Applications of Deep Learning to Deception Detection in Speech

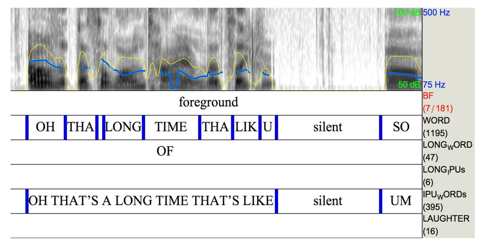
Kai-Zhan Lee (kl2792@columbia.edu), Sarah Ita Levitan, Julia Hirschberg

Spoken Language Processing (SLP) Group – Columbia University in the City of New York

Background

- **Deception** is the deliberate choice to mislead, in order to achieve personal gain or to avoid a penalty.
- The **Columbia Cross-culture Deception Corpus** (CxD) is a collection of transcribed and recorded interviews, each consisting of 24 questions; interviewees lie in response to exactly 12 of these questions, indicating truthfulness with a set of keys, and are rewarded monetarily for successful lies.
- Previous papers have attained attained accuracies up to **9.95%** above majority-class baseline by using random forest classifiers. (Levitan, et al.)
- **Deep learning**, which uses neural networks as classifiers, is a machine learning method that was made possible by the recent rise in computational power.


Experimental setup

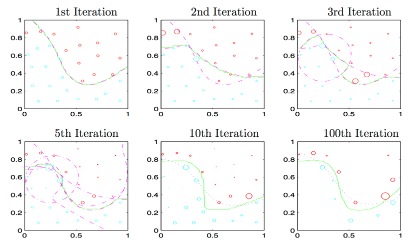

Transcript segment

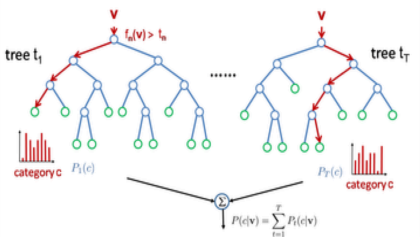
Research Question

How can we optimize neural networks with CxD to best improve on the accuracy of previously-used deception detection classifiers?

Methodology

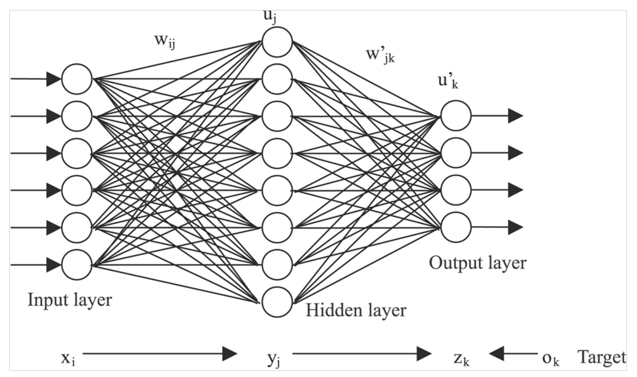
- Interviews were recorded in a sound-proof box and sourced to Amazon Mechanical Turk (MTurk) for transcripts. Transcripts consist of time-stamped intonational phrase units (IPU) for both interview participants. Participants also completed the NEO Five-Factor Inventory personality test and a demographics form.
- These IPUs were merged into 'turns', IPU sequences that are uninterrupted by another speaker. A question-matching script was created to identify questions from the interviewer and extract the turn from the interviewee directly after. Keypresses were used to determine each turn's truthfulness.
- The acoustic feature extractor openSMILE was used to extract 6373 features from each turn, and these were combined with language, gender, and 5 personality scores to form a 6380-feature data set.
- Finally, testing on this data set was performed with ensemble classifiers and neural networks, using various optimizers.

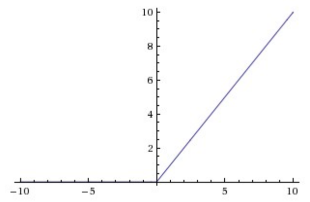

AdaBoost Iterations

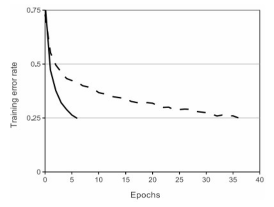

Mechanisms of the random forest classifier

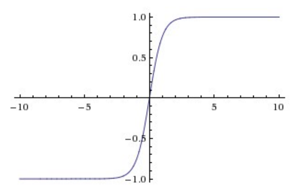
Methodology (cont.)

