The benefits of unified frameworks for language understanding

Colin Raffel
UNC Chapel Hill and Hugging Face
Text-to-Text Transfer Transformer

T5
"translate English to German: That is good."

T5

"Das ist gut."
"cola sentence: The course is jumping well."

T5

"not acceptable"
"stsB sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi."

"six people hospitalized after a storm in attala county."
"translate English to German: That is good."

"cola sentence: The course is jumping well."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."

"stsib sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"Das ist gut."

"not acceptable"

"6.8"

"six people hospitalized after a storm in attala county."
How to Fine-Tune BERT for Text Classification?

Chi Sun, Xipeng Qiu*, Yige Xu, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, China
{sunc17,xpqiu,ygxu18,xjhuang}@fudan.edu.cn

ON THE STABILITY OF FINE-TUNING BERT: MISCONCEPTIONS, EXPLANATIONS, AND STRONG BASELINES

Marius Mosbach
Spoken Language Systems (LSV)
Saarland Informatics Campus, Saarland University
mmosbach@lsv.uni-saarland.de

Maksym Andriushchenko
Theory of Machine Learning Lab
École polytechnique fédérale de Lausanne
maksym.andriushchenko@epfl.ch

Dietrich Klakow
Spoken Language Systems (LSV)
Saarland Informatics Campus, Saarland University
dietrich.klakow@lsv.uni-saarland.de

Fine-Tuning Pretrained Language Models:
Weight Initializations, Data Orders, and Early Stopping

Jesse Dodge 1 2  Gabriel Ilharco 3  Roy Schwartz 2 3  Ali Farhadi 2 3 4  Hannaneh Hajishirzi 2 3  Noah Smith 2 3

REVISITING FEW-SAMPLE BERT FINE-TUNING

Tianyi Zhang*△§  Felix Wu†  Arzoo Katiyar△  Kilian Q. Weinberger‡  Yoav Artzi‡
†ASAPP Inc.  △Stanford University  §Penn State University  †Cornell University
tz58@stanford.edu  {fwu, kweinberger, yoav}@asapp.com arzoo@psu.edu
The Natural Language Decathlon: Multitask Learning as Question Answering

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, Richard Socher
Salesforce Research
{bmccann,nkeskar,cxiong,rsocher}@salesforce.com

Examples

<table>
<thead>
<tr>
<th>Question</th>
<th>Context</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is a major importance of Southern California in relation to California and the US?</td>
<td>...Southern California is a major economic center for the state of California and the US....</td>
<td>major economic center</td>
</tr>
<tr>
<td>What is the translation from English to German?</td>
<td>Most of the planet is ocean water.</td>
<td>Der Großteil der Erde ist Meerwasser</td>
</tr>
<tr>
<td>What is the summary?</td>
<td>Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...</td>
<td>Harry Potter star Daniel Radcliffe gets £320M fortune...</td>
</tr>
<tr>
<td>Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?</td>
<td>Premise: Conceptually cream skimming has two basic dimensions – product and geography.</td>
<td>Entailment</td>
</tr>
<tr>
<td>Is this sentence positive or negative?</td>
<td>A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.</td>
<td>positive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Context</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What has something experienced?</td>
<td>Areas of the Baltic that have experienced eutrophication.</td>
<td>eutrophication</td>
</tr>
<tr>
<td>Who is the illustrator of Cycle of the Werewolf?</td>
<td>Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.</td>
<td>Bernie Wrightson</td>
</tr>
<tr>
<td>What is the change in dialogue state?</td>
<td>Are there any Eritrean restaurants in town?</td>
<td>food: Eritrean</td>
</tr>
<tr>
<td>What is the translation from English to SQL?</td>
<td>The table has column names... Tell me what the notes are for South Australia</td>
<td>SELECT notes from table WHERE 'Current Slogan' = 'South Australia'</td>
</tr>
<tr>
<td>Who had given help? Susan or Joan?</td>
<td>Joan made sure to thank Susan for all the help she had given.</td>
<td>Susan</td>
</tr>
</tbody>
</table>
**Zero-shot**

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

1. Translate English to French:  
   task description

2. cheese =>  
   prompt

**One-shot**

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

1. Translate English to French:  
   task description

2. sea otter => loutre de mer  
   example

3. cheese =>  
   prompt

**Fine-tuning**

The model is trained via repeated gradient updates using a large corpus of example tasks.

