Commonsense Reasoning in the Wild

Xiang Ren

Joint work w/ Bill Lin, Pei Zhou, Qinyuan Ye, Jay Pujara
Wangchunshu Zhou, Yejin Choi, Chandra Bhagavatula

Department of Computer Science & Information Science Institute
University of Southern California
http://inklab.usc.edu
Language is often ambiguous / underspecified

Hey, let’s hoop at 10. Same park.

Q: Here, what does “10” mean?

Caption: Her voice is amazing!
Q: Who does “her” refer to?
Making proper presumptions is important!

Hey, let’s hoop at 10. Same park.

Q: Here, what does “10” mean?

- When meeting, people usually specify place and time
- Time can be referred by numbers

A: 10 refers to time of day.
(Still not clear if it is 10AM or 10PM!)

Caption: Her voice is amazing!

Q: Who does “her” refer to?

- A person holding a microphone would have more prominent voice
- A person standing on a stage in front of an audience is likely singing/speaking

A: “Her” refers to the girl in red dress.
Common sense knowledge are shared across tasks

A person feels happy and excited after getting a pet

Dialogue Response Generation

I am getting a dog today
Congrats! You must be super excited!

Q: How is the boy feeling right now?
A: stressed, happy, sad, confused

VQA
LM pretraining is not the answer

Lin et al., 2020

A bird usually has [MASK] legs. 1st: four (44.8%) 2nd: two (18.7%)
A car usually has [MASK] wheels. 1st: four (53.7%) 2nd: two (20.5%)
A car usually has [MASK] round wheels. 1st: two (37.1%) 2nd: four (20.2%)

Agrawal et al., 2016

Q: What color are the safety cones?
GT A: green
Predicted A: orange

Premise: The judge by the actor stopped the banker.
Hypothesis: The banker stopped the actor.
Answer: Entailment ✗

Most cones were orange in training set

McCoy et al., 2019

Lexical overlaps usually indicate entailment in training data
LM pretraining is not the answer

Reporting bias of commonsense knowledge $\leftrightarrow$ pretraining of massive language models

Premise: The judge by the actor stopped the banker.
Hypothesis: The banker stopped the actor.
Answer: Entailment $\times$

McCoy et al., 2019

Lexical overlaps usually indicate entailment in training data

Most cones were orange in training set
CSR Models on Research Benchmarks

I’m looking for a cheap hotel in Los Angeles

Ok, what date do you prefer?
Hey, I’m going skydiving tomorrow. It’s my first time!

Sorry I don’t know what that means.
Performs well on a benchmark

Collect dataset → i.i.d. → Training → Evaluation

97% Acc.
A person performing in front of people might be nervous
People performing in front of people find it harder to be relaxed
It can be hard for someone to be calm when they’re about to perform

Linguistically-varied statements of the same inference rule

RICA (Zhou et al., 2021)

Performs well on a benchmark
• Model learns dataset shortcuts

Performs well in the wild
• Robust to linguistic variations

INV: Swap one character with its neighbor (typo)

DIR: Paraphrase of question should be duplicate

Behavioral Testing 🌍

Ribeiro et al., 2020
Perform well on a benchmark
- Model learns dataset shortcuts
- Struggles with underspecified/adversarial inputs

Perform well in the wild
- Robust to linguistic variations
- Resolves ambiguity/noise with presumptions

When is the Super Bowl?

Do you mean When is the Super Bowl 2022?

Super Bowl 2022 will be at 3:30 PM on February 13.

Underspecified Inputs

Adversarial Inputs

Levinson, 2000

Jia and Liang, 2017
Wallace et al., 2019
Performs well **on a benchmark**
- Model learns dataset shortcuts
- Struggles with underspecified/adversarial inputs
  - Customized to a narrow task

Performs well **in the wild**
- Robust to linguistic variations
- Resolves ambiguity/noise with presumptions
Performs well on a benchmark
• Model learns dataset shortcuts
• Struggles with underspecified/adversarial inputs
• Customized to a narrow task

