Motivation

A deep neural network is a structure of neural layers, where each neuron receives as an input the activations from the previous layer, computes a weighted sum and applies a nonlinear activation. All the neurons perform a nonlinear mapping from the input data to the output. To learn the mapping the weights of neurons are adapted by backpropagation [1].

It has been proven that after training an artificial neural network on a large data set, the network can recognize a category of unseen before objects and do it even better than a human [2]. One of the broadly used examples nowadays is the ImageNet annual challenge [3], where the goal is to make a classification algorithm for 1000 categories of objects presented on photographs. The ImageNet challenge provides a huge data set of so-called natural images. This data set became a widely used benchmark for evaluation of neural network performance, but also for analysis of networks' decision process [4]. With the expanding applicability of neural networks, there is an increasing need to explain their decisions.

Approaches

The currently available approaches for understanding neural networks' reasoning can be divided into two categories [5]: (1) techniques that interpret models and parts of networks, and (2) methods explaining decisions focusing on data. Model interpretation techniques search for a prototypical example of a category or a pattern maximizing the activity of a neuron. Decision explaining methods aim at finding the reason why the model makes the particular prediction and verify that the model outputs reasonable decisions.

Model interpretation

Many model interpreting techniques are based on activation maximization (AM) [6]. AM is an optimization approach which synthesizes an input pattern that maximizes the activation of a chosen neuron. The activation of a neuron is a function of parameters, namely weights and biases, and the input sample. The parameters are assumed to be fixed after training the network. The activation is maximized by gradient ascent in the image space: the gradient of the activation is computed and the input is shifted in the image space in the direction of the gradient.

The Deconvolutional Network [7] [8] (DeconvNet) projects patterns exciting a neuron on the input image. The DeconvNet is basically any convolutional neural network (CNN) exploited in the reverse order and each layer of the DeconvNet is attached to the original net. To project an activation of a CNN’s neuron, the activations of the other neurons in the same layer are set to zero and the feature maps produced by this layer are passed to the DeconvNet, where the activations of the layer beneath are reconstructed and passed to the next layer below. This reconstruction is repeated for all layers under the examined one and produces an image showing the part of the input that activated the neuron. In 2013 Simonyan et al. [4] modified AM method by adding a regularization and applied to the ImageNet dataset. In 2016 Nguyen et al. [9] proposed the Deep Generator Network, which uses the Generative Adversarial Network architecture to improve the AM algorithm by generating realistic images that visualize activations. The above methods not only show the features that the neuron looks at, but also create a class prototype if the neuron is located in the last fully connected layer.

Decision explaining

One of the first approaches to understanding the decisions of classification nets is called saliency maps, or sensitivity maps, or pixel attribution maps [4]. It finds regions of an image that were important for the classification and highlights them [7] [10] [11] [12] [13] [14]. The authors use gradients or occlusion [7] to assign a measure of importance to individual pixels. In the occlusion method, parts of the image are systematically covered with a gray square and the changes in the activity of the top feature map and in the classifier output are measured as an importance of the area.

For image classification, a saliency map is a gradient of the class activation function with respect to the input image. The class activation function is the function computed for each possible class, and the class with the highest score is selected [4] [6] [10]. The saliency map shows how much a little change in each pixel would alternate the score for the predicted class. Recently improved versions of this algorithm such as Layerwise Relevance Propagation [15], Integrated Gradients [14], Guided Backpropagation [11], CAM [12], GradCAM [13], patternNet [16], PatternAttribution [16] and SmoothGRAD [17].

Although there are many methods available, none of them is perfect, and the evaluation metric, which mostly based on a human notion, is arguable [16] [18] [19].

References

[1] Rumelhart, D. E.,Hinton, G. E., Williams, R. J.,, *Learning representations by back-propagating errors.,* Nature 323 (6088), 533–536., 1986.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Deep Residual Learning for Image Recognition,* CVPR, 2016.

[3] Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei., *ImageNet Large Scale Visual Recognition Challenge,* IJCV, 2015.

[4] Simonyan, K., Vedaldi, A., Zisserman, A.,, *Deep inside convolutional networks: Visualising image classification models and saliency maps.,* CoRR abs/1312.6034., 2013.

[5] W. Samek & K.-R. Müller, *Tutorial on Interpretable Machine Learning,* GCPR 2017 Tutorial.

[6] Erhan, D., Bengio, Y., Courville, A., Vincent, P., *Visualizing higher-layer features of a deep network.,* Tech. Rep. 1341, 2009.

[7] Zeiler, M. D., Fergus, R.,, *Visualizing and understanding convolutional networks.,* Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham, 2014.

[8] Zeiler, M., Taylor, G., Fergus, R., *Adaptive deconvolutional networks for mid and high level feature learning,* ICCV 2011.

[9] Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J., *Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.,* Advances in Neural

Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. pp. 3387–3395., 2016.

[10] Baehrens, David, Schroeter, Timon, Harmeling, Stefan, Kawanabe, Motoaki, Hansen, Katja, and Muller, Klaus-Robert, *How to explain individual classification decisions.,* Journal of Machine Learning Research, 11 (Jun):1803–1831, 2010.

[11] Springenberg, Jost Tobias, Dosovitskiy, Alexey, Brox, Thomas, and Riedmiller, Martin., *Striving for simplicity: The all convolutional net.,* arXiv preprint, arXiv:1412.6806, 2014.

[12] Zhou, Bolei, Khosla, Aditya, Lapedriza, Agata, Oliva,Aude, and Torralba, Antonio., *Learning deep features for discriminative localization.,* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929, 2016.

[13] Selvaraju, Ramprasaath R, Das, Abhishek, Vedantam, Ramakrishna, Cogswell, Michael, Parikh, Devi, and Batra ,Dhruv., *Grad-cam: Why did you say that?,* ICCV, 2017.

[14] Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi., *Axiomatic attribution for deep networks.,* arXiv preprint, arXiv:1703.01365, 2017.

[15] Bach, Sebastian, Binder, Alexander, Montavon, Gregoire, Klauschen, Frederick, Muller, Klaus-Robert, and Samek, Wojciech., *On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation.,* PloS one, 10(7):e0130140, 2015.

[16] Kindermans, P., Schutt, K.T., Alber, M., Muller, K., Erhan, D., Kim, B., & Dahne, S., *Learning how to explain neural networks: PatternNet and PatternAttribution.,* arXiv preprint, arXiv:1705.05598, 2017.

[17] Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., & Wattenberg, M., *SmoothGrad: removing noise by adding noise.,* CoRR, abs/1706.03825., 2017.

[18] D. Bau*, B. Zhou*, A. Khosla, A. Oliva, and A. Torralba., *Network Dissection: Quantifying Interpretability of Deep Visual Representations.,* Computer Vision and Pattern Recognition (CVPR), 2017.

[19] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres, *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).,* ICML, 2018.

Please cite the repository if you use this text in your research.

```
@misc{Vinogradova2019,  author = {Vinogradova, K.},  title = {Your CNN Interpretation},  year = {2019},  publisher = {GitHub},  journal = {GitHub repository},  howpublished = {\url{https://github.com/kiraving/cnn_interpret_kv}}}
```