Right for the Right Reason: Making Image Classification Robust

Anna Nguyen
Karlsruhe Institute of Technology
Karlsruhe, Germany
anna.nguyen@kit.edu

Adrian Oberföll
Karlsruhe Institute of Technology
Karlsruhe, Germany
adrian.oberfoell@student.kit.edu

Michael Färber
Karlsruhe Institute of Technology
Karlsruhe, Germany
michael.faerber@kit.edu

ABSTRACT

Convolutional neural networks (CNNs) have achieved astonishing performance on various image classification tasks. Although such models classify most images correctly, they do not provide any explanation for their decisions. Recently, there have been attempts to provide such an explanation by determining which parts of the input image the classifier focuses on most. It turns out that many models output the correct classification, but for the wrong reason (e.g., based on irrelevant parts of the image). In this paper, we propose a new score for automatically quantifying to which degree the model focuses on the right image parts. The score is calculated by considering the degree to which the most decisive image regions – given by applying an explainer to the CNN model – overlap with the silhouette of the object to be classified. In extensive experiments using VGG16, ResNet, and MobileNet as CNNs, Occlusion, LIME, and Grad-Cam/Grad-Cam++ as explanation methods, and Dogs vs. Cats and Caltech 101 as data sets, we can show that our metric can indeed be used for making CNN models for image classification more robust while keeping their accuracy.

ACM Reference format:
Anna Nguyen, Adrian Oberföll, and Michael Färber. 2016. Right for the Right Reason: Making Image Classification Robust. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Deep convolutional neural networks (CNNs) are the state-of-the-art method for image classification. Despite achieving a high accuracy in numerous scenarios, these models do not provide any explanation on why a decision was made (i.e., what the decisive features for classification were). This circumstance often limits the interpretability and therefore the users' trust in the model and its application. For users to trust a model, we assume that it should focus on the relevant features a user would also focus on [7]. In the case of image classification, these features are the image regions that are decisive for determining the class. The phenomenon that an image is correctly classified but due to irrelevant features is known as "classifying right for the wrong reason" [8]. An example is given in Figure 1, where snow is taken by the CNN as the decisive feature for recognizing wolfs. In case many images of wolfs with snow are in the training data set, a CNN might exploit such correlations and misclassify a husky as a wolf [7]. However, a CNN should still be robust enough to classify images of such cases correctly.

(a) Husky classified as wolf (b) Explanation

Figure 1: Raw data and explanation of a bad model's prediction in the *Husky vs. Wolf* task. Figure from [7].

In the past, several methods have been proposed that provide visual explanations of CNN outputs by highlighting image regions that are most likely to contribute to the predicted class. Zeiler and Fergus [12] showed with occlusion experiments that a classification model is sensitive to local structures in an image while training. Ribeiro et al. [7] introduced a local interpretable model-agnostic explanation (LIME) that approximates the classifier locally in an interpretable way. Grad-Cam [10] and Grad-Cam++ [1] use gradients of the last layer to get the importance weights for the predicted class. Thus, these methods allow users to manually investigate decisive features on the image. Other methods use the insights of those explanation methods to improve the classification model. For example, Ross et al. [8] penalize the gradients that lie outside of an object mask which indicates relevant features of the input. Schramowski et al. [9] approach the task of correcting a model by including a human in the loop who revises the explanation. Jia et al. [5] remove protected concepts (e.g., gender, race, background) in their approach to obtain a better model by learning an agnostic representation without those information. However, in general, all these approaches rely on human experts or artificially generated data sets to show their improvement via explanation methods such as the above mentioned approaches. Approaches for exploiting such visual explanations in an automated way, e.g., for calculating degrees of being *right for the right reason* without human interaction, are to the best of our knowledge missing so far.

