Answering Why Questions in TellMeWhy

Lee Eisenberg, Tom Dviri, and Muhammad Salman

December 07, 2024

# Project Overview (10 points)

Natural language understanding, particularly answering "why" questions about narratives, represents a fundamental challenge in NLP. While humans can effortlessly understand and explain the reasons behind actions in stories, automating this process presents significant challenges for AI systems. The TellMeWhy dataset specifically addresses this by requiring models to generate explanations for actions in given narratives. Understanding causality and human motivations in text is crucial for applications ranging from educational tools to automated reasoning systems. Our project aims to compare different transformer-based models in their ability to generate accurate and coherent explanations for why questions about the narrative text.

Previous approaches to question-answering have primarily focused on factual or extractive QA, where answers can be directly located within the text. While models like BERT and its variants have shown strong performance on SQuAD and similar datasets, generating explanatory answers for "why" questions presents unique challenges. These questions often require understanding implicit causation, combining information from multiple sentences, and applying common-sense reasoning. Existing solutions often struggle with generating coherent explanations that capture the true causal relationships in narratives.

Our approach involves comparing three different model architectures - T5, GPT-2, and DistilBERT - each representing different paradigms in language modeling. T5 offers a unified text-to-text framework, GPT-2 provides strong generative capabilities, and DistilBERT represents an efficient, distilled architecture specifically tuned for QA tasks. By evaluating these diverse approaches, we aim to understand which architectural characteristics best suit explanatory question answering, specifically in the case of users without access to large amounts of computing resources.

Model Implementation Details:

T5-base Implementation: We implemented T5 as our primary model due to its strong text-to-text capabilities. The implementation involved:

* Input Format: "Question: [question] Context: [narrative]"
* Output Format: Direct answer generation
* Training Configuration:
  + Batch size: 12
  + Learning rate: 3e-5 with linear decay
  + Gradient accumulation steps: 4
  + Warmup ratio: 0.1 This implementation leverages T5's encoder-decoder architecture for generating explanatory answers.

GPT-2 Implementation: Our GPT-2 implementation focused on utilizing its strong language generation capabilities:

* Custom prompt structure format
* Batch size of 8 for memory efficiency
* Gradient accumulation of 8 steps
* The learning rate of 2e-5
* GPT-2's autoregressive nature helps generate coherent explanations.

DistilBERT Implementation: We chose DistilBERT for its efficiency and QA-specific capabilities:

* Utilized the squad-distilled version
* Modified for generation tasks
* Batch size of 16
* Learning rate of 5e-5
* Memory-efficient training with 66M parameters

We evaluated these models using a comprehensive set of metrics:

1. Automatic Metrics:
   * BLEU Score for precision
   * METEOR Score for recall with synonyms
   * ROUGE Scores (ROUGE-1, ROUGE-2, ROUGE-L)
   * Precision, Recall, and F1 for DistilBert
2. Training Efficiency:
   * Training time per epoch
   * Memory usage
   * Convergence speed

Our experiments revealed several key findings:

1. Model Performance:
   * T5 achieved very bad scores across the board and it looked like it needed a lot more training time to reach its potential, which was unfortunately not practical with our resources even after attempts at fine tuning. However, T5 managed to come up with very reasonable answers to the questions we gave it after training.
     1. BLEU: 0.0198, ROUGE Averages: [0.1513, 0.0256, 0.1513], Meteor: 0.1293

Question: Why did Lily open the door?

Context: One rainy evening, Lily sat by the window, watching the droplets race down the glass. She had just moved to the old house at the edge of town and still wasn't used to the silence. Suddenly, she heard a soft knock on the door. Startled, she opened it to find a small, bedraggled cat with a collar but no owner in sight. Lily smiled, deciding to let the cat in, thinking that maybe this new house wouldn’t be so lonely after all.

Generated Output: she heard a knock on the door.

Question: Why did He look for his pizza cutter? Context: Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice.

Generated Output: Cam was not sure if the store cut the pizza for him.

