## Movie Review Sentiment Classification Using Pre-trained NLP Models

*Megan Egbert, Arman Karimian, Noushin Mehdipour, Tayler Pauls, Athar Roshandelpoor*

[*{megbert,armandok,noushinm,tayler,athar}@bu.edu*](mailto:athar@bu.edu)

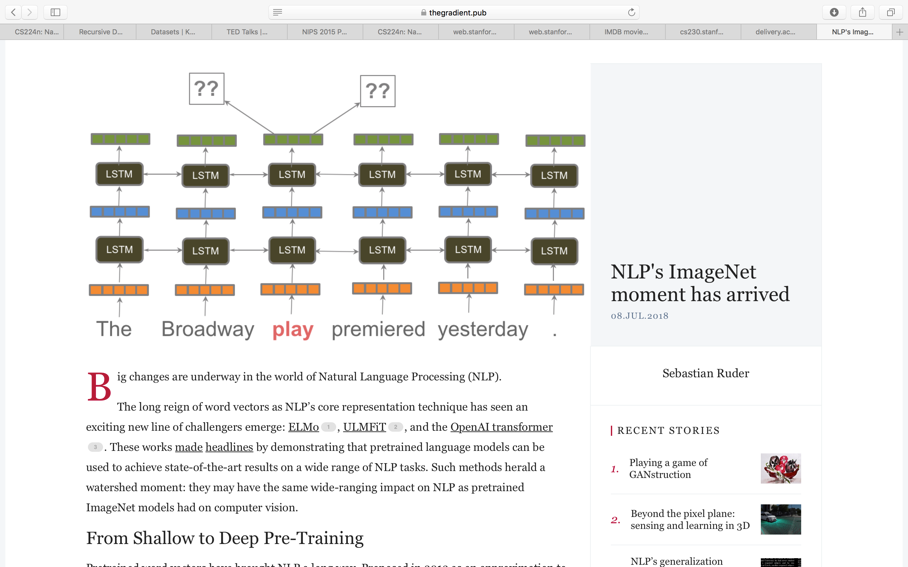


Figure 1.

**1. Task**

Distributed representations pre-trained with models like *Word2Vec* [1] and *GloVe* [2] have become common initializations for the word vectors of deep learning models. These methods are utilized to initialize the first layer of NN, the rest of which is then trained on the data for a particular task. Although deep learning has achieved modern models on many NLP tasks, those models are trained from scratch, needing huge datasets and days to converge. Inductive transfer learning has largely impacted computer vision, but approaches in NLP still needs task-specific considerations and training from scratch.

We study Universal Language Model Fine-tuning (ULMFiT) and Embeddings from Language Models (ELMo), which can be used to achieve Computer Vision (CV) like transfer learning for any task in NLP [3,4].

**2. Related Work**

To address the aforementioned issues in NLP, four general ideas used in CV and NLP are combined. The first enhancement is inspired by transferring the first layers of model similar to ImageNet methods used in CV [5]. Although in recent works, it is suggested to fine tune the last layer(layers) of the pre-trained model without changing the other layers, which is proved to have better performance [6]. Another related work is Hypercolumns, which is pretraining the embeddings that contain additional context via other tasks. These embeddings can be later used as features and concatenated with the word embeddings or inputs at intermediate layers [7]. This method has also been replaced by end-to-end fine-tuning [6]. Multi-task learning (MTL) is also similar to the proposed approach in the sense that it adds a language modeling objective to the model that is trained jointly with the main task model [8]. However, using MTL, tasks must be trained from scratch every time[9]. Previous works on fine-tuning were able to transfer between similar tasks (QA for instance [10]), but failed to successfully transfer between unrelated tasks [11].

Fine-tuning a language model was also studied in [12], but it had overfitted in a 10k labeled data. The ULMFiT approach, however, provides a general-domain pretraining and removes overfitting in a small (100 labeled) datasets.

**3. Approach**

The focus of this project will be the sentiment analysis task (in general classification task) through NLP. Rich features will be extracted from textual information in order to quantify subjective information.

More specifically, we begin with predicting movie review scores. We compare the error rates of our approach with other models such as CoVe[2] and oh-LSTM[3]. Depending on how hard the implementation will be, we may add more interesting datasets. Furthermore, we may experiment with unidirectional RNN models, regular RNN cells or LSTM cells.

**4. Dataset and Metric**

We are using three data sets. First, IMDB dataset. This dataset contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification. The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg). In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing. It is possible that we will change the ratio between test and training data based on performance.

Second, we will use Stanford Sentiment Treebank (SST1 and SST2) which consists of 11,855 single sentences extracted from movie reviews. It was parsed with the Stanford parser and includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges. This new dataset allows us to analyze the intricacies of sentiment and to capture complex linguistic phenomena.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | #Train | #Test | Avg | Max | #Class |
| IMDB | 25000 | 25000 | 265 | 3k | 2 |
| SST1 | 11855 | 2210 | 18 | NA | 5 |
| SST2 | 9613 | 1821 | 19 | NA | 2 |

average/maximum length of documents (#words) of the training/test data.

* <http://ai.stanford.edu/~amaas/data/sentiment/>
* <https://github.com/undertheseanlp/underthesea/wiki/DATA-SST>

We hope to show that the error rate of this model on our dataset is less than former NLP models. [3].

**5. Proposed Timeline and Roles**

Each teammate should be assigned some non-trivial coding task.

|  |  |  |
| --- | --- | --- |
| **Task** | **Deadline** | **Lead** |
| Literature Review | 10/28/18 | all |
| Preprocessing data/look for new datasets | 10/28/18 | Tayler |
| Preprocessing data/look for new datasets | 10/28/18 | Megan |
| Implement ULMFit on data | 10/30/18 | Athar |
| Implement ELMo on data | 10/30/18 | Arman |
| Performance analysis and comparisons | 11/05/18 | Noushin |
| Report and presentation | 12/15/18 | all |

**References**

1. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.
2. Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
3. J. Howard, S. Ruder: Universal Language Model Fine-tuning for Text Classification, arXiv.org, 2018.
4. Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).
5. Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. Cnn features off- the-shelf: an astounding baseline for recognition. InProceedings of the IEEE conference on computer vision and pattern recognition. pages 806–813.
6. Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015a. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pages 3431–3440.
7. Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In Advances in Neural Information Processing System
8. Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In Proceedings of AAAI 2018.
9. Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2017. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks pages 1–10.
10. Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question Answering through Transfer Learning from Large Fine-grained Supervision Data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers).
11. Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications? *Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing*.
12. Andrew M. Dai and Quoc V. Le. 2015. Semi- supervised Sequence Learning. Advances in Neural Information Processing Systems (NIPS’15) http://arxiv.org/abs/1511.01432.