Interpretability of a Deep Learning Model for Rodents Brain Semantic Segmentation

Leonardo Nogueira Matos, Mariana Fontainhas Rodrigues, Ricardo Magalhães, Victor Alves, Paulo Novais

Interpretability of a Deep Learning Model for Rodents Brain Semantic Segmentation

Leonardo Nogueira Matos[1][0000−0002−6302−3299], Mariana Fontainhas Rodrigues[2], Ricardo Magalhães[2][0000−0001−6279−2195], Victor Alves[3][0000−0003−1819−7051], and Paulo Novais[3][0000−0002−3549−0754]

[1] Computer Science Department, Federal University of Sergipe, Brazil.
leonardo@dcomp.ufs.br
[2] Life And Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal. {marianafontainhas,ricardo.lazarus}@gmail.com
[3] Algoritmi Center, University of Minho, Braga, Portugal.
{valves,pjon}@uminho.pt

Abstract. In recent years, as machine learning research has become real products and applications, some of which are critical, it is recognized that it is necessary to look for other model evaluation mechanisms. The commonly used main metrics such as accuracy or F-statistics are no longer sufficient in the deployment phase. This fostered the emergence of methods for interpretability of models. In this work, we discuss an approach to improving the prediction of a model by interpreting what has been learned and using that knowledge in a second phase. As a case study we have used the semantic segmentation of rodent brain tissue in Magnetic Resonance Imaging. By analogy with what happens to the human visual system, the experiment performed provides a way to make more in-depth conclusions about a scene by carefully observing what attracts more attention after a first glance in en passant.

Keywords: Deep Learning Model · Magnetic Resonance Imaging · Interpretability

1 Introduction

A few years ago, deep learning technology achieved state-of-the-art in several areas of Artificial Intelligence and sparked a growing interest from the academic community, especially after the article [3], with more than 34,000 citations, has been published. The authors described the model used to win the Imagenet Large Scale Visual Recognition Challange 2012 (ILSVRC2012) [10]. The Alexnet model, used to win the competition, achieved an error rate of 15.3%, a classification error in a set of more than 100 thousand test images organized in 1000 categories. In the previous year, the winning model, the last one that did not use deep learning, achieved an error rate of 26.2%. From 2015, the models have an error rate of less than 5% [1], exceeding the performance of the classification of objects of this base by humans that, according to [1], is 5.1%.

Deep neural networks are also being used successfully in other domains, such as image captioning [14], visual question ansewer [11], music recommendation [6], language translation [13], speech recognition [2], medical image analysis [5], among others. The list goes beyond the classification and segmentation of images, which were the main objectives of Imagenet's challenge.

Despite the unprecedented breakthroughs in a variety of computer vision tasks, model understanding and interpretation is of utmost importance in some critical areas. In criminal justice, for example, a decision that affects the life of a legal individual can not be taken into account if it is based on a method without transparency. Medicine is another critical area that mainly rely on interpretable models. Medical staff also requires decisions made by computers provide some sort of explanation, so interpretability is not only desirable but also necessary. Nevertheless, a deep model with a deep and complex architecture is naturally hard to interpret. On the other hand, shallow models such as decision trees and linear regressors are easy to interpret, and therefore their decisions are easy to explain. The more complex the model, the more difficult it becomes to interpret it. The medical area therefore needs mechanisms that make complex models interpretable in order to accept them as trustworthy.

In this work we explore a method for interpretability of deep neural networks, called Guided Backpropagation (GBP), which provides insights to interpret decision making by showing parts, artifacts or patterns in the input that were relevant to the model. We used as a case study a CNN trained to segment different parts that form the brain of a rodent on an MRI basis. Then we try to take advantage of the knowledge learned, and identified by the GBP, doing a feedback of the system, that is, once we identify relevant parts, we refine the prediction by discarding irrelevant parts that, in this case, correspond to artifacts that do not belong to the animal's brain. For a successful use of the gradient, we find that this signal can be modeled by a log normal pdf. This is one of the contributions of the work, since the hypothesis on the distribution can be used in a wider field of scientific applications.

