Harvard John A. Paulson
School of Engineering
and Applied Sciences

Explain Yourself! Leveraging Language Models for Commonsense Reasoning

Authors: Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, Richard Socher
Presenters: Karly Hou, Eshika Saxena, Leonard Tang, Kat Zhang

Introduction

- **Commonsense reasoning**: making human-like presumptions and judgements about ordinary situations
- Modern ML methods struggle with commonsense reasoning
- Explanations help verbalize reasoning that models learn while training
- Common sense Question Answering (CQA) dataset

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion

How do these models perform reasoning and to what extent is that reasoning based on world knowledge?

Key Contributions

Common Sense Explanations (CoS-E): Collected human explanations (annotations and natural language explanations) to build on top of CQA

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.

Key Contributions

1. Common Sense Explanations (CoS-E)
2. Commonsense Auto-Generated Explanations (CAGE)
3. CAGE outperforms best baseline by 10% and produces explanations to justify its predictions
4. Explanation transfer on two out-of-domain datasets

Related Work: Commonsense reasoning

- Commonsense reasoning datasets:
  - Story Cloze: predicting story ending from a set of plausible endings
  - Situations with Adversarial Generations (SWAG): predicting next scene based on initial event
- Models achieve human-level performance on some datasets
- Models struggle with understanding how pronouns resolve between sentences and world knowledge

- CQA addresses this by requiring models to infer from the question
- Language models perform poorly compared to human participants on CQA

Unclear: Do models actually do common-sense reasoning?

Related Work: Natural language explanations

- Rationale generation by highlighting complete phrases in input text that are sufficient to predict desired output (Lei et al., 2016)
- Human-generated natural language explanations to train a semantic parser to generate noisy labeled data and train a classifier for generating explanations (Hancock et al., 2018)
- Interpretability comes at the cost of loss in performance on Stanford Natural Language Inference dataset (Camburu et al., 2018)
- Multi-modal: Ensemble explanations and visual explanations improve performance (Rajani and Mooney, 2018 and 2017)

Do explanations for CQA lead to improved performance?

Related Work: Knowledge transfer in NLP

- Reliance on transfer of knowledge through pre-trained word vectors (e.g. Word2vec, GloVe) and contextualized word vectors (more refined with general encoding)
- Language models trained from scratch on large amounts of data and fine-tuned on specific tasks perform well
  - Only a few parameters need to be learned from scratch
  - Perform well on small amounts of supervised data
- Gap: Fine-tuned language models don’t perform as well on CQA

Can we leverage these models to generate explanations and show that these explanations capture common sense?

Common Sense Explanations (CoS-E)

Dataset Structure/Creation

- Based on the CQA dataset
- CoS-E provides natural-language explanations for the correct answer choice and highlights important words in the question
- Explanations generated using MTurk
- Goal: To show whether models are performing reasoning correctly

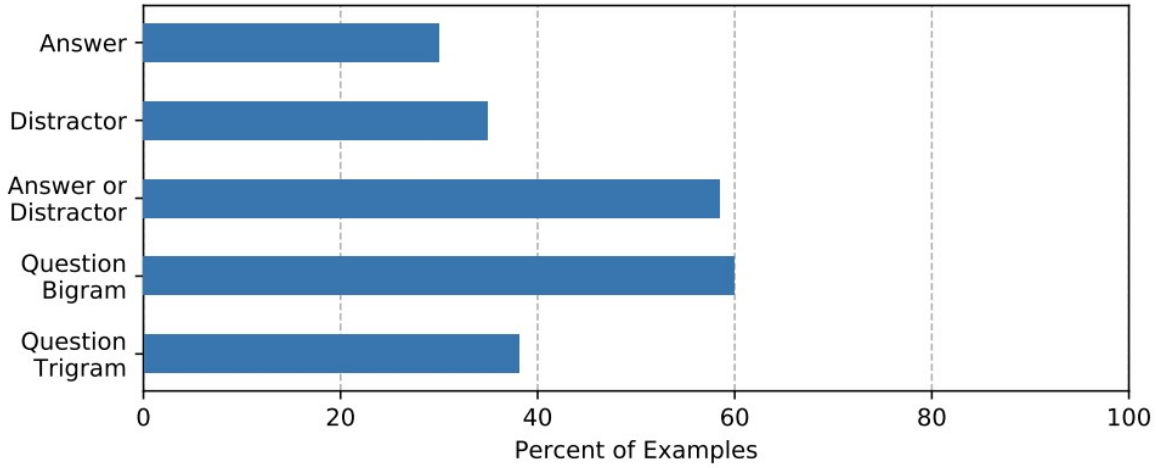
Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.

