True/False Question Answering: Comparing simple transformers with state-of-the-art mega-models

Andrew Higgins
Datasci 266: Natural Language Processing
UC Berkeley School of Information

Project overview

Boolean Question Answering

Dataset:

- BoolQ - Yes/No Questions

Models:

- BERT (base & finetuned)
- T5 (base & finetuned)
- ST-MoE (state-of-the-art)

Background

Background - Question Answering

Domain of information: Closed-Domain or Open-Domain

- **Closed**: Models are restricted to only information in BoolQ dataset

Task type: Extractive or Abstractive:

- BERT - Binary classification task
- T5 - **Extract** “Yes” or “No” from answer options

Background: QA benchmarking: SuperGLUE

SuperGLUE is a **Language Understanding benchmark**.

- For assessing model performance on various tasks:
  - Recognizing textual entailment
  - Reading comprehension with commonsense reasoning
  - Words in context
  - Multiple-choice QA
  - BoolQ binary QA
  - More
- ST-MoE is 2nd on the leaderboard overall
  - Highest BoolQ performance overall

Background - ST-MoE, a QA state-of-the-art

Google Brain, 2022.

MoE - “Mixture of Experts”

- Multiple expert networks divide the problem space into specialized regions.

ST - “Switch Transformer”

- An auxiliary network that efficiently routes vectors through relevant experts
- Allows for specialization while maintaining reasonable compute demand

Background - ST-MoE

- Beats SOTA at many tasks, even when scaled down to equal compute demand
- Full sized version (shown) is SOTA for most tasks tested
  - SQuAD
  - SuperGLUE
  - ARC (gradeschool MC tests)

			Previous Best (↑)			Ours (↑)
Name	**Metric**	**Split**	Zero-Shot	One-Shot	Fine-Tune	Fine-Tune
SQuADv2	F1	dev	68.3[e]	70.0[e]	96.2[a]	**96.3**
SQuADv2	acc	dev	62.1[e]	64.6[e]	**91.3**[a]	90.8
SuperGLUE	avg	test	–	–	90.9	**91.2**
BoolQ	acc	dev/test	83.0[e]	82.8[e]	92.0	**92.4**
Copa	acc	dev/test	91.0[d]	92.0[e]	98.2	**99.2**
RTE	acc	dev/test	68.8[e]	71.5[e]	**94.1**	93.5
WiC	acc	dev/test	50.5[e]	52.7[e]	**77.9**	77.7
MultiRC	F1	dev/test	72.9[d]	72.9[d]	88.6	**89.6**
WSC	acc	dev/test	84.9[e]	83.9[e]	**97.3**	96.6
ReCoRD	acc	dev/test	90.3[e]	90.8[e]	**96.4**	95.1
CB	acc	dev/test	46.4[d]	73.2[e]	**99.2**	98.0
XSum	ROUGE-2	test	–	–	24.6[h]	**27.1**
CNN-DM	ROUGE-2	test	–	–	21.6[a]	**21.7**
WinoGrande XL	acc	dev	73.4[e]	73.2[d]	–	96.1
ANLI R3	acc	test	40.9[e]	40.8[e]	53.4	**74.7**
ARC-Easy	acc	test	71.9[e]	76.6[e]	92.7[g]	**95.2**
ARC-Challenge	acc	test	51.4	53.2	81.4[g]	**86.5**
CB TriviaQA	em	dev	68.0[e]	**74.8**[e]	61.6[b]	62.3
CB NatQA	em	test	21.5[e]	23.9[e]	41.5[c]	**41.9**
CB WebQA	em	test	38.0[f]	25.3	42.8[b]	**47.4**

Table 12: **ST-MoE-32B versus previous best for inference-only techniques and fine-tuned models.** A split of "dev/test" refers to dev split for Zero-Shot and One-Shot and test split for Fine-Tune quality. Data not available filled in with "–". Superscript letters denote the result: [a]: Raffel et al. (2019) [b]: Roberts et al. (2020) [c]: Karpukhin et al. (2020), [d]: Brown et al. (2020), [e]: Du et al. (2021), [f]: Wang et al. (2021), [g]: UnifiedQA + ARC MC/DA + IR, [h]: Zhang et al. (2020).

