News Classification Model

Xiaobo Zhang, Yuhao Zhang
Stanford CS230 Students
{bobz, njzyh}@stanford.edu
SCPD Video Link: https://youtu.be/6Mb5UWcPYoI

Motivation

News feed applications are growing exponentially in the big data area. Even though there are a lot of news to explore, there are challenges to find news that meet the users' interest. One of the major tasks for those intelligent applications is to recommend personalized feeds to users. The core of a recommendation system is usually a ML-based model that categorizes news feeds based on its content and metadata to make recommendations based on a user's activities. This project proposes a news content classifier that classifies text news content that sets predefined labels. The input to the system is text format news content (a combination of title and content) and the output is one of the predefined labels of the news categories.

Data

The dataset is General News Category Dataset from Kaggle. The Json file contains 202,372 records and a total of 41 distinct categories. They are Json files with pure text attributes.

Features

We extracted headline and short_description and catenate them together to form the input texts. The category attribute is used for labels.

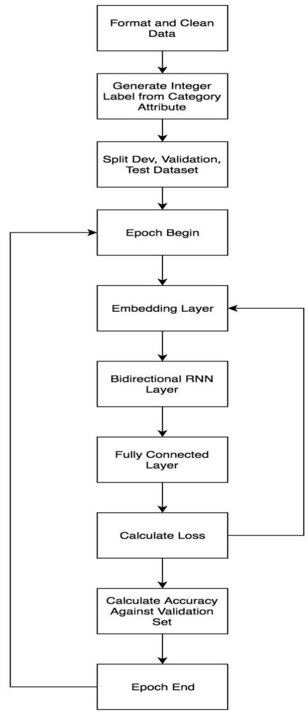
Models

Models we have explored: Bidirectional RNN, Bidirectional RNN with Attention and MT-DNN. We calculate the loss with tensorflow sparse_softmax_cross_entropy_with_logits loss function.

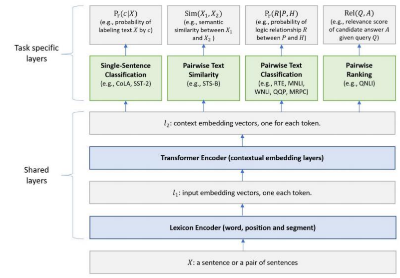
Result

Model	Training Accuracy (180,000 Records, 90%)	Test Accuracy (20,000 Records, 10%)
Bidirectional RNN Accuracy	0.5453	0.5352
Bidirectional RNN with Attention Accuracy	0.6028	0.5987
MT-DNN	0.3760	0.3349

Bidirectional RNN with Attention



MT-DNN [7]



Future

MT-DNN is a novel and effective model for NLP tasks. We did not have enough bandwidth and computing resources to run it with original hyperparameters. We recommend migrating it from Docker to the AWS platform for future experiments.

References

[1] Hingmire, S.; Chougule, S.; Palshikar, G. K.; and Chakraborti, S. 2013. Document classification by topic labeling. In SIGIR, 877–880.
[2] Lewis, D. D. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In SIGIR, 37–50.
[3] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
[4] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329, 2014.
[5] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In ICML, 2011.
[6] Schuster, M., and Paliwal, K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 11 (Nov. 1997), 2673–2681.
[7] Liu, X., He, P., Chen, W., and Gao, J. Multi-task deep neural networks for natural language understanding. CoRR, abs/1901.11504, 2019.
[8] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: A system for large-scale machine learning. Tech. rep., Google Brain, 2016. arXiv preprint.
[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation
[10] R. JeffreyPennington and C. Manning. Glove: Global vectors for word representation. 2014.
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.