ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Evaluating the Quality of Saliency Maps for Distilled Convolutional Neural Networks

Jasper Wilfling[1], Matias Valdenegro-Toro[1], and Marco Zullich[1]

1- Department of AI - Faculty of Science and Engineering - University of Groningen
Nijenborgh 9, 9747AG - Groningen, the Netherlands

Abstract. Knowledge Distillation (KD) is a popular technique to compress Deep Neural Networks. Studies on KD often evaluate it on the basis of accuracy and time-complexity; however, there exist other facets of a model performance, like *explainability* and *fairness*. In the present work, we evaluate the quality of saliency maps in terms of *faithfulness* and *coherence* in the context of KD and compare the results obtained with the uncompressed model. Our findings indicate how KD is potentially decreasing the accuracy of the saliency maps, thus acting as a warning on the usage of KD when high-quality explanations are required.

1 Introduction

Knowledge Distillation (KD) [1] is a Model Compression (MC) technique whereby a (usually) large machine learning (ML) model, called *teacher*, transfers its *knowledge* onto a (usually) smaller model called *student*. This is paramount in today's ML world, whereas size and energy consumption of Deep Neural Networks (DNNs) pose a threat to sustainability and accessibility of these models. The goal of KD is to have the *task-level* performance (e.g., accuracy in case of classification) of the student to be as close as possible to the teacher's.

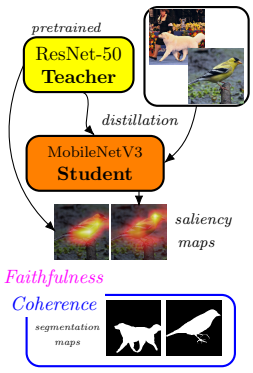

Fig. 1: Illustration of our methodology. We compare the quality of segmentation maps produced by student and teacher models on faithfulness and coherence—the latter making use of precomputed segmentation maps.

Often MC techniques are evaluated in terms of task-level performance and time complexity; however, rarely other facets of performance, such as *explainability* and *fairness*, are considered. Explainable AI (XAI) is the field that studies how to make *understandable* a specific aspect of an AI system [2]. Saliency maps are a common XAI tool used to outline *important* features in the prediction of a single data point by a ML model. They can be functionally evaluated on facets such as *faithfulness*—capability of identifying all and only the features which are relevant to the model—and *coherence*—adherence of the map to a predetermined ground truth.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

In the present work, we present an analysis on the quality of saliency maps produced in the context of KD. We apply KD on Convolutional NNs (CNNs) on a subset of the popular Imagenet [3]. We then produce saliency maps using several variants of the popular Grad-CAM technique (Grad-CAM [4], Grad-CAM++ [5], HiResCAM [6], XGradCAM [7]), and proceed to evaluate these techniques according to faithfulness and coherence, using different metrics elicited from the literature. A diagram of our methodology is depicted in Figure 1. Our findings suggest that KD might significantly impact the quality of the segmentation maps, both on faithfulness and coherence.

In the literature, there are not a lot of works investigating the quality of the *explanations* produced by compressed DNNs, likely due to functional evaluation being a recent research line [2]. Alharbi et al. [8] propose a KD method that tries to imitate saliency maps produced by the teacher; however, their work does not delve into the assessment of the quality of these saliency maps. Termritthikun et al. [9] operate KD on X-ray imaging classification. They create saliency maps from student and teacher and visually compare the two, concluding that student models produce more *compact* maps.

To the best of our knowledge, ours is the first work proposing a thorough, albeit small-scale, functional comparison of input attribution explanations produced by student and teacher models in the context of KD. Our code is available on this GitHub repository: https://github.com/JasperWi/Explainable-KD-CNN.

2 Materials and Methods

Dataset In our experiments, we made use of the Imagenet-1k [3] and its Imagenet-S_{50} subset [10], which contains 615 images and fine-grained segmentation maps over 50 categories. For the distillation of the student model, we made use of a subset Imagenet-1k on these 50 classes, which yielded about 65 000 training and 2500 test images. We used Imagenet-S_{50} for the XAI part, using the segmentation maps as ground truth for computing coherence.