2 Machine Learning Interpretability

There is a vast literature addressing machine learning interpretability, especially in recent years [16], but, as discussed in [4], the term interpretability is ill-defined. Therefore, we assume a consensual and yet subjective meaning to that concept as well as some desired characteristics for them as depicted on [8] and [7]. We assume interpretability as being a prerequisite for trust. Ribeiro *et al* argue that humans would trust in a machine learning if predictions could be explained and the explanations are faithful and intelligible. They continue to state that explaining a prediction is related to present textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. patches in an image) and the model's prediction.

According to Pereira's line of reasoning, methods can be categorized as model dependent when, unlike deep neural networks, the model is restricted to an

inherently easy to interpret family, or model agnostic, a more comprehensive and flexible case, when model is treated as a "black box".

The methods can be global, when the ability to interpret is concentrated on how the model learn the data from a population, i.e., it does not concern the prediction of an individual sample on isolation. A popular approach is the visualization of high-dimensional data t-SNE (Van der Maaten & Hinton, 2008), a form to project high dimensional data on to the Cartesian plane preserving the notion of proximity. Pereira used the global interpretability to test the coherence between the proposed method and the a priori knowledge of specialists. They developed a method to segment brain lesions on MRI and were able to observe through global interpretability what MRI sequences were most appropriate for different tasks such as normal tissue segmentation versus lesion segmentation.

Another important desired characteristic covered by Pereira is explicability which deals with the reasoning about a particular decision. It is based under the assumption that it is possible to explain the reason for a given activation, identifying artifacts and regions in the input responsible for this activation. Some authors describes methods that propagate the signal back from the end to the beginning throwgouht the model. Techniques that adhere to that concept are saliency maps.

3 Methods

The methods discussed in this section are addressed in two blocks: machine learning system and interpretability system.

3.1 Machine Learning System

A Convolutional Neuronal Network architecture was used, specifically the U-Net architecture [9]. This architecture presents two major novelties, it begins with a spatial contraction phase, where a combination of convolutional and max-pooling layers are used to highlight the information on condensed feature maps. This is followed by an expansion phase using Up-sampling layers, which are concatenated with the matching down-sampling layers.

The architecture was trained using a data-set of rodent MRI data, aiming to classify different tissue classes within these images. A supervised training method was used, using the dice coefficient as the loss function, which was used to evaluate the performance of the model at each step. The optimizer adjusted the weights of the model at each step, using a stochastic gradient algorithm, with a learning rate of 0.0003, a decay of 1.5×10^{-6} and a momentum of 0.9, with a batch size of 5.80% of the data-set was used for training the model (further divided at $80\% - 20\%$ for training and validation) and 20% for testing the model.

3.2 Interpretability System

We will analyzed the model in Sec. 3.1 from the perspective of local interpretability, based on the work of Springenberg *et al.* [12], which is refered in literature as Guided Backpropagation.

Saliency Maps In this work, we try to present visually relevant aspects that allow to interpret the model's functioning. In this case, since we are going to analyze the prediction of an isolated pattern, not a set, and we tried to explain what a network sees to perform a prediction, our approach is based on a saliency map, i.e., an attentional map made by the network. Our analyzes is based on methods that propagate the signal retroactively from the last to the first layer, such as those proposed by Zeiler and Fergus [15], called Deconvnet, and Springenberg et al. [12], called Guided Brackpopagation.

Zeiler and Fergus [15] were the winners of the ILSVRC2013 contest. By publishing the model used, they also published a method to visually present the knowledge the network had learned. This method, called Deconvnet, has become quite popular and follow by others. The idea was to propagate the signal through the network in the reverse order, that is, from end to beginning, until reaching the input level. The signal is the activation of an isolated neuron, usually the highest activation in a layer, although in the original article they used the 10 higher activations. There are two important aspects to consider in this approach: the weights the network had learned during training are the same used to backpropagate the signal throughout convolutional layers, 2) the method has a forward phase in which a pattern is presented and the positions of the maxima in max pooling layers, called switches, are saved and a backward phase when the salience map is properly composed.