1. sea otter => loutre de mer  
   example #1

   gradient update

1. peppermint => menthe poivrée  
   example #2

   gradient update

   ...

1. plush giraffe => girafe peluche  
   example #N

   gradient update

1. cheese =>  
   prompt
Learning via SGD during unsupervised pre-training

**inner loop**

sequence #1
1. $5 + 8 = 13$
2. $7 + 2 = 9$
3. $1 + 0 = 1$
4. $3 + 4 = 7$
5. $5 + 9 = 14$
6. $9 + 8 = 17$

**In-context learning**

sequence #2
1. goat $\Rightarrow$ goat
2. sakne $\Rightarrow$ snake
3. brid $\Rightarrow$ bird
4. fish $\Rightarrow$ fish
5. duck $\Rightarrow$ duck
6. chim $\Rightarrow$ chimp

sequence #3
1. thanks $\Rightarrow$ merci
2. hello $\Rightarrow$ bonjour
3. mint $\Rightarrow$ menthe
4. wall $\Rightarrow$ mur
5. otter $\Rightarrow$ loutre
6. bread $\Rightarrow$ pain

**outer loop**
1. In what year was the first-ever Wimbledon Championship held? **Answer: 1877.**

2. Hg is the chemical symbol of which element? **Answer: Mercury.**

3. Which email service is owned by Microsoft? **Answer: Hotmail.**

4. Which country produces the most coffee in the world? **Answer: Brazil.**

5. In which city was Jim Morrison buried? **Answer: Paris.**

6. Which song by Luis Fonsi and Daddy Yankee has the most views (of all time) on YouTube? **Answer: “Despacito.”**

7. What was the first state? **Answer: Delaware.**

8. What is the capital city of Spain? **Answer: Madrid.**

9. What is the painting “La Gioconda” more usually known as? **Answer: The Mona Lisa.**

from [https://www.scarymommy.com/best-trivia-questions-answers/](https://www.scarymommy.com/best-trivia-questions-answers/)
A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks

Nikunj Saunshi, Sadhika Malladi & Sanjeev Arora
Princeton University
{nsaunshi, smalladi, arora}@cs.princeton.edu

Theorem 4.1. Let \( \{p_s\} \) be a language model that is \( \epsilon \)-optimal, i.e. \( \ell_{xent}(\{p_s\}) - \ell_{xent}^* \leq \epsilon \), for some \( \epsilon > 0 \). For a classification task \( T \) that is \( (\tau, B) \)-natural, we have

\[
\ell_T (\{p_s\}) \leq \tau + \sqrt{2B^2\epsilon (\gamma(p_T))^{-1}}
\]
UNIFIEDQA: Crossing Format Boundaries with a Single QA System

Daniel Khashabi\textsuperscript{1}  Sewon Min\textsuperscript{2}  Tushar Khot\textsuperscript{1}  Ashish Sabharwal\textsuperscript{1}  Oyvind Tafjord\textsuperscript{1}  Peter Clark\textsuperscript{1}  Hannaneh Hajishirzi\textsuperscript{1,2}

\textsuperscript{1}Allen Institute for AI, Seattle, U.S.A.  \textsuperscript{2}University of Washington, Seattle, U.S.A.

