Medical Imaging & Reconstruction Lab

Interpreting Deep Neural Networks for Medical Imaging using Concept Graphs

Avinash Kori, Parth Natekar, Balaji Srinivasan, and Ganapathy Krishnamurthi

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Introduction

Interpretability in abstract sense involves finding answers to the following questions:

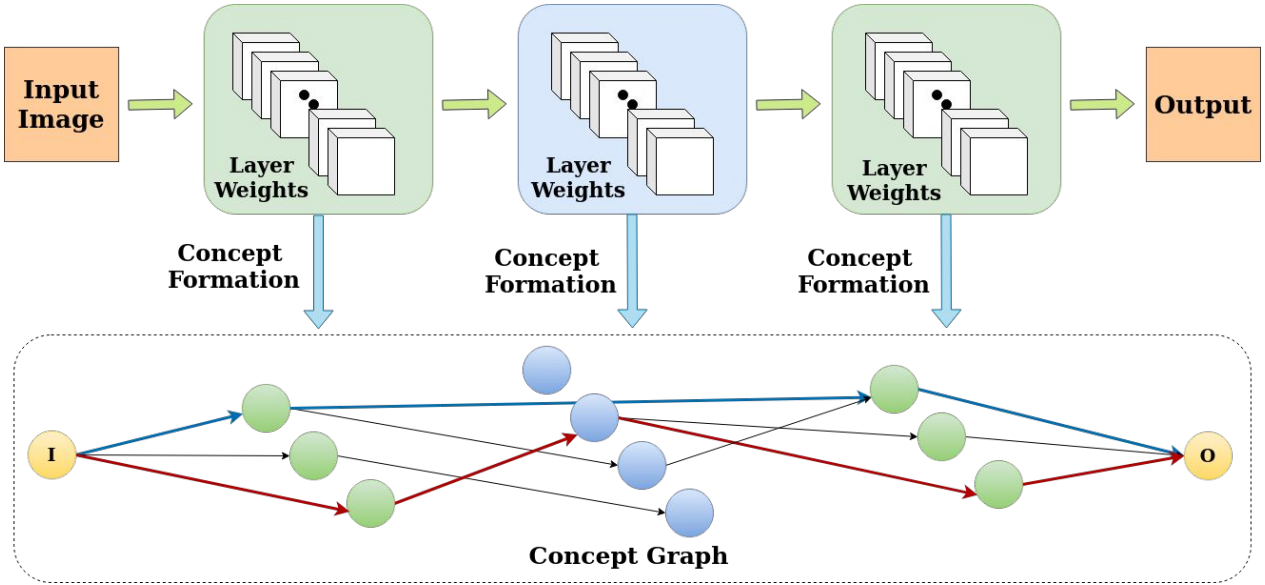
- Why did the model make that prediction?
- When can we trust the predictions of the model?
- How can we correct the errors made by the model?

Related Work

Methods	Attention maps	Concept Based	Hierarchy of steps
Dissection[1]	✔	✘	✘
GradCAM[2]	✔	✘	✘
SHAP[3]	✔	✘	✘
LIME[4]	✔	✘	✘
Ghorbani et.al[5]	✔	✔	✘
Ours	✔	✔	✔

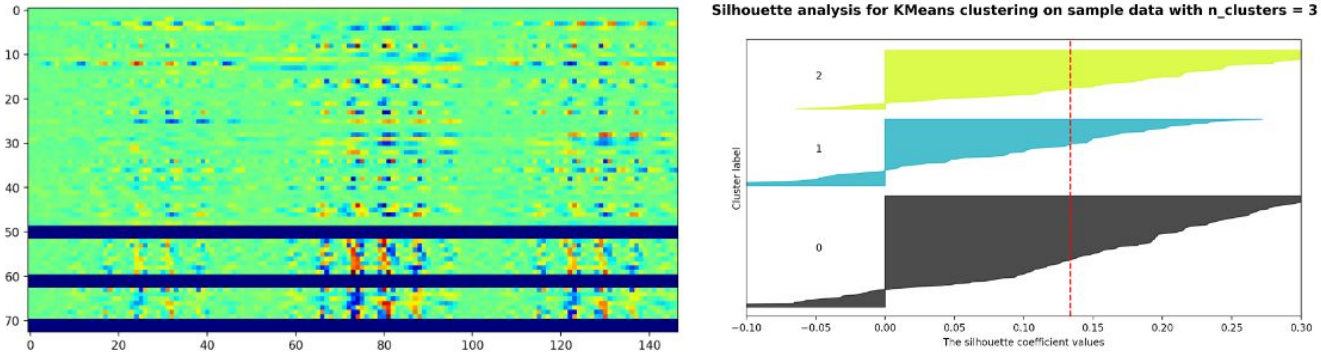
[1] Network Dissection: Quantifying Interpretability of Deep Visual Representations; [2] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization; [3] A Unified Approach to Interpreting Model Predictions; [4] "Why Should {I} Trust You?": Explaining the Predictions of Any Classifier; [5] Towards automatic concept-based explanations

