テキストを通して世界を見る：機械読解における常識的推論のための画像説明文の評価

Diana Galvan-Sosa[1], 西田京介[2], 松田耕史[3,1], 鈴木潤[1,3], 乾健太郎[1,3]

[1]東北大学　[2]NTTメディアインテリジェンス研究所
[3]理化学研究所

Commonsense knowledge

- Set of background information an individual is assumed to know.

Food is cooked in a kitchen.

Food is served on a plate.

⋮

Food is eaten.

Example: Cooking

Known facts

Commonsense knowledge

- Usually acquired through crowdsourced annotations of events.
  - E.g. ConceptNet (Liu and Singh, 2004), ATOMIC (Sap et al., 2018), Event2Mind (Rashkin et al., 2018)

Cooking

Food is needed for cooking.

Cooking is the main activity of chefs.

People cook when they are hungry.

Cooking is a good hobby.

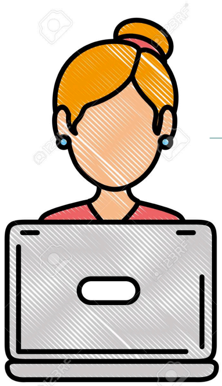
…

Pro: Large-scale data

Con: Reporting bias

Commonsense knowledge

- Usually acquired through crowdsourced annotations of events.
  - E.g. ConceptNet (Liu and Singh, 2004), ATOMIC (Sap et al., 2018), Event2Mind (Rashkin et al., 2018)



Cooking

Food is needed for cooking.

...

Cooking is a good hobby.

A kitchen is a room for cooking.

Stoves are used for cooking.

Some information can be overlooked

Pro: Large-scale data

Con: Reporting bias

Commonsense knowledge

- **Our premise**: Understanding is a process by which people match what they **see** and hear to what they have already experienced (Schank and Abelson, 1977).

Cooking

Food is needed for cooking.

Women cooking in a kitchen.

A pan being used for cooking.

...

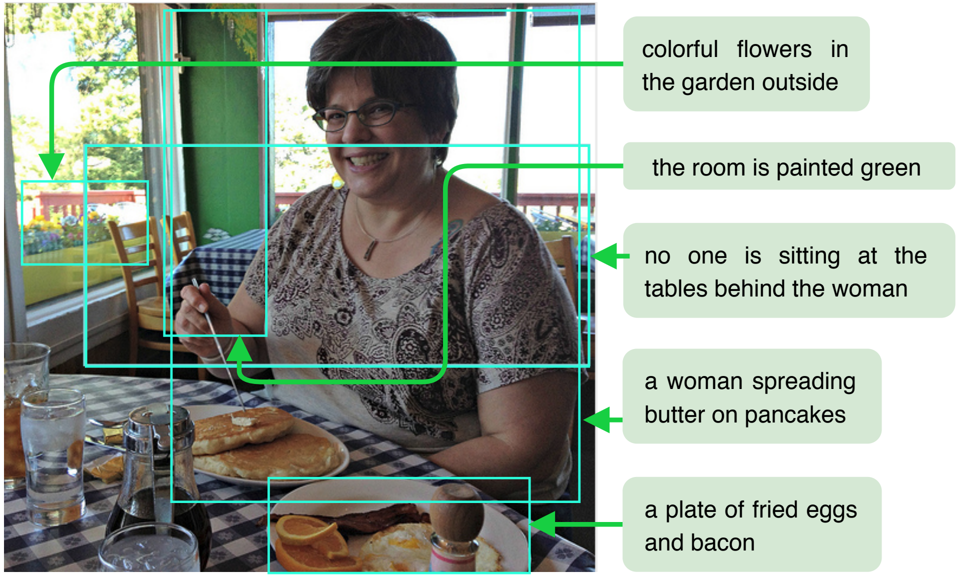
Woman chopping vegetables.

✓ Richer information

✓ Already available through an existing image dataset

Visual Genome (Krishna et al., 2017)

- Over 108K images with an average of 50 region descriptions each.



Our goal: Use Visual Genome as a source of commonsense knowledge.

