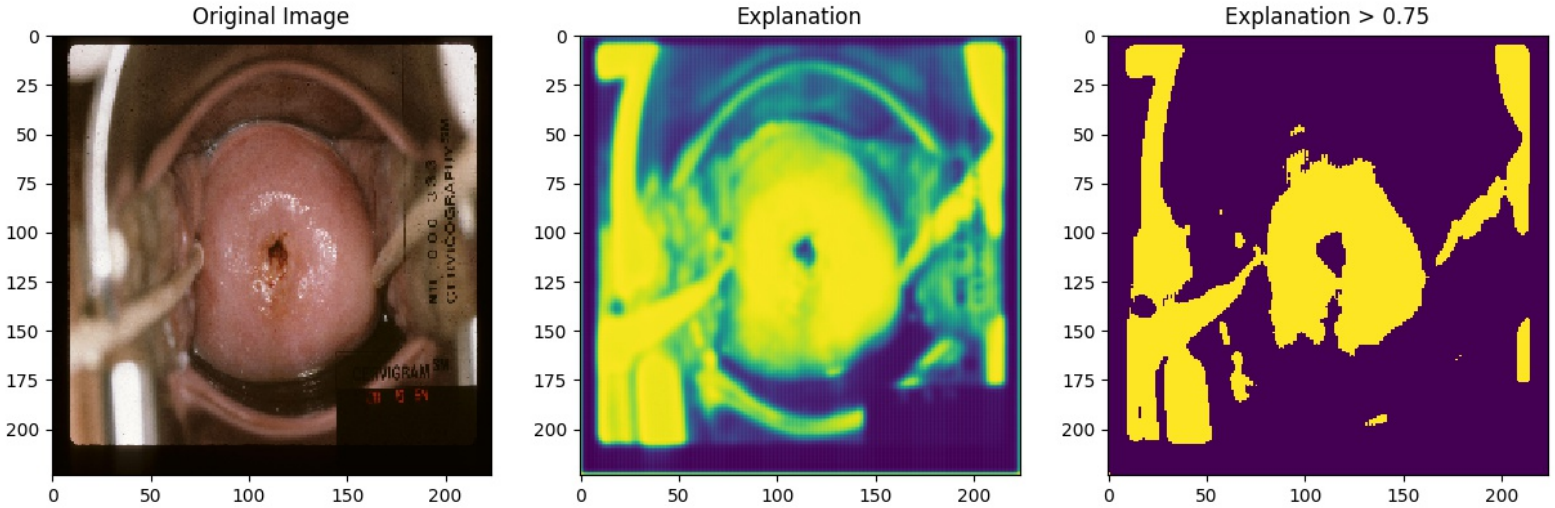
Unsupervised Approach



Explanation obtained with $\alpha = 0.99$ and $\beta = 0.9$
Cancer case correctly classified (55.47%)

Results on the Cervical Cancer Dataset (3/5)

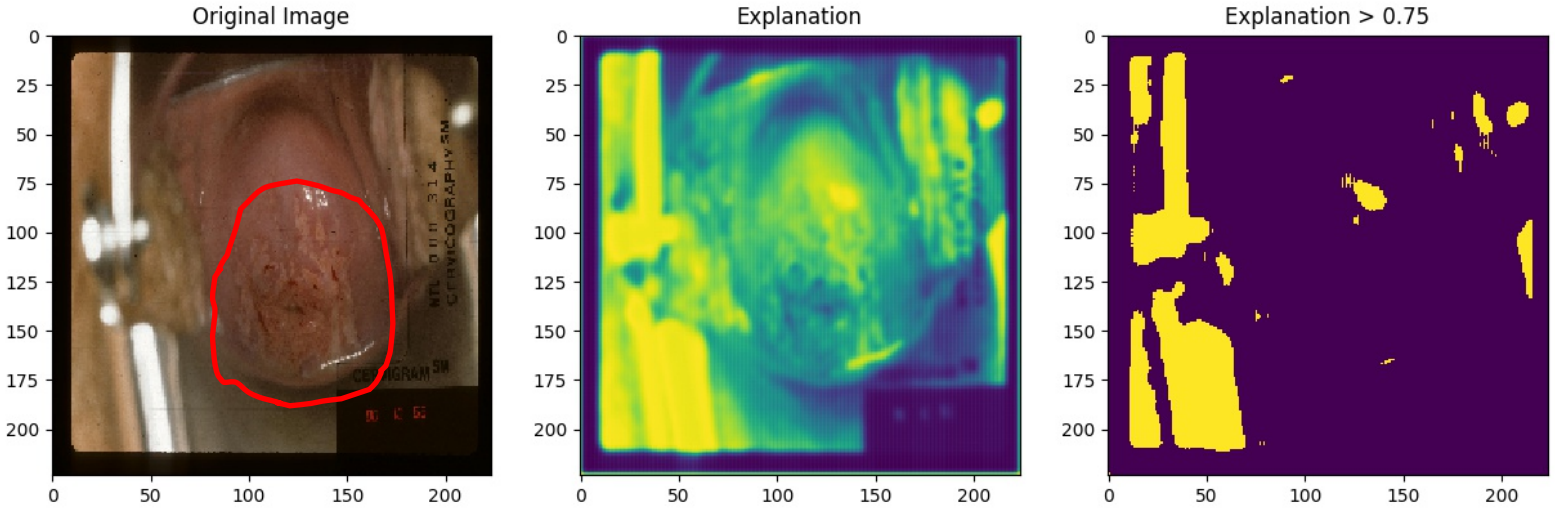
Weakly Supervised Approach



Explanation obtained with $\alpha = 0.88$
Normal case correctly classified (99.78%)

Results on the Cervical Cancer Dataset (5/5)

Weakly Supervised Approach



Explanation obtained with $\alpha = 0.88$
Normal case misclassified (86.87%)

Main Contribution

- A novel in-model joint approach:
  - Generates visual explanations
  - Applicable to different CNN-based classifiers
  - Custom training procedure and loss functions → no supervision or additional labelling
  - Validation on a real world medical scenario

Publications

- Isabel Rio-Torto, Kelwin Fernandes, and Luís F. Teixeira, “Towards a Joint Approach to Produce Decisions and Explanations Using CNNs”, In 9th Iberian Conference on Pattern Recognition and Image Analysis, Springer International Publishing, 2019.
  - Selected for oral presentation
  - One of the best ranked papers in the ML category
  - Honourable mention winner
  - Invited for extended version on Pattern Recognition Letters (1/2)

Future Work

- Further expert evaluation.
- Robustness assessment.
- Combination of the proposed approaches.
- Loss function with population-level properties.
- Comparison between different levels of annotation granularity.
- Application to other domains and modalities.

Producing Decisions and Explanations: A Joint Approach Towards Explainable CNNs

MASTER'S DEGREE IN ELECTRICAL AND COMPUTERS ENGINEERING

ISABEL RIO-TORTO DE OLIVEIRA

27 SEPTEMBER 2019

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

SUPERVISORS:
LUÍS TEIXEIRA
KELWIN FERNANDES