Explainability for regression CNN in fetal head circumference estimation from ultrasound images

Jing Zhang, Caroline Petitjean, Florian Yger, Samia Ainouz

Explainability for regression CNN in fetal head circumference estimation from ultrasound images

Jing Zhang[1], Caroline Petitjean[1], Florian Yger[2], and Samia Ainouz[1]

[1] Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS, Rouen, France
{jing.zhang,samia.ainouz}@insa-rouen.fr,caroline.petitjean@univ-rouen.fr
[2] LAMSADE, Université Paris-Dauphine, France florian.yger@dauphine.fr

Abstract. The measurement of fetal head circumference (HC) is performed throughout the pregnancy to monitor fetus growth using ultrasound (US) images. Recently, methods that directly predict biometric from images, instead of resorting to segmentation, have emerged. In our previous work, we have proposed such method, based on a regression convolutional neural network (CNN). If deep learning methods are the gold standard in most image processing tasks, they are often considered as black boxes and fails to provide interpretable decisions. In this paper, we investigate various saliency maps methods, to leverage their ability at explaining the predicted value of the regression CNN. Since saliency maps methods have been developed for classification CNN mostly, we provide an interpretation for regression saliency maps, as well as an adaptation of a perturbation-based quantitative evaluation of explanations methods. Results obtained on a public dataset of ultrasound images show that some saliency maps indeed exhibit the head contour as the most relevant features to assess the head circumference and also that the map quality depends on the backbone architecture and whether the prediction error is low or high.

Keywords: Saliency maps · Explanation evaluation · regression CNN · biometric prediction · medical imaging

1 Introduction

The measurement of fetal head circumference (HC) is performed throughout the pregnancy as a key biometric to monitor fetus growth and estimate gestational age. In clinical routine, this measurement is performed on ultrasound (US) images (Fig. 1), via manually tracing the skull contour and fitting it into an ellipse. Automated segmentation approaches have been proposed, lately based on CNN in order to solve this tedious task, but these models require large dataset of manually segmented data. In our previous work [21], we departed from the mainstream approach of segmentation and instead proposed a regression network, in order to directly predict the head circumference.

Fig. 1. Ultrasound images of fetus head with head circumference in millimeters

Compared to a classification model, the last layer of a regression CNN model is a linear or sigmoid activation function, instead of the softmax layer. Also, the regression loss function is metric-inspired, for instance, it can be the Mean Absolute Error (MAE) or the Mean Squared Error (MSE). It is known that the high accuracy of deep learning methods comes at the cost of a low interpretability, i.e. the model is seen as a black box, which does not provide explanations along with the prediction. In this paper, our goal is to investigate how explanation methods can help us to get some insights on the regression network and to appreciate its behavior [12]. In classification networks, explanations may take the form of saliency or sensitivity maps [10], highlighting the areas that particularly contributed to a decision. The saliency maps have been applied on different neural networks such as CNN, LSTM, and in various tasks, for example classification, detection and image segmentation [16]. To the best of our knowledge, this paper is the first interpretation of a regression CNN that is dedicated to the estimation of biometric from medical images.

In this paper, our contributions are the following: we adapt explanation methods in regression CNN and provide an interpretation of what a saliency map is, in the regression case. We are thus able to gain insight into the CNN regression model for our HC prediction problem, and see what pixels contribute the most to the estimation of the HC: we expect them to be those of the head contour. We also address the problem of evaluating the explanation methods, in the regression case. Adebayo's sanity checks consist in performing randomization tests, in the data or in the model, and evaluate the changes in the produced saliency maps [1]. Another example is Samek's proposal, that has particularly inspired us [11], to compare and assess different explanation methods. The principle is to inject noise gradually in the image, in locations that have been highlighted by the saliency maps, and see how the prediction is affected by this perturbation. However, the method is designed for classification networks and requires some adaptation.