In [12], Springenberg *et al.* established a different method of identifying the artifacts in the input that most influence decision making. They observed that by propagating the gradient in the reverse order and projecting it at the input, it is also possible to identify patterns that influenced the activation of an output unit. In addition, they realized that the backpropagation of the gradient could be guided only by the higher values, neglecting the values of smaller amplitude, which justifies the name of the method — Guided Backpropagation. The control of the propagation proposed by Springenberg *et al.* was imposed on the relu layers. In these layers, the signal was propagated only in places where the relu function allows the signal to pass in both forward and backward phase. To illustrate this concept, consider the example in Figure 1, adapted from [12]. Let f be the propagated signal in the forward phase and b, the signal propagated in the backward. The superscript index indicates the layer and the subscript index its position. Springenberg *et al.* have shown that in the deeper layers the reconstructed image is cleaner than that obtained by the Deconvnet method.

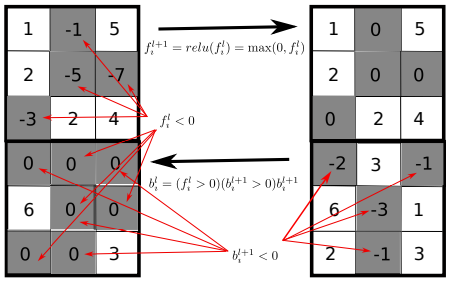

Fig. 1. Gradient guided backpropagation. Adapted from [12]

4 Experiments

4.1 Database

A data-set with 144 images were used to train the network. Data was acquired on a 11.7T using an SE-EPI diffusion sensitive acquisition with TR=5s, $TE = 20$, voxel resolution of $0.375 \times 0.375 \times 0.5mm^3$. Data was pre-processed using FSL, correcting movement and averaging B0 weighted images. The ground truth was generated using SPM segment tool to create a semantic classification of the brain. All in-vivo experiments were done in the context of the FCT-ANR co-founded SIGMA project and were conducted in accordance with the recommendations of the European Community (2010/63/EU) and the French legislation (decree nº2013-118) for use and care of laboratory animals and were approved by the "Comité d'Éthique en Expérimentation Animale du Comissariat à l'Énergie Atomique et aux Énergies Alternatives – Direction des Sciences du Vivant, Ile-de-France (CETEA/CEA/DSV IdF, protocol number ID 13-023).

4.2 Rodents' brain semantic tissue segmentation

The architecture was inspired by U-NET. It is a full convolutional network, which means that it does not contain full connected layers and, since it performs a multiclass segmentation, it contains a softmax activation function in the last layer. The network segments three different classes: white matter (WM), gray matter (GM) and cerebrospinal liquid. There is also an extra class that corresponds to the background, i.e., it refers to the negation of the others, since the softmax output must sum one. Hence, the last layer of the network contains four binary channels (featuremaps), each one related to one class, Fig 2.

4.3 Background removal

The model was trained with sectioned brain images. That is, the images were preprocessed before being presented to the network. Preprocessing consisted of the removal of existing artifacts that were not part of rat brain tissue by the application of a mask. The pixels inside the mask were preserved, the outer ones

