Improving Dropout with Attention

Anonymous EurNLP submission

Abstract

Attention Dropout is introduced to improve model accuracy per iteration and model interpretability.

1 Introduction

Dropout (Srivastava et al., 2014) is a simple and efficient regularization technique. However, its specific effect on Natural Language Processing (NLP) tasks has not been thoroughly explored. Larger models in deep learning have been shown to provide better test scores (Simonyan and Zisserman, 2014) but their accuracy per iteration and interpretability raises a significant problem in NLP.

2 Attention Dropout and Results

The probability of dropping a unit is denoted as p_{drop} and the number of words in the sequence is denoted as N. A vector of hidden size is associated with each layer which we call the layer vector. On the forward pass, N dot product operations are computed with the layer vector across the input's word embeddings to obtain attention values α. The attention distribution is inverted and the softmax operation is applied to satisfy the probability axioms as follows: $\alpha'_i = softmax(max(\alpha) - \alpha_i)$. We then sample K indices from α', where $K = max(p_{drop} * N, 1)$. The K word embeddings corresponding to the sampled indices are set to zero.

We compared Attention Dropout to Dropout on 3 binary classification NLP datasets; Large Movie Review Dataset (IMDB) (Maas et al., 2011), Quora Question Pairs (QQP) (Chen et al., 2018), and Question Natural Language Inference (QNLI) (Wang et al., 2018). Attention Dropout outperforms Dropout in test accuracy on all experiments in Table 1. The architecture used is the same as BERT (Devlin et al., 2018) but without position and segmentation embeddings. Both models (AD, D) are the same except for their dropout type.

L	Dataset	Datasize	E	AD	D
6	IMDB	25K	1	84.0	83.5
12	IMDB	25K	1	**85.0**	49.5
6	QQP	30K	1	67.1	63.0
12	QQP	30K	1	**70.1**	63.0
6	QNLI	20K	2	**43.4**	41.8

Table 1: Model test accuracy after E epochs of Attention Dropout (AD) and Dropout (D) trained on a subset (Datasize) of the datasets, using L layers and L attention heads.

3 Discussion and Conclusion

Magnitude based pruning methods (Gomez et al., 2019; Poernomo and Kang, 2018) are extended by incorporating layer cosine similarity. Unlike research in Serrà et al. 2018, layers dynamically attend to embeddings by ignoring others. Entire embedding vectors are dropped to preserve latent space information, in contrast to all other Dropout techniques which drop individual units. Moreover, tests show that the number of units dropped with Attention Dropout is significantly closer to empirical expectation which is likely to improve training, for example, AD achieved a mean difference of -281, while D showed a mean difference of 16096. Furthermore, Attention Dropout improves model interpretability by only managing N attention coefficients compared to the Transformer's (Vaswani et al., 2017) N^2.

Attention Dropout improves accuracy per iteration by maintaining latent information and attending to relevant word embeddings. Moreover, relative to sequence length we improve model interpretability by providing linear explanations, compared to previous research's quadratic.

References

Z. Chen, H. Zhang, X. Zhang, and L. Zhao. 2018. Quora question pairs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aidan N Gomez, Ivan Zhang, Kevin Swersky, Yarin Gal, and Geoffrey E Hinton. 2019. Learning sparse networks using targeted dropout. *arXiv preprint arXiv:1905.13678*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Alvin Poernomo and Dae-Ki Kang. 2018. Biased dropout and crossmap dropout: Learning towards effective dropout regularization in convolutional neural network. *Neural Networks*, 104:60–67.

Joan Serrà, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.