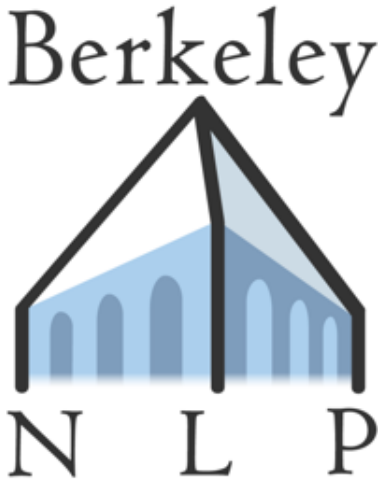
Adapting Language Model for Zero-Shot Learning by Meta-Tuning on Dataset and Prompt Collections



Ruiqi Zhong, Kristy Lee*, Zheng Zhang*, and Dan Klein
{ruiqi-zhong, kristylee, zhengzhang1216, klein}@berkeley.edu

What's Zero-Shot Learning?

What's Zero-Shot Learning?

Traditional Setup

Training

I love this movie → **1**
The action part is great → **1**
Total waste of time → **0**
Really boring → **0**

Testing

Highly recommended → **1**

What's Zero-Shot Learning?

	Traditional Setup	Zero-Shot Setup
Training	*I love this movie* → 1 *The action part is great* → 1 *Total waste of time* → 0 *Really boring* → 0	**Any training task except sentiment classification.**
Testing	*Highly recommended* → 1	*Highly recommended* ***Is this a positive review?*** → yes

What's Zero-Shot Learning?

[Input] [Task Description] → [Answer]

Based on **task description**, perform a new task that was unseen at training time

Language Model Prompting

Language Model Prompting

Language Model
Training

Berkeley NLP is a group of ____ → *faculty and graduate*
my own happiness was ____ → *entirely identified with this object*
behind three concentric ____ → *walls and enjoyed nearly …*
rethinking both the theoretical and ____ → *empirical paradigms …*
…

Language Model Prompting

Language Model
Training

Berkeley NLP is a group of ____ → *faculty and graduate*
my own happiness was ____ → *entirely identified with this object*
behind three concentric ____ → *walls and enjoyed nearly …*
rethinking both the theoretical and ____ → *empirical paradigms …*
…

Zero-Shot
Testing

Highly recommended. ***Is this a positive review?*** *____* → *yes*

Language Model Prompting

Language Model
Training

Berkeley NLP is a group of ____ → *faculty and graduate*
my own happiness was ____ → *entirely identified with this object*
behind three concentric ____ → *walls and enjoyed nearly …*
rethinking both the theoretical and ____ → *empirical paradigms …*
…

Bridge the **Misalignment** with Clever “Prompting"

Zero-Shot
Testing

*Highly recommended. **Is this a positive review?** ____* → *yes*

Machine Learning 101

"Directly optimize the target objective
(e.g. **zero-shot** learning)"

Meta-Tuning

Zero-Shot
Training

?

Zero-Shot
Testing

Highly recommended. ***Is this a positive review?*** ____ → ***yes***

Meta-Tuning

Zero-Shot
Training

Learn to perform a task based on task descriptions;

Zero-Shot
Testing

Highly recommended. ***Is this a positive review?*** ____ → ***yes***

Meta-Tuning

Zero-Shot
Training

The market is fluctuating. ***Is this sports news?*** ____ → ***no***
Olympics started yesterday. ***Is this sports news?*** ____ → ***yes***
Utilitarian means XXX. ***Does it contain a definition?*** ____ → ***yes***
I were hungry. ***Is it grammatical?*** ____ → ***no***

Manually unified datasets from 43 different sources.
~200 unique labels and hand-wrote 440 **task descriptions**.

Zero-Shot
Testing

Highly recommended. ***Is this a positive review?*** ____ → ***yes***

Abstract View

Abstract View

Fine-tune on
Task A, B, C

[Input A1] [Task Description A] → [Answer A1]
[Input A2] [Task Description A] → [Answer A2]
[Input B1] [Task Description B] → [Answer B1]
[Input B2] [Task Description B] → [Answer B2]
[Input C1] [Task Description C] → [Answer C1]

. . .

Test on
Task D, E

[Input D1] [Task Description D] → [Answer D1]
[Input D2] [Task Description D] → [Answer D2]
[Input E1] [Task Description E] → [Answer E1]

...

Abstract View

Fine-tune on
Task A, B, C

[Input A1] [Task Description A] → [Answer A1]
[Input A2] [Task Description A] → [Answer A2]
[Input B1] [Task Description B] → [Answer B1]
[Input B2] [Task Description B] → [Answer B2]
[Input C1] [Task Description C] → [Answer C1]

. . .

Different enough to be
counted as “zero-shot”

Test on
Task D, E

[Input D1] [Task Description D] → [Answer D1]
[Input D2] [Task Description D] → [Answer D2]
[Input E1] [Task Description E] → [Answer E1]

...

Cross Validation Split

Movie Review Classification
Hotel Review Classification
Airline Review Classification

Review | Good vs. Bad

Stance Classification
Liberal/Conservative Classification

Social Media | Societal

Question Paraphrase Detection
Answer Type Classification

Question Categorization

Hate Speech Detection
Offensive Speech Detection

Social Media | Societal | Emotion

Aligned Objective —> Better

Cross Validation Average AUC ROC

T5 Large **Meta-tuned**	**73.8%**
T5 Large **UnifiedQA**	**70.5%**
T5 Medium Meta-tuned	67.5%

Aligned Objective —> Better

Cross Validation Average AUC ROC

T5 Large **Meta-tuned**	**73.8%**
T5 Large **UnifiedQA**	**70.5%**
T5 Medium Meta-tuned	67.5%

ICML 2021 Zhao et. al
GPT-3 (**175B parameter**)
after calibration
80% accuracy on SST-2

Aligned Objective —> Better

Cross Validation Average AUC ROC

T5 Large **Meta-tuned**	**73.8%**
T5 Large **UnifiedQA**	**70.5%**
T5 Medium Meta-tuned	67.5%

ICML 2021 Zhao et. al
GPT-3 (**175B parameter**)
after calibration
80% accuracy on SST-2

Ours
RoBERTa-Large (**500x smaller**)
without calibration
88% accuracy on SST-2

Larger —> Better

Cross Validation
Average AUC ROC

T5 **Large** Meta-tuned	**73.8%**
T5 Large UnifiedQA	70.5%
T5 **Medium** Meta-tuned	**67.5%**

ICML 2021 Zhao et. al
GPT-3 (175B parameter)
after calibration
80% accuracy on SST-2

Ours
RoBERTa-Large (500x smaller)
without calibration
88% accuracy on SST-2

Scaling: Aligned & Larger -> Better

Scaling: Aligned & Larger -> Better

- We only experimented with small models (700M)

Scaling: Aligned & Larger -> Better

- We only experimented with small models (700M)
- FLAN (Google, 137B) outperforms GPT-3 (173B)

Scaling: Aligned & Larger -> Better

- We only experimented with small models (700M)
- FLAN (Google, 137B) outperforms GPT-3 (173B)
- T0 (13B) outperforms GPT-3 (173B)

Thank you!

Paper: https://arxiv.org/abs/2104.04670

Code: https://github.com/ruiqi-zhong/Meta-tuning