Assignment 2: Build and Compare Small Language Models

Learning Goals

- Implement and train small sequence models for language modeling (RNN, LSTM, Transformer).
- Compare models using cross-entropy and perplexity.
- Conduct simple ablation studies and analyze training stability.
- Demonstrate generation ability with different sampling strategies.

Task Overview

Train at least two models for next-token prediction on a small corpus:

1. A recurrent model (RNN or LSTM).
2. Another variant (RNN, LSTM, or a small Transformer).

Compare results both quantitatively and qualitatively.

Datasets

You can use any datasets. Use an 80/10/10 train-validation-test split.

Model and Training Guidelines

For example:

- Embedding size: 128–256, hidden size: 256, layers: 1–2.
- Dropout 0.1–0.3, gradient clipping at 1.0.
- Sequence length: 128 (word/subword) or 256 (character).
- Optimizer: Adam or AdamW, learning rate 10^{-3} to 3×10^{-4}.

Evaluation

- Report validation and test perplexity.
- Generate samples with different temperatures ($T = 0.7, 1.0, 1.3$).
- Plot training and validation loss curves.
- Record training time.

Cross-entropy loss:

$$H = -\frac{1}{N} \sum_{t=1}^{N} \log p_\theta(x_t \mid x_{<t})$$

Perplexity:

$$\mathrm{PPL} = \exp(H)$$

Ablation Studies

Include at least two comparisons, such as:

- Dropout: 0.0 vs. 0.2
- Context length: 128 vs. 256
- Tokenization: word vs. subword

Deliverables

1. **Report** (up to 5 pages): dataset, models, results, analysis, and generation samples.
2. **Code**: notebook or repo with clear instructions.

Deadline: Submit report and code by 23:59 (Singapore time) on Sep 12th.