

FineTuning of LLMs

Accuracy

The pre-trained Phi model achieves an accuracy of 36% on the SNLI task, which is close to the baseline accuracy of 33% (random guessing among three classes). A deeper analysis reveals that the model is highly biased towards predicting that the premise and hypothesis are unrelated (class label 1), resulting in an inflated accuracy near 33% due to frequent incorrect predictions in that class.

After fine-tuning, the model's accuracy improves to 69%. This improvement is primarily due to the reduction in the model's bias toward outputting class 1. Fine-tuning allows the model to learn a more balanced representation, enabling it to distinguish more accurately between the relatedness classes, which explains the significant boost in accuracy.

Training Arguments

```
training_args = TrainingArguments(  
    output_dir=output_dir,  
    per_device_train_batch_size=1,  
    per_device_eval_batch_size=1,  
    num_train_epochs=5,  
    evaluation_strategy="epoch",  
    save_strategy="epoch",  
    logging_dir='./logs',  
    logging_steps=10,  
    load_best_model_at_end=False,  
    fp16=True,  
    gradient_checkpointing=True,  
    max_grad_norm=1.0  
)
```

QLoRA config

```
model = prepare_model_for_kbit_training(model)  
peft_config = LoraConfig(  
    r=32, #Rank  
    lora_alpha=32,  
    target_modules=[  
        'q_proj',  
        'k_proj',  
        'v_proj',  
        'dense'  
    ],  
    bias="none",  
    lora_dropout=0.05, # Conventional  
    task_type="CAUSAL_LM",  
)
```

Time taken

Total time taken 2.0 hours, 30.0 minutes and 33.17526078224182 seconds

Fine tuning the model on 1001 samples of the SNLI dataset for 5 epochs takes 2 hours 30 minutes and 33 seconds.

Training Parameters

trainable params: 20,971,520 || all params: 2,800,655,360 || trainable%: 0.7488

The trainable parameters in the QLoRa adapter are 20,971,520 which represents only 0.74% of the total trainable parameters in the model(2,800,655,360). This significantly reduces the computational and memory constraint compared to full fine-tuning.

Hardware

Training used up a average load of 99.99981834695731 and a average memory of 4052.87356948 2289 over the 5 epochs

The fine tuning process uses 99.99% of the GPU over the 5 epochs as calculated using the `gpu.load` function from `gpuUtils`.
The average memory load over the 5 epochs is 4052 MB calculated using the `gpu.memoryUsed` utility from `gpuUtils`.

Explanation of the results

We go over 3 examples that were seemingly corrected through fine-tuning and 3 examples that were not corrected.

Example 1

Before Fine Tuning

Given the premise This church choir sings to the masses as they sing joyous songs from the book at a church. and the hypothesis The church has cracks in the ceiling., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

2

2 1

After Fine-Tuning

Given the premise This church choir sings to the masses as they sing joyous songs from the book at a church. and the hypothesis The church has cracks in the ceiling., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2. Answer: 2

The hypothesis entails the premise.

Answer: 2

Explanation: The hypothesis states that the church has cracks in the ceiling. The premise states that this church choir sings to the masses as they sing joyous songs from the

2 1

Explanation:

Here the two statements look quite related since they both talk about a church with a choir. It is entirely possible that the two statements could be related and the fact that the church building has cracks could make it unsafe and contradict the fact that a church choir is singing to the masses there. The premise and hypothesis in this example may require nuanced understanding of language to correctly infer that the two statements are unrelated.

Example 2

Before Fine Tuning

74

Given the premise A tattooed woman clicking on a mouse on a desk. and the hypothesis A tattooed man clicking on a mouse on a desk., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

1

1 2

After Fine-Tuning

74

Given the premise A tattooed woman clicking on a mouse on a desk. and the hypothesis A tattooed man clicking on a mouse on a desk., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

Answer: 1

The hypothesis entails the premise.

Answer: 1

Explanation: The hypothesis states that a tattooed man is clicking on a mouse on a desk. The premise states that a tattooed woman is clicking on a mouse on

1 2

Explanation:

The model might focus on the shared elements ("tattooed," "clicking on a mouse," "desk") and therefore consider the statements to be closely related. Since the overall context is highly similar, it could easily miss the single contradictory detail (gender) that changes the label. Very subtle changes in the gender (man vs woman). The model might need more training examples that emphasize how small word changes (like pronouns or gender nouns) can entirely alter the meaning and turn an entailment into a contradiction.

Example 3

Before Fine Tuning

8

Given the premise A female softball player wearing blue and red crouches in the infield, waiting for the next play. and the hypothesis the player is flying planes, determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

1

1 2

After Fine-Tuning

8

Given the premise A female softball player wearing blue and red crouches in the infield, waiting for the next play. and the hypothesis the player is flying planes, determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2. Answer: 2

The hypothesis entails the premise.