Performs well in the wild
• Robust to linguistic variations
• Resolves ambiguity/noise with presumptions
• Generalizable across a wide range of tasks

applicable to a wide range of tasks

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>🏀</td>
<td>🏀</td>
</tr>
<tr>
<td>🎧</td>
<td>🎧</td>
</tr>
<tr>
<td>🎳</td>
<td>🎳</td>
</tr>
<tr>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>🟡</td>
<td>🟡</td>
</tr>
</tbody>
</table>

generalizes well to new tasks

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>🏁</td>
<td>🏁</td>
</tr>
<tr>
<td>🍼</td>
<td>🍼</td>
</tr>
<tr>
<td>🏔️</td>
<td>🏔️</td>
</tr>
<tr>
<td>🍺</td>
<td>🍺</td>
</tr>
<tr>
<td>🍻</td>
<td>🍻</td>
</tr>
</tbody>
</table>

DecaNLP ([McCann et al., 2018](https://www.nature.com/articles/s41591-018-0202-x))
T5 ([Raffel et al., 2019](https://arxiv.org/abs/1910.10683))
ExT5 ([Aribandi et al., 2021](https://www.ijcai.org/Proceedings/2021/0988))
Muppet ([Aghajanyan et al., 2021](https://www.ijcai.org/Proceedings/2021/0988))

CrossFit ([Ye et al., 2021](https://arxiv.org/abs/2110.15150))
Natural Instructions ([Mishra et al., 2021](https://arxiv.org/abs/2103.01419))
FLEX ([Bragg et al., 2021](https://arxiv.org/abs/2111.13813))
FLAN ([Wei et al., 2021](https://arxiv.org/abs/2111.13813))
T0 ([Sanh et al., 2021](https://arxiv.org/abs/2111.13813))
This talk - New ways of formulating CSR challenges

**Discriminative (closed-ended) reasoning**

Alex spilled the food she just prepared all over the floor and it made a huge mess.

**Q** What will Alex want to do next?  
**A** (a) taste the food  
(b) mop up ✓  
(c) run around in the mess

Social IQA (Sap et al. 2019)

Towards more **open-ended** reasoning

A boy throws a frisbee and a dog catches it in the air.

(Lin et al., Findings of EMNLP’20)  
(Lin et al., NAACL’21)  
(Wang et al., ICLR’22)
This talk - New ways of formulating CSR challenges

Reasoning in a logically robust/consistent manner

Apples and oranges grow on trees
Oranges and apples grow on trees
Fruits grow on trees
Apples and oranges grow on plants

Trees grow on apples and oranges
Apples and trees grow on oranges

(Zhou et al., EMNLP’21)
This talk - New ways of formulating CSR challenges

Study the cross-task generalization ability of NLP models

Directly Learning Calculus...

Machine: ? ? ?
Baby: ? ? ?

Learning Math, History, Geography, Chemistry, ... in high school

Learning Calculus in undergrad
Student: Yes I can do it!

(Ye et al., EMNLP’21)
An intelligent behavior possessed by humans that demonstrates common sense

{dog, frisbee, catch, throw}

A boy throws a frisbee and a dog catches it in the air.
Can machines learn to describe a daily scene using concepts?

{dog, frisbee, catch, throw}
Generative Commonsense Reasoning

**Input:** A set of concept words (objects / actions)

{dog, frisbee, catch, throw}

**Output:** A sentence describing everyday scenes using all the concepts.

A dog catches a frisbee when a boy throws it.

Humans

Machines

GPT2: A dog throws a frisbee at a football player.

T5: Dog catches a frisbee and throws it at a dog.

