Towards Interpretability of Segmentation Networks by Analyzing DeepDreams

Vincent Couteaux[1,2(✉)], Olivier Nempont[2], Guillaume Pizaine[2], and Isabelle Bloch[1]

[1] LTCI, Télécom Paris, Institut polytechnique de Paris, Paris, France
vincent.couteaux@telecom-paristech.fr
[2] Philips Research Paris, Suresnes, France

Abstract. Interpretability of a neural network can be expressed as the identification of patterns or features to which the network can be either sensitive or indifferent. To this aim, a method inspired by DeepDream is proposed, where the activation of a neuron is maximized by performing gradient ascent on an input image. The method outputs curves that show the evolution of features during the maximization. A controlled experiment shows how it enables to assess the robustness to a given feature, or by contrast its sensitivity. The method is illustrated on the task of segmenting tumors in liver CT images.

Keywords: Interpretability · Deep Learning · DeepDream · Segmentation · Liver CT images

1 Introduction

Interpretability of deep neural networks is becoming more and more crucial as deep learning algorithms perform critical tasks such as driving a car or assisting a physician in establishing a diagnosis. In this work we are interested in interpreting segmentation networks by appraising their sensitivity to high-level features. Indeed, segmenting anatomical structures in medical images is one of the tasks that hugely benefited from Convolutional Neural Networks (CNNs), to the point that this framework is now state-of-the-art in most segmentation tasks [5,6,8].

Research on interpretable Deep Learning has been very active for a few years now. Thorough reviews [1,7] extensively describe the field, among which so-called saliency methods are especially popular [4,14,16,17]. The understanding of these methods has grown recently, with some works examining their limitations [11,18]. More generally, saliency methods address the problem of *feature attribution* which, in the case of a segmentation network, boils down to pixel attribution and is thus of limited value.

Another class of interpretability methods consists in visualizing patterns that activate a particular neuron in the network. Most of them consist in maximizing

K. Suzuki et al. (Eds.): ML-CDS 2019/IMIMIC 2019, LNCS 11797, pp. 56–63, 2019.
https://doi.org/10.1007/978-3-030-33850-3_7

the activation in the input space [13,17,19,20]. These visualizations are insightful when the network is trained on natural images, as they generate natural structures and appearances, but they are harder to interpret on medical images.

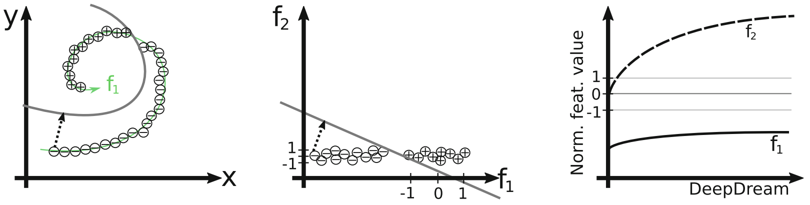


Fig. 1. Illustration of the method with a 2-dimensional classifier. Left: input space, ⊕ and ⊖ are resp. positive and negative samples; the classification function is the grey line; the data is described by features f_1 (green arrow) and f_2 (orthogonal to f_1). Middle: features space; features are normalized w.r.t. the set of positive samples. Left and middle: the path of steepest slope (or *DeepDream path*) is represented as a dotted arrow. Right: projection of this path on f_1 and f_2 (DeepDream analysis). (Color figure online)

The method in [10] is closer to our motivation, *i.e.* to analyze the influence of human-understandable features on the output of a network. Using abstract concepts defined by sets of example images is appealing, especially for complex concepts that would be difficult to model. But this transfers the burden to the creation of concept-labelled databases, which can be challenging in medical imaging. On the other hand, image domain features such as radiomic features can be used to directly evaluate relevant concepts in medical images when a segmentation mask is available, and seems therefore well suited to the interpretation of segmentation networks.

We detail our method in Sect. 2, starting by giving an intuitive definition of what the *sensitivity* and *robustness* to a feature might be for a network. Then we describe our method based on activation maximization to highlight features that the network is sensitive to (Sect. 2.2). We show in a controlled setting that the method correctly assesses the robustness of a network to a specific feature. Other experiments show how we can get insights about what a network has learned using our method (Sect. 3).

