Interpretability of a Deep Learning Models

Avinash Kori, Ganapathy Krishnamurthi, Balaji Srinivasan
Robert Bosch Center for Data Science and Artificial Intelligence, IIT Madras, India
koriavinash1@gmail.com, {gankrish, sbalaji}@iitm.ac.in

Objective

Current deep learning models are deep-rooted in correlation. Understanding their behavior and interpreting them is essential for deploying any model (especially in the biomedical domain).

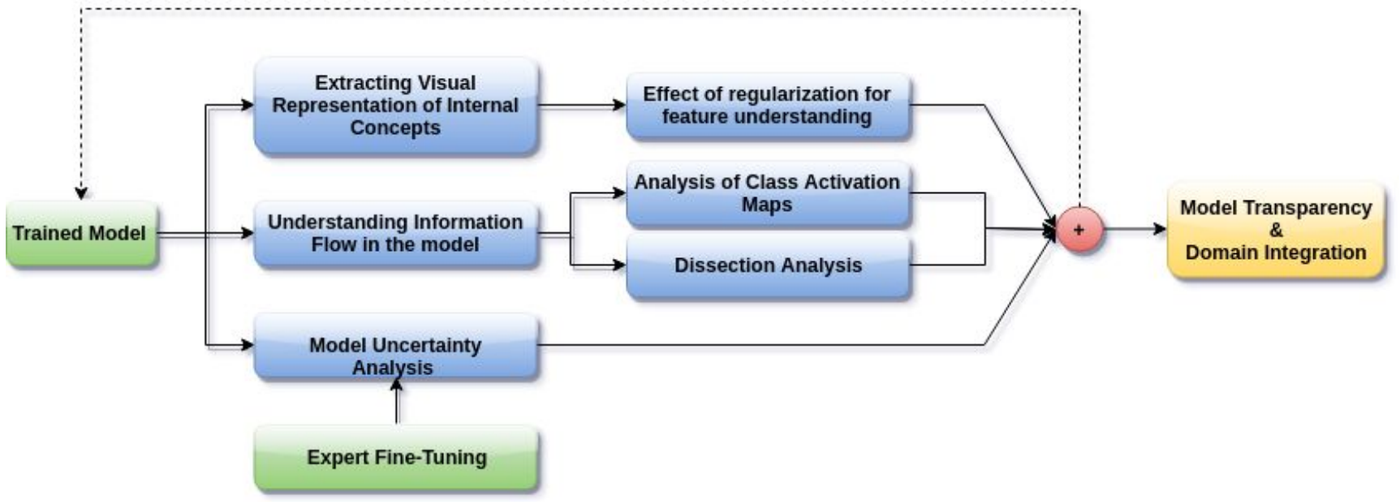
This work aims to develop a framework for the explainability of these models.

Main Questions in interpretability:

- **Why** did the model make that prediction?
- **When** can we trust the predictions of the model?
- **When** will the model fail?
- **How** can we correct the errors?