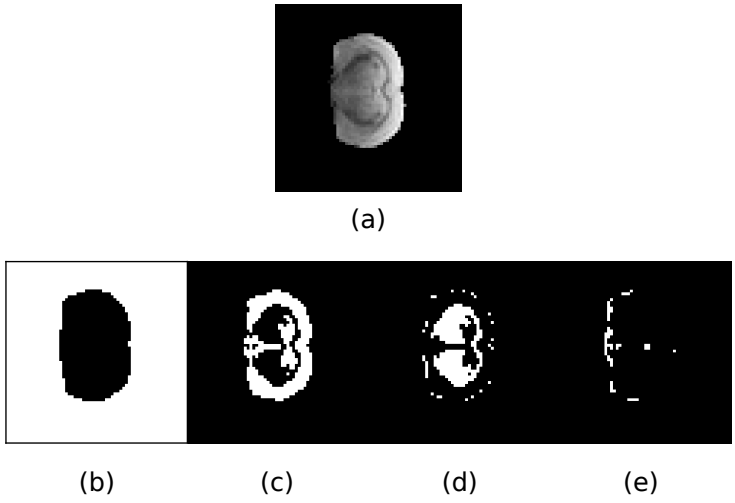

Fig. 2. Input and output. (a) Input; (b) Background. (c) White matter; (d) Gray matter; (e) Cerebrospinal liquid

were zeroed. This step allowed to treat with clean images, without presence of foreign objects, and produce an output with high precision. In the test phase, however, it is not possible to use the same mask used in the training, because during image acquisition procedure small displacements can occur which leads to translated images, not allowing the mask to be fitted on the target.

Model explanation When applying GBP, it is possible to see that the edges are the most importants parts the network takes into account to make a decision. Even when segmenting gray matter, Fig 2 (d), which is the innermost area, GBP is higher at the edges. In Fig 3 we can also observe that, if a translation of the input occurs, a corresponding translation will also occur at the output. This can be explained by the fact that convolution is translation invariant, in this way convolutional layers can identify the presence of patterns in any place. Another important fact is that, since the background used in the training images is fairly uniform, since they always have a dark background, the presence of any artifacts in the input can easily confuse the network, leading to a wrong segmentation, Fig. 4. The use of the mask, therefore, is imperative because without it the model can not produce the expected results.

Image enhancement To identify the location of the mask, we use a method to enhance the projected gradient image, or simply gradient image, as we will call it from now on. This is the central element in this decision-making. That is, the gradient image which corresponds to what the network is seeing, identifies the site where the mask should be positioned. When dealing with images of the gradient, more specifically of its magnitude, we assume that its histograms are

