Impact of Transformer Depth on Small Datasets for Chatbot Models

Anonymous EMNLP submission

Abstract

This study investigates the impact of transformer layer variations (2, 4, 8, and 16 layers) on the performance and efficiency of chatbot models trained on a small curated dataset. From the original DailyDialog dataset containing 76,052 samples, 10,000 instances were selected after preprocessing, including duplicate removal and filtering by token count (3–60). Evaluation metrics include BLEU, ROUGE, METEOR, and qualitative assessments. Results indicate that the 8-layer transformer achieves the best balance between response quality and computational efficiency. These findings highlight optimal configurations for transformer models in low-resource conversational AI.

1 Introduction

Modern NLP models, particularly transformers, have demonstrated exceptional performance across tasks with large-scale datasets. However, challenges arise when applying such architectures to small datasets. This study evaluates the impact of varying transformer depth (2, 4, 8, and 16 layers) on model performance and computational efficiency in a small-data setting.

2 Background and Related Works

TBU

3 Method

3.1 Dataset Preprocessing

The experiments were conducted using the DailyDialog dataset, a widely-used datasets for dialog generation tasks. The original dataset consists of 76,052 multi-turn conversational samples. To dapt the data for this study, the following preprocessing steps were applied:

1. **Duplicate Removal**: All duplicate samples were removed to ensure diversity in the training data and minimize inference of perplexity in model training.

2. **Token-length Filtering**: Dialog samples were filtered to retain only those with token counts between 3 and 60, ensuring meaningful yet manageable input lengths. 95% of total training datasets are included in this range.

3. **Sample Selection**: A subset of 10,000 instances was randomly selected from the preprocessed dataset to simulate a small-scale training scenario.

4. **Tokenization**: Tokenization was performed using the NLTK punkt tokenizer and a vocabulary size of 15,000 tokens was constructed.

3.2 Transformer Model Configuration

The configuration of the model follows the standard transformer design with a slight modification in embedding dimension size and variation in encoder and decoder layer numbers, ranging from 2, 4, 8 to 16.

- **Embedding Dimension (d_{model}):** 256. The weights of the embedding layers in the encoder and decoder are shared, since the source and target languages are the same (English).
- **Number of Attention Heads (n_{heads}):** 8.
- **Feedforward Layer Dimension (d_{ff}):** 1024.
- **Dropout Rate:** 0.1.
- **Positional Encoding:** Maximum sequence length of 128 tokens.

3.3 Training Setup

The models were trained using the Adam optimizer with the following hyperparameters:

- **Epochs per model**: 10

- **Learning Rate Scheduler**: Warm-up steps of 4,000, followed by inverse square root decay.
- **Optimizer Parameters**:

$$\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e - 9$$

- **Batch Size**: 64
- **Loss Function**: Sparse Categorical Crossentropy, with a padding mask applied to ignore <pad> tokens in the input and target sequences.

3.4 Evaluation Metrics

To assess model performance and training efficiency, following evaluations were performed:

1. **Quantitative Metrics for Text Generation** The 3 following evaluation metrics have scores between 0 to 1, and score above 0.2 is suggested for generative models.
    - **BLEU**: Measures n-gram overlap between generated responses and ground truth.
    - **ROUGE**: Evaluates overlap of longer sequences, and ROUGE-L was used for the experiments.
    - **METEOR**: Consider semantic similarity, penalizing over-reliance on exact matches
2. **Computational Efficiency Metrics**:
    - Training Time Per Epoch: Time taken to complete one training epoch
    - Convergence Speed: Number of epochs required to achieve optimal performance

4 Result

5 Conclusion

TBU

6 Acknowledgements

TBU

7 References

This is an appendix.