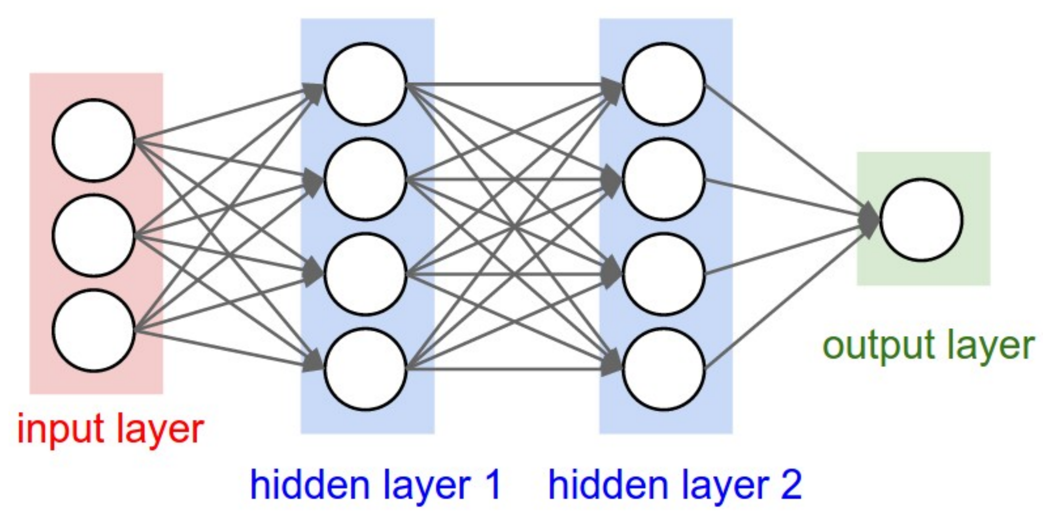
Applied Generative Models for Creating Humorous Photo Captions

Tyler Dohrn,[1] Dragomir Radev PhD[2]

[1]Yale College and [2]Yale University

Introduction

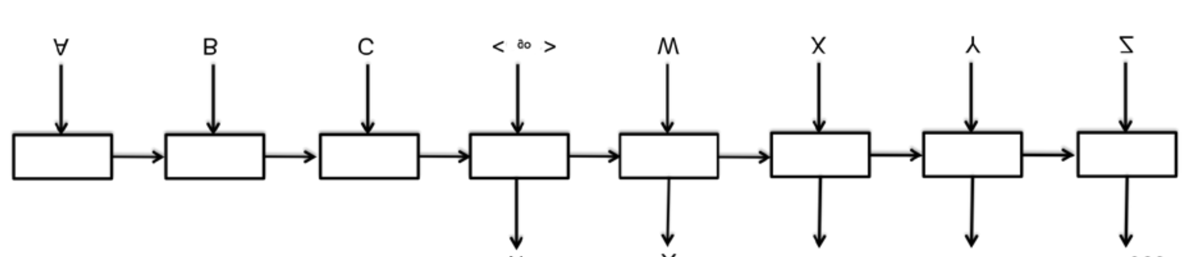
We present a method for combining generative and discriminative models to produce humorous content with a low rate of error. In this work, we use Sequence to Sequence for our generative models, and we use Support Vector Machines, Naive Bayes Estimation, Deep Neural Networks, and Bootstrapped Deep Neural Ensembles for our discriminative error-correcting models.



Materials and Methods

We trained our models on a data-mined corpus of 4,000 labeled sentences (2,000 humorous, 2,000 non-humorous). The 4,000-sentence corpus was split equally into training data and testing/validation data in each trial.

Sequence to Sequence (Seq2Seq)