**Extractive [SQuAD]**

**Question:** At what speed did the turbine operate?

**Context:** (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...

**Gold answer:** 16,000 rpm

**Multiple-Choice [ARC-challenge]**

**Question:** What does photosynthesis produce that helps plants grow?

**Candidate Answers:** (A) water (B) oxygen (C) protein (D) sugar

**Gold answer:** sugar

**Abstractive [NarrativeQA]**

**Question:** What does a drink from narcissus's spring cause the drinker to do?

**Context:** Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to "Grow dotingly enamored of themselves." ...

**Gold answer:** fall in love with themselves

**Yes/No [BoolQ]**

**Question:** Was America the first country to have a president?

**Context:** (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...

**Gold answer:** no
# Measuring Massive Multitask Language Understanding

<table>
<thead>
<tr>
<th>Model</th>
<th>Humanities</th>
<th>Social Science</th>
<th>STEM</th>
<th>Other</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>27.9</td>
<td>28.8</td>
<td>27.0</td>
<td>27.7</td>
<td>27.9</td>
</tr>
<tr>
<td>ALBERT</td>
<td>27.2</td>
<td>25.7</td>
<td>27.7</td>
<td>27.9</td>
<td>27.1</td>
</tr>
<tr>
<td>GPT-2</td>
<td>32.8</td>
<td>33.3</td>
<td>30.2</td>
<td>33.1</td>
<td>32.4</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>45.6</td>
<td>56.6</td>
<td>40.2</td>
<td>54.6</td>
<td>48.9</td>
</tr>
<tr>
<td>GPT-3 Small (few-shot)</td>
<td>24.4</td>
<td>30.9</td>
<td>26.0</td>
<td>24.1</td>
<td>25.9</td>
</tr>
<tr>
<td>GPT-3 Medium (few-shot)</td>
<td>26.1</td>
<td>21.6</td>
<td>25.6</td>
<td>25.5</td>
<td>24.9</td>
</tr>
<tr>
<td>GPT-3 Large (few-shot)</td>
<td>27.1</td>
<td>25.6</td>
<td>24.3</td>
<td>26.5</td>
<td>26.0</td>
</tr>
<tr>
<td>GPT-3 X-Large (few-shot)</td>
<td>40.8</td>
<td>50.4</td>
<td>36.7</td>
<td>48.8</td>
<td>43.9</td>
</tr>
</tbody>
</table>

The following are multiple choice questions about high school mathematics:

- How many numbers are in the list 25, 26, ..., 100?
  - (A) 75 (B) 76 (C) 22 (D) 23
  - Answer: B
- Compute \(i + i^2 + i^3 + \cdots + i^{258} + i^{259}\).
  - (A) -1 (B) 1 (C) \(i\) (D) -\(i\)
  - Answer: A
- If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?
  - (A) 28 (B) 21 (C) 40 (D) 30
  - Answer: C

Collin Burns, Steven Basart, Andy Zou: Columbia University, UC Chicago, UC Berkeley

Dan Hendrycks, Mantas Mazeika, Dawn Song, Jacob Steinhardt: UC Berkeley, UIUC
Muppet: Massive Multi-task Representations with Pre-Finetuning

Armen Aghajanyan  
Facebook  
amenag@fb.com

Anchit Gupta  
Facebook  
ananchit@fb.com

Akshat Shrivastava  
Facebook  
akshats@fb.com

Xilun Chen  
Facebook  
xilun@fb.com

Luke Zettlemoyer  
Facebook  
lsz@fb.com

Sonal Gupta  
Facebook  
sonalgupta@fb.com

RoBERTa Pre-Finetuning Scale Ablation
Meta-tuning Language Models to Answer Prompts Better

Ruiqi Zhong  Kristy Lee*  Zheng Zhang*  Dan Klein
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

Standard Classification Format
A total waste of time. classify 0
Great movie, must see! classify 1

Convert

Question Answering Format
[Question] Is the review positive?
[Context] A total waste of time. answer “No”
[Context] Great movie, must see! answer “Yes”

Hate Speech Detection
Question Categorization
Topic Classification
Sentiment Classification
Stance Classification

Meta-tune
Evaluate on Unseen Tasks
CROSSFIT 🎯: A Few-shot Learning Challenge for Cross-task Generalization in NLP

Qinyuan Ye  Bill Yuchen Lin  Xiang Ren
University of Southern California
{qinyuany, yuchenc.lin, xiangren}@usc.edu
Can we obtain a model that can readily perform tons of NLP tasks at human-level performance without further fine-tuning?

Maybe we should try training a giant model in a massively multi-task setup with an extremely diverse set of task formats.

https://bigscience.huggingface.co/