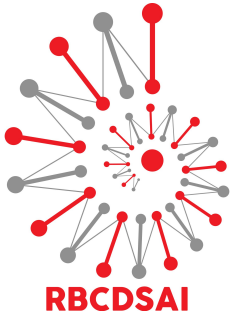


Interpretability of a deep learning models

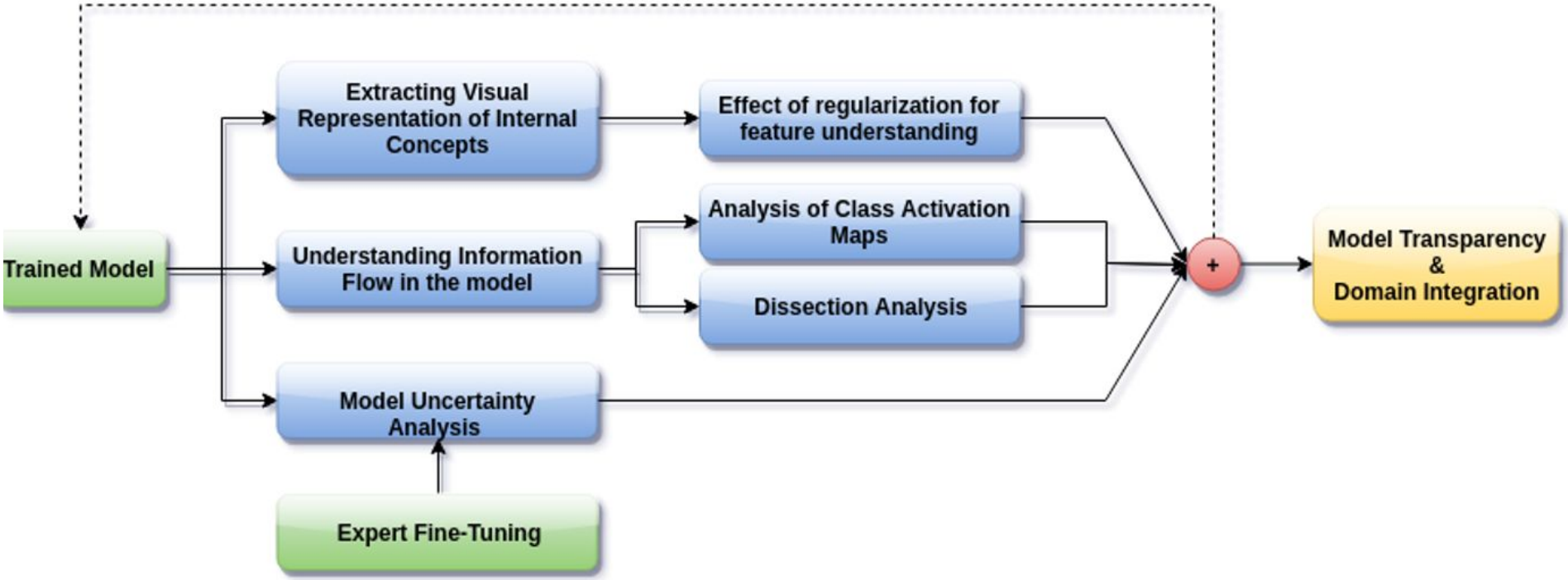
Avinash Kori, Ganapathy Krishnamurthi, Srinivasan Balaji

