GPT Limitations

GPT & Benchmarks

Learning goals

- Understand architectural limitations of GPT-3
- Understand differences to human learning and thinking

LIMITATIONS OF GPT3: TEXT GENERATION

- Repetitions
- Lack of coherence
- Contradictions

LIMITATIONS OF GPT3: COMMON SENSE

- Common sense physics
- E.g., “If I put cheese in the fridge, will it melt?”
- See below

LIMITATIONS OF GPT3: COMPARISON TASKS

- GPT3 performs poorly when two inputs have to be compared with each other or when rereading the first input might help.
- E.g., is the meaning of a word the same in two sentences (WiC).
- E.g., natural language inference, e.g., ANLI
- Not a good match for left-to-right processing model.
- Possible future direction: bidirectional models

LIMITATIONS OF GPT3: SELF-SUPERVISED PREDICTION ON TEXT

- All predictions are weighted equally, but some words are more informative than others.
- Text does not capture the physical world.
- Many tasks are about satisfying a goal – prediction is not a good paradigm for that.

LIMITATIONS OF GPT3: LOW SAMPLE EFFICIENCY

- Humans experience much less text than GPT3, but perform better.
- We need approaches that are as sample-efficient as humans, i.e., need much less text for same performance.

LIMITATIONS OF GPT3: SIZE/ INTERPRETABILITY CALIBRATION

- Difficult to use in practice due to its size.
- Behavior hard to interpret
- Probability badly calibrated

DISCUSSION: DOES GPT3 “LEARN” FROM CONTEXT?

- GPT3 learns a lot in pretraining.
- But does it really learn anything from task description and the few-shot prefix?
- Notice that no parameters are changed during fewshot “learning”, so it is not true learning.
- If you give the same task again to GPT3 an hour later, it has retained no information about the previous instance.
- How much of human learning is “de novo”, how much just uses existing scales.