Providing Rational For Deep Learning Predictions
CS224N Project Proposal

Kevin Shaw, Henry Neeb, Aly Kane

February 8, 2017

Mentor

Ignacio Cases

Problem Description

Sentiment analysis via deep learning has the potential to revolutionize customer service across all industries. However, adoption of this powerful technology will be accelerated or slowed by humans who own the customer relationships. A system that provides both a quantitative measure (rating) and qualitative example (brief description) of a review would bridge this gap and increase trust.

Such a method is described in *Rationalizing Neural Predictions*. The paper prescribes a two-part model which predicts a multi-sentiment analysis (called encoder) and extracts summary phrases (called generator). We wish to re- implement this process in tensorflow, attempting to match the authors' results.

If time permits, we wish to explore:

- Applicability of algorithm on new datasets
- Combinations of other encoders and generators to improve upon results
  - Specifically, generator methods that allow for a more general independence assumption (see related work for more details)

Data

We will use the BeerAdvocate review data set also used in the paper. This data represents a multi-sentiment analysis problem with user provided labeled sentiment. There are approximately 1.5 million reviews, with each review on average containing 144.9 words [2]. Each review contains a numeric rating target for various aspects of the beer being reviewed (e.g. aroma, head, and taste).

Methodology/Algorithm

The methodology is specified in detail in the paper *Rationalizing Neural Predictions*. In general, the model is proficient at predicting multiple sentiments (encoder) and generating a short summary of the sentiment using words from the text (generator). This is accomplished by:

- Encoding Step: Abstract an encoding process on each review x to minimize L_2 loss between $enc(x)$ and true rating y, using a $RCNN$
- Generator Step:
  - Define z, a l dimensional binary vector that "picks" words from x to represent a summary. We wish to estimate $P(z|x)$.
  - We wish to estimate $P(z|x)$ with $\prod_{i=1}^{n} p(z_i|x, z_{1,i-1})$, using a RCNN
- Jointly optimize encoding and generating by enforcing that the encoding of the generated summary must also predict as well as the original text
- Add regularization parameters for z to enforce small z (number of words) and distance between words in z

Related Work

Our primary source is the paper *Rationalizing Neural Predictions*, which describes the general algorithm for extracting rationals from predictions.

One possible improvement to their process is allowing for a more general independence assumption in their rational generator model. The current generator assumes that words picked are conditionally independent of all future words picked given the review and the prior picked words. A possible improvement is specified in the paper "Generating Sentences from Continuous Space", which recommends a methodology for generating sentences rather than just words using a RNN-based variational autoencoder [1].

Evaluation Plan

The primary evaluation will be a precision metric that will be based on sentence-level annotations and whether selected words are in the sentences describing the target aspect (e.g. appearance, smell, palate.). If time permits, we wish to try to improve their model on the same accuracy metric.

References

[1] Oriol Vinyals Andrew M. Dai Rafal Jozefowicz Samuel R. Bowman, Luke Vilnis and Samy Bengio. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*, 2016.

[2] Regina Barzilay Tao Lei and Tommi Jaakkola. Rationalizing neural predictions. *EMNLP*, 2016.