Using Deep Neural Inspector to Evaluate Predictive Embeddings in Gang-Affiliated Tweets

Alyssa Hwang

ahh2143@columbia.edu

Background

- Deep Neural Inspector (DNI): evaluates how much a machine learning model has learned by comparing hypothesis functions to the output of neuron/layer
- Use a model that classifies gang-related Tweets as aggression, loss, or other
- Represent each Tweet as a vector of values from the Dictionary of Affect in Language (scale of 1-3)
  - Pleasantness
  - Activation
  - Imagery

Research Questions

- How can we make the DNI work on the NLP model?
- What does the NLP model learn in relation to the pleasantness, activation, or imagery of each word in a Tweet?

Methodology

- Collect dataset: Tweets posted by affiliates of a Chicago gang
- Define four models: random/trained aggression and loss classifiers
- Write feature functions to produce vectors of DAL values for each input text
- Run DNI on first (unigrams) and second (bigrams) layers separately, measuring F1 scores with absolute correlation and logistic regression

Results

Table 1: Top F1 Scores, First Convolutional Layer (Unigrams)

Model	Metric	Node (hypothesis)		Random	Trained
Aggression	Absolute Correlation	21	(imagery)	0.0168	0.329
		10	(imagery)	0.0370	0.328
		124	(activation)	0.164	0.311
	Logistic Regression	(Imagery)		0.645	0.843
		(Pleasant)		0.587	0.836
		(Activation)		0.543	0.832
Loss	Absolute Correlation	155	(imagery)	0.0818	0.385
		96	(imagery)	0.0250	0.321
		66	(imagery)	0.102	0.312
	Logistic Regression	(Imagery)		0.645	0.853
		(Pleasant)		0.561	0.826
		(Activation)		0.541	0.827

Table 2: Top F1 Scores, Second Convolutional Layer (Bigrams)

Model	Metric	Node (hypothesis)		Random	Trained
Aggression	Absolute Correlation	28	(activation)	0.0397	0.190
		190	(imagery)	0.0748	0.185
		124	(activation)	0.0505	0.174
	Logistic Regression	(Imagery)		0.396	0.457
		(Pleasant)		0.398	0.444
		(Activation)		0.355	0.413
Loss	Absolute Correlation	37	(imagery)	0.0150	0.200
		23	(imagery)	0.0193	0.193
		2	(imagery)	0.0772	0.165
	Logistic Regression	(Imagery)		0.367	0.456
		(Pleasant)		0.370	0.450
		(Activation)		0.309	0.391

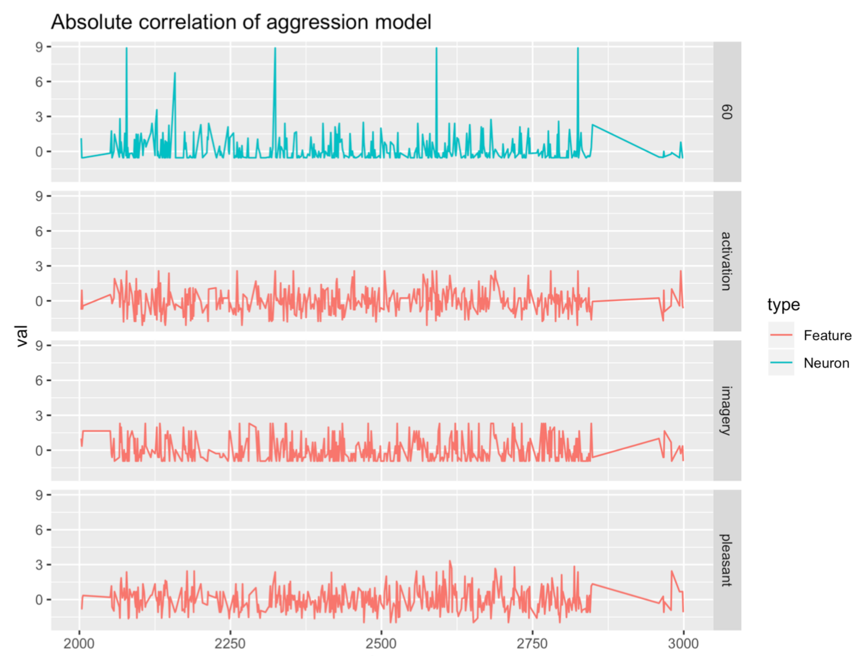

Figure 1: Absolute correlation of aggression model for neuron 60 of the first convolutional layer (unigrams)

