




A Dashboard for Interactive Convolutional Neural Network Training And Validation Through Saliency Maps

Tim Cech², Furkan Simsek¹, Willy Scheibel¹, and Jürgen Döllner¹

¹University of Potsdam, Digital Engineering Faculty, Hasso Plattner Institute, Germany

²University of Potsdam, Digital Engineering Faculty, Germany

Abstract

Quali-quantitative methods provide ways for interrogating Convolutional Neural Networks (CNN). For it, we propose a dashboard using a quali-quantitative method based on quantitative metrics and saliency maps. By those means, a user can discover patterns during the training of a CNN. With this, they can adapt the training hyperparameters of the model, obtaining a CNN that learned patterns desired by the user. Furthermore, they neglect CNNs which learned undesirable patterns. This improves users' agency over the model training process.

CCS Concepts

• *Computing methodologies* → *Artificial intelligence*;

1. Introduction

Today, *Machine Learning* (ML) systems are apparent in high-stakes contexts such as medical diagnostics [BSCA20], autonomous driving [PYCM20], or criminal justice [TPB*22]. In ML systems concerned with image data, often, *Convolutional Neural Networks* (CNN) are applied, which encode complex decision functions rendering the resulting ML system opaque [Fu94]. In high-stakes contexts, this opaqueness and the subsequent lack of accountability is not desirable, because neither users nor developers can infer why the ML system made a specific decision. On the contrary, ML systems must be accessible to external audits [RSW*20]. On that account, developers need technical means to understand what their ML system actually learned.

One opaque part when training a CNN is the choice of the training hyperparameters, e.g., the choice of an optimizer. Training hyperparameters are such parameters that have to be defined *a priori* before the actual training starts. This choice, in general, is neither trivial nor universally applicable and is usually done by evaluating the quantitative performance of the CNN on a test dataset [AAPV19]. Additionally, this choice influences if the model learns a meaningful abstraction from the data.

To address this problem, we pick up a research direction that encourages the usage of quali-quantitative methods for obtaining a qualitative small-data view as well as a quantitative big-data view on a given ML system [BP14]. For it, we present SalienCNN, an interactive dashboard that combines classic quantitative measures and visualization techniques, e.g. a confusion matrix, with qualitative methods for a guided interrogation of a CNN using *Saliency Maps* (SM). The dashboard allows the user to dynamically adapt

training hyperparameters of a CNN if either the qualitative monitoring with SMs or the monitoring of quantitative measurements give the user a reason to adjust the model for re-training. With this, they obtain a CNN which learned desirable patterns while also showing strong performance and filter models, which either learned undesirable patterns or perform poorly. Thus, by using SalienCNN, we increase users' agency in the training process of a CNN. Previously, several dashboards were proposed to understand CNN models. We identify *CNNPruner* [LWS*21] and *ConceptExplainer* [HMKB23] as most closely related to our work. We share the goal of CNNPruner to enable expert users to improve a given CNN during training time. In contrast to this method, we allow users to monitor CNNs even more closely by adapting its training hyperparameters. With ConceptExplainer we share the idea of discovering complex patterns to enable users to assess the model quality. In contrast to ConceptExplainer, we refrain from defining complex patterns ourselves and let the user assess whether or not the marked regions by the SMs constitute an important pattern as the user would do in a qualitative interview.

2. Dashboard

SalienCNN contains three major views. First, the default view contains visualizations of standard quantitative measurements, e.g., a bar chart showing the accuracy development over epochs or a confusion matrix [ZGCH21]. A side-by-side view enables the user to compare SMs side by side. Thereby, we consider different SM techniques. Basically, SMs are techniques that utilize the metaphor of showing where a CNN model "actually looks" when classifying an image [OSCW19]. In particular, we consider