Problem and Importance

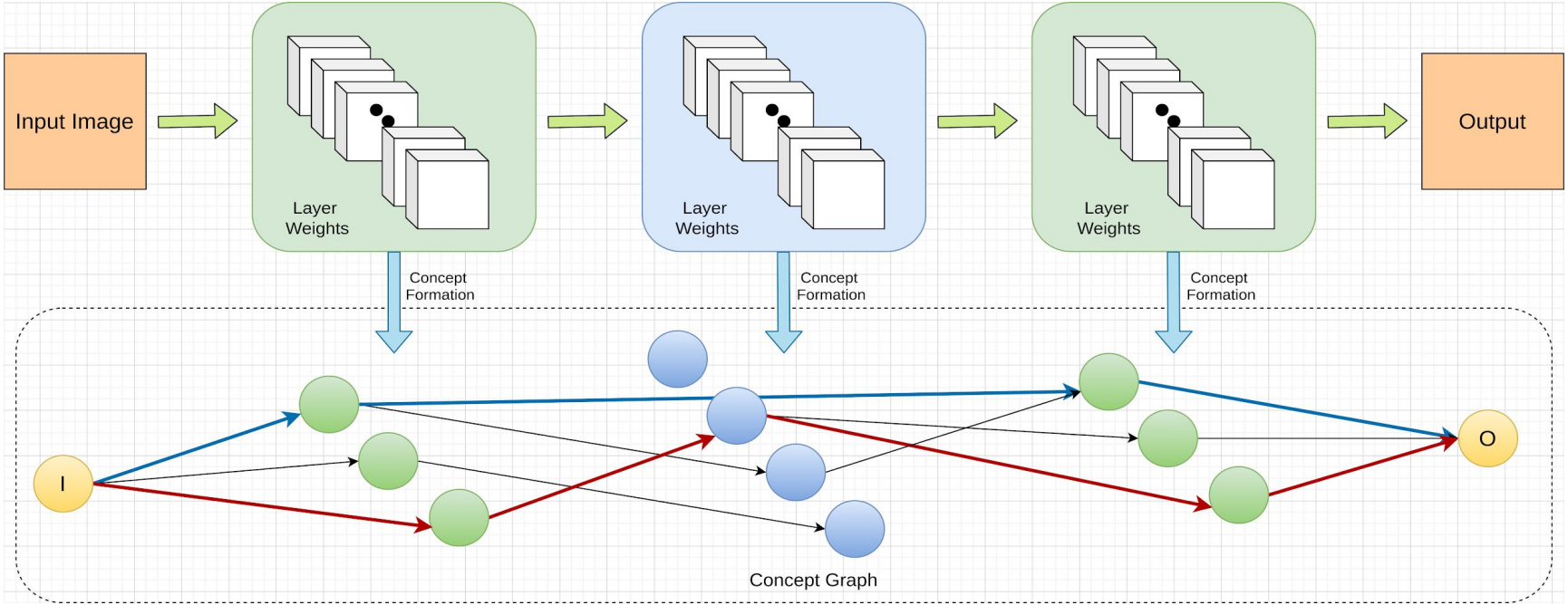
Current deep learning models are correlation based models. Explainability/ Interpretability of these models requires us to answer these following questions:

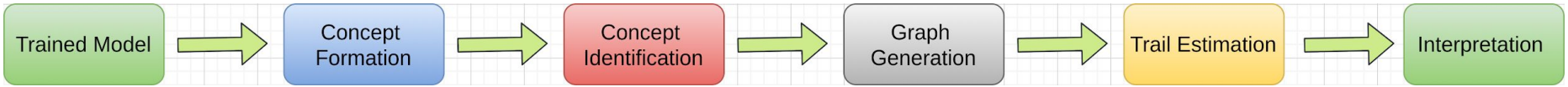
- **Why** did the model make that prediction?
- **When** can we trust the predictions of the model?
- **When** will the model fail?
- **How** can we correct the errors?

Why?


Trained Model
Extracting Visual Representation of Internal Concepts
Understanding Information Flow in the model
Model Uncertainty Analysis
Expert Fine-Tuning
Effect of regularization for feature understanding
Analysis of Class Activation Maps
Dissection Analysis
+
Model Transparency & Domain Integration