Dataset

BoolQ

Yes/No question answering database created by Google in 2019

- Built by humans from yes/no-like Google queries
- Answered with Wikipedia
- Context passage copied from the relevant Wikipedia article

BoolQ - Dataset constituents

12697 Examples (9427 test, 3270 validation)

Each entry consists of 4 parts:

- Title - Not used
- Question
- Passage - context paragraph
- Answer - boolean
  - Yes - 62.3%
  - No - 37.7%

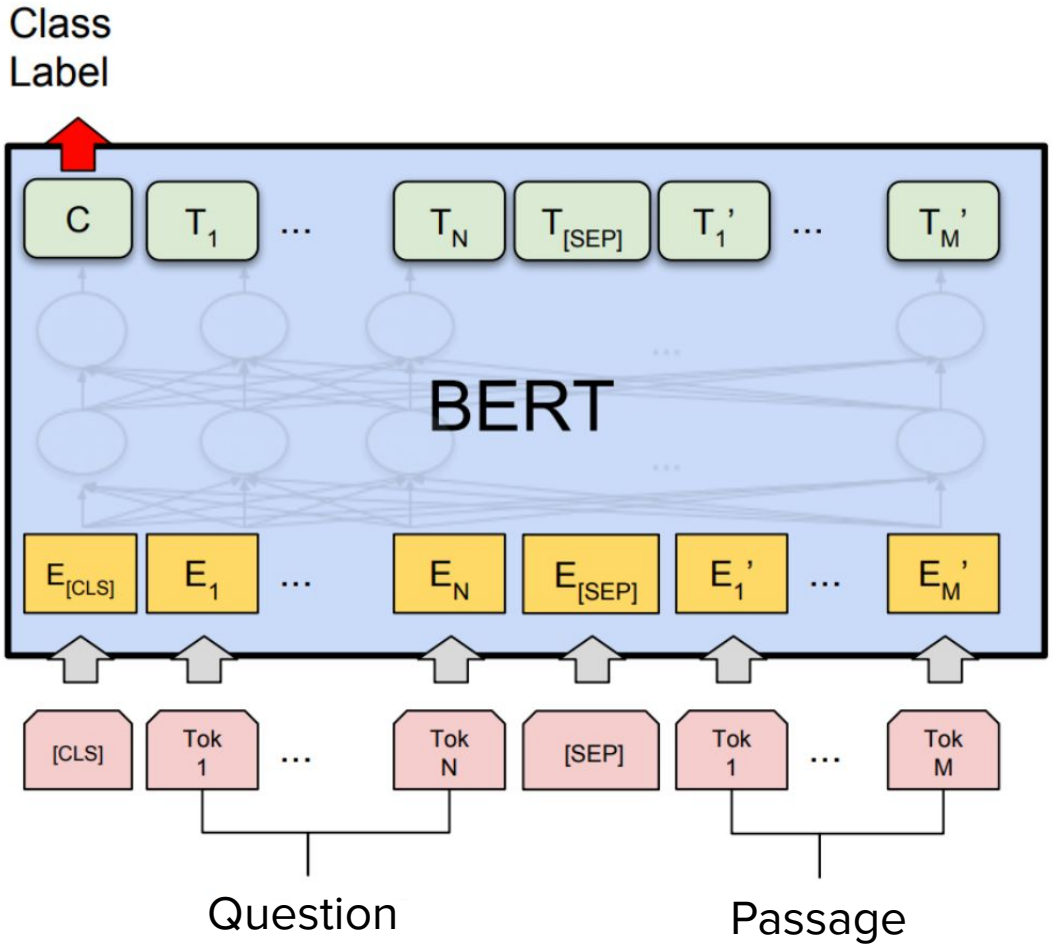
BoolQ - Examples

Answer	Passage	Question	Title
False	Both the violin and viola are played under the jaw. The viola, being the larger of the two instruments, has a playing range that reaches a perfect fifth below the violin's. The cello is played sitting down with the instrument between the knees, and its playing range reaches an octave below the viola's. The double bass is played standing or sitting on a stool, with a range that typically reaches a minor sixth, an octave or a ninth below the cello's.	is a cello and a bass the same thing	Violin family
True	Open Water is a 2003 American survival horror thriller film. The story concerns an American couple who go scuba diving while on vacation in the Caribbean, only to find themselves stranded miles from shore in shark-infested waters when the crew of their boat accidentally leaves them behind. The film is loosely based on the true story of Tom and Eileen Lonergan, who in 1998 went out with a scuba diving group, Outer Edge Dive Company, on the Great Barrier Reef, and were accidentally left behind because the dive-boat crew failed to take an accurate headcount. The film	is open water based on a true story	Open Water (film)