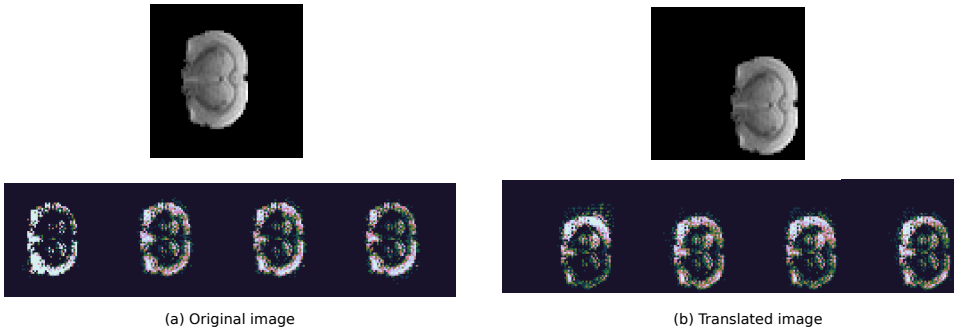

Fig. 3. Saliency maps of translated images

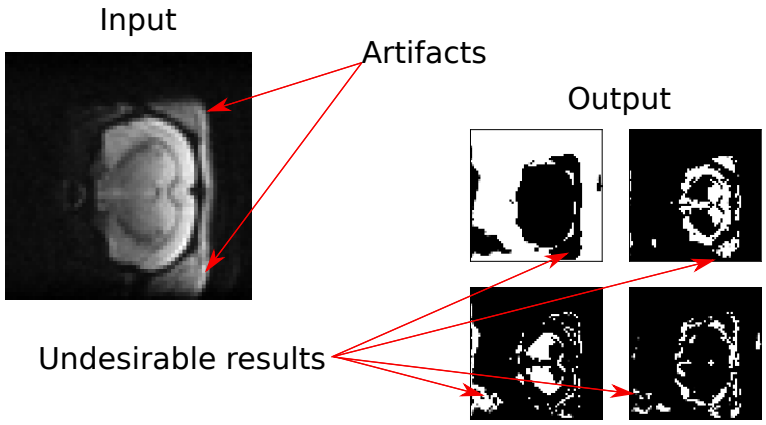

Fig. 4. Noisy segmentation

governed by a log normal distribution. Based on this hypothesis, we can adjust the histogram distribution of the gradient to that of a Gaussian distribution with $\mu = 0$ and $\sigma = 1$, which would allow us to identify a threshold for distinguish between foreground and background.

The transformation of the histogram to fit it into a Gaussian curve can be done by applying a linear expansion, followed by a logarithmic compression, Fig 5. The next step is to normalize the values by the linear transformation $z = (x - \mu_x)/\sigma_x$. The final step consist to apply a threshold function, Eq. 1, to z, in replacement of step function as usually is done. This allows us to leave some intermediate values, which are not labeled as background or foreground, postponing to the next step, which takes into account the geometry of the shape, the decision on which region these pixels belong to.

The maximum in Eq. 1 occurs when $x = 2$. It means that positive values far from zero are mapped to 1, nevertheless, if they are too high, i.e. grather than 2, they decrease in size, which favor to reduce the influence of artifacts since they also have pixels associated to high gradients. In this work, by making the

denominator of exponent in (1) equals to 0.5, we narrow the interval where the maximum activations are located.

$$f(x) = e^{-(\frac{x-2}{.5})^2} \tag{1}$$

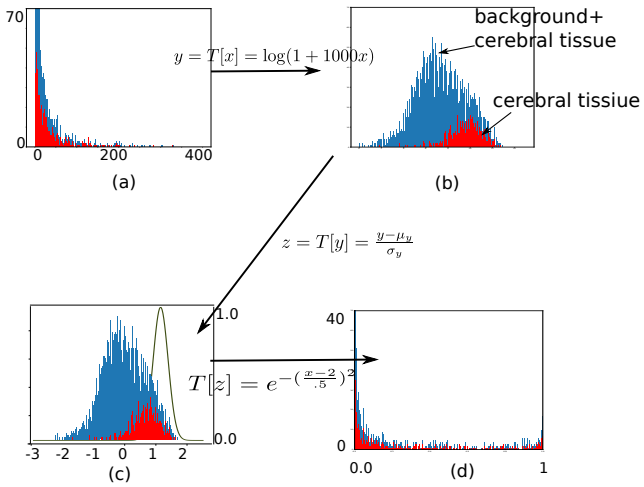

Fig. 5. Histogram transform. (a) Histogram of gray level magnitude; (b) Normal histogram; (c) Thresholding using sigmoid function; (d) Final histogram

