frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fncom.2020.00006

Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis

*Parth Natekar, Avinash Kori and Ganapathy Krishnamurthi**

Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India

The accurate automatic segmentation of gliomas and its intra-tumoral structures is important not only for treatment planning but also for follow-up evaluations. Several methods based on 2D and 3D Deep Neural Networks (DNN) have been developed to segment brain tumors and to classify different categories of tumors from different MRI modalities. However, these networks are often black-box models and do not provide any evidence regarding the process they take to perform this task. Increasing transparency and interpretability of such deep learning techniques is necessary for the complete integration of such methods into medical practice. In this paper, we explore various techniques to explain the functional organization of brain tumor segmentation models and to extract visualizations of internal concepts to understand how these networks achieve highly accurate tumor segmentations. We use the BraTS 2018 dataset to train three different networks with standard architectures and outline similarities and differences in the process that these networks take to segment brain tumors. We show that brain tumor segmentation networks learn certain human-understandable disentangled concepts on a filter level. We also show that they take a top-down or hierarchical approach to localizing the different parts of the tumor. We then extract visualizations of some internal feature maps and also provide a measure of uncertainty with regards to the outputs of the models to give additional qualitative evidence about the predictions of these networks. We believe that the emergence of such human-understandable organization and concepts might aid in the acceptance and integration of such methods in medical diagnosis.

Keywords: interpretability, CNN, brain tumor, segmentation, uncertainty, activation maps, features, explainability

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
Fan Zhang,
Harvard Medical School,
United States
Hongming Li,
University of Pennsylvania,
United States

***Correspondence:**
Ganapathy Krishnamurthi
gankrish@iitm.ac.in

Received: *07 September 2019*
Accepted: *17 January 2020*
Published: *07 February 2020*

Citation:
Natekar P, Kori A and Krishnamurthi G (2020) Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis. Front. Comput. Neurosci. 14:6. doi: 10.3389/fncom.2020.00006

1. INTRODUCTION

Deep learning algorithms have shown great practical success in various tasks involving image, text and speech data. As deep learning techniques start making autonomous decisions in areas like medicine and public policy, there is a need to explain the decisions of these models so that we can understand *why* a particular decision was made (Molnar, 2018).

In the field of medical imaging and diagnosis, deep learning has achieved human-like results on many problems (Esteva et al., 2017; Weng et al., 2017; Kermany et al., 2018). Interpreting the decisions of such models in the medical domain is especially important, where transparency and a clearer understanding of Artificial Intelligence are essential from a regulatory point of view and to make sure that medical professionals can trust the predictions of such algorithms.

Understanding the organization and knowledge extraction process of deep learning models is thus important. Deep neural networks often work in higher dimensional abstract concepts. Reducing these to a domain that human experts can understand is necessary—if a model represents the underlying data distribution in a manner that human beings can comprehend and a logical hierarchy of steps is observed, this would provide some backing for its predictions and would aid in its acceptance by medical professionals.

However, while there has been a wide range of research on Explainable AI in general (Doshi-Velez and Kim, 2017; Gilpin et al., 2018), it has not been properly explored in the context of deep learning for medical imaging. Holzinger et al. (2017) discuss the importance of interpretability in the medical domain and provide an overview of some of the techniques that could be used for explaining models which use the image, omics, and text data.

In this work, we attempt to extract explanations for models which accurately segment brain tumors, so that some evidence can be provided regarding the process they take and how they organize themselves internally. We first discuss what interpretability means with respect to brain tumor models. We then present the results of our experiments and discuss what these could imply for machine learning assisted tumor diagnosis.

2. INTERPRETABILITY IN THE CONTEXT OF BRAIN TUMOR SEGMENTATION MODELS

Interpreting deep networks which accurately segment brain tumors is important from the perspectives of both transparency and functional understanding (by functional understanding, we mean understanding the role of each component or filter of the network and how these relate to each other). Providing glimpses into the internals of such a network to provide a *trace of its inference steps* (Holzinger et al., 2017) would go at least some way to elucidating exactly how the network makes its decisions, providing a measure of legitimacy.

There have been several methods explored for trying to look inside a deep neural network. Many of these focus on visual interpretability, i.e., trying to extract understandable visualizations from the inner layers of the network or understanding what the network looks at when giving a particular output (Zhang and Zhu, 2018).

For a brain tumor segmentation model, such methods might provide details on how information flows through the model and how the model is organized. For example, it might help in understanding how the model represents information regarding the brain and tumor regions internally, and how these representations change over layers. Meaningful visualizations of the internals of a network will not only help medical professionals in assessing the legitimacy of the predictions but also help deep learning researchers to debug and improve performance.

In this paper, we aim to apply visual interpretability and uncertainty estimation techniques on a set of models with different architectures to provide human-understandable visual interpretations of some of the concepts learned by different parts of a network and to understand more about the organization of these different networks. We organize our paper into mainly three parts as described in **Figure 1**: (1) Understanding information organization in the model, (2) Extracting visual representations of internal concepts, and (3) Quantifying uncertainty in the outputs of the model. We implement our pipeline on three different 2D brain tumor segmentation models—a Unet model with a densenet121 encoder (Henceforth referred to as the DenseUnet) (Shaikh et al., 2017), a Unet model with a ResNet encoder (ResUnet) (Kermi et al., 2018), and a simple encoder-decoder network which has a similar architecture to the ResUnet but without skip or residual connections (SimUnet). All models were trained till convergence on the BraTS

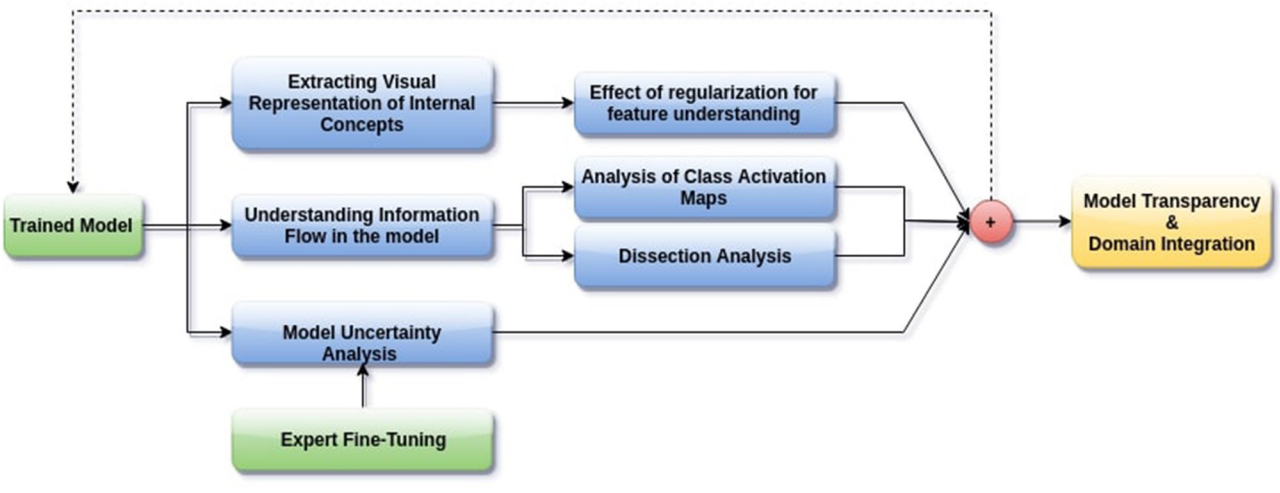

FIGURE 1 | Proposed pipeline for interpreting brain tumor segmentation models to aid in increasing transparency. The dotted backward arrow shows the possibility of using the inferences from such an experiment to enhance the training process of networks.

