Visual Debates

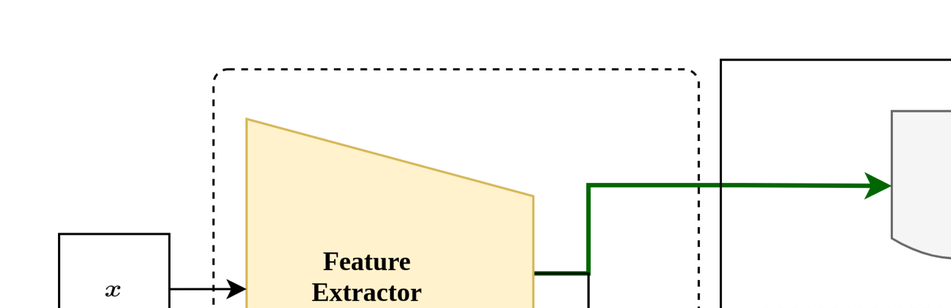
Avinash Kori[1,2], Ben Glocker[1], and Francesca Toni[2]

[1]Biomedical Image Analysis Group, Imperial College London | [2]Computational Logic and Argumentation Group, Imperial College London

1 Aim

*Given an input **x** and a pre-trained model (**Φ**), aim is to understand model's reasoning behind making certain prediction via computational argumentation.*

2 Existing Approaches



3 Our Approach

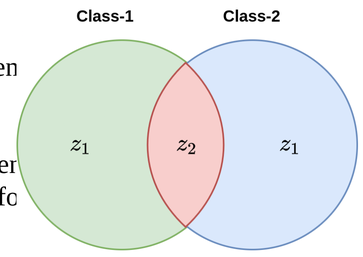
We model our framework as a multiplayer sequential zero-sum game, where players aim to maximize their utilities by adjusting their arguments with respect to other players' counterarguments in the process of understanding the classifier's reasoning.

4 Advantages

- The contrastive nature of our framework encourages players to put forward diverse arguments, picking up the reasoning trails missed by their opponents.
- The Debate framework focuses on interaction between arguments and not just feature importance.

7 Hypothesis

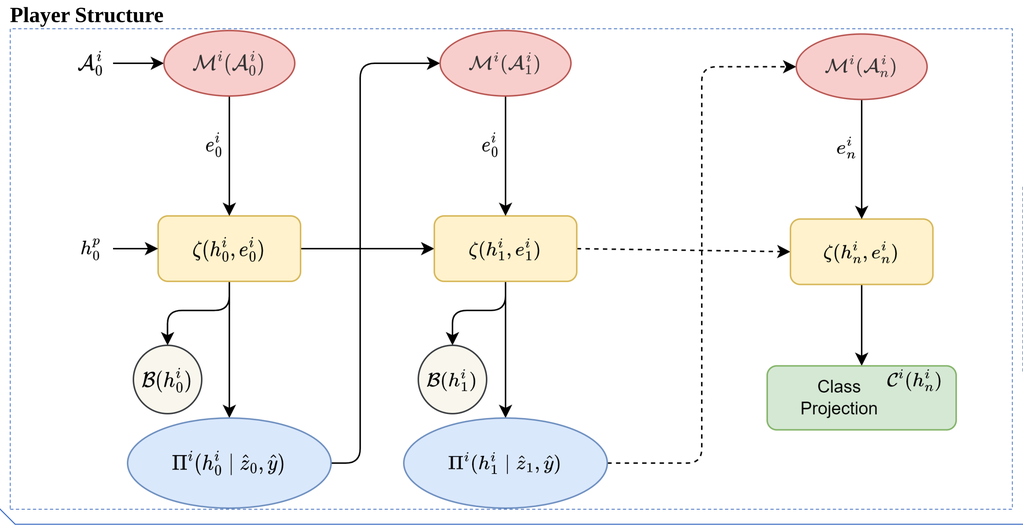
- In the proposed debate game, both the players converge at NE, making true and honest argumen about the given environment [2].
- At any NE, sampled features z for any given image can be divided into z_1 and z_2, such that $z_1, z_2 \subseteq z$, $z_1 \cup z_2 = z$ and $z_1 \cap z_2 = \varphi$, where z_1 is a set of features uniquely observed for a given class of images (semi-factual set of features) while z_2 is a set of features that can be observed fo multiple classes (counter-factual set of features), as described in the figure.