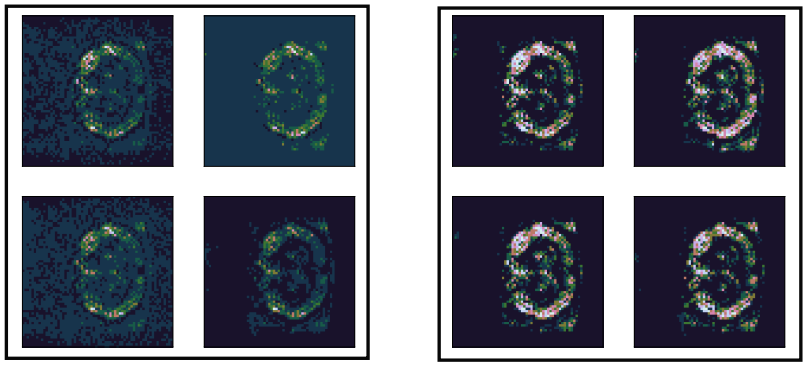

Fig. 6. Image enhancement

Image correlation The second step in mask positioning is given by obtaining its location from the threshold applied to the gradient images. This is done by calculating the 2D cross correlation between the enhanced gradient image and the mask. Cross-correlation involving binary images, or roughly binary, has maxima in places where there is greater alignment between the shapes. Because gradient images generally have shapes and contours that are repeated in different achievements of magnetic resonance imaging, the use of cross-correlation may be successful in this type of application. To mitigate the effect caused by the presence of artifacts, the mask is preliminarily multiplied by the image gradient of the model, Fig 7 (a), which makes the resulting image more sparse, reducing cross-correlation with parts belonging to the artifacts without changing the value of the correlation with the parts of the brain tissue, Fig. 7 (c).

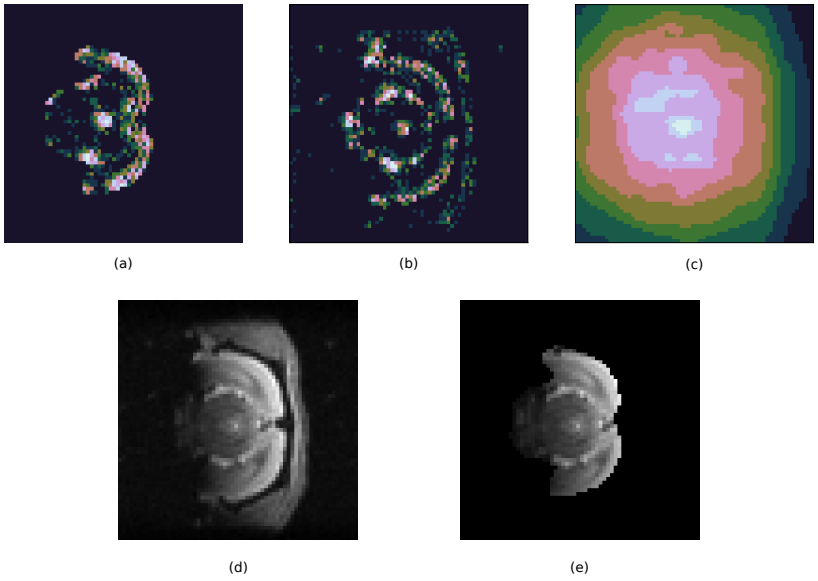

Fig. 7. Shape alignment. (a) Shape mask; (b) Gradient image; (c) Cross correlation 2D; (d) Input; (e) Input after background removal

4.4 Results

We discuss the background removal on the magnetic resonance of rodents by analysing the semantic segmentation of the tissues WM, GM and cerebrospinal fluid as well as the segmentation of the brain over 40 MRI slices of an individual. We compare the results of background removal based on GBP against a ground truth obtained by manual segmentation. The metrics accuracy, specificity, sensitivity and dice similarity coefficient (DSC), commonly adopted in segmentation analysis, are presented in tables 1 and 2.

Table 1. Network performance after background removal (with GBP)

Class	Accuracy	Specificity	Sensitivity	DSC
Brain segmentation	**98.52**	**99.17**	**99.07**	**99.12**
White Matter	98.38	87.9	84.75	86.05
Gray Matter	98.41	81.83	73.1	76.76
Cerebrospinal Liquid	98.41	53.61	68.55	60.15

Table 2. Network performance after background removal (manually)

Class	Accuracy	Specificity	Sensitivity	DSC
White Matter	98.92	92.67	89.65	91.05
Gray Matter	98.96	89.45	81.02	84.65
Cerebrospinal Liquid	99.12	71.79	84.59	77.61

It can be observed that the segmentation of the brain was performed very efficiently, as presented in first row in Table 1. However, the semantic segmentation reached low performance, especially for cerebrospinal fluid. This is mainly due to the fact that the presence of artifacts related to the exterior of the animal's brain harm the perfect mask alignment. As a result, the metrics associated with small parts, such as those occupied by the cerebrospinal fluid, become impaired.