In this paper, we use automatically generated explanations to evaluate the degree to which a given CNN model classifies images *right for the right reason* and is therefore robust in real-world settings (in which features available during training might not occur anymore). This robustness evaluation is based on a novel *quality score* that quantifies the model's performance concerning its explanation. The score is calculated by comparing the regions which are decisive according to an explanation method (e.g., [1, 7, 10, 12])

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

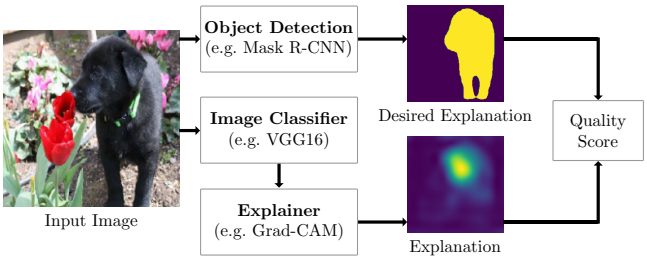


Figure 2: Overview of our approach.

with the image regions of the object to be detected (determined automatically by applying object masking). By setting the object mask of the object to be classified as the desired explanation, we imitate the humanfis capability of considering and classifying objects without exploiting irrelevant image regions, such as distracting background information (without overemphasizing detailed object parts). We argue that our proposed quality score for robustness can be a natural extension for evaluating CNN models alongside existing metrics such as accuracy. Furthermore, our proposed score can also be used to measurably improve the CNN model's quality (i.e., robustness against noise) in case of low-resource training. Using VGG16, ResNet, and MobileNet as CNNs, Occlusion, LIME, and Grad-Cam/ Grad-Cam++ as explanation methods, and Dogs vs. Cats and Caltech 101 as data sets, we show through experiments that the quality score can be used to quantify the robustness of models. Overall, our score has the potential to generate models that can generalize better on unseen data and thus increase the userfis trust in the model.

Overall, our main contributions are as follows:

(1) We propose a quality score metric for quantifying the robustness of classification models automatically on large scale data sets. With our metric, models can be evaluated regarding their explanation next to accuracy.
(2) We perform extensive experiments on various data sets and CNN models.[1] We can show that our metric can be used for making CNN models for image classification more robust while keeping their accuracy.

2 APPROACH

An overview of our approach is outlined in Figure 2. Given an input image on which objects should be detected robustly, we first apply an object detection method (e.g., Mask R-CNN) to obtain the image regions of the object itself (i.e., image silhouette). Our underlying assumption is that parts of the explanation that lie outside of the silhouette are indicative of a classification for the wrong reasons, and that, consequentially, parts of the explanation that lie inside of the silhouette are indicative of a classification for the right reasons. In this way, the obtained object masks serve as a lower bound of an ideal explanation. Note that the silhouette of objects on images can be obtained with a high quality nowadays (see Section 3).

Simultaneously, a CNN model (e.g., pretrained VGG16) is applied to obtain labels of recognized objects (e.g., "dog"). An explanation method (e.g., Grad-Cam) then outputs the image regions which are most influential given the CNN model and the input image. Both the object mask and the explanation output is then used to compute the quality score with respect to robustness. Since the explanation methods support different highlighting levels, our score is constructed in such a way that the score is the higher the more the highlighted explanation lies in the object mask. In the following, we describe the computation of the quality score in more detail.

Given a data set D with correctly classified images and an image $d \in D$ with pixels p^d_{ij}, width w^d, and height h^d, let A^d denote the matrix whose values a^d_{ij} equals the activation of the pixels of the object mask, where $i \in \{1, \ldots, h^d\}, j \in \{1, \ldots, w^d\}$, $h^d, w^d \in \mathbb{N}$. We regard A^d as a fuzzy set, i.e. whose values have degrees of membership depicted as a^d_{ij}. We define $a^d_{ij} \in \mathbb{R}$ with $0 \leq a^d_{ij} \leq 1$ where $a^d_{ij} = 1$ if the pixel p^d_{ij} of the input image belongs to the object and $a^d_{ij} = 0$ otherwise. In similar manner, let B^d be the matrix whose values b^d_{ij} equals the activation of the pixels of the explanation. We additionally normalize the values b^d_{ij} between zero and one, i.e. $0 \leq b^d_{ij} \leq 1$ where $b^d_{ij} = 1$ if the pixel p^d_{ij} of the input image belongs to the highest activation and $b^d_{ij} = 0$ otherwise. Our quality score is, then, defined as follows:

$$\text{Score}(A^d, B^d) = \frac{\sum_{i,j} a^d_{ij} b^d_{ij}}{\sum_{i,j} b^d_{ij}} \in [0, 1] \tag{1}$$

The score measures the relative value of the explanation lying in the object mask. Our score differs from the weighted Jaccard index in the fact that the weighted Jaccard index would be high only if the entire explanation intersects with the whole object (as it measures the similarity of the explanation and the object).