Proposed Framework



Concept Formation

- **Weight Clustering** : Grouping of weights based on Silhouette coefficient[5] to identify optimal number of clusters in a given layer of CNN



(a) Initial layers of UNet, unrolled weights: 64, (3x3x16) weight tensor, **(b)** Silhouette analysis of the unrolled weight layer

[5] Silhouettes: a graphical aid to the interpretation and validation of cluster analysis

Concept Identification

$$y_p^l(x) = \frac{1}{Z} \sum_i \sum_j \left(\mathbb{E}_{k \sim idx_p} \Phi_{l,k}(x) \right)$$

$$\beta_{m,p}^l(x) = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_p^l(x)}{\partial \Phi_{l-1,m}(x)}$$

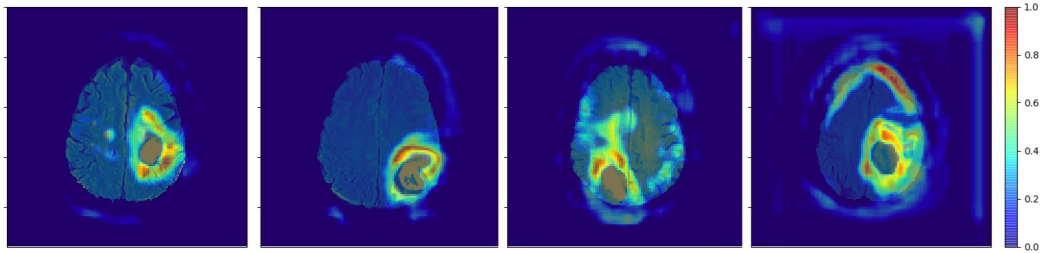
$$CAM_p^l = ReLU \left(\sum_m \beta_{m,p}^l(x) \Phi_{l-1,m}(x) \right)$$

CAM^l_p is concept attention map of cluster **p** in layer **l**, **β** is feature importance, and **y** is concept representative variable

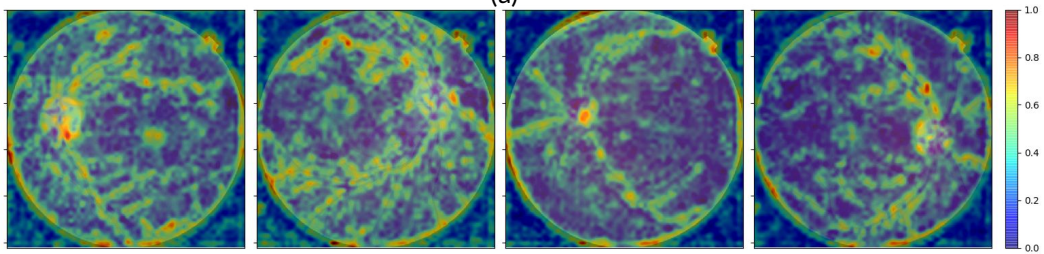
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Concept Completeness

- **Concept Consistency**: attention obtained by single concept over multiple images in a dataset

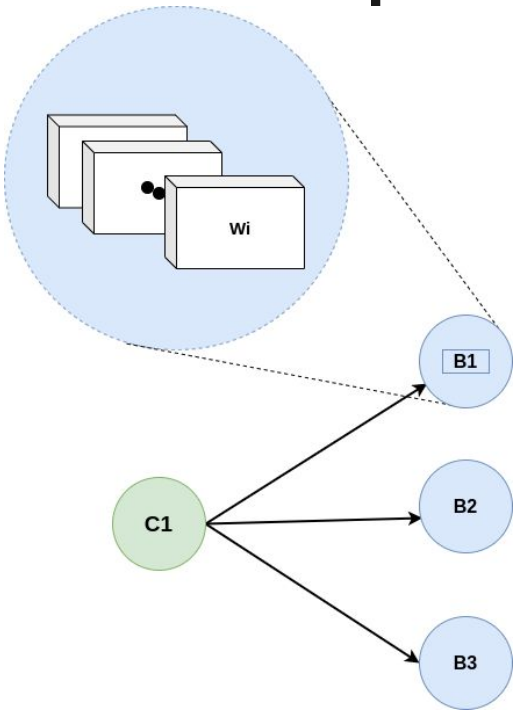


(a)

