Making Sense of Common Sense

How can language models perform better at commonsense reasoning?

Sonali Serro, Haerang Lee
W266, Fall 2020, UC Berkeley
Github

Why is it OK to leave the **door of a wardrobe** open, but not the **door of a refrigerator**?

A question so obvious to us is not so easy for machines that lack innate common sense.

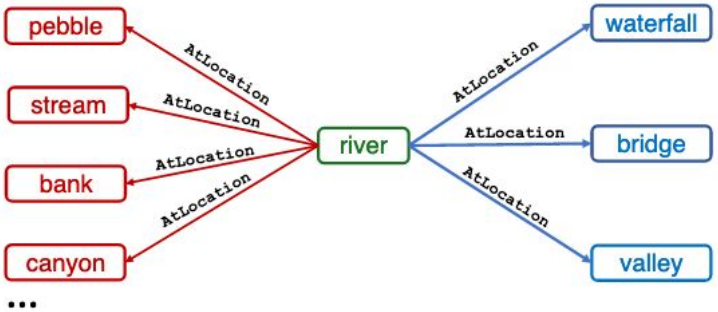
Main Task

CommonsenseQA

Multiple-choice QA. Aims to capture commonsense reasoning beyond associations.

BERT-Large 55.9%
Human 88.9%

a) Sample ConceptNet for specific subgraphs



b) Crowd source corresponding natural language questions and two additional distractors

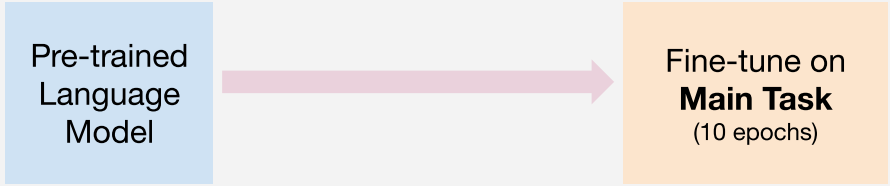
*Where on a **river** can you hold a cup upright to catch water on a sunny day?*
✓ **waterfall**, ✗ bridge, ✗ valley, ✗ pebble, ✗ mountain

*Where can I stand on a **river** to see water falling without getting wet?*
✗ waterfall, ✓ **bridge**, ✗ valley, ✗ stream, ✗ bottom

*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*
✗ waterfall, ✗ bridge, ✓ **valley**, ✗ bank, ✗ island