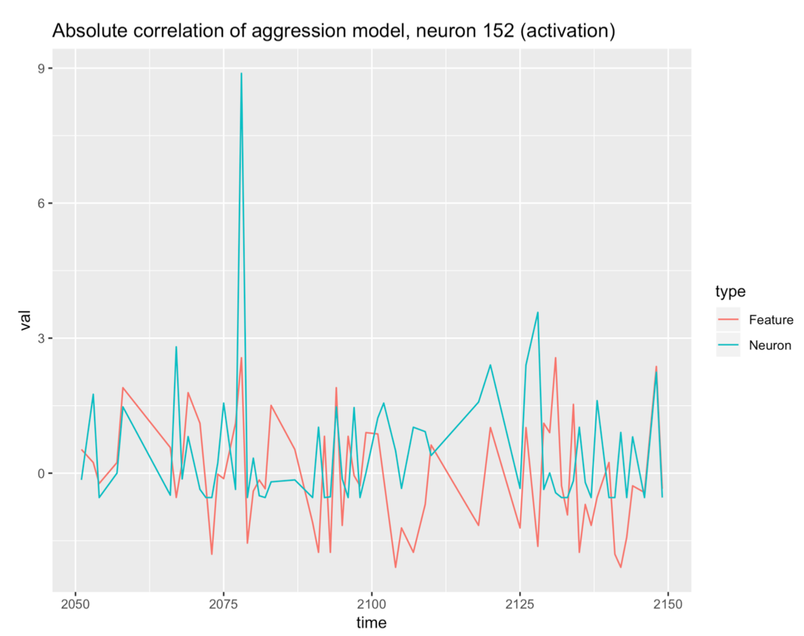

Figure 2: Absolute correlation of aggression model, neuron 152 (activation hypothesis)

Discussion

- F1 scores were higher for trained models, as expected, confirming that DNI does work
- Model does learn, as indicated by the higher absolute correlation to the hypothesis functions
- Larger logistic regression score → layer works better as a whole than individual neurons
- The loss models showed lower scores for both layers, suggesting that their performance is lower or that it is easier to classify aggression
- The second convolutional layer showed much lower scores than the first convolutional layer, suggesting that unigrams are more advantageous for learning than bigrams

Further Studies

- Learning more about the NLP model
  - Higher performance for low pleasantness and higher activation?
  - Particular keywords or sentiment?
- Lexical Inquiry Word Count: word families
- Evaluating past ML models and building more efficient ones in the future

Literature

Detecting Gang-Involved Escalation on Social Media Using Context. For publication.

Predictive Embeddings for Hate Speech Detection on Twitter. For publication.

Langman, Peter. School Shooters: Understanding High School, College, and Adult Perpetrators. N.p.:Rowman & Littlefield, 2015. Print.

Oatley G., Crick T. (2014) Changing Faces: Identifying Complex Behavioural Profiles. In: TryfonasT., Askoxylakis I. (eds) Human Aspects of Information Security, Privacy, and Trust. HAS 2014.Lecture Notes in Computer Science, vol 8533. Springer, Cham

Oksanen, Atte, Pekka Rasanen, and James Hawdon. "Hate Groups: From Offline to Online SocialIdentifications." The Causes and Consequences of Group Violence: From Bullies to Terrorists.Lanham, MD: Lexington, 2014. N. pag. Print.

Sumner, C., Byers, A., Boochever, R., Park, G.J.: Predicting Dark Triad Personality Traits fromTwitter Usage and a Linguistic Analysis of Tweets. In: Proceedings of the 11th InternationalConference on Machine Learning and Applications (ICMLA 2012). IEEE Press (2012)

Acknowledgments

I would like to thank Kathy McKeown, Eugene Wu, Thibault Sellam, Ruiqi Zhong, and Michelle Yang for their generous support and feedback.

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science