ABSTRACTING DEEP NEURAL NETWORKS INTO CONCEPT GRAPHS FOR CONCEPT LEVEL INTERPRETABILITY

Avinash Kori
Department of Engineering Design
Indian Institute of Technology, Madras
koriavinash1@gmail.com

Parth Natekar
Department of Engineering Design
Indian Institute of Technology, Madras
patnat26@gmail.com

Ganapathy Krishnamurthi
Department of Engineering Design
Indian Institute of Technology, Madras
gankrish@iitm.ac.in

Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
sbalaji@iitm.ac.in

ABSTRACT

The black-box nature of deep learning models prevents them from being completely trusted in domains like biomedicine. Most explainability techniques do not capture the concept-based reasoning that human beings follow. In this work, we attempt to understand the behavior of trained models that perform image processing tasks in the medical domain by building a graphical representation of the concepts they learn. Extracting such a graphical representation of the model's behavior on an abstract, higher conceptual level would unravel the learnings of these models and would help us to evaluate the steps taken by the model for predictions. We show the application of our proposed implementation on two biomedical problems - brain tumor segmentation and fundus image classification. We provide an alternative graphical representation of the model by formulating a *concept level graph* as discussed above, which makes the problem of intervention to find active inference trails more tractable. Understanding these trails would provide an understanding of the hierarchy of the decision-making process followed by the model. [As well as overall nature of model]. Our framework is available at `https://github.com/koriavinash1/BioExp`.

keywords: Interpretation, Causality, Active-trails, Concepts, Concept-level-graph, Concept-identification

1 Introduction

Understanding the decision process of deep learning models is essential in domains such as biomedicine. Deep learning models are black boxes and as they are integrated into medical diagnosis, it becomes necessary to give a clear explanation of the concepts learnt by the model in a form understandable to medical professionals [1]. Clinicians also prefer upfront information about the global properties of a model, such as its known strengths and limitations [2].

For this, internal model logic and concepts need to be identified, and the relationships between these concepts needs to be shown in a human-understandable form. Previous interpretability techniques are visualization based, such as saliency map based methods or perturbation based methods to identify regions important regions in the input image. However, these techniques do not reflect the 'concept-based thinking' that human-reasoning shows [3], neither do they allow us to uncover the model's understanding of the relationship between such concepts.

Graphical models provide a tractable way to depict concepts and the relationships between these objects or concepts. However, there is a clear tug-of-war between model performance and transparency in this context [1]. Consider, for example, that we build a simple Bayesian Model for predicting the severity of Diabetic Retinopathy. Each node in the Bayesian Model is a human-understandable concept, such as microanuerisms, dark spots, exudates, and hemorrhages, and the output is the severity of Diabetic Retinopathy. Assuming we learn the structure and parameters of such a model,

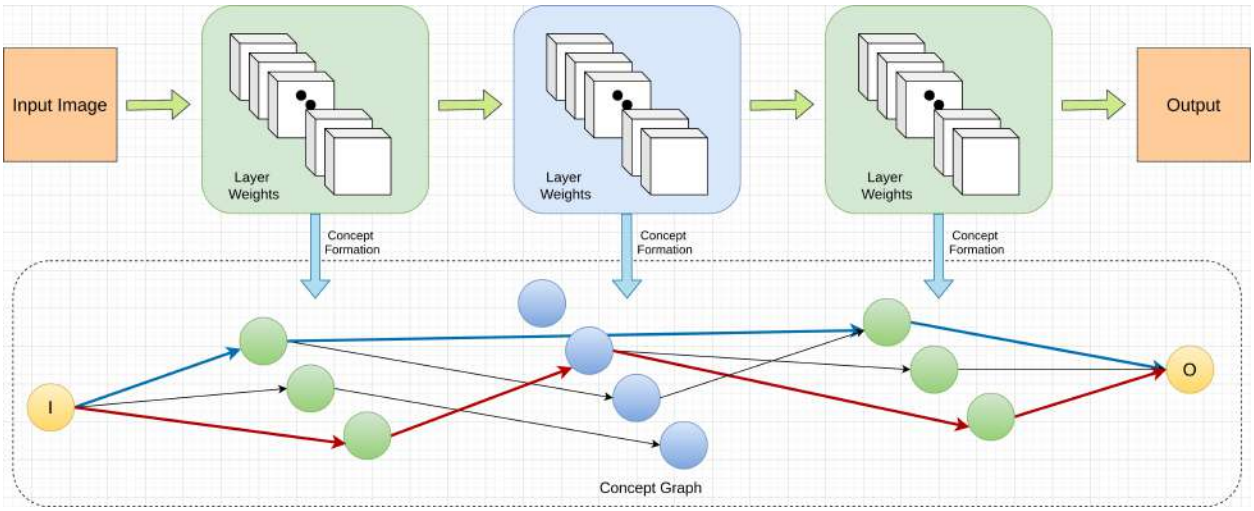


Figure 1: In the proposed framework, we construct a concept graph for any given trained model, which makes the problem more tractable. To generate concept graphs, we cluster weights in some user-defined layers of the network, use them as concepts, and later estimate links based on mutual information and information theory. Trails represented in red and blue show the interpretable steps for a network to predict the final result

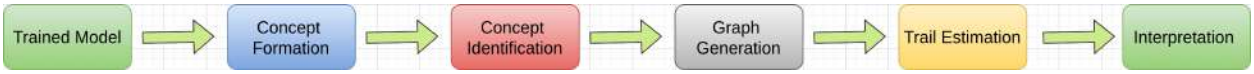


Figure 2: Above figure describes all the steps followed for interpreting as proposed in above framework fig, 1

we would have a completely transparent technique for our task. However, this comes at the cost of significantly reduced performance - a Bayesian Model built in this manner has neither the representational capacity to handle our data, nor the complexity to generalise.

Deep Neural Networks, on the other hand, provide a much more efficient way to represent and learn from image data. However, they do not lend themselves to the simple conceptual analysis that graphical models like a Bayesian Network do. We propose a method to massage a trained Deep Learning model into an equivalent graphical structure at the level of abstract, human-understandable concepts. This provides us with a simple, transparent representation of the model's logic and allows us to determine the pathway it takes for making a prediction. Such a concept level representation of a deep network is similar to that in deep probabilistic models, where the depth of the graph is considered over concepts instead of the depth of the computational graph [4].

We posit that such an abstraction is possible in a deep network since individual filters may be specialised to learn individual concepts. In the context of representation learning, it is hypothesized that deeper representation learning algorithms tend to discover representations that are more disentangled [5]. For example, experiments in Network Dissection show that individual filters learn disentangled visual concepts [6]. This behaviour has also been shown in the context of brain tumor segmentation models [7]. It is also known that filters in a network learn progressively more complex concepts, with initial filters acting as edge detectors and later filters learning more complex concepts [8]. Grouping filters which detect the same concept within a layer would then enable us to build a graphical representation of such visual concepts inherent in the network.

