SuperGLUE

A Stickier Benchmark for General-Purpose Language Understanding Systems

Alex Wang*, Yada Prukaschatkun*, Nikita Nangia*, Amanpreet Singh*, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman

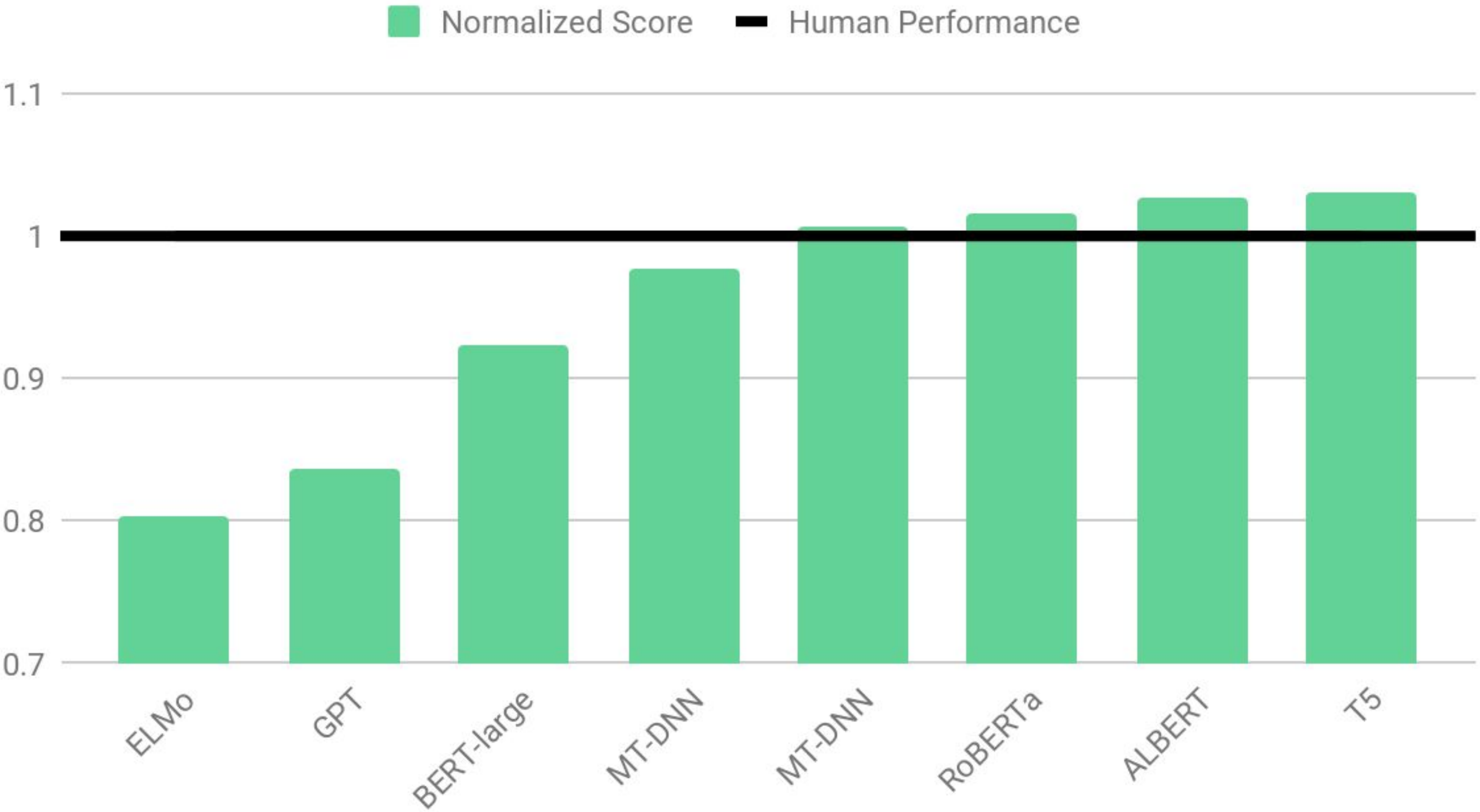
Motivation

- High-level: want robust, general-purpose NLU systems
- SuperGLUE goals
    - Standardize evaluation
    - Provide single-number metric that reflects NLU ability
- Make it easy for non-domain experts to work on these problems