2 Method

2.1 Overview

Segmentation networks achieve state-of-the-art performance on most segmentation tasks. They can extract complex features at multiple scales and successfully perform challenging segmentation tasks where modeling approaches using handcrafted features would have failed. To interpret this complex decision function, we want to determine how *sensitive* or *robust* a neural network is to a set of

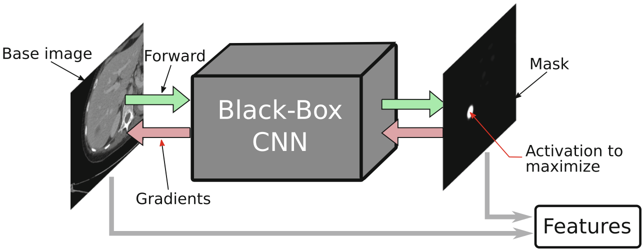


Fig. 2. Representation of an iteration as described in Sect. 2.2: the current image is forwarded in a segmentation CNN. We retrieve the output map and the gradient w.r.t. an arbitrary neuron activation from the output map. We compute the features from the image and segmentation mask and update the image following the gradient for the next iteration.

high-level features $\{f_k\}_{1 \leq k \leq K}$, such as the size of the object, statistics on its intensity distribution or its shape.

We consider that a network is sensitive to a feature f_k if its alteration impacts the network decision. Conversely, we say that the network is robust - or indifferent - to a feature if it is not sensitive to it. However a feature of an object cannot in general be modified without modifying others characteristics, therefore such properties cannot be directly evaluated. Starting from a baseline producing a negative response, we can find a minimal alteration that produces a positive response by following the path of steepest slope in the input space (the arrow in Fig. 1), using the network gradients. This procedure is similar to activation maximization, also known as *DeepDream* [15]. If the features f_k are smooth functions, we can assume that the path of steepest slope in the input space will favor features to which the network is the most sensitive.

In Fig. 1 we provide a schematic view of this process in two dimensions. Intuitively, a network should be indifferent to a feature that is useless (here f_2) for characterizing an object, and sensitive to a feature that is essential (here f_1).

2.2 Algorithm

Being given a trained binary segmentation network S of any architecture, we compute the *DeepDream analysis* with an iterative algorithm, illustrated in Fig. 2. It starts from an image X_0 with no foreground (an image with no lesion in the case of lesion segmentation for instance), and pick a neuron i we want to maximize. At each iteration j and until convergence:

- We forward the image X_j through the network and retrieve the segmentation mask $M_j = S(X_j)$, as well the gradient of the neuron activation $\frac{\partial i}{\partial X}$.
- We update the image for the next iteration $X_{j+1} = X_j + \alpha \frac{\partial i}{\partial X}$.
- We compute features $f_k(X_j, M_j)$.

The output is a plot of the curves $j \rightarrow f_k(X_j, M_j)$. These curves can be interpreted to assess the sensitivity of the network to those features.

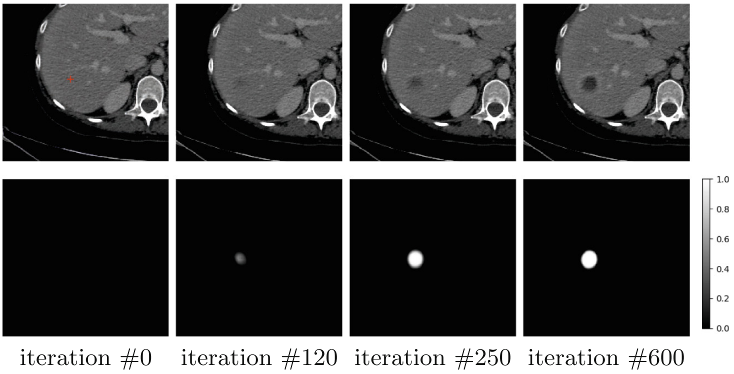


Fig. 3. Different steps of gradient ascent performed on a CT slice showing a healthy liver, with a network trained to segment liver tumors from CT slices. The top row shows the image being "DeepDreamed", while the bottom row shows the output of the network (high probabilities of a pixel being part of a tumor are white, low probabilities are black). The red cross on the leftmost image shows the pixel maximized by gradient ascent. We observe that a responding area appears during the procedure. (Color figure online)

This procedure, derived from activation maximization also known as DeepDream, has been shown to work on many classification network architectures [13,17,19,20] and we found that it was easily applicable on several segmentation architectures. Figure 3 shows how the image and segmentation mask respond to the activation maximization.