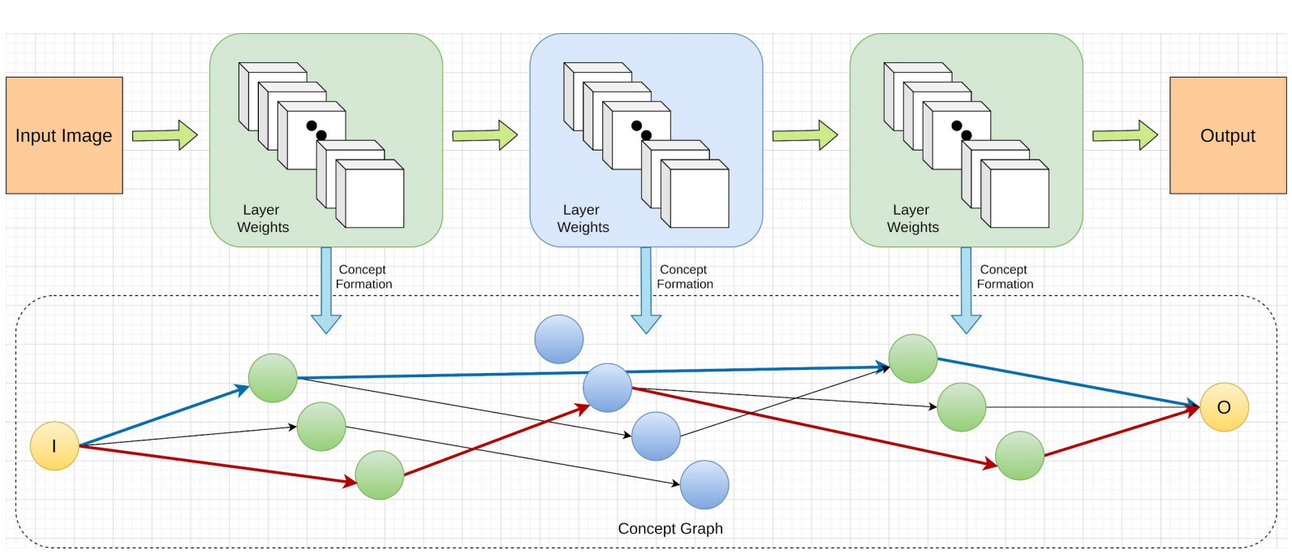
Here, in this work, we address the first two questions, 'Why' and 'When' can the model be trusted.

'Why?' Framework



Activation Maximization	Dissection	Uncertainty
$x^* = argmax_x(\Phi_{k,l}(x) - R_\theta(x) - \lambda\|\|x\|\|_2^2))$	$M_{k,l}(x) = \Phi_{k,l}(x) \geq T_{k,l}(x)$	$Uncertainty \approx \frac{1}{T}\sum_{t=1}^{T}\Phi(x \mid w^t)^T\Phi(x \mid w^t) - \mathbb{E}(\Phi(x \mid w^t))^T\mathbb{E}(\Phi(x \mid w^t))$
Generates representative image which maximizes the activation of specific filter	Generates implicit and explicit concepts learned by the network	Provides the information about uncertain regions in the prediction

'When?' Framework



3) Graph Formation

$$NMI(\mathbb{Q}(\Phi(x \mid do(C^p_{-i} = 0), do(C^q_{-j} = 0))), \mathbb{P}(\Phi(x \mid do(C^p_{-i} = 0)))) -$$

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

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

1) Concept Formation

- The idea of Concept is to group of all the weights responsible in the formation of a particular feature which encodes a particular human-understandable or non-understandable concept
- Networks weights are transformed using non-symmetrical transformation function, later they are clustered using a hierarchical clustering method using distance-based thresholding to form concepts

2) Concept Identification

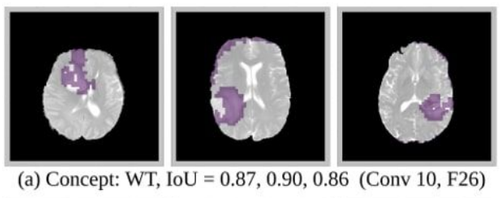
$$y(l,k,x) = \frac{1}{P}\sum_i\sum_j \Phi'_{l,k}(x)$$

$$\beta(l,k,x) = \frac{1}{N}\sum_i\sum_j \frac{\partial y(l,k,x)}{\partial \Phi'^{ij}_{l,k}(x)}$$

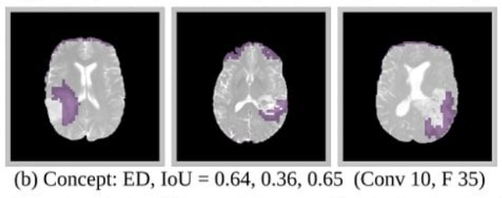
$$Concept = \mathbb{E}_{k \sim idx_p}\left(ReLU\left(\beta(l,k,x)\Phi'_{l,k}(x)\right)\right)$$