Models

BERT - Encoder

- <CLS> header token passed to a binary classification layer.
- Input length truncated/padded to 128
- BERT-base-uncased
  - 11 transformer layers
- DistilBERT-base-cased-distilled-SQuAD
  - 5 transformer layers



BERT - variants used

BERT-base-uncased

- 11 transformer layers
- No task-specific pretraining

DistilBERT-base-cased-distilled-SQuAD

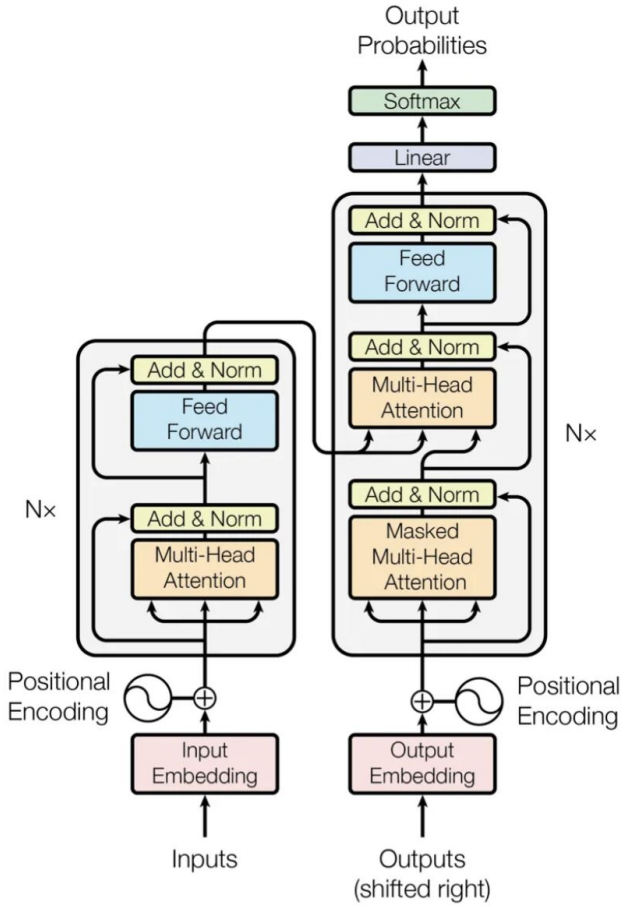
- 5 transformer layers
- Pretrained on Stanford QA Database (SQuAD)

T5 - Encoder-Decoder

- Could not directly use T5 for binary classification.
- Prompt design:

“question: {BoolQ question} (A) Yes (B) No context: {BoolQ passage}”

- Output post-processing
  - Convert generated “yes” or “no” to Boolelan
  - This became complicated...



T5 - variants used

T5-base

- 220 million parameters
- No task-specific fine-tuning

T5-small-finetuned-BoolQ - From user “mrm8488” on HuggingFace

- 60 million parameters
- Fine tuned on BoolQ

Results

Accuracy

- Models fall between baseline and SOTA, as expected
- T5-base is the exception, as the model failed to output in the correct format of “yes/no”

	Model	Accuracy (↑)
	Baseline (predict “yes”)	62.3%
BERT	base-uncased	64.3%
	BERT-base-cased-distilled-SQuAD	65.3%
T5	T5-base	~
	T5-small-finetuned-BoolQ	85%
	State-of-the-art (ST-MoE)	**92.4%**

Results - BERT

- Both models only performed slightly better than baseline
- SQuAD fine-tuned model only performed marginally better than base model
  - Accuracy was unstable throughout training
  - Epoch 1 had the highest validation accuracy
- Future work
  - Hyperparameter tuning to prevent overfitting
  - Fine tune on much larger dataset

Results - T5 base

Output behavior

```
['Philadelphia', 'No', 'Hal', 'No', 'Richard', 'No', 'No', 'No', 'No', 'No']
```

- No “yes”, only “no” and other words
- Most common outputs (out of 100 examples) ➜
- Hypothesis: model is trying to justify “yes” answer, instead of just outputting it
  - Result: Not valid
  - Accuracy: 47%
- Attempts to change prompt structure did not help
- Considered a failed model

No	45
	2
third	2
Tom	2
The	2
no	2
comedy	2
in	1
double	1
each	1

Results - T5-small-finetuned-BoolQ

Output behavior `['yes', 'no', 'yes', 'yes', 'no', 'yes', 'yes', 'yes', 'yes', 'yes']`

- Major improvement over T5-base
- Model only outputs “yes” or “no”
- Accuracy of 85% is close to SOTA (92.4%)
- Impressive for a model less than 10% the size of ST_MoE

End