8 Game Structure

$$\Gamma =< \{\mathcal{Q}, z\}, \{\mathcal{P}^2, \mathcal{P}^2\}, \{\mathcal{A}^1, \mathcal{A}^2\}, \{\mathcal{C}^1, \mathcal{C}^2\}, \{\mathcal{U}^1, \mathcal{U}^2\} >$$

Player Structure



Game Objective

$$V(\mathcal{P}^1_{\theta^1}, \mathcal{P}^2_{\theta^2}) = \min_{\theta^1} \max_{\theta^2}$$
$$\mathbb{E}\Big[\sum_t \log \Pi_{\theta^1}(h_t^1 \mid \hat{z}_t, \hat{y}) \mathcal{U}_t^2(\mathcal{S}^1, \mathcal{S}^2)\Big]$$
$$- \mathbb{E}\Big[\sum_t \log \Pi_{\theta^2}(h_t^2 \mid \hat{z}_t, \hat{y}) \mathcal{U}_t^2(\mathcal{S}^1, \mathcal{S}^2)\Big]$$

5 Debate Framework

6 Quantization

- The quantization process initially requires us to define a codebook, with *K* discrete embeddings where each embedding is a *D* dimensional vector.
- We then define a discrete uniform prior and learn a categorical distribution as follows [1]:

$$\mathbb{P}(z = k \mid x) = \begin{cases} 1 & \text{for} \quad k =_i \|\Phi_e(x) - l_i\|_2 \\ 0 & \text{otherwise} \end{cases}$$

11 Results

n / $Z_R \rightarrow$ \ Dataset ↓	Non Committed			Split Ratio (Z_R)			Committed			Split Ratio (Z_R)		
	4	6	10	4	6	10	4	6	10	4	6	10
SHAPE	0.55	0.79	0.89	0.58	0.60	0.60	0.59	0.82	0.93	0.59	0.60	0.58
MNIST	0.58	0.61	0.75	0.44	0.45	0.56	0.52	0.64	0.73	0.40	0.47	0.58
AFHQ	0.60	0.77	0.80	0.43	0.58	0.54	0.61	0.81	0.79	0.53	0.59	0.59

Table 1: Ablation results on debate accuracy and split ratio wrt debate length

9 Properties

Arguments behavior at convergence for the model trained on SHAPE dataset, the plots
from left to right corresponds to 4, 6, and 10 argument debates.

10 Considered Models

We tested our framework on three different pre-trained models:

1. 5 layered sequential CNN, trained on MNIST dataset with an image resolution 32x32
2. 5 layered sequential CNN, trained on SHAPES dataset with an image resolution 32x32
3. Densenet 121 trained on AFHQ dataset with an image resolution 128x128

12 Conclusion

- We propose and justify the argumentative framework to demystify the reasoning process of any pre-trained CNN classifiers.
- Contestable/Argumentative approaches provides multiple different perspectives on the model's reasoning process.
- One main limitation in this work is to address human understandability of the generated player arguments, which we plan to address in future work.

13 References

- Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems 30 (2017)
- Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv preprint arXiv:1805.00899.
- Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. Advances in neural information processing systems, 27.
- Kori, A., Glocker, B. and Toni, F., 2022. Visual Debates. *arXiv preprint arXiv:2210.09015*.
- Github code: koriavinash1/VisualDebates (github.com)

Acknowledgements: This work was supported and funded by UKRI centre for Doctoral Training in Safe and Trusted AI (EP/S023356/1).

@koriavinash001 koriavinash1.github.io