In Section 2 we briefly recall the state-of-the-art in saliency maps algorithm for classification CNN and their meaning in case of a regression network; we also presented the evaluation methodology used to assess the explanation methods. Experimental results are presented in Section 3 and conclusions are drawn in Section 4.

2 Saliency map methods for regression CNN

In this section, we briefly describe 8 explanation methods from the state-of-the-art that are used to produce saliency maps in classification CNN. [12, 22, 16]. Then, we present the evaluation method of perturbation analysis [11] and adapt it to the regression CNN to evaluate the performance of these methods.

2.1 State-of-the-art saliency maps in CNN

Two categories of saliency maps are generally considered, perturbations-based or propagation-based. In perturbation-based approaches, the goal is to estimate how perturbation applied to the input image, such as blurring or injecting noise, changes the predicted class [5, 22]. In propagation-based techniques, the idea is to backpropagate a relevance signal from the output to the input. In this paper, we will focus on the latter category of methods that actually encompass (i) sensitivity (or gradient-based) analysis, (ii) deconvolution methods, and (iii) Layer-wise Relevance Propagation (LRP) variants.

The sensitivity analysers include the **Gradient** [14] method, that simply computes the gradient of the output w.r.t. input image, and expresses how much the output value changes w.r.t. a small change in input; the **SmoothGrad** [17], that averages the gradient over random samples in a neighborhood of the input with added noise, and which is an improvement of Gradient method that can sharpen the saliency map; the **Input*Gradient** [13] technique, that strengthens the saliency map by multiplying Gradient with input information; and the **Integrated Gradients** [19], that computes the integration of the gradient along a path from the input to a baseline black image.

Deconvolution methods are the **DeConvNet** [20] that acts equivalently as a decoder of CNN models, which reverses the CNN layers, and the **Guided BackProp** [18] that combines backpropagation and DeConvNet.

The core idea of **Layer-wise Relevance Propagation (LRP)** [3] is to compute a relevance score for each input pixel layer by layer in backward direction. It first forward-passes the image so as to collect activation maps and backpropagates the error taking into account the network weights and activations. The **DeepTaylor** [9] method identifies the contribution of input features as the first-order of a Taylor expansion, through Taylor decomposition, then it can estimate the attribution of each neuron one by one.

In the classification setting, a saliency map provides an estimation of how much each pixel contributes to the class prediction. In the regression setting, the saliency map will provide an estimation of how much each pixel is impacting the model, and is contributing to decrease the prediction error, as measured by the loss function, that is in general the MAE or MSE.

2.2 Evaluation of explanation methods based on perturbation

Explanation methods (also called analyzers) perform differently depending on the model, the task at hand, the data, etc. In order to quantitatively evaluate

those analyzers, we build upon the perturbation analysis of [11], originally designed to assess explainability methods in classification networks. Let us first describe the perturbation process and then the evaluation metric.

First, the input image to be analyzed is subsampled by a grid. Each subwindow of the grid is ranked according to its importance w.r.t. to the pixel-wise saliency scores assigned by the analyzers. Then, the information content of the image is gradually corrupted by adding perturbation (Gaussian noise) to each subwindow, starting with the most relevant subwindow, w.r.t. the ranking just mentioned. The effect of this perturbation on the model performance is measured with the prediction error. This procedure is repeated for each subwindow. Generally, the accuracy of model will drop quickly when important information is removed and remains largely unaffected when perturbing unimportant regions. Thus, the analyzers can be compared by measuring how quickly their performance drops. That is to say, the quicker the model performance drops after introducing perturbation, the better the analyzer is capable of identifying the input components responsible for the output of the model.

The quantitative evaluation proposed in [11] for classification network, consists in computing the difference between the score $f(x)$ indicating the certainty of the presence of an object in the image x, in the presence and in the absence of perturbation. This difference is called Area over Perturbation Curve (AOPC) and defined more precisely defined in in [11] as:

$$\mathrm{AOPC}_{Analyzer} = \frac{1}{N}\sum_{n=0}^{N}(f(x_n)^{(0)} - \frac{1}{K}\sum_{k=0}^{K} f(x_n)^{(k)}) \tag{1}$$

where N is the number of images, K is the number of perturbation steps, x is the input image.