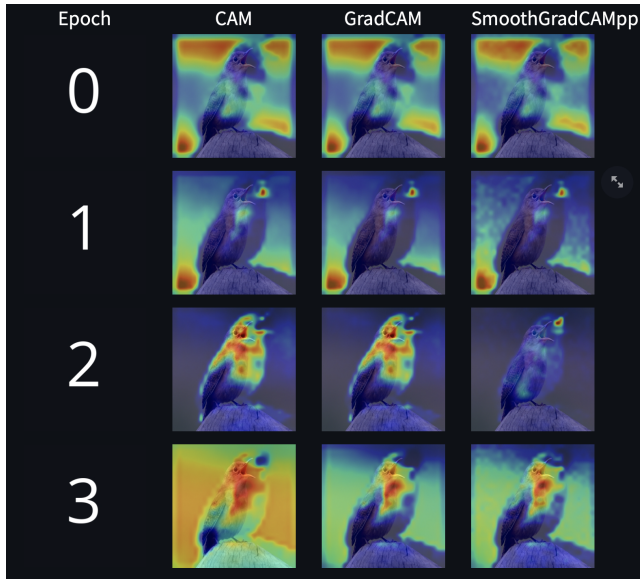


Figure 1: The side-by-side view from our dashboard comparing the result of different SM techniques side by side per epoch

the *Class Activation Mapping* (CAM) algorithm [ZKL*16], *Gradient CAM* (GradCAM) [SCD*20], and *Smoothed GradCAM++* (SmoothGradCAMpp) [OSCW19] to allow the user to compare the strength and weaknesses of different SM techniques. The fallacy view informs the user about one specific weakness that is the result of a sanity check for SM techniques called *input variance* [AGM*18]. The input variance describes the phenomenon that the result of an SM algorithm may change when adding a noise signal in the input layer and subtracting it in the first hidden layer so that the CNN is not changed effectively. Therefore, mathematically completely identical models may result in different SMs. We present SalienCNN on the example of the CUB200-2011 dataset [WBW*11] because this dataset was used in previous studies on model architecture understanding (e.g. [RCC*22, SRKL20]) but SalienCNN is also applicable on other datasets.

In the side-by-side view, as shown in Figure 1, SMs based on CAM, GradCAM, and SmoothGradCAMpp per epoch are shown side by side to allow users to assess the model quality in terms of the more complex patterns. For it, a user chooses a class they wish to investigate. Then, we recommend to the user an SM progression over time for each of the following cases according to the training result of the most current epoch: (1) A correctly classified image, (2) an image wrongly attributed to the chosen class, and (3) an image of the chosen class that was misclassified. This allows the user to assess whether, in one of three cases, the CNN learned an abstraction that is not applicable to the intended use-case. Furthermore, the user is also able to obtain the SM progression for a custom image. Using SMs, users are encouraged to identify patterns themselves as they would in a qualitative interview where the interviewer would be free to focus on any part of the response of the interviewee. As showcased in Figure 1, the user may observe

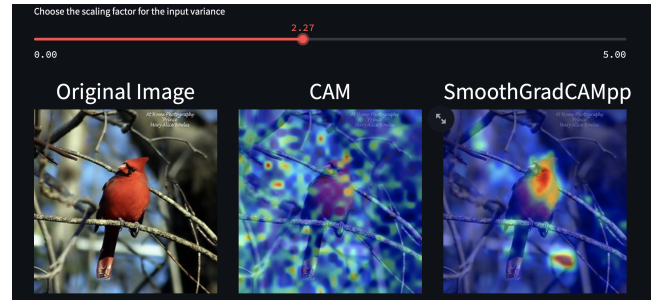


Figure 2: In the fallacy view, we allow the user to test the effect of the input variance sanity check. Here, the right SmoothGradCAMpp SM is a lot more resilient against input variance.

that the CNN slowly learns to focus on the neck and head of the bird like humans would usually do to identify a bird.

The fallacy view, as shown in Figure 2, informs the user about the input variance sanity check and how strongly it affects the different SM techniques. As shown there, for the given image, the SM produced by the CAM algorithm becomes noisy when it is subject to input variance, while the SM produced by the SmoothGradCAMpp method stays relatively consistent. Hereby, the user may gain deeper insights, e.g., that they may trust the visualization provided by the SmoothGradCAMpp method more than from the basic CAM method. For it, the user obtains a more in-depth understanding of the provided visualizations.

After assessing the quantitative quality through performance measurements and the qualitative quality through SMs, the user may conclude that a model has not learned a meaningful abstraction from the data. Then, the dashboard enables the user to choose a different set of training hyperparameters and re-train the model. Thus, they iteratively obtain a more accountable CNN which learned desired patterns and is therefore auditable.

