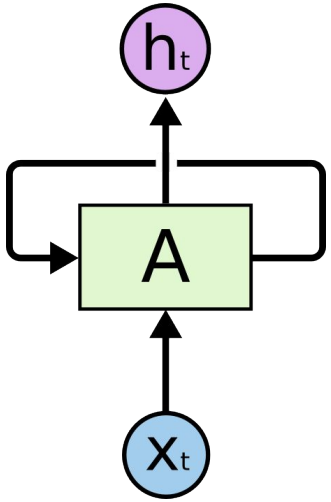
Hate Speech Detection with Neural Networks

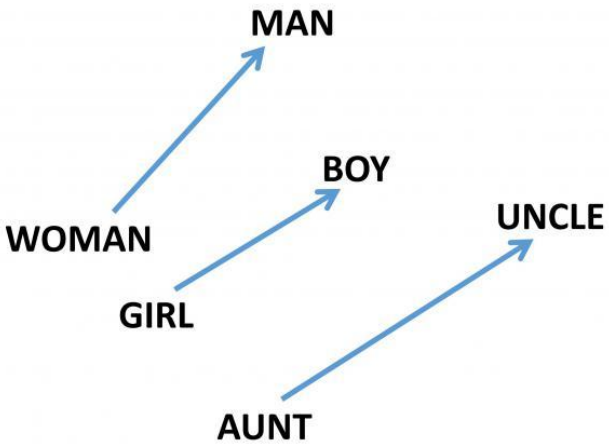
Adyasha Maharana, Abhinav Gupta

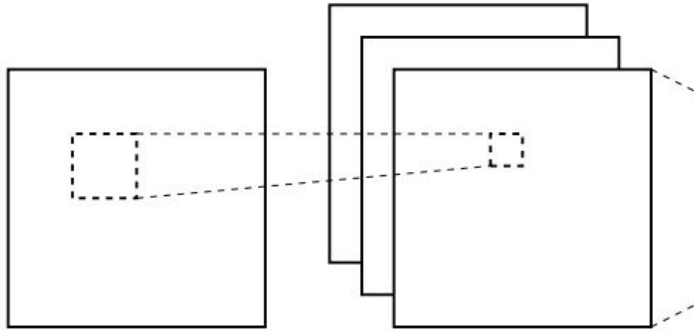
PyCon 2017


Cyber
Bullying

Deep Learning has revolutionized the way we think about Natural Language Processing







Distributed Semantics: to aid language understanding
Recurrent Neural Networks: for language generation
Convolutional Neural Networks: for classification tasks

Online Abusive Speech Dataset

Curated and distributed by Emily Spahn @eyspahn

May 2015 comments from 24 sub-reddits

# Non-hate comments	106509
# Gender-hate comments	5628
# Size-hate comments	39461
# Race-hate comments	6139
# Religion-hate comments	673

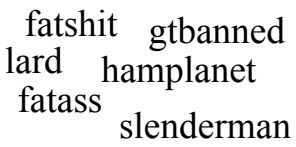
kaggle

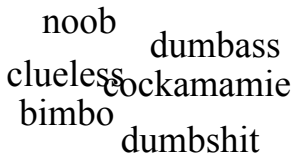
Data preprocessing techniques for hate speech