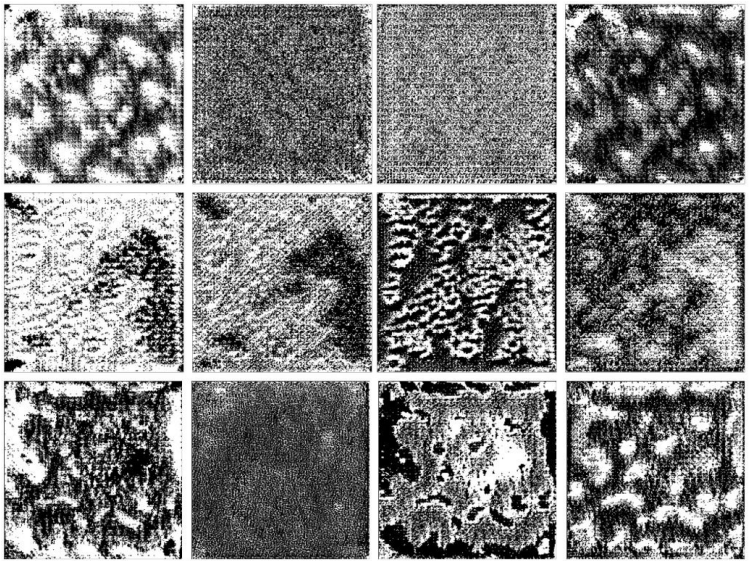
When?


Input Image
Layer Weights
Layer Weights
Layer Weights
Output
Concept Formation
Concept Formation
Concept Formation
I
O
Concept Graph

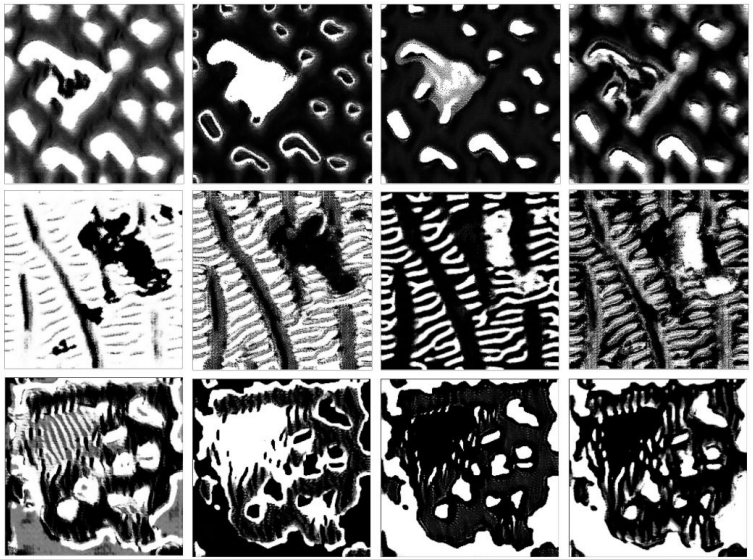

Trained Model
Concept Formation
Concept Identification
Graph Generation
Trail Estimation
Interpretation

Activation Maximization

$$argmax_x(\Phi_{k,l}(x) - R_\theta(x) - \lambda||x||_2^2)$$



(a) No Regularization

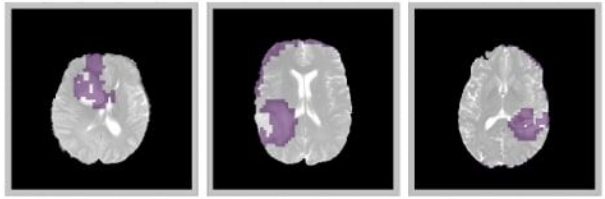


(b) With Regularization

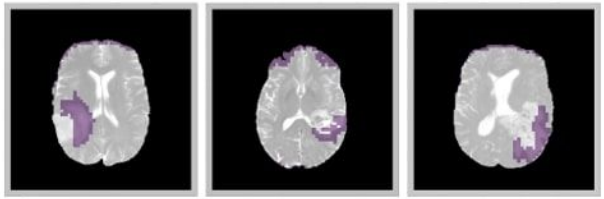
Network Dissection

$$M_{k,l}(x) = \Phi_{k,l}(x) \geq T_{k,l}(x)$$

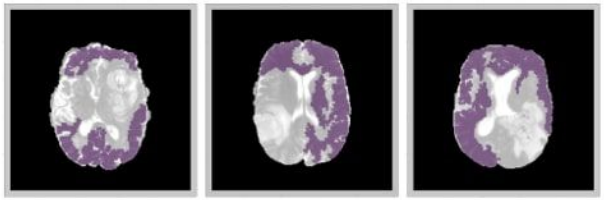
$$IoU(M_{k,l}(x), gt) = \frac{|M_{k,l}(x) \cap gt|}{|M_{k,l}(x) \cup gt|} \geq c$$