The quality score is applied on all images in a data set D. We use the average of all quality scores for an image collection as the aggregated score:

$$\text{AvgScore}(D) = \frac{1}{n} \sum_{d=1}^{n} \text{Score}(A^d, B^d) \in [0, 1], \tag{2}$$

where $n \in \mathbb{N}$ is the number of images in data set D. We use this average quality score for evaluating a CNN model w.r.t. its explanation (i.e., its robustness). AvgScore only considers the scores of images classified correctly by the model because we want to evaluate if images are classified right for the right reasons. Therefore, images which are classified wrong are excluded.

The quality score is a relative value and significantly depends on the explanation method and the classifier used. Since the score depends on the specific architecture of a CNN, it only allows to compare different training states within the model. Based on the change of the score during training, it can be evaluated if a certain training strategy leads to an improvement or deterioration of the model's robustness given by the explanation. If the quality score improves for a certain explanation method the improvement should be similar to the improvement of the other explanation methods, even though the absolute values vary.

Besides using our quality score for measuring the robustness of already given CNN models, our quality score enables developers to make models more robust (i.e., classifying right for the right reason)

[1] We provide the source code online at https://www.dropbox.com/s/7lusyqu6qapvh92/CIKM2020-code.zip?dl=0 and will publish it on GitHub after acceptance.

via additional training. By using transfer learning techniques (i.e. freezing certain layers by fixing their weights and further training the remaining layers with the same or another data set), we can systematically monitor the robustness alongside established metrics like accuracy and use this strategy to obtain classifiers which generalize well.

3 EVALUATION

3.1 Evaluation Data Sets

We use the following data sets in the evaluation:

Dogs vs. Cats. We use the Dogs vs. Cats[2] data set, which is provided by Microsoft Research and has served as data set for a widely used kaggle challenge. The data set contains 3,000 dog and cat images, 1,500 per class. We can assume that this data set is representative to a real-world data set with respect to the number of classes and images. We use Mask R-CNN [2] to create the object masks on images. Given the data set size, we used 70% of the images for training and 30% for testing. The quality of the object masking is essential for the validity of the proposed quality score. We thus manually evaluated the quality of the computed object masks for 200 randomly chosen images. It turned out that 182 out of the 200 images had an excellent quality. We thus argue that Mask R-CNN performs well for our purpose.

Caltech 101. A more sophisticated and widely used data set is Caltech 101 [6] with 101 object categories built by the California Institute of Technology. We create a uniform distributed data set by drawing random sampling from the categories resulting in a total of 6,060 images with 60 images per class. We use a test split of 0.25. This data set is provided with hand-labeled object masks for all images. Thus, we use those labeled object masks as desired explanation.

3.2 Evaluation Setting

Our experiments are executed on a server with 12 GB of GPU RAM. We use TensorFlow and the Keras deep learning library to build and train deep neural networks.

To demonstrate our score, we re-use trained CNN models. In particular, we rely on transfer learning. Transfer learning is a technique to adapt a trained neural network for a problem to a similar problem. Several layers from the trained model are re-used on a new model. For our experiments, we focus on three state-of-the-art image classification models: *VGG16* [11], *ResNet50* [3], and *MobileNet* [4]. The models are pre-trained on the *Large Scale Visual Recognition Challenge 2012* (ILSVRC2012) data set. We adapt each model's upper output dense layers to the specific data set (i.e., number of categories in the used image classification data sets *Dogs vs. Cats* and *Caltech 101*, respectively). In our experiments, the AvgScore settled around a fixed value after 50 images. For that reason and due to high computing power costs in case of LIME, we calculate the AvgScore for 50 images per epoch in the following experiments.

