MSAN 631 - Deep Learning

1. Team

Harrison Mamin

2. Project Title

Predicting Author Demographics from Anonymous Online Posts

3. Background and Motivation

Over the course of my practicum at Ultimate Software, a company in the HR Analytics domain, I've seen that some incredible insights can be extracted from text data. It is possible, for example, to classify emotions, identify subjective vs. objective statements, or detect toxic language. Combined with the quantity of easily accessible text data online, natural language processing has become an important component of many companies' data science efforts. However, one angle that is not always considered is the source of this data. Take the classic example of sentiment classification in movie reviews: rather than simply finding whether a review is positive or negative, it might be beneficial to know if the author of the review was a 14 year old boy or a 45 year old woman. It would be interesting and useful to determine how results differed across demographics.

4. Project Objectives

I believe this is a task that could be useful in many different situations and industries. Text data is widely available online, but much of it is anonymous. By identifying author demographics, it could help companies better understand how different users are responding to or interacting with their product. This could help shape marketing campaigns or decisions regarding product strategy. More generally, this provides context to other analyses or models. For example, rather than just learning that users responded negatively to a new feature, we might find that 20-somethings responded negatively while teenagers liked it.

5. Data

The raw dataset consists of nearly 700,00 blog posts from roughly 20,000 different authors. It was collected from blogger.com in August 2004. Data is stored in XML files, which I downloaded from the Blog Authorship Corpus website, hosted by Bar-Ilan University. I plan to start by experimenting with a smaller subset of the data, and if the model is underfitting I can increase the size. If necessary, I've also found a Twitter dataset which contains gender (but not age), which could augment the current dataset. That being said, the blog corpus is large enough that I don't foresee that being necessary.

6. Techniques Overview

I'm tentatively planning to start with a simple bag of words model as a baseline, followed by a bidirectional LSTM, a common architecture for sequence modeling. Time allowing, I am hoping to implement some other models as well as a means of comparison. In particular, I am curious to evaluate how CNN's perform on an NLP problem as this is something I have not tried yet.

Since the dataset contains labels for both age and gender, I'm also interested in comparing two different methodologies: an end-to-end approach where one model aims to place users in a single demographic bucket, vs. a two-model approach where one model aims to predict age and the other predicts gender.

7. Optional Outcomes

Some other areas I'm interested to explore (that may or may not be implemented during this class) are building an attention-based model or comparing deep learning based approaches to a simpler linear model after training new fastText embeddings. Eventually, I'd also like to build up some more useful tools to help with future PyTorch projects (for example, creating live graphs that update to show loss, other performance metrics, or even diagnostics like visualizing gradients or weights during training).

8. Evaluation

This dataset is actually relatively balanced, so accuracy is a reasonable metric here for gender classification. Mean squared error can measure success at predicting age. For the potential case where I frame this as an end-to-end problem, only a classification metric would be needed. Some further research will be required to find what level of performance is being achieved by others, either on this dataset or similar ones. I'm also curious to see how this model would perform on data from different sources – while the dataset is quite large, it is all from one website and has a limited age range.

9. References

I found the dataset at the following link: http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm (full citation below). As I progress further, I should have a better idea of what papers might be helpful. If I do end up going the route of comparing my model to a linear classifier using fastText embeddings, the 2016 paper *Bag of Tricks for Efficient Text Classification* (https://arxiv.org/pdf/1607.01759.pdf) from Facebook research could prove useful.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker (2006). Effects of Age and Gender on Blogging in *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

10. Schedule

With a 1 person team, delegation of responsibilities will not be an issue. I'm hoping to wrap up the data pre-processing stage within the next day or two. By the end of week 3, I would like to have fit the first model. By the end of week 4, I plan to implement a second architecture for comparison.

Depending on my results, the last week should consist of hyperparameter tuning, running experiments by tweaking training methods or model architectures, and writing the report.