Here, we propose to adapt the AOPC to the regression case, and if we denote by $\epsilon(x)^{(0)}$ the prediction error of initial image evaluated by the analyzer and $\epsilon(x_n)^{(k)}(1 \leq k \leq K)$ the prediction error of the perturbed image $(x_n)^{(k)}$ at step k, we can define the $\mathrm{AOPC}^{\mathrm{regression}}_{\mathrm{Analyzer}}$ as:

$$\mathrm{AOPC}^{regression}_{Analyzer} = \frac{1}{N}\sum_{n=0}^{N}(\epsilon(x_n)^{(0)} - \frac{1}{K}\sum_{k=0}^{K} \epsilon(x_n)^{(k)}) \tag{2}$$

A larger AOPC in absolute value means that an analyzer has a steep decrease while the perturbation steps is increasing.

3 Experiments

3.1 Experimental setup

We analyse two regression models that we proposed in our previous work [21], namely the regression ResNet50 and regression VGG16 (implemented using Keras). As their names show, the backbone architectures are ResNet50 [6] and

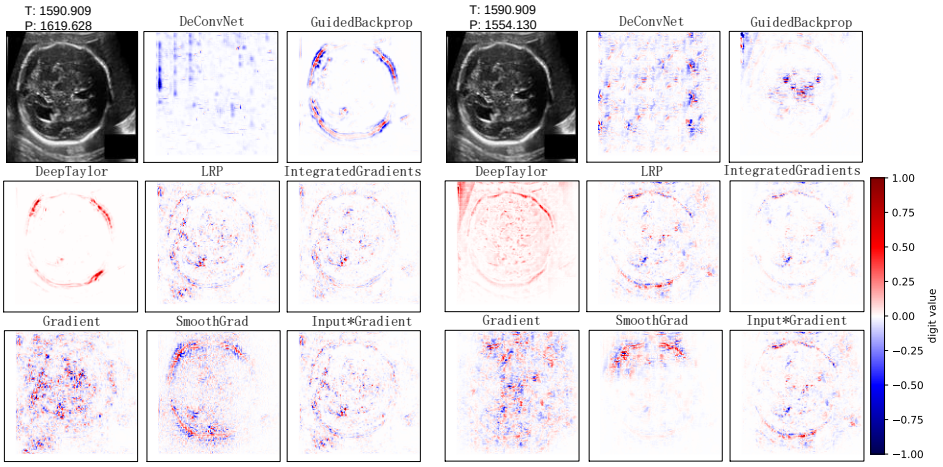
VGG16 [15] resp., and the loss is the mean absolute error. Both models are pre-trained on ImageNet; subsequently the last (softmax) layer is replaced by a linear layer and the network is fully retrained on a public dataset of ultrasound fetal head images called HC18 [7]. The HC18 dataset contains 999 US images, along with the corresponding head circumference, that we randomly split into a training (600), a validation (200) and a test set (199). We augment the data of the training set to 1800 images, and perform resizing of the images to the size 128×128 pixels. With a 5-fold cross validation, the mean absolute errors (MAE) that we obtained on the test set were 37.34 ± 37.46 pixels (4.78 ± 4.41 mm) in reg-ResNet50 and 40.17 ± 40.99 pixels (5.46 ± 5.99 mm) in reg-VGG16.

In the following, we will compute the saliency maps on the test set images. We first show the saliency maps of various explanation methods for our regression problem, for both architectures Reg-ResNet50 and Reg-VGG16, the quantitative evaluation of explanation methods, and a more in-depth study of prediction results, with the best ranked methods, namely Input*Gradient and LRP. We have used the iNNvestigate toolbox to perform our experiments [2].

