Question-Answering using RoBERTa Model

Shriya Sravani Y
UCSC
sy4@ucsc.edu

Abstract

This report investigates the robustness of the RoBERTa model fine-tuned on the SQuAD2.0 dataset for question answering tasks. The study explores the model's behavior across three types of passages: in-domain, edited, and out-of-domain. We analyze its performance in scenarios designed to test its understanding, adaptability, and ability to avoid providing answers when the context lacks sufficient information. Results indicate that the model performs well on in-domain data but can be tricked with edited or out-of-domain contexts, showcasing limitations in robustness and transferability.

1 Introduction

1.1 Question Answering

Question Answering (QA) is a critical task in natural language processing, requiring models to retrieve and generate accurate responses from text passages. RoBERTa, a pre-trained transformer model, has demonstrated significant success in this domain, especially when fine-tuned on datasets like SQuAD2.0. SQuAD2.0 adds complexity to the QA task by including unanswerable questions, forcing models to not only answer correctly but also recognize when an answer does not exist in the provided context.

This report aims to evaluate the robustness of the deepset/roberta-base-squad2 model across varying contexts. The evaluation comprises three parts:

1. Testing in-domain passages found "in the wild."
2. Testing passages edited to mislead the model.
3. Testing out-of-domain passages from diverse sources.

1.2 RoBERTa Model ans SQuAD2.0

RoBERTa: A robustly optimized version of BERT, RoBERTa employs larger mini-batches, longer training, and removes the next-sentence prediction task to enhance performance. Fine-tuned on SQuAD2.0, it becomes adept at handling unanswerable questions alongside traditional QA tasks.
SQuAD2.0 Dataset: The Stanford Question Answering Dataset v2.0 expands the original dataset with over 50,000 unanswerable questions. The model is trained to identify these and avoid giving incorrect answers, making it ideal for robustness evaluation.

2 Part - I

The first part deals with the robustness of RoBERTa model for question answering. This model has been trained on SquAD 2.0 dataset, which contains questions about passages from English Wikipedia.

The goal is to assess the model's ability to correctly answer questions, identify when the answer is not present, and understand how difficult it is to mislead the model.

2.1 In-Domain Passages

Passage 1: Eiffel Tower

Question	Answer	Confidence
Who designed the Eiffel Tower?	Gustave Eiffel	0.92
When was the Eiffel Tower constructed?	between 1887 and 1889	0.58
Where is the Eiffel Tower located?	Paris, France	0.91
What is the Eiffel Tower made of?	wrought-iron	0.53

Passage 2: Albert Einstein

Question	Answer	Confidence
Who developed the theory of relativity?	Albert Einstein	0.95
When did Einstein win the Nobel Prize?	1921	0.97
What is the photo-electric effect?	Albert Einstein	0.12

Observations

- RoBERTa successfully answered straightforward questions with high confidence.
- For ambiguous or unanswerable questions, such as "What is the photoelectric effect?", the model returned irrelevant answers with low confidence (0.1286).
- The model struggled with unanswerable questions, returning incorrect answers.
- For explicit factual questions, the model performed well, demonstrating robustness.

2.2 Edited Passages

Passage 1: Altered Eiffel Tower

Question	Answer	Confidence
Who designed the Eiffel Tower?	Karl Benz	0.98
Where is the Eiffel Tower located?	Berlin, Germany	0.94

Observations

- The model confidently returned incorrect answers (e.g., "Karl Benz" and "Berlin, Germany") based on the misleading context.
- This indicates that the model relies entirely on the provided passage and lacks external validation or reasoning capabilities.
- Edited passages successfully misled the model, even with high confidence.

2.3 Out-of-Domain Passages

Passage 1: COVID-19

Question	Answer	Confidence
What causes COVID-19?	SARS-CoV-2 virus	0.67
Who developed the vaccine?	Pfizer-BioNTech	0.35
How does the virus spread?	through respiratory droplets	0.70

Passage 2: Interstellar Movie

Question	Answer	Confidence
Who directed Interstellar?	Christopher Nolan	0.86
What movie did User1 watch?	Interstellar	0.63
What did User2 think of the visuals?	incredible	0.46

Observations

- On structured scientific text (e.g., COVID-19), the model provided partially correct answers with moderate confidence.
- On conversational data (e.g., Reddit-style chat), confidence was lower (e.g., 0.4698 for casual language like "incredible").
- The model handled structured scientific text better than informal conversational text but struggled with ambiguous questions and contextually rich out-of-domain data.

Overall,

- **In-Domain Passages**: RoBERTa is robust for explicit factual questions but it struggles with ambiguous or unanswerable questions.
- **Edited Passages**: The model is highly vulnerable to misleading information and cannot verify the validity of the context.
- **Out-of-Domain Passages**: RoBERTa generalizes reasonably well to structured texts but struggles with informal or conversational language.

3 Part II

3.1 Dataset

The Covid-QA dataset is structured similarly to SQuAD, containing context paragraphs and question-answer pairs. The dataset is divided as follows:

- Train Split: 104 articles
- Dev Split: 21 articles
- Test Split: 22 articles

The baseline model is a zero-shot RoBERTa model without any additional fine-tuning on Covid-QA. I used the `transformers` library to generate predictions and evaluate them against the test split.

Next, I fine-tuned RoBERTa on the Covid-QA train split to adapt it to the domain using the Hugging Face Trainer API with the following hyperparameters:

- **Batch Size:** 8
- **Learning Rate:** 2e-5
- **Epochs:** 5
- **Warmup Steps:** 500
- **Weight Decay:** 0.01
- **Early Stopping:** Enabled (patience = 2)

Instead of full fine-tuning, I trained an Adapter, which introduces a small set of task-specific parameters to each transformer layer, reducing computational costs. The model is trained on the Covid-QA train split and evaluated on the dev and test splits.

4 Results and Analysis

Model	Dev EM	Dev F1	Test EM	Test F1
RoBERTa-SQuAD2 (Zero-shot)	18.2	19.8	20.5	24.1
Adapter-based RoBERTa	23.7	25.8	34.2	42.1

Table 6: Performance Comparison of Different Models

4.1 Observations

- **Baseline Performance:** The **zero-shot EM score** indicates that RoBERTa trained on SQuAD struggles with biomedical-specific questions.
- **Impact of Full Fine-tuning:** Fine-tuning RoBERTa on Covid-QA train split improves EM and F1 significantly).
- **Adapter-based Performance:** Adapter-based fine-tuning achieves comparable results with fewer trainable parameters, making it more efficient.