This representation of the model has many advantages. It can tell us about the model's biases - for example, if it relies heavily on one concept for one class of predictions. It also allows us to determine causal concept trails inherent in the model, as we have shown in this work. Our main contributions in this works are the following: (i) A method to represent a Deep Neural Network as a graphical model over abstract, high level concepts, encouraging concept-based explainability, (ii) Identification of inference trails from this graphical representation that help us understand the model's decision-making logic.

2 Proposed Framework

This work aims to abstract the model into an equivalent graphical model representation where concepts learnt by the network become nodes, and edges depict relationships between them. We take a clustering based approach to identify weights which may be detecting similar concepts in the input image. We perform our clustering on the weights instead of the features so that our explanations are independent of the input sample. Previous experiments show that for state-of-the-art DNNs trained on large-scale datasets like ImageNet [9], euclidian distance in the activation space of final layers is an effective perceptual similarity metric [10]. [It is not unreasonable that such behaviour extends to deep learning models in the medical domain]. We use the euclidian distance between weight vectors averaged across the channel dimension as our similarity metric.

We posit that the weight clusters thus identified are responsible for detecting individual concepts in the input image, and thus form the concept nodes in the abstracted graphical model. We visualize the concept detected by the clusters formed using a modification of Grad-CAM [11]. Such an approach works well since Grad-CAM basically visualizes attention of a weight layer on the input image. By zeroing out weights from other clusters and only keeping weights from a particular cluster before obtaining concept attention maps, we can find what the weights in a cluster look at and hence what concept they correspond to. Potential causal trails are then found from the generated graphical model using a normalized mutual information based approach.

The proposed framework for understanding the deep learning models consists of the following steps: (i) concept formation, (ii) concept significance analysis, (iii) concept identification, (iv) network formation, and (v) trail estimation. Figure 2 and 1 provide a detailed overview of the described framework. Next, we go over each section of this framework in detail.

2.1 Concept Formation

We posit that layers, especially in later stages of the network, learn a set of concepts (which are independent of each other). Essentially, groups of weight vectors in a layer are responsible for detecting a particular concept in the input image. Weight vectors can be clustered using a suitable metric and their attention over the input image can be used to determine the concept they are specialized to detect. Such an analysis can be performed at any level of granularity, for example one could perform the analysis choosing, say, only the first, fifth, ninth, and eleventh layers of a deep network so that a high level understanding can be gained of the concepts learnt by these layers and the relationships between them.

Let the trained network be $\Phi(W, X)$, and the layers chosen for analysis be $\{..., l-n, l, l+m, ...\}$. The clusters $\{C_p^l, C_q^l, C_r^l, ...\}$ are formed as a result of clustering weights at layer l in the network Φ. Let $W = \{w_1, w_2, ..., w_n\}$, where $W \in \mathbb{R}^{f \times f \times inc \times outc}$ and $w_i \in \mathbb{R}^{f \times f \times inc}$. Due to high dimensionality of the weight tensor, we take the mean of the weight tensor across the $outc$ dimension to obtain a representative tensor $w_i^{rep} = \frac{1}{inc} \sum_c w_i^c \in \mathbb{R}^{f \times f}$. To amplify the difference between symmetric weights we encode position information [12] along with weights.

Clusters are formed using a hierarchical clustering method [13] using distance-based thresholding. This provides additional degrees of freedom to group weights into as many numbers of significantly different concepts. After obtaining the clusters, we visualize the flattened weight vector to observe similarity among the clustered weights. Figure 3 depicts the same. Since direct visual interpretation is hard, to quantify the effectiveness of our clustering method we use $\mathbb{E}(SilhouetteScore)$ [14] as a metric. The Silhouette score measures the similarity of datapoints within the same cluster as opposed to the datapoints in other clusters.

2.2 Concept Identification

In the Concept identification step, we try to associate formed weight clusters with some region in the input image which corresponds to a human-understandable *concept*.

Consider cluster C_p^l. To identify the the concept that is learnt by the cluster and to depict this in a human understandable fashion, we first modify the trained network by dissecting the network at layer l, the outputs of which are denoted by Φ_l. Then, we perform a variation of Grad-CAM (which we will simply refer to as concept attention maps), using the filters in the cluster C_p^l as the outputs for which attention is to be computed, as described in equation 3.

In practice, this is done as follows. The dissected network Φ_l is modified by adding (1×1) convolution at the end, the weights of which are set to one. We then set the weights of all filters in the layer l which do not belong to the cluster p to zero (i.e., $do(C_{-p}^l = 0)$). The effective operation performed by the added convolutional layer Φ_{l+1} is then equivalent to taking the mean across the channel dimension of only those filters which belong to the cluster. We denote the output of this layer by $\mathbb{E}_{k \sim idx_p} \Phi_{l,k}$, where idx_p are the set indices in a layer l belonging to cluster C_p^l, equation 1.

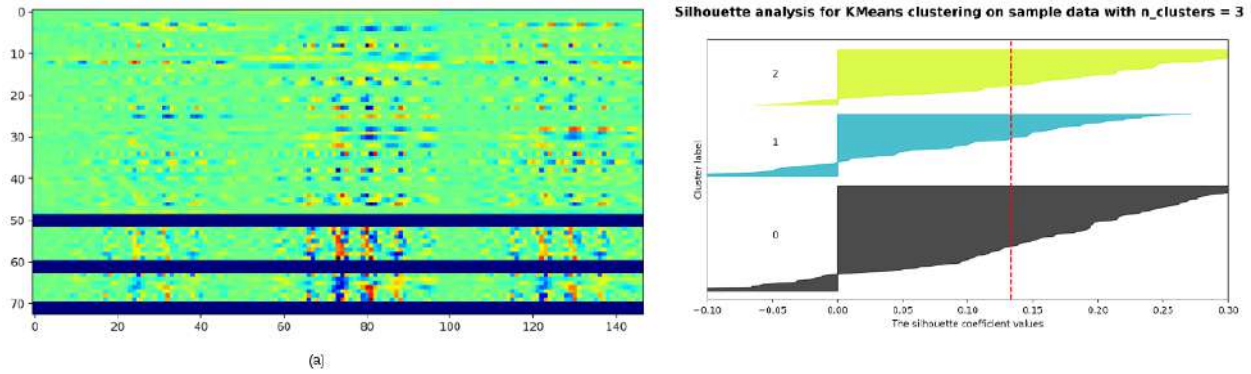


Figure 3: Above image describes the effectiveness of clustering. Sub-figure (a) describes the initial layer weights from ResNet50 trained on APTOS [15] data, in the figure dark blue horizontal bands seperates the weights among multiple clusters (provided figure has 3 clusters). Sub-figure (b) quantifies the effectiveness of clusters obtained as the result of proposed method

Concept identification then amounts to finding the gradient attention maps of this output with respect to the activations of the penultimate layer in the dissected network, i.e. Φ_{l-1} as described in equation 2.