Dogs vs. Cats. To demonstrate the benefits of our score, we evaluate several transfer learning strategies. We first adjust the output layer of all models to the two categories (dog and cat) and train them for 10 epochs on the Dogs vs. Cats data set (where all layers except output layer are frozen). After that, we freeze different combinations of layers for further training. In the original papers of the models, the convolutional layers are divided in five blocks. For simplification and comparability, we use this convention for our strategies. Thus, we always set whole blocks of layers to either be trainable or non-trainable during training. We also summarize the last dense layers to one block. We train every strategy a further ten epochs. We investigate the following strategies for further training:

(a) train the last dense layers which we denote as dense block,
(b) train the last two convolutional blocks, i.e. the fourth and fifth convolutional block,
(c) train the first three convolutional blocks, i.e. the first, second and third convolutional block,
(d) train all layers, i.e. all convolutional and dense blocks.

Caltech 101. In addition to the experiment above, we perform another experiment inspired by [8, 9]. To actively force the model to be more robust and thus to provide a better explanation, we followed a naive approach by using artificial images in the transfer learning process. We edit the images in a way that it only contains the object to classify and masked out the background with random pixels. This should force the model to focus more on the object and increase the quality score.

3.3 Evaluation Results

Dogs vs. Cats. Figure 3 shows the results for VGG16 with training strategies (a) and (b). We can see that the performance of the model measured with accuracy did not change within ten epochs (see Figure 3 (a)/(b) left graph). However, we observed a change in AvgScore (see Figure 3 (a)/(b) right graph). The quality score after ten epochs computed with any explanation method for strategy (b) is significantly higher than the score for strategy (a). This fits to the general knowledge that complex structures in the input images are learned in the later convolutional blocks and are therefore more decisive for the classification. The results of strategy (d) and (b) and the results of strategy (c) and (a) are similar to each other respectively, which emphasizes that the last convolutional layers are important since they focus more on the important features. Without using the proposed quality score this improvement would not be evident since the accuracy of all models is about the same.

In Figure 4 (a), we provide examples of the explanations visualized with Grad-Cam before (upper images) and after transfer learning (bottom images) with strategy (b) on VGG16. We can see that the score increases for both examples after training and that the visualized explanation has a stronger focus on the object. With only ten epochs of additional training, we were able to improve the model so that it utilizes more important features such as the face of the animal. Without our score, it would be obvious to not train the model any further due to the non-changing accuracy. We can see in the examples that the bottom images focus on more specific features (i.e., the face) to classify dogs and cats. Thus, the model generalizes better on unseen data.

We also performed the described experiment on the CNN models ResNet50 and MobileNet and on the Caltech 101 data set. We observed similar results and do not show them due to page limitations.

[2] https://www.kaggle.com/c/dogs-vs-cats, last accessed: 2020-04-19

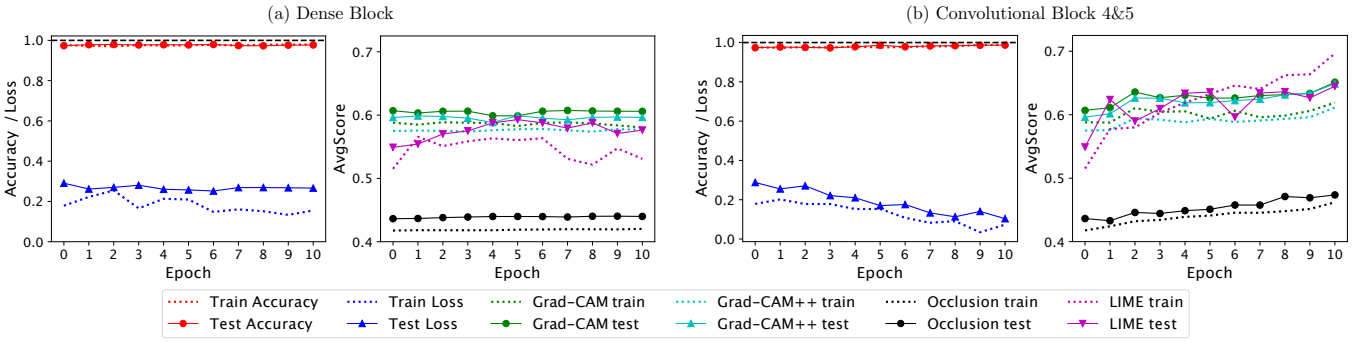
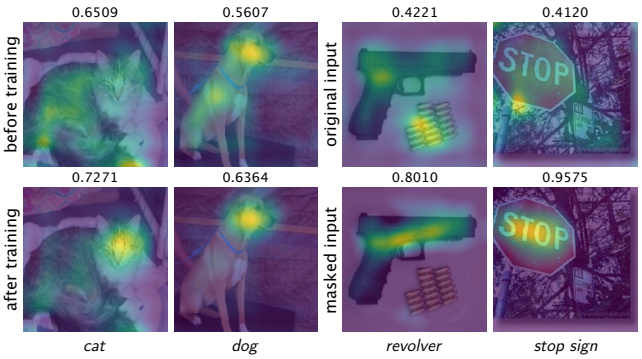