(CommonGen, Findings of EMNLP 2020)
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>BLEU-4</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KFCNet</td>
<td>43.819</td>
<td>18.845</td>
<td>33.911</td>
</tr>
<tr>
<td>(Jun 08, 2021)</td>
<td>MSRA and Microsoft Ads Email Paper (EMNLP'21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>KGVAE</td>
<td>42.818</td>
<td>18.423</td>
<td>33.564</td>
</tr>
<tr>
<td>(May 18, 2021)</td>
<td>Alibaba and Xiamen University Email Paper (AAAI 2022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>KFC (v1)</td>
<td>42.453</td>
<td>18.379</td>
<td>33.277</td>
</tr>
<tr>
<td>(Mar 23, 2021)</td>
<td>MSRA and Microsoft Ads Email Paper (EMNLP'21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>R3-BART</td>
<td>41.954</td>
<td>17.706</td>
<td>32.961</td>
</tr>
<tr>
<td>(April 26, 2021)</td>
<td>Anonymous (under review) Email Document (placeholder)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>WittGEN + T5-large</td>
<td>38.233</td>
<td>18.036</td>
<td>31.682</td>
</tr>
<tr>
<td>(July 1, 2021)</td>
<td>Anonymous (under review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Imagine-and-Verbalize</td>
<td>40.565</td>
<td>17.716</td>
<td>31.291</td>
</tr>
<tr>
<td>(Jan 26, 2022)</td>
<td>USC/ISI Email Paper (ICLR22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>RE-T5 (Retrieval-Enhanced T5)</td>
<td>40.863</td>
<td>17.663</td>
<td>31.079</td>
</tr>
<tr>
<td>(Jan 12, 2021)</td>
<td>Microsoft Cognitive Services Research Group Email Paper (ACL21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A* Neurologic (T5-large)</td>
<td>39.597</td>
<td>17.285</td>
<td>30.130</td>
</tr>
<tr>
<td>(Oct 19, 2021)</td>
<td>UW and AI2 Email Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>VisCTG (BART-large)</td>
<td>36.939</td>
<td>17.109</td>
<td>29.973</td>
</tr>
<tr>
<td>(Aug 1, 2021)</td>
<td>CMU-LTI Email Paper (arXiv)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CommonGen Leaderboard: [https://inklab.usc.edu/CommonGen/leaderboard.html](https://inklab.usc.edu/CommonGen/leaderboard.html)
Gehrmann et al., 2021)
Sanh et al., 2021)
Wei et al., 2021)

CommonGen Leaderboard: https://inklab.usc.edu/CommonGen/leaderboard.html
Externalizing scene imagination: Structured Knowledge Representation
Externalizing scene imagination: Structured Knowledge Representation

Scene Knowledge Graph (SKG)

<table>
<thead>
<tr>
<th>Relation types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG1</td>
<td>(play, ARG1, guitar)</td>
</tr>
<tr>
<td>ARG0</td>
<td>(play, ARG0, man)</td>
</tr>
<tr>
<td>ARG2</td>
<td>(ask, ARG2, girl)</td>
</tr>
<tr>
<td>Location</td>
<td>(play, Location, stage)</td>
</tr>
<tr>
<td>Time</td>
<td>(play, Time, sing)</td>
</tr>
<tr>
<td>Op1</td>
<td>(down, Op1, stair)</td>
</tr>
<tr>
<td>Part</td>
<td>(dog, Part, ear)</td>
</tr>
</tbody>
</table>
Externalizing scene imagination: Imagine-and-verbalize

(Wang et al., ICLR’22)
Externalizing scene imagination: **Imagine-and-verbalize**

“A man throws a frisbee and a dog catches it.”

(Wang et al., ICLR’22)
Externalizing scene imagination: Imagine-and-verbalize

(Wang et al., ICLR’22)
Externalizing scene imagination: Imagine-and-verbalize

(Wang et al., ICLR’22)
How can imagination help?

**Input:** \{hit, bottle, open, shoe, wall\}

**Output:**

without imagination

- Someone **opens** his **shoes** and hits a bottle on a wall.

with imagination

- A man opens a bottle and hits his shoes against a wall.
Results on CommonGen (leaderboard)

- SOTA (KFCNet) uses a much larger corpus (>700M)
- Imagination > Prototype-based (Except KFCNet)
  - VisCTG (Image)
  - KG-BART
  - Node2Text
How do we reply in conversations?

I’m going to sing in front of hundreds tomorrow...

Performing in front of audience can cause anxiety

Deep breaths, you’ll do great!
Grounding in Communications

Effective communications require reaching mutual beliefs and knowledge among participants (called *grounding*).

**Common Sense** plays a critical role in grounding in communications.

How do we reply in conversations?

I’m going to **perform in a piano recital** tomorrow...

Performing in front of audience can cause **anxiety**

Recalling & Combining common sense with information expressed in NL to **make inferences**

Deep breaths, you’ll do great!
How do we reply in conversations?

I’m going to perform in a piano recital tomorrow...

Performing in front of audience can cause anxiety.

Deep breaths, you’ll do great!

Recalling & Combining common sense with information expressed in NL to make inferences.