Visualization of explanation methods We visualize the saliency maps provided by the 8 selected explanation methods in Fig.2. From these images, we can barely see the features retrieved by explanation method DeConvNet and Gradient in both models, that is to say these two methods seem somehow insensitive to the models. This may be explained by the gradient shattering problem [4] for the gradient method. Regarding DeConvNet's saliency map, it may be due to the the architecture of deconvolution network which reconstructs the convolution networks reversely. In addition, for Reg-ResNet50, methods Gradient, GuidedBackprop and SmoothGrad fail to highlight the head contour. We will see that these observations are confirmed by the quantitative evaluation.

Quantitative evaluation of explanation methods based on perturbation Here, we compare the explanation methods through perturbation analysis. In this experiment, the input image of size 128×128 pixels is divided into a grid of 4×4 subwindows of size 32×32 pixels. Gaussian noise with mean value 0 and standard deviation 0.3 is added to each subwindow, according to their importance assigned by analyzers during the 16 steps. Fig. 3 is an example of the perturbation process of Gradient analyzer.

In Fig. 4, we show the evolution of the prediction error w.r.t. the quantity of noise added at each perturbation steps, on first the most significant subwindow in the analyzer's sense, to the least significant one. One can observe that consistently, the prediction error is increasing, as the level of noise increases. Methods with the steepest curve, LRP and Input*gradient, exhibit the largest sensitivity to perturbations, and as such, should highlight the contributing pixels, in the sense of this criterion. Interestingly the Integrated gradient analyzer seems to be relevant for VGG16, but not for Reg-ResNet50. In the future, it will be interesting to vary the subwindow size to see if results are affected. We expect that a finer grid will be better suited to a thin structure like the head skull.



(a) Reg-VGG16 (b) Reg-ResNet50

Fig. 2. Comparison of different saliency maps with Reg-VGG16 and Reg-ResNet50. P: predicted HC value, T: ground truth HC value (in pixels).

Table 1. Performance (AOPC scores) of different analysis methods after perturbation, with two regression models. G: Gradient, SG: SmoothGrad, DCN: DeConvNet, DT: DeepTaylor, GB: GuidedBackprop, I*G: Input*Gradient, IG: IntegratedGradients. Lower is better. Best scores in bold.

Model	G	SG	DCN	DT	GB	I*G	IG	LRP
Reg_VGG16	-7.312	-7.398	-2.869	-7.401	-1.663	**-9.189**	**-9.490**	**-9.175**
Reg_ResNet50	-11.533	-11.841	-9.249	-9.890	-9.717	**-14.748**	-5.603	**-14.577**

In Table 1, we compared AOPC scores on regression VGG16 and regression ResNet50 models respectively. Since the AOPC is the difference between the prediction error with and without perturbation, we expect that the analyzer that are indeed perturbed by the noise will return a large AOPC score, in absolute value. We can see that the regression ResNet50 has higher AOPC score than regression VGG16 model. Again we can gather from this table that both the LRP and Input*Gradient methods perform well in those two models.

Note that other explanation methods have inconsistent performance depending on the model. This highlights the necessity to choose the proper explanation method before analyzing a specific model.

Comparison of regression models As shown in Fig 2, both regression VGG16 and regression ResNet50 are successful in learning the features from ultrasound images to assess the HC. From Table 1, we can gather that the regression ResNet50 has slight better performance on the whole, since AOPC values are larger in absolute value.

