Interpretability of Deep Learning

What is interpretability?

Definition of interpretability is not strictly formalized, but there are two distinguishable views of this concept [1, 2, 8]:

- model transparency (as the opposite of black-box) - understanding model internals
- justification of model predictions

Importance of the problem

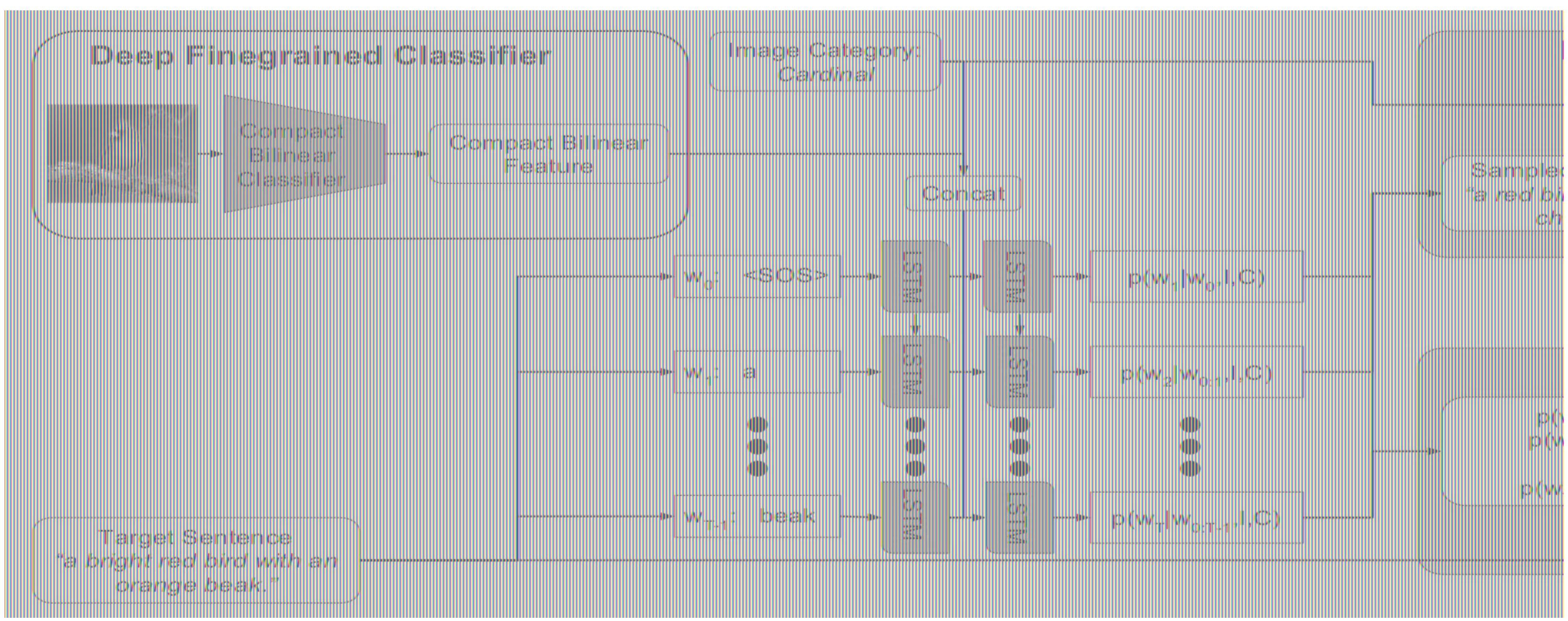
- Better interpretability helps with testing NNs for unexpected behavior which is crucial for critical applications.
  (difficulties of testing are compounded by the vulnerability to adversarial examples [4]
  Although, techniques like in [3] can address issues of testing without improving interpretability)

- If used data falls under EU General Data Protection Regulation, data owner can use his “right to explanation” [5]

- Better interpretability allows to leverage human prior knowledge in case of selecting between different models, while a performance metric can be potentially misleading (as a result of overfitting or discrepancies between a chosen optimized metric and desired model properties) [8]

Generating Visual Explanations

- Authors combined image recognition and image-captioning approaches and proposed a model for generating *visual explanations* which are both image-specific and class-discriminative.



