Interpreting Deep Neural Networks for Medical Imaging using Concept Graphs

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

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

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

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

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 help us to unravel 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, and find active inference trails in the model. We work with radiologists and ophthalmologists to understand the obtained inference trails from a medical perspective and show that medically relevant concept trails are obtained which highlight the hierarchy of the decision-making process followed by the model. Our framework is available at `https://github.com/koriavinash1/BioExp`.

Introduction

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 (Holzinger et al. 2017a). Clinicians also prefer upfront information about the global properties of a model, such as its known strengths and limitations (Cai et al. 2019).

For this, semantic concepts internal to the model and their relationships need to be identified and represented in a human-understandable form. Previous interpretability techniques are example based or attention based (Molnar 2020), such as attribution, saliency, or feature visualization, and do not reflect the 'concept-based thinking' that human-reasoning shows (Armstrong, Gleitman, and Gleitman 1983), neither do they allow us to uncover the model's understanding of the relationship between such concepts. In the related work section we detail where our method stands in relation to current work in this area.

Graphical models provide a tractable way to depict concepts and the relationships between these concepts. However, there is a clear tug-of-war between model performance and transparency in this context (Holzinger et al. 2017a). Consider, for example, that we build a simple Bayesian Model for predicting the severity of Diabetic Retinopathy, where each node in the Bayesian Model is a human-understandable concept, such as microanuerisms, dark spots, exudates, and hemorrhages. Assuming we learn the structure and parameters of such a model, we would have a completely transparent technique for our task. However, it is difficult and computationally taxing to achieve the same level of performance with a Bayesian model as a deep neural network. This also requires an explicit differentiation or concept-level labelling of all relevant concepts expected to be in the Bayesian model, which is generally unavailable.

While Deep Neural Networks provide a much more efficient way to represent and learn from image data, they do not lend themselves to the simple conceptual analysis that graphical models like Bayesian Networks do. We propose a method to repurpose 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 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 (Goodfellow, Bengio, and Courville 2016).

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 more disentangled representations (Bengio 2013). For example, experiments in Network Dissection show that individual filters learn disentangled visual concepts (Bau et al. 2017). This behaviour has also been shown in the context of brain tumor segmentation models (Natekar, Kori, and Krishnamurthi 2020). Grouping filters which detect the same concept within a layer would then enable us to build a graphical representation of such 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 re-

lies heavily on one concept for one class of predictions. It also allows us to determine active inference 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, and, (ii) Identification of inference trails from this graphical representation that help us understand the model's decision-making logic.

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. Such a method ensures that our explanations are independent of the input sample and that our formulations are computationally practical. Previous experiments show that for state-of-the-art DNNs trained on large-scale datasets like ImageNet (Deng et al. 2009), euclidian distance in the activation space of final layers is an effective perceptual similarity metric (Zhang et al. 2018). 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 (Selvaraju et al. 2017). 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 Grad-CAM attention maps, we can find what the weight cluster corresponds to in the input space. Potential active inference 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 identification, (iii) concept significance analysis, (iv) graph formation, and (v) trail estimation. Figure 1 and 2 provide a detailed overview of the described framework. Next, we go over each section of this framework in detail.

Concept Formation

We posit that groups of weight vectors in a layer are responsible for detecting a particular concept in the input image. Weight clustering has been used before in the context of network compression (Han, Mao, and Dally 2015; Son, Nah, and Mu Lee 2018). We show that a clustering based approach can be used to identify weights which 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.

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 $W = \{w_1, w_2, ..., w_n\}$ be the set of weights in a layer, 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 (Kori, Krishnamurthi, and Srinivasan 2018; Palop, Mucke, and Roberson 2010) along with weights.

Clusters are formed using a hierarchical clustering method (Johnson 1967) 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, for visual verification we view the flattened weight vector to observe similarity among the clustered weights. Since direct visual interpretation is insufficient, to quantify the effectiveness of our clustering method we use $\mathbb{E}(SilhouetteScore)$ over all weights (Rousseeuw 1987) as a metric. Figure 11 in the Appendix depicts this for a sample layer.

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 concept 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 a (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. 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, providing a single-channel condensation of the cluster which can be used for finding the concept-attention map. 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, as formulated in equation 1.

Concept identification then amounts to finding the concept 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.

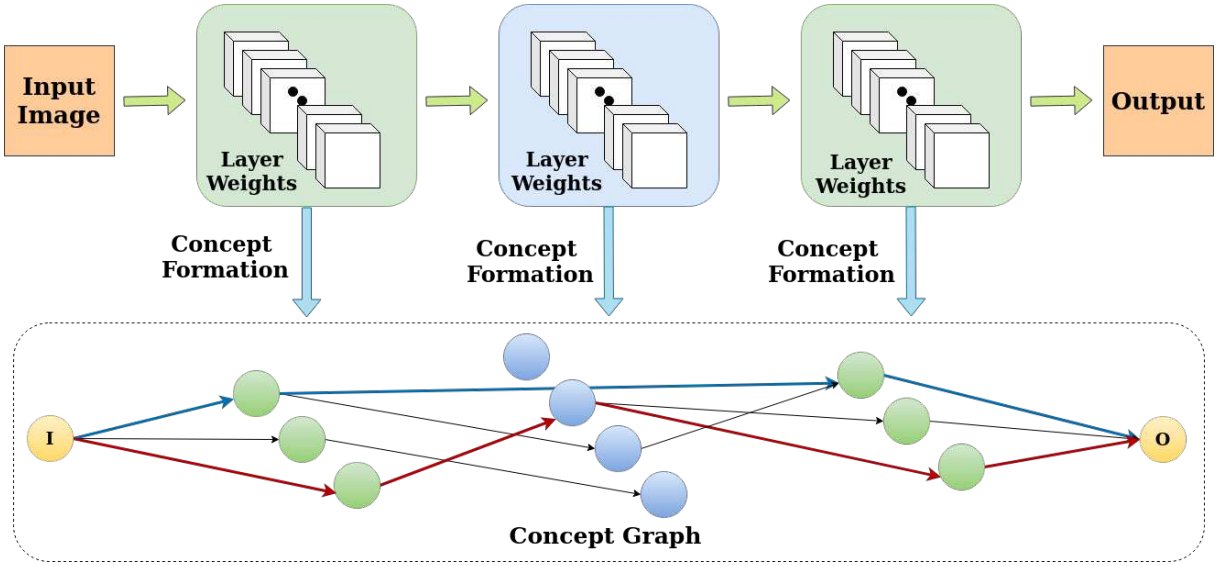


Figure 1: In the proposed framework, we construct a concept graph for a trained deep model. To generate concept graphs, we cluster weights in user-defined layers of the network, use them as concepts, and later estimate links based on a mutual information based metric. For example, trails represented in red and blue show active concept-level inference trails a network uses to predict the final result

Figure 2: The above figure describes all the steps in the proposed concept-based interpretability framework visualized in Figure 1

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

$$\beta_{m,p}^l(x) = \frac{1}{Z}\sum_i\sum_j \frac{\partial y_p^l(x)}{\partial \Phi_{l-1,m}(x)} \tag{2}$$

$$CAM_p^l = ReLU\left(\sum_m \beta_{m,p}^l(x)\Phi_{l-1,m}(x)\right) \tag{3}$$

Where, m is the index of a filter in layer $l-1$ and k is index of filter in layer l, β are the Grad-CAM importance weights, i, j are the indices for the height and width dimensions of the feature map of the additional convolutional layer, and CAM is the output concept-attention map for concept p of layer l.

Once the concepts are identified, we conduct significance tests to ensure that the concepts formed are consistent, robust, and localized. These procedures are detailed next. Figures 8, 9, and 10 show the results of the conducted consistency and robustness tests for our identified concepts, which provide further evidence to support our hypothesis that groups of weight vectors in the model are responsible for detecting different semantic concepts.

Consistency: To evaluate the consistency of clusters generated by the proposed method, we examine their regularity over multiple input samples in our datasets. Figure 10 illustrates the same, where each row corresponds to the concept attention map for an identified cluster over different images in the input dataset. It can be observed that identified clusters have similar concept attention maps for multiple input samples, irrespective of tumor location or optic disk location.

Robustness: Here, we try to evaluate the robustness of the formed clusters. Weights belonging to a specific layer in a neural network can be considered as i.i.d (Giryes, Sapiro, and Bronstein 2016). 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. Figure 4 depicts this graphically.