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

$$\beta(l,k,x) = \frac{1}{Z}\sum_i\sum_j \frac{\partial y(l,k,x)}{\partial \Phi_{l-1}(x)} \tag{2}$$

$$Concept = ReLU\left(\sum_m \beta(l,k,x)\Phi_{l-1,m}(x)\right) \tag{3}$$

Where, m is the index of a filter in layer $l-1$.

Once the concepts are identified, we conduct a significance test to ensure that the concepts formed are consistent and are localized. To accomplish this, we first generate the distribution for each concept. Weights belonging to a specific layer in a neural network can be considered as i.i.d [16]. We posit that after learning, all the weights belonging to a particular cluster come from an underlying distribution and are i.i.d. We assume a gaussian generating distribution for weights in the cluster and approximate this using the first and second order moment of the weights in the cluster.

Consider an identified cluster $C_p^l \in \mathbb{R}^{f\times f\times inc\times n}$, where f is the filter size, inc is the number of in-channels, and n is the number of weights in the cluster. Let $w_i \in C_p^l$ be a weight belonging to the cluster C_p^l. Then, $w \in \mathbb{R}^{f\times f\times inc}$, i.e. the cluster C_p^l contains n weight tensors w_i of size $f \times f \times inc$. We generate a gaussian distribution for each pixel x_j at position j in the flattened weight w_i,

$$x \sim \mathcal{N}(\mu,\sigma) \tag{4}$$

$$\mu = \mathbb{E}_i(x_j), \sigma = \mathbb{E}_i(x_j - \mathbb{E}_i(x_j)) \tag{5}$$

We then sample n number of weights as detailed above, replace all n weights in the cluster C_p^l by the sampled weights, and recompute our concept attention maps.

We observe that recomputed concept attention maps are quite similar to our original concept attention maps. We also generate recomputed attention maps using a uniform sampler taken over the range of all weights in the layer, and compare this with the results of the gaussian samples over cluster. Results are described in appendix, section 6.1.

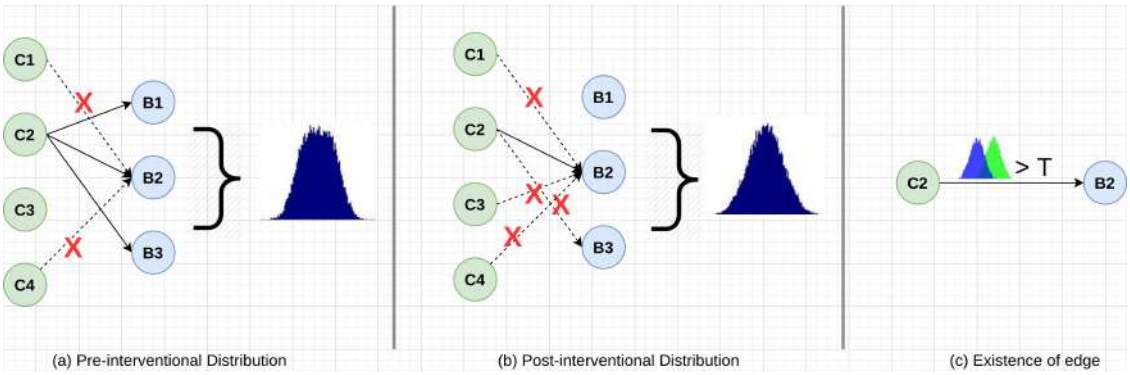


Figure 4: Above image describes the process of link formation. Sub-figure (a) describes how pre-interventional distribution is formed, sub-figure (b) describes how post-interventional distribution is formed, (c) exibits the condition for the existence of edge.

2.3 Network Formation and Information Flow

Once concepts and the related parameters have been identified for the given set of layers, we have the means to construct our equivalent graphical representation. We use a template based representation, shown in Figure 5. Given such a model, we could now identify inference trails that lead to a particular decision. Note that this graphical mode is not intended to be complete, only representative, i.e. it is not meant to completely explain the decision of the network. Since our graph can be constructed over any set of layers chosen by the user, there could be multiple inference trails that denote relationships between different concepts.

However, given a model, we can still identify relationships between chosen concepts to generate a human-understandable trace of inference which augments model predictions and provides medical professionals with a more complete diagnosis. In order to identify the relationship between two concepts, we use a mutual information based measure. We compute the normalized mutual information between the pre-interventional and post-interventional feature map distribution.

For the directed link between two concepts in layer p and q, $C_i^p \rightarrow C_j^q$, the pre-interventional distribution $\mathbb{P}(\Phi(x \mid do(C_{-i}^p = 0)))$ is the feature map distribution obtained on zeroing out the weights belonging to all concepts other than C_i^p in layer p (i.e., $do(C_{-i}^p = 0)$). Similarly, the post-interventional distribution $\mathbb{Q}(\Phi(x \mid do(C_{-i}^p = 0), do(C_{-j}^q = 0)))$ is the feature map distribution obtained at the layer q by zeroing out the weights belonging to all the clusters other than i in layer p as well as the weights belonging to the cluster C_j^q in layer q (i.e., $do(C_{-i}^p = 0)$ and $do(C_{-j}^q = 0)$). In this formulation the distributions are considered as pre and post interventional only with respect to layer q (layer of child node). Figure 4 shows this process graphically.

Based on our formulation, the directed link $C_i^p \rightarrow C_j^q$, exists only if equation 6 is satisfied.

$$NMI(\mathbb{Q}(\Phi(x \mid do(C_{-i}^p = 0), do(C_{-j}^q = 0))), \mathbb{P}(\Phi(x \mid do(C_{-i}^p = 0)))) > T \tag{6}$$

This basically states that the link exists only if the mutual information between pre and post interventional distribution is higher than a set threshold. High mutual information implies that a significant portion of the information flowing from the layer p to layer q occurs through that specific link $C_i^p \rightarrow C_j^q$.

2.4 Trail Estimation

Given our graphical representation and the existence of links between concepts, we now have a method to track inference steps taken by the model. The obtained concept graph is tree with depth m, where m is number of layers specified by the user for interpretability. The trails are all the paths running from input to a particular node used in an inference. The obtained trails encode the flow of concept level information used in making prediction.

We now have a graphical model representing our deep network that allows medical professionals to investigate inference trails within it. For example, consider the trail $X \rightarrow C_1 \rightarrow C_4 \rightarrow C_8 \rightarrow Y$. Medical professionals can then highlight whether or not such an inference trail makes sense from a biomedical perspective. This would allow them to understand the model's characteristics - its biases, pitfalls, and the common logical trails it uses to make an inference.

