CS229 Project Proposal

Haiyuan Mei, Ranjani Iyer

TOTAL POINTS

5 / 5

QUESTION 1

1 Project Proposal 5 / 5

✓ **+ 5 pts** **All looks good**

+ 0 pts Did not submit

+ 4 pts Project not properly scoped

+ 4 pts Not enough detail

+ 3 pts Insufficient work -- please come see project TA

+ 5 pts Project does not make use of any technique taught in class. Full marks are being given but please consider using more CS229 techniques.

Your mentor: Jaspreet

Good work!

Hybrid Distributional and Definitional Word Vectors

Team members: Haiyuan Mei, Ranjani Venkatesh Iyer

Mentor: Andrey Kurenkov

Motivation

Word vectors are typically computed by implementing distributional statistics (such as co-occurrence), but it is surprising that the most logical source of words' meanings - dictionaries - are not leveraged in the process. We want to investigate the ability to integrate word definitions with distributional statistics in the process of creating word vectors.

This project is a continuation of Andrey Kurenkov and Tony Duan's previous work, Def2Vec, in which the authors quantitatively and qualitatively demonstrated that leveraging definitions alone can be used to embed words into a semantically meaningful space comparable to GloVe embeddings; they also demonstrated the utility of Def2Vec in improving the performance of a Neural Machine Translation model when the pre-trained vectors vocabulary is limited and there are several out-of-vocabulary words.

However, since the definitions were just being used to recreate GloVe vectors, the model was of limited value since GloVe is already computed for a large portion of the english vocabulary. The prior team realized that an intriguing direction to explore is creating an entirely new form of word vector that does not rely on distributional statistics and is instead only based on definitions - a 'definitional' word vector'. Intuitively, this should be doable by autoencoding the definitions, since a word's definition encodes its meaning and so the latent vector of the encoded definition should be a valid word vector. It may be that these vectors contain information that that distributional vectors do not capture, which motivates the introduction of a combined distributional and definitional word vectors - Hybrid Distributional and Definitional Word Vectors. Including

both types of representation can capture complementary aspects of a given word's meaning, so the combined vector may outperform either one alone.

Method

As a continuation of a previous project, our plan is to extend the work that has already been done on creating 'definitional' word vectors. We will first iterate and improve an Seq2Seq autoencoder implementation that can act as a baseline method, and then attempt to implement a variational autoencoder.[1] Lastly, we will evaluate various possible ways of combining the two types of vectors. If the combined vectors do well on intrinsic metrics, we will also attempt to evaluate them on extrinsic metrics via downstream tasks such as NMT.

Experiments

We plan to apply an intrinsic evaluation and an extrinsic evaluation for HybridVec. For intrinsic evaluation, we will start with similarity[2] and relatedness[3] benchmarks. In the extrinsic evaluation, we can use supervised machine translation tasks, and test with Bleu metrics.

---

[1] **Generating Sentences from a Continuous Space** Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, Samy Bengio

[2] SimLex999 [Hill et al., 2016] and SimLex333

[3] RG [Rubenstein and Goodenough, 1965], WS353 [Finkelstein et al., 2001], SCWS Huang et al. [2012] and MTurk Radinsky et al., [2011] Halawi et al. [2012].