HIERARCHICAL INTERPRETATIONS FOR NEURAL NETWORK PREDICTIONS

-Chandan Singh, W. James Murdoch, Bin Yu

-ICLR 2019

Presenter: Rishab Bamrara

https://qdata.github.io/deep2Read/

April 10, 2020

Experiment Summary:

- **Stanford Sentiment Treebank (SST):** Standard NLP benchmark which consists of movie reviews ranging from 2 to 52 words long. In addition to labels of reviews, it also has labels for each phrase in the review.
- **Word Embeddings:** Glove (glove.6B.300d)
- **Model:** Bi-LSTM

Experiment Summary:

- Epochs: 50
- Training Accuracy: 99.9592
- Training Loss: 0.000024
- Dev Accuracy: 80.2752 → 86.2% or 85.8% mentioned in paper
- Dev Loss: 1.410839

Experiment Summary:

PyTorch. For SST, we train a standard binary classification LSTM model[2], which achieves 86.2% accuracy. On MNIST, we use the standard PyTorch example[3], which attains accuracy of 97.7%. On ImageNet, we use a pre-trained VGG-16 DNN architecture Simonyan & Zisserman (2014) which attains top-1 accuracy of 42.8%. When using ACD on ImageNet, for computational reasons, we start the agglomeration process with 14-by-14 superpixels instead of individual pixels. We also smooth the computed image patches by adding pixels surrounded by the patch. The weakened models for the human experiments are constructed from the original models by randomly permuting a small percentage of their weights. For SST/MNIST/ImageNet, 25/25/0.8% of weights are randomized, reducing test accuracy from 85.8/97.7/42.8% to 79.8/79.6/32.3%.

Results:

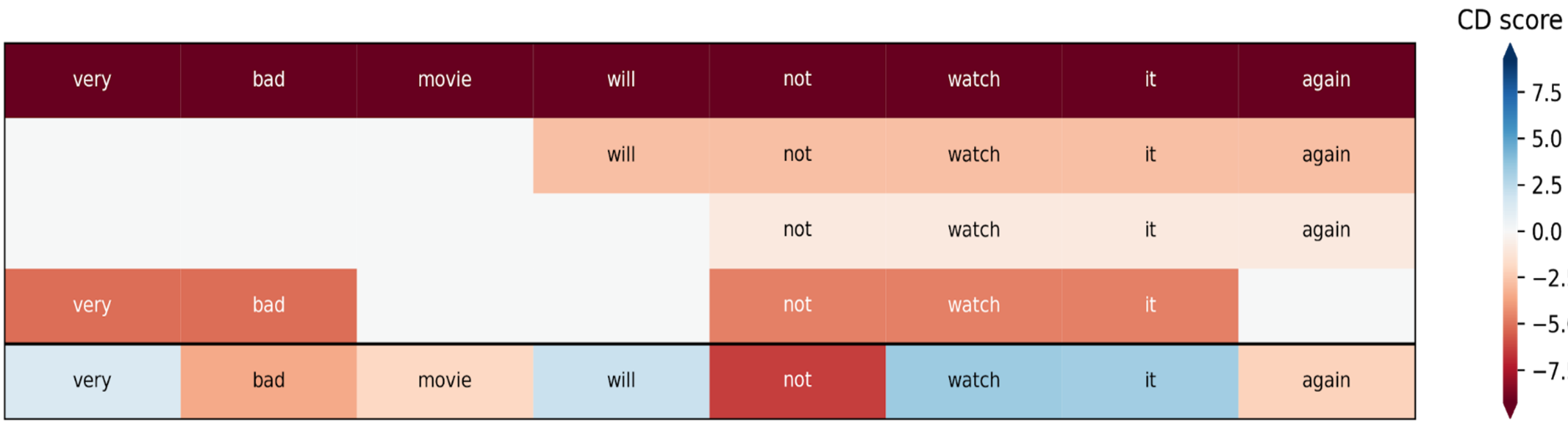

Figure 2: Showing ACD Hierarchical interpretations for a sentence.

References:

- Sebastian Bach, Alexander Binder, Gr´egoire Montavon, Frederick Klauschen, Klaus Robert M¨uller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

- W James Murdoch and Arthur Szlam. Automatic rule extraction from long short term memory networks. ICLR, 2017.

- Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. CoRR, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. CoRR, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

- Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078, 2015.