KD KD [1] is a generic technique to *transfer* knowledge from a trained ML model—called *teacher*—to another model—termed *student*, which is usually smaller in size. Specifically, in *output-based* distillation, the teacher provides its output in form of *soft labels*, which can be combined with the *hard labels*, represented by the ground truth, to train the student. In the context of C-way classification, NNs produce a vector of logits $\mathbf{z} \in \mathbb{R}^C$. The authors then propose the following loss function for training the student:

$$\mathcal{L} = \lambda \cdot \overbrace{KL\left[\text{softmax}_T(\mathbf{z}_{\text{student}}), \text{softmax}_T(\mathbf{z}_{\text{teacher}})\right]}^{\text{distillation loss}} + (1-\lambda) \cdot \overbrace{CCE\left[\text{softmax}(\mathbf{z}_{\text{student}}), y\right]}^{\text{classification loss}}. \quad (1)$$

KL indicates the Kullback-Leibler divergence; $\lambda \in [0, 1]$ is a hyperparameter which mediates between distillation and classification loss. $\text{softmax}_T(\mathbf{z})$ is the softmax with temperature T: $\frac{\exp(\mathbf{z}/T)}{\sum_{k=1}^{C} \exp(\mathbf{z_k}/T)}$. $T > 0$ is a hyperparameter that flattens the simplex produced by the softmax; CCE indicates the Categorical-Cross Entropy.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

XAI: Input Attribution and Functional Evaluation In the present work, we specifically target *saliency maps* as the XAI method we study. The goal is to elicit the *saliency* of the features within the input. Since our data are images, these tools rank the importance of (groups of) pixels in determining a specific prediction produced by a CNN.

We consider explanations which are variations of Grad-CAM [4]. Its main idea is to select a convolutional layer in a CNN (usually the last one before the classification head) and multiply its activations times a function of the gradient of the output w.r.t. these activations. The output is then (i) averaged per-channel, (ii) normalized between 0 and 1, and (iii) upscaled to match the original size of the image. In the present work, we consider Grad-CAM alongside three variations: Grad-CAM++ [5], HiRes-CAM [6], and XGrad-CAM [7]. Their only difference with Grad-CAM is the way they compute the aforementioned function of the gradients.

Faithfulness indicates the adherence of the saliency map with the prediction dynamics of the model. Ideally, we would like an input attribution technique to elicit *all* of the features which are relevant for the model, while avoiding to highlight irrelevant features. Faithfulness can be measured with several metrics [2], usually based off of *incremental deletion*. The underlying idea is that, by *deleting* progressively more salient features, there should be a noticeable drop in the accuracy or confidence of the model. Monotonicity score (MS) [11] and Faithfulness estimate (FE) [12] estimate faithfulness as the correlation between the accuracy drop and the proportion of deleted pixels.

Coherence calculate the level of *overlap* between the features highlighted as salient and a corresponding ground truth. Since we make use of Imagenet-S_{50}, we use, as ground truth, the segmentation maps proposed in the dataset. We consider two metrics: (a) Pointing Game (PG) [13], which measures whether the max pixel in the saliency maps falls within the segmentation map, and (b) Attribution Localization (AL) [14], which computes the *precision* of the saliency map w.r.t. the segmentation map.

CNN Models In our work, we make use of two popular CNN architectures. As teacher, we employ a ResNet-50 [15] pretrained on the Imagenet-1k datasets. This model has 50 convolutional layers and around 25 million parameters. As a student, we use MobileNetV3-small [16], which employs 2.5 million parameters, thus roughly $\frac{1}{10}$-th of ResNet-50. We trained the student on solving the image classification task on Imagenet-1k subset on the 50 classes in the dataset Imagenet-S_{50}. A schematization of our methodology is presented in Figure 1.

3 Experimental Settings and Results

We ran all of the experiments using Python with the PyTorch and torchvision libraries. For producing the XAI tools, we made use of the `pytorch-grad-cam` tool [17]. For the evaluation of the explanations, we used Quantus [18]. For KD, we used as a teacher the ResNet-50 pretrained on ImageNet-1k available