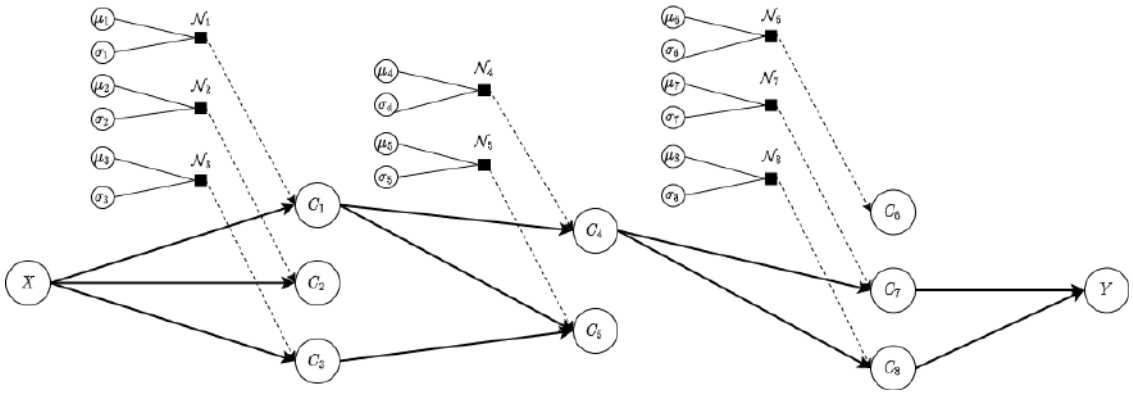


Figure 5: The constructed graphical representation for the network given the set of layers to analyse. Each pixel in a concept is drawn from its own gaussian distribution, using the mean and variance of the pixel over the cluster as parameters. Dotted arrows show the concept is sampled from its corresponding normal distribution. Dark arrows show links between concepts.

3 Experiments

We illustrate the working of our proposed framework on both classification and segmentation tasks. For the classification task, we considered the Diabetic Retinopathy problem, and for segmentation, we considered the Brain Tumor Segmentation problem. In both the experiments, the aim was to explain the building blocks of the model, and understand the hierarchy of decision making in deep learning models. All the experiments and results can be reproduced by using notebooks provided in the code repository `https://github.com/koriavinash1/BioExp_Experiments`.

3.1 Brain Tumor Segmentation

In the past decade, there has been significant development of image processing algorithms for segmenting tumors and intra-tumoral structures in brain MRI images [17]. Deep Learning has shown great potential in this context, with the BraTS challenge [18, 19, 20, 21] setting the benchmark for research in this area. The BraTS dataset contains nearly 300 brain MRI volumes annotated by experts for tumor regions. Various deep learning algorithms have shown great performance in segmenting tumor core, enhancing tumor, and edema regions from these MRI volumes. However, the black-box nature of deep learning algorithms are a hurdle the their integration in medical diagnosis.

We implement our algorithm on a ResNet based model [22] for brain tumor segmentation to cluster filters learning similar concepts, our model achieves 0.743, 0.693, and 0.523 of dice on whole tumor, tumor core and enhancing tumor segmentation respectively. We than visualize and identify these concepts, and try to build a graphical representation of the network's understanding of the problem. Finally, we estimate inference trails from the graphical structure seen.

3.1.1 Concepts

The $\mathbb{E}(SilhouetteScore)$ over all the data-points is 0.241, indicating the formation of some absolute clusters. Figure 6 describes the identified concepts for our brain tumor segmentation model. Initial layers (convolutional layers 3 and 5) correspond to edges in a specific direction. For example, in Figure 6(b) corresponds to concave edges, while 6(c) and 6(d) correspond to convex and approximately horizontal edges, respectively. In higher layers, filters start capturing more global information. It can be observed that some concepts capture brain boundary, while some capture tumor boundary. Figure 6 contains a description of all the global and local level concepts obtained from a network. The atlas [23] was used to formulate appropriate descriptions for every concept. This behaviour is in line with the understanding that filters in shallower layers learn simple patterns while deeper layers learn progressively more complex concepts [7].

3.1.2 Trails and Discoveries

Figure 7(a) and Figure 7(b) describe visual trails involved in predicting whole tumor boundary, and tumor core region respectively. These trails also indicate the structural hierarchy in estimating features from local to global features. They also show lateral to medial prediction patterns, which can also be observed in radiologist's analysis. Description of figure 7 also provide description trails along with visual trails in an image based on the predefined concept description.

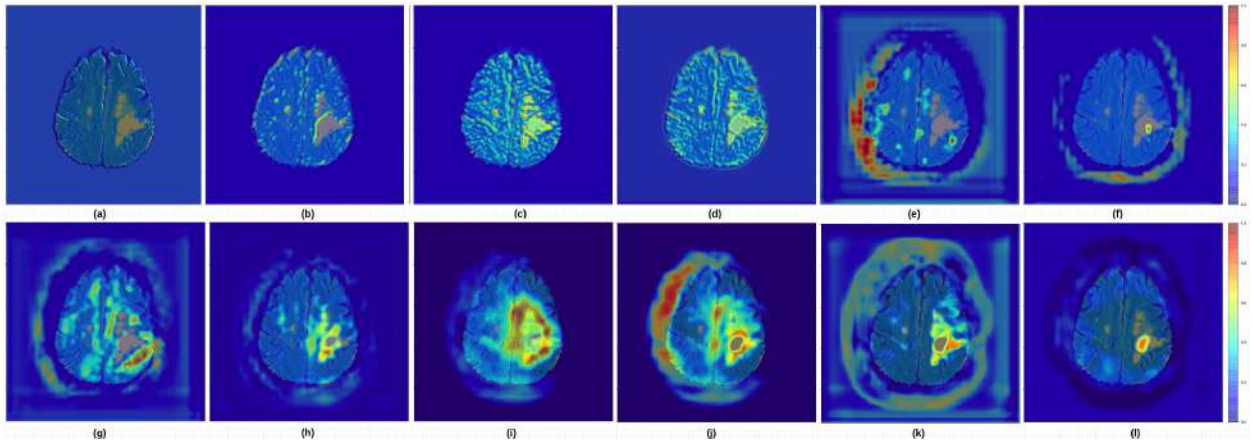
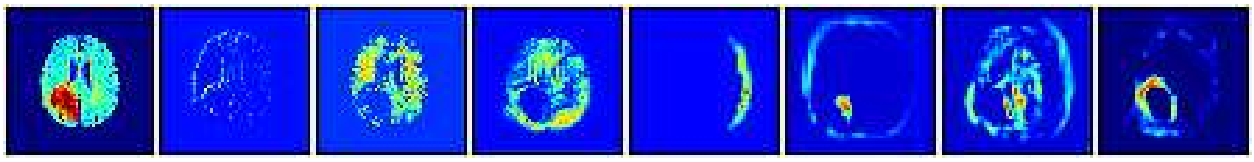
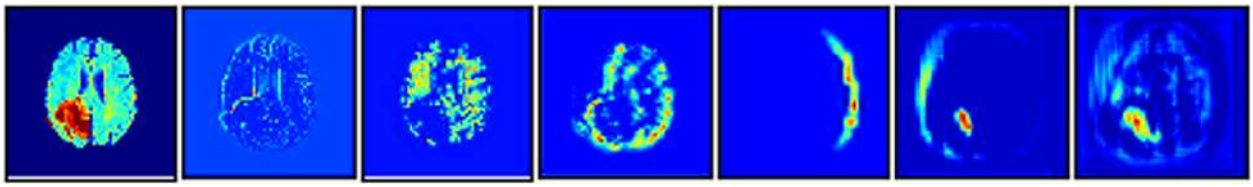