TABLE 1 | Performance metrics of our networks.

Model type	WT dice	TC dice	ET dice
DenseUnet	0.830	0.760	0.685
ResUnet	0.788	0.734	0.649
SimUnet	0.743	0.693	0.523

WT, Whole Tumor; TC, Tumor Core; ET, Enhancing Tumor.

2018 dataset (Menze et al., 2014; Bakas et al., 2017a,b, 2018). A held out validation set of 48 volumes (including both LGG and HGG volumes) was used for testing. **Table 1** shows the performance of the three models on this test set.

Our models are not meant to achieve state of the art performance. Instead, we aim to demonstrate our methods on a set of models with different structures commonly used for brain tumor segmentation and compare them to better understand the process they take to segment the tumors. In this primary study, we do not use 3D models, since the visualization and analysis of interpretability related metrics is simpler for 2D models. Also, it is not clear how some of our results would scale to 3D models and whether it would be possible to visualize these. For example, disentangled concepts observed by performing network dissection might not be meaningful when visualized slice wise and would have to be visualized in 3D. This and the related analysis poses an additional layer of difficulty.

We now give a brief introduction of each interpretability techniques in our pipeline. *Network Dissection* aims to quantify to what extent internal information representation in CNNs is human interpretable. This is important to understand what concepts the CNN is learning on a filter level, and whether these correspond with human level concepts. *Grad-CAM* allows us to see how the spatial attention of the network changes over layers, i.e., what each layer of the network looks at in a specific input image. This is done by finding the importance of each neuron in the network by taking the gradient of the output with respect to that neuron. In *feature visualization*, we find the input image which maximally activates a particular filter, by randomly initializing an input image and optimizing this for a fixed number of iterations, referred to as *activation maximization*. Such an optimized image is assumed to be a good first order representation of the filter, which might allow us to understand how a neural network "sees." *Test-time dropout* is a computationally efficient method of approximate Bayesian Inference on a CNN to quantify uncertainty in the outputs of the model.

In the following sections, each element of the proposed pipeline is implemented and its results and implications are discussed.

3. UNDERSTANDING INFORMATION ORGANIZATION IN THE MODEL

3.1. Network Dissection

Deep neural networks may be learning explicit disentangled concepts from the underlying data distribution. For example, Zhou et al. (2014) show that object detectors emerge in networks trained for scene classification. To study whether filters in brain tumor segmentation networks learn such disentangled concepts, and to quantify such functional disentanglement (i.e., to quantify to what extent individual filters learn individual concepts), we implement the Network Dissection (Bau et al., 2017) pipeline, allowing us to determine the function of individual filters in the network.

In-Network Dissection, the activation map of an internal filter for every input image is obtained. Then the distribution α of the activation is formulated over the entire dataset. The obtained activation map is then resized to the dimensions of the original image and thresholded to get a concept mask. This concept mask might tell us which individual concept a particular filter learns when overlaid over the input image.

For example, in the context of brain-tumor segmentation, if the model is learning disentangled concepts, there might be separate filters learning to detect, say, the edema region, or the necrotic tumor region. The other possibility is that the network somehow spreads information in a form not understandable by humans - entangled and non-interpretable concepts.

Mathematically, Network Dissection is implemented by obtaining activation maps $\Phi_{k,l}$ of a filter k in layer l, and then obtaining the pixel level distribution α of $\Phi_{k,l}$ over the entire dataset.

A threshold $T_{k,l}(x)$ is determined as the 0.01-quantile level of $\alpha_{k,l}(x)$, which means only 1.0% of values in $\Phi_{k,l}(x)$ are greater than $T_{k,l}(x)$. (We choose the 0.01-quantile level since this gives the best results qualitatively (visually) and also quantitatively in terms of dice score for the concepts for which ground truths are available). The concept mask is obtained as:

$$M_{k,l}(x) = \Phi_{k,l}(x) \geq T_{k,l}(x) \tag{1}$$

A channel is a detector for a particular concept if:

$$IoU(M_{k,l}(x), gt) = \frac{|M_{k,l}(x) \cap gt|}{|M_{k,l}(x) \cup gt|} \geq c \tag{2}$$

In this study, we only quantify explicit concepts like the core and enhancing tumor due to the availability of ground truths gt and recognize detectors for other concepts by visual inspection. We post-process the obtained concept images to remove salt-and-pepper noise and keep only the largest activated continuous concept inside the brain region in the image. The IoU between the final concept image and the ground truth for explicit concepts is used to determine the quality of the concept.

The results of this experiment, shown in **Figures 2–4**, indicate that individual filters of brain-tumor segmentation networks learn explicit as well as implicit disentangled concepts. For example, **Figure 2E** shows a filter learning the concept *whole tumor region* i.e., it specifically detects the whole tumor region for any image in the input distribution, the filter in **Figure 2B** seems to be learning the *edema region*, while **Figure 2A** shows a filter learning the *white and gray matter region*, an implicit concept which the network is not trained to learn. Similar behavior is seen in all networks (**Figures 2–4**). This means that we can make