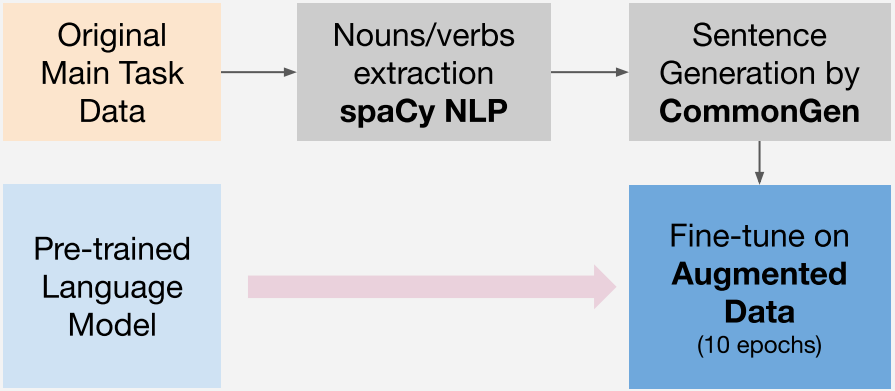
Source

Baseline



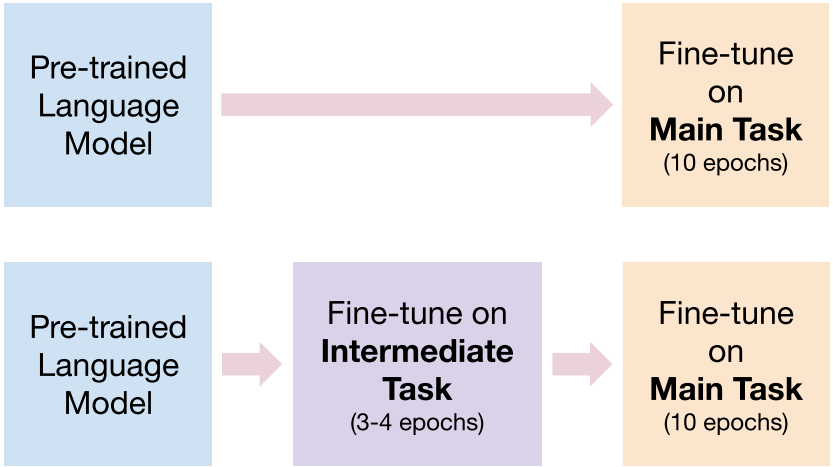
Intermediate-Task Transfer Learning

Generative Data Augmentation



Intermediate Task Transfer Learning

Transfer Learning Pipeline



Fine-tuning a pretrained model on **a different intermediate task** before fine-tuning it on the **main task.**

We fine-tune on the intermediate task for just 3-4 epochs, as over-fitting on the intermediate task can impair target task performance.

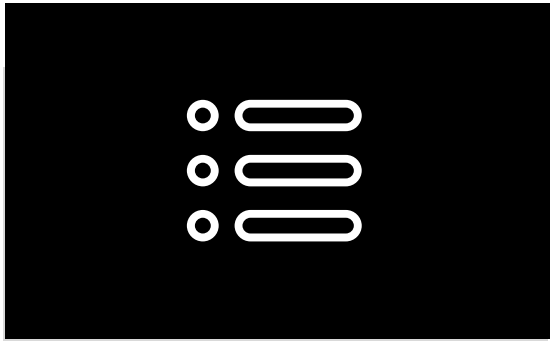
Intermediate Tasks

SocialIQA

Multiple-choice QA about social situations. Questions probe **emotional and social intelligence** from various everyday situations.

CosmosQA

Multiple-choice QA requiring the interpretation of the likely **causes and effects** of events. Emphasizes reading between the lines and reasoning beyond the text.

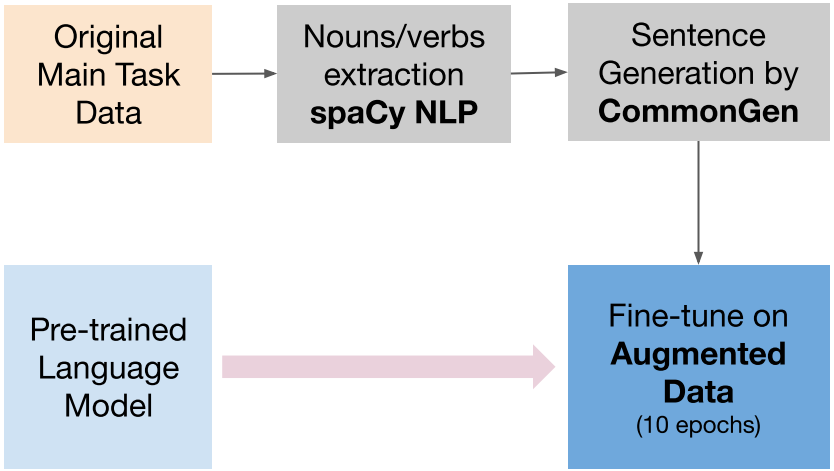


HellaSwag

Natural language inference task, generated via *adversarial filtering* that tests a model's ability to choose the most plausible **continuation to a sequence of events or steps.**

Generative Data Augmentation

Data Augmentation Pipeline



Augment the CommonsenseQA dataset using context generated by a model trained on a generative commonsense reasoning task.

Data Augmentation Task

CommonGen

Constrained text generation benchmark dataset that aims to evaluate a model's generative commonsense reasoning abilities.