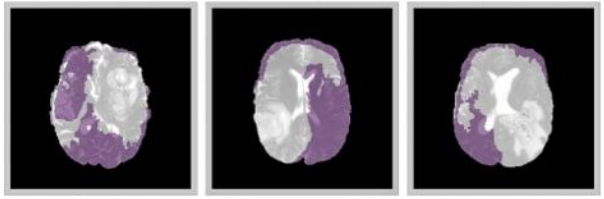
(a) Concept: WT, IoU = 0.87, 0.90, 0.86 (Conv 10, F26)



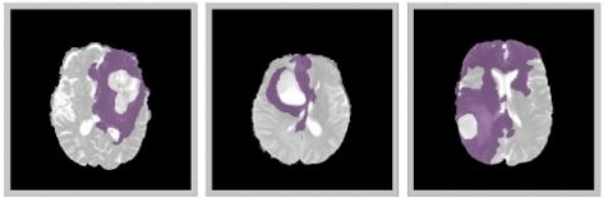
(b) Concept: ED, IoU = 0.64, 0.36, 0.65 (Conv 10, F 35)



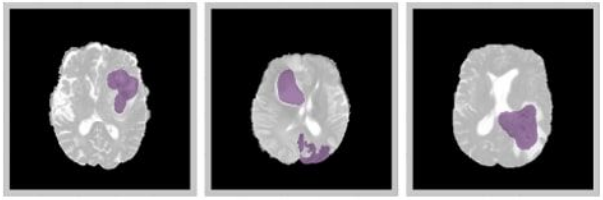
(c) Concept: Non Tumor Region, (Conv 57, F12)



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



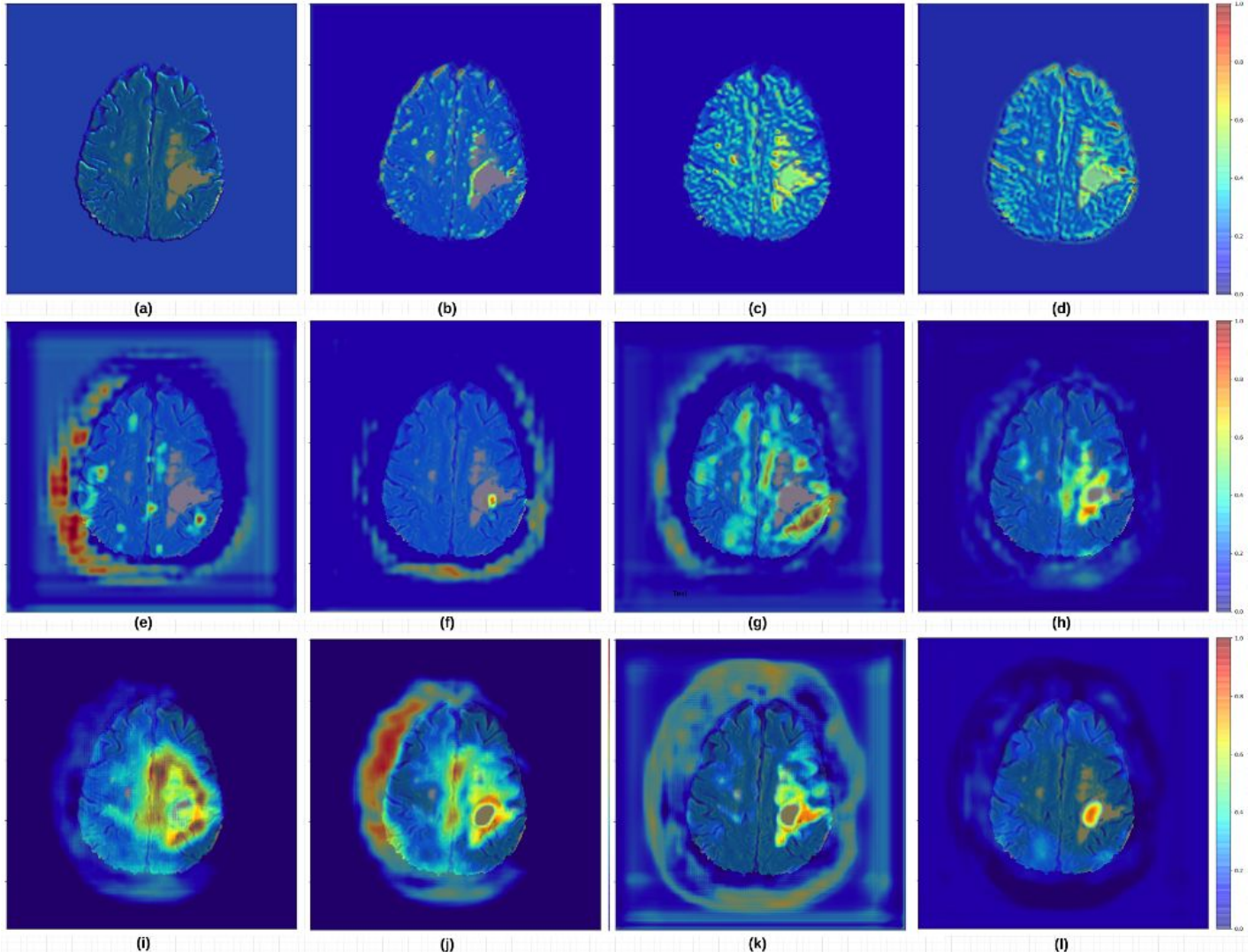
(e) Concept: Tumor Boundary, (Conv 63, F32)