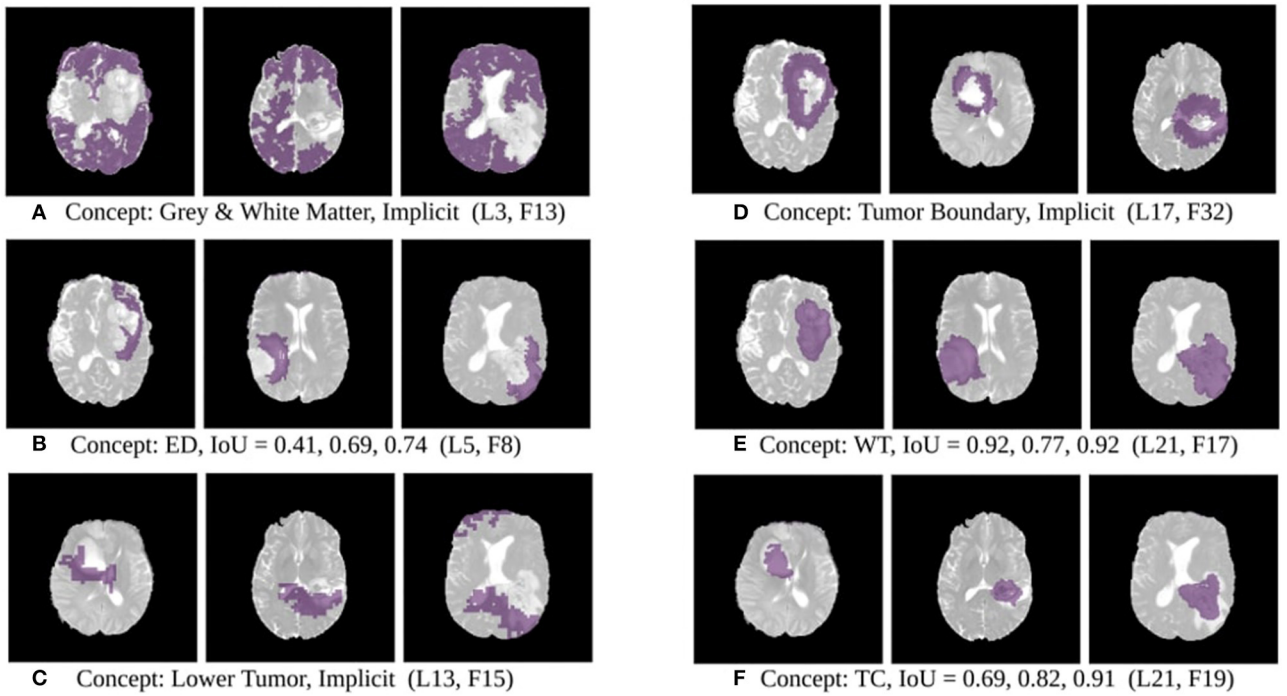

FIGURE 2 | Disentangled concept mask *M* learned by individual filters of the ResUnet overlaid over brain image. This includes explicit concepts for which ground truth labels are available as well as implicit concepts for which their are no labels. IoU scores are mentioned in the sub-captions for all 3 images. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED, Edema.

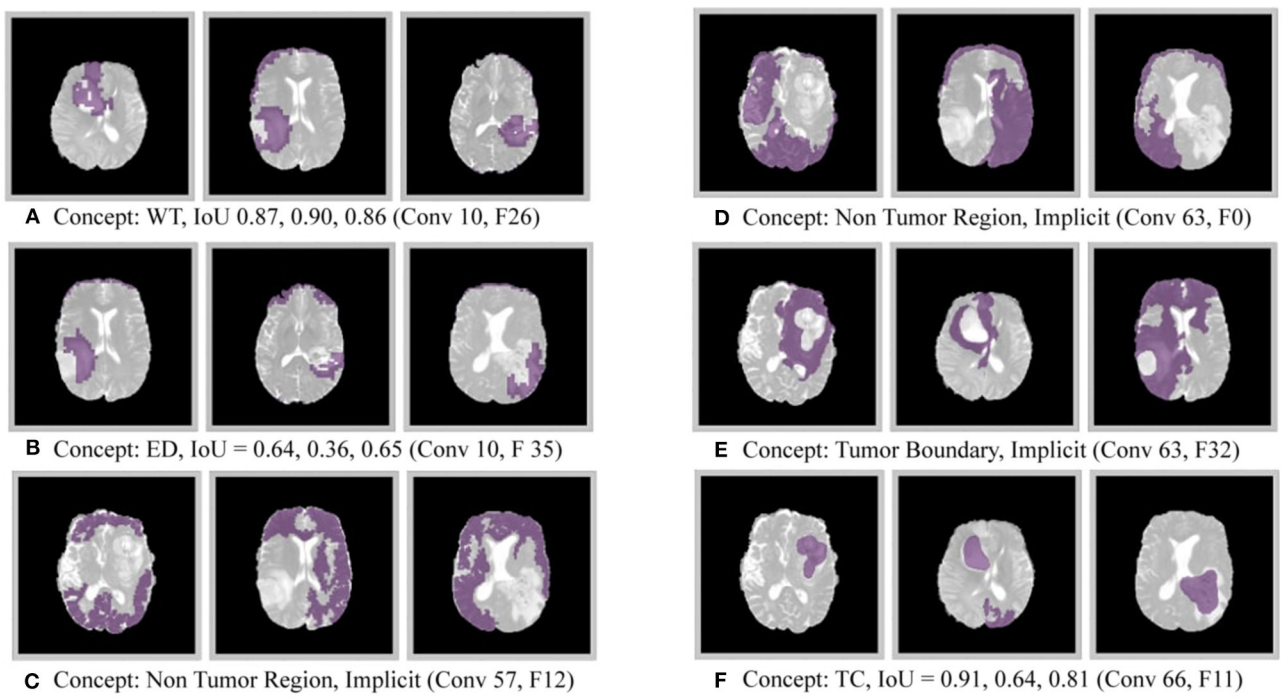

FIGURE 3 | Disentangled concepts learned by filters of the DenseUnet. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED: Edema.

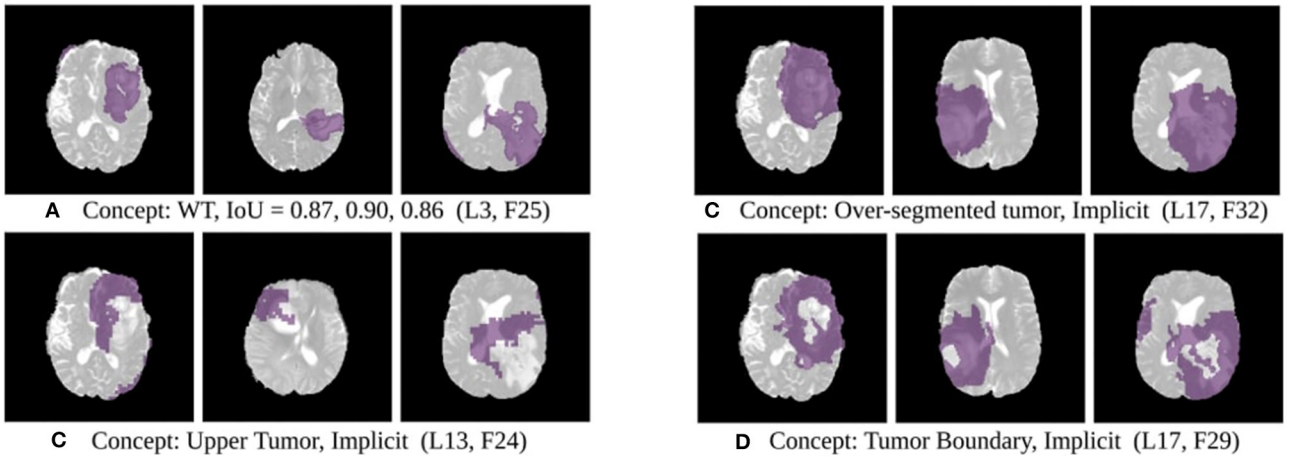

FIGURE 4 | Disentangled concepts learned by filters of the SimUnet. L, Layer; WT, Whole Tumor; TC, Tumor Core; ED: Edema.

attributions based on function to the network at a filter level—indicating a sort of functional specificity in the network i.e., individual filters might be specialized to learn separate concepts.