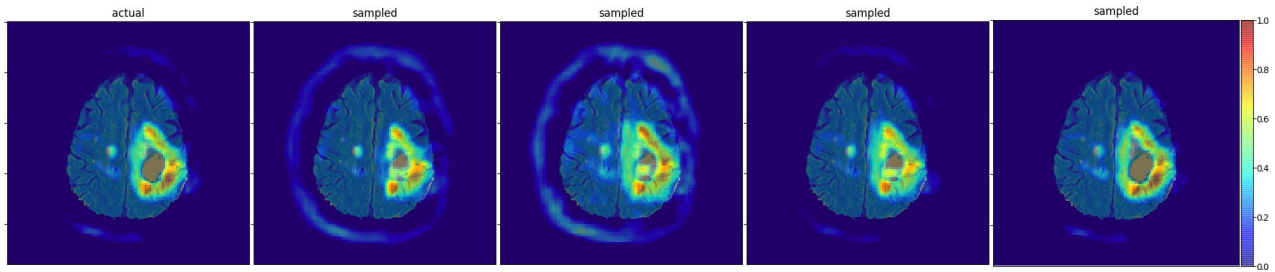


(b)

Concept Completeness

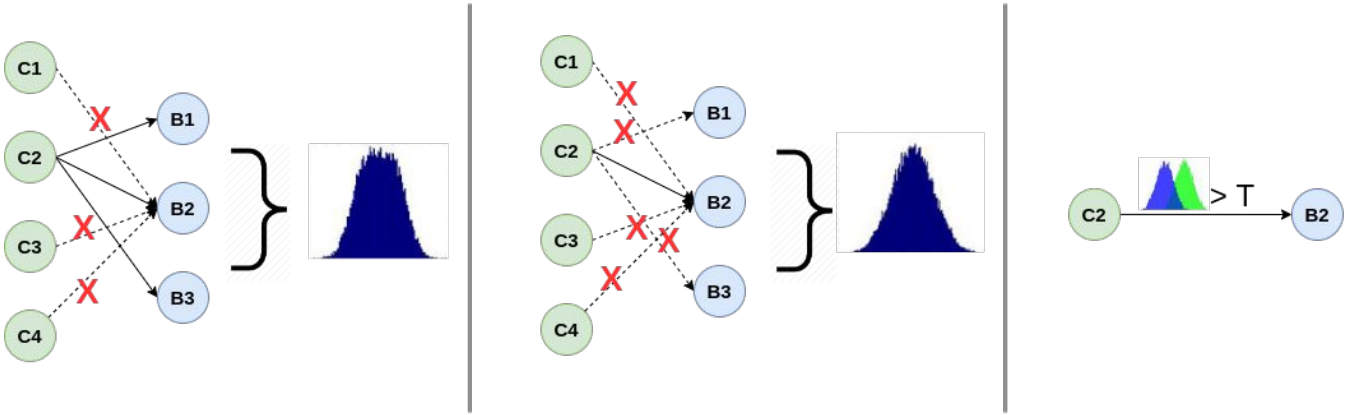


- **Concept Robustness**: attention obtained by sampled concept over an image in a dataset
- Aim of robustness is to analyse how spreaded the weights are in a concept



$\mathbf{C}_i$s are concepts in layer *(l-1)* and $\mathbf{B}_i$s are concepts in layer *l*