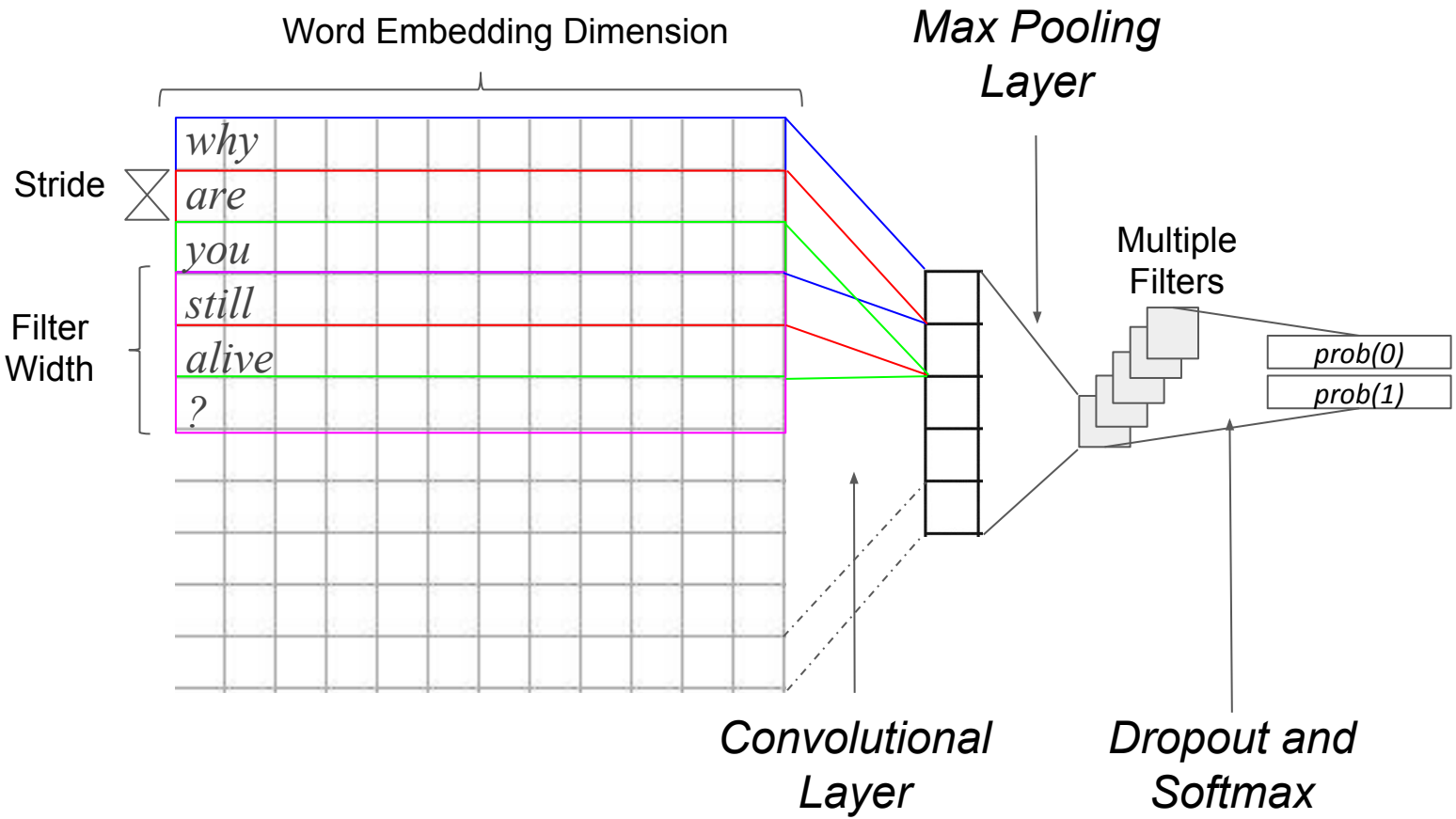
Use of Minimum Edit Distance to unmask derogatory words
@$$h@le → asshole