Neural Networks are inspired by neuroscientific principles. What does this functional specificity mean in this context? Debates are ongoing on whether specific visual and cognitive functions in the brain are segregated and the degree to which they are independent. Zeki and Bartels (1998) discuss the presence of spatially distributed, parallel processing systems in the brain, each with its separate function. Neuroscientific studies have shown that the human brain has some regions that respond specifically to certain concepts, like the face fusiform area Kanwisher and Yovel (2006)—indicating certain visual modularity. Studies based on transcranial magnetic stimulation of the brain also show separate areas of the visual cortex play a role in detecting concepts like faces, bodies, and objects (Pitcher et al., 2009).

The emergence of concept detectors in our study indicates that brain-tumor segmentation networks might show a similar modularity. This indicates that there is some organization in the model similar to the process a human being might take to recognize a tumor, which might have an implications with regards to the credibility of these models in the medical domain, in the sense that they might be taking human-like, or at least human understandable, steps for inference.

The extracted disentangled concepts can also be used for providing contextual or anatomical information as feedback to the network. Though we do not explore this in this study, 3D concept maps obtained from networks can be fed back as multi-channel inputs to the network to help the network implicitly learn to identify anatomical regions like the gray and white matter, tumor boundary etc. for which no labels are provided, which might improve performance. This would be somewhat similar to the idea of feedback networks discussed by Zamir et al. (2017), where an implicit taxonomy or hierarchy can be established during training as the network uses previously learned concepts to learn better representations and increase speed of learning.

3.2. Gradient Weighted Class Activation Maps

Understanding how spatial attention of a network over an input image develops might provide clues about the overall strategy the network uses to localize and segment an object. Gradient weighted Class Activation Maps (Grad-CAM) (Selvaraju et al., 2017) is one efficient technique that allows us to see the networks attention over the input image. Grad-CAM provides the region of interest on an input image which has a maximum impact on predicting a specific class.

Segmentation is already a localization problem. However, our aim here is to see *how attention changes over internal layers of the network*, to determine how spatial information flows in the model. To understand the attentions of each layer on an input image, we convert segmentation to a multi-label classification problem by considering class wise global average pooling on the final layer. The gradient of the final global average pooled value is considered for attention estimation in Grad-CAM. To understand the layer-wise feature map importance, Grad-CAM was applied to see the attention of every internal layer.

This mathematically amounts to finding neuron importance weights $\beta^c_{l,k}$ for each filter k of a particular layer l with respect to the global average pooled output segmentation for a particular channel c:

$$y(c) = \frac{1}{P}\sum_i\sum_j \Phi^c(x) \tag{3}$$

$$\beta^c_{l,k} = \frac{1}{N}\sum_i\sum_j \frac{\partial y(c)}{\partial A^{ij}_{l,k}(x)} \tag{4}$$

$$O_{GradCAM}(c) = ReLU\left(\sum_k \beta^c_{l,k} A_{l,k}(x)\right) \tag{5}$$

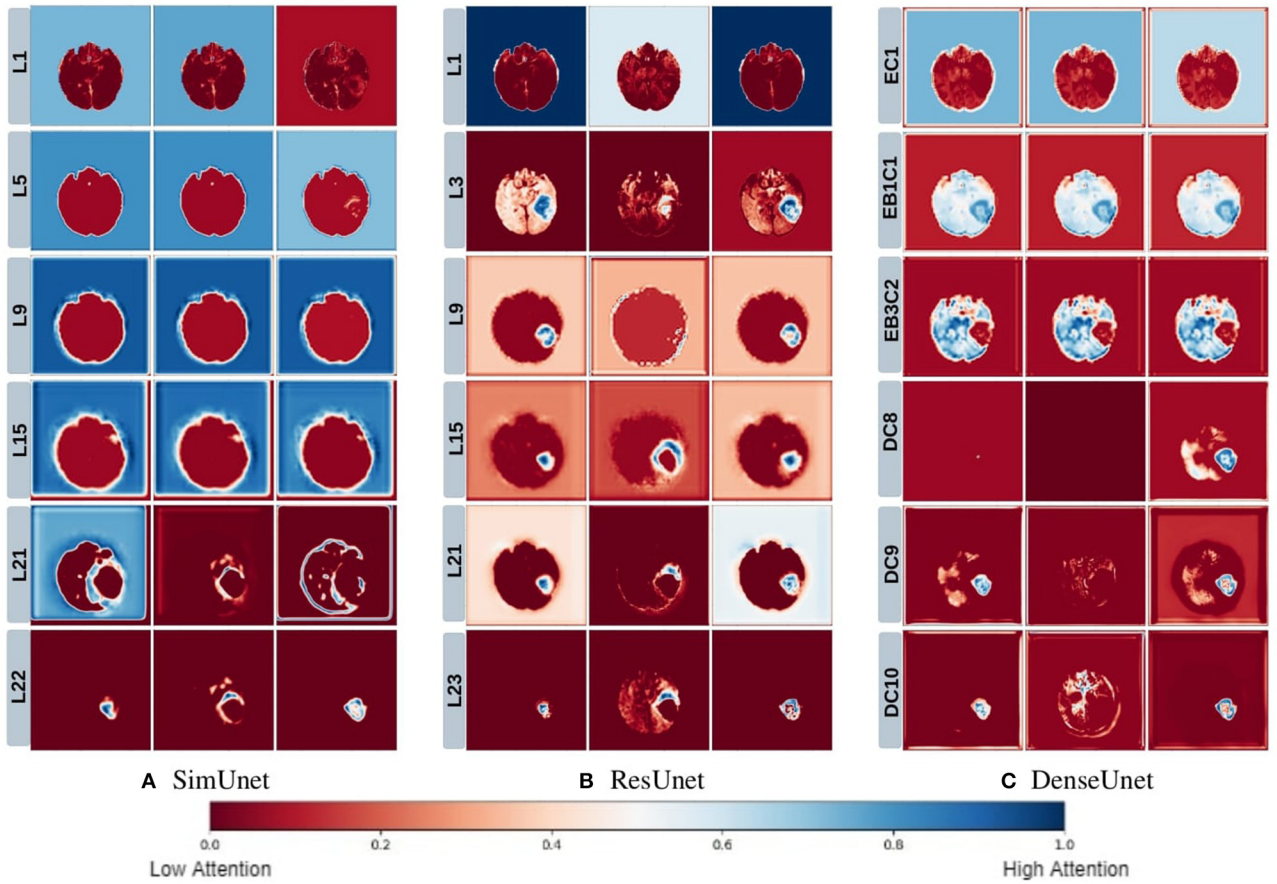

FIGURE 5 | This figure depicts the gradient based class activation maps obtained at selected intermediate layers of all the three networks in ascending order. L, Layer; E, Encoding; B, Block; D, Decoding.