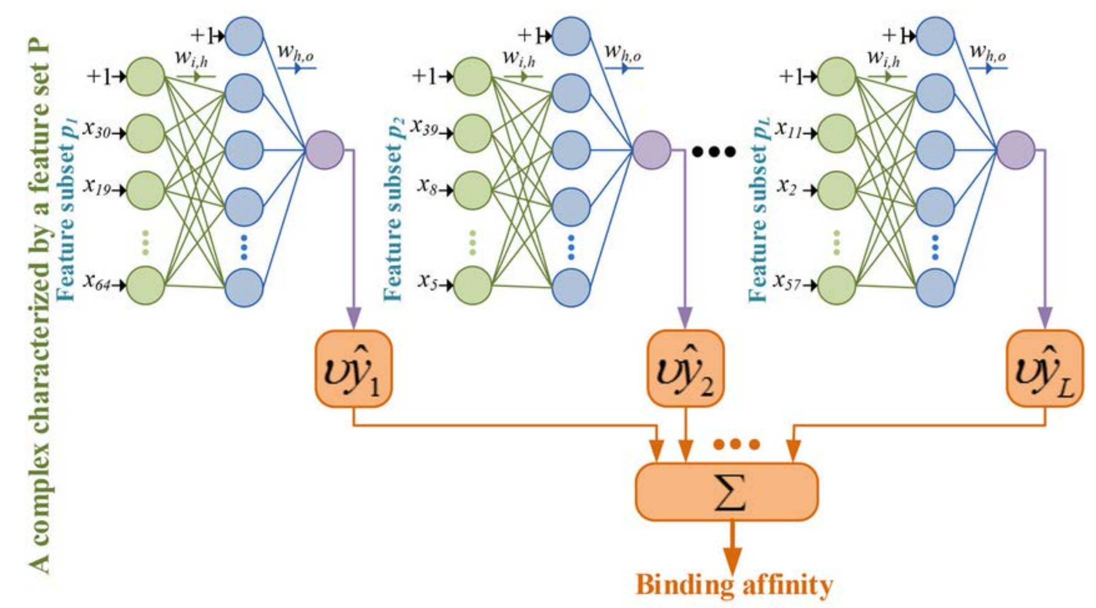
5-tuples (A,B,C,D,E) were generated from the humorous subset of the training corpus. Seq2Seq transitions (A,B,C,D) -> (A,B,C,D,E) and (A,B,C,D) -> E were tested. The intuition was that if this model trained successfully, an unseen 4-tuple (A,B,C,D) would generate novel humorous content. Unfortunately, our experimentats produced negative results for using Seq2Seq in this manner.

Discriminative Model Results

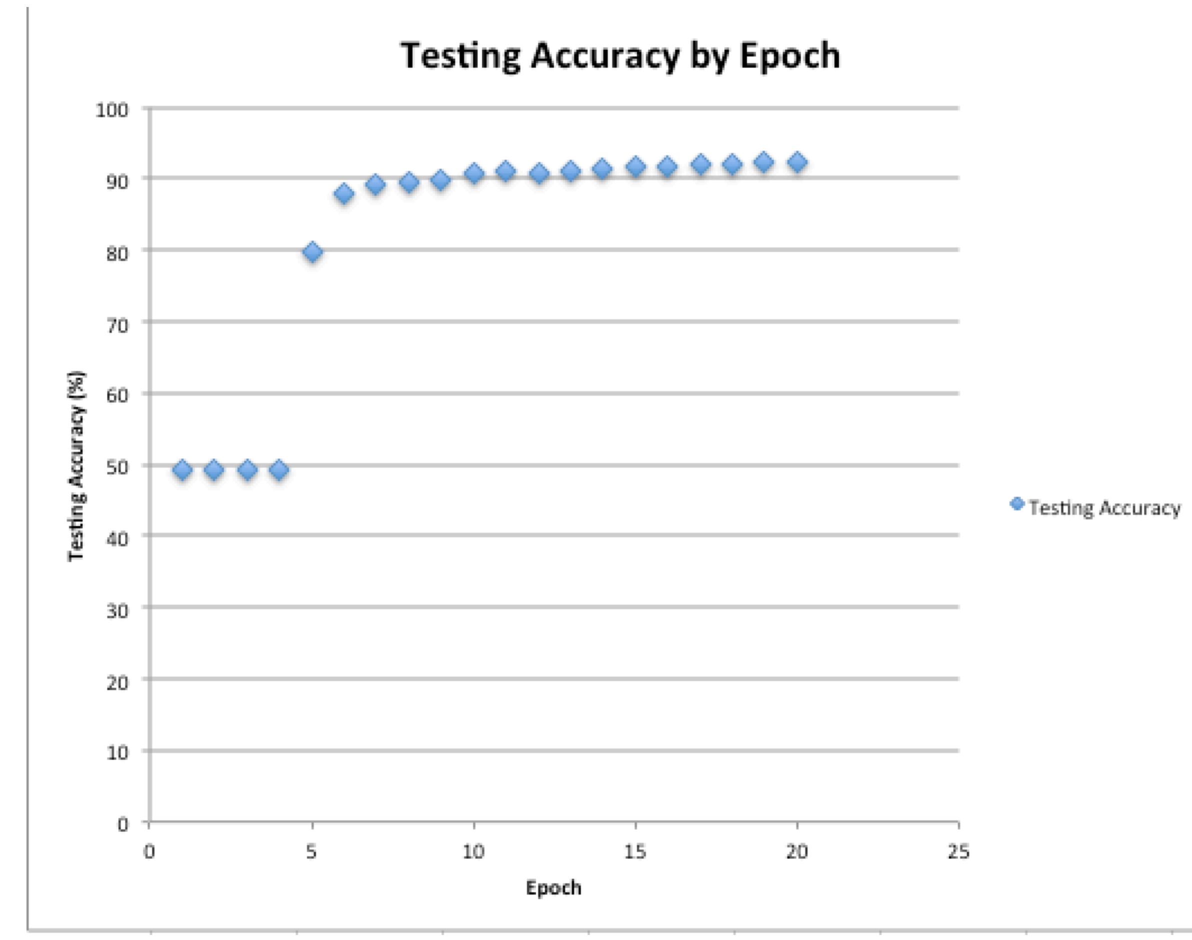
Model	Accuracy
SVM	55.1%
NBE	54.4%
DNN	**93.5%**

