Interpretability vs. Complexity: The Friction in Deep Neural Networks

José P. Amorim
IPO-Porto Research Centre and CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
jpamorim@dei.uc.pt

Pedro H. Abreu
CISUC, Department of Informatics Engineering University of Coimbra, Portugal
pha@dei.uc.pt

Mauricio Reyes
Medical Image Analysis University of Bern, Switzerland
mauricio.reyes@med.unibe.ch

João Santos
IPO-Porto Research Centre
Porto, Portugal
joao.santos@ipoporto.min-saude.pt

***Abstract*—Saliency maps have been used as one possibility to interpret deep neural networks. This method estimates the relevance of each pixel in the image classification, with higher values representing pixels which contribute positively to classification.**

The goal of this study is to understand how the complexity of the network affects the interpretabilty of the saliency maps in classification tasks. To achieve that, we investigate how changes in the regularization affects the saliency maps produced, and their fidelity to the overall classification process of the network.

The experimental setup consists in the calculation of the fidelity of five saliency map methods that were compare, applying them to models trained on the CIFAR-10 dataset, using different levels of weight decay on some or all the layers.

Achieved results show that models with lower regularization are statistically (significance of 5%) more interpretable than the other models. Also, regularization applied only to the higher convolutional layers or fully-connected layers produce saliency maps with more fidelity.

***Index Terms*—Convolutional neural network, interpretability, complexity, saliency map**

I. Introduction

As neural networks grow in complexity [1] their capacity to learn mappings from the input data to the classification label increases. Explanations are provided to understand this mapping and the predictions made by the network. Different post-hoc explanations [2] have been proposed, from prototype explanations [3], to local approximations [4]. One of the proposed explanations are called saliency maps, which produce an estimation of each pixels' relevance in the overall prediction of the network for each input image. While small relevance scores correspond to pixels which do not contribute to the classification of the image, higher values correspond to the pixels which contribute the most to the prediction. There are many saliency maps methods [5]–[8] which give different estimations of the relevance scores. In order to quantitatively evaluate if a given relevance score is suitable, we need to assess if the method truly discriminates between relevant and irrelevant pixels. Due to the non-existence of ground-truth relevance scores, the quantitative evaluation of saliency maps poses a demanding problem. Fidelity [9] is a concept which determines how well a relevance score agrees with how the model works. **However, a direction that remains to be explored is if it exists a relationship between the fidelity of saliency maps and the complexity of the network.**

With the increase of computational power and data available to train deep neural networks we have seen a rapid increase of network complexity resulting on a increase capacity to describe the data. But, only increasing the capacity of the networks results in the overfitting of the model to the training data, resulting on a low test set performance. For this reason, regularization approaches are used in order to decrease the capacity of the network to fit the data and avoid overfitting. Some of these methods, such as L^2 regularization, reduce the complexity of the network, limiting the values of the network parameters. **However, it is not yet understood what is the impact that certain changes of the network parameters has on the different saliency map methods and consequently in its interpretation.**

In order to address the two open research problems previously identified, the goal of this work consists on finding the relationship between the complexity of the network and the quality of its explanations.

In order to achieve that, we will try to answer the following questions:

- How does the regularization of the deep neural network affect the quality of saliency maps?
- Does the layer regularized affect the interpretability of the network? And if it does, how should one choose to regularize the network in order to obtain greater interpretability?

In this work, we measured the interpretability of different convolutional neural networks (CNNs) trained on CIFAR-10 dataset [10] with different L^2 regularization values (0, 0.0001,

This work was supported in part by the FCT Research Grant SFRH/BD/136786/2018 and the project NORTE-01-0145-FEDER-000027, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

978-1-7281-6926-2/20/$31.00

0.001, 0.005, 0.01, 0.05). We compared the quality of saliency maps produced by five different saliency map methods. The methods chosen were gradient [5], DeConvNet [6], Guided Backpropagation [5], Deep Taylor Decomposition [8] and Layer-wise Relevance Propagation [7] and are described in the following section. In order to answer the second research question we also applied regularization on different layers and measure its impact on interpretability.