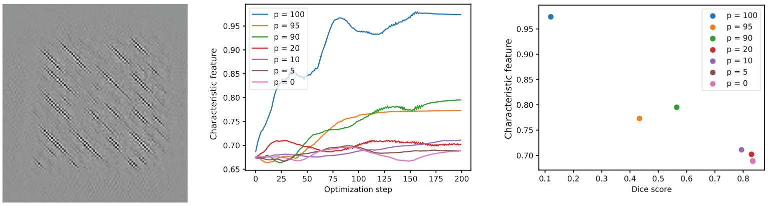
Although any kind of features can be used, we chose to use radiomic features as they are specifically designed to characterize segmented tissues in medical images [2,9,21], and have shown to capture enough information for Computer-Aided Diagnosis [3,9].

Our DeepDream analysis consists in computing a set of features $f_k(X_j, M_j)$ at each step j of the DeepDream path. As activation maximization produces small changes in input but decisive changes in output, we expect the features to be tweaked according to the sensitivity of the network to those features. To interpret the evolution of feature values observed during the DeepDream analysis, we normalize a particular feature with respect to the distribution of this feature computed on the validation dataset used during training.

3 Experiments

We conduct three experiments to assess the potential of a DeepDream analysis to interpret a segmentation network. We show that the sensitivity computed from the DeepDream analysis is associated with the performance of the network, as expected (Sect. 3.1). The second experiment shows how our method highlights the difference of sensitivities between networks trained on different databases

(Sect. 3.2). Finally we show what kind of insight we can get with our method by applying it to the the real-world use case of liver tumor segmentation (Sect. 3.3).



(a) (b) (c)

Fig. 4. Controlled experiment. We trained 7 networks with different probabilities p of marking the positively labeled zones. (a) DeepDream of the network for $p = 100\%$. (b) Evolution of the characteristic feature during the gradient ascent process. (c) Dice score on the unmarked test set and characteristic feature at the end of the gradient ascent process for the 7 networks. Networks that performed poorly on the unmarked test set and thus relied on the marking showed a high characteristic feature in their dream.

For all experiments we use basic contracting-expanding architectures with 3×3 convolutions, max-pooling or up-convolution every 2 convolution layers and number of filters doubling at each level, trained with an Adam optimizer until convergence.

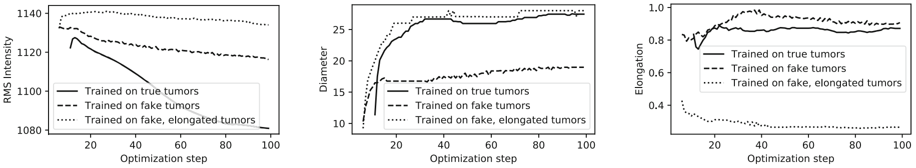
3.1 DeepDream Sensitivity and Segmentation Performance

In order to get a setting where the actual sensitivity to a feature is known, we ran the following experiment: For *cat* and *dog* classes from the COCO database [12], each image is augmented with a marking with a probability p. We chose a synthetic texture made of 135° line segments of random positions and intensities as the marking. Then, for different values of p, we train several networks G_p to segment cats and dogs on this training dataset.

A simple, intuitive way to assess the robustness of a network with respect to the marking is then to compute its score on a test dataset with no marking. Given the score of G_0 as the baseline, a similar score indicates that a network is robust to the marking.

We assess the presence of the marking in any DeepDream generated as described in Sect. 2.2 by computing the maximum response of the convolution of the dream with a 135° line segment. We call this feature the *characteristic feature* of the marking. Starting from the same realization of white noise, we then compute the characteristic feature at each optimization step, for all networks G_p. Results are illustrated in Fig. 4.

Networks reaching a Dice score close to the baseline ($p \leq 20\%$) did not see the characteristic feature evolve during DeepDream, in contrast to those which relied on the marking ($p \geq 90\%$). This shows that we are able to correctly assess the sensitivity of a network to a particular feature by analyzing its DeepDreams.



(a) (b) (c)

Fig. 5. DeepDream analysis of 3 networks trained on different datasets. (a) Root mean squared intensity along the DeepDream Path. (b) Maximum diameter of the dreamed tumor. (c) Elongation (1 means round, and 0 means elongated in the standard definition of elgontation in radiomics.)