Where, P and N are the number of pixels in the output segmentation map and the activation map of the relevant layer for channel c respectively, Φ^c is the output segmentation map for class c of network Φ, $y(c)$ describes the spatially pooled final segmentation map, $A_{l,k}(x)$ is the activation map for the k^{th} filter of the l^{th} layer, and $O_{GradCAM}(c)$ represents an output map which is the result of *GradCAM* for channel c.

We posit that model complexity and residual connections might have an impact on how early a model can localize the tumor region. For example, the DenseUnet and ResUnet localize the tumor region in the first few layers, while the SimUnet, which has no skip or residual connections, localizes the tumor region only in the final few layers (**Figure 5**). This indicates that skip and residual connections help learn and propagate spatial information to the initial layers for faster localization. While previous literature indicates that skip connections allow upsampling layers to retain fine-grained information from downsampling layers (Drozdzal et al., 2016; Jégou et al., 2017), our results indicate that information might also be flowing in the other direction i.e., skip and residual connections help layers in the downsampling path to learn spatial information earlier.

Drozdzal et al. (2016) also discuss that layers closer to the center of the model might be more difficult to train due to the vanishing gradient problem and that short skip or residual connections might alleviate this problem. Our results support this as well - middle layers of the SimUnet, which does not have residual or skip connections, seem to learn almost no spatial information compared to the other two networks (**Figure 5A**).

Our results in **Figure 5** also show that models take a largely top-down approach to localizing tumors - they first pay attention to the entire brain, then the general tumor region, and finally converge on the actual finer segmentation. For example, attention in all three models is initially in the background region. In the DenseUnet and ResUnet, attention quickly moves to the brain and whole tumor within the first few layers. Finer segmentations are done in the final few layers. The *necrotic tumor* and *enhancing tumor* are often separated only in the last few layers for all models, indicating that segregating these two regions might require a lesser number of parameters.

This top-down nature is consistent with theories on visual perception in humans–the global-to-local nature of visual perception has been documented. Navon (1977) showed through experiments that larger features take precedence over smaller

features, called the *Global Precedence Effect*. While this effect has its caveats (Beaucousin et al., 2013), it is generally robust (Kimchi, 2015). Brain tumor segmentation models seem to take a similar top-down approach, and we see in our experiments that such behavior becomes more explicit as model performance improves.

While the results from the last two sections are not unexpected, they are not trivial either—the models do not need to learn disentangled concepts, especially implicit ones like the whole brain or the white matter region for which no explicit labels have been given, nor do they need to take a hierarchical approach to this problem. The fact that such human-understandable traces of inference can be extracted from brain tumor segmentation models is promising in terms of their acceptance in the medical domain.

4. EXTRACTING VISUAL REPRESENTATIONS OF INTERNAL CONCEPTS

4.1. Activation Maximization

Visualizing the internal features (i.e., the representations of the internal filters obtained on activation maximization) of a network often provides clues as to the network's understanding of a particular output class. For example, visualizing features of networks trained on the ImageNet (Deng et al., 2009) dataset shows different filters maximally activated either by textures, shapes, objects or a combination of these (Olah et al., 2018). However, this technique has rarely been applied to segmentation models, especially in the medical domain. Extracting such internal features of a brain-tumor segmentation model might provide more information about the qualitative concepts that the network learns and how these concepts develop over layers.

We use the Activation Maximization (Erhan et al., 2009) technique to iteratively find input images that highly activate a particular filter. These images are assumed to be a good first-order representations of the filters. Mathematically, activation maximization can be seen as an optimization problem:

$$x^* = \arg\max_x (\Phi_{k,l}(x) - R_\theta(x) - \lambda||x||_2^2) \tag{6}$$

Where, x^* is the optimized pre-image, $\Phi_{k,l}(x)$ is the activation of the k^{th} filter of the l^{th} layer, and $R_\theta(x)$ are the set of regularizers.

In the case of brain-tumor segmentation, the optimized image is a 4 channel tensor. However, activation maximization often gives images with extreme pixel values or random repeating patterns that highly activate the filter but are not visually meaningful. In order to prevent this, we regularize our optimization to encourage robust images which show shapes and patterns that the network might be detecting.

4.2. Regularization

A number of regularizers have been proposed in the literature to improve the outputs of activation maximization. We use three regularization techniques to give robust human-understandable feature visualizations, apart from an L2 bound which is included in Equation (6).

4.2.1. Jitter

In order to increase translational robustness of our visualizations, we implement Jitter (Mordvintsev et al., 2015). Mathematically, this involves padding the input image and optimizing a different image-sized window on each iteration. In practice, we also rotate the image slightly on each iteration. We find that this greatly helps in reducing high-frequency noise and helps in crisper visualizations.

4.2.2. Total Variation

Total Variation (TV) regularization penalizes variation between adjacent pixels in an image while still maintaining the sharpness of edges (Strong and Chan, 2003). We implement this regularizer to smooth our optimized images while still maintaining the edges. The TV regularizer of an image I with (w, h, c) dimension is mathematically given as in Equation (7):

$$R_{TV}(I) = \sum_{k=0}^{c}\sum_{u=0}^{h}\sum_{v=0}^{w}([I(u,v+1,k) - I(u,v,k)] + [I(u+1,v,k) - I(u,v,k)]) \tag{7}$$

4.2.3. Style Regularizer

In order to obtain visualizations which are similar in style to the set of possible input images, we implement a style regularizer inspired from the work of Li et al. (2017). We encourage our optimization to move closer to the style of the original distribution by adding a similarity loss with a template image, which is the average image taken over the input data distribution. In style transfer, the gram matrix is usually used for this purpose. However, we implement a loss which minimizes the distance between the optimized and template image in a higher dimensional kernel space, as implemented in Li et al. (2017), which is computationally less intensive.

Mathematically, Equation (6) is modified to the following:

$$x^* = \arg\max_x (\Phi_{k,l}(x) - \zeta R_{TV}(x) + \gamma L(x,s) - \lambda||x||_2^2) \tag{8a}$$

$$L(x,s) = \sum_i \sum_j (k(x_i,x_j) + k(s_i,s_j) - 2k(x_i,s_j)) \tag{8b}$$

$$k(x,y) = \exp(-\frac{||x-y||_2^2}{2\sigma^2}) \tag{8c}$$

Where $L(x,s)$ it the style loss between the optimized pre-image and the template image s, $k(x,y)$ is the Gaussian kernel, $\Phi_{k,l}(x)$ is the filter for which activations need to be maximized, $R_{TV}(x)$ is the Total Variation Loss, and $||x||_2^2$ is an upper bound on the optimized pre-image $x*$. Approximate values of the regularization coefficients are $\lambda \sim 10^{-4}$, $\gamma \sim 10^{-2}$, and $\zeta \sim 10^{-5}$. For jitter and rotation, the image is randomly shifted by $\sim$8 pixels, and rotated by $\sim$10 degrees.