CommonsenseQA Question-Answer choice:

What do people aim to do at work? A: learn from each other B: wear hats

SpaCy NLP EXTRACTED FEATURES:

{people, aim, work, learn, from, each, other}
{people, aim, work, wear, hats}

CommonGen GENERATED SENTENCES:

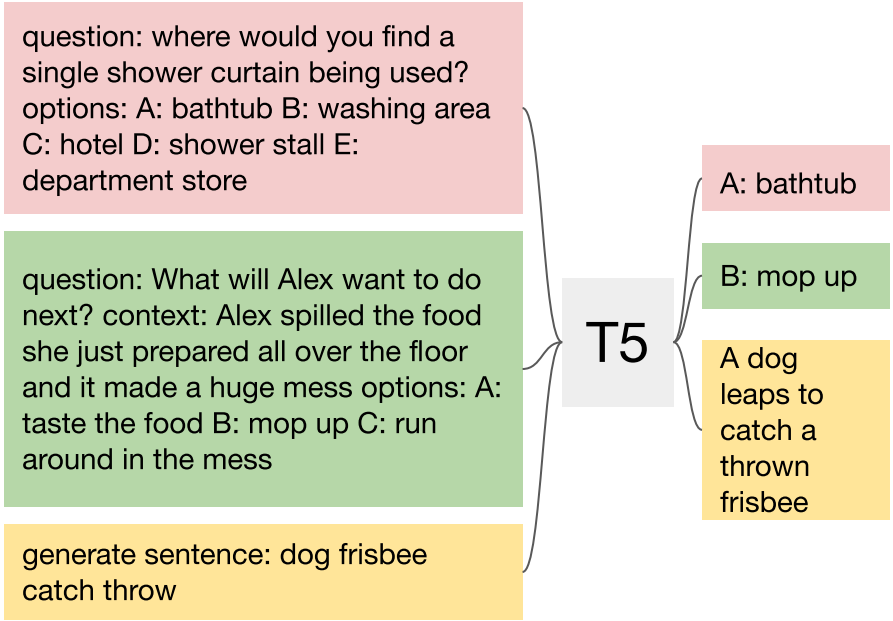
people work together to learn from each other
people are wearing hats to work

AUGMENTED CommonsenseQA:

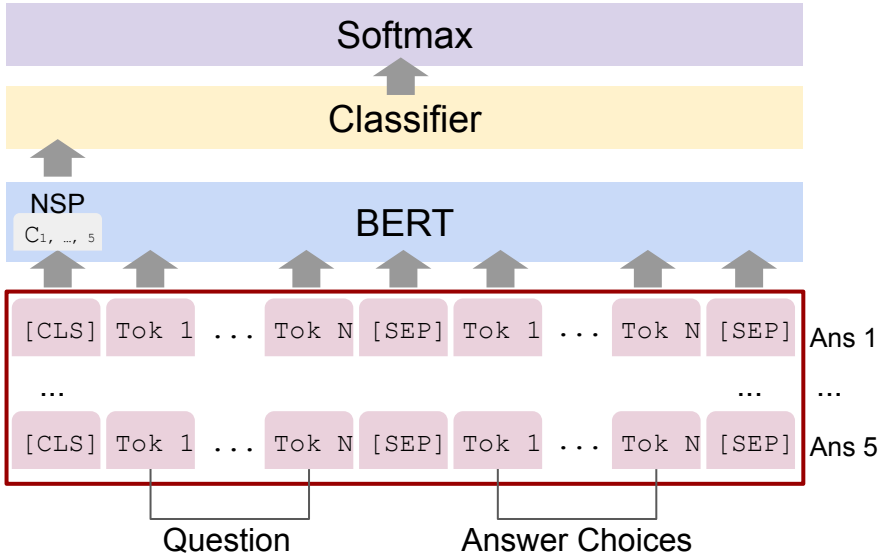
What do people aim to do at work? A: learn from each other | ***people work together to learn from each other*** B: wear hats | ***people are wearing hats to work***