3.2 Sensitivity to Intensity and Shape Features

In the LiTS database[1], tumors appear as hypointense areas in the liver parenchyma. In this experiment we compare a network trained on real tumors to a network trained on synthetic tumors, to test how our method highlights the differences of two networks trained on seemingly similar tasks.

We generate synthetic tumors by lowering the intensities in random areas of healthy livers. The DeepDream analysis shows that the network trained on real tumors is more sensitive to low intensities in the liver (Fig. 5a) than the network trained on synthetic tumors. This indicates that the synthetic network focuses on other features than the intensity.

To determine if the DeepDream analysis is also able to assess the sensitivity to shape features, we train a network to segment only synthetic elongated tumors, as opposed to the overall round shape of real tumors, as observed in clinical environments. We observe that the network trained on elongated tumors is indeed more sensitive to elongation (Fig. 5c).

3.3 Analysis of a Tumor Segmentation Network

To illustrate how one can use DeepDream analysis with radiomic features, we analyze a network trained to segment liver tumors in CT scans. We visualize the evolution of 6 relevant radiomic features, normalized so that 0 is the mean value of the feature computed on the validation dataset, and 1 is one standard deviation above the mean (Fig. 6).

The values of intensity and sphericity quickly evolve towards the normal range, indicating that the network is sensitive to both features. By contrast, the Grey-Level Co-occurrence Matrix (GLCM) Contrast, a texture feature that measures intensity disparity among neighboring pixels, as well as the entropy of the intensities distribution, stay below the normal range, indicating that the network is robust to heterogeneity. This is coherent with our intuition that the network should react to flat hypointense areas in the liver, without significant texture

[1] https://competitions.codalab.org/competitions/17094.

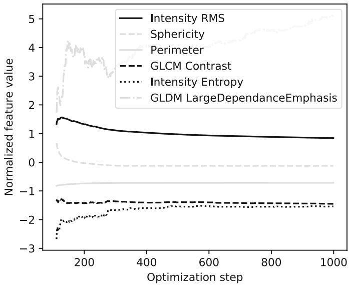


Fig. 6. Evolution of features along the DeepDream path of a liver tumors segmentation network, starting from a healthy liver. Images and masks are shown in Fig. 3.

information. However we also notice that the value of the Large Dependence Emphasis feature goes rapidly and strongly out of normal range, suggesting a lack of robustness to this feature.

4 Conclusion

In this paper, we proposed a new approach to interpret segmentation networks. We generate and analyze fake positive objects using a gradient ascent method. This provides insights on the sensitivity and robustness of the trained network to specific high-level features.

Future work will focus on formulating theoretically grounded definitions of sensitivity and robustness and on providing theoretical guarantees that DeepDream primarily modifies the most sensitive features. Other state-of-the-art segmentation architectures (such as U-Nets, DeepLab or PSPNet) will also be tested, as well as multiclass segmentation networks.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Aerts, H.J.W.L., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**, 4006 (2014)
3. Avanzo, M., Stancanello, J., Naqa, I.M.E.: Beyond imaging: the promise of radiomics. Phys. Med. Eur. J. Med. Phys. **38**, 122–139 (2017)
4. Bach, S., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS ONE **10**(7), e0130140 (2015)
5. Christ, P.F., et al.: Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. CoRR abs/1702.05970 (2017)
6. Couteaux, V., et al.: Kidney cortex segmentation in 2D CT with U-Nets ensemble aggregation. Diagn. Intervent. Imaging **100**, 211–217 (2019)
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

8. Erden, B., Gamboa, N., Wood, S.: 3D convolutional neural network for brain tumor segmentation. Computer Science, Stanford University, USA, Technical report (2017)
9. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2015)
10. Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: ICML (2018)
11. Kindermans, P.J., et al.: The (un) reliability of saliency methods. arXiv preprint arXiv:1711.00867 (2017)
12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. Int. J. Comput. Vision **120**(3), 233–255 (2016)
14. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
15. Mordvintsev, A., Olah, C., Tyka, M.: Inceptionism: Going deeper into neural networks. Google Research Blog (2015)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": explaining the predictions of any classifier. In: HLT-NAACL Demos (2016)
17. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
18. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D., Ravikumar, P.: How sensitive are sensitivity-based explanations? arXiv preprint arXiv:1901.09392 (2019)
19. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
20. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
21. Zwanenburg, A., Leger, S., Vallières, M., Löck, S., et al.: Image biomarker standardisation initiative. arXiv preprint arXiv:1612.07003 (2016)