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,

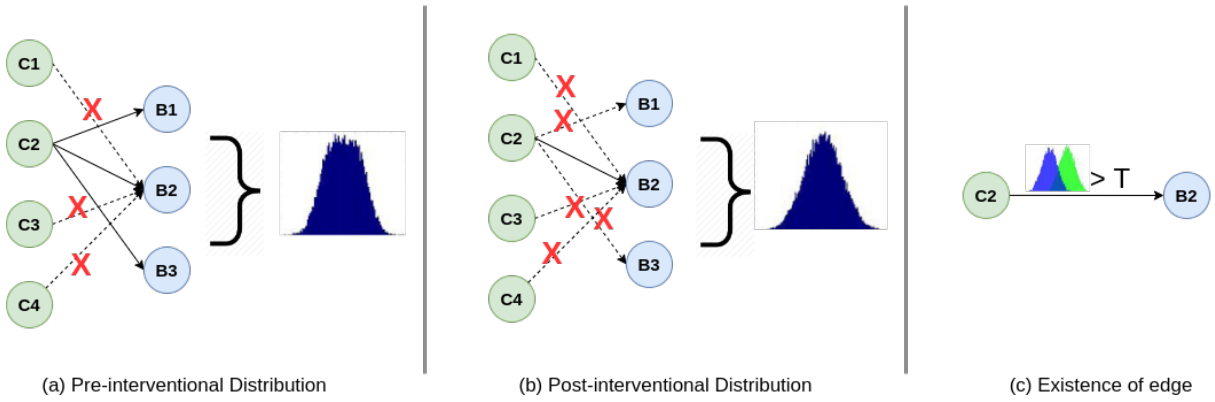


Figure 3: Above image describes the process of link formation. Sub-figure (a) describes how pre-interventional distribution is formed, sub-figure (b) describes how post-interventional distribution is formed, (c) exibits the condition for the existence of edge.

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

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

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. Figures 8 and 9 show the results of this experiment.

We observe that recomputed concept attention maps correspond to the same region in the input space as the original concept attention maps. We also generate recomputed concept attention maps using a uniform prior over the cluster weights as well as a gaussian prior taken over the range of all weights in the layer, and compare this with the results of using a gaussian prior over only the cluster weights. It can be observed that concept attention maps (CAMs) formed by using gaussian priors over only the weights belonging to that particular cluster are visually similar to the originals for each sampling run, while CAMs formed using uniform priors or CAMs formed using gaussian priors over all the weights do not encode the same concept in the input space and show high variability for each sampling run. This behaviour is seen consistently over all input samples. Thus, we empirically justify that our identified concepts come from the same underlying distribution, and that the gaussian is a reasonable proxy for this distribution.

Network Formation and Information Flow

Once concepts and have been identified for the given set of layers, we have the means to construct our equivalent graphical representation.

Given these concepts, we can identify relationships between them to generate a human-understandable trace of inference which augments model predictions. In order to identify the relationship between two concepts, we compute the normalized mutual information between the pre-interventional and post-interventional feature map distribution, as described below.

For the directed link between two concepts in layer p and q, $C_i^p \rightarrow C_j^q$, the pre-interventional distribution $\mathbb{P}(\Phi_j(x \mid do(C_{-i}^p = 0)))$ 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., $do(C_{-i}^p = 0)$, where the do operator indicates a manual intervention on the argument to set it to a particular value, which is 0 in this case). This distribution tells us about information flowing from C_i^p to all concepts in the succeeding layer q. Similarly, the post-interventional distribution $\mathbb{Q}(\Phi_j(x \mid do(C_{-i}^p = 0), do(C_{-j}^q = 0)))$ 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 all the clusters other than j in layer q (i.e., $do(C_{-i}^p = 0)$ and $do(C_{-j}^q = 0)$). This distribution tells us about the information flowing only from C_i^p to C_j^q. In this formulation the terms *pre* and *post* interventional are considered only with respect to layer q. Figure 3 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.

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

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 concept C_i^p to layer q occurs through that specific link $C_i^p \rightarrow C_j^q$. This results in the formation of a concept graph, an example visualization of which is shown in Figure 4. Note that this graphical model is not intended to be complete, only representative. 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.

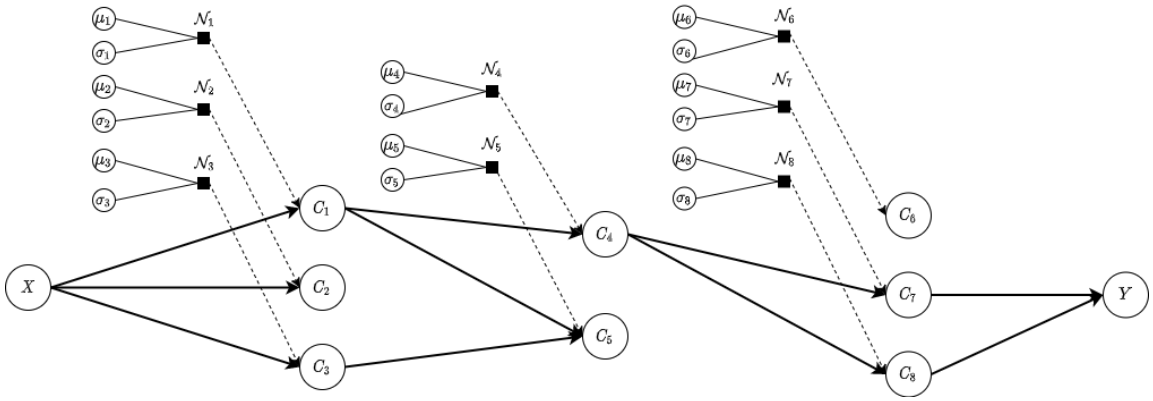


Figure 4: A visual depiction of the constructed graphical representation for the network given the set of layers to analyse. Each pixel in a concept can be imagined to be drawn from its own gaussian distribution, using the mean and variance of the pixel over the cluster as parameters. Dotted arrows show the concept is sampled from its corresponding normal distribution. Dark arrows show links between concepts.

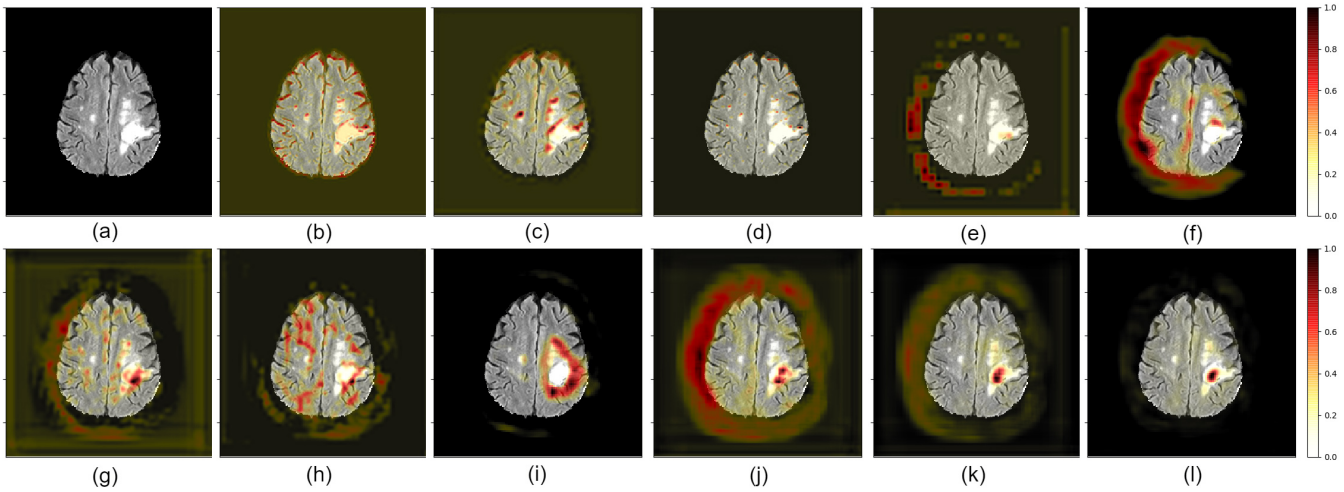


Figure 5: Concepts obtained from various layers of a trained U-net model superposed over the MRI Flair channel. (a) C_0^3: doesn't capture any input region, (b) C_1^3: concave edges, (c) C_2^3: linear edges, (d) C_2^5: interior key points. (e) C_0^{13}: Lateral left hemispherical brain boundary, (f) 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 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 a DAG 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 a prediction.