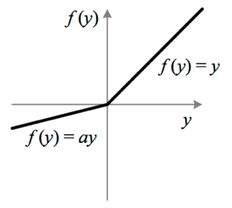
- Choice of optimizer was critical in determining the performance of neural nets.
- Activation function plays a large role as well in rate of convergence.

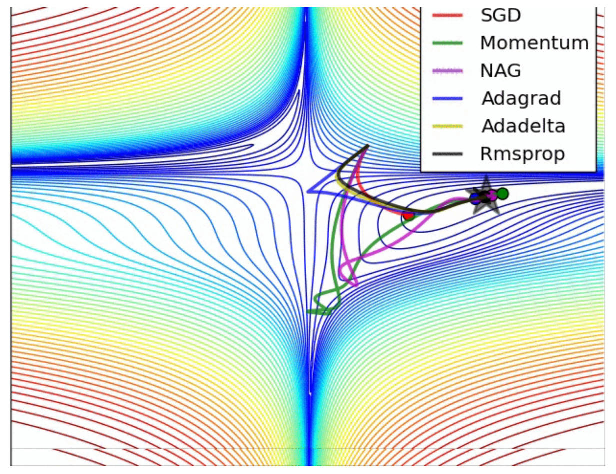

Simplified representation of a neural net


ReLU activation


ReLU vs Tanh

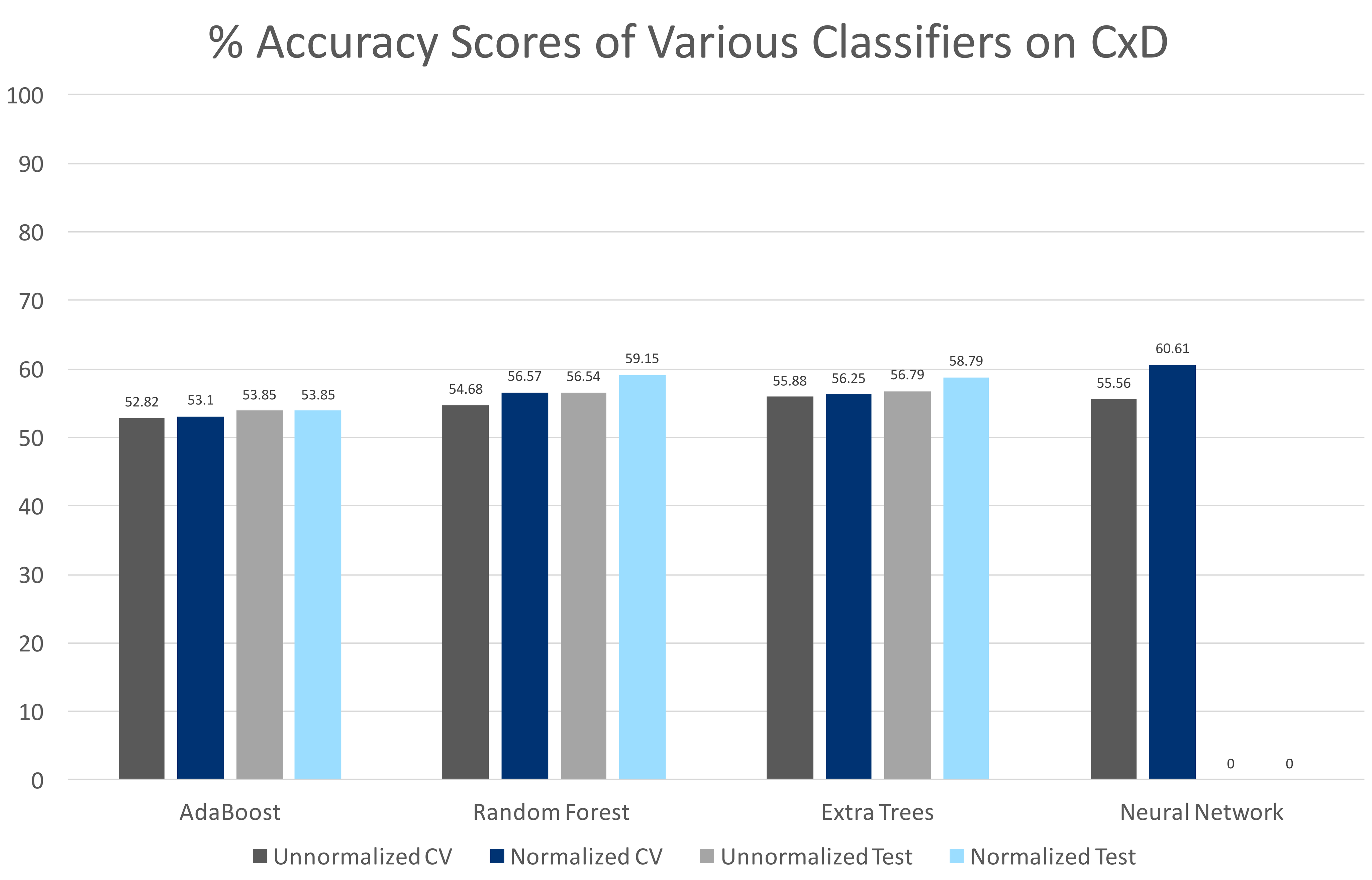

Tanh activation


ReLU vs PReLU



An illustration of optimizers

Results



Notes:

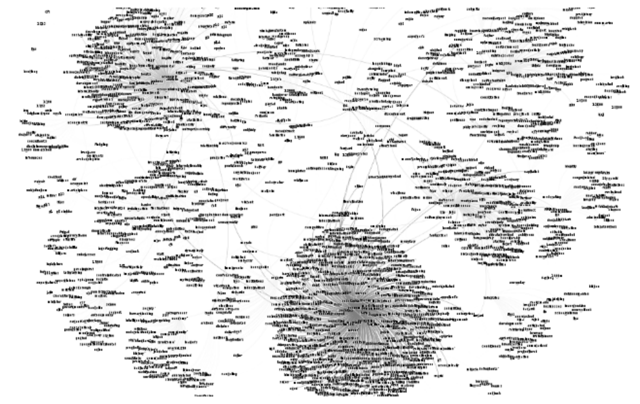
- Majority-class baseline is 51.62% for the training set and 51.15% for the test set.
- When using Nesterov-accelerated AdaDelta, accuracy fluctuated wildly, even at a glacial learning rate of 10^{-8}.
- The neural net was optimized with a Nesterov-accelerated stochastic gradient descent optimizer at a learning rate of 10^{-6}.
- Corroborating the results of Krizhevsky et al., PReLU was the activation layer that resulted in fastest convergence.
- Neural net optimization for CxD is a work in progress. Test accuracies are coming soon!

Summary of Results

- Normalization tends to increase accuracy, regardless of classifier.
- The best ensemble classifier performs at 15.64% above baseline, while the best neural net performs at 18.49% above baseline.
- The neural net's improvement from the random forest classifier is 18.22%.

Discussion

- Neural networks are more than capable of outperforming the best ensemble classifiers.
- openSMILE acoustic features are very effective for determining the veracity of a segment of audio
- Clearly, there is great potential for neural networks in SLP, and in the field of deception detection overall.