Results of the experiments show that regularization does affect the quality of saliency maps, as models trained with lower regularization show higher interpretability than models trained with high regularization. Also, based on results from experiments performed with different regularization values based on layers, it was possible to conclude that regularization applied only to the higher convolutional layers or fully-connected layers produce saliency maps with more fidelity.

To the best of the authors' knowledge there isn't any study which investigates the relationship between the complexity and the interpretability of a neural network.

The paper is organised as follows: first, Section II introduces important concepts such as capacity and complexity, as well as related work on saliency maps methods and metrics. Then, Section III describes the experiments, including the data, models and metrics used. Section IV presents and discusses the experimental results. Finally, Section V summarizes conclusions of the paper as well as promising research directions.

II. Background Knowledge

The capacity of a model can be seen as its ability to fit a wide variety of functions [11] and it has an impact in whether it underfits or overfits the data. When a model underfits the data, it is unable to reduce the training set error, while when it overfits the data, it is able to reduce the training set error but not the test set error. By increasing the model's capacity, the model is able also to memorizing properties of the training data that do not serve to the test data.

The capacity of the deep neural network can be controlled by varying their depth and breadth [1]. The complexity of the network, which can be measured in different ways from its depth to the number of connections, is related to the capacity of the network to learn from the data. The question now becomes, what is the relationship between the complexity of the network and the quality of its explanations.

The problem which was here presented involve three aspects which will be described in this section, that includes the regularization of the network, the saliency map methods and the interpretability metrics.

A. Regularization

Regularization is a modification to a learning algorithm intended to reduce its generalization error [11]. Many regularization approaches limit the capacity of models, and its complexity, by adding a parameter norm penalty.

a) L^2 Regularization: commonly known as weight decay, it's a regularization strategy that drives the weights closer to zero, by adding a regularization term to the objective function:

$$\Omega(\theta) = \frac{1}{2}\|w\|_2^2 \tag{1}$$

Components of the weight vector corresponding to directions that do not contribute to reducing the objective function are decayed away through the use of the regularization throughout training [11].

b) Early Stopping: When training models with sufficient capacity to overfit the task, it is common to see that after constant decrease in the training error and validation error, the validation error often starts to rise. Early stopping works by keeping a copy of the model parameters every time the validation error improves, in order to return the setting when the validation error was the lowest.

B. Saliency Map Methods

In our formal description, an *input* corresponds to an image and is represented by a tensor $x \in \mathbb{R}^d$. A model describes a function $f : \mathbb{R}^d \to \mathbb{R}^c$, which maps the d-dimensional images to a prediction vector where c corresponds to the number of classes of the classification problem.

Below, the explanation methods which produce saliency maps and which were used in this work, will be briefly described.

a) Gradient: The gradient quantifies how much a change in each feature would a change the predictions $f(x)$ in a small neighborhood around the input [5].

$$Grad(x) = \frac{\partial f(x)}{\partial x} \tag{2}$$

b) DeConvNet: The DeConvNet associates the architecture of the model, with a corresponding architecture that reverses the computations and produces an image as the output [6]. To do this, each layer is associated with a corresponding layer that reverses the computation.

c) Guided Backpropagation: a combination of the previous two methods, guided backpropagation prevents the backward flow of negative gradients, corresponding to the neurons which decrease the activation of the units we are inspecting [5]. Negative gradient are set to zero while backpropagating.

d) Deep Taylor Decomposition: DTD is obtained by propagating the model output through the network using redistribution rules, until the input features are reached [8]. The propagation rules are derived from a Taylor decomposition performed at each unit of the network.

e) Layer-wise Relevance Propagation: similar to DTD, LRP is obtained by propagating the model output though the network using redistribution rules [7]. LRP find its mathematical foundations in Deep Taylor Decomposition [8]. The redistribution rules proportional decompose the relevance score of upper layers to obtain lower layer relevance scores, based on the forward mappings between layers.