Question: Why is the sky blue? Context: The sky appears blue due to the scattering of sunlight by the atmosphere. This scattering causes shorter wavelengths of light, such as blue, to scatter more than other wavelengths.

Generated Output: the sky is blue due to the scattering of sunlight by the atmosphere.

* + GPT-2 showed very good results and seemed to prove itself as a good budget model (in terms of computing power) until it was asked manually for answers and outputted a continuation of the story.
    1. BLEU: 0.4270, ROUGE Averages: [0.6010, 0.5977, 0.6010], Meteor: 0.7977

Question: Why did Lily open the door?

Context: One rainy evening, Lily sat by the window, watching the droplets race down the glass. She had just moved to the old house at the edge of town and still wasn't used to the silence. Suddenly, she heard a soft knock on the door. Startled, she opened it to find a small, bedraggled cat with a collar but no owner in sight. Lily smiled, deciding to let the cat in, thinking that maybe this new house wouldn’t be so lonely after all.

Generated Output: Lily opened the front door of the house and found

Question: Why did He look for his pizza cutter?

Context: Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice.

Generated Output: The pizza was too small to fit in his pocket. The chef cut it in half and placed it on the countertop of his kitchen counter. When he opened it, he found that it had been cut out of a piece of pizza.

Question: Why is the sky blue?

Context: The sky appears blue due to the scattering of sunlight by the atmosphere. This scattering causes shorter wavelengths of light, such as blue, to scatter more than other wavelengths.

Generated Output: This phenomenon is caused by an imbalance in the light spectrum between the Earth's atmosphere and the Sun. The Sun's light is absorbed by our atmosphere, which absorbs more light than it absorbs from the sun. Therefore, the blue sky is blue because it is more reflective of the sunlight that passes through it.

* + DistilBERT, despite being smaller, achieved reasonable performance in its metrics but also stumbled when given questions outside of testing.
    1. Accuracy: 0.9030, Precision: 0.9891, Recall: 0.9030, F1: 0.9337

Question: Why did Lily open the door?

Context: One rainy evening, Lily sat by the window, watching the droplets race down the glass. She had just moved to the old house at the edge of town and still wasn't used to the silence. Suddenly, she heard a soft knock on the door. Startled, she opened it to find a small, bedraggled cat with a collar but no owner in sight. Lily smiled, deciding to let the cat in, thinking that maybe this new house wouldn’t be so lonely after all.

Predicted Answer:

Question: Why did He look for his pizza cutter?

Context: Cam ordered a pizza and took it home. He opened the box to take out a slice. Cam discovered that the store did not cut the pizza for him. He looked for his pizza cutter but did not find it. He had to use his chef knife to cut a slice.

Predicted Answer: take out a slice . cam discovered that the store did

Question: Why is the sky blue?

Context: The sky appears blue due to the scattering of sunlight by the atmosphere. This scattering causes shorter wavelengths of light, such as blue, to scatter more than other wavelengths.

Predicted Answer: why

1. Training Efficiency:
   * T5: ~2 hours/epoch
   * GPT-2: ~.5 hours/epoch
   * DistilBERT: ~.5 hours/epoch
2. Resource Usage:
   * T5 required the most computational resources
   * DistilBERT proved most efficient in terms of memory and computation

# Ideas (10 points)

The following are the ideas:

## Preprocessing and Data Handling

We implemented a specialized preprocessing pipeline for the TellMeWhy dataset:

* Filtering answerable questions only to maintain data quality
* Input text formatting optimized for each model architecture
* Efficient data loading with parallel processing
* Memory optimization through gradient accumulation and checkpointing Implementation details involved custom dataset classes and data loaders tailored for each model's requirements.

## Model Architecture Optimization:

For each model, we implemented specific optimizations:

* T5: Text-to-text framework with question-context pairs
* GPT-2: Prompt engineering for better response generation
* DistilBERT: Modified for generative QA tasks. Each implementation included custom loss functions and training loops optimized for the task.