For example, consider the sample trail $X \rightarrow C_1 \rightarrow C_4 \rightarrow C_8 \rightarrow Y$ in Figure 4. Medical professionals can then highlight whether or not such an inference trail makes sense from a biomedical perspective, and understand the model's biases and its common logical steps of inference. The next section details the application of the above framework on benchmark biomedical image datasets.

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`.

Brain Tumor Segmentation

In the past decade, there has been significant development of image processing algorithms for segmenting intra-tumoral structures in brain MRI images (Bakas et al. 2018). Deep Learning has shown great potential in this context, with the BraTS challenge (Kamnitsas et al. 2017; Wang et al. 2017; Myronenko 2018; Kori et al. 2018) 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.

We implement our algorithm on a UNet based model for brain tumor segmentation, which is a popular segmentation architecture in the medical context (Ronneberger, Fischer, and Brox 2015). Our model also has residual connections as per (Kermi, Mahmoudi, and Khadir 2018), and achieves a dice score of 0.788, 0.743, and 0.649 on whole tumor, tumor core and enhancing tumor segmentation respectively on a held-out validation set of 48 volumes. Our model is not meant to achieve state of the art performance. Instead, we aim to demonstrate our method on a commonly used architecture for brain-tumor segmentation. The next sections detail the concepts and active inference trails obtained as a result of our framework on this task.

Concepts The $\mathbb{E}(SilhouetteScore)$ over all the datapoints is 0.241, indicating the formation of weak but significant clusters. Figure 5 describes the various concepts identified from our model. Initial layers (convolutional layers 3 and 5) correspond to edges in a specific direction or brain boundaries. In higher layers, filters start capturing more local information. It can be observed that some concepts capture brain boundary, while some capture tumor boundary. Figure 5 contains a description of the various concepts obtained from out network. This behaviour is in line with the understanding that filters in shallower layers of brain tumor segmentation models learn simple patterns while deeper layers learn progressively more complex concepts (Natekar, Kori, and Krishnamurthi 2020). The brain atlas described in (Ding et al. 2016) was used to formulate appropriate descriptions.

Trails and Discoveries Figure 6 describes inference trails involved in predicting the enhancing tumor region (Trails for other classes are available in the Appendix). These show the model's attention is initially on the outer edges and keypoints of the brain, then moves to the white and grey matter region, then the tumor boundary, and finally the internal tumor region. The caption of Figure 6 also provides a description of the visual trails for an image based on the predefined concept description. In the discussion section, we analyse these trails with feedback from a certified radiologist.

Diabetic Retinopathy classification

Diabetic Retinopathy (DR) is frequent in individuals suffering from diabetes (Fong et al. 2004). Deep Learning algorithms have shown great promise in detecting the severity of diabetic retinopathy 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 validation set of the APTOS dataset (Society 2019). The APTOS dataset contains around 5500 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, exudates, and hemorrhages. Thus, it becomes necessary to see whether deep learning models process and identify these features, and to see the model's understanding of relationships between these and the predicted severity of DR. We follow a similar process as that for brain tumor segmentation, detailed below.

Concepts The $\mathbb{E}(SilhouetteScore)$ over all the datapoints is 0.2, which again indicates the formation of weak but significant clusters. Figure 12 describes the identified local and global level concepts, encoding blood vessels, hard and soft exudates, dot-blot hemorrhages, etc.

Trails and Discoveries Similar to the trails obtained for the BraTS dataset, we show example inference trails obtained for the APTOS dataset in Figure 7 and Figures 13 and 14 in the Appendix. These describe visual trails involved in predicting 'Severe', 'Moderate', and 'Proliferative' classes of diabetic retinopathy respectively. An ophthalmologist's feedback was obtained on the concept trails, which is elaborated in the discussion section. Once again, we see the emergence of medically relevant concepts in a hierarchical manner, which may provide additional support to medical professionals apart from just the output classification.

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 (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Ustun and Rudin 2014). In contrast, ante-hoc methods try to build an explainable model while training itself (Caruana et al. 2015; Holzinger et al. 2017b;).

Current research directions in post-hoc interpretability focus mainly on visualizing network attributions or illustrative samples in the input space (Selvaraju et al. 2017; Bau et al. 2017; Olah, Mordvintsev, and Schubert 2017; Kim et al. 2018). 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 (Bau et al. 2017). Previous concept based interpretability methods (Ghorbani et al. 2019; Kim et al. 2018) 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 (Yeh et al. 2019). However, neither of these methods consider the relationship between concepts learnt by the model and do not provide a trace of inference steps. Also, these methods either require a pre-processed set of input samples as concepts (Kim et

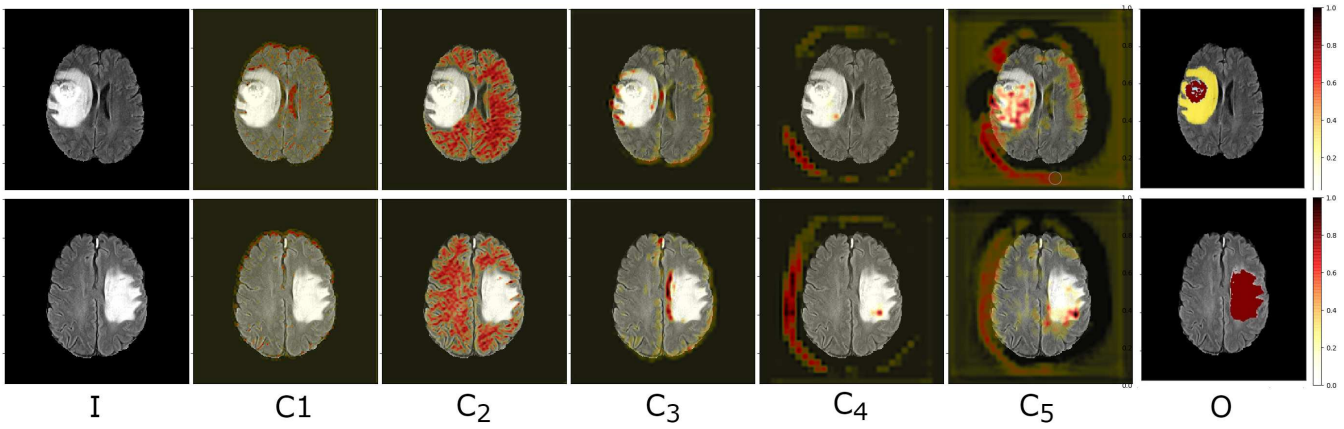


Figure 6: Active inference trail for enhancing tumor (Each row is a trail for one input sample, red regions are high attention): *(**I**: Input image to a network) $->$ (C_1: Concave edges) $->$ (C_2: White matter region) $->$ (C_3: Tumor boundary) $->$ C_4: (Lateral brain boundary) $->$ (C_5: Inferior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

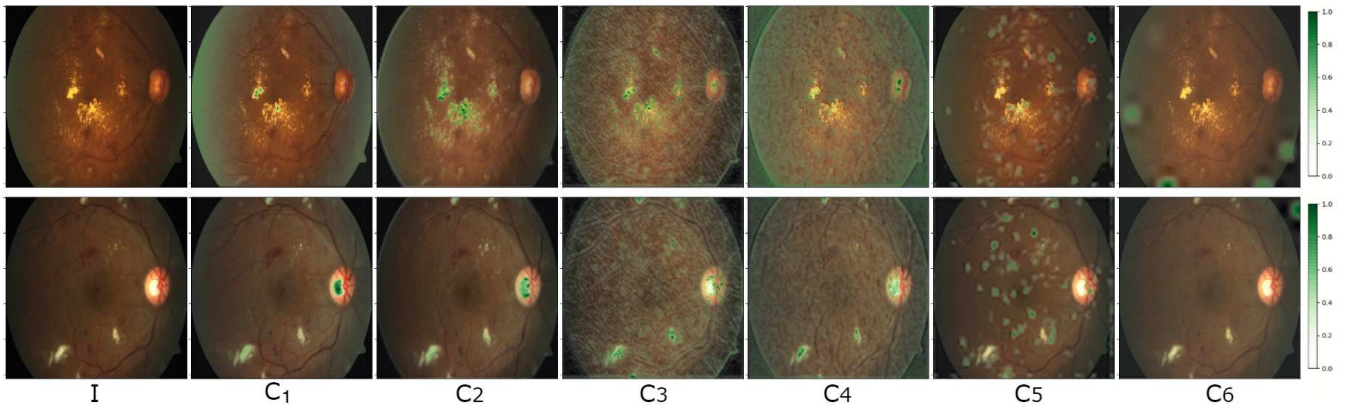


Figure 7: Active inference trail for severe DR (green regions are high attention): *(I: Input Image) $->$ (C_1: Optic Cup/Hard exudates) $->$ (C_2: Hard Exudates) $->$ (C_3: Blood vessels, soft exudates) $->$ (C_4: Blood vessel, soft exudates) $->$ (C_5: dot-blot Hemorrhages/laser scar marks of retinal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