C. Interpretability Metrics

Although a large number of saliency map methods have been proposed, relatively few metrics to evaluate their fidelity have been proposed. Fidelity is a concept that should capture how well the relevance given to each pixel represents the process of the model. We will now describe two interpretability metrics that can be considered a proxy for fidelity and evaluate the quality of the saliency maps.

Confidence drop tracks the decrease of confidence in the model's classification when removing a percentage of the most relevant pixels given by a saliency map. If the saliency maps present a high fidelity to the model, the confidence should drop faster than when the fidelity is low.

$$Drop(k) = f(x^{(0)}) - f(x^{(k)}) \tag{3}$$

In Equation 3 we can see that the confidence drop corresponds to the difference in confidence when perturbating the k higher relevant pixels.

Another interpretability metric is called Area Over the Perturbation Curve (AOPC) [9]. The AOPC tracks the decrease of confidence in the model's classification when iteratively removing the most relevant pixels given by a saliency map.

The AOPC equation is described in the following equation:

$$AOPC = \frac{1}{L+1}\langle \sum_{k=1}^{L} f(x^{(0)}) - f(x^{(k)}) \rangle \tag{4}$$

In Equation 4, L is the number of pixel perturbation steps, $f(x)$ is the output value of the classifier for input image x (i.e. the confidence assigned to the class), $x(0)$ is the original image and $x(k)$ is the image after k perturbations.

III. Experiments

A. Dataset and Models

Experiments were performed on the CIFAR-10 dataset [10] as it is a well-known image classification dataset with suitable complexity. A description of the dataset is present in Table I.

TABLE I
DESCRIPTION OF THE DATASET USED IN THE STUDY

name	samples	classes	width	height	channels
CIFAR10	60000	10	32	32	3

In this experiments, a standard convolutional neural network (CNN) containing three convolutional blocks and two fully-connected layers, was used. Each convolutional block is composed by two convolutional layers followed by a max-pooling layer. The classification is done using a softmax layer after the fully-connected layers. A figure representing the architecture of the network is presented in Figure 1.

The training set and test set were combined and 10-fold cross-validation was used to train and evaluate the model. During the training of the model, different levels of L^2 weigh decay were used (0, 0.0001, 0.001, 0.005, 0.01, 0.05) as well

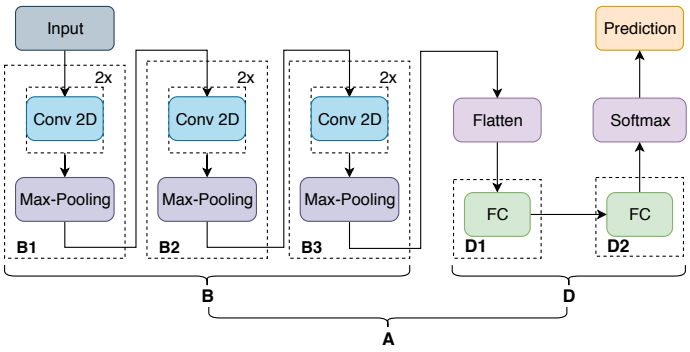

Fig. 1. Architecture of the neural network.

early stopping to prevent overfitting, stopping training after 10 epochs of no improvement in the loss function. In order to understand the changes in interpretability that were caused by changing the complexity in different layers of the network, we separated regularization in different group of layers. $B1$, $B2$ and $B3$ corresponds to regularization on the first, second or third convolutional block respectably; $D1$ and $D2$ correspond to regularization on the first or second fully-connected layer. Finally, B corresponds to regularization on all convolutional blocks, D corresponds to regularization on all fully-connected layers and A corresponds to regularization on all layers of the network. In order to evaluate the interpretability of a model, we take the test samples and the trained classifier, and apply the saliency map method to the samples, resulting in saliency maps which are used to calculate the interpretability metric. These concepts are visually illustrate in Figure 2.

