IS 6733 - Fall 2024 Final Project

Numeral-aware Headline Generation

Kyle Bailey

Introduction

The primary goal of this project, based on SemEval-2024 Task 7, is to create a system that generates **numeral-aware headlines** from news articles. This involves:

1. Extracting key numerical information from the text.
2. Performing necessary **numerical reasoning** to ensure that numbers in the generated headlines are factually accurate.
3. Evaluating generated headlines on both **numerical accuracy** and **headline quality** using a mix of automated metrics.

Dataset used:

The **Num-HG dataset** will be used for the project, as it includes examples of numeral-aware headlines and numerical content from news articles. This will guide the model toward understanding how numerals are typically used and presented in headline contexts.

Approach

Evaluated model performance on the 1st fold of the NumHG data set using the following three methods:

- The out-of-box configuration of Meta's Llama 3.1-8B-Instruct model.
- A QLoRA fine-tuned version that took the complete headline as the target variable.
- A version that utilized both the QLoRA fine-tuned model and chain-of-thought prompting.

Data Preprocessing

Because Meta's Llama 3.1-8B-Instruct model was tuned to accept input in the form of user/assistant chat templates our data required massaging to prepare it for fine-tuning and generating headlines.

This included:

- Framing input and target variables as a list of message dictionaries.
- Format list of message dictionaries for tokenizing and labeling according to Llama 3.1-8B-Instruct chat template specs.

Fine Tuning Methodology

Applied supervised fine tuning using QLoRA and unsloth package to update vector weights for 41,943,040 parameters.

Parameter selections include:

- LoRA Rank = 16
- LoRA Alpha = 16
- Target Modules = ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]
- Epoch = 2
- Batch Size = 32 and Gradient Accumulation = 8 (for an effective Batch Size = 256)
- Learning Rate = 2e-5
- Weight Decay = 0.01

Chain-Of-Thought Methodology

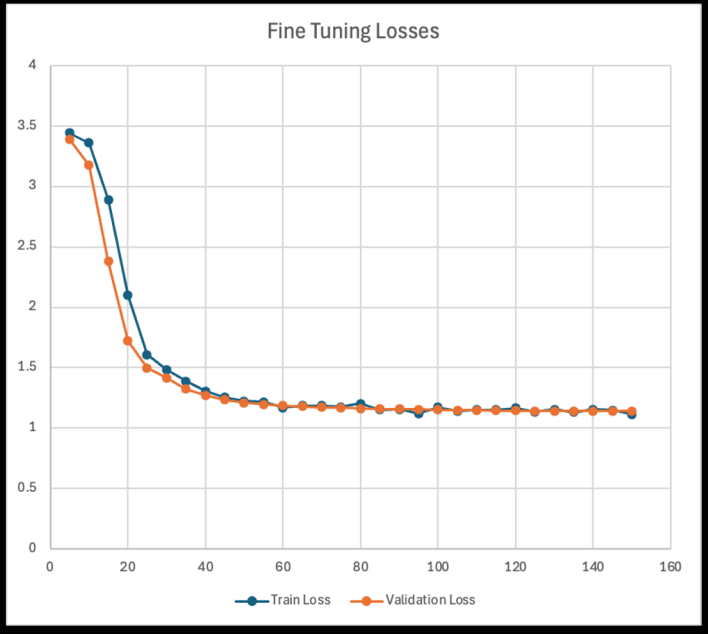
Generate a single headline for this news article that includes at least one key numeric feature:

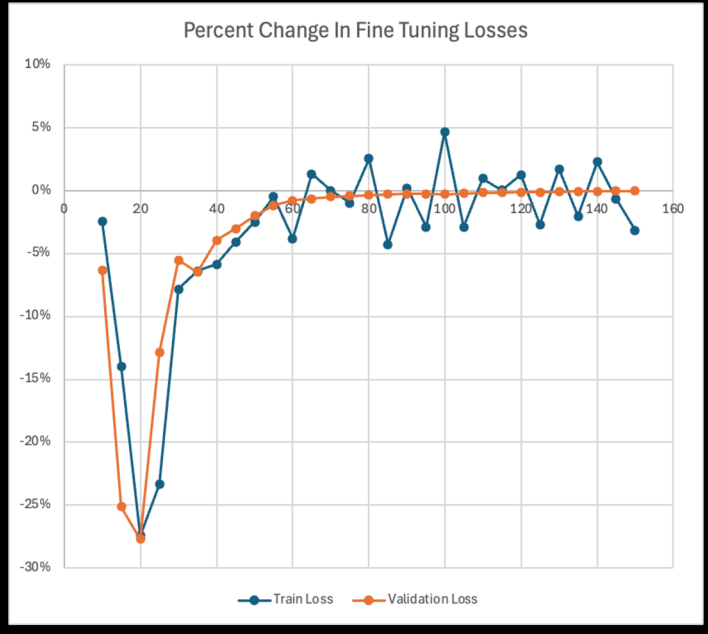
{article body}

When generating the headline consider the following:
1. What is the subject of the article?
2. What is the sentiment of the article?
3. Does the headline accurately portray the subject and sentiment of the article?
4. Is the headline an appropriate length?
5. Is the key numeric feature formatted as a number?

Be sure to return only the generated headline without any enclosing quotation marks.

Fine Tuning Progress





Performance Comparisons

Base Model

Rouge Score
- **rouge1:** 0.320807
- **rouge2:** 0.110107
- **rougeL:** 0.273208

Accuracy
- **All Accuracy:** 0.314291
- **Copy Accuracy:** 0.433953
- **Cal Accuracy:** 0.037059

MoverScore: 0.144228

BERTScore:
- **P:** 0.276058
- **R:** 0.455564
- **F1:** 0.364521

Fine Tuned Model

Rouge Score
- **rouge1:** 0.44197
- **rouge2:** 0.191655
- **rougeL:** 0.433951

Accuracy
- **All Accuracy:** 0.530726
- **Copy Accuracy:** 0.572497
- **Cal Accuracy:** 0.433951

MoverScore: 0.257495

BERTScore:
- **P:** 0.476129
- **R:** 0.454180
- **F1:** 0.465213

FT + Chain-of-Thought Model

Rouge Score
- **rouge1:** 0.435587
- **rouge2:** 0.186737
- **rougeL:** 0.385779

Accuracy
- **All Accuracy:** 0.538115
- **Copy Accuracy:** 0.571465
- **Cal Accuracy:** 0.460849

MoverScore: 0.251104

BERTScore:
- **P:** 0.475043
- **R:** 0.441438
- **F1:** 0.458261

Improvement Opportunity

Now that we have a better foundational understanding of the process and the quirks of this particular task these are possible next steps to refine the model:

- Fine tuning with different available inputs (e.g. calculation or masked headline)
- Attempt multi-stage fine tuning (e.g. fine tune first with number as the target and calculation as the input and then second fine tune with calculation as the target and the article body as the input)
- Iterating through different chain-of-thought prompts programmatically to determine optimal instruction
- Fine tune over all five folds in the NumHG dataset for better performance metrics