Name: David Dinkevich

ID: 666226238

Email: david.dinkevich@mail.huji.ac.il

GitHub Repo: https://github.com/DavidDinkevich/ANLP-ex1

Advanced NLP: Homework 1 Report

Section 1:

Question 1:

1. SQuAD

SQuaD measures reading comprehension and sentence-level inference, since given text and a question, the model has to locate and understand relevant information. So this forces it to learn the intrinsic ability to parse text and understand paraphrasing, etc.

2. BoolQ

BoolQ tests the model's ability to reason over a text and predict a Boolean answer. It requires semantic entailment, negation handling, and inference which are intrinsic tasks.

3. DROP

DROP requires models to perform arithmetic and reason across sentences to answer questions. The model needs to be capable of compositional reasoning and understanding time, which are intrinsic tasks.

Question 2a:

1. Prompt Ensembling: you run the model with different versions of the same prompt (different wordings) and pick the most popular answer
    a. Advantages: reduces the chances of "falling on bad wording", more likely to leverage the model's knowledge. Less dependent on the exact wording of the prompt.
    b. Computational Bottlenecks: you have to run the entire model multiple times
    c. Parallelizable: yes, run the model in parallel—there is no conflict here
2. Self-Consistency: when using Chain of Thought, we sample different "reasoning paths" and take the answer that is reached by the most reasoning paths. One way to do this is stochastic decoding (sample the next token according to the softmax

probabilities)—this way you get different answers for each prompt, and therefore different reasoning chains.

   a. Advantages: this makes reasoning even more effective, because if multiple different reasoning paths all reach the same answer, than that answer is likely to be correct, as opposed to following a single reasoning path that might be flawed.
   b. Computational Bottlenecks: need to run multiple forward passes, which can be expensive if the reasoning chains are long
   c. Parallelizable: yes, do the sampling on different GPUs

3. Verifiers: first we generate responses with the model (these can be CoT), and then we train a separate verifier model to pick the most likely correct answer.
   a. Advantages: helps to avoid hallucination, can do unit tests for code, eval math expressions. We can also use Retrieval Augmented Generation.
   b. Computational Bottlenecks: have to run an additional model
   c. Parallelizable: yes—verification of each candidate answer can be done in separately

Question 2b:

I would choose Self Consistency + Chain of Thought. I only have one GPU, so I probably don't have enough vram to store both the model and another verifier model (this rules out Verifiers).

Also, since we're dealing with complex scientific facts, Chain of Thought is really important and will probably make a big difference, so Prompt-Ensembling is probably not enough, we need that extra power. And since I have a lot of vram, I can afford to generate long answers and get several reasoning paths, and then pick the best one.

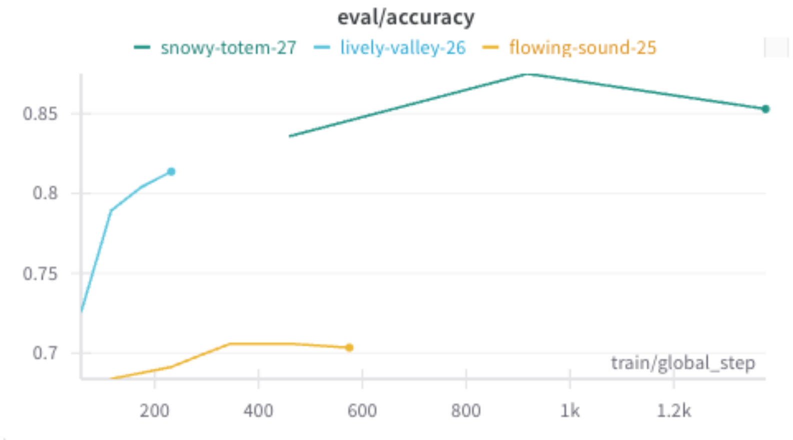
Section 2.1:

Did the configuration that achieved the best validation accuracy also achieve the best test accuracy?

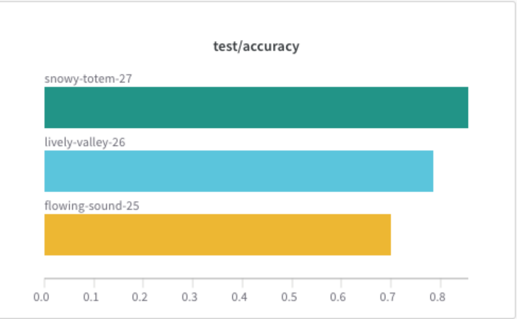
We ran 3 trained checkpoints as instructed:

1. "flowing-sound-25": epoch_num: 5, lr: 1e-06, batch_size: 32, eval_acc: 0.7034
2. "lively-valley-26": epoch_num: 4, lr: 1e-05, batch_size: 64, eval_acc: 0.8137
3. "snowy-totem-27": epoch_num: 3, lr: 3e-05, batch_size: 8, eval_acc: 0.8529

It can be seen that snowy-totem-27 got the highest evaluation accuracy:

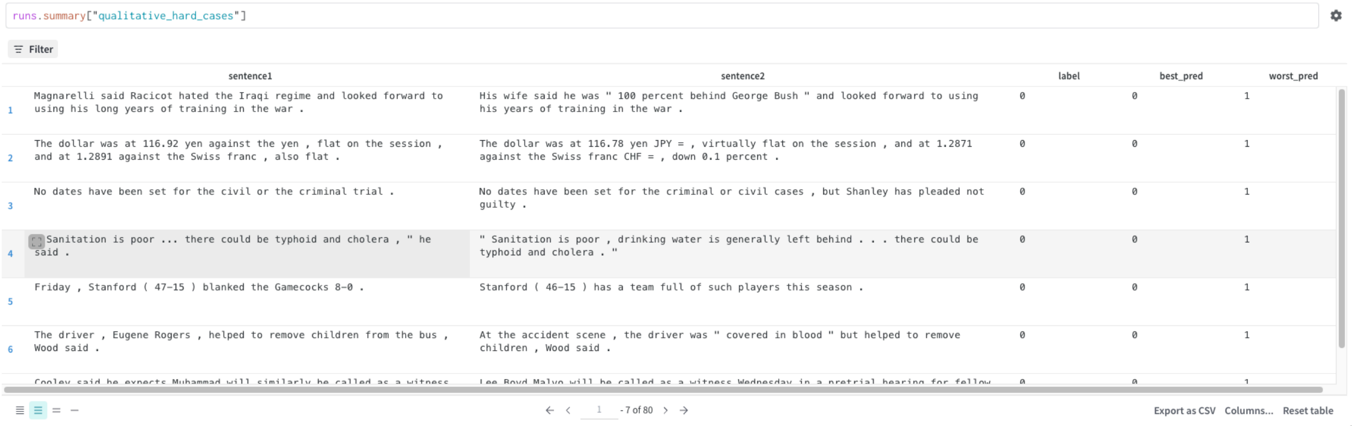


It also got the highest test accuracy:



Qualitative Analysis:

In order to do the qualitative analysis, we took the best run, “snowy-totem-27” and the worst run, “flowing-sound-25”, and analyzed examples in which the model predicted correctly in snowy-totem-27 and wrongly in flowing-sound-25:



```
runs.summary["qualitative_hard_cases"]
```

	sentence1	sentence2	label	best_pred	worst_pred
1	Magnarelli said Racicot hated the Iraqi regime and looked forward to using his long years of training in the war .	His wife said he was " 100 percent behind George Bush " and looked forward to using his years of training in the war .	0	0	1
2	The dollar was at 116.92 yen against the yen , flat on the session , and at 1.2891 against the Swiss franc , also flat .	The dollar was at 116.78 yen JPY = , virtually flat on the session , and at 1.2871 against the Swiss franc CHF = , down 0.1 percent .	0	0	1
3	No dates have been set for the civil or the criminal trial .	No dates have been set for the criminal or civil cases , but Shanley has pleaded not guilty .	0	0	1
4	Sanitation is poor ... there could be typhoid and cholera , " he said .	" Sanitation is poor , drinking water is generally left behind . . . there could be typhoid and cholera . "	0	0	1
5	Friday , Stanford (47-15) blanked the Gamecocks 8-0 .	Stanford (46-15) has a team full of such players this season .	0	0	1
6	The driver , Eugene Rogers , helped to remove children from the bus , Wood said .	At the accident scene , the driver was " covered in blood " but helped to remove children , Wood said .	0	0	1
	Cooley said he expects Muhammad will similarly be called as a witness	Lee Boyd Malvo will be called as a witness Wednesday in a pretrial hearing for fellow	0	0	1

We noticed some common patterns that led the lower performing model to guess wrongly:

1. <u>Common phrases:</u> sometimes when the same phrase or sequence of words is used in both sentences, the model will guess that the sentences are equivalent, even though the other part of each sentence makes them inequivalent.
    a. Example 1:
        i. Sentence 1: *"Magnarelli said Racicot hated the Iraqi regime and looked forward to using his long* ***years of training in the war*** *."*
        ii. Sentence 2: *"His wife said he was " 100 percent behind George Bush " and looked forward to using his* ***years of training in the war .****"*
    b. Example 2:
        i. Sentence 1: *"Their contract will expire at 12 : 01 a.m. Wednesday instead of 12 : 01 a.m. Sunday , said Rian Wathen , organizing director for* ***United Food and Commercial Workers Local 700 .****"*
        ii. Sentence 2: *"It has outraged the membership , " said Rian Wathen , organizing director of* ***United Food and Commercial Workers Local 700*** *."*
2. <u>Extra information:</u> sometimes sentence 2 contains all of the information in sentence 1, and adds extra information to it. In this case, the lower-performing model will predict that the statements are equivalent, even though they technically aren't because the second sentence added new information:
    a. Example 1:
        i. Sentence 1: *"No dates have been set for the civil or the criminal trial ."*
        ii. Sentence 2: *"No dates have been set for the criminal or civil cases , but Shanley has pleaded not guilty ."*
    b. Example 2:
        i. Sentence 1: *"He tried to fight off officers and was taken to a hospital after a police dog bit him but was later released ."*
        ii. Sentence 2: *"Cruz tried to fight off officers and was hospitalized after a police dog bit him , Sgt. Steve Dixon said ."*

Sometimes the lower performing model will classify two statements as inequivalent, even though they are equivalent:

3. <u>Similar phrase beginning:</u>
    a. Sentence 1: *"****The central bank is*** *expected to announce a new round of interest rate hikes during its upcoming policy meeting."*
    b. Sentence 2: *"****The central bank is*** *likely to unveil another set of interest rate increases at its next policy gathering."*