[6] L. A. Hendrics, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darell, “Generating Visual Explanations”, in proc. of the European Conference on Computer Vision, 2016

Visualizing Recurrent Networks

- Character-level language model (CLM) predicts next character based on previous sequence
- Authors have trained different CLMs (LSTM, RNN, GRU) on English version of Leo Tolstoy’s War and Peace novel and source code of the Linux Kernel to study RNN behavior
- Authors have identified multiple interpretable long-range LSTM cells

Several examples of cells with interpretable activations discovered in best Linux Kernel and War and Peace LSTMs.
Text color corresponds to tanh(c), where -1 is red and +1 is blue.

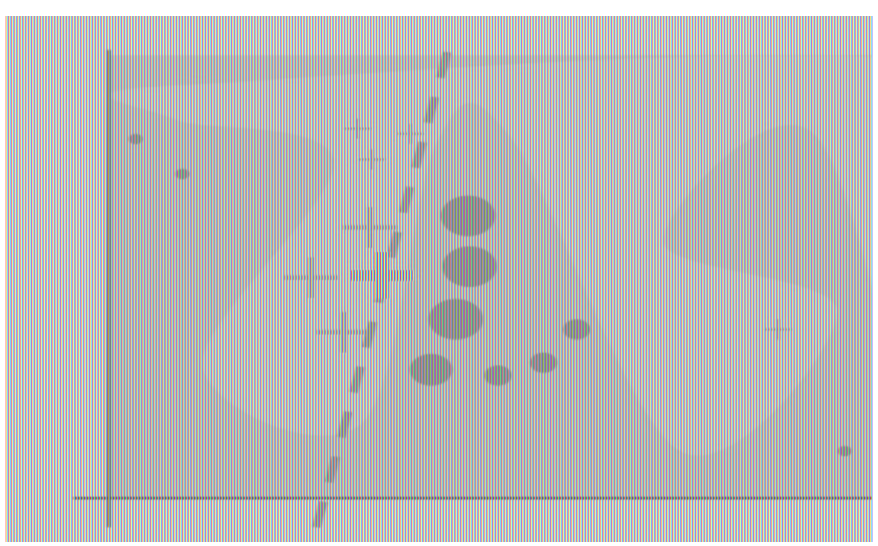
[7] A. Karpathy, J. Johnson, Fei-Fei Li, “Visualizing and Understanding Recurrent Networks”, CoRR, vol. abs/1506.02078, 2015

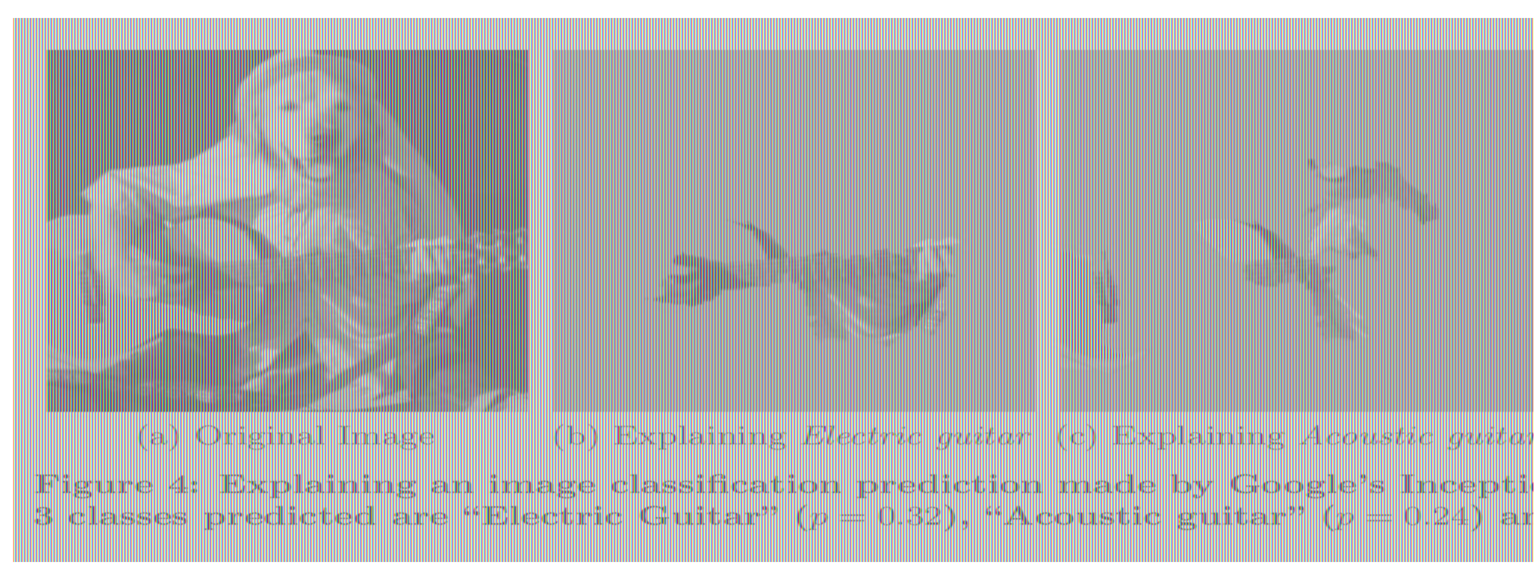
Also Karpathy did a talk on introduction to RNNs and results obtained in this paper
https://skillsmatter.com/skillscasts/6611-visualizing-and-understanding-recurrent-networks

