Lab 09: Vector Semantics
Part – II Language Model in Recurrent Neural Network with Regularizing & Optimizing LSTM

Adnan Irshad, University of Trento
adnan.irshad@unitn.it

1. Introduction

The objective of Exercise-2 was to further optimize the best model obtained from Exercise-1 by incorporating additional regularization techniques based on the paper titled "Regularizing and Optimizing LSTM Language Models" by Merity et al. The focus was on exploring three regularization methods: weight tying, variational dropout, and non-monotonically triggered Average-SGD (AvSGD) and evaluating their impact on language modeling performance.

2. Baseline

In Exercise-2, I built upon the best model obtained from Exercise-1, which incorporated modifications such as replacing RNN with LSTM, adding dropout layers, and using the AdamW optimizer. After changing RNN with LSTM layer and applying dropout layers on embedding and output the best model were recorded with PPL 171.15 in Exercise-1.

3. Implemented Improvements

To augment the baseline model, I adopted the following techniques described in the research paper:

1. Weight Tying: Weight tying is a technique that constrains the model's parameters by sharing the weights between the input and output embeddings. By doing so, the model benefits from regularization and dimensionality reduction. I applied weight tying by ensuring that the input and output embeddings have the same weight matrices. This regularization technique aimed to enhance the model's generalization capabilities.

2. Variational Dropout: Variational dropout is a dropout variant that injects noise during training and sampling. It helps the model to better handle uncertainty and improves robustness. I applied variational dropout by implementing a Variational Dropout class & replacing the traditional dropout layers in the model with variational dropout layers. This regularization technique aimed to enhance the model's ability to capture and model different sources of variation.

3. Non-monotonically Triggered AvSGD: Non-monotonically triggered AvSGD is an optimization technique that combines average SGD with adaptive learning rate scheduling. It helps the model achieve better convergence and avoids overfitting. I applied non-monotonically by implementing the algorithm (Non-monotonically Triggered AvSGD (NT-AvSGD)) mentioned in paper which triggered AvSGD by incorporating the AvSGD optimizer into the training process and adjusting the learning rate based on performance on a held-out validation set. This optimization technique aimed to improve the training dynamics and prevent overfitting.

4. Results

After implementing the additional regularizations, I evaluated the performance of the optimized language model by computing the perplexity (PPL) metric on following best Hyperparameters as per different experiments:

- Learning Rate: 1.0
- Clip: 5
- Hidden Size: 200
- Embedding Size: 300

I did multiple experiments with Combined Variational Dropout + NT-ASGD + Weight Tying and found the following as the best output.

- **Exp #1:**
  - Test PPL: 250.85144823554143
- **Exp #2:**
  - Test PPL: 220.98253596525882
- **Exp #3:**
  - Test PPL: 204.16782099553143

The best LSTM model achieved a perplexity of 204.16782099553143, indicating improved performance compared to the baseline.

5. Observations

The incorporation of weight tying, variational dropout, and non-monotonically triggered AvSGD resulted in substantial enhancements to the baseline model. The model achieved lower perplexity values and demonstrated an improved understanding of semantic relationships between words, as evidenced by the analogy task results. These findings underscore the effectiveness of the proposed techniques in augmenting the model's performance and its ability to capture word associations accurately.

6. References:

https://openreview.net/pdf?id=SyyGPP0TZ