Language Model

T5 vs. BERT



T5 input and output formats



BERT input format and architecture

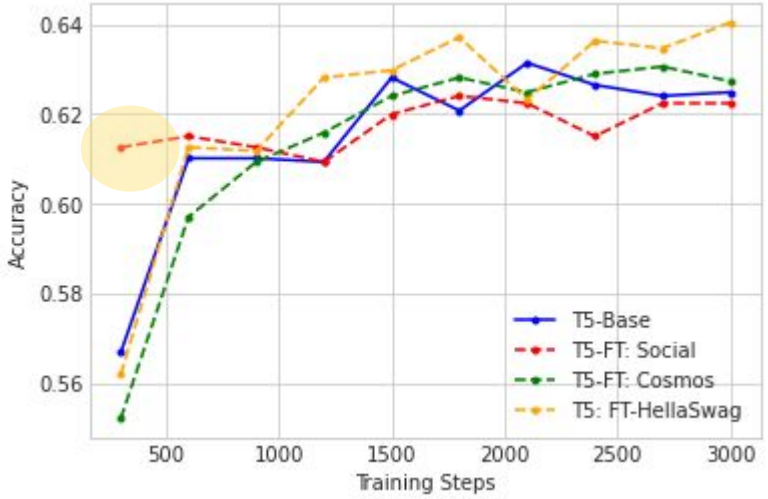
Results

Intermediate Task Transfer Learning

- **Exceeded baseline performance:**
  - T5-FT$_{\text{HellaSwag}}$
  - T5-FT$_{\text{Cosmos}}$
- **Did not exceed baseline:**
  - T5-FT$_{\text{Social}}$ (but converges quickly)

Model	Dev Acc (%)	LR	WS
T5-Base	62.57	1e-4	0
T5-FT$_{\text{Social}}$	62.00	1e-4	200
T5-FT$_{\text{Cosmos}}$	62.82	5e-5	0
T5-FT$_{\text{HellaSwag}}$	**64.04**	1e-4	200

Top: Table 2: Model accuracy is measured against the official CommonsenseQA development set.



Bottom: Figure 3: CommonsenseQA dev accuracy over training steps over 10 epochs

Performance in Low-Resource Settings

- Intentionally **limited** main task (CommonsenseQA) training data
- **Helpful datasets when commonsense data scarce:**
  - SocialIQA
  - CosmosQA
- **Unhelpful in low-resource settings**
  - HellaSwag (possibly due to very different data format.)

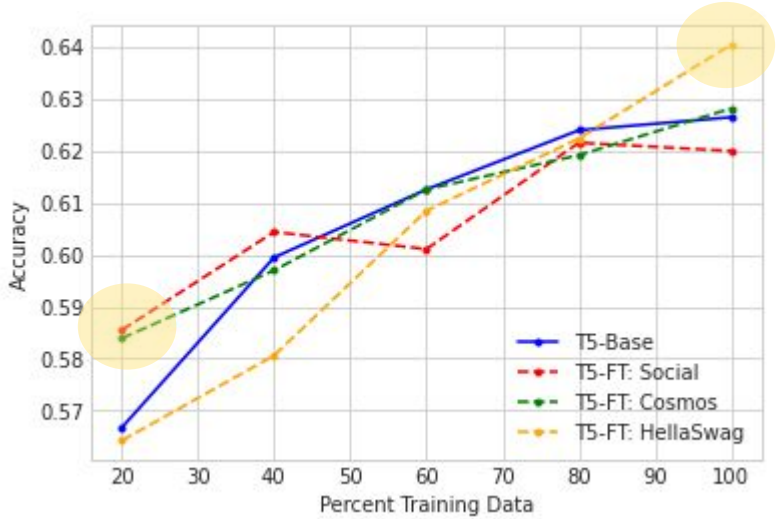


Figure 4: Accuracy per fraction of training data.

Generative Data Augmentation

- Converges considerably faster (2x).
- Does *not* exceed baseline performance
- Still, promising signs of improving a model's commonsense reasoning abilities.

Model	Dev Accuracy (%)	
	3 epochs	10 epochs
T5-Base	61.83	**62.73**
T5-Base$_{CG}$	**62.48**	62.48

Table 3: Generative Data Augmentation Results using CommonGen. Model accuracy is measured against the official CommonsenseQA development set. Batchsize=8 and lr=1e-4.

Error Analysis: Transfer Learning

Quality of Responses

- Higher quality means more common-sensical. A human may have made a similar choice.
- **HellaSwag** and **CosmosQA** tend to improve quality.
- **SocialIQA** cuts both ways. Limited common sense knowledge transfer.