The effect of varying the hyperparameters for each of the regularizers is shown in **Supplementary Figure 6**. The effect of jitter is most pronounced—adding jitter by just 2-3 pixels helps reduce high frequency noise and clearly elucidate shapes in the image. Increasing total variation regularization increases smoothness while maintaining shapes and boundaries, reducing

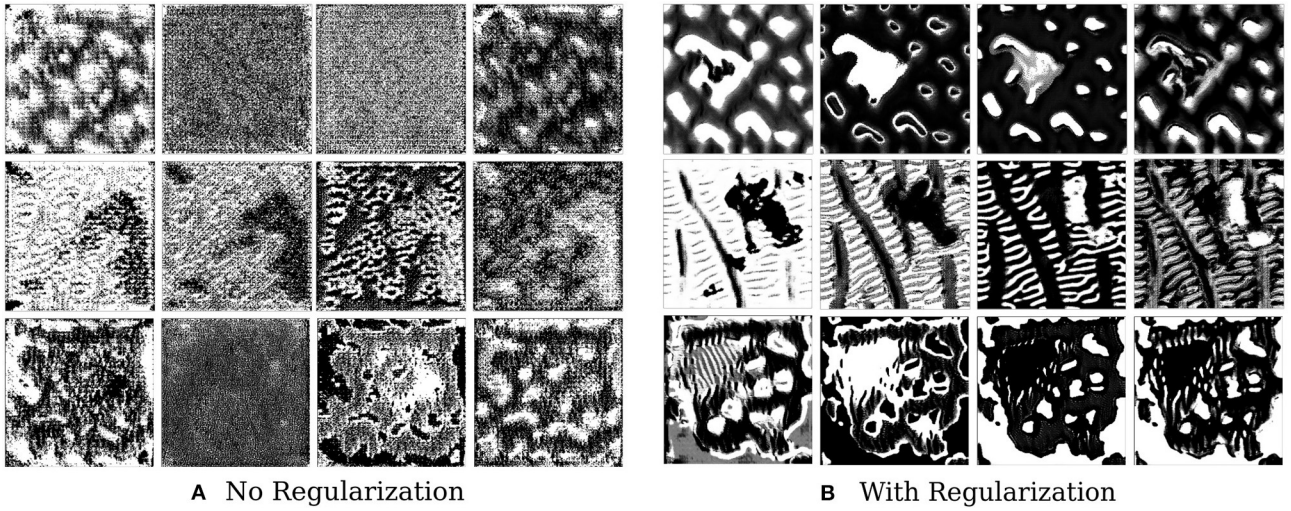

FIGURE 6 | This figure depicts the effect of regularizers on visualized features of brain tumor segmentation models. The four columns on the left show the four channel feature map obtained on optimizing without regularization, while the columns on the right show the effect of adding regularizers.

salt and pepper noise. Increasing style regularization brings the image closer to an elliptical shape similar to a brain. The effect of changing the regularization hyperparameters from a medical perspective in the context brain-tumor segmentation, however, is not clear and further studies would be required in this direction.

We find that style constraining the images and making them more robust to transformations does help in extracting better feature visualizations qualitatively–optimized pre-images do show certain texture patterns and shapes. **Figure 6** shows the results of such an experiment. The effect of regularizers is clear–not regularizing the image leads to random, repeating patterns with high-frequency noise. Constrained images show certain distinct shapes and patterns. It is still not clear, however, that these are faithful reflections of what the filter is actually detecting.

Not a lot of prior work has been done in this area in the context of medical imaging, and our results are useful in the sense that they show that constrained optimization generates such patterns and shapes as compared to noisy unregularized images, which has also been seen in the domain of natural images. In the natural image domain, the resulting pre-images, after regularization, have less high frequency noise and are more easily identifiable by humans. As discussed in the work of Olah et al. (2017) and Nguyen et al. (2016), jitter, L2 regularization, Total Variation, and regularization with mean images priors are shown to produce less noisy and more useful objects or patterns. In medical imaging, however, the resulting patterns and shapes are harder to understand and interpret.

In order to extract clinical meaning from these, a comprehensive evaluation of which regularizers generate medically relevant and useful images based on collaboration with medical professionals and radiologists would be required. This could provide a more complete understanding of what a brain tumor segmentation model actually detects qualitatively. However, this is out of scope of the current study. As we have mentioned in section 7, this will be explored in future work.

5. UNCERTAINTY

Augmenting model predictions with uncertainty estimates are essential in the medical domain since unclear diagnostic cases are aplenty. In such a case, a machine learning model must provide medical professionals with information regarding what it is not sure about, so that more careful attention can be given here. Begoli et al. (2019) discuss the need for uncertainty in machine-assisted medical decision making and the challenges that we might face in this context.

Uncertainty Quantification for deep learning methods in the medical domain has been explored before. Leibig et al. (2017) show that uncertainties estimated using Bayesian dropout were more effective and more efficient for deep learning-based disease detection. Yang et al. (2017) use a Bayesian approach to quantify uncertainties in a deep learning-based image registration task.

However, multiple kinds of uncertainties might exist in deep learning approaches–from data collection to model choice to parameter uncertainty, and not all of them are as useful or can be quantified as easily, as discussed below.

Epistemic uncertainty captures uncertainty in the model parameters, that is, the uncertainty which results from us not being able to identify which kind of model generated the given data distribution. Aleatoric uncertainty, on the other hand, captures noise inherent in the data generating process (Kendall and Gal, 2017). However, Aleatoric Uncertainty is not really useful in the context of this work—we are trying to explain and augment the decisions of the model itself, not the uncertainty in the distribution on which it is fit.