al. 2018), or automatically segment the input image at various resolutions to create concepts (Ghorbani et al. 2019). However, in the medical domain, obtaining such concepts is difficult - manual concept curation is time consuming and would require medical experts, while segmenting the input image may not lead to the formation of coherent anatomical concepts which add interpretability value, especially in cases where the task itself is image segmentation. In such domains, interpretability needs to emerge organically from the model itself and provide an understanding of the model's decision making logic.

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.

Discussion

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 show that consistent, distinct trails that lead to a particular classification made up of anatomically relevant concepts can be identified.

While in previous work on interpretability in the medical domain (Natekar, Kori, and Krishnamurthi 2020), the existence of disentangled concepts is shown in brain-tumor segmentation networks, in this work we create a concept-level graph that depicts the relationships between these concepts and provides an understanding of inference trails in the model. As opposed to previous concept-based approaches (Ghorbani et al. 2019; Kim et al. 2018), no manual extraction of concepts from the input dataset is required, which

is a challenging task in the medical domain. In this initial work, we demonstrate the potential of our technique on two medical datasets - the BraTS dataset for brain tumor segmentation and the APTOS dataset for diabetic retinopathy classification.

For brain-tumor segmentation, a certified radiologist's comments on the extracted concept trail was solicited. They noted the lateral to medial and anterior to superior nature of attention of the model, as well as the hierarchical approach to segmentation which is in line with a radiologist's thought process. They commented that tumour boundary delineation as seen in Figure 6 concept C_3 has value for neurosurgeons when obtaining biopsy or resecting the tumour since this helps prevent damage to unaffected brain tissue. They also noted that a neuroradiologist would be able to immediately perceive the presence of gliomas in the flair sequence and it is in general not possible to break down that perception in terms of the trails obtained from the concept graphs. However, the visualization of concepts that are focused on tumour boundaries and the tumour core would help in improving confidence and trust in the deep learning model. The tumor core and characteristics are also defined which will aid in diagnosis and grading of the tumor.

For Diabetic Retinopathy, an ophthalmologist's feedback was obtained on the output trail described in Figure 7. Various features, such as hard and soft exudates, dot-blot haemorrhages, optic cup, and laser scar marks of retinal photocoagulation were identified. In the case of DR, it is interesting that features like this, which ophthalmologists look at to classify DR images, emerge implicitly from the model, even though it has not been explicitly trained to learn these.

Acknowledgments. We would like to acknowledge help from Dr. Ravikanth Balaji and Dr. Devika Joshi for providing clinician (radiological and opthalmological) feedback on the inference trails obtained.

References

Armstrong, S. L.; Gleitman, L. R.; and Gleitman, H. 1983. What some concepts might not be. *Cognition* 13(3):263–308.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. 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*.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

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

Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. " 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.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. 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*, 1721–1730.

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

Ding, S.-L.; Royall, J. J.; Sunkin, S. M.; Ng, L.; Facer, B. A.; Lesnar, P.; Guillozet-Bongaarts, A.; McMurray, B.; Szafer, A.; Dolbeare, T. A.; et al. 2016. Comprehensive cellular-resolution atlas of the adult human brain. *Journal of Comparative Neurology* 524(16):3127–3481.

Fong, D. S.; Aiello, L.; Gardner, T. W.; King, G. L.; Blankenship, G.; Cavallerano, J. D.; Ferris, F. L.; and Klein, R. 2004. Retinopathy in diabetes. *Diabetes care* 27(suppl 1):s84–s87.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, 9277–9286.

Giryes, R.; Sapiro, G.; and Bronstein, A. M. 2016. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing* 64(13):3444–3457.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Holzinger, A.; Plass, M.; Kickmeier-Rust, M.; Holzinger, K.; Crişan, G. C.; Pintea, C.-M.; and Palade, V. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence* 49(7):2401–2414.

Holzinger, A.; Biemann, C.; Pattichis, C. S.; and Kell, D. B. 2017a. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Holzinger, A.; Plass, M.; Holzinger, K.; Crisan, G. C.; Pintea, C.-M.; and Palade, V. 2017b. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. *arXiv preprint arXiv:1708.01104*.

Johnson, S. C. 1967. Hierarchical clustering schemes. *Psychometrika* 32(3):241–254.

Kamnitsas, K.; Ledig, C.; Newcombe, V. F.; Simpson, J. P.; Kane, A. D.; Menon, D. K.; Rueckert, D.; and Glocker, B. 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis* 36:61–78.

Kermi, A.; Mahmoudi, I.; and Khadir, M. T. 2018. Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri vol-

umes. In *International MICCAI Brainlesion Workshop*, 37–48. Springer.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677.

Kori, A.; Soni, M.; Pranjal, B.; Khened, M.; Alex, V.; and Krishnamurthi, G. 2018. Ensemble of fully convolutional neural network for brain tumor segmentation from magnetic resonance images. In *International MICCAI Brainlesion Workshop*, 485–496. Springer.

Kori, A.; Krishnamurthi, G.; and Srinivasan, B. 2018. Enhanced image classification with data augmentation using position coordinates. *arXiv preprint arXiv:1802.02183*.

Molnar, C. 2020. *Interpretable Machine Learning*. Lulu. com.

Myronenko, A. 2018. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, 311–320. Springer.

Natekar, P.; Kori, A.; and Krishnamurthi, G. 2020. Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. *Frontiers in Computational Neuroscience* 14:6.

Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature visualization. *Distill*. https://distill.pub/2017/feature-visualization.

Palop, J. J.; Mucke, L.; and Roberson, E. D. 2010. Quantifying biomarkers of cognitive dysfunction and neuronal network hyperexcitability in mouse models of alzheimer's disease: depletion of calcium-dependent proteins and inhibitory hippocampal remodeling. In *Alzheimer's Disease and Frontotemporal Dementia*. Springer. 245–262.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

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

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Society, A. P. T.-O. 2019. Asia pacific tele-ophthalmology society 2019, dataset.

Son, S.; Nah, S.; and Mu Lee, K. 2018. Clustering convolutional kernels to compress deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 216–232.

Ustun, B., and Rudin, C. 2014. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*.

Wang, G.; Li, W.; Ourselin, S.; and Vercauteren, T. 2017. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, 178–190. Springer.

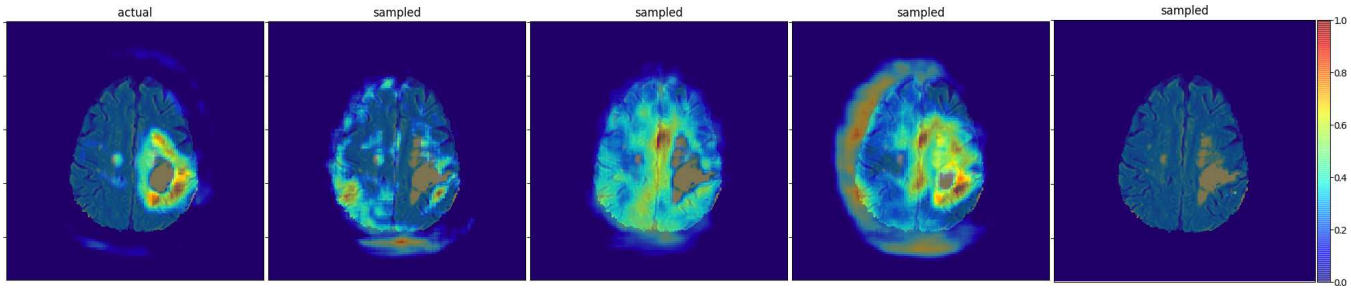
Yeh, C.-K.; Kim, B.; Arik, S. O.; Li, C.-L.; Ravikumar, P.; and Pfister, T. 2019. On concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

