Untrained XLNet with Doc2Vec in Movie Sentiment Classification

Eric Li and Varun Gudibanda
Carnegie Mellon University

Introduction

The IMDB Movie Review Dataset[3] is used as a benchmark for advancements in Natural Language Processing. We propose combining the two following language models to achieve state of the art results in reduced training time:

- XLNet[1]: the current state of the art autoregressive pretraining method. Requires finetuning on the dataset which can be expensive.
- Doc2Vec[2]: Document embedding model which distinguishes word embedding between different documents. Word embeddings are shallow semantic-level.

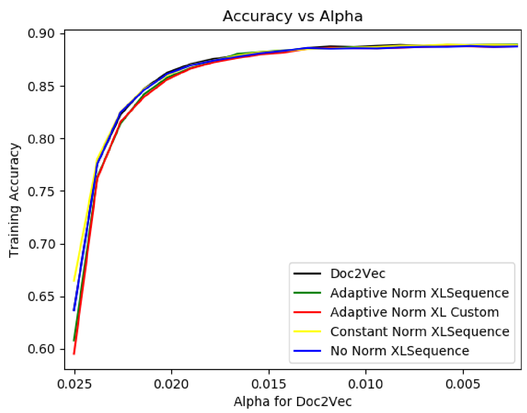
XLNet provides an extremely robust sentiment encoding for individual words while Doc2Vec provides document-level features which XLNet fails to capture. Further, XLNet is expensive to train. As such, we propose using the features of an untrained XLNet in combination with Doc2Vec.

Results

Accuracy was only evaluated on training sets due to time constraints. Doc2Vec was trained in the following manner:

1. Iterate over data.
2. Shuffle the data differently for each iteration.
3. Manually control learning rate and reduce each iteration.

This is in opposition to using a fixed learning-rate which is the default. Training Doc2Vec in this manner tends to produce better results[4].
This led to the following results where alpha is the learning rate.



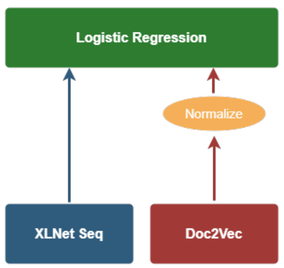
So far there do not appear to be any statistical improvements in accuracy when XLNet features are included.

Model Architecture

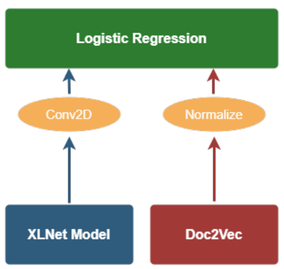
Preprocessing:

- Adding separator and close tokens
- Convert text into tokenized text, then into IDs for XLNet
- Creating attention masks for XLNet
- Cleaning text for Doc2Vec

We propose two architectures that both use a DBoW implementation of Doc2Vec:
The first uses XLNetForSequence Classification. This outputs two values for binary sequence classification. We append these to the Doc2Vec features along with a normalization factor for Doc2Vec.



The second uses a raw XLNet model without any top layer. This returns the hidden states of the final hidden layer. In order to make these useable by the Logistic Regression, we apply a simple convolution with constant weights of a normalization factor determined to make the final features less than 1.



We consider three normalization methods:

1. No normalization: Features are kept as they are from both models and simply concatenated.
2. Constant normalization: Doc2Vec features are multiplied by a constant factor such that they have order similar to that of XLNet features.
3. Adaptive normalization: Doc2Vec features are scaled according to each mini-batch such that the maximum feature has value 1.

Conclusions

When evaluated separately, we derive the following best test accuracy rates:

- Untrained XLNet features: 0.5276
- Base Doc2Vec: 0.8886
- Doc2Vec + XLNet: 0.89

The difference in accuracy is statistically insignificant.

Further Work

- It is possible that the XLNet features fail to assist linear classifiers. Thus, we will consider exploring non-linear methods such as Random Forest classifiers.
- Investigate using XLNetForTokenClassification. This should provide more token-specific results which is more relevant in our problem and better matches the features provided by Doc2Vec.
- Formalize results using training/testing set. It is possible that training error does not accurately represent robustness of features due to to overfitting.

Acknowledgements

We would like to thank our project mentor, Dinc Basar for helping us formalize our proposal. We would also like to thank our instructors, Leila Wehbe and Tom Mitchell for giving us the resources and background to pursue this topic.

References

[1] Zhilin Yang et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Carnegie Mellon University* (2020).

[2] Quoc Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents”. In: *Stanford University* (2014).

[3] Andrew L. Maas et al. “Learning Word Vectors for Sentiment Analysis”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.

[4] Radim Rehurek. “Doc2Vec Tutorial”. In: *RARE Technologies* (2014).