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

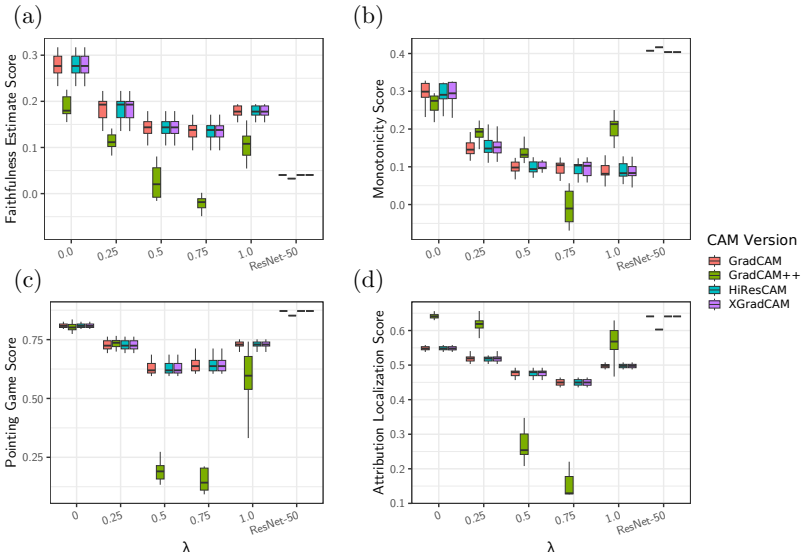

Fig. 2: Results for (a) FE, (b) MS, (c) PG, (d) AL, on the student models for various values of λ and on the teacher ("ResNet-50"). Each boxplot (excluding teacher) is computed over 5 runs. $\lambda = 0$ indicates a student trained with hard labels, while $\lambda = 1$ indicates a model trained using soft labels only.

on torchvision. We trained the students for 160 epochs using the RADAM optimizer [19] with a learning rate of 0.003, a batch size of 128, and gradient clipping to norm 1. We set the temperature parameter in Equation (1) to 50 and trained with $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. $\lambda = 0$ indicates a regular training without distillation, while, for $\lambda = 1$, we only train with the distillation loss. For each value of λ, we distilled 5 different models, accounting for the stochasticity in the training process. We rescaled the images to a resolution of 224×224 and standardized them using the mean values $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. In each batch, we randomly erased rectangular patches of the images, replacing them with black pixels. This step is necessary for the computation of the fairness metrics, since erasing the least/most salient pixels may lead to out-of-distribution data [2].

The pretrained teacher model had an accuracy of 81.45%, while all student models attained a mean accuracy between 80.46% and 84.43%, with the students trained with $\lambda = 1$ recording the highest mark. The results for faithfulness and coherence are shown in Figure 2. We can notice how the student model trained without distillation ($\lambda = 0$) consistently scores the best values in the different metrics. For the student models with KD, faithfulness seems to be impacted by a higher level of distillation; however, the purely distilled models seem to perform better in the case of FE; the latter trend is also noticeable for both coherence metrics, where the models using a mixture of distillation and classification loss consistently score lower results. Finally, for MS, PG, and AL, the teacher model seems to record better scores than the students, while, on

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

FE, its results seem very underwhelming. Considering the single XAI tools, we can notice how GradCAM, HiResCAM, and XGradCAM tend to produce very similar results, while GradCAM++ seems to generically underperform w.r.t. the other three, while also being quite unstable.

4 Discussion and Conclusions

In the present work, we evaluated the quality of the saliency maps produced from a MobileNetV3-small Convolutional Neural Network. We trained the model using Knowledge Distillation (KD) on the Imagenet-S_{50} dataset using a pretrained ResNet-50 teacher. We used Grad-CAM and some other variations for producing the saliency maps and we evaluated them using faithfulness and coherence metrics. The results suggest that, despite KD helping in getting students which are more accurate than their teacher, the latter are able to produce more faithful and coherent segmentation maps. When considering the student models, a regular, hard-labels only training seems to yield better results than KD. Also, mixing hard and soft labels seems to produce worse faithfulness and coherence scores. The results hint at the possibility that KD might impact significantly the quality of the segmentation maps, with distilled models generically recording lower values than their counterpart trained with classification loss. Also, especially regarding coherence, both this model and the teacher seem to score better results in both faithfulness and coherence. Finally, considering the specific XAI tools, it seems that GradCAM++ is consistently the least coherent, and is also very unstable, thus it should possibly be dropped in favor of the other three.

Our work is, though, a smaller-scale experiment which only considers (a) one data modality (images), (b) one choice of student-teacher pair, and (c) one specific dataset, which, despite being a subset of the very famous Imagenet, might still not be representative of the domain of images. In addition, there are different techniques for producing saliency maps, such as LIME and SHAP, and we only considered faithfulness and coherence as metrics for functionally assessing the saliency maps; other metrics could be considered for future work.

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336â359, October 2019. ISSN 1573-1405.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

[6] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

[7] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[8] Raed Alharbi, Minh N. Vu, and My T. Thai. Learning interpretation with explainable knowledge distillation. *CoRR*, abs/2111.06945, 2021.

[9] Chakkrit Termritthikun, Ayaz Umer, Suwichaya Suwanwimolkul, Feng Xia, and Ivan Lee. Explainable knowledge distillation for on-device chest x-ray classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, page 1â12, 2023.

[10] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7457â7476, June 2023. ISSN 1939-3539.

[11] Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021.

[12] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[13] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[14] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[17] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[18] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[19] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.