Figure 3: VGG16 Results. Transfer learning strategies with VGG16 with explanation methods Occlusion, LIME and Grad-Cam/Grad-Cam++.



(a) Dogs vs. Cats (b) Caltech 101

Figure 4: Example images of explanation via Grad-Cam on VGG16 after transfer learning using (a) Dogs vs. Cats and (b) Caltech 101 data set. Quality scores are shown above the images, class labels at the bottom.

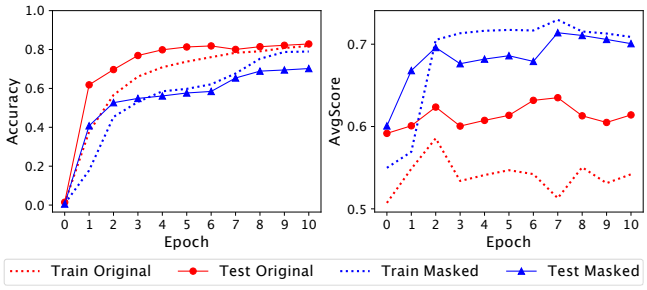


Figure 5: Transfer Learning on Caltech 101 with original and masked images.

Caltech 101. Figure 5 shows the results for ten epochs of transfer learning VGG16 on Caltech 101 with the original and masked images as input. As we can observe in the left graph, training with the original images results in a higher accuracy than training with the masked images. However, the quality score (computed with Grad-Cam as explainer, see graph on the right) of the model trained with masked input images is significantly higher than the quality score of the model trained with the original input images. This indicates that confounding factors are important to consider with respect to classification and that evaluating image classifiers beyond accuracy can be very fruitful.

Figure 4 (b) shows example images with explanations. Despite high accuracy, we can see that the explanations for the images where we masked out the background before explanation (images at the bottom) are more intuitive and more focused on the actual objects than the original input images. We also examined this approach on the Dogs vs. Cats data set but with no worth mentioning results. The reason might be that the models used already use the object's silhouette for classification.

4 CONCLUSION

In this paper, we focused on evaluating CNN image classifiers regarding their explainability. We introduced a novel quality score to support the training process besides the accuracy and loss function. We have shown in our experiments that our quality score can be used to counteract cases where a model makes its predictions based on wrong features. Overall, our quality score enables us to train models that can generalize better (i.e., are more robust) and that can furthermore increase the user's trust in the model by focusing the model on the object of interest.

REFERENCES

[1] Aditya Chattopadhyay, Anirban Sarkar, et al. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *Proc. of WACV'18*. 839–847.
[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2020. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2 (2020), 386–397.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. of CVPR'16*. 770–778.
[4] Andrew G. Howard et al. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR* abs/1704.04861 (2017).
[5] Sen Jia, Thomas Lansdall-Welfare, et al. 2018. Right for the Right Reason: Training Agnostic Networks. In *Proc. of IDA'18*. 164–174.
[6] Fei-Fei Li, Robert Fergus, et al. 2007. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* 106, 1 (2007), 59–70.
[7] Marco Túlio Ribeiro, Sameer Singh, et al. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. of SIGKDD'16*. 1135–1144.
[8] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proc. of IJCAI'17*. 2662–2670.

[9] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, et al. 2020. Right for the Wrong Scientific Reasons: Revising Deep Networks by Interacting with their Explanations. *CoRR* abs/2001.05371 (2020).
[10] Ramprasaath R. Selvaraju, Abhishek Das, et al. 2016. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *CoRR* abs/1610.02391 (2016).
[11] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. of ICLR'15*.
[12] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Proc. of ECCV'14*, Vol. 8689. 818–833.