Predicting Virality on Reddit with Bidirectional LSTMs

Kristy Duong, Henry Lin

{kristy5, henryln1}@stanford.edu

Department of Computer Science, Stanford University

Overview

- Reddit is a popular site where users post submissions and users can vote and comment on them
- Goal: to predict submission score based on the title of the thread using RNNs
- Explored loss functions, regression vs. classification
- Used embedding layer and bidirectional LSTM with fully connected layer
- Classification model was better suited for this task than regression due to the large number of very high scored and very low scored submissions

Dataset

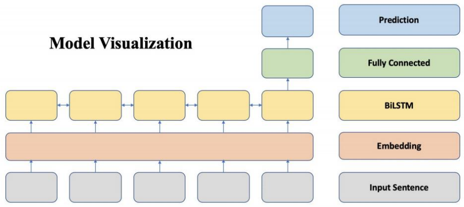
- ~235,000 data points split 70/15/15 from September 2018 r/AskReddit
- Each data point is (title, score)
- Viral posts (score over 1000) are rare → data augmentation performed to increase model exposure to viral posts
- GloVe word embeddings to transform words into RNN input

Features

- GloVe word embeddings of Reddit submission titles
- Input padded according to longest Reddit submission title
- Uses GloVe to convert words into meaningful vectorized representation for the model
- LSTM used for processing time series data which is in this case sequential words in title

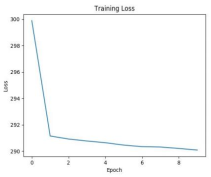
Regression Model

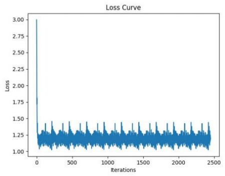
- Regression model using MSE Loss / Huber loss
- Heavily biased by presence of outliers, did not perform well
- Strict margins may have impacted performance negatively
  - Predictions rounded to nearest whole number


Model Visualization

Classification Model

- 20-class classification model using Cross Entropy loss
- Performed much better than regression by grouping scores into buckets
- Buckets broken down to favor lower scores attempting to reflect distribution in dataset


Training Loss


Loss Curve

Results/Analysis

Train	Dev	Test
40.5%	31.0%	30.5%

- Presence of contradictory data points made it difficult for model to learn
- Other factors such as number of comments or time of submission post may be better indicators of virality
- Most scores fell into two buckets so binary classification may be a better model

Title	Score
Have you ever called 911? If yes why?	1
Why did you call 911?	1279

Future Work

In our future work, we would like to explore the use of attention in our model to provide insights about what particular words are weighted highly and may be good indicators of virality. We would also like to explore an ensemble model with both classification and regression components or the incorporation of metadata in the prediction.

References

[1] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In Proceedings of the 23rd international conference on World wide web, pages 925–936. ACM, 2014. [2] Maria Glenski and Tim Weninger. Predicting user-interactions on reddit. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pages 609–612. ACM, 2017. [3] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5-6):602–610, 2005. [4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. [5] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In Seventh International AAAI Conference on Weblogs and Social Media, 2013. [6] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014. [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.