First Attempt: GLUE

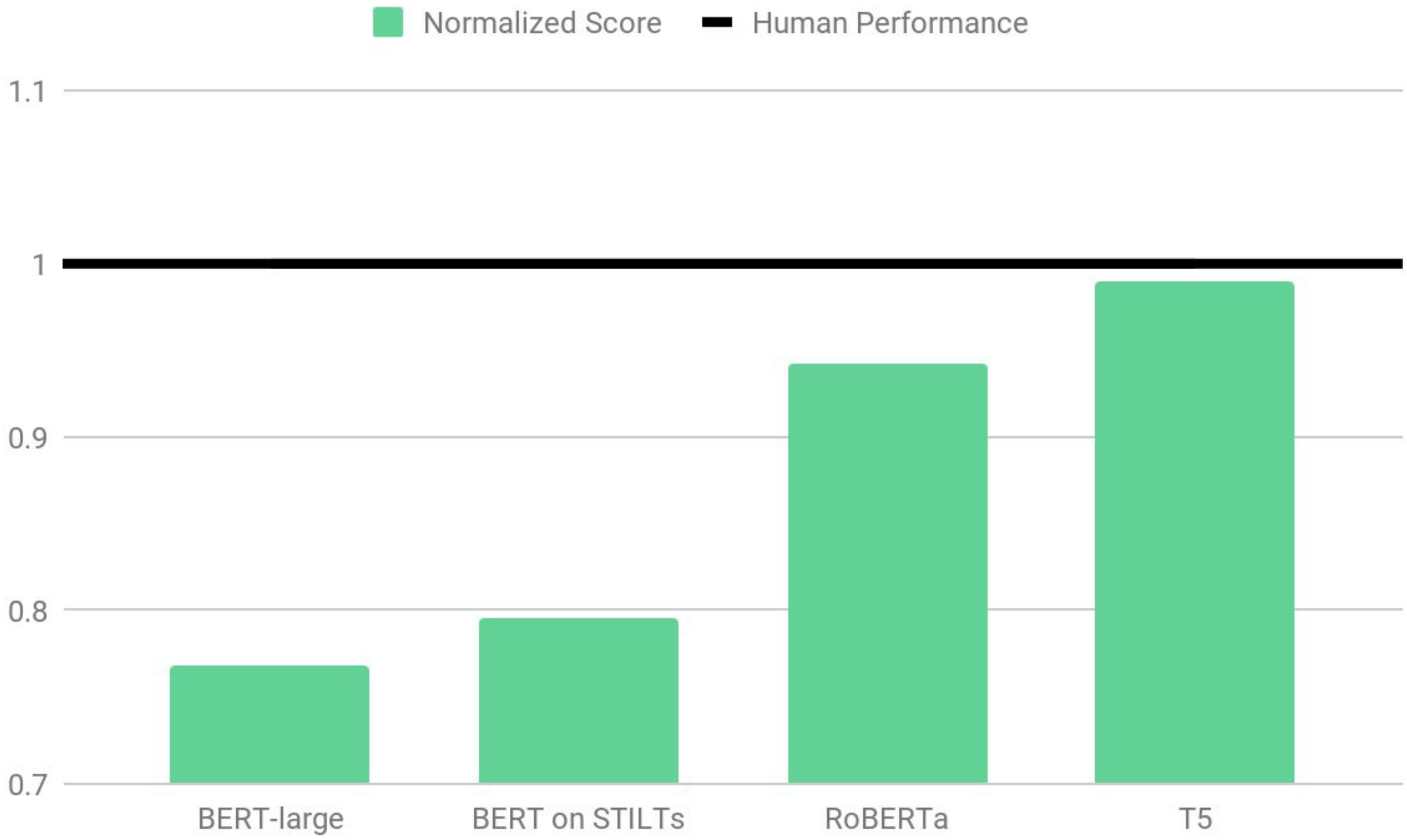
- Benchmark of 9 sentence- and sentence-pair classification tasks
  - Different tasks (sentiment analysis, paraphrase detection, etc.), genre, amount of data
  - Evaluate system on all nine tasks; overall score is average across tasks
- Released May 2018

ELMER'S
Glue-All
MULTI-PURPOSE GLUE
For Household Repairs, Craft Projects, & more!
Extra Strong Formula
Dries Fast
Safe & Non-Toxic
WARNING: CHOKING HAZARD—Small parts. Not for children under 3 years.
16 fl oz (473 mL)


🟩 Normalized Score
Human Performance
1.1
1
0.9
0.8
0.7
ELMo
GPT
BERT-large
MT-DNN
MT-DNN
RoBERTa
ALBERT
T5

SuperGLUE

- New benchmark of 8 NLU tasks
- Also:
  - Additional diagnostics
  - Rules updates
  - Starter code
- Tasks were selected from an open call to the NLP community
  - Screen each proposed task to be easy for humans, hard for machines
  - Emphasized tasks with little training data
  - More diverse set of task formats, e.g. QA, coreference
- Released May 2019



🟩 Normalized Score
▬ Human Performance
1.1
1
0.9
0.8
0.7
BERT-large
BERT on STILTs
RoBERTa
T5

Takeaways

- Real, robust recent progress in NLP
- NLU is not solved!
  - Models are susceptible to adversarial inputs (e.g., Jia et al. 2017)
  - Models rely on shortcut heuristics (e.g., McCoy et al., 2019)
- SuperGLUE is a good testbed for:
  - Sample-efficient learning
  - Multi-task learning
  - Learning w/ limited data
  - Model distillation and compression
- SustaiNLP workshop @ EMNLP

super.gluebenchmark.com