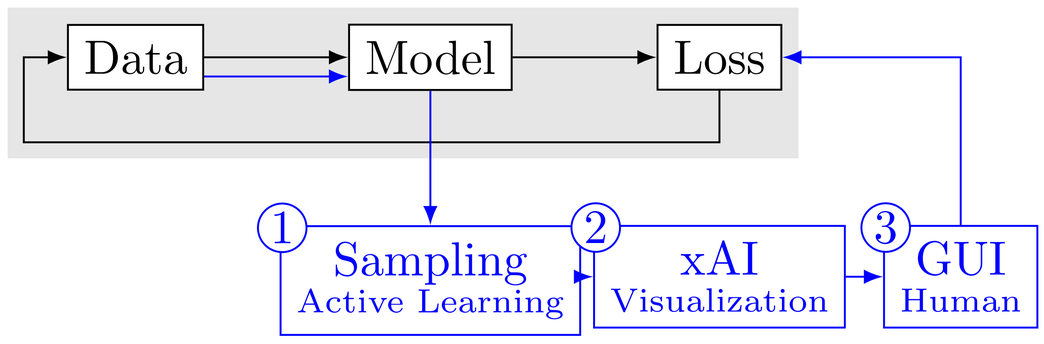
Interpretability-Guided Human Feedback During Neural Network Training

Pedro Serrano e Silva[1,2], Ricardo Cruz[2,3], Tiago Gonçalves[2,3], ASM Shihavuddin[4]

[1]NILG.AI, [2]Faculdade de Engenharia da Universidade do Porto, [3]INESC TEC, [4]Green University of Bangladesh

Motivation and Proposal

When models make wrong predictions, a typical solution is to acquire more data related to the error: an expensive process known as active learning. Our supervised classification approach combines active learning with interpretability so the user can correct such mistakes during the model's training. At the end of each epoch, our training pipeline shows examples of mistaken cases to the user, using interpretability to allow the user to visualise which regions of the images are receiving the model's attention. The user can then guide the training through a regularisation term in the loss function. This approach differs from previous works where the user's role was to annotate unlabelled data since, in this proposal, the user directly influences the training procedure through the loss function.



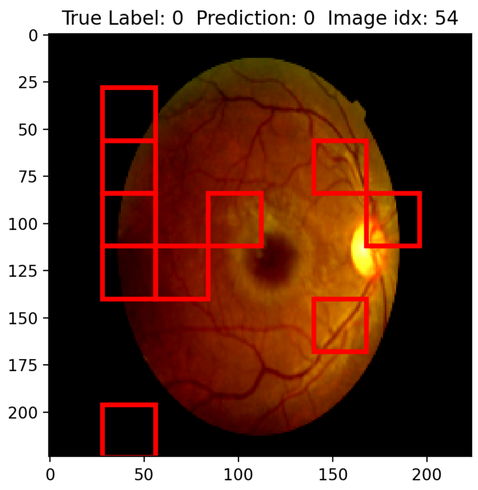
Implementation

Algorithm 1 Pseudocode of our training method.

function TRAIN(model, images, labels)
  preds $\leftarrow$ model(images)
  images $\leftarrow$ EntropySample(images)
  saliencies $\leftarrow$ xAI(images, preds)
  $W_{i,j} \leftarrow$ UserInterface(saliencies)
  loss $\leftarrow$ CE(preds, labels) $+ \sum_{i,j} \lambda W_{i,j} \frac{\partial \text{preds}}{\partial \text{images}_{i,j}}$
end function

We use entropy-based methods to sample the relevant images that will be input to the xAI algorithms. For each, saliency maps are generated and displayed to the user over the original image. The user can then click on the shown features. This process returns a weights tensor, later integrated into the loss function.

User Interface



The users can click on the squared regions they perceive as less relevant for the classification task.

Results

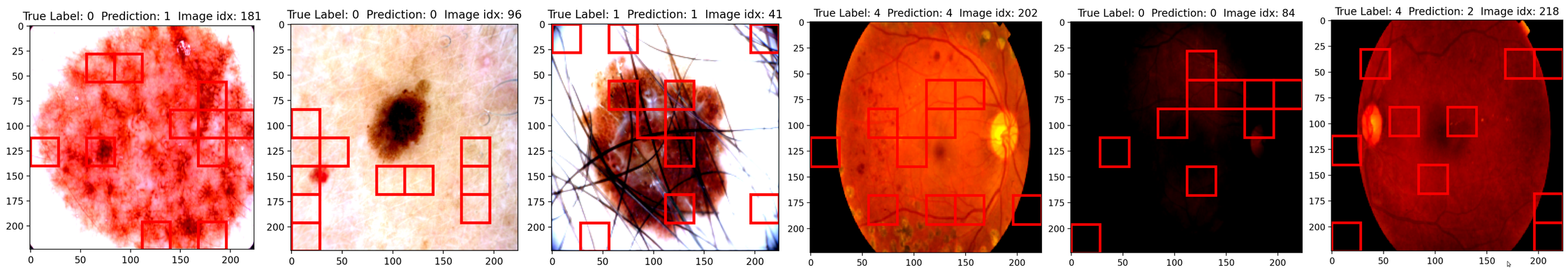
Results obtained for the different databases regarding the baseline and our approach:

Dataset	Method	Accuracy	Min Loss
ISIC2017	Baseline 10%	**85.91**	0.52
	HITL 10%	83.87	**0.47**
	Baseline 100%	88.59	0.34
APTOS2019	Baseline 10%	78.41	0.63
	HITL 10%	**81.14**	**0.61**
	Baseline 100%	85.11	0.47
NCI	Baseline 0.5%	92.52	0.41
	HITL 0.5%	**92.58**	**0.36**
	Baseline 50%	**93.18**	0.23
	HITL 50%	93.16	**0.22**

Querying Problems

We tested the pipeline on a multi-class classification task using different databases. We also identified several problems related to the querying step.

- On the ISIC2017 database, we highlight the following cases: ambiguous images (a), creating difficulties in their interpretation (b); poor summarization of the DeepLIFT attributions; the presence of hair in the image (c).
- On the APTOS2019 database, we point to the following cases: unclear attributions, although with a correct prediction (d); darkness hiding important features (e); possibly wrong label, and a prediction that seems to respect patterns found in images of the predicted label (f).



(a) (b) (c) (d) (e) (f)

Conclusions

- Experiments performed in three datasets showed some loss reduction – 0.47, 0.61, and 0.36 for the proposed pipeline versus 0.52, 0.63, and 0.41 for the baseline, respectively, for each database.
- There can still be complications when annotating the data, which can be due to the database or a consequence of the limitations of the xAI algorithms.

Future Work

- Fine-tune hyper-parameters such as the number of training epochs in which we ask humans for feedback, the type of sampling, the number and shape of squares to display, and the loss function.
- Explore other geometric shapes besides the current squared grid and change how to show the model's explanations to the user.

Acknowledgments: Supported by the Portuguese Foundation for Science and Technology - FCT within PhD grant 2020.06434.BD.