Producing consistent inferences amidst logically-equivalent yet linguistically-varied paraphrases.
RICA: Robust Inference on Commonsense Axioms

- Sets of natural language statements in the \textit{“premise-conclusion”} format that express the same commonsense axiom but linguistically varied

- Examples:
  - \textbf{Original}: “A is heavier than B, so A is \texttt{<better>} at sinking than B.”
  - \textbf{Negation}: “A is heavier than B, so A is \texttt{not <worse>} at sinking than B.”
  - \textbf{Entity Swap}: “B is heavier than A, so A is \texttt{<worse>} at sinking than B.”
  - \textbf{Antonym}: “A is heavier than B, so A is \texttt{<worse>} at \texttt{floating} than B.”
  - …

\textbf{Recalling & Combining common sense} with information expressed in NL to \textit{make inferences}

(Zhou et al., EMNLP’21)
RICA: Robust Inference on Commonsense Axioms

- Probe model’s *robustness against linguistic variations* (of the same commonsense axiom)

- Masked word prediction task: Choose `<better>` or `<worse>`:
  - **Original**: “A is heavier than B, so A is `<MASK>` at sinking than B.”
  - **Perturb1**: “A is heavier than B, so A is *not* `<MASK>` at sinking than B.”
  - **Perturb2**: “B is heavier than A, so A is `<MASK>` at sinking than B.”
  - **Perturb3**: “A is heavier than B, so A is `<MASK>` at *floating* than B.”
  - …

Producing *consistent* inferences amidst *logically-equivalent yet linguistically-varied* paraphrases

(Zhou et al., EMNLP’21)
RICA: Overview of the probe construction

Define logical primitives

Mine common sense

Represent commonsense in logical form

Create commonsense statements that can be used to probe language models

Perturb and convert logical form to text

(Zhou et al., EMNLP’21)
Define logical primitives

1. Base Predicates
   - Property(A,p)
   - Relation(A,B,r)
   - Comparator(x,y)

2. Logical Template
   - Rel(A,B,r) →
   - Comp(Prop(A,p), Prop(B,p))

- Define three basic first-order logic predicates
- Connect predicates to form abstract logical templates
  - A is B’s <r>, so A is more/less <p> than B
Goal: Fill the abstract templates with concrete common sense

A is B’s <r>, so A is more/less <p> than B
  - <r> → “lawyer”
  - <p> → “knowledge of law”

Crawl from knowledge bases
  - Step 1: Get a list of occupations
  - Step 2: Query ConceptNet for triples, such as <Occupation, HasProperty, p>
Probe construction III

- Fill logical templates with crawled common sense

- Apply perturbation operators and convert to text

4. Created Axiom
\[
\text{Rel}(A, B, \text{lawyer}) \Rightarrow \\
\text{Comp(Prop}(A, \text{knowledge of law}), \text{Prop}(B, \text{knowledge of law}))
\]

Represent commonsense in logical form

Perturb and convert logical form to text

5. Commonsense Statement Set
A is B’s lawyer, so A is more knowledgeable about law than B
B is A’s lawyer, so A is not more knowledgeable about law than B
A is B’s lawyer, so A is less clueless about law than B
A is B’s lawyer, so B is less informed on the law than A

Replace A and B with Novel Entities: A \rightarrow\text{prindag} B \rightarrow\text{fluberg}
Probe construction III

Goal: create perturbed forms that preserve the commonsense axiom

- **Linguistic Operators:**
  - Negation: “knowledgeable” → “not knowledgeable”
  - Antonym: “knowledgeable” → “clueless”
  - Paraphrase: “knowledgeable” → “informed”

- **Composition:**
  - negation + paraphrase → “not informed”
  - ...

- **Asymmetry Operators:** “A is B’s lawyer” → “B is A’s lawyer”

- 24 types in total
Experiments

Masked Word Prediction (MWP)
1. BERT / RoBERTa
2. ERNIE (KG-enhanced LM)
3. BART (Seq2seq)

Testing Set: **1.6k human-curated**

Evaluation Settings:
1. **Zero-Shot**: without fine-tuning
2. **Low-Resource**: fine-tune on **1k** of all verified probes
3. **High-Resource**: fine-tune on all verified probes (**9k**)
4. Large-Scale on Raw Data: **100k** from the machine generated set

Metric: Average accuracy
Results: Human-Curated Set

- Random-guessing like performance on all settings for all models.

- Training on similar data does **not** help achieve real robustness.

<table>
<thead>
<tr>
<th>Setting</th>
<th>BERT etc.</th>
<th>Average Accuracy on Human-Curated Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>91.7%</td>
<td>~50%</td>
</tr>
<tr>
<td>Zero-Shot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low-Resc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Resc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis: Positivity Bias

- Heavy bias towards positive-valence words such as “more”, “better”, “easier”.

