Interpreting and Explaining Deep Models Visually

Jose Oramas M.[1,2], Kaili Wang[1], and Tinne Tuytelaars[1]

KU Leuven, ESAT-PSI, Belgium   UAntwerpen, IDLab, Belgium

Methods based on deep neural networks (DNNs) have achieved impressive results for several computer vision tasks, such as image classification, object detection, and image generation, etc. Combined with the general tendency in the community of developing methods with a focus on high quantitative performance, this has motivated the wide adoption of DNN-based methods, despite the initial skepticism due to their black-box characteristics. Our goal is to bridge the gap between methods aiming at model *interpretation*, i.e., understanding what a given trained model has actually learned, and methods aiming at model *explanation*, i.e., justifying the decisions made by a model.

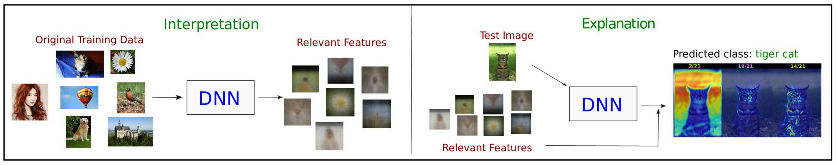
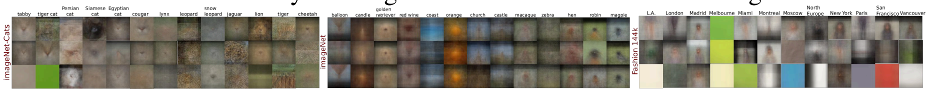
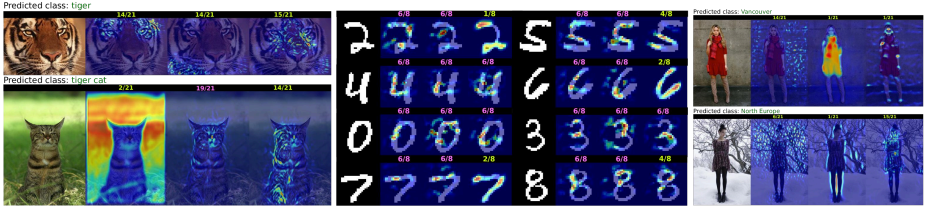


Fig. 1: Proposed Interpretation / Explanation pipeline.

Model Interpretation. Interpretation of DNNs is commonly achieved in two ways: either by a) manually inspecting visualizations of every single filter (or a random subset thereof) from every layer of the network ([7,8]) or, more recently, by b) exhaustively comparing the internal activations produced by a given model w.r.t. a dataset with pixel-wise annotations of possibly relevant concepts ([1,3]). Here we reduce the amount of computations by identifying a sparse set of features encoded by the model (internally) which could serve as indicators for the semantic concepts modelled by the network. More specifically, through a μ-Lasso formulation, a set of relevant layer/filter pairs are identified for every class of interest j. This results in a relevance weight w_j, associated to class j, for every filter-wise response x computed internally by the network (Fig.1). As shown on the image below, we produce average visualizations of these features to enable visual interpretation of the model. Moreover, we remove the dependence on external additional annotated data by re-using the same data use to train the original model.



Model Explanation. We "explain" the predictions made by a deep model, by accompanying its predicted class with a a set of heatmap visualizations (Fig.). Having identified a set of relevant features (indicated by W) for the classes of interest, we generate feedback visualizations by taking into account the response of these features on the content of a tested images. A test image I is pushed through the network producing the class prediction $\hat{j}=F(I)$. Then, taking into account the internal responses x, and relevance weights $w_{\hat{j}}$ for the predicted class $\hat{j}$, we generate visualizations indicating the image regions that contributed to this prediction.



Evaluating Visual Explanations We propose, *an8Flower*, a synthetic dataset where the feature defining the classes of interest is controlled by design. This allows to compute binary masks indicating the regions that should be highlighted by an explanation heatmap. This enables objective means to assess the accuracy of these visualizations. In Tab. 1, we show quantitative performance of the proposed method. In the figure below, we show some qualitative examples of the visual explanations and interpretations from our method. Overall our method achieves a better balance between level of detail and coverage of the relevant features than those produced by existing methods. Please see [4] for more details.

Table 1: Area under the IoU curve (in percentages) on an8Flower over *5-folds*.

Method	**single-6c**	**double-12c**
Upsam. Act.	16.8±2.63	16.1±1.30
Deconv+GB, [6]	21.3±0.77	21.9±0.72
Grad-CAM, [5]	17.5±0.25	14.8±0.16
Guided Grad-CAM, [5]	19.9±0.61	19.4±0.34
Grad-CAM++, [2]	15.6±0.57	14.6±0.12
Guided Grad-CAM++, [2]	19.6±0.65	19.7±0.27
Ours	**22.5±0.82**	**23.2±0.60**

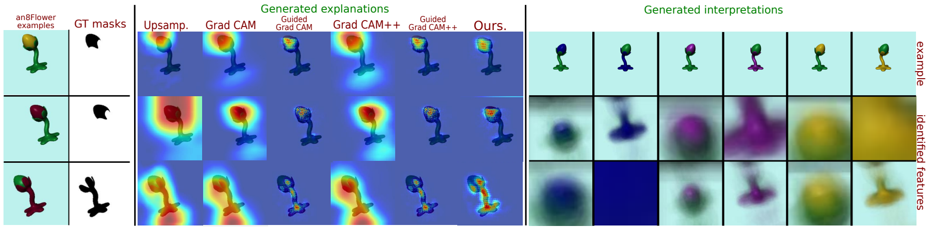


Fig. 2: Left: Examples and GT-masks from the proposed *an8FLower* dataset. Center: Comparison of generated visual explanations. Right: Examples of the generated visual interpretations.

Acknowledgments: This work was supported by the FWO SBO project Omnidrone, the VLAIO R&D-project SPOTT, the KU Leuven PDM Grant PDM/16/131, and a NVIDIA GPU grant.

1. Bau, D., et al.: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR (2017)
2. Chattopadhyay, A., et al.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)
3. Fong, R., et al.: Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: CVPR (2018)
4. Oramas M, J., et al.: Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In: ICLR (2019)
5. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
6. Springenberg, J.T., et al.: Striving for simplicity: the all convolutional net. In: ICLR WS (2015)
7. Yosinski, J., et al.: Understanding neural networks through deep visualization. In: ICML Workshops (2015)
8. Zeiler, M., et al.: Visualizing and understanding convolutional networks. In: ECCV (2014)