Bonus Task

In this bonus task, a transformer-based model was fine-tuned for multi-class toxic text classification using parameter-efficient fine-tuning. Specifically, DistilBERT was adapted to the dataset using Low-Rank Adaptation (LoRA), allowing efficient training while updating only a small subset of model parameters.

The pretrained model **DistilBERT (distilbert-base-uncased)** was used as the base architecture. DistilBERT is a lightweight transformer model that retains most of BERT's language understanding capabilities while being computationally efficient.

To reduce training cost and memory usage, **Low-Rank Adaptation (LoRA)** was applied to the self-attention layers of the model. LoRA freezes the original pretrained parameters and introduces small trainable adapters within the attention mechanism, enabling effective task adaptation with minimal parameter updates.

The input text was constructed by concatenating the *query* and *image description* fields. Toxic category labels were encoded into numerical class identifiers. The dataset was split into training and testing sets using stratified sampling to preserve class distribution.

Text data was tokenized using the DistilBERT tokenizer with padding and truncation to a maximum sequence length of 128 tokens. The tokenized inputs were converted into PyTorch tensors to enable transformer-based training.

The model was fine-tuned for **15 epochs** using the AdamW optimizer with a learning rate of **2e-5**. Due to class imbalance in the dataset, **Macro F1 score** was selected as the primary evaluation metric, as it assigns equal importance to all classes.

LoRA was configured with a low-rank dimension of 8 and applied to the query and value projection layers of the attention mechanism, ensuring parameter-efficient learning.

Epoch	Training Loss	Validation Loss	Macro F1
1	No log	0.143396	0.947809
2	No log	0.134816	0.947809
3	No log	0.143424	0.947809
4	0.112349	0.132940	0.947809
5	0.112349	0.138876	0.947809
6	0.112349	0.131326	0.947809
7	0.110692	0.137220	0.947809
8	0.110692	0.142725	0.947809
9	0.110692	0.137442	0.947809
10	0.105999	0.140899	0.947809
11	0.105999	0.136187	0.947809
12	0.105999	0.137205	0.947809
13	0.105999	0.136865	0.947809
14	0.101420	0.140121	0.947809
15	0.101420	0.139686	0.947809

{'eval_loss': 0.14339616894721985, 'eval_macro_f1': 0.9478090482858236, 'eval_runtime': 2.1433, 'eval_samples_per_second': 279.945, 'eval_steps_per_second': 17.73, 'epoch': 15.0}

Evaluation Results

The final evaluation on the test set produced the following results:

- **Validation Loss:** 0.14
- **Macro F1 Score:** 0.95
- **Evaluation Runtime:** approximately 2.1 seconds

The high macro F1 score demonstrates strong and balanced classification performance across all toxicity categories, including minority classes. These results confirm the effectiveness of LoRA-based fine-tuning in adapting a pretrained DistilBERT model to the toxic text classification task while maintaining computational efficiency and robust generalization.