B. Saliency Map Methods

In this experiments, five different saliency map methods were compared. The criteria for choosing these methods was based on their proven applicability in the literature as well as its properties. The methods chosen were gradient [5], DeConvNet [6], Guided Backpropagation [5], Deep Taylor Decomposition [8] and Layer-wise Relevance Propagation [7]. Some methods only estimate positive relevance while other methods estimate positive and negative relevance. For example, Guided Backpropagation and Deep Taylor Decomposition, produce only positive relevance. The implementation of the explanation methods was done using the iNNvestigate Toolbox v1.0.8 [12].

C. Saliency Metrics

We use two different perturbations, one by deleting the most relevant pixels given the relevance score provided by the saliency map, and the other by deleting a random pixel. In random perturbation method, the value is a gray-scale value sampled from a uniform distribution in the case of gray-scale images, and RGB value in the case of colored images. This approach attempts to destroy the information contained in the pixel.

To measure the confidence drop caused by the perturbation, we have segmented different percentages of most relevant

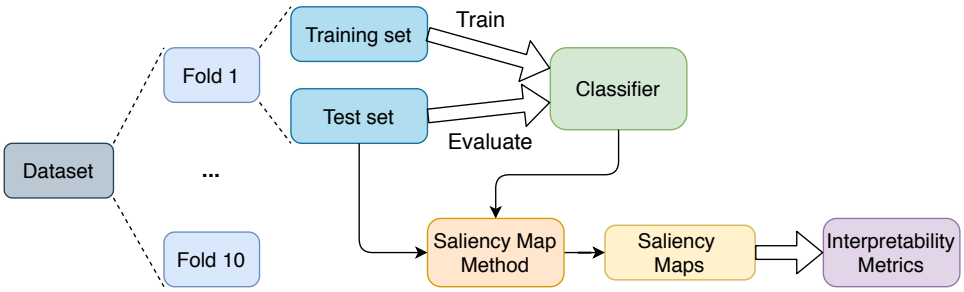

Fig. 2. Architecture of experimental setup.

pixels. We have chosen to group the pixels in the 5%, 10%, 20%, 30%, 40%, 50% and 75% most relevant pixels.

In order to calculate the AOPC metric we used perturbation steps corresponding to 10% of the dataset.

IV. Results

In this study, two experiments were conducted. In the first, regularization was applied in all layers of the network (convolutional and fully-connected). In the second experiment we separated regularization by different groups of layers: regularization on only one layer or block of layers ($B1$, $B2$, $B3$, $D1$, $D2$) and regularization applied to multiple blocks of layers (B, D, A).

A. How does the regularization of the deep neural network affect the quality of saliency maps?

The first experiment measures the interpretability of models regularized in all layers with different L^2 weight decay values.

The Table II is composed by the saliency map methods (first column), and the weight decay values (first row). The values of the table represent the mean AOPC value for the 10 folds. The highest interpretability values for each method are highlighted in bold.

The results in Table II show that the methods that display highest values of interpretability and that produce saliency maps with more fidelity to the model's decision are the LRP and Gradient methods.

The results in Table II show a substantial difference between interpretability and saliency map methods. In general, the quality of such methods is higher when the network is trained with smaller regularization values, although the exact value is not consistent between methods.

In order to assess the statistical significance of the interpretability metric AOPC when the regularization values in all layers is changed, the Friedman's test was applied with a significance level of 5%. We considered the different weight decays as well as the different saliency map methods. It was determined that the regularization does have an statistical significance on the interpretability metric. Statistical significance were detected between the lower regularization values (0, 0.0001, and sometimes 0.001) and the higher regularization values (0.01 and 0.05) consistently in all methods.

The same statistical significance tests were also conducted with the experiments using random perturbation which found no statistical difference between neither of the regularization values in all saliency methods.