3. Conclusions & Future Work

To summarize, SalienCNN allows the human assessment of Convolutional Neural Networks during training time. We support users' agency in the training process by providing them with information with quantitative metrics and qualitative SM pattern recognition. We consider SalienCNN as a starting point for future work. For one, we only considered one sanity check, which could be extended with more comparison of the reliability of Saliency Map techniques. Furthermore, we plan to perform a user study to measure how effective our dashboard would be in a real-case scenario.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was partially funded by the Federal Ministry for Education and Research (Germany) through grant 01IS22062 ("AI research group FFS-AI").

References

- [AAPV19] ADAM S. P., ALEXANDROPOULOS S.-A. N., PARDALOS P. M., VRAHATIS M. N.: No free lunch theorem: A review. In *Approximation and Optimization: Algorithms, Complexity and Applications*. Springer, 2019, pp. 57–82. doi:10.1007/978-3-030-12767-1_5. 1
- [AGM*18] ADEBAYO J., GILMER J., MUELLY M., GOODFELLOW I., HARDT M., KIM B.: Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (2018), vol. 31, Curran Associates, Inc. 2
- [BP14] BLOK A., PEDERSEN M. A.: Complementary social science? quali-quantitative experiments in a big data world. *Big Data & Society* 1, 2 (2014), 1–6. doi:10.1177/2053951714543908. 1
- [BSA20] BATTINENI G., SAGARO G. G., CHINATALAPUDI N., AMENTA F.: Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine* 10, 2 (2020), 1–11. doi:10.3390/jpm10020021. 1
- [Fu94] FU L. M.: Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics* 24, 8 (1994), 1114–1124. doi:10.1109/21.299696. 1
- [HMKB23] HUANG J., MISHRA A., KWON B. C., BRYAN C.: ConceptExplainer: Interactive explanation for deep neural networks from a concept perspective. *Transactions on Visualization and Computer Graphics* 29, 1 (2023), 831–841. doi:10.1109/TVCG.2022.3209384. 1
- [LWS*21] LI G., WANG J., SHEN H.-W., CHEN K., SHAN G., LU Z.: CNNPruner: Pruning convolutional neural networks with visual analytics. *Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1364–1373. doi:10.1109/TVCG.2020.3030461. 1
- [OSCW19] OMEIZA D., SPEAKMAN S., CINTAS C., WELDERMARIAM K.: Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models. *CoRR cs.CV* (2019). arXiv:1908.01224. 1, 2
- [PYCM20] PENG Z., YANG J., CHEN T.-H. P., MA L.: A first look at the integration of machine learning models in complex autonomous driving systems: A case study on apollo. In *Proc. of ESEC/FSE 2020* (2020), ACM, pp. 1240–1250. doi:10.1145/3368089.3417063. 1
- [RCC*22] RUDIN C., CHEN C., CHEN Z., HUANG H., SEMENOVA L., ZHONG C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* 16 (2022), 1–85. doi:10.1214/21-SS133. 2
- [RSW*20] RAJI I. D., SMART A., WHITE R. N., MITCHELL M., GEBRU T., HUTCHINSON B., SMITH-LOUD J., THERON D., BARNES P.: Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proc. of FAT* '20* (2020), ACM, pp. 33–44. doi:10.1145/3351095.3372873. 1
- [SCD*20] SELVARAJU R. R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D., BATRA D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, 2 (2020), 336–359. doi:10.1007/s11263-019-01228-7. 2
- [SRKL20] SAGAWA S., RAGHUNATHAN A., KOH P. W., LIANG P.: An investigation of why overparameterization exacerbates spurious correlations. In *Proc. of the 37th International Conference on Machine Learning* (2020), vol. 119, PMLR, pp. 8346–8356. 2
- [TPB*22] TRAVAINI G. V., PACCHIONI F., BELLUMORE S., BOSIA M., DE MICCO F.: Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *International Journal of Environmental Research and Public Health* 19, 17 (2022), 1–13. doi:10.3390/ijerph191710594. 1
- [WBW*11] WAH C., BRANSON S., WELINDER P., PERONA P., BELONGIE S.: *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011. 2
- [ZGCH21] ZHOU J., GANDOMI A. H., CHEN F., HOLZINGER A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 1–19. doi:10.3390/electronics10050593. 1
- [ZKL*16] ZHOU B., KHOSLA A., LAPEDRIZA A., OLIVA A., TORRALBA A.: Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), IEEE Computer Society, pp. 2921–2929. doi:10.1109/CVPR.2016.319. 2