Another important aspect that should be considered is the fact that removing the background by applying a mask coming from of another animal, although well positioned, may leave residues or remove parts of the foreground since the brain sizes are not necessarily the same. In Fig. 8 we present a particular case (slice 25) where precision is high for brain area and low for cerebrospinal fluid. Despite the fact that there is a tight aligment between manually and GBP background removal, small displacements in the brain area afects severly the metrics evaluation of cerebrospinal liquid due the fact that it occupies a small portion of the image.

5 Conclusion

We discuss in this work an approach for interpretability of CNN models. We show that the prediction of a model can be increased using the knowledge acquired by the model itself. In a case study involving semantic segmentation of rodent brain tissues, we have shown that the important parts of the input, identified by the GBP method, can be used to enhance the prediction in a second phase. More specifically, we proposed to use what the network learn to find the place to put a mask for background removal. In analogy to what occurs with the visual system of humans, the experiment carried out presents a way to draw more accurate conclusions about a scene by observing carefully what draws more attention after a first interpretation in *en passant*. We also show a way to enhance the output of GBP method, exploring the hypothesis that the gray level distribution of the images is governed by log normal pdf. This is an interesting finding, as it may

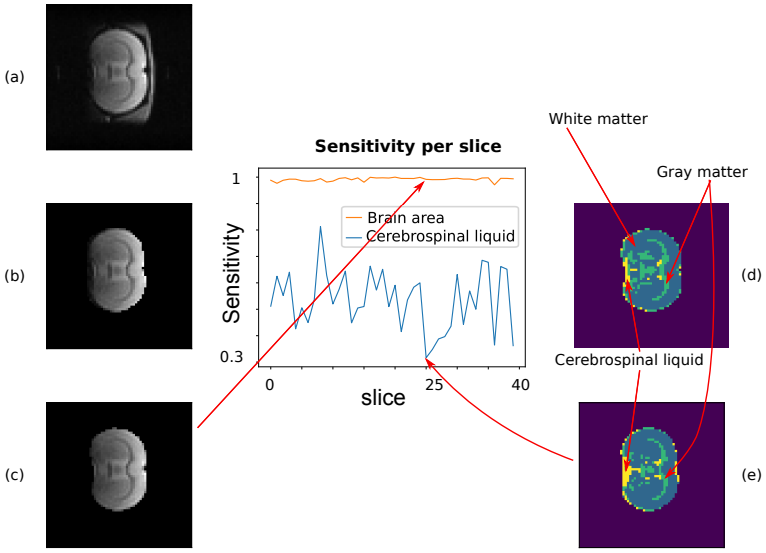

Fig. 8. Precision on segmenting cerebrospinal liquid. (a) Input image; (b) Background manually removed; (c) Background removal with GBP; (d) Semantic segmentation ground trouth; (e) CNN semantic segmentation based on GBP background removal

help support other works exploring the nature of the probability distribution of the gradient-propagated signal. As future work, we intend to explore more precise techniques for performing segmentation of the brain, ie the removal of the background, using the information provided by the GBP method. This can be done with the help of another connectionist system replacing the 2D cross-correlation approach explored in this article.

Acknowlegdments. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019. We gratefully acknowledge the support of the NVIDIA Corporation with their donation of a Titan V board used in this research.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
2. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine **29**(6), 82–97 (2012)
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

4. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
5. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
6. Van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Advances in neural information processing systems. pp. 2643–2651 (2013)
7. Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C.A., Reyes, M.: Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation. Medical image analysis **44**, 228–244 (2018)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144. ACM (2016)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
11. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. CoRR, abs/1610.02391 **7** (2016)
12. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
14. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
16. Zhang, Q.s., Zhu, S.C.: Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering **19**(1), 27–39 (2018)