Questions	Answer Choices (Green for target, *red* for incorrect prediction) A	B	C	D	E
HIGH-QUALITY					
Reading newspaper one of many ways to practice your what?	**literacy**	*knowing how to read*	money	buying	money bank
LOW-QUALITY					
What are people likely to do when an unexpected decent outcome occurs?	kill each other	**thank god**	*experience pain*	hatred	talk to each other

Figure 8: Sample errors and corresponding quality evaluation.

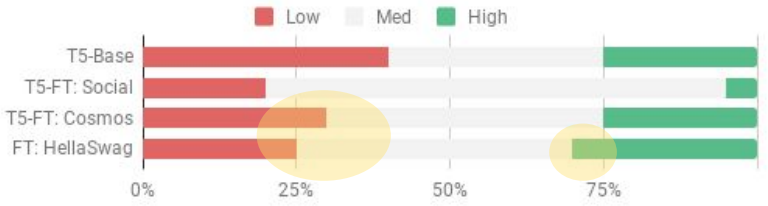
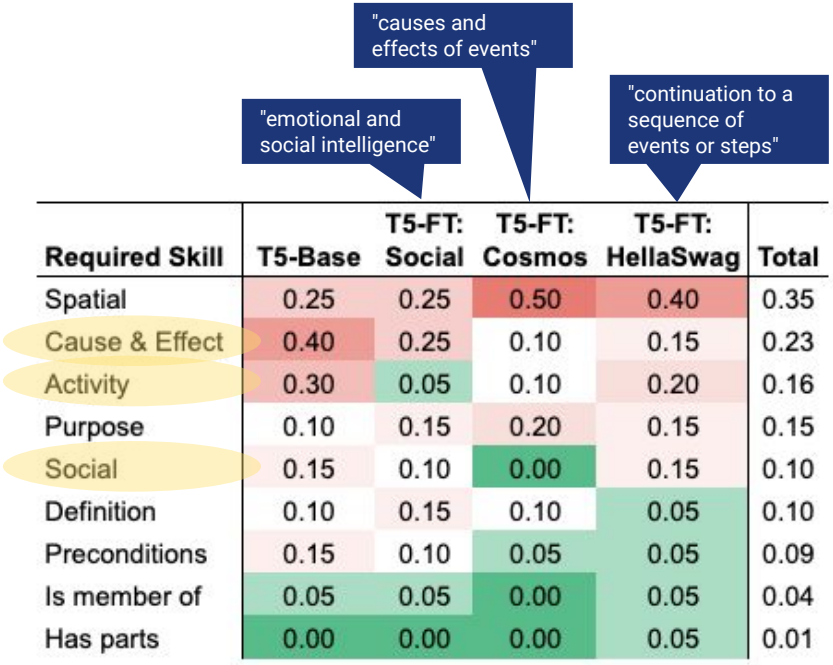


Figure 6: Response quality of sample errors by model.

Commonsense Reasoning Skills Required

- What skills would a human have used to solve the problems?
- **Decreased error types**: *Cause & effect*, *activity*, and *social* reasoning skills. in line with nature of datasets.
- **Increased error type**: *Spatial* (41% of dev data, though. Not concerning.)



Required Skill	T5-Base	T5-FT: Social	T5-FT: Cosmos	T5-FT: HellaSwag	Total
Spatial	0.25	0.25	0.50	0.40	0.35
Cause & Effect	0.40	0.25	0.10	0.15	0.23
Activity	0.30	0.05	0.10	0.20	0.16
Purpose	0.10	0.15	0.20	0.15	0.15
Social	0.15	0.10	0.00	0.15	0.10
Definition	0.10	0.15	0.10	0.05	0.10
Preconditions	0.15	0.10	0.05	0.05	0.09
Is member of	0.05	0.05	0.00	0.05	0.04
Has parts	0.00	0.00	0.00	0.05	0.01

Figure 7: Types of commonsense reasoning skills required by the incorrect examples.

Takeaways & Questions

1. Both **intermediate task transfer learning** and **generative data augmentation** can be useful in teaching common sense to a model.
2. Based on the input's contents and format, the benefits may manifest in different ways.
3. Future work could include incorporating generated commonsense knowledge into a language model.

References

- Templates, shapes, and elements from http://freegoogleslidestemplates.com
- Images: See at the bottom of each slide
- For full references, please see our paper on Github