B. How does the layer regularized affect the interpretability of the network?

Another question we had was to learn what layers regularization is more appropriate in order to produce the saliency maps with better fidelity to the model. To answer this question we run an experiment training CNN's with regularization only on specific layers and we extracted the saliency maps in order to measure their fidelity using the AOPC metric.

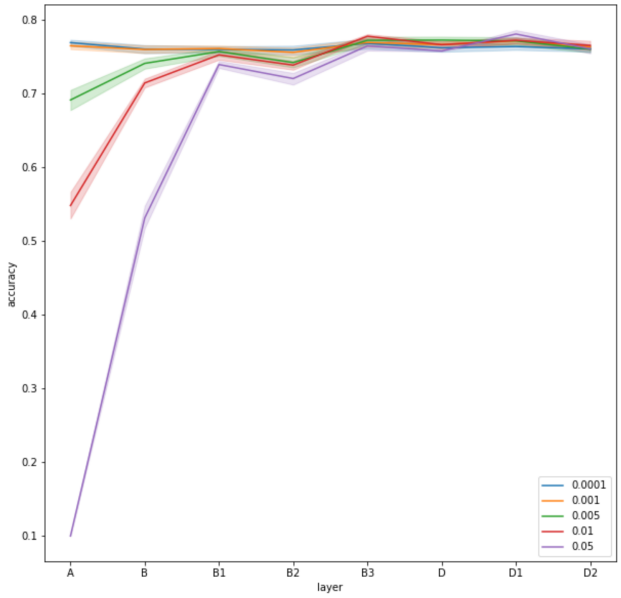

Fig. 3. Accuracy of the different models based on layer and weight decay of regularization.

In Figure 3 its possible to see the accuracy of the different models based on the layer which was regularized with the specified weight decay values. The models which were regularized in all layers have lower performance than the other models, especially with higher regularization values. From Figure 3 we can see that there is a performance benefit to use regularization on only some layers, and not in all of them.

TABLE II

RESULTS COMPARING INTERPRETABILITY (AOPC) OF SALIENCY MAP METHODS ON MODELS WITH DIFFERENT REGULARIZATION VALUES.

	L^2 weight decay					
Method	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
DeConvNet	**0.314**	0.293	0.253	0.232	0.134	9.3E-12
Deep Taylor	0.218	**0.230**	0.229	0.201	0.114	9.3E-12
Gradient	**0.419**	0.413	0.417	0.356	0.226	9.3E-12
Guided Backprop	0.269	**0.281**	0.263	0.224	0.132	9.3E-12
LRP	0.423	0.421	**0.427**	0.362	0.231	9.3E-12

TABLE III

RESULTS USING THE DECONVNET METHOD. THE FIRST COLUMN CORRESPONDS TO THE LAYER REGULARIZED, AND THE FIRST ROW THE L^2 WEIGHT DECAY. THE VALUES REPRESENT THE MEAN AOPC, AND THE HIGHEST VALUES FOR EACH REGULARIZATION VALUE ARE HIGHLIGHTED IN BOLD.

	L^2 weight decay					
Layer	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
B1	**0.3142**	0.3082	0.3004	0.2810	0.2772	0.2292
B2	**0.3142**	**0.3368**	0.3065	0.2882	0.2651	0.2328
B3	**0.3142**	0.3107	0.3135	0.3037	0.2710	0.3228
B	**0.3142**	0.2878	0.2651	0.2451	0.2603	0.1600
D1	**0.3142**	0.3253	0.3265	**0.3298**	**0.3425**	0.3666
D2	**0.3142**	0.3247	0.3238	0.3197	0.3249	0.3155
D	**0.3142**	0.3262	**0.3344**	**0.3298**	0.3395	**0.3503**
A	**0.3142**	0.2934	0.2526	0.2316	0.1343	0.0000

TABLE IV

RESULTS USING THE DEEP TAYLOR METHOD. THE FIRST COLUMN CORRESPONDS TO THE LAYER REGULARIZED, AND THE FIRST ROW THE L^2 WEIGHT DECAY. THE VALUES REPRESENT THE MEAN AOPC, AND THE HIGHEST VALUES FOR EACH REGULARIZATION VALUE ARE HIGHLIGHTED IN BOLD.