Results

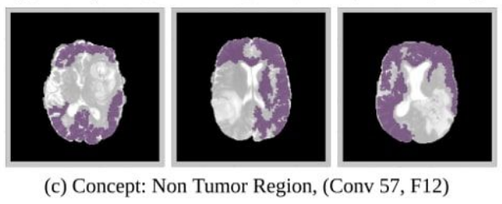
Network Dissection



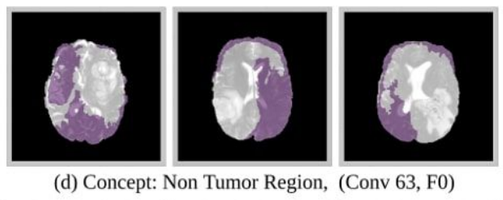
(a) Concept: WT, IoU = 0.87, 0.90, 0.86 (Conv 10, F26)



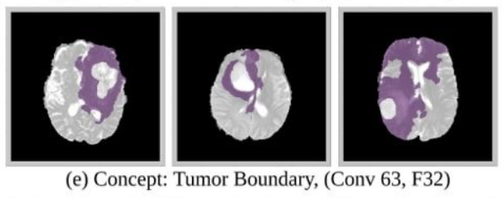
(b) Concept: ED, IoU = 0.64, 0.36, 0.65 (Conv 10, F 35)



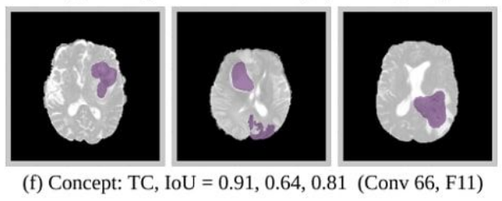
(c) Concept: Non Tumor Region, (Conv 57, F12)



(d) Concept: Non Tumor Region, (Conv 63, F0)

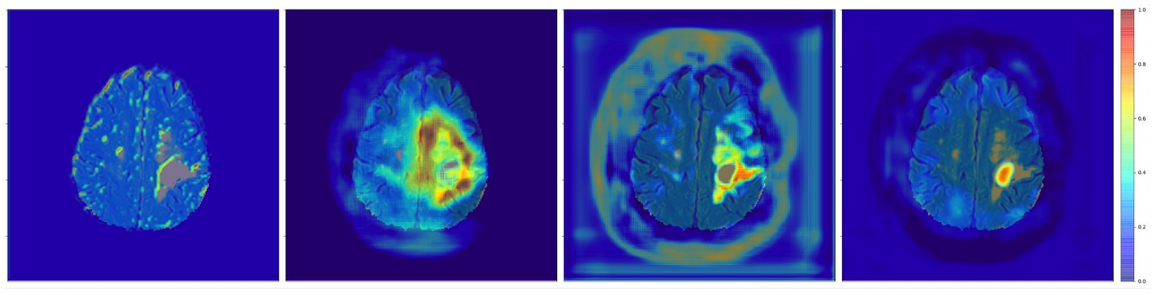


(e) Concept: Tumor Boundary, (Conv 63, F32)

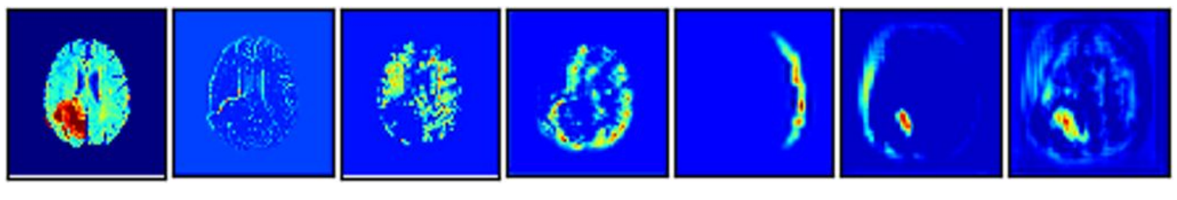


(f) Concept: TC, IoU = 0.91, 0.64, 0.81 (Conv 66, F11)

Concept Identification



Trail Estimation and Interpretation



(Input Image to a network) -> (Concave edge detector) -> (Corner keypoints all over the brain) -> (Anterior brain boundary and inner brain corner keypoints) -> (Lateral right hemispherical brain boundary) -> (Lateral left hemispherical brain boundary) -> (Lateral tumor region)

Code Available at: https://github.com/koriavinash1/BioExp