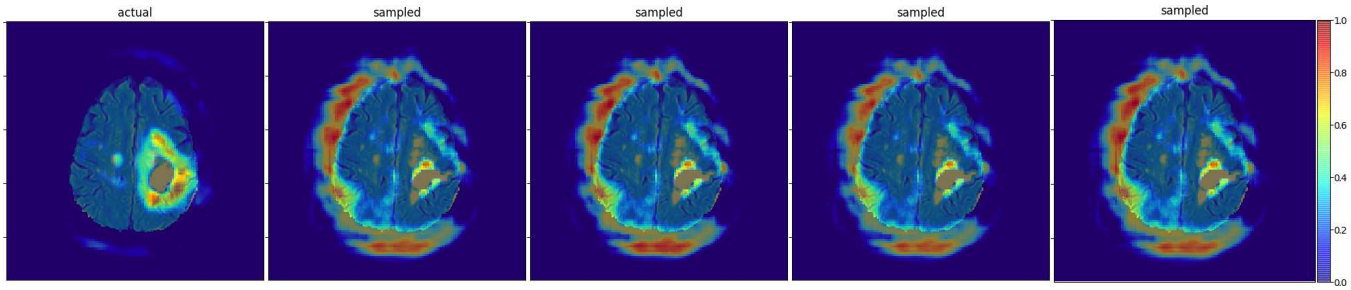
Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Appendix I

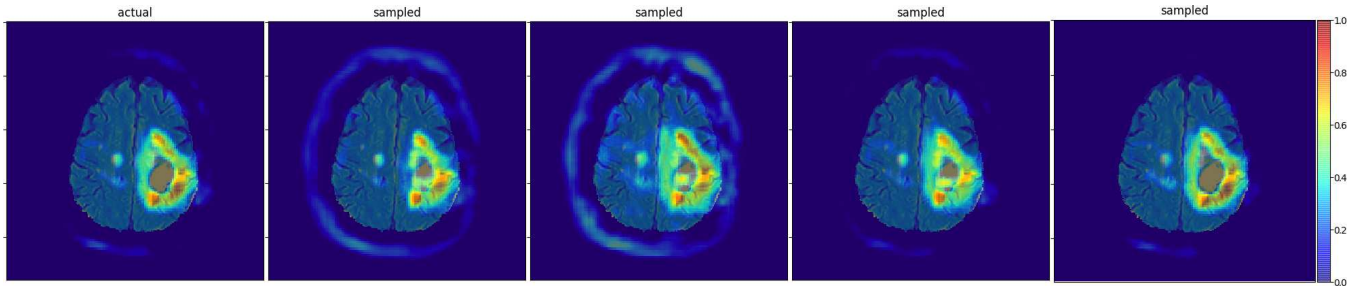
Here we show additional figures and examples which result from our primary analysis above. First, the results for cluster significance tests are shown - robustness and consistency. Then we show additional examples for brain-tumor segmentation and diabetic retinopathy classification, as well as other supporting images.



(a) Layer: 19, Gaussian Prior over entire weight layer

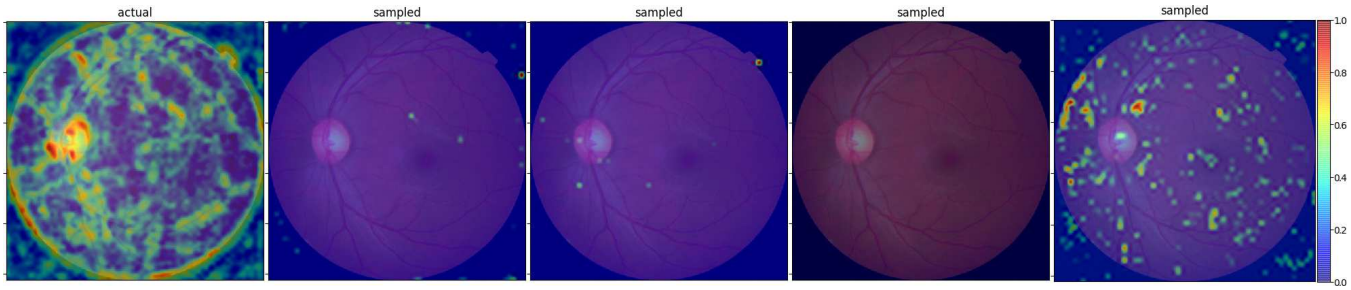


(b) Layer: 19, Uniform Prior over only the weight cluster

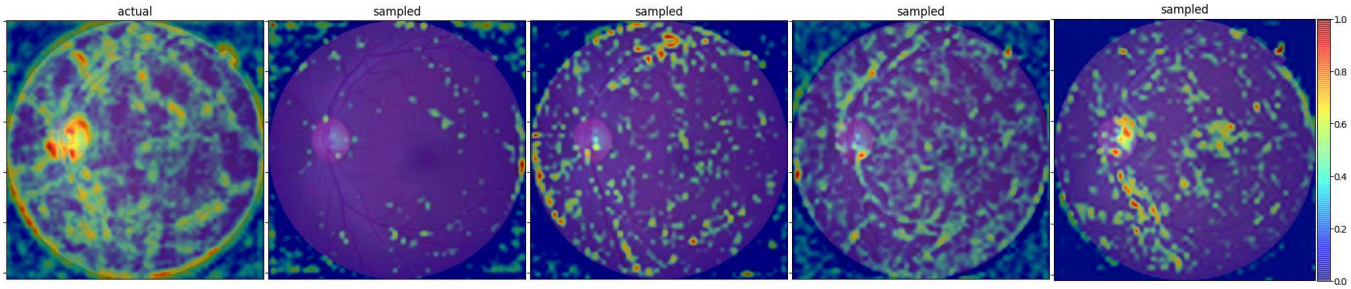


(c) Concept: C_2^{19}, Gaussian Prior over only the weight cluster

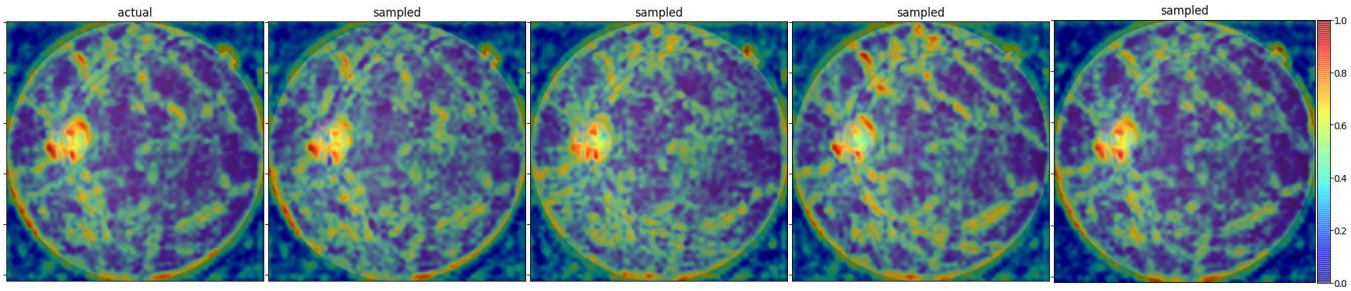
Figure 8: This figure illustrates results of robustness experiments on BraTs data, (a) Concept attention maps by assuming Gaussian distribution over all the weights in a layer, (b) Concept attention maps by assuming Uniform distribution over only the cluster weights, and (c) Concept attention maps by assuming Gaussian distribution over only the cluster weights. Note that using a gaussian prior over only the cluster gives most consistent concept attention maps.