Figure 6: This figure illustrates the concepts obtained from various layers of a trained U-net model. Based on the region of activation we provide description of the concepts as follows: (local feature information) (a) C_0^3: doesn't capture any input region, (b) C_1^3: concave edges, (c) C_2^3: convex edges, (d) C_2^5: approximately horizontal edges. (global feature information) (e) C_0^{13}: Lateral left hemispherical brain boundary, (f) C_3^{13}: Lateral right hemispherical 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 left hemispherical brain boundary and tumor core boundary, (k) C_1^{21}: Edema region, (l) C_2^{21}: Tumor core region.



(a) Trail for whole tumor boundary



(b) Trail for tumor core

Figure 7: Above figure depicts visual trails followed by network, (a) shows the steps followed by the network in predicting whole tumor boundary and edima region, trail description is: *(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)*, (b) shows the steps followed in predicting tumor core, trail description is: *(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) -> (Tumor core region)*

3.2 Diabetic Retinopathy classification

Diabetic Retinopathy is frequent in individuals suffering from diabetes [24]. Deep Learning algorithms have shown great promise in detecting the severity of diabetic retinopathy or macular edema, and have the potential to greatly simplify diagnosis and detection. We implement our framework on a ResNet50 based network which achieves a Cohen Kappa Score of 0.71 on the APTOS dataset [15]. The APTOS dataset contains around 5000 human retina images taken using fundus photography. The severity of diabetic retinopathy has been rated for each image on a scale of 0 (no DR) to 4 (Proliferative DR).

Each stage of DR is characterized by certain features - such as microanuerisms, dark spots, exudates, and hemorrhages. These are concepts which medical professionals look for to determine severity of diabetic retinopathy. Thus, it becomes necessary to see how deep learning models process and identify them, and to see the model's understanding of relationships between these and the predicted severity of DR. These relationships and concepts are described in appendix, section 6.2

4 Related Work

Explainability is generally categorized into post-hoc and ante-hoc methods, where post-hoc explainability methods try to analyze and make inferences on trained models [25, 26, 27, 7]. In contrast, ante-hoc methods try to build an explainable model while training itself [28, 29, 30].

Current research directions in post-hoc interpretability focus mainly on visualizing network attributions or illustrative samples in the input space [11, 6, 31, 32]. Our work is related to methods involving disentangled latent representations and concept based explanations. For example, previous experiments on network dissection show that deep networks learn disentangled latent concepts [6]. Previous concept based interpretability methods [33, 32] use input patches to identify salient concepts that lead to a particular output. This has been extended to include a completeness measure for identified concepts [34]. However, neither of these methods considers the relationship between concepts learnt by the model. They also require a pre-processed set of input samples and explainability does not emerge organically from the model itself.

Our work introduces a post-hoc interpretability method, by abstracting the trained model into interpretable *concept graphs*, where concepts and their relationships emerge implicitly from the model, doing away with the need for user-curated input concepts. Our concept graphs allow easy visualization of the model's logic on an abstract, human-understandable level.

5 Discussion

We note a couple of limitations of our method. The experiments were performed on 2D data, which in the future, we plan to test on 3D data as well. The general idea of abstracting the network makes it more tractable in visualizing and understanding the network. We also assume the effectiveness of representative weight in concept formation phase. We experimentally show that averaging over channels also works (In appendix 6.1 we show the robustness and consistency of this method). This assumption helps us to reduce the dimensionality of weight tensors and conduct effective clustering. We plan to extend this work on 3D data and perform counterfactual analysis on the obtained trails in the future. This may provide a more statistical perspective of these concepts in a trail.

In conclusion, this work aims to provide concept-based interpretability for deep neural networks, demonstrating the results on medical data. We use a clustering technique to extract a graphical representation of concepts in the network, and visualize the clustered concepts using a variation of Grad-CAM. We then use an information-theoretic measure to determine relationships between concepts and build concept level inference trails within our network. Our results clearly show the formation of distinct concepts within layers that are consistent over the input dataset. We build concept level inference trails for deep networks trained for the two chosen tasks - brain tumor segmentation and diabetic retinopathy classification.

References

[1] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[2] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

[3] Sharon Lee Armstrong, Lila R Gleitman, and Henry Gleitman. What some concepts might not be. *Cognition*, 13(3):263–308, 1983.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[5] Yoshua Bengio. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing*, pages 1–37. Springer, 2013.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[7] Parth Natekar, Avinash Kori, and Ganapathy Krishnamurthi. Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Frontiers in Computational Neuroscience*, 14:6, 2020.

[8] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[12] Avinash Kori, Ganapathy Krishnamurthi, and Balaji Srinivasan. Enhanced image classification with data augmentation using position coordinates. *arXiv preprint arXiv:1802.02183*, 2018.

[13] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[14] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[15] Asia Pacific Tele-Ophthalmology Society. Asia pacific tele-ophthalmology society 2019, dataset, 2019.

[16] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016.

[17] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[18] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[19] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer, 2017.

[20] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[21] Avinash Kori, Mehul Soni, B Pranjal, Mahendra Khened, Varghese Alex, and Ganapathy Krishnamurthi. Ensemble of fully convolutional neural network for brain tumor segmentation from magnetic resonance images. In *International MICCAI Brainlesion Workshop*, pages 485–496. Springer, 2018.

[22] Adel Kermi, Issam Mahmoudi, and Mohamed Tarek Khadir. Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes. In *International MICCAI Brainlesion Workshop*, pages 37–48. Springer, 2018.

[23] Song-Lin Ding, Joshua J Royall, Susan M Sunkin, Lydia Ng, Benjamin AC Facer, Phil Lesnar, Angie Guillozet-Bongaarts, Bergen McMurray, Aaron Szafer, Tim A Dolbeare, et al. Comprehensive cellular-resolution atlas of the adult human brain. *Journal of Comparative Neurology*, 524(16):3127–3481, 2016.

