Hyperparameter Optimization for Joke Generation Model

Aayush patil(22220)

November 18, 2024

Introduction

- `max_length`: Maximum input sequence length.
- `embed_dim`: Embedding dimension size.
- `num_layers`: Number of Transformer layers.
- `num_heads`: Number of attention heads.
- `ff_dim`: Feed-forward network dimension.
- `dropout`: Dropout rate.
- `lr`: Learning rate.
- `batch_size`: Batch size.

Hyperparameter Variations and Results

I conducted experiments across 5 hyperparameter configurations.

Config	`max_len`	`embed_dim`	`num_layers`	`num_heads`	`ff_dim`	`dropout`	`batch_size`	**Val Los**
1	64	128	2	4	256	0.2	16	6.15
2	128	256	4	8	512	0.1	8	5.75
3	128	512	4	8	512	0.1	16	5.77
4	256	256	6	8	768	0.2	8	5.80
5	64	128	2	4	512	0.3	8	6.25

Table 1: Validation loss for various hyperparameter configurations.

Key Findings

1. Best Configuration: The optimal configuration is:

- `max_len`: 128

- `embed_dim`: 256
- `num_layers`: 4
- `num_heads`: 8
- `ff_dim`: 512
- `dropout`: 0.1
- `lr`: 3e-4
- `batch_size`: 8
- `Number of epochs`: 11

This achieved a validation loss of **5.759552019525627**.