Analyzing the Properties of Bag-of-n-gram via Sentence Reconstruction

Qi Huang, Zhanghao Chen, Zijie Lu, Ye Yuan

March 6, 2018

Problem Definition and Motivation

Bag-of-n-gram representation has been found to excel in many NLP tasks, especially in sentiment classification. A simple multi-layer perceptron model with bag-of-n-gram as input described in (Cho, 2017) outperforms more sophisticated models based on neural networks. This result suggests that bag-of-n-gram representation may contain most information of a sentence's content. The main goal of this project is to investigate the properties of bag-of-n-grams representation, specifically whether it contains enough information to reconstruct the original sentence.

Tentative Approach

We break down the main goal into two subproblems:
1) With respect to text genre and language, find how large n should be for the reconstruction to be close to optimal;
2) Analyze the correlation between the performance of sentence reconstruction and sentiment classification, a representative downstream NLP task.

The reconstruction problem can be viewed as a translation from a sentence to itself. Hence, we propose to approach this problem in 2 ways: 1) apply the RNN Encoder-Decoder model described in (Cho, et al. 2014), where we replace the vector representation generated by the Encoder with the bag-of-n-gram representation as the input for the decoder; 2) replace hidden vectors produced in the encoder network in (Dzmitry, et al. 2014) with vector representation of n-grams, and tune the decoder network to align its attention to n-grams representations directly. We will use the metric described in (Li, et al. 2015) to evaluate the reconstruction performance. We will further attempt different n to investigate the optimal choice of n for the reconstruction with respect to 2 text genres (Twitter and Newspaper, with short and relatively long dependenc respectively) and 3 different languages (English, French and Dutch).

For the second subproblem, we plan to compare the performance of sentence reconstruction and sentiment analysis based on the "Build It, Break It" dataset. Our hypothesis is that 1) for every n, the performance of both tasks are positively correlated and 2) the increase in size of n leads to improvement in performance of both tasks. This hypothesis awaits confirmation by empirical results.

Timeline

$3/15 \sim 3/31$	Experiment sentence reconstruction on toy dataset.
$4/1 \sim 4/15$	Experiment on larger dataset and improve model. Analyze the optimal choice of n with respect to different languages and genres.
$4/16 \sim 4/30$	Analyze correlation between performance of sentence reconstruction and sentiment classification. Explore mapping bag-of-n-gram to bag-of-target-n-gram for machine translation if time allows.
$5/1 \sim 5/11$	Summarize the result into a paper.

Relative Experience

All of us have previously worked on deep learning projects. One of the team member audited most of DS-GA 1011 in Fall semester. All of us have gone through online deep learning courses, such as Stanford CS231n and Deep Learning on Coursera.

References

[1] Cho, Kyunghyun. Strawman: an Ensemble of Deep Bag-of-Ngrams for Sentiment Analysis. In arXiv preprint arXiv:1707.08939, 2017.

[2] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In arxiv preprint arXiv:1406.1078, 2014.

[3] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In arXiv preprint arXiv:1409.0473, 2014.

[4] Li, Jiwei, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In arXiv preprint arXiv:1506.01057, 2015.