ANLP- Ex1 - Inbar Bahnof 207009721

Link to Git Repo

Question 1 - Datasets that use QA to annotate intrinsic concepts

1. **nvidia - OpenMathReasoning**

   OpenMathReasoning is a large-scale math reasoning dataset for training large language models. It annotates an intrinsic concept because reasoning requires the model to deeply comprehend the semantic content, structure, and implications of the text - including understanding mathematical terminology, recognizing logical relationships, and executing multi-step reasoning based on the text input, rather than relying on surface level understanding.

2. **google - BoolQ**

   BoolQ is a question answering dataset for yes/no questions that are paired with short passages from Wikipedia. The dataset annotates an intrinsic concept because answering requires understanding the context of the question at a semantic level and performing reasoning to determine whether the text implies that the answer to the question is true or false - testing comprehension of the text.

3. **Stanford Question Answering Dataset - SQuAD**

   Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. The dataset annotates an intrinsic concept because it tests deeper comprehension skills by requiring models not only to locate relevant information but also to perform reasoning operations, understanding intrinsic logical and general understanding of text, and also to understand if the information that requires to answer the question is even in the text.

Question 2 (a)

Self-consistency

- Description - Sampling a diverse set of reasoning paths - chain of thought prompts, and choose the most consistent or majority answer among them.
- Advantages - Encouraging the model to explain the reasoning behind its answer can ultimately lead to a better answer. In addition, comparing a few different answers can also lead to better and more accurate and stable answers.
- Computational Bottlenecks - Sampling many outputs dramatically increases the compute required at inference time,and storing and evaluating multiple outputs can slow down response time and increase memory usage.
- Can the computation be parallelized? - Yes, since this method samples independent paths of chain-of-thought, each of them can be computed seperatly.

Verifiers

- Description - Using small models that are trained to answer good/bad on a certain answer to a task to verify the answers - Instead of taking a majority of all generated chains, select the majority\best of verified answers.
- Advantages - Verifiers help to estimate if the answer is correct or incorrect. and by that improve the model's answers accordingly. The fact that verifiers can be trained separately, means they can be used to post-process outputs from different base models.
- Computational Bottlenecks - Using verifiers requires additional forward passes through the verifier model, which is increasing total inference cost. In addition, high quality verifiers may be a big model on their own.
- Can the computation be parallelized? - Yes, since verifier evaluations of different candidate outputs are independent, so each of them can be computed seperatly.

Increasing Compute Budget

- Description - Increasing the budget of inference time can improve the results. We can either increase the model parameters, or have a few smaller models that compute the same thing and taking their average/the answer that most of them agree on.

- Advantages - Larger or multiple models generally lead to better performance because they bring more capacity and diverse reasoning. In addition, multiple models reduce variance and are less likely to make the same mistake, improving reliability.
- Computational Bottlenecks - Larger models or multiple models require significantly more memory, compute, and energy, during Inference time.
- Can the computation be parallelized? - Yes, independent model can be computed separately, and even a very large model can be computed simultaneously.

Extending Reasoning Depth with Chain-of-Thought Techniques

- Description - This approach involves increasing the length and complexity of the reasoning process of the model during inference. Techniques such as Planning - make the model plan what it needs to do before doing it, Backtracking - revising incorrect reasoning paths, and Self-Evaluation - assessing the quality of generated reasoning, are used to produce more thoughtful and accurate answers.
- Advantages - This method encourages the model to break down complex problems into understandable steps, improving factual and logical correctness. Also long and complex reasoning makes model outputs more transparent and easier to debug or trust.
- Computational Bottlenecks - Longer reasoning chains mean more tokens to generate, which increases the computational power it requires - techniques like backtracking or planning often involve custom logic or repeated forward passes.
- Can the computation be parallelized? - Partly - if planning or self-evaluation are independent, they can be generated and computed in parallel. However, techniques like backtracking may require sequential steps or dependency tracking, which limits parallelism.