Concept Graph Formation

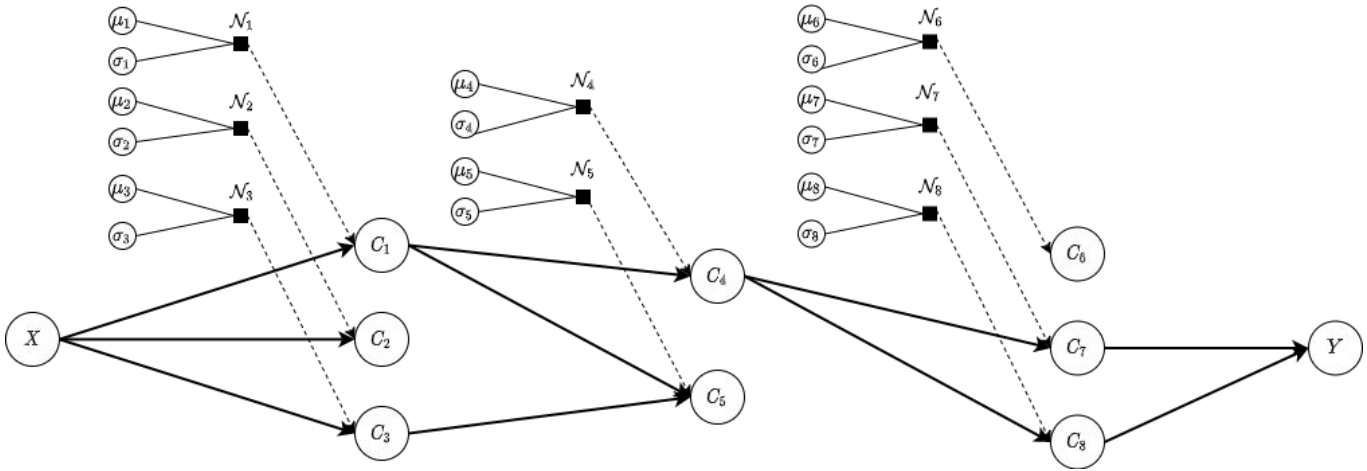


(a) Pre-interventional Distribution (b) Post-interventional Distribution (c) Existence of edge

$$\text{NMI}\left(\mathbb{Q}\left(\Phi_j(x \mid do(C^p_{-i} = 0), do(C^q_{-j} = 0))\right), \mathbb{P}\left(\Phi_j(x \mid do(C^p_{-i} = 0))\right)\right) > T$$

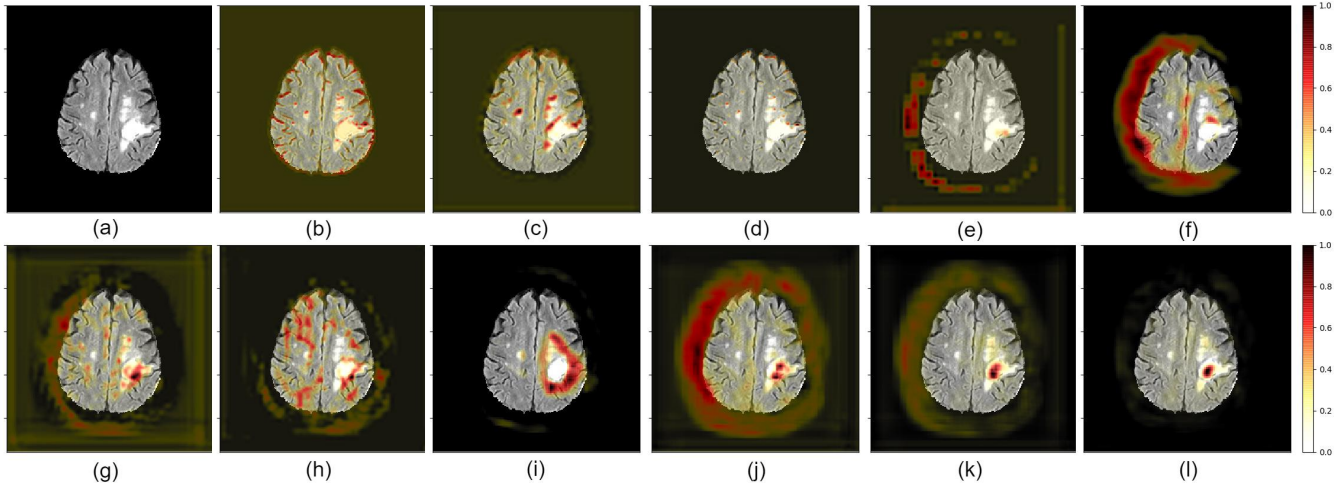
ℙ: pre-interventional distribution, **ℚ:** post-interventional distribution, **Φ:** trained model, and $\mathbf{C}_{-i}^{p}$ corresponds to all concepts other than **i** in layer **p**

Concept Graph Formation



Concepts

(a) C_0^3: doesn't capture any input region, **(b)** C_1^3: concave edges, **(c)** C_2^3: linear edges, **(d)** C_2^5: interior key points, **(e)** C_0^{13}: Lateral left hemispheric brain boundary, **(f)** C_3^{13}: Lateral left hemispherical and tumor core brain boundary, **(g)** C_2^{15}: Anterior tumor boundary, **(h)** C_3^{15} : Tumor core boundary, **(i)** C_2^{19}: Whole tumor boundary, **(j)** C_0^{17}: Lateral brain boundary and tumor core boundary, **(k)** C_1^{21}: Diffused tumor core region, **(l)** C_2^{21}: Tumor core region.



