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

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

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

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

Results/Analysis

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>40.5%</td>
<td>31.0%</td>
<td>30.5%</td>
</tr>
</tbody>
</table>

- 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

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.

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

References