Epistemic uncertainty can, in theory, be determined using Bayesian Neural Networks. However, a more practical and

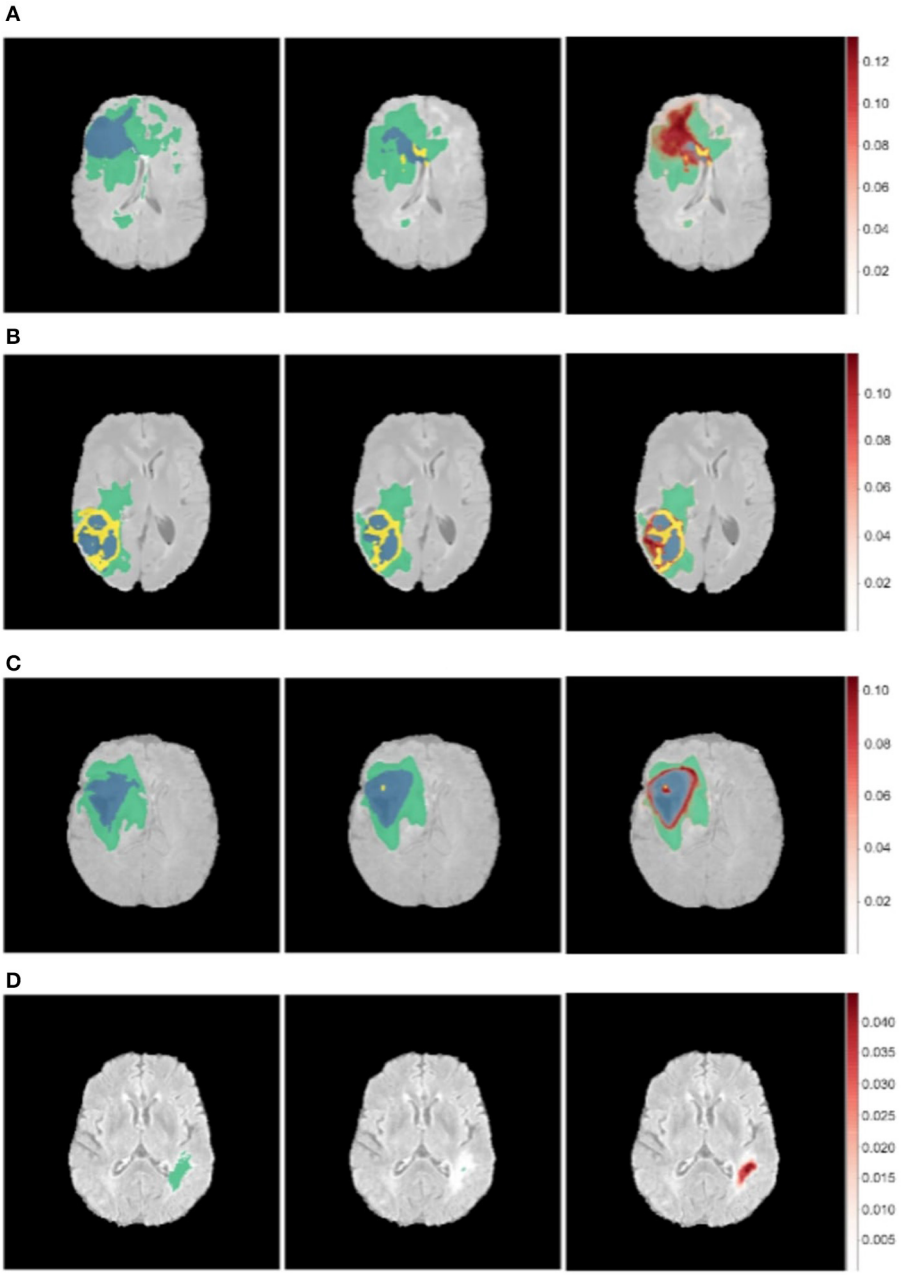

FIGURE 7 | Uncertainty estimations (shown in red) for the DenseUnet using TTD for a selected set of images. Ground Truth (Left), Model Prediction (Middle), and Uncertainty (Right). Misclassified regions are often associated with high uncertainty. **(A)** Misclassified Core Tumor Region which is associated with high model uncertainty. **(B)** Misclassified Enhancing/Core Tumor Region which is associated with high model uncertainty. **(C)** High model uncertainty at class borders. **(D)** Tumor region completely missed by model, captured in the model uncertainty map.

computationally simple approach is to approximate this Bayesian inference by using dropout at test time. We use test time dropout (TTD) as introduced in Gal and Ghahramani (2016) as an approximate variational inference. Then,

$$p(y|x,w) \approx \frac{1}{T}\sum_{t=1}^{t}\Phi(x|w^t) \qquad (9a)$$

$$var_{epistemic}(p(y|x,w)) \approx \frac{1}{T}\sum_{t=1}^{T}\Phi(x|w^t)^T\Phi(x|w^t) - \mathrm{E}(\Phi(x|w^t))^T\mathrm{E}(\Phi(x|w^t)) \qquad (9b)$$

Where $\Phi(x|w^t)$ is the output of the neural network with weights w^t on applying dropout on the t^{th} iteration. The models are retrained with a dropout rate of 0.2 after each layer. At test

time, a posterior distribution is generated by running the model for 100 epochs for each image. We take the mean of the posterior sampled distribution as our prediction and the channel mean of the variance from Equation 9 as the uncertainty (Kendall et al., 2015). The results of this are shown in **Figure 7**.

We find that regions which are misclassified are often associated with high uncertainty. For example, **Figure 7A** shows a region in the upper part of the tumor which is misclassified as *necrotic tumor*, but the model is also highly uncertain about this region. Similar behavior is seen in **Figure 7B**. In some cases, the model misses the tumor region completely, but the uncertainty map still shows that the model has low confidence in this region (**Figure 7D**), while in some cases, boundary regions are misclassified with high uncertainty (**Figure 7C**). In a medical context, these are regions that radiologists should pay more attention to. This would encourage a sort of collaborative effort—tumors are initially segmented by deep learning models and the results are then fine-tuned by human experts who concentrate only on the low-confidence regions, **Figure 1** shows.

More sample images as well as uncertainty for other networks can be found in the **Supplementary Material**.

6. CONCLUSION

In this paper, we attempt to elucidate the process that neural networks take to segment brain tumors. We implement techniques for visual interpretability and concept extraction to make the functional organization of the model clearer and to extract human-understandable traces of inference.
From our introductory study, we make the following inferences:

- Disentangled, human-understandable concepts are learnt by filters of brain tumor segmentation models, across architectures.
- Models take a largely hierarchical approach to tumor localization. In fact, the model with the best test performance shows a clear convergence from larger structures to smaller structures.
- Skip and residual connections may play a role in transferring spatial information to shallower layers.
- Constrained optimization helps to extract feature visualizations which show distinct shapes and patterns which may be representations of tumor structures. Correlating these with the disentangled concepts extracted from Network Dissection experiments might help us understand how exactly a model detects and generalizes such concepts on a filter level.
- Misclassified tumor regions are often associated with high uncertainty, which indicates that an efficient pipeline which combines deep networks and fine-tuning by medical experts can be used to get accurate segmentations.

As we have discussed in the respective sections, each of these inferences might have an impact on our understanding of deep learning models in the context of brain tumor segmentation.

