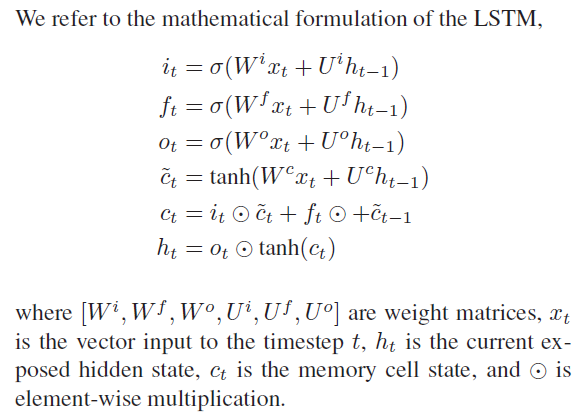
Regularizing and Optimizing LSTM Language Models

1. Introduction

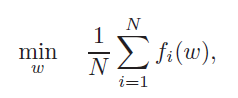
ASGD carries out iterations similar to SGD, but instead of returning the last iterate as the solution, returns an average of the iterates past a certain, tuned, threshold T . This threshold T is typically tuned and has a direct impact on the performance of the method. We propose a variant of ASGD where T is determined on the fly through a non-monotonic criterion and show that it achieves better training outcomes compared to SGD.

2. Weight-dropped LSTM

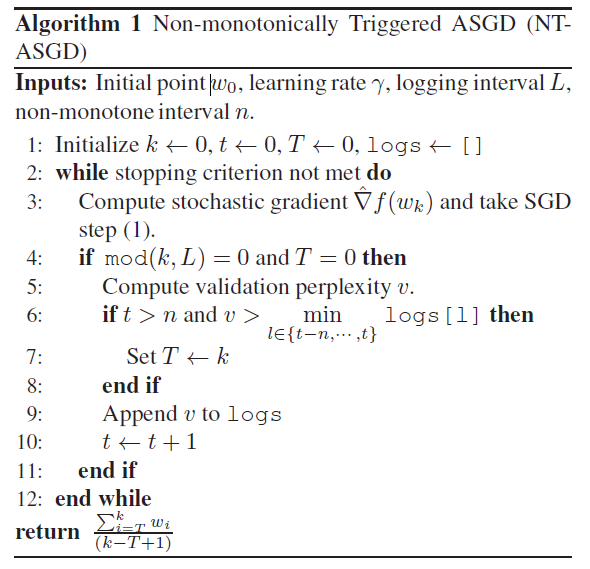


3. Optimization

SGD is among the most popular methods for training deep learning models across various modalities including computer vision, natural language processing, and reinforcement learning.

Motivated by this observation, we investigate averaged SGD (ASGD) to further improve the training process.



Ideally, averaging needs to be triggered when the SGD iterates converge to a steady-state distribution.

4. Extended regularization techniques

4.1. Variable length backpropagation sequences

To prevent such inefficient data usage, we randomly select the sequence length for the forward and backward pass in two steps.

4.2. Variational dropout

4.3. Embedding dropout

4.4. Weight tying

4.5. Independent embedding size and hidden size

4.6. Activation Regularization (AR) and Temporal Activation Regularization (TAR)

5. Experiment Details

6. Experimental Analysis

7. Pointer models

In past work, pointer based attention models have been shown to be highly effective in improving language modeling.

8. Model Ablation Analysis

9. Conclusion

In this work, we discuss regularization and optimization strategies for neural language models. While the regularization and optimization strategies proposed are demonstrated on the task of language modeling, we anticipate that they would be generally applicable across other sequence learning tasks.