- There were only 2160 train samples and 648 test samples; more samples are needed for more robust results; need to improve script for identifying interviewer questions.
- Next step: adding lexical (text) features with word embeddings, multidimensional feature-vectors (sets of numerical values) that represent words.


Simplified depiction of word embeddings

Relationship	Example 1	Example 2
France - Paris	Italy: Rome	Japan: Tokyo
big - bigger	small: larger	cold: colder
Miami - Florida	Baltimore: Maryland	Dallas: Texas
Einstein - scientist	Messi: midfielder	Mozart: violinist
Sarkozy - France	Berlusconi: Italy	Merkel: Germany
copper - Cu	zinc: Zn	gold: Au
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev
Microsoft - Windows	Google: Android	IBM: Linux
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy
Japan - sushi	Germany: bratwurst	France: tapas

Examples of relational equivalence

References

Introduction:
Levitan et al., Cross-Cultural Production and Detection of Deception from Speech

Methodology (left to right):
R. Meir and G. Rätsch. An introduction to Boosting and Leveraging
http://www.iis.ee.ic.ac.uk/icvl/iccv09_tutorial_files/random_forest_new2.png

Methodology (cont.):
http://www.extremetech.com/wp-content/uploads/2015/07/NeuralNetwork.png
http://cs231n.github.io/neural-networks-1/
Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks
http://sebastianruder.com/content/images/2016/01/contours_evaluation_optimizers.gif

Discussion (left to right):
https://media.licdn.com/mpr/mpr/AAEAAQAAAAAAAAamAAAAJDA5NWZlMWFlLTQzZjEtNDVmOS1hMWIyLTNiOGU2YTc3NTY3Nw.png
http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/img/Mikolov-AnalogyTable.png