	L^2 weight decay					
Layer	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
B1	**0.2183**	0.2216	0.2252	0.2274	0.2211	0.2188
B2	**0.2183**	0.2283	0.2232	0.2226	0.2224	0.1988
B3	**0.2183**	**0.2368**	0.2310	0.2323	0.2245	0.2416
B	**0.2183**	0.2269	0.2312	0.2326	0.2236	0.1453
D1	**0.2183**	0.2279	**0.2465**	0.2427	**0.2499**	**0.2526**
D2	**0.2183**	0.2267	0.2302	0.2405	0.2235	0.2354
D	**0.2183**	0.2250	0.2348	**0.2452**	0.2473	0.2516
A	**0.2183**	0.2301	0.2292	0.2014	0.1138	0.0000

TABLE V

RESULTS USING THE GRADIENT METHOD. THE FIRST COLUMN CORRESPONDS TO THE LAYER REGULARIZED, AND THE FIRST ROW THE L^2 WEIGHT DECAY. THE VALUES REPRESENT THE MEAN AOPC, AND THE HIGHEST VALUES FOR EACH REGULARIZATION VALUE ARE HIGHLIGHTED IN BOLD.

	L^2 weight decay					
Layer	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
B1	**0.4192**	0.4212	0.4249	0.4167	0.4145	0.4153
B2	**0.4192**	0.4242	0.4244	0.4101	0.4153	0.3898
B3	**0.4192**	**0.4301**	0.4252	0.4295	0.4230	0.4279
B	**0.4192**	0.4279	0.4272	0.4149	0.3963	0.2309
D1	**0.4192**	0.4227	0.4252	0.4287	0.4226	**0.4358**
D2	**0.4192**	0.4181	**0.4291**	**0.4303**	**0.4273**	0.4341
D	**0.4192**	0.4211	0.4248	0.4289	0.4266	0.4201
A	**0.4192**	0.4127	0.4174	0.3564	0.2260	0.0000

In Tables III- VII is visible that, for each saliency map method, the interpretability metric based on the L^2 weight decay used to regularize the specific layer of the model.

Regarding the results presented, we can see that interpretability appears to be higher with lower regularization values. Additionally, interpretability appears to be higher when regularization happens in higher convolutional layers or in fully-connected layers.

The methods that display highest values of interpretability and that produce saliency maps with more fidelity to the model's decision are the LRP and Gradient methods.

Following further analysis to these results, we can plot the number of times that regularization in a specific layer has produced the best interpretability values for each method. This plot is presented in Figure 4, and as we can see, it once again shows that regularization is more effective in higher convolutional layers or in fully-connected layers.

TABLE VI

RESULTS USING THE GUIDED BACKPROP METHOD. THE FIRST COLUMN CORRESPONDS TO THE LAYER REGULARIZED, AND THE FIRST ROW THE L^2 WEIGHT DECAY. THE VALUES REPRESENT THE MEAN AOPC, AND THE HIGHEST VALUES FOR EACH REGULARIZATION VALUE ARE HIGHLIGHTED IN BOLD.