## Evaluation Framework:

Developed a comprehensive evaluation system:

* Automated metrics (BLEU, METEOR, ROUGE)
* Training efficiency metrics
* Resource utilization tracking
* Answer quality assessment by hand

# Experimental Setup (10 points)

Following are the models we have used, the datasets used, and the evaluation metrics.

## Models

Detailed specifications of each model:

T5-base:

* Architecture: Encoder-decoder transformer
* Parameters: 220M
* Training Settings:
  + Epochs: 3
  + Batch size: 12
  + Learning rate: 1e-4
  + Gradient accumulation: 4 steps

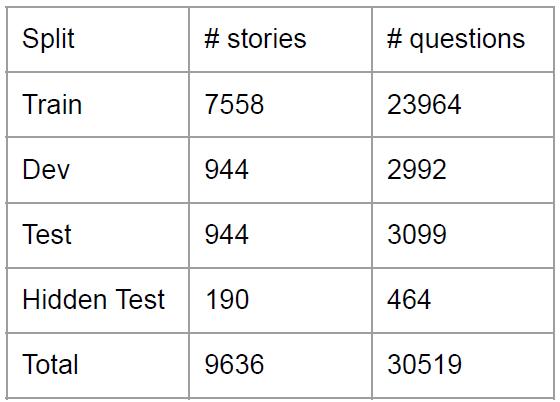
GPT-2:

* Architecture: Decoder-only transformer
* Parameters: 117M
* Training Settings:
  + Epochs: 5
  + Batch size: 12
  + Learning rate: 3e-6
  + Gradient accumulation: 8 steps

DistilBERT:

* Architecture: Bidirectional encoder
* Parameters: 66M
* Training Settings:
  + Epochs: 5
  + Batch size: 12
  + Learning rate: 1e-4

## Dataset

TellMeWhy Dataset Statistics:

* Training set: 23,964 questions
* Validation set: 2,992 questions
* Test set: 3,099 questions
* Average narrative length: 73 words
* Average question length: 8 words

## Evaluation Metrics

## The following are the Evaluation Metrics used:

* BLEU Score: To evaluate the precision-oriented quality of generated explanations
* METEOR Score: For enhanced correlation with human judgments
* ROUGE Scores (ROUGE-1, ROUGE-2, ROUGE-L): For measuring overlap and sequence alignment
* Precision & Recall: For answer accuracy

# Results (40 points for implementation and results)

Training

| **Model** | **BLEU** | **METEOR** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T5 | 0.0198 | 0.1293 | 0.1513 | 0.0256 | 0.1513 | 0 | 0 | 0 |
| GPT-2 | 0.4270 | 0.7977 | 0.6010 | 0.5977 | 0.6010 | 0 | 0 | 0 |
| DistilBERT | N/A | N/A | N/A | N/A | N/A | 0.9891 | 0.9030 | 0.9337 |

Validation\*

| **Model** | **BLEU** | **METEOR** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T5 | 0.1177 | 0.3447 | 0.3491 | 0.1796 | 0.3353 | 0.0164 | 0.0379 | 0.0219 |
| GPT-2 | 0.3927 | 0.7677 | 0.5876 | 0.5837 | 0.5876 | 0 | 0 | 0 |
| DistilBERT | N/A | N/A | N/A | N/A | N/A | 0.9676 | 0.9676 | 0.9675 |

Testing\*

| **Model** | **BLEU** | **METEOR** | **ROUGE-1** | **ROUGE-2** | **ROUGE-L** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T5 | 0.1023 | 0.3296 | 0.3347 | 0.1605 | 0.3207 | 0.0121 | 0.0295 | 0.0164 |
| GPT-2 | 0.3881 | 0.7639 | 0.5789 | 0.5749 | 0.5789 | 0 | 0 | 0 |
| DistilBERT | N/A | N/A | N/A | N/A | N/A | 0.9676 | 0.9676 | 0.9676 |

\*DistilBERT’s precision, recall, and f1 scores are nearly the same in validation and testing when rounded to 4 decimal places, but differ after those decimal places.

