Amazon Complainers

CNNs + Word Embeddings to Predict Star Rating from Review Text

Colin Cassady, Brady Fowler, Tom Molinari 2017.03.02

INTRODUCTION

The goal of this project is to train a convolutional neural network to classify untagged raw text reviews into Amazon.com's 5 star rating scale. In DSI-6016 we explored the application of ANN structures to learn complicated target concepts from massive amounts of semi-structured data. As exhibited in the ALVINN system, artificial neural networks (and their extended versions, convolutional and recurrent neural networks) are capable of learning deep representations and modelling complex systems.

We are inspired by a recent lecture from the Symanto Group Director of Innovation, Yassine Benajiba who discussed the use of ConvNets to manipulate text data. In his talk, Yassine referenced several papers by Zhang and Wallace, LeCun, and Kim which each propose structures that take raw text as input and perform classification tasks. We intend to replicate the underlying structure (joining a word embedding model and CNN) to predict star rating from review text using the Amazon Product Data dataset. Our models will be assembled and tested in the newly released PyTorch library and will require GPU acceleration to train model weights (which will be tuned with batch gradient descent and a binary cross entropy loss objective function).

Data Collection

The Amazon Product Data is made available by UC San Diego and consists of "142.8 million reviews spanning May 1996 - July 2014". A review consists of the following data: Reviewer ID, reviewer name, the textual content of the review, the star rating awarded by the review, the 'summary' (which is placed above a review like a title), and the time at which it was posted. Please see the following link for more detailed documentation: http://jmcauley.ucsd.edu/data/amazon/.

References

[1] "Text Understanding from Scratch." Xiang Zhang, Yann LeCun, arXiv:1502.01710 [cs.LG] https://arxiv.org/abs/1502.01710

[2] "Distributed Representations of Words and Phrases and their Compositionality." Tomas Mikolov, et. al. "https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[3] "Convolutional Neural Networks for Sentence Classification" Y. Kim, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Oct. 2014. http://emnlp2014.org/papers/pdf/EMNLP2014181.pdf

[4] "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." Ye Zhang and Byron C. Wallace. CoRR. 2015. https://arxiv.org/pdf/1510.03820v2.pdf.