- Fine-tuning on RICA mitigates the imbalance issue (but still fails)

<table>
<thead>
<tr>
<th></th>
<th>Average Accuracy without Fine-Tuning</th>
<th>Average Accuracy after Fine-tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human (both positive and negative)</strong></td>
<td>91.7%</td>
<td>~50%</td>
</tr>
<tr>
<td>Pos. Words</td>
<td>BERT etc. 87.2%</td>
<td>BERT etc.</td>
</tr>
<tr>
<td>Neg. Words</td>
<td>BERT etc. 12.5%</td>
<td>BERT etc.</td>
</tr>
</tbody>
</table>
Analysis: Robustness Issue

- Severe variation among different linguistic perturbation operators
Combining common sense with information expressed in NL to make inferences

Producing consistent inferences amidst logically-equivalent yet linguistically-varied paraphrases.

https://sites.google.com/usc.edu/rica
Cross-task generalization in NLP

Learning at the instance-level

Generalize from a few seen training instances, to multiple unseen test instances.

**Train**
- This movie is extraordinary. Positive
- Watching it is a waste of time. Negative

**Test**
- It’s such a wonderful movie! ?
- I’m so disappointed! ?

*Task: Movie Review Sentiment Classification*
Cross-task generalization in NLP

Learning at the **instance-level**

Generalize from a few *seen training instances*, to multiple *unseen test instances*.

**Train**
- This movie is extraordinary. **Positive**
- Watching it is a waste of time. **Negative**

**Test**
- It’s such a wonderful movie! **?**
- I’m so disappointed! **?**

**Task: Movie Review Sentiment Classification**

Learning at the **task-level**

Generalize from a few *seen training tasks*, to multiple *unseen test tasks*.

**Train**
- Movie Review Sentiment Classification
- Reading Comprehension on News
- Biomedical Relation Extraction

**Test**
- Paraphrase Identification
- Commonsense Multiple-choice QA
- Mining Alpha Factors from News Corpora

**Goal:** Achieve competitive performance on the test task with fewer annotations.
CrossFit 🏋️: A Few-shot Learning Challenge for Cross-task Generalization

- Humans can learn a new task efficiently with only few examples, by leveraging their knowledge obtained when learning prior tasks.

- We refer to this ability as cross-task generalization.

- How such ability can be acquired, and further applied to build better few-shot learners across diverse NLP tasks.

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

- Gather 160 diverse few-shot tasks in text-to-text format

(Datasets)

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

NLP Few-shot Gym

• Gather 160 diverse few-shot tasks in text-to-text format
• Manually group the tasks into categories and sub-categories.

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

NLP Few-shot Gym

- Gather 160 diverse few-shot tasks in text-to-text format
- Manually group the tasks into categories and sub-categories.
- Design 8 partitions of the tasks to test cross-task generalization in different scenarios

| Partition 1: Random | | Partition 2.1: 45non-class | |
|---------------------|----------------|---------------------------|
| Randomly split 160 tasks into 120/20/20 for train/dev/test tasks. | | Train: 45 non-classification tasks Dev/Test: 10 classification tasks |

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

**NLP Few-shot Gym 🌊**

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios

**CrossFit Setting**

Large-scale Pre-training

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

NLP Few-shot Gym

• Gather 160 diverse few-shot tasks in text-to-text format
• Manually group the tasks into categories and sub-categories.
• Design 8 partitions of the tasks to test cross-task generalization in different scenarios

CrossFit Setting

Large-scale Pre-training
+ Upstream Learning on a set of seen tasks ($T_{train}$)

Using multi-task learning and meta-learning methods (e.g., MAML, Reptile)

(Ye et al., EMNLP 2021)
CrossFit: Quick Summary

**NLP Few-shot Gym 💪**

- Gather **160 diverse few-shot tasks** in text-to-text format
- Manually **group the tasks** into categories and sub-categories.
- Design **8 partitions** of the tasks to test cross-task generalization in different scenarios

**CrossFit 🏋️ Setting**

**Large-scale Pre-training**

+ Upstream Learning on a set of seen tasks ($T_{train}$)
+ Downstream Fine-tuning on an unseen target task ($T_{test}$)

(Ye et al., EMNLP 2021)
Evaluation Metric

- We define **Average Relative Gain** (ARG), to measure the overall performance gain on all unseen tasks.

- ARG is the relative performance changes before and after the upstream learning stage for each test task, and averaged across all test tasks.