Question 2 (b)

I would extend Reasoning Depth with Chain-of-Thought Techniques - Because of the fact that not like self-consistency or ensemble methods, that are the most efficient when using a few cores simultaneously, chain-of-thought reasoning can be done sequentially, making it more efficient for a single-GPU setup. In addition, a large-memory GPU enables generating longer reasoning chains without reaching memory limits, which is crucial for complex tasks that require detailed step-by-step

thinking. Also chain-of-thought reasoning helps verify correctness or catch errors in logic - which is very important when solving a complex task.

Practical Part - 2

Results on validation:

1. epoch_num: 2, lr: 0.1, batch_size: 16, eval_acc: 0.6838
2. epoch_num: 3, lr: 0.001, batch_size: 16, eval_acc: 0.6838
3. epoch_num: 2, lr: 2e-05, batch_size: 16, eval_acc: 0.8333
4. epoch_num: 4, lr: 3e-05, batch_size: 32, eval_acc: 0.8529

Results on test:

1. epoch_num: 2, lr: 0.1, batch_size: 16, test_acc: 0.664927536231884
2. epoch_num: 3, lr: 0.001, batch_size: 16, test_acc: 0.664927536231884
3. epoch_num: 2, lr: 2e-05, batch_size: 16, test_acc: 0.7953623188405797
4. epoch_num: 4, lr: 3e-05, batch_size: 32, test_acc: 0.8127536231884058

As we can see, the configuration that achieved the best validation accuracy did also achieve the best test accuracy.

Examples of sentences the worst and best models didn't agree on:

1. Sentence 1: The company didn't detail the costs of the replacement and repairs .
   Sentence 2: But company officials expect the costs of the replacement work to run into the millions of dollars .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1
2. Sentence 1: Air Commodore Quaife said the Hornets remained on three-minute alert throughout the operation .
   Sentence 2: Air Commodore John Quaife said the security operation was unprecedented .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

3. Sentence 1: The broader Standard & Poor 's 500 Index < .SPX > was 0.46 points lower , or 0.05 percent , at 997.02 .
   Sentence 2: The technology-laced Nasdaq Composite Index .IXIC was up 7.42 points , or 0.45 percent , at 1,653.44 .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

4. Sentence 1: Hong Kong was flat , Australia , Singapore and South Korea lost 0.2-0.4 percent .
   Sentence 2: Australia was flat , Singapore was down 0.3 percent by midday and South Korea added 0.2 percent .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

5. Sentence 1: Last year , Comcast signed 1.5 million new digital cable subscribers .
   Sentence 2: Comcast has about 21.3 million cable subscribers , many in the largest U.S. cities .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

6. Sentence 1: " It was a little bit embarrassing the way we played in the first two games , " Thomas said .
   Sentence 2: " We 're in the Stanley Cup finals , and it was a little bit embarrassing the way we played in the first two games .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

7. Sentence 1: Against the Japanese currency , the euro was at 135.92 / 6.04 yen against the late New York level of 136.03 / 14 .
   Sentence 2: The dollar was at 117.85 yen against the Japanese currency , up 0.1 percent .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

8. Sentence 1: Oracle 's $ 5 billion hostile bid for PeopleSoft is clearly not sitting well with PeopleSoft 's executives .
   Sentence 2: Oracle on Friday launched a $ 5.1 billion hostile takeover bid for PeopleSoft .
   True Label: 0

Best Model Prediction: 0
Worst Model Prediction: 1

9. Sentence 1: Against the Japanese currency , the euro was at 135.92 / 6.04 yen against the late New York level of 136.03 / 14 .
   Sentence 2: The dollar was at 117.85 yen against the Japanese currency , up 0.1 percent .
   True Label: 0
   Best Model Prediction: 0
   Worst Model Prediction: 1

Key Patterns in Harder Examples for the Lower-Performing Model

- A few pairs mention different indices or numerical values - The lower-performing model seems to misinterpret these as contradictory, likely due to surface-level differences in numbers.
- A few paris require knowledge about the same entity or event being described in different ways - this suggests that the weaker model is less able to understand deeper meanings about the world.