(a) Layer: $3d$ Gaussian Prior over entire weight layer

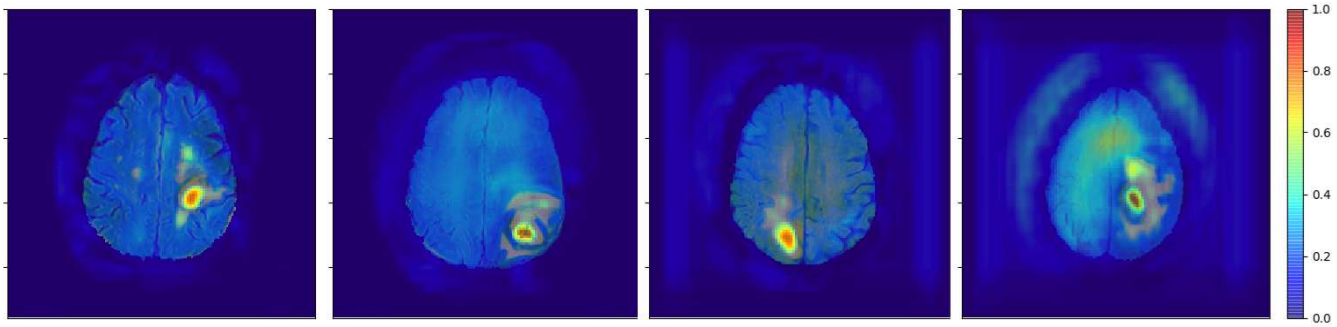


(b) Layer: $3d$ Uniform Prior over only the cluster weights

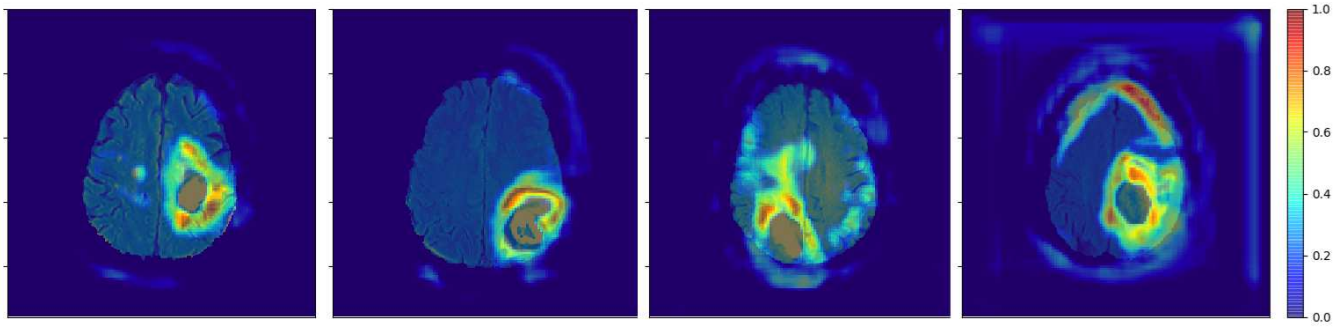


(c) Concept: C_4^{3d} Gaussian Prior over only the cluster weights

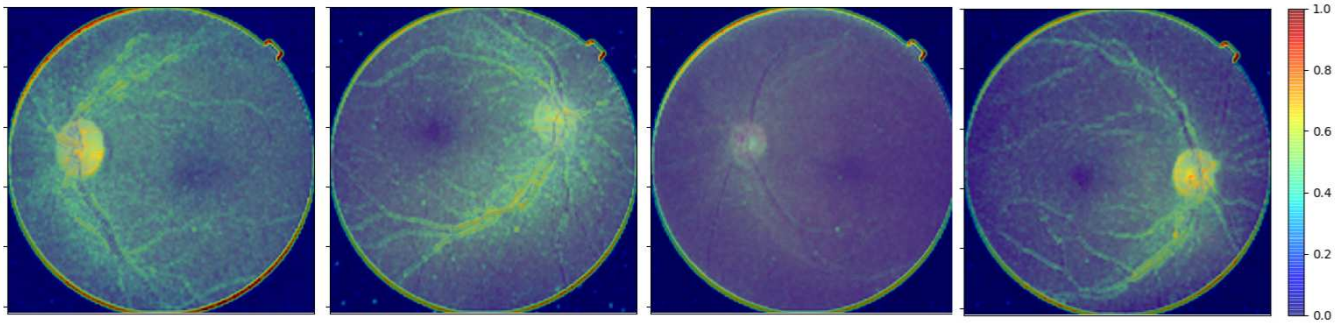
Figure 9: This figure illustrates results of robustness experiments on APTOS data, (a) Concept attention maps by assuming Gaussian distribution over all the weights in a layer, (b) Concept attention maps by assuming Uniform distribution over only the cluster weights, and (c) Concept attention maps by assuming Gaussian distribution over only the cluster weights. Note that using a gaussian prior over only the cluster gives most consistent concept attention maps.



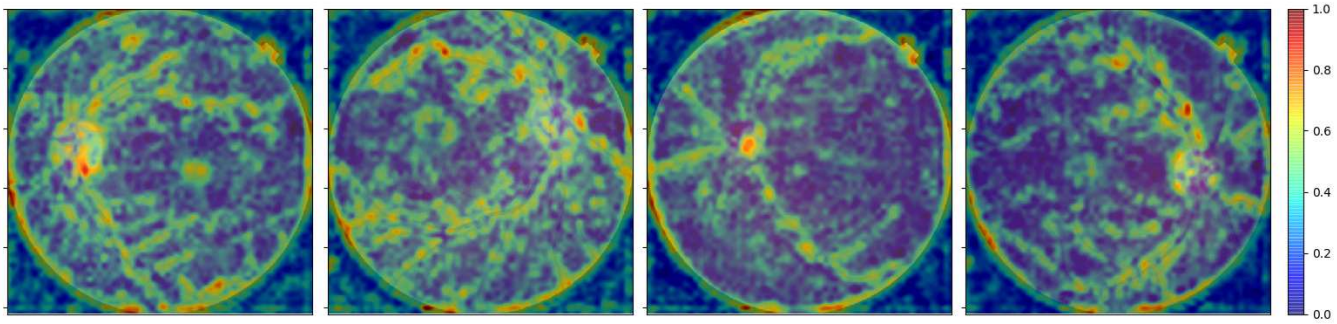
(a) BraTS Concept: C_2^{21} Tumor Core region



(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 10: The above figure shows the consistency of concept formation; each row indicates shows the concept-attention map for a cluster for different input samples

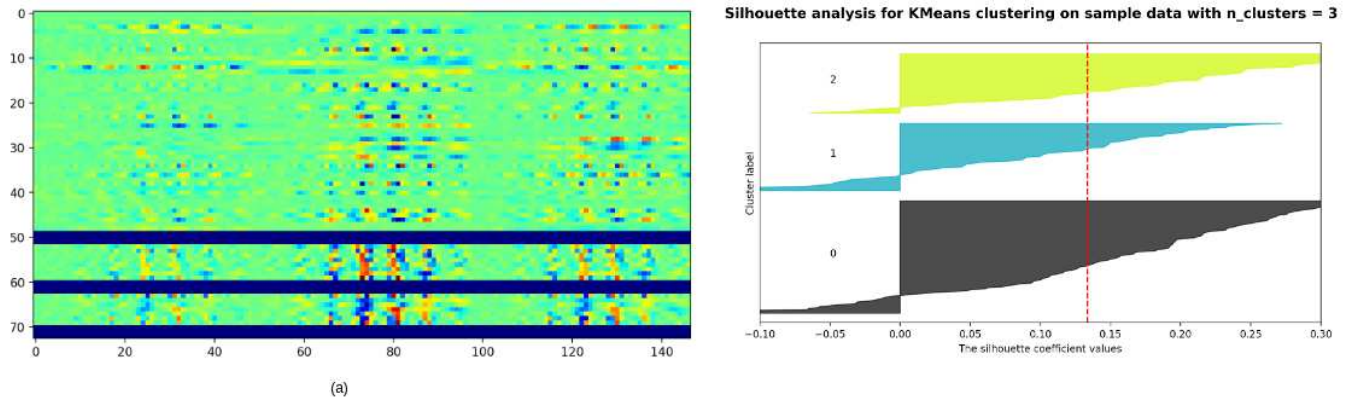


Figure 11: Above image describes the effectiveness of clustering. Sub-figure (a) describes the initial layer weights from ResNet50 trained on APTOS (Society 2019) data, in the figure dark blue horizontal bands seperates the weights among multiple clusters (provided figure has 3 clusters). Sub-figure (b) quantifies the effectiveness of clusters obtained as the result of proposed method using a silhouette plot