- *This is not a perfect metric*, but it helps us to get a general sense. We still plot and report relative gain for individual tasks.

**Example**

<table>
<thead>
<tr>
<th>Task</th>
<th>Direct FT</th>
<th>Upstream + FT</th>
<th>Rel. Gain</th>
<th>ARG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A</td>
<td>50% F1</td>
<td>70% F1</td>
<td>40%</td>
<td>7.5%</td>
</tr>
<tr>
<td>Task B</td>
<td>40% Acc.</td>
<td>30% Acc.</td>
<td>-25%</td>
<td></td>
</tr>
</tbody>
</table>

\[(40\% - 25\%) / 2 = 7.5\%\]
Experiments

- We mainly use BART-Base (Lewis et al., 2020) as the main model for our analysis.
  - Also we verify some of our findings with BART-Large and T5-v1.1-Base (Raffel et al., 2019)

- Methods for comparison
  - **Downstream Fine-tuning** (also used as the baseline for computing ARG)

![Diagram showing experiments setup]

For each task in $T_{test}$
- Fine-tune on $D_{train}$
- Validate on $D_{dev}$
- Report performance on $D_{test}$

Test Tasks $T_{test}$
- Task 1
- Task 2
- Task 3
Experiments

- We mainly use BART-Base (Lewis et al., 2020) as the main model for our analysis.
  - Also we verify some of our findings with BART-Large and T5-v1.1-Base (Raffel et al., 2019)

- Methods for comparison
  - Downstream Fine-tuning
  - Upstream Learning then Downstream Fine-tuning
    - Multi-task Learning

Train Tasks $T_{train}$

- Task 1 $D_{train}$ $D_{dev}$
- Task 2 $D_{train}$ $D_{dev}$
- Task 3 $D_{train}$ $D_{dev}$
- Task 4 $D_{train}$ $D_{dev}$
- Task 5 $D_{train}$ $D_{dev}$
- Task 6 $D_{train}$ $D_{dev}$

Concat into a big training set

$M_0$ BART-Base

Upstream Learning

$M_1$ $D_{train}$

Fine-tune

$M$
Experiments

- We mainly use **BART-Base** (Lewis et al., 2020) as the main model for our analysis.
  - Also we verify some of our findings with **BART-Large** and **T5-v1.1-Base** (Raffel et al., 2019)

- Methods for comparison
  - **Downstream Fine-tuning**
  - **Upstream Learning** then **Downstream Fine-tuning**
    - Multi-task Learning
    - Model Agnostic Meta-learning (Finn et al., 2017)

Variants of MAML
- First-order MAML
- Reptile (Nichol et al., 2017)
**Question 1**
Is upstream learning using seen tasks helpful?

**Method**
We applied multi-task learning and meta-learning algorithms during upstream learning.

**Findings**
Yes! Upstream learning methods do help LMs to acquire cross-task generalization.

The conclusion holds on different splits of seen/unseen tasks, and with different upstream learning methods.

**Evidence 1**
ARG (defined earlier) is **positive** for all 8 partitions and all 4 upstream learning methods

<table>
<thead>
<tr>
<th>No.</th>
<th>Shorthand</th>
<th>ARG(Multi)</th>
<th>ARG(MAML)</th>
<th>ARG(FoMAML)</th>
<th>ARG(Rept.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Random</td>
<td>35.06%</td>
<td>28.50%</td>
<td>22.69%</td>
<td>25.90%</td>
</tr>
<tr>
<td>2.1</td>
<td>45cls</td>
<td>11.68%</td>
<td>9.37%</td>
<td>10.28%</td>
<td>13.36%</td>
</tr>
<tr>
<td>2.2</td>
<td>23cls+22non-cl</td>
<td>11.82%</td>
<td>9.69%</td>
<td>13.75%</td>
<td>14.34%</td>
</tr>
<tr>
<td>2.3</td>
<td>45non-cl</td>
<td>11.91%</td>
<td>9.33%</td>
<td>11.20%</td>
<td>14.14%</td>
</tr>
<tr>
<td>3.1</td>
<td>Held-out-NLI</td>
<td>16.94%</td>
<td>12.30%</td>
<td>12.33%</td>
<td>14.46%</td>
</tr>
<tr>
<td>3.2</td>
<td>Held-out-Para</td>
<td>18.21%</td>
<td>17.90%</td>
<td>21.57%</td>
<td>19.72%</td>
</tr>
<tr>
<td>4.1</td>
<td>Held-out-MRC</td>
<td>32.81%</td>
<td>27.28%</td>
<td>28.85%</td>
<td>28.85%</td>
</tr>
<tr>
<td>4.2</td>
<td>Held-out-MCQA</td>
<td>12.20%</td>
<td>4.69%</td>
<td>6.73%</td>
<td>7.67%</td>
</tr>
</tbody>
</table>