	L^2 **weight decay**					
Layer	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
B1	**0.2690**	0.2722	0.2720	0.2698	0.2563	0.2454
B2	**0.2690**	**0.2884**	0.2696	0.2700	0.2561	0.2312
B3	**0.2690**	0.2860	0.2844	0.2825	0.2615	0.2881
B	**0.2690**	0.2777	0.2723	0.2611	0.2678	0.1629
D1	**0.2690**	0.2867	0.2852	0.2912	0.2903	0.2993
D2	**0.2690**	0.2792	0.2716	**0.2922**	0.2878	0.2755
D	**0.2690**	0.2788	**0.2919**	0.2867	**0.2951**	**0.3022**
A	**0.2690**	0.2813	0.2630	0.2243	0.1324	0.0000

TABLE VII

RESULTS USING THE LRP METHOD. THE FIRST COLUMN CORRESPONDS TO THE LAYER REGULARIZED, AND THE FIRST ROW THE L^2 WEIGHT DECAY. THE VALUES REPRESENT THE MEAN AOPC, AND THE HIGHEST VALUES FOR EACH REGULARIZATION VALUE ARE HIGHLIGHTED IN BOLD.

	L^2 **weight decay**					
Layer	**0**	**0.0001**	**0.001**	**0.005**	**0.01**	**0.05**
B1	**0.4226**	0.4274	0.4304	0.4202	0.4203	0.4196
B2	**0.4226**	0.4295	0.4289	0.4154	0.4184	0.3947
B3	**0.4226**	**0.4340**	0.4284	0.4344	0.4256	0.4302
B	**0.4226**	0.4309	0.4312	0.4206	0.4055	0.2336
D1	**0.4226**	0.4282	0.4278	**0.4351**	0.4288	**0.4439**
D2	**0.4226**	0.4227	**0.4333**	0.4337	**0.4351**	0.4387
D	**0.4226**	0.4273	0.4301	0.4334	0.4317	0.4281
A	**0.4226**	0.4211	0.4265	0.3622	0.2313	0.0000

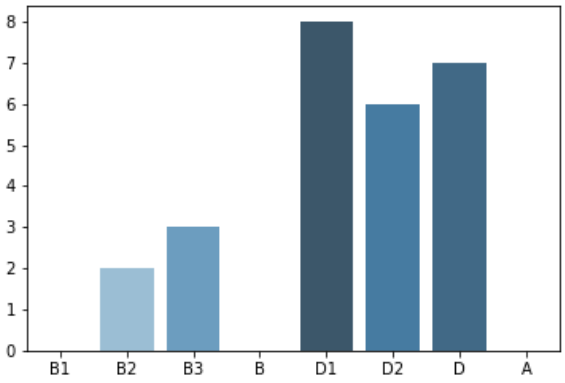

Fig. 4. Plot showing the number of times the regularization in a layer produced the best interpretability value for each method.

V. CONCLUSIONS

In this work we studied the relationship between regularization and interpretability in a CNN context. From the results obtained with the experimental data, the following main conclusions may be derived:

- The quality of saliency maps is higher when the network is trained with smaller regularization values;
- LRP and Gradient produce saliency maps with higher fidelity to the model's decision;
- Overall, in order to obtain higher interpretability, regularization should be applied on later convolutional layers or in fully-connected layers;
- Models in which all layers were regularized display lower interpretability than models in which only the fully-connected layers were regularized.

To the best of the authors' knowledge this was the first study in this context.

In the future, new experiments will be conducted to test other mechanisms of regularization such as dropout, as well as extend our work to other datasets. We will also compare the saliency maps produced by the different methods in order to understand different in their distributions.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[2] Z. C. Lipton, "The Mythos of Model Interpretability," 2016.

[3] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for Interpretability," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2280–2288.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, ser. KDD '16. New York, NY, USA: ACM, aug 2016, pp. 1135–1144. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939778

[5] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014.

[6] M. D. Zeiler, e. D. Fergus, Rob", T. Pajdla, B. Schiele, and T. Tuytelaars, "Visualizing and Understanding Convolutional Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, 2014, pp. 818–833.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[8] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, p. 211–222, May 2017.

[9] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining Recurrent Neural Network Predictions in Sentiment Analysis," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 159–168.

[10] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Master's thesis, University of Toronto, Canada, 2009.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[12] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks!" 2018.