Referring dictionary to separate agglutinated terms
uglyduckling → ugly duckling

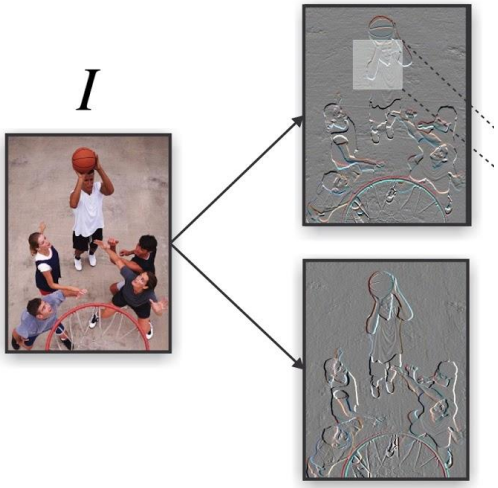
Word embeddings trained on hate speech





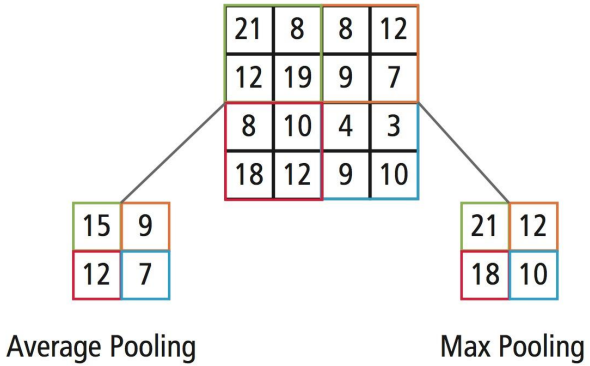


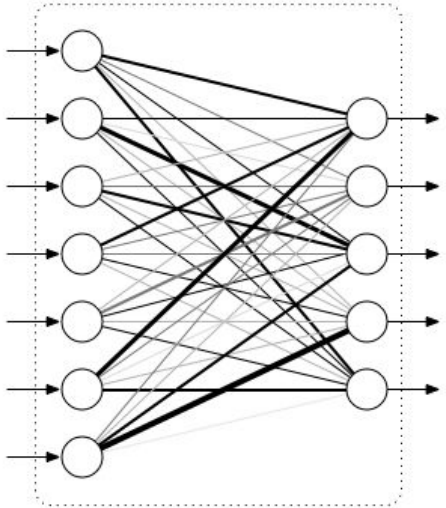
Convolutional Neural Network for Text Classification



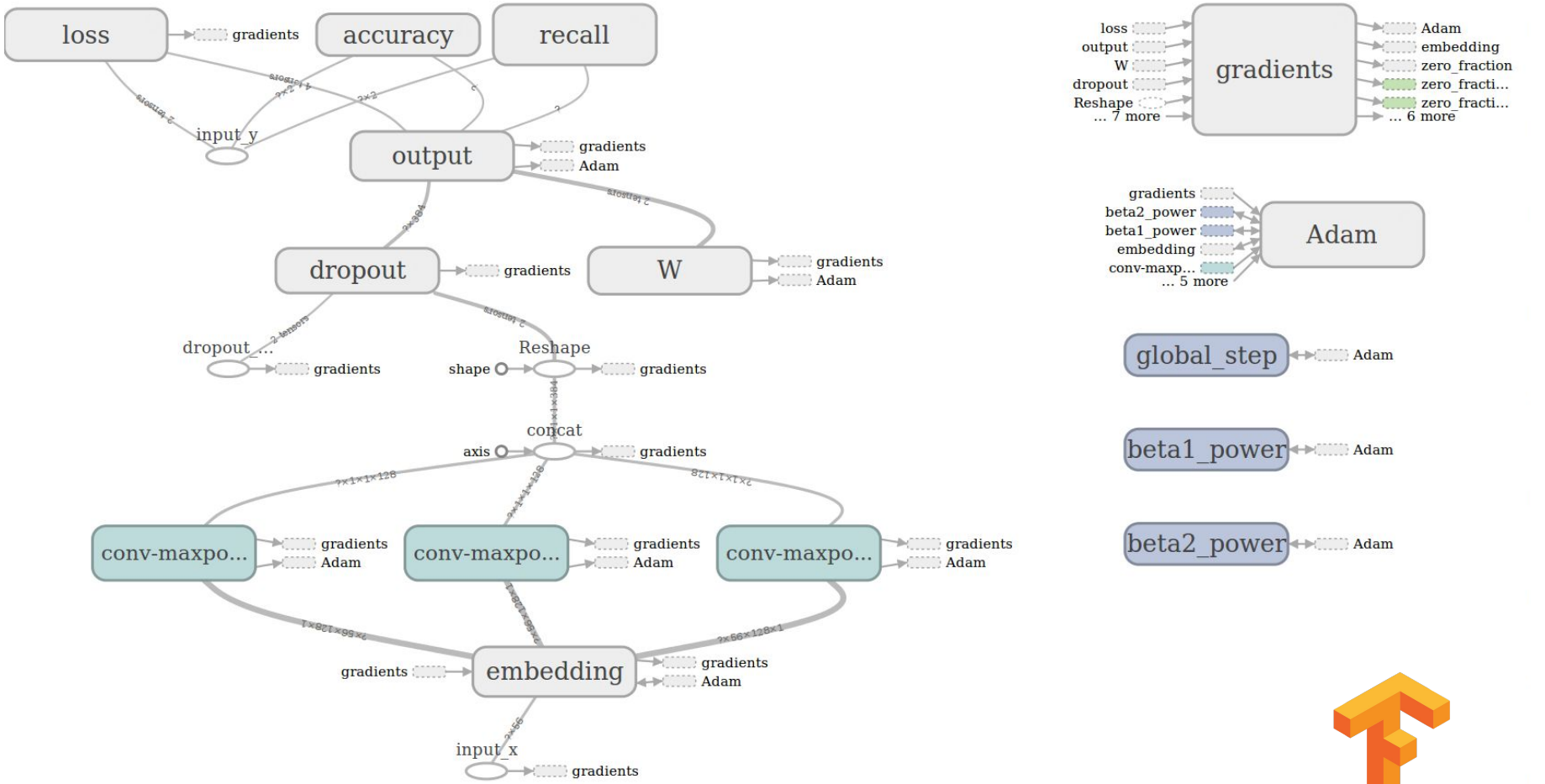
Convolutional Layers are filters that extract position invariant features from the input. Deeper the layer, more complex the feature