While more experiments on a broader range of models and architectures would be needed to determine if such behavior is consistently seen, the emergence of such human-understandable concepts and processes might aid in the integration of such methods in medical diagnosis–a model which seems to take human-like steps is easier to trust than one that takes completely abstract and incoherent ones. This is also encouraging from a neuroscience perspective - if model behavior is consistent with visual neuroscience research on how the human brain processes information, as some of our results indicate, this could have implications in both machine learning and neuroscience.

7. FUTURE WORK

Future work will be centered around gaining a better understanding of the segmentation process for a greater range of models (including 3D models) and better constrained optimization techniques for extracting human-understandable feature visualizations which would allow an explicit understanding of how models learn generalized concepts. For instance, it would be worth-wile to understand what set of regularizers generates the most medically relevant images. Textural information extracted from the optimized pre-images can also be analyzed to determine their correlation with histopathological features.

Further exploration regarding how these results are relevant from a neuroscience perspective can also be done, which might aid in understanding not just the machine learning model, but also how the brain processes information. The inferences from our explainability pipeline can also be used to integrate medical professionals into the learning process by providing them with information about the internals of the model in a form that they can understand.

DATA AVAILABILITY STATEMENT

Publicly available data sets were used for this study. The data sets can be found at the BRATS 2018 challenge (https://www.med.upenn.edu/sbia/brats2018/data.html) (Bakas et al., 2017a,b).

AUTHOR CONTRIBUTIONS

PN and AK developed the pipeline, performed the analysis, implementation, revised the manuscript, and generated the visualizations. PN wrote the first draft. GK edited the manuscript, supervised, and funded the study.

ACKNOWLEDGMENTS

This work was funded by the Robert Bosch Center for Data Science and Artificial Intelligence (RBCDSAI), under project number CR1920ED617RBCX008562 (Interpretability for Deep Learning Models in Healthcare).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00006/full#supplementary-material

Supplementary Figure 1 | Network Architectures used in our study.

Supplementary Figure 2 | Concepts learned by filters of a particular layer of the ResUnet for an input image (Conv Layer 21).

Supplementary Figure 3 | Concepts learned by filters of a particular layer of the DenseUnet for an input image (Encoding Block 1, Conv 2).

Supplementary Figure 4 | Grad-CAM results for consecutive layers of the ResUnet [view: top to bottom, column **(A)**, followed by top to bottom, column **(B)**].

Supplementary Figure 5 | Activation maps for layers of the ResUnet.

Supplementary Figure 6 | Effect of independently changing hyperparamaters for each regularizer. **(Top)** Jitter coefficient increases [0 pixels, 1p, 6p, 12p, 20p]. **(Middle)** Style Coefficient increases [10^{-2}, 10^{-1}, 1, 5, 10]. **(Bottom)** Total Variation regularization increases [10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}] to smoothen image.

Supplementary Figure 7 | Uncertainty estimations (shown in red) for the DenseUnet **(a–d)** and ResUnet **(e,f)**. Ground Truth (Left), Model Prediction (Middle), and Uncertainty (Right).

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the Tcga-gbm Collection*. The cancer imaging archive (2017).

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017b). Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., 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* (Honolulu, HI), 6541–6549.

Beaucousin, V., Simon, G., Cassotti, M., Pineau, A., Houdé, O., and Poirel, N. (2013). Global interference during early visual processing: Erp evidence from a rapid global/local selective task. *Front. Psychol* 4:539. doi: 10.3389/fpsyg.2013.00539

Begoli, E., Bhattacharya, T., and Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* 1:20. doi: 10.1038/s42256-018-0004-1

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* (IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, eds G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise (Athens: Springer), 179–187. doi: 10.1007/978-3-319-46976-8_19

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *Univ. Montreal* 1341:1. Available online at: https://www.semanticscholar.org/paper/Visualizing-Higher-Layer-Features-of-a-Deep-Network-Erhan-Bengio/65d994fb778a8d9e0f632659fb33a082949a50d3#paper-header

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a bayesian approximation: representing model uncertainty in deep learning," in *International Conference on Machine Learning* (New York, NY), 1050–1059.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (Turin: IEEE), 80–89.

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

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI), 11–19.

Kanwisher, N., and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B Biol. Sci.* 361, 2109–2128. doi: 10.1098/rstb.2006.1934

Kendall, A., Badrinarayanan, V., and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint* arXiv:1511.02680.

Kendall, A., and Gal, Y. (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Long Beach, CA), 5574–5584.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131. doi: 10.1016/j.cell.2018.02.010

Kermi, A., Mahmoudi, I., and Khadir, M. T. (2018). "Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multimodal mri volumes," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Granada: Springer), 37–48. doi: 10.1007/978-3-030-11726-9_4

Kimchi, R. (2015). *The Perception of Hierarchical Structure*. Oxford handbook of perceptual organization, 129–149.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* 7:17816. doi: 10.1038/s41598-017-17876-z

Li, Y., Wang, N., Liu, J., and Hou, X. (2017). Demystifying neural style transfer. *arXiv preprint* arXiv:1701.01036.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imag.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Molnar, C. (2018). *Interpretable Machine Learning*. A Guide for Making Black Box Models Explainable, 7.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google AI Blog* (Google). Available online at: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cogn. Psychol.* 9, 353–383.

Nguyen, A., Yosinski, J., and Clune, J. (2016). Multifaceted feature visualization: uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint* arXiv:1602.03616.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*. Available online at: https://distill.pub/2017/feature-visualization (accessed August 30, 2019).

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., et al. (2018). The building blocks of interpretability. *Distill*. Available online at: https://distill.pub/2018/building-blocks (accessed August 28, 2019).

Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., and Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Curr. Biol.* 19, 319–324. doi: 10.1016/j.cub.2009.01.007

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* (Venice), 618–626.

Shaikh, M., Anand, G., Acharya, G., Amrutkar, A., Alex, V., and Krishnamurthi, G. (2017). "Brain tumor segmentation using dense fully convolutional neural network," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes (Quebec City, QC: Springer), 309–319. doi: 10.1007/978-3-319-75238-9_27

Strong, D., and Chan, T. (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems* 19:S165. doi: 10.1088/0266-5611/19/6/059

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12:e0174944. doi: 10.1371/journal.pone.0174944

Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: fast predictive image registration–a deep learning approach. *NeuroImage* 158, 378–396. doi: 10.1016/j.neuroimage.2017.07.008

Zamir, A. R., Wu, T.-L., Sun, L., Shen, W. B., Shi, B. E., Malik, J., et al. (2017). "Feedback networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (San Juan, Puerto Rico), 1308–1317.

Zeki, S., and Bartels, A. (1998). The autonomy of the visual systems and the modularity of conscious vision. *Philos. Transact. R. Soc. Lond. Ser. B Biol. Sci.* 353, 1911–1914. doi: 10.1098/rstb.1998.0343

Zhang, Q.-S., and Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Front. Inform. Techn. Electr. Eng.* 19, 27–39. doi: 10.1631/FITEE.1700808

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint* arXiv:1412.6856.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.