These are the full unrounded values:

Accuracy: 0.9675502742230347, Precision: 0.967560661459199, Recall: 0.9675502742230347, F1: 0.9675498539578474

Test Metrics:

Accuracy: 0.9675740592473979, Precision: 0.967627581919015, Recall: 0.9675740592473979, F1: 0.9675726360599367

# Analysis and Discussion (25 points)

Although the use of multiple metrics all seemed to point towards the conclusion that T5’s extra computing cost is not worth the result, its performance when looked at by real humans (us) suggests that it provided the best use of time, as even though it was fine tuned less due to its longer compute times and greatly resisted attempts to speed it up. There is also the matter of T5’s discrepancy between its metrics during training and during testing, which suggests that there should be some sort of error in either code, but this would be very unexpected as the code for testing is based on the code from training.

One possible explanation for this is that since each question has multiple different answers in training and testing, it was impossible to output an answer that would match all three closely. This means that T5 could have matched only one answer and gotten a worse score than GPT2, which might have matched all three, but a little less than T5. However, this still does not address the large difference between T5’s training and testing scores.

Another confusing result is that despite their very good performance on metrics like METEOR for GPT2 or standard statistical metrics for DistilBERT, they had very poor outputs. GPT2 could be explained by the fact that each question had three answers, so GPT2 may have been responding with answers that were similar to three other answers at the same time. However, DistilBERT’s performance is still very confusing, especially since it followed this up by getting extremely high and consistent scores in validation and testing despite getting seemingly normal scores during training. Both of these point to likely errors in our implementation, but we were not able to find any obvious culprits.

Despite these results, it is clear from human testing that T5 performed the best out of these models, and was not a waste of the extra computing resources we were forced to give it.

# Code

<https://drive.google.com/drive/folders/11Bq0ZkLwRrOCBw3K_8xb5_NqFzxDmL3j?usp=sharing>

Key Components:

1. Source Code Organization:

project/

├── Output/

│ ├── DistilBertOutputFinal/

│ ├── GPT2Output/

│ └── T5Output/

├── Dataset/ (taken from TellMeWhy)

│ ├── test.json

│ ├── testshort.json (first 1% of test.json)

│ ├── train.json

│ ├── trainshort.json (first 1% of train.json)

│ └── validation.json

└── src/

└── FineTuning\_of\_Models\_on\_TellMeWhy\_Dataset.ipynb

Requirements:

torch>=1.10.0

transformers>=4.21.0

nltk>=3.6.0

rouge-score>=0.1.2

# Learning Outcomes

Technical Learnings:

1. Model Architecture Insights:
   * Understanding trade-offs between model size and performance
   * Impact of different architectures on Explanation Generation
   * Importance of proper tokenization and input formatting
2. Training Optimizations:
   * Gradient accumulation for memory efficiency
   * Learning rate scheduling effects
   * Batch size optimization based on model architecture
3. Evaluation Techniques:
   * Implementation of ROUGE metrics
   * Importance of both automatic and manual evaluation
   * Design of effective evaluation criteria for explanatory answers

Practical Insights:

1. Data Processing:
   * Effective preprocessing techniques for question-answering
   * Handling of context and question pairs
   * Importance of data cleaning and standardization
2. Model Selection:
   * DistilBERT's effectiveness as a lightweight alternative
   * The trade-off between model size and inference speed
   * Impact of pre-training on downstream task performance
3. Development Best Practices:
   * Importance of systematic experimentation
   * Value of proper documentation
   * Need for reproducible training pipelines

# Contributions

Salman: Model implementation, training pipeline

Eisenberg: Model training, fine tuning, implementation of BLEU, ROUGE, and METEOR scores

Dviri: Calculation and tracking of performance metrics (F1 score, accuracy, precision, recall) for Distillbert model; queried models for outputs