CQA

CoS-E

Dataset Considerations

- *CoS-E-selected* refers to the highlighted words, *CoS-E-open-ended* refers to the explanations
- Quality control was performed on the annotations and explanations
- Even explanations that don’t discuss the ground truth answer are useful

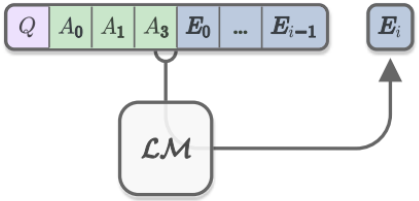


Commonsense Auto-Generated Explanations (CAGE)

CAGE Phase 1

Commonsense Auto-Generated Explanations (CAGE) Phase 1:

- Provide CQA example alongside corresponding CoS-E explanation to a language model
- Train model to generate the CoS-E explanation

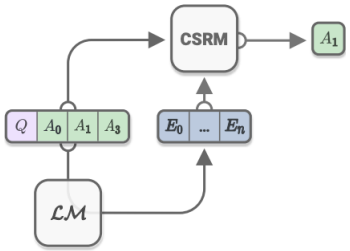


(a) One time-step of training a CAGE language model to generate explanations from CoS-E. It is conditioned on the question tokens Q concatenated with the answer choice tokens A_1, A_2, A_3 and previously generated tokens $E_1, \ldots, E_{i-1}$. It is trained to generate token E_i.

CAGE Phase 2

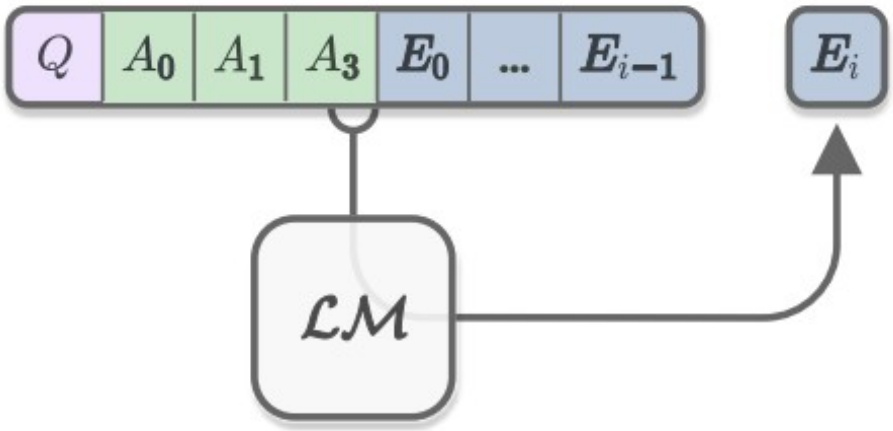
Commonsense Auto-Generated Explanations (CAGE) Phase 2:

- Use language models to generate explanations for each example in the training and validation sets of CQA
- Provide CAGE explanations to a second model by concatenating it to the original input (question, answer choices, and language model output)



(b) A trained CAGE language model is used to generate explanations for a downstream commonsense reasoning model (CSRM), which itself predicts one of the answer choices.

CAGE Phase 1: Intuition



One training step for CAGE

How is CAGE trained?

- Language Model trained to generate explanations from question-answer choice pairs
- Use pretrained OpenAI GPT
- Fine-tuned on the CQA and CoS-E dataset combination
- Two possible settings: explain-then-predict **(reasoning)** and predict-then-explain **(rationalization)**

CAGE Notation

- Question q
- Answer choices c_0, c_1, c_2
- Correct answer $a \in c_0, c_1, c_2$
- CoS-E explanation e_h
- CAGE predict explanation e

Reasoning

- Model is fine-tuned on the question, answer choices, and explanation tokens, but not the actual label.