**Evidence 2**
When we aggregate test performance gain from all upstream learning methods and partitions...

- **>5% relative gain** 51.47%
- **within ±5%** 35.93%
- **<-5% relative gain** 12.60%
Question 2
How does the selection of seen tasks influence the performance?

Method – Controlled Experiments
Seen tasks: (1) 100% classification
(2) 50% class + 50% non-class
(3) 100% non-classification
Unseen tasks: 100% classification

Findings
Classification tasks and non-classification tasks seem to be equivalently helpful.
Our understanding of tasks may not align with how models learn transferable skills!
Question 3
Does the improved cross-task generalization ability go beyond few-shot settings?

Method
Increase the amount of training data for downstream/unseen tasks (32, 64, \(\rightarrow\) 4.1k, 8.7k)

Findings
Cross-task generalization helps most on CommonsenseQA, ROPES and MNLI.

On these three datasets, the benefits brought by upstream learning methods extend into medium resource cases with up to 2048 training examples.
We found that …

- **Upstream learning methods** such as multi-task learning and meta-learning help pre-trained LMs to **acquired cross-task generalization**.
- Task similarity in terms of task format **does not** align with how models learn transferable skills.

We envision the **CrossFit 🏋️** Challenge and the **NLP Few-shot Gym💦** to serve as the testbed for many interesting “**meta-problems**”

- Generating Prompts? ([Shin et al., 2020; Gao et al., 2020](#))
- Select appropriate upstream tasks? ([Zamir et al., 2018; Standley et al., 2020; Vu et al., 2020](#))
- Apply task augmentation? ([Murty et al., 2021](#))
- Continual Learning? ([Jin et al., 2021](#))
- Task decomposition? ([Andreas et al., 2016; Khot et al., 2021](#))
Massive Multi-tasking
Ye et al., 2021

Neuro-Symbolic Reasoning
Lin et al., 2019; Wang et al., 2022

Commonsense Reasoning

Explainability & Interpretability
Jin et al., 2020; Kennedy et al., 2020

Instructions & Interactions
Ye et al., 2020; Yao et al., 2021

Trustworthy AI

Fluid Human-machine Communication
Symbolic knowledge helps create trustworthy NLP models

Prior work:
1. + Path for CSQA (EMNLP’20 Findings)
2. + Triplets for KG completion (ACL’21 Findings)
3. + Graph for GCSR (ICLR’22)

Next step:

Adding knowledge

Symbolic knowledge as the backbone of model explanation

“Why does a model make a particular decision?”

knowledge for refining model for continual learning

“Can we debug a model?”
How should we use commonsense reasoning to achieve better cross-task generalization?

Diverse Commonsense Reasoning Tasks

- CommonGen
- WinoGrande
- PIQA
- CSQA

Input: Given the options below, select the most suitable answer for the following question: What place is not interesting to children? Options: classroom, toy store, soccer game, dinner. Output: classroom

A Unified CSR Dataset

Representations of Reasoning Skills

Skill-Fusion based generalization

Task A 😵

A new task w/ very limited labels.

Skill Retriever

Retrieved Skills

A multi-task LM (e.g., T0, FALN, GPT-3, etc.)

By (re-)learning a few CSR skills, I can now do Task A better! 😄
Questions?

Solving a Commonsense Reasoning Dataset

Goal: Perform well on a test set

Commonsense Question Answering

- UnifiedQA* Khashabi et al. (2020)
- XLNet+GraphReason
- RoBERTa Liu et al. (2019)
- CAGE-reasoning
- KagNet
- BERT-LARGE

Paper With Code: CommonsenseQA 1.1

Solving Commonsense Reasoning

Goal: Satisfy the real-world needs

- well-rounded
- learns fast
- data efficient
- robust to variations
- can resolve ambiguity

People
- A person performing in front of people might be nervous
- find it hard to relax

Search

Do you mean *when is the super bowl 2022?*

And more...