[24] Donald S Fong, Lloyd Aiello, Thomas W Gardner, George L King, George Blankenship, Jerry D Cavallerano, Fredrick L Ferris, and Ronald Klein. Retinopathy in diabetes. *Diabetes care*, 27(suppl 1):s84–s87, 2004.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[26] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[27] Berk Ustun and Cynthia Rudin. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*, 2014.

[28] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[29] Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M Pintea, and Vasile Palade. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. *arXiv preprint arXiv:1708.01104*, 2017.

[30] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49(7):2401–2414, 2019.

[31] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[32] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677, 2018.

[33] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9277–9286, 2019.

[34] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019.

6 Appendix

6.1 Concept Identification

In this section we empirical results justifying our claims in clustering and concept formation. We evaluate consistence and robustness of the proposed method to justify our claims.

Consistency: To evaluate the consistency of the proposed method, we examine the regularity of concepts over multiple input samples in our datasets. Figure 8 illustrates the same, where each row corresponds to a different concept, and each column corresponds to a visualization of the concept over different images in the input dataset. It can be observed that similar concepts encode similar regions in the images irrespective of tumor location or optic disk location.

Robustness: Here, we try to justify our claim of assuming Gaussian priors over the weights belonging to a particular cluster. We do the same by comparing the results obtained by considering Uniform and Gaussian prior over layer weights. Comparative results are described in figure 9 and figure 10. It can be observed that distribution formed by using the statistics of cluster weights reproduces similar concept attention maps compared to the other two experiments. These results justify that the weights are inherently grouped into different clusters responsible for identifying certain concepts.

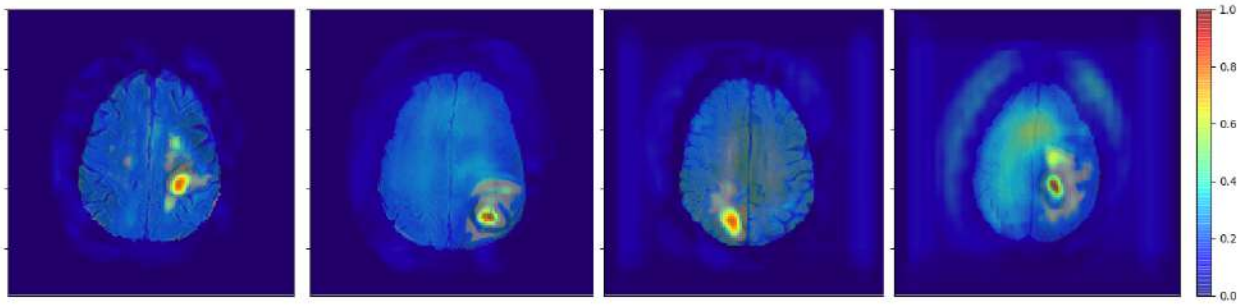
6.2 APTOS Results

6.2.1 Concepts

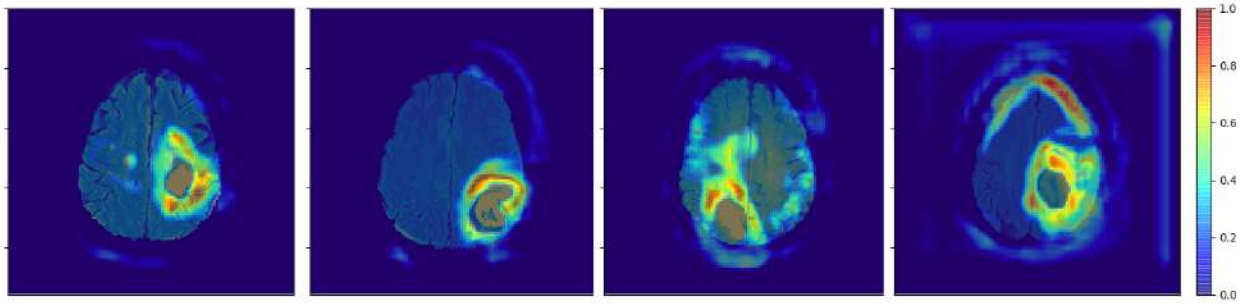
The $\mathbb{E}(SilhouetteScore)$ over all the data-points is 0.2, which again indicates the formation of some absolute clusters. Figure 11 describes the identified concepts; Figure 11 indicates local concepts and global level concepts, encoding blood vessels, exudates, dark spots, etc.

6.2.2 Trails and Discoveries

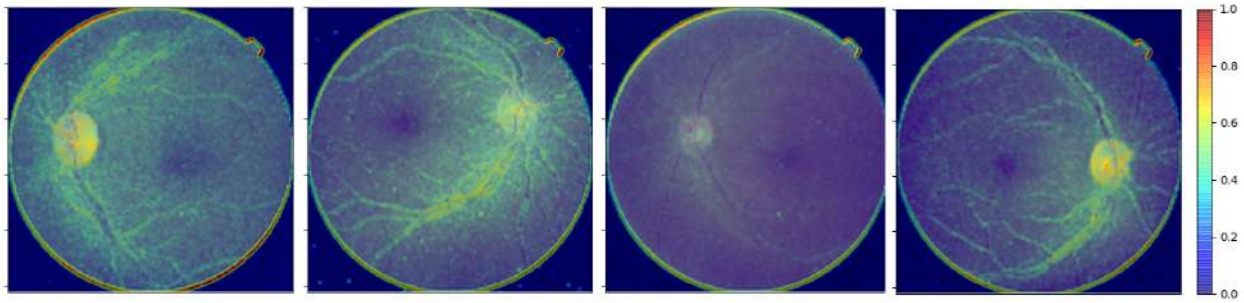
Similar to the trails obtained in BraTS experiment 3.1.2, we can observe local to global hierarchy and lateral to medial prediction patterns can even be in diabetic retinopathy. Figure 12(a) and Figure 12(b) describes visual trails involved in predicting 'Mild', and 'Proliferative' classes of diabetic retinopathy respectively.