$$C_{RE} = \text{“}q, c0, c1, \text{or } c2? \text{ commonsense says ”}$$

- Objective Function (canonical conditional language modeling objective):

$$\sum_i \log P(e_i | e_{i-k}, \ldots, e_{i-1}, C_{RE}; \Theta)$$

Rationalization

- Model is now also given the ground truth label a:

$$C_{RA} = \text{“ } q, c0, c1, \text{ or } c2? \ a \text{ because ”}$$

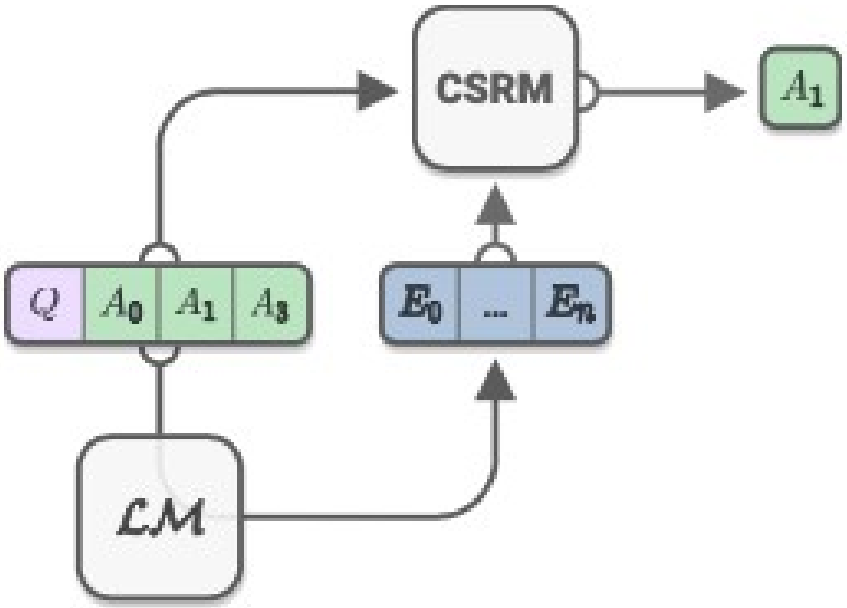
- Objective Function is the same as before but is also conditioned on the *label a*
- Thus, the explanations create rationalization that makes the model more interpretable

Training Parameters

- Generate sequences of maximum length 20
- Batch Size: 36, Epochs: 10
- Selected the best model using BLEU and perplexity scores

Commonsense Predictions with Explanations

CAGE Phase 2 Intuition



CAGE Inference

- Given human explanation from CoS-E or LM reasoning, can then perform predictions on CQA
- Simply concatenate Question, [Sep], Explanation, [Sep], Answer Choice as input to downstream CSRM (classifier)
- Use binary classification head on top of BERT backbone
  - 3 answer choices → 3 input sequences
  - Take sequence yielding highest confidence as output

CSRM (BERT) Training HP

- Train batch size: 24
- Test batch size: 12
- 10 training epochs
- Max sequence length of 50 for labels-only; 175 including explanations

Experimental Results

Experimental Results

Method	Accuracy (%)
BERT (baseline)	63.8
CoS-E-open-ended	65.5
CAGE-reasoning	**72.6**

Table 2: Results on CQA dev-random-split with CoS-E used during training.

Experimental Results

- Google search “question + answer choice” for each example and collected 100 top snippets per answer as context for **Reading Comprehension model**
- Extra data did not improve accuracy
- CAGE-reasoning resulted in **10% accuracy gain** over previous SOTA

Method	Accuracy (%)
RC (Talmor et al., 2019)	47.7
GPT (Talmor et al., 2019)	54.8
CoS-E-open-ended	60.2
CAGE-reasoning	**64.7**

Experimental Results

Method	Accuracy (%)
CoS-E-selected w/o ques	53.0
CoS-E-limited-open-ended	67.6
CoS-E-selected	70.0
CoS-E-open-ended w/o ques	84.5
CoS-E-open-ended*	**89.8**

Table 4: Oracle results on CQA dev-random-split using

- Oracle upper-bound: **human-generated explanations** from CoS-E provided during training and validation
- Unfair setting bc human that provided explanation had **ground truth answer**
- **“CoS-E selected”**: explanation consists of words humans selected as justification for model

Transferring Explanations Across Domains

- How well do natural language explanations transfer from CQA to SWAG and Story Cloze Test?
- Use GPT CAGE model fine-tuned on CQA train/dev to generate explanations on SWAG and Story Cloze Spring 2016 train/val
- Rinse and repeat using BERT with classifier head

Experimental Results

Method	SWAG	Story Cloze
BERT	84.2	89.8
+ expl transfer	83.6	89.5

- Camburu et al (2018) show that transferring explanations from SNLI to MultiNLI performs very poorly
- Transfer of explanations on commonsense reasoning tasks
- NLI problem has small fixed set of pre-defined labels unlike commonsense reasoning tasks like CQA, SWAG, Story Cloze
- **Adding explanations led to very small decrease in performance**

