Towards Frequency-Based Explanation for CNN

Zifan Wang (zifanw) [1] **Yilin Yang (yiliny2)** [1] **Ankit Shrivastava (ashriva2)** [1] **Varun Rawal (vrawal)** [1]

Overview

The gap between human's understanding and the logic behind Deep Neural Networks (DNNs) are receiving more attention as DNNs show competitive inference power in areas where humans used to be indispensable, e.g. medical diagnosis, auto-piloting, and credit systems. Entrusting users by explaining models' behavior becomes as necessary as promoting the performance. Recent work of explanations focuses on associating importance of input features either to the model's output (e.g. Karpathy et al. (2015), Sundararajan et al. (2017), Selvaraju et al. (2019)) or to the distribution of data with which the model is trained (e.g. Koh & Liang (2017), Yeh et al. (2018), Leino et al. (2018)); however, the community may overlook the another facet of input data – the distribution of frequency components. Treating input as signals in R^n space, we can decompose them into basis signals of diverse frequencies. While humans may only respond to a specific range of frequencies, DNNs are capable of using information from all frequencies. We aim to explore how DNNs respond to the distribution of input frequencies to develop its inference logic. **Specifically, we want to answer the following questions**: i) given an input x, can we quantify the contribution of each basis component to a model's output? ii) given a dataset D, can we increase the alignment of the frequency components of the model that aligns with those humans can respond to.

In this project, we focus on Convolutional Neural Networks (CNNs), a portion of DNNs that are most frequently studied and applied for image-related tasks. **Challenges to this project include**: i) A numerical evaluation of the contribution of each different frequency component (frequency attribution). ii) Formation of reasonable hypotheses from plenty of experiments and verify the hypotheses. Understanding the network behaviors are different from promoting the performance on a specific task. Each proposition requires careful justifications and sufficient experiments. iii) Visualization of the importance of frequencies in the time domain. As the most effective approach to convince the human about the model's logic, visualizing frequencies can be harder than visualizing a portion of input features with heatmap.

[1]Carnegie Mellon University, Pittsburgh, PA 15213, USA.

10708 Class Project, Spring, 2019.

Related Work

We provide a survey of related work in DNNs explanations and using frequency analysis to interpret the concepts learned by CNNs.

One of the instance-based explanations is attribution methods, which explain models prediction by computing a contribution score for each feature in the input. An intuitive method is using the gradient of some user-defined quantity of interests, e.g. the prediction class, for the input (Simonyan et al., 2013). However, the vanilla gradient method suffers from vanishing gradient issues. *Integrated Gradients*((Sundararajan et al., 2017) demonstrates a way to solve the vanishing gradient by aggregating gradient along a path from a user-defined baseline to the input. *Integrated Gradients* is designed to satisfy *completeness* criteria, the sum of attribution scores is equal to the model's output change from the chosen baseline. Instance-based attribution methods associate the model's prediction for a particular input back to the relevant input features, but they are not able to illustrate high-level concepts learned by the model. *Influence-directed explanations* (Leino et al., 2018) explains the behavior of CNNs by introducing *distributional influence*, a quantity measuring how the distribution of inputs affects the hidden layers. In this way, *influence-directed explanations* identifies expert neurons in the hidden layer that can illustrate the learned concepts shared by a given distribution of instances.

Besides in the original feature space, several works provide new insights into the CNN behaviors from the aspect in the frequency domain. As the theory behind Discrete Fourier Transform (DFT), input signals can be decomposed into multiple base components with different frequencies. Human's understanding of an image may rely on the low frequencies components since high-frequency components are sensed more like noises or details. Wang et al. (2019) shows that unlike human beings, high-frequency components play significant roles in promoting CNN's accuracy. Adopting information from high-frequency components may cause the model to form very different concepts in learning as humans do. By observing adversarial defended models, Wang et al. (2019) concludes that

smoothing the CNN kernels helps to enforce the model to use features of low frequencies. While the conclusion remains questionable due to the lack of theoretical proof is discussed, the paper proposes a novel view of attributing the frequencies components to the model's predictions.

An alternative view of understanding the importance of frequency components is to observe a model's behavior when specific frequency components are modified. This area is studied known as the frequency components analysis on adversarial examples. Guo et al. (2018) proposes an adversarial attack only targeting the low-frequency components in an image, which shows that the model does utilize the features in the low-frequency domains for predictions instead of only learning from high-frequency components. Sharma et al. (2019) demonstrated that state-of-the-art defenses are nearly as vulnerable as undefended models under low-frequency perturbations, which implies current defense techniques are only valid against adversarial attack in the high-frequency domain. On imperfection in Sharma et al. (2019)'s work is that there is no comparison between the average distortion required by the low-frequency attacks and the average distortion required by the high-frequency attacks, leaving the result being valid adversarial questionable. The discussion of whether a model is relying more on low or high-frequency components to form generalize concepts may eventually fall into the discussion of *robust* and *non-robust features* (Ilyas et al., 2019), where *non-robust features* along are sufficient good generalization while *robust features* align better with human's perceptions.

Dataset

We will use the CIFAR-10 dataset (Krizhevsky et al.) for our experiment. It consists of 60000, 32×32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. We will also use MNIST (LeCun & Cortes, 2010) for testing and development, and may involve a small portion of ImageNet (Deng et al., 2009) for comparisons.

Plan of Activities

In summary, the goal of our project is to understand how different frequency perturbations in input affect the model's decision of classification using the attribution method and then using this information design appropriate defense. To achieve this goal we have divided the work in following components and assigned their responsibility:

- **Design attribution methods by Yilin and Zifan.** Attribution methods will be used to explain the effect of different frequency components on CNN models. The methods will be based on (Simonyan et al., 2013) with planned improvements that compute the contribution score of features for frequency adversarial perturbation to the input. We expect this method to tell us which features of the adversarial image generated using one of the frequency components are most important while classifying the example
- **Design of adversarial attacks by Ankit and Varun.** We will design frequency adversarial perturbations based on the methods discussed in (Sharma et al., 2019) and (Zhang et al., 2019) as an verification of which part of frequency components are more robust.
- **Design of defense** The adversarial defense will be designed based on the information provided by attribution methods. The design will be similar in principle to (Sharma et al., 2019) but for higher frequency components with required modification.

Timeline

By 19th, march we planned to design adversarial attacks and attribution methods and discuss the conclusions in our mid report. Later, we will plan to design the defense for different frequency contributions.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Guo, C., Frank, J. S., and Weinberger, K. Q. Low frequency adversarial perturbation, 2018.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features, 2019.

Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions, 2017.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Leino, K., Sen, S., Datta, A., Fredrikson, M., and Li, L. Influence-directed explanations for deep convolutional networks, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, Oct 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Sharma, Y., Ding, G. W., and Brubaker, M. A. On the effectiveness of low frequency perturbations. *CoRR*, abs/1903.00073, 2019. URL http://arxiv.org/abs/1903.00073.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328. JMLR. org, 2017.

Wang, H., Wu, X., Yin, P., and Xing, E. P. High frequency component helps explain the generalization of convolutional neural networks. *CoRR*, abs/1905.13545, 2019. URL http://arxiv.org/abs/1905.13545.

Yeh, C.-K., Kim, J. S., Yen, I. E. H., and Ravikumar, P. Representer point selection for explaining deep neural networks, 2018.

Zhang, H., Avrithis, Y., Furon, T., and Amsaleg, L. Smooth adversarial examples. *arXiv preprint arXiv:1903.11862*, 2019.