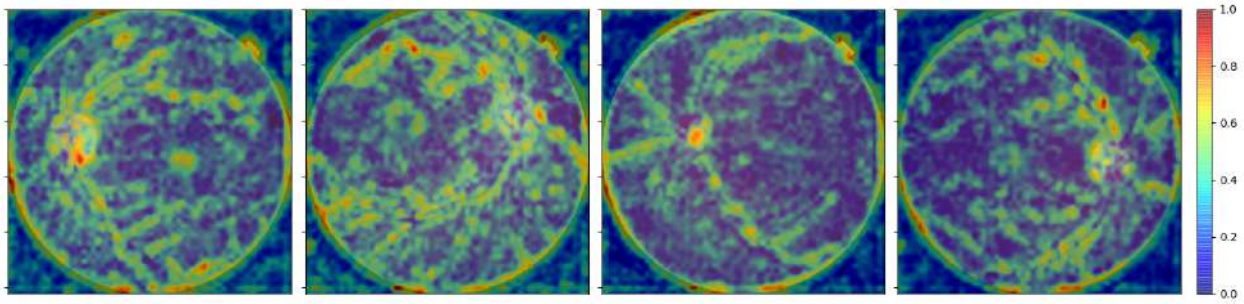
(a) BraTS Concept: C_2^{21} Tumor Core region



(b) BraTS Concept: C_2^{19} Whole Tumor boundary

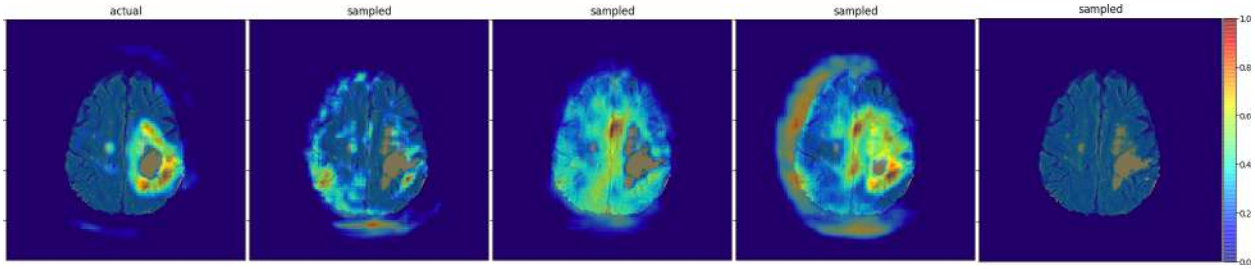


(c) APTOs Concept: C_2^{2a} Lateral Eye boundary

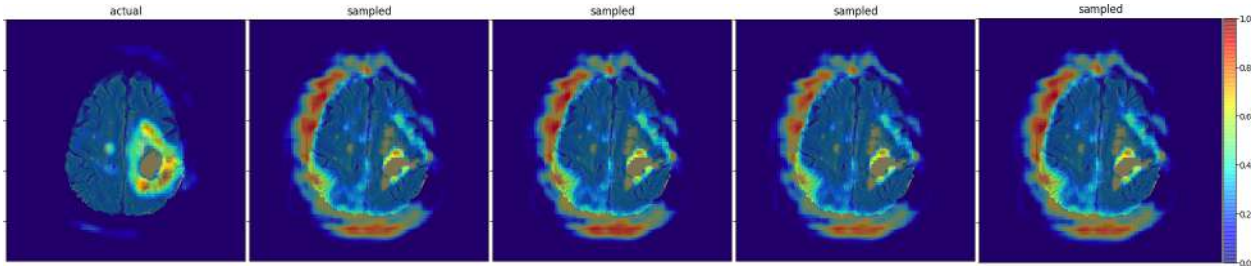


(d) APTOs Concept: C_4^{3d} Major Blood vessels

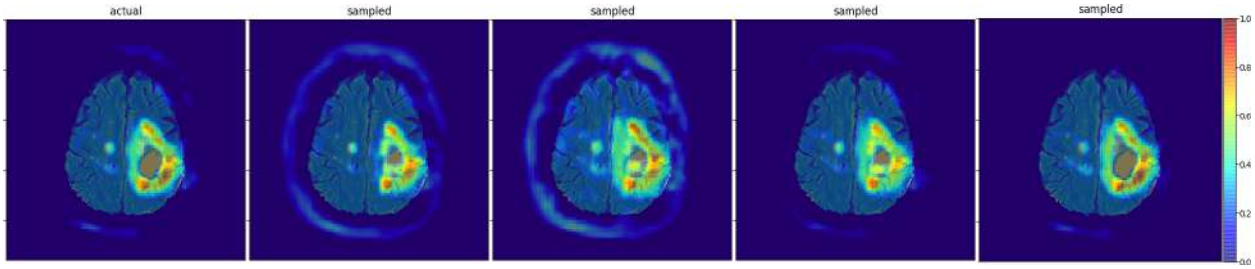
Figure 8: Above figure shows the consistency in concept formation; each row indicates different concepts over different input data point (along each column)



(a) Layer: 19 Gaussian Prior (control)