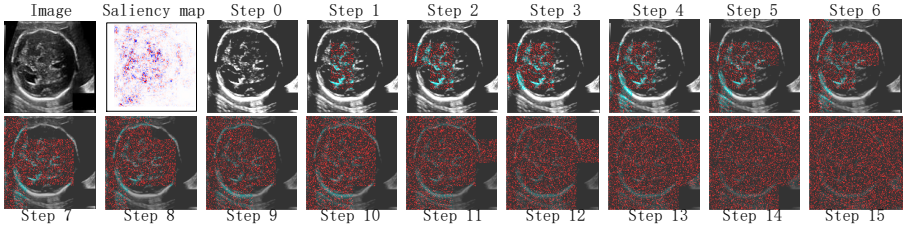
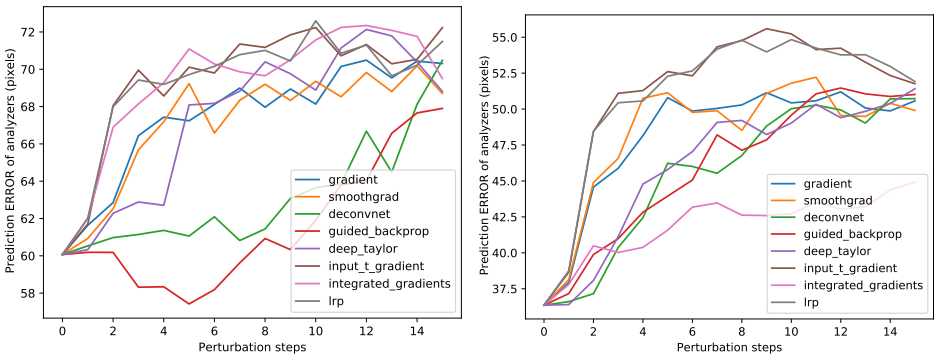


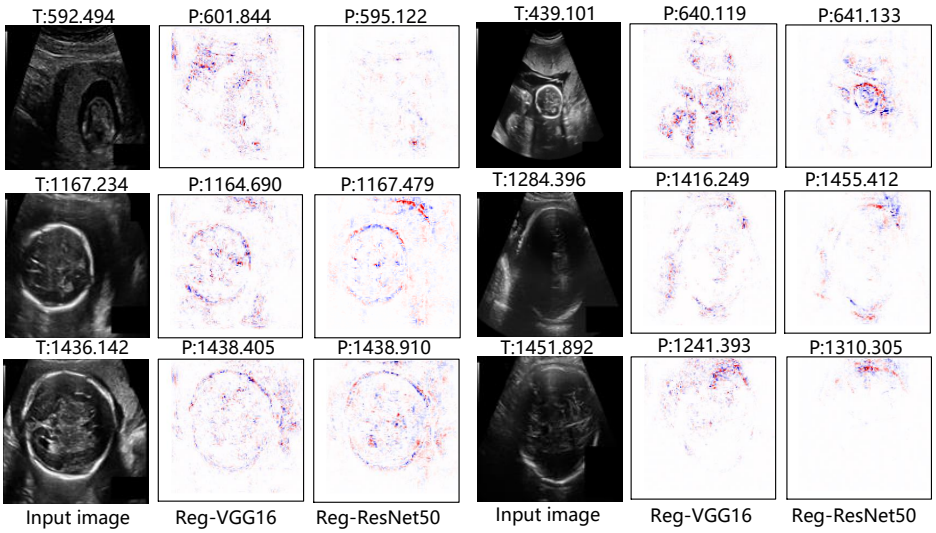
Fig. 3. Perturbation process for the saliency map produced by the Gradient method. Step 0 is the original input image. From step 1 to step 15, Gaussian noise is added gradually on the image subwindows. The perturbation order of these subwindows corresponds to the saliency scores assigned by the Gradient method analysis, i.e. the most contributing pixels are perturbed first. Red: noise, blue: original image pixels.



(a) Reg-VGG16 (b) Reg-ResNet50

Fig. 4. Prediction error (in pixels) of different analyzers during each perturbation step based on Regression VGG16 and Regression ResNet50 model. The horizontal axis is the perturbation steps.