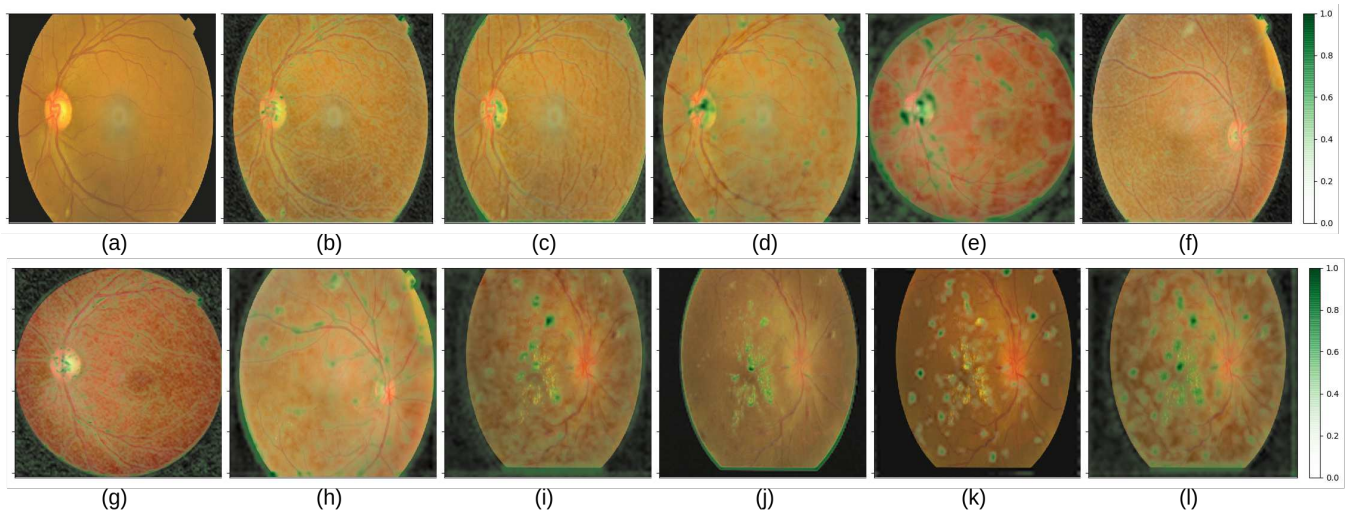


Figure 12: This figure illustrates the concepts obtained from various layers of a trained ResNet50 model. Based on the region of activation we provide description of the concepts as follows: (a) C_1^1 : doesn't capture any input region, (b) C_2^1 : Right lateral edges, (c) C_1^{2a}: Lateral edges, (d) C_2^{2a}: Optic disk + lateral edges, (e) C_2^{2c}: Optic disk + blood vessels, (f) C_2^{3a}: All blood vessels (tiny), (g) C_4^{3d}: Major blood vessels, (h) C_5^{3d}: Blood vessels (eroded), (i) C_2^{4a}: Yellow spots (may be hard exodates), (j) C_1^{4f}: Yellow spots (may be hard exodates), (k) C_3^{4a}: Pale Yellow (may be hard exodates), (l) C_2^{5c}: Hard/Soft exodates

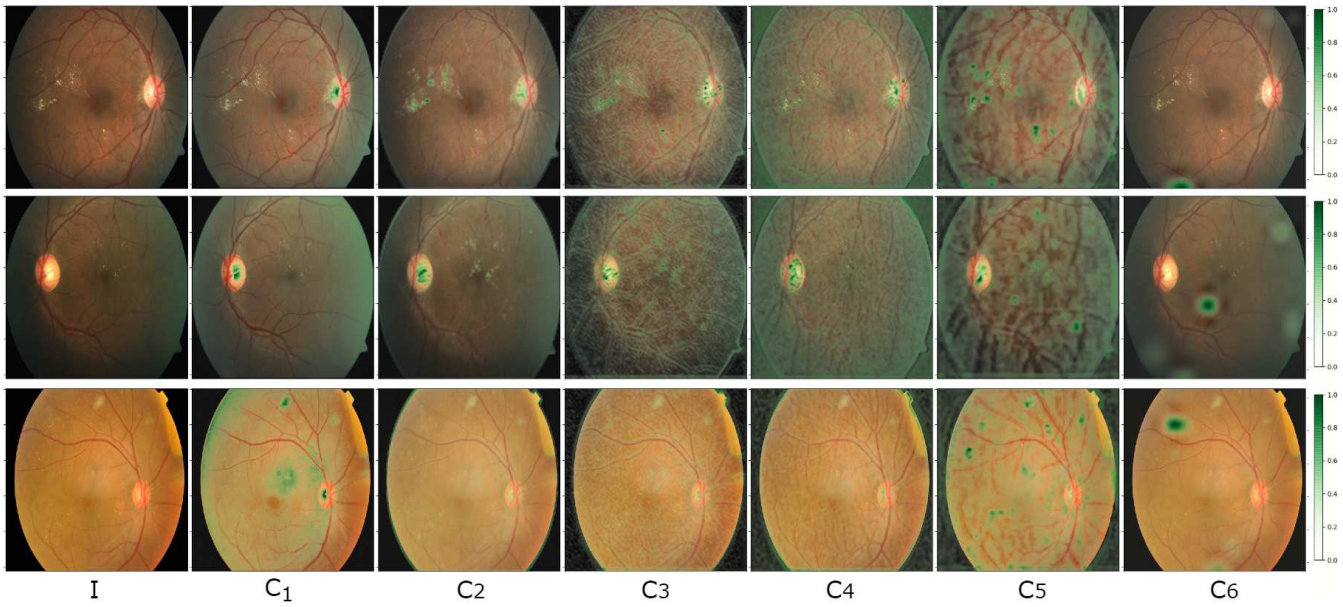


Figure 13: Active inference trail for Moderate DR (Green regions are high attention): *(I: Input Image to a network)* $->$ *(C_1: Soft exudates + Optic Cup)* $->$ *(C_2: Hard exudates)* $->$ *(C_3: All blood vessels)* $->$ *(C_4: Optic disk and blood vessels)* $->$ *(C_5: Inverted Blood vessel (eroded) Image)* $->$ *(C_6: Dark spots)*

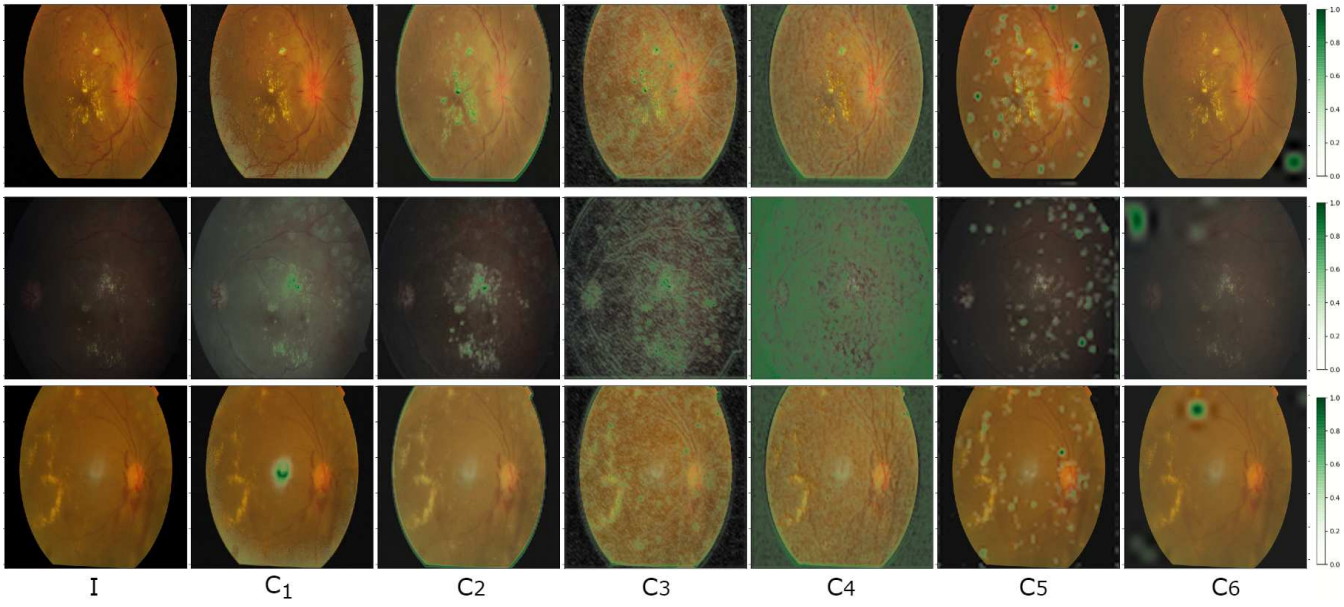


Figure 14: Active inference trail for Proliferative DR (Green regions are high attention): *(I: Input Image to a network)* $->$ *(C_1: Pale areas, due to attenuated artery endings + macula)* $->$ *(C_2: Hard exudates)* $->$ *(C_3: All blood vessels + key points)* $->$ *(C_4: Optic disk and blood vessels)* $->$ *(C_5: Laser scar marks of retinal photocoagulation + blot haemorrhages* $->$ *(C_6: Dark spots)*