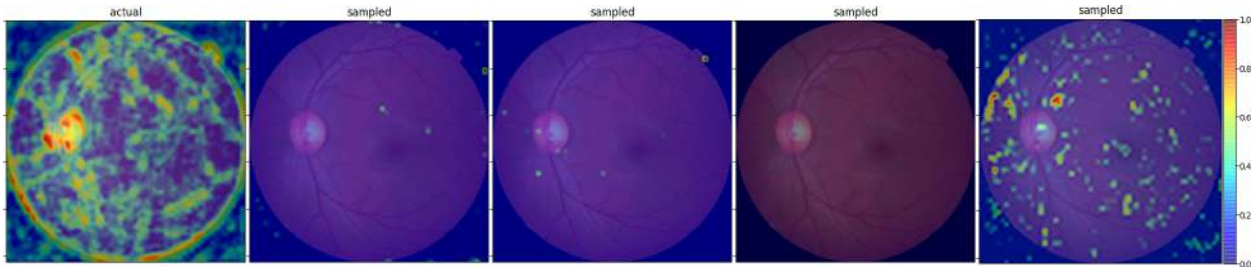


(b) Layer: 19 Uniform Prior

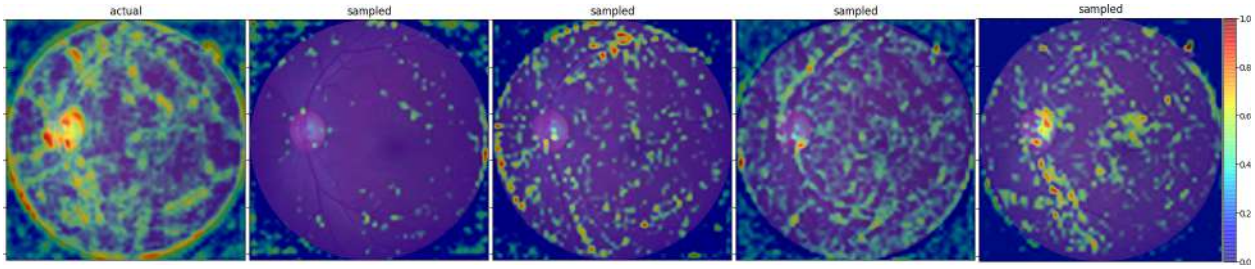


(c) Concept: C_2^{19} Gaussian Prior

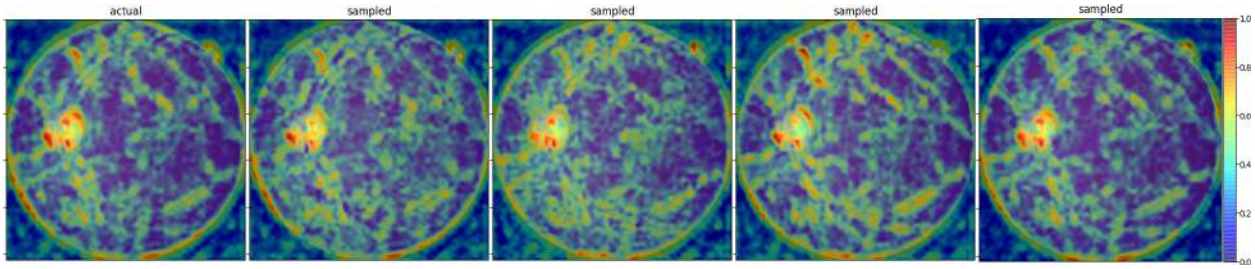
Figure 9: This figure illustrates results of robustness experiments on BraTs data, (a) Concept attention maps by assuming Gaussian distribution over all the weights in a layer, (b) Concept attention maps by assuming Uniform distribution over all the weights in a layer, and (c) Concept attention maps by assuming Gaussian distribution only over the cluster weights



(a) Layer: $3d$ Gaussian Prior (control)



(b) Layer: $3d$ Uniform Prior



(c) Concept: C_4^{3d} Gaussian Prior

Figure 10: This figure illustrates results of robustness experiments on APTOS data, (a) Concept attention maps by assuming Gaussian distribution over all the weights in a layer, (b) Concept attention maps by assuming Uniform distribution over all the weights in a layer, and (c) Concept attention maps by assuming Gaussian distribution only over the cluster weights

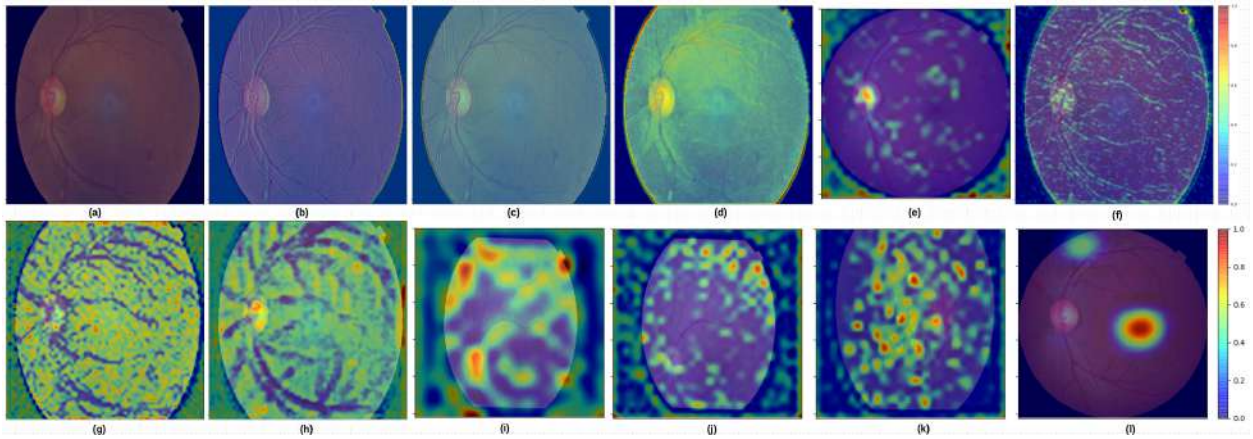
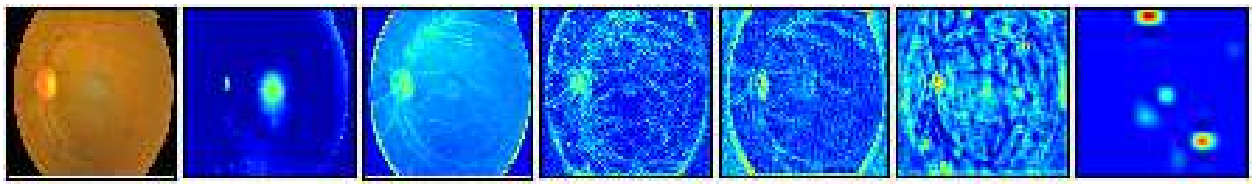
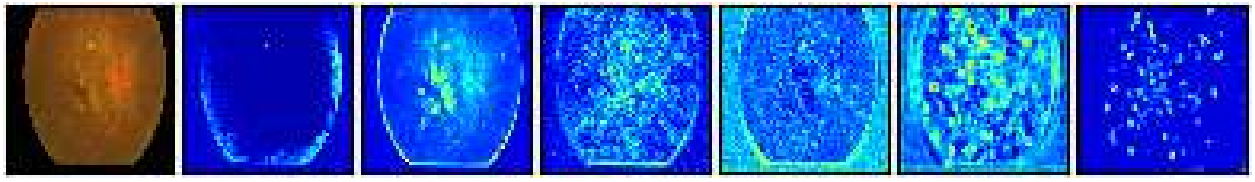


Figure 11: This figure illustrates the concepts obtained from various layers of a trained ResNet50 model. Based on the region of activation we provide description of the concepts as follows: (a) C_1^1 : doesn't capture any input region, (b) C_2^1 : Local feature; right lateral edges, (c) C_1^{2a}: Local feature; left lateral edges, (d) C_2^{2a}: Local feature; lateral edges, (e) C_2^{2c}: Global feature; optic disk, (f) C_2^{3a}: All blood vessels (major and tiny), (g) C_4^{3d}: Major blood vessels, (h) C_5^{3d}: Blood vessels (eroded), (i) C_2^{4a}: White spots (may be exodates) and optic disk, (j) C_1^{4f}: White spots (may be exodates), (k) C_3^{4a}: Hard exodates, (l) C_2^{5c}: Dark spots



(a) Trail for Mild image classification



(b) Trail for Proliferative image classification

Figure 12: Above figure depicts visual trails followed by network, (a) shows the steps followed by the network in classifying image as Mild, trail description is: *(Input Image to a network) -> (Local feature: left lateral edges) -> (Hard exodates in proliferative, and optic disk in Normal) -> (All blood vessels) -> (Optic disk and blood vessels) -> (Inverted Blood vessel (eroded) Image) -> (Dark spots) -> (Mild)*, (b) shows the steps followed to classify image as Proliferative: *(Input Image to a network) -> (Local feature: left lateral edges) -> (Hard exodates in proliferative, and optic disk in Normal) -> (All blood vessels) -> (Optic disk and blood vessels) -> (Inverted Blood vessel (eroded) Image) -> (White spots (may be exodates)) -> (Proliferative)*