Generating Reviews and Discovering Sentiment

Radford et. al.

5.14.17

Summary by Justin Chen

Summary:

- Trained a supervised character-level recurrent language model on Amazon reviews and used this as an unsupervised feature extractor for sentiment analysis and for generating sentences with a positive or negative sentiment.
- Tasks:
    - Semantic relatedness, text classification, paraphrase detection
- Datasets:
    - Stanford Sentiment Treebank, Hutter Prize Wikipedia, Amazon Product Review
- Hypothesis for why weaker performance of purely unsupervised approaches
    1. Skip-thought vectors train on corpus of books and distribution does not match distribution of reviews of consumer goods evaluation data
    2. Limited capacity of models results in underfitting
        a. Models are **lossy** - apt at capturing overall ideas, but not at more fine-grained semantic or syntactic details
- Considered byte-level (character-level) for generality
- Used a fuck ton of data - Amazon product review dataset
    - 82 M reviews = 38 B bytes
    - Data split into 1000 shards
    - Training: 38 M*998 shards = 37.924 B bytes
    - Validation: 38 M
    - Testing: 38 M
- Architecture:
    - Single layer multiplicative LSTM 4096 units wide
    - Converges faster than vanilla LSTM
    - Logistic regression classifier on top of learned features
- Training:
    - 1 epoch
    - mini-batch size: 128 with sequences of length 256
    - States set to 0 when training on each shard
    - Adam with learning rate of 5e-4 and linear decay
    - Weight normalization
    - 1 month to completely train with 4 Titan X
    - L1 for text classification and L2 for all other tasks
- Feature extraction:
    - Newline replaced with single space to avoid model resetting state
    - Leading spaces replaced with newline+space start token

    - Trailing whitespaces replaced with a space to simulate end token
    - UTF-8 encoding
    - State initialized to zeros
    - Final cell states of mLSTM used as features and transformed with tanh
- Finds that a single neuron controls binary sentiment
    - Suggests that all information about sentiment in the model is compact represented by a single scalar.
- Ceiling capacity:
    - Diminishing returns as training data increases (tried four orders of magnitude more data) observed in test accuracy - only a bit more than 1% increase
    - Drop in accuracy when testing on other types of text like document datasets instead of sentence level data
- Discussion and Future Work:
    - Dataset is unbalanced with more positive than negative reviews
    - Subsets of dimensions may correspond to different tasks
    - Hierarchical/ multi-timescale architectures could help with document-level data
    - Could train on more diverse data to address disparity in generalization from short reviews to long documents

Comments:

- The results were completely disappointing. They weren't even state-of-the-art
- We know that each neuron learns something and so a single neuron learning sentiment is not surprising Karpathy showed this in *Visualizing and Understanding Recurrent Networks* in 2015.
- Their conclusion that domain of training data dictates domain model can generalize across was pretty obvious from the start
    - They mention in section 5 that a model trained on fantasy and romance novels will not be able to generalize to sentiment of reviews. This was obvious. The distributions are completely different. Consider the motivation for writing a novel versus writing a review. Reviews are typically short and so people try to convey more meaning per unit (per character, you could think about it that way), where as in novels, the meaning can be conveyed over multiple sentences, so the meaning per unit is less dense, so the underlying distributions are completely different.
    - Additionally, people (just an observation from being human) typically only write a review if they're extremely pleased with the product or service or if they're strongly compelled because they're unhappy. Again this affects the density of meaning per unit. Not surprising.
    - Novels can also contain a myriad of sentiment. Overall, novels could intentionally convey contradicting dynamics, whereas reviews the intention in much more direct, hence the meaning per unit.

Questions:

- What do they mean by “...trained as a language model and not as a supervised feature extractor.” on page 4?
- How is it learning unsupervised features?
- Sentiment neuron (section 4.2)
    - What causes this phenomenon to arise?
    - Is it the overall configuration of the topology and weights, or the immense amount of data, or both?
    - If this was retrained and randomly initialized, would the sentiment neuron consistently reappear albeit in a different location?