Comparison of saliency maps for correct vs incorrect prediction In this experiment, we arbitrarily pick one of the best performing methods from the previous results, and thus the use Input*Gradient explanation method to generate saliency maps from images with small prediction error (Fig. 5 (a)), and with large prediction error (Fig. 5 (b)). We can see that the well predicted images have obvious head contour, at least in the 2 last rows of Fig 5 (a). The models are able to learn the features from these images, therefore the saliency maps show key features. However, it is not always the case: the first row shows a small prediction error, and the head contour are not specifically highlighted. For the badly predicted images, the saliency maps highlight features that are spread and not localized into meaningful segments. The models can not learn the features from these images.



(a) Good prediction (b) Bad prediction

Fig. 5. Saliency map of Reg-VGG16 and Reg-ResNet50 with Input*Gradient explanation method. P and T: resp. predicted and ground truth HC values (pixels).

4 Conclusion

Understanding whether the model can learn the relevant features in images and take the right decision is crucial in the medical domain. Whereas there have been a wealth of works in classification networks, there is a void for interpreting regression networks. In this paper, we address the problem of estimating the head circumference in fetal head directly from US images. We use several post-hoc explanation techniques that produce saliency maps and adapt a perturbation based quantitative evaluation method, to assess the relevance of the saliency maps. The experimental results proved that the regression CNN models are able to learn the key features from the input ultrasound fetus images, and in particular, the head circumference. One finding is that for this application, Gradient and DeConvNet method are particularly insensitive to different CNN models or data, and that ResNet50 seem to have better learnt the head features. Thus so far, we have extended the model property from classification to regression and explored a specific regression task. Future works also include investigating the explainability of other regression losses: in this paper, we used the MAE, but the mean square error or the Huber loss are alternatives, and there is no heuristics yet to decide which loss is better [8]. This will allow us to adapt or design new loss functions, that can account for an enhanced learnability of the regression CNN, to further improve the HC prediction. In addition to investigate individual image-wise explanations, we also intend to explore the generation of meta-explanations by aggregating individual explanations, to gain additional insight into the model behavior. Other regression applications will also be interesting to explore.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems. pp. 9505–9515 (2018)
2. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K., Dähne, S., Kindermans, P.: innvestigate neural networks! CoRR **abs/1808.04260** (2018), http://arxiv.org/abs/1808.04260
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7) (2015)
4. Balduzzi, D., McWilliams, B., Butler-Yeoman, T.: Neural taylor approximations: Convergence and exploration in rectifier networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 351–360. JMLR. org (2017)
5. Fong, R., Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 3449–3457 (Oct 2017). https://doi.org/10.1109/ICCV.2017.371, http://arxiv.org/abs/1704.03296, arXiv: 1704.03296
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE CVPR. pp. 770–778 (2016)
7. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PLOS ONE **13**(8), 1–20 (08 2018), https://doi.org/10.1371/journal.pone.0200412
8. Lathuilière, S., Mesejo, P., Alameda-Pineda, X., Horaud, R.: A comprehensive analysis of deep regression. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 1–17 (2019)
9. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
10. Morch, N.J., Kjems, U., Hansen, L.K., Svarer, C., Law, I., Lautrup, B., Strother, S., Rehm, K.: Visualization of neural networks using saliency maps. In: Proceedings of IEEE International Conference on Neural Networks. vol. 4, pp. 2085–2090 (1995)
11. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
12. Samek, W., Müller, K.R.: Towards explainable artificial intelligence. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 5–22. Springer (2019)
13. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. CoRR **abs/1605.01713** (2016), http://arxiv.org/abs/1605.01713
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2014)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
16. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. Journal of Imaging **6**(6:52) (2020)

17. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. In: Workshop on Visualization for Deep Learning, ICML (2017)
18. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015), http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a
19. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3319–3328. JMLR. org (2017)
20. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
21. Zhang, J., Petitjean, C., Lopez, P., Ainouz, S.: Direct estimation of fetal head circumference from ultrasound images based on regression cnn. In: Medical Imaging with Deep Learning (2020)
22. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis (2017), http://eprints.gla.ac.uk/214152/