Fine-Tuning Flan-T5 for Neutralizing Trump Phrases

Your Name

August 22, 2025

1 Dataset Preparation

I utilized strict style parameters for selecting the phrases from the dataset, since it was quite large. I settled on a selection pipeline which favored medium length phrases with meaningful context, without being overwhelmingly long.

Initially, I tried to use the DeepSeek R1 model for constructing the neutral counterparts, but I encountered issues due to the presence of negative phrases against China. Therefore, I opted for ChatGPT-5. I provided it with the JSON file of Trump quotes and requested neutral counterparts for a selected number of Trump phrases (between 1000 to 3000), which were deemed particularly representative of Trump's style.

2 Fine-Tuning Process

For fine-tuning, **Flan-T5-Large** was chosen as the base model. This decision was primarily driven by hardware limitations, as training large models on a Google Colab GPU is constrained in terms of runtime and memory. Using **LoRA (Low-Rank Adaptation)** allowed efficient fine-tuning without updating all model parameters.

During implementation, tokenization and padding issues were encountered because T5 models require a pad token. This was resolved by setting the tokenizer's **pad_token** to the model's **eos_token**. Data preprocessing was done on a local laptop, which required negligible runtime.

2.1 Computational Environment

- **GPU:** 15 GB VRAM (Colab-provided GPU NVIDIA Tesla T4)
- **CPU / RAM:** 12.7 GB system RAM
- **Disk:** 112.6 GB
- **Software Environment:** Python 3.12, PyTorch, Transformers, PEFT
- **Training Time:** 1842 seconds

2.2 Hyperparameters

Parameter	Value
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.05
Target modules	[”q”, ”v”]
Learning rate	2e-4
Epochs	5
Max sequence length	60
Precision	FP16

Table 1: Hyperparameters used for fine-tuning Flan-T5-Large with LoRA.

3 Qualitative Assessment

Below are illustrative examples of model outputs before and after fine-tuning:

- **Original:** Joe Biden is presiding over the most corrupt administration in American history
  Neutralized: Joe Biden is the most corrupt president in the United States.
- **Original:** We cannot let him tell us what to do with our country. He has no clue. He can't put two sentences together.
  Neutralized: He is not able to tell us what to do with our country.
- **Original:** So, I just want to tell you, I'm an innocent man. I did nothing wrong.
  Neutralized: I'm an innocent man. I did nothing wrong.

While the model has certainly learned to avoid utilizing some of the exaggerations and jargon of Donald Trump, the LLM has too few parameters to actually excel at normalizing the text. Improvement could certainly be made by utilizing a bigger model, longer training epochs and a bigger dataset.