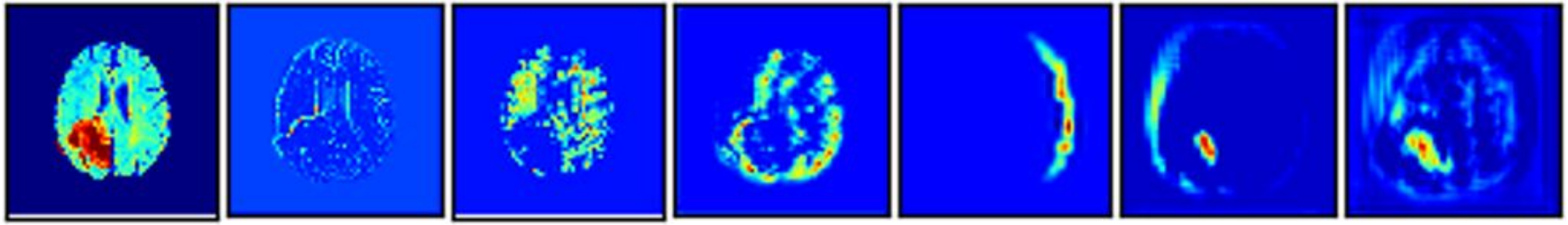
(f) Concept: TC, IoU = 0.91, 0.64, 0.81 (Conv 66, F11)

Concepts

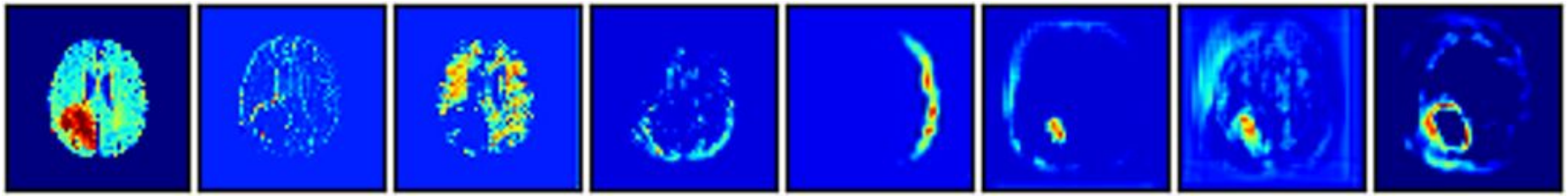
This figure illustrates the concepts obtained from various layers of a trained U-net model.



Trails



(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)



(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) -> (Superior tumor boundary) -> (Lateral tumor region) -> (Whole tumor boundary and edima region)

Thank You