Final solution report

Introduction

The goal of the material Detoxification Task is to turn material written in a harmful way into neutrally styled language that has the same meaning. In this study, I use the T5 model, a sequence-to-sequence model, to offer a solution for this job. We may make use of a dataset that is extracted from the ParaNMT corpus that includes phrase pairs along with their corresponding length differences, similarity scores, and toxicity levels. Developing a model that can successfully remove harmful language while maintaining its original meaning is the aim.

Data Analysis

I make the selection to 2,000 sentence pairs from this corpus. Each pair in the dataset has the following columns:

- reference: The original text with a toxic style.
- translation: A paraphrased version of the reference text with a neutral style.

The dataset is split into a training set and a validation set for model training and evaluation.

Model Specification

I use the T5 (Text-to-Text Transfer Transformer) model, specifically the "t5-small" variant, for text detoxification. The T5 model is a versatile sequence-to-sequence model capable of handling a wide range of natural language processing tasks.
The model is trained to convert toxic text into neutral text while maintaining the same meaning. It is fine-tuned for the text detoxification task by providing pairs of toxic reference text and their neutral translations.

Training Process

The training process involves several steps:

- Data Preprocessing: The toxic reference text is prefixed with "detoxify:" to guide the model. Both the reference and translation text are tokenized and padded to a maximum length of 128 tokens.
- Model Initialization: I initialize the T5 model and tokenizer, and specify the device (GPU if available) for training.
- Training Loop: The model is trained for multiple epochs. We use the AdamW optimizer with a learning rate of 1e-4. During training, we iterate through the dataset in batches, calculate the loss, and update the model's parameters using backpropagation.
- Validation: After each epoch, we evaluate the model's performance on the validation set to monitor progress and detect overfitting. We calculate the validation loss.

Evaluation

To evaluate the model's effectiveness in text detoxification, we use the validation loss and paradetox metric (from references/paper_3.pdf) as a primary metric. A lower validation loss indicates better performance in transforming toxic text into neutral text.

By contrasting the model's produced detoxified text with the original poisonous language, I can also assess the model's detoxification quality. I evaluate how much of the original meaning is preserved in the model while eliminating toxicity.

Results

The validation loss is tracked over the course of the model's five epochs of training. Better results in text detoxification are shown by lower validation loss numbers. By giving the input sentence, the final model may be used to detoxify harmful language; it will produce a neutral version of the same sentence while maintaining its meaning.
The model's capacity to convert toxic language into a more neutral style while maintaining the original text's semantics is indicative of its efficacy in text detoxification.