Problems with BERT

GPT & Benchmarks

Learning goals

- Recap BERT-like models
- Understand problems models following this paradigm have

RECAP: BERT, ROBERTA ETC.

- Transformer
- Training: Masked language modeling (MLM)
- BERT learns an enormous amount of knowledge about language and the world through MLM training on large corpora.
- Application: finetune on a particular task
- Great performance!
- What's not to like?
- (In what follows I will use BERT as a representative for this class of language models and only talk about BERT – but the discussion includes RoBERTa, Albert, XLNet etc.)

PROBLEMS WITH BERT (1)

- You need a different model for each task.
- (Because BERT is differently finetuned for each task.)
  - Not realistic in many real deployment scenarios, e.g., on mobile devices.
- Human learning: we arguably have a single model that solves all tasks!
- Question: Is there a framework that allows us to create a single model that solves all tasks?

PROBLEMS WITH BERT (2)

- BERT has two training modes, first (MLM) pretraining, then finetuning.
- Finetuning is supervised learning, i.e., learning from labeled examples.
- Arguably, learning from labeled examples is untypical for human learning.
- You never learn a task solely by being presented a bunch of examples, without explanation.
- Instead, in human learning, there is almost always a task description.
- Example: How to boil an egg. “Place eggs in the bottom of a saucepan. Fill the pan with cold water. Etc.”
- (Notice that this is not an example.)
- Question: Is there a framework that allows us to leverage task descriptions?

PROBLEMS WITH BERT (3)

- BERT has great performance, but . . .
- . . . it only has great performance if the training set is fairly large, generally 1000s of examples.
- This is completely different from human learning!
- We do use examples in learning, but in most cases, only a few.
- Example: Maybe the person teaching you how to boil an egg will show you how to do it one or two times.
- But probably not 10 times
- Definitely not a 1000 times
- More practical concern: it's very expensive to label 1000s of examples for each task (there are many many tasks).
- Question: Is there a framework that allows us to learn from just a small number of examples?
- This is called few-shot learning.

PROBLEMS WITH BERT (4)

- More subtle aspect of the same problem (i.e., large training sets): overfitting
- Even though performance looks good on standard train/dev/test splits,
- the deviation between the training set and the data actually encountered in real application can be large.
- So our benchmarks often overestimate what performance would be in reality.