5-tuples (A,B,C,D,E) were generated from the humorous subset of the training corpus. Seq2Seq transitions (A,B,C,D) -> (A,B,C,D,E) and (A,B,C,D) -> E were tested. The intuition was that if this model trained successfully, an unseen 4-tuple (A,B,C,D) would generate novel humorous content. Unfortunately, our experiments produced negative results for using Seq2Seq in this manner.

Bootstrapped Neural Ensembles

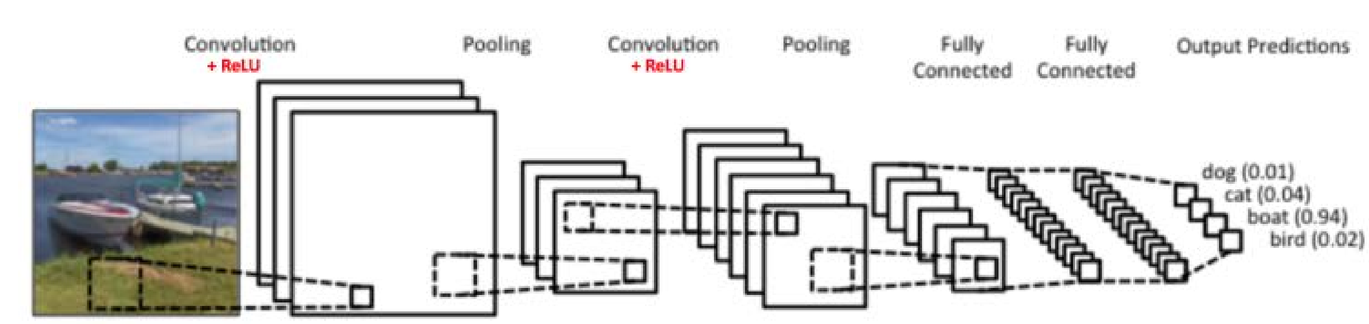


Train neural ensembles by bootstrapping equivalent-cardinality datasets by sampling with replacement, and training one neural network per bootstrap dataset. The output of the ensemble is treated as a statistical distribution, and the classification decision is a function of this statistical distribution (typically mean, standard deviation)



Neural Network Architecture

I began by experimenting with complicated models (deep, wide convolutional nets beginning with a learned word embedding, then two convolutional layers, max pooling, fully connected layers with dropout. This more complicated approach led to difficulty training and poor generalization out-of-sample.



I replaced this architecture with a much simpler one to great success. The final neural network architecture takes a Google Word2Vec 300-dimensional real-valued word representation as input. The Word2Vec representation of a sentence is given by the coordinate-wise average over the 300 coordinate values of constituent words in the sentence.

$$R(s_i) = \frac{1}{L(s_i)} \sum_{j=1}^{L(s_i)} E(w_{ij})$$

This sentence representation is fed into two fully connected 128-node hidden layers and then to two output nodes. Increasing the width of the hidden layers did not aid performance. Increasing the depth of the network created problems in training. I believe that given time to adjust the network architecture (implementing dropout, adding more noise to training gradient, etc.), a deeper network could be trained successfully. However, in this application, it does not appear necessary.

Results

Sentence	DNN Tag
All pro athletes are bilingual. They speak English and profanity.	Humor
The four most beautiful words in our common language: I told you so.	Humor
Force is mass times acceleration	Non-humor
When I lost my rifle, the Army charged me; that's why in the Navy, the captain goes down with the ship	Humor

Conclusion

This semester was a fantastic introduction to Natural Language Processing and applied research. I hope to continue with the LILY Lab, and cannot wait to see what we will all accomplish together. I am optimistic about continuing my research with combined generative and discriminative models for the purpose of generating humor. In particular, I am excited about experimenting further with the discriminative model, which really seemed to perform quite well.

Acknowledgements

I would like to thank Professor Dragomir Radev for his time and expertise. I would also like to thank all of the members of the LILY Lab for their time and support. I would like to specifically thank Rachel Prince for sharing her humor datasets and supervising the effort to manually label New Yorker caption photos. I would also like to especially thank Daniel Keller for his help and expertise.