Answer: 2

Explanation: The hypothesis is that the player is flying planes, which is not directly implied by the premise. The premise only states that the player is a female softball

2 2

Explanation:

The premise describes a "female softball player wearing blue and red crouches in the infield, waiting for the next play," which implies the context of a sporting event. The hypothesis, "the player is flying planes," introduces a different, incompatible activity unrelated to the original context.

This discrepancy makes it a straightforward contradiction since the two statements describe fundamentally different activities. Thoroughly studying the test samples would indicate that there are several examples of an individual doing such different activities that are tagged as a contradiction which may improve performance on these samples.

The reason the pre-trained might have perceived these statements as unrelated is because the two activities are so fundamentally different they have no shared context except the reference to 'the' player in the hypothesis.

Example 4

Before Fine Tuning

58

Given the premise A light technician man with tribal tattoos aiming a spotlight over a balcony. and the hypothesis An actress takes the stage., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

1

1 2

After Fine-Tuning

58

Given the premise A light technician man with tribal tattoos aiming a spotlight over a balcony. and the hypothesis An actress takes the stage., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2. Answer: 1

The hypothesis entails the premise.

Answer: 1

Explanation: The hypothesis is about an actress taking the stage, while the premise is about a light technician man. There is no direct implication between the two statements. Output

1 2

Explanation:

The premise and hypothesis describe two different activities (lighting work vs. acting) involving two different people. However, the model might be interpreting the lack of explicit conflict as neutral rather than a contradiction. To correctly label this as a contradiction, the model needs to infer that a light technician and an actress are not likely to be the same individual performing different actions. This requires more nuanced understanding of social roles, which may not have been fully learned during fine-tuning. The two statements are not exactly in direct contradiction to each other. The lack of direct negation or opposition (like "not" or "never") can make it challenging to categorize this as a contradiction without understanding contextual nuance.

Example 5

Before Fine-Tuning

65

Given the premise A group of men playing rugby on the sand. and the hypothesis A group of woman play volleyball in a court., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

1

1 2

After Fine-Tuning

65

Given the premise A group of men playing rugby on the sand. and the hypothesis A group of woman play volleyball in a court., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2. Answer:

2

The hypothesis entails the premise.

Answer: 2

Explanation: The hypothesis states that a group of women are playing volleyball in a court. The premise states that a group of men are playing rugby on the sand. There

2 2

Explanation:

The premise discusses a group of men playing rugby on sand, while the hypothesis mentions a group of women playing volleyball in a court. These two activities are not only different in gender and location but also in the sport itself. This creates a straightforward contradiction. However much like the previous example the premise and hypothesis have very little shared context as nearly all the elements are different- actor, instrument and location. It is also possible that the model might encounter neutral relationships more frequently than explicit contradictions, especially if the contradictions aren't stark. As a result, without fine-tuning, the model might lean towards predicting **1** (neutral) rather than **2** (contradiction).

Example 6

Before Fine-Tuning

44

Given the premise A man is sitting on a blue bench with a blue blanket covering his face. and the hypothesis A man is playing chess while wearing a green tutu., determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2.

INPUT

##OUTPUT

1

1 2

After Fine-Tuning

70

Given the premise A man in a white t-shirt and jeans is holding a mallet and chisel next to his abstract sculpture which stands on several bricks. and the hypothesis A man is wearing a white shirt, determine whether the hypothesis entails the premise. Output 0 in case of entailment, 1 in case the statements are unrelated or not directly implied, 2 in case of a contradiction. Please answer with only the number: 0, 1, or 2. Answer: 0
The hypothesis entails the premise.

Solution:

The hypothesis states that a man is wearing a white shirt. The premise states that a man is holding a mallet and chisel next to his abstract sculpture. There is no
0 0

Explanation:

The premise describes a man sitting on a bench with a blanket covering his face. The hypothesis, on the other hand, describes a man engaged in the action of playing chess while wearing a green tutu. The two actions are very different and do not logically connect. Before fine-tuning, the model might have struggled with such specific contradictions, especially when a hypothesis introduces unrelated or outlandish details (like the green tutu). It may not have been able to identify the contradiction due to the lack of shared context between the two statements.

Note:

It was noticed that the model saved after the 4th epoch seemed to have much better performance than the model checkpoint after the 5th epoch. This could suggest that the model was overfitting. Further the reduction in the training loss while the validation loss shows no significant difference with each epoch is also characteristic of severe overfitting. Moreover, changing the learning rate also showed no improvement.

I have used the Causal LM setup since sequence classification would lead to loss of 0 after one epoch. I found others had also faced similar difficulties:

<https://github.com/huggingface/transformers/issues/26498>

<https://github.com/huggingface/transformers/issues/18730>