Local Interpretable Model-agnostic Explanations (LIME)

- Authors have proposed a method for justifying prediction of any classifier by finding an interpretable model (e.g. linear) over interpretable representation which is locally faithful

- Interpretable representation needs to be chosen for every task individually and can differ from features used for prediction
(e.g. super-pixel for image classification or bag-of-words for text classification)





[8] M. T. Ribeiro, S. Singh, C. Guestrin "Why Should I Trust You?: Explaining the Predictions of Any Classifier", in proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016

Summary

- Today most of the approaches focus on interpretability of predictions for supervised learning task, especially in the domain of computer vision

- Models for other tasks are still considered as black boxes

References

- [1] Z. C. Lipton "The Mythos of Model Interpretability",
  in proc. of the 2016 ICML Workshop on Human Interpretability in Machine Learning, New-York, NY, June 2016
- [2] Chakraborty, Supriyo, Tomsett, Richard, Raghavendra, Ramya, Harborne, Daniel, Alzantot, Moustafa, Cerutti, Federico, Srivastava, Mani, Preece, Alun David, Julier, Simon, Rao, Raghuveer M., Kelley, Troy D., Braines, David, Sensoy, Murat, Willis, Christopher J. and Gurram, Prudhvi, "Interpretability of deep learning models: a survey of results",
  IEEE Smart World Congress 2017 - Workshop on Distributed Analytics Infrastructure and Algorithms for Multi-Organization Federations, San Francisco, CA, USA, August 2017
- [3] K. Pei, Y Cao, J. Yang, S. Jana, "DeepXplore: Automated Whitebox Testing of Deep Learning Systems",
  in proc. of the 26th Symposium on Operating Systems Principles, pp. 1-18, Shanghai, China, October 2017
- [4] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples",
  in proc. of 3rd International Conference on Learning Representations, San-Diego, CA, USA, May 2015
- [5] B. Goodman, S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"",
  in proc. of the 2016 ICML Workshop on Human Interpretability in Machine Learning, New-York, NY, June 2016
- [6] L. A. Hendrics, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darell, "Generating Visual Explanations",
  in proc. of the European Conference on Computer Vision, pp. 3-19, Amsterdam, Netherlands, October 2016
- [7] A. Karpathy, J. Johnson, Fei-Fei Li, "Visualizing and Understanding Recurrent Networks",
  CoRR, vol. abs/1506.02078, 2015
- [8] M. T. Ribeiro, S. Singh, C. Guestrin "Why Should I Trust You?: Explaining the Predictions of Any Classifier",
  in proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, San Francisco, CA, USA, August 2016
- [9] N. Tsopze, E. Mephu Nguifo, G. Tindo, "Towards a generalization of decompositional approach of rule extraction from multilayer artificial neural network",
  in proc. IJCNN 2011, pp. 1562-1569