Qualitative Analysis

Analysis of CAGE

- CAGE-reasoning at train + validation → 72% accuracy
- CoS-E-open-ended performance at 90%—why the gap?
- Measure quality of CAGE
  - Human evaluation (42% CAGE vs 52% CoS-E-open-ended)
  - BLEU score measures syntactical precision by n-gram overlap
  - Perplexity: token-level measure of how well language models predict next word
- Result: beneficial to fine-tune the LM, but humans and LMs have widely varying ways of providing useful explanations

Analysis of baseline BERT model

- Error analysis on baseline BERT w/o explanations: performs poorly on longer / more compositional questions → explanations help w this
- CAGE reasoning typically simpler construction than CoS-E-open-ended, but adds meaningful context
- However, CAGE still provides “incorrect” answers often

Question:	A child wants to play, what would they likely want?
Choices:	**play tag**, breathe, fall down
CoS-E:	A child to play tag
Reason	Children want to play tag, and they want to play tag with their friends.
Rationale:	Children want to play tag, what would they want to do?

Question:	What is the main purpose of having a bath?
Choices:	**cleanness**, use water, exfoliation, hygiene, wetness
Explanation:	the only purpose of having a bath is to clean yourself.
Question:	Where can you store you spare linens near your socks?
Choices:	cabinet, chest, hospital, **dresser drawers**, home
Explanation:	dresser drawer is the only place that you can store linens.
Question:	Where do you find the most amount of leafs?,
Choices:	**forrest**, floral arrangement, compost pile, field, ground
Explanation:	the most likely place to find leafs is in a garden.

Domain transfer: SWAG + Story Cloze

	SWAG
Question:	Men are standing on motorbikes getting ready for a motocross competition.
Choices:	man places the ladders onto a fence and winds up a marching wall, high with hammer and a stone., **man is talking to the camera and standing on a podium.**, man stands outside in the field going at arms of people and leading a long jumping calf in front., man drops the javelin to the ground and jumps it very high.
Explanation:	man is talking to the camera and not the crowd.
Question:	The man examines the instrument in his hand.
Choices:	The person studies a picture of the man playing the violin., **The person holds up the violin to his chin and gets ready.**, The person stops to speak to the camera again., The person puts his arm around the man and backs away.
Explanation:	the person is holding the instrument in his hand.
Question:	The woman is seated facing the camera while another woman styles her hair.
Choices:	The woman in purple is wearing a blue dress and blue headband, using the pits to style her hair., The woman begins to cut the hair with her hair then serves it and begins brushing her hair and styling it., The woman puts some right braids on his., **The woman continues to have her hair styled while turned away from the camera.**
Explanation:	the woman is using the braids to trim her hair.
	Story Cloze (ROCStories)
Question:	My friends all love to go to the club to dance. They think it's a lot of fun and always invite. I finally decided to tag along last Saturday. I danced terribly and broke a friend's toe.
Choices:	My friends decided to keep inviting me out as I am so much fun., **The next weekend, I was asked to please stay home.**
Explanation:	the next weekend, i would be asked to stay home
Question:	Ari spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. Ari waits 2 weeks for his pickles to get sour.
Choices:	**Ari opens the jar to find perfect pickles.**, Ari's pickles are sweet.
Explanation:	pickles are the only thing that can be found in a jar.
Question:	Gina sat on her grandpa's bed staring outside. It was winter and his garden was dead until spring. Her grandpa had passed away so there would be no one to tend it. The weeds would take over and strangle the flowers.
Choices:	Gina asked her grandpa what kind of flowers he liked best., **Gina decided to go outside and pick some of the weeds.**
Explanation:	the weeds would take over and strangle the flowers.

Conclusion & Group Discussion

Conclusion

- CoS-E on top of CommonsenseQA
- CAGE framework → LM leverages explanations
- Classifier on top of explanations
- SOTA performance on a difficult commonsense reasoning task
- Opens further avenues for studying explanation as it relates to interpretable commonsense reasoning

Discussion questions

- BLEU and perplexity as measures of goodness?
  - Has been shown multiple times to correlate poorly with human judgement
- Joint training of explanation and label? Not one prior to other
- Can commonsense reasoning help general reasoning (e.g. mathematics, Fermi, counterfactual) in other domains?
- How to align human/machine explanations?
- Research merit of this work?
- Recent work (RLPrompt, AutoPrompt) has shown that optimal prompts for LMs (with respect to downstream task performance) are often times gibberish
  - What does this say about the validity of using SOTA LMs for explanations?
  - Can we regularize LM training to better align with human reasoning?