Trail Visualization

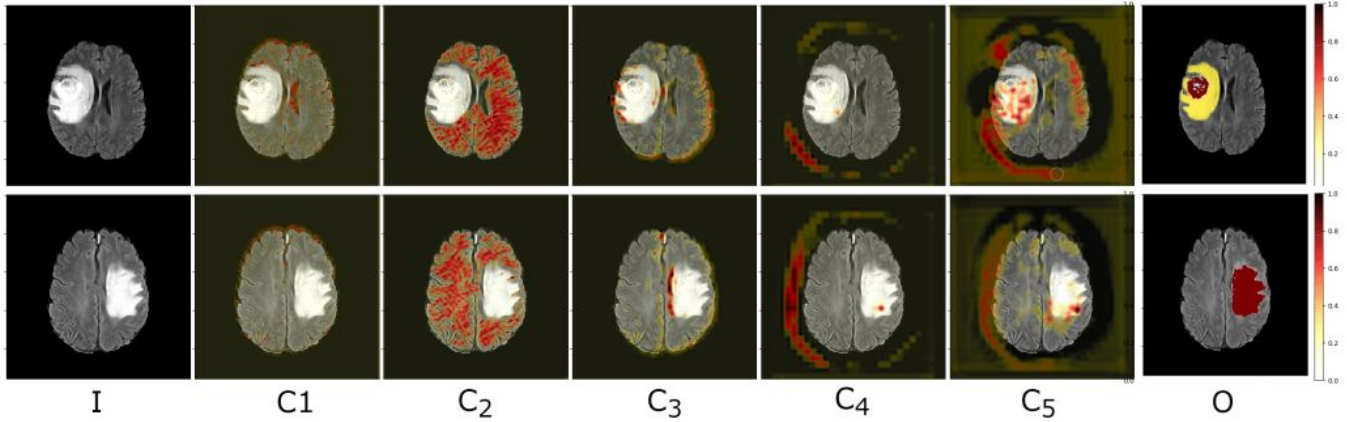


Fig. 6: Active inference trail for enhancing tumor (Each row is a trail for one input sample, red regions are high attention): *(**I**: Input image to a network) − > (C_1: Concave edges) − > (C_2: White matter region) − > (C_3: Tumor boundary) − > C_4: (Lateral brain boundary) − > (C_5: Inferior tumor boundary) − > (Enhancing Tumor)*

Trail Visualization

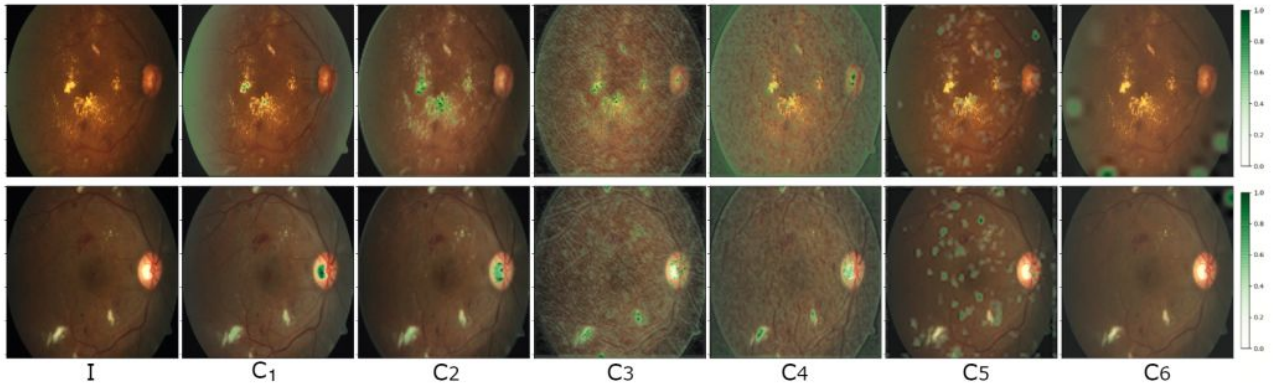


Fig. 7: Active inference trail for severe DR (green regions are high attention): *(I: Input Image) − > (C_1: Optic Cup/Hard exudates) − > (C_2: Hard Exudates) − > (C_3: Blood vessels, soft exudates) − > (C_4: Blood vessel, soft exudates) − > (C_5: dot-blot Hemorrhages/laser scar marks of retinal photocoagulation)*

Future Work

- Trail importance estimation
- Extension of work in ante-hoc interpretability(HAI), to use estimated trails in training phase
- Extension of approach on 3D networks and RNNs

Thank you

Github : https://github.com/koriavinash1/BioExp

Contact: koriavinash1@gmail.com

Acknowledgements: Dr. Ravikanth Balaji (Radiologist), Dr. Devika Joshi (Ophthalmologist), RBCDSAI, Reviewers, and Organizers