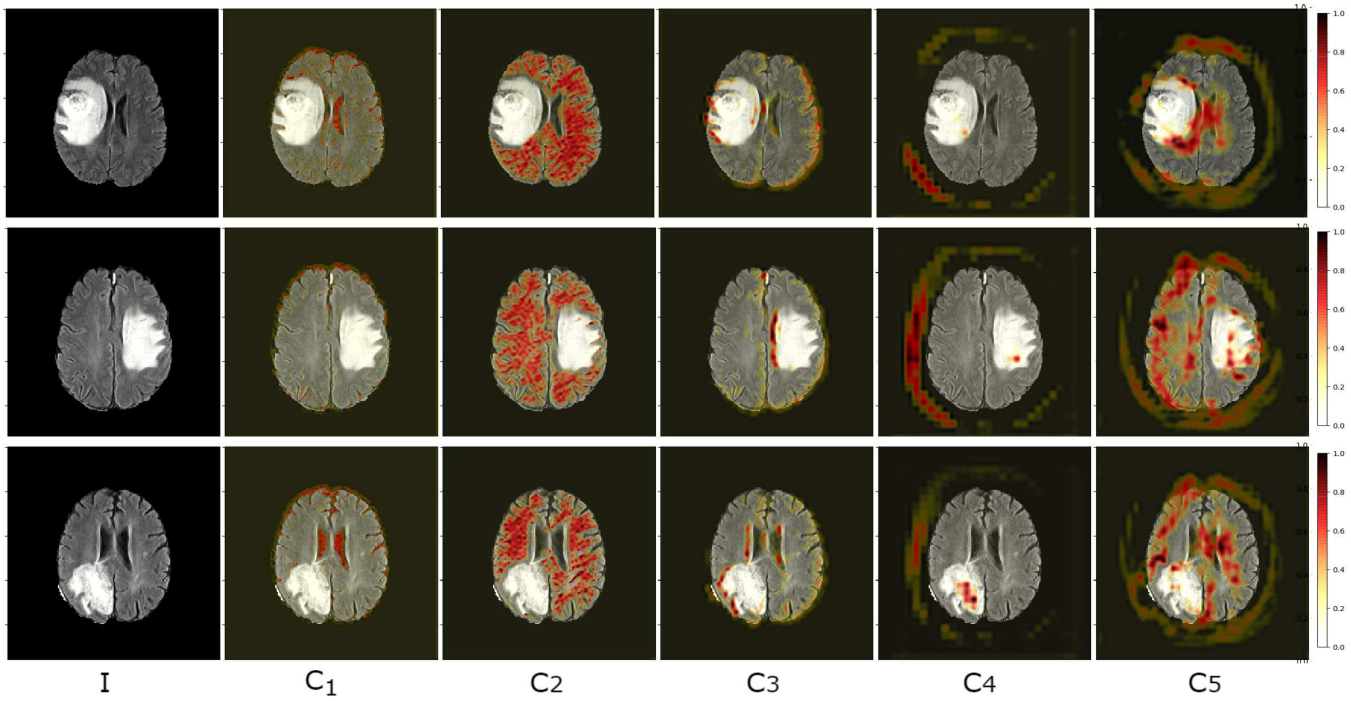


Figure 15: Active inference trail for Edema (Each row is a trail for one input sample, red regions are high attention): *(**I**: Input Image to a network) $->$ (C_1: Concave edges) $->$ (C_2: White matter) $->$ (C_3: Brain and tumor boundary) $->$ C_4: (Lateral brain boundary) $->$ (C_5: Lateral tumor boundary and mid brain) $->$ (Edema region)*

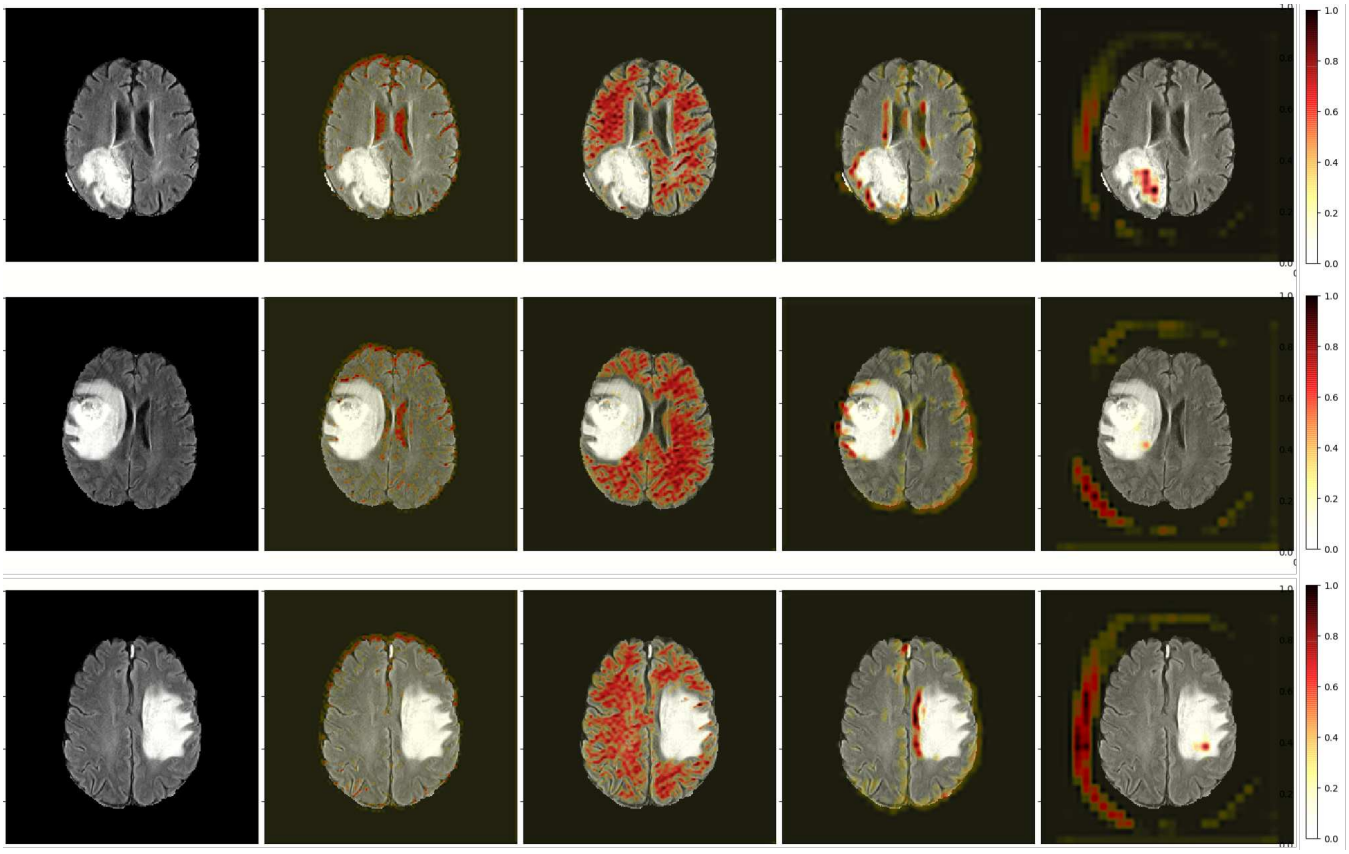


Figure 16: Active inference trail for Tumor Core (Each row is a trail for one input sample, red regions are high attention): ***(I**: Input Image to a network)* $->$ *(C_1: Concave edges)* $->$ *(C_2: White matter)* $->$ *(C_3: Brain and tumor boundary)* $->$ *C_4: Tumor Core)*