Pooling Layers control size of representation, prevent overfitting and reduce number of parameters for learning



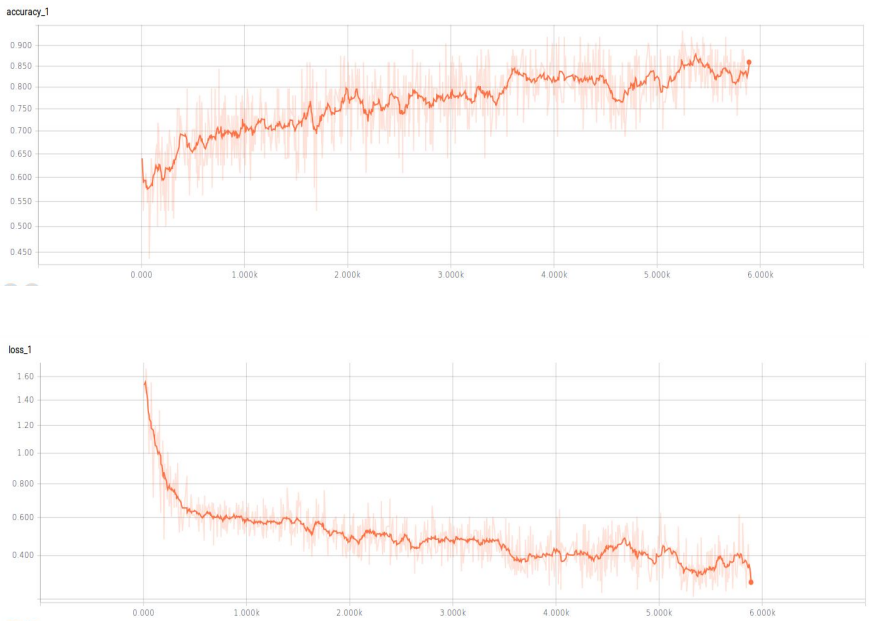


Fully Connected Layers are the last layer in a CNN. Every neuron in previous layer is connected to all neuron in current layer



TensorFlow Graph for the Neural Network

TensorFlow



Real-time visualization of Accuracy and Loss in TensorBoard

	Non Hate	Gender Hate	Race Hate	Size Hate	Religion Hate
Non Hate	92508	6043	2145	5611	1003
Gender Hate	845	3932	339	430	95
Race Hate	106	912	4519	564	54
Size Hate	2111	1402	328	35011	190
Religion Hate	56	98	126	45	343

Confusion Matrix of Results

This model has problems...

Upper cap on sentence length: Sometimes, insults are spread across more than one sentences. They have to be processed together
Solution: paragraph2vec (paragraph as one vector)

Has to be updated with evolving online jargon

False Positives are harmful to the ecosystem of free speech

Cannot handle sarcasm: Artificial Intelligence is yet to be able to comprehend the nuances of human conversation
Solution: NLP Researchers toiling away at their desks day and night

Nevertheless..
The vast amount of online content makes it impossible for manual moderation. Artificial Intelligence has to be step in at some point and help moderate online visual and textual content - with human supervision.

Make Internet Great Again!