Evaluation approaches

- **Intrinsic:** Evaluate knowledge independently.
  - Physical commonsense extraction from image datasets (Yatskar et al., 2018), (Mukuze et al., 2018)
  - Language models as knowledge bases (Petroni et al., 2019)
- **Extrinsic:** Measure knowledge on a real application.
  - Machine reading comprehension datasets: Visual QA (Goyal et al., 2017), MCScript (Ostermann et al., 2018, 2019), CosmosQA (Huang et al., 2019), etc.

Our approach: Evaluate the knowledge in Visual Genome through a reading comprehension task.

MCScript (Ostermann et al., 2018, 2019)

- A reading comprehension dataset of stories about general everyday activities (i.e. making breakfast)

MCScript

	Text	*CS
Train	7,032	2,699
Dev	1,006	405

*Commonsense

MCScript 2.0

	Text	CS	Text/CS
Train	5,685	7,091	1,415
Dev	844	966	210

T Today I woke up and decided to make bacon and eggs for breakfast. I walked to the kitchen and got out all of the ingredient I needed, which included, eggs, bacon, cheese, onion, and green pepper. ... Once the bacon was cooked, I poured the veggies and egg mixture into the pan, stirring occasionally, until the mixture set up and was solid . ... I put the plate on the table and poured out a glass of orange juice to go with my meal . It was delicious!

Q1 What did they peel?

a. Onion b. Bacon

Q2 What did they set on the plate?

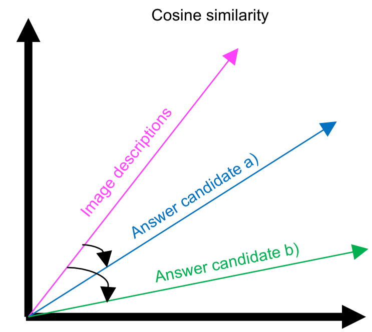
a. The orange juice b. The eggs and bacon

Key idea

5 Choose the answer with the highest cosine similarity

1 Extract key words

T	I wanted to plant a tree. I went to the home and garden store and picked a nice oak. Afterwards, I planted it in my garden.
Q1	What was used to dig the hole?
	a. a shovel b. his bare hands



4 Match with answer candidates

- Square gray shovel with handle
- Large blue and white airplane with blue symbol
- Two workers next to an airport
- A worker digs in the snow
- Two workers digging a hole
- Man using a shovel to dig

3 Get image descriptions

2 Query VisualGENOME

Similarity baseline

- **Hypothesis**: If the **region description** has meaningful commonsense information, its cosine similarity should be greater for the **correct answer**.
- **Answer-region similarity score**
  1. Calculate the cosine similarity of an answer candidate with each of the retrieved region descriptions.
  2. Average all the cosine similarities in 1)
- **Sentence vector representations**
  - TF-IDF vectors
  - BERT sentence embeddings (SBERT) (Reimers and Gurevych, 2019)

Results

Selecting the answer that has the highest cosine similarity with the image descriptions was good for commonsense questions

Data	MCScript			MCScript2.0			
Model	Text	Commonsense	Total	Text	Commonsense	Text/Commonsense	Total
Similarity baseline							
TF-IDF vectors	52.4	50.4	51.8	53.8	54.2	55.2	54.2
SBERT embeddings	55.6	**56.0**	**55.7**	**58.8**	**60.5**	**59.5**	**59.7**
SBERT embeddings (all regions)	**55.8**	54.1	55.3	50.4	55.0	53.8	52.9

The cosine similarity between the correct answer and the regions descriptions **decreased** when using **all** the region descriptions in an image.

Fine-tuned BERT

- **Hypothesis**: Region descriptions can help a state-of-the-art model to further improve its perfomance on commonsense questions.

- **BERT (base) (Devlin et al., 2018):** Fine-tuned on MCScript. Two different input formats.
  - Vanilla BERT

CLS	Question tokens	[unused0]	Answer tokens	SEP	Text tokens	SEP

  - Visually Enhanced BERT

CLS	Question tokens	[unused0]	Answer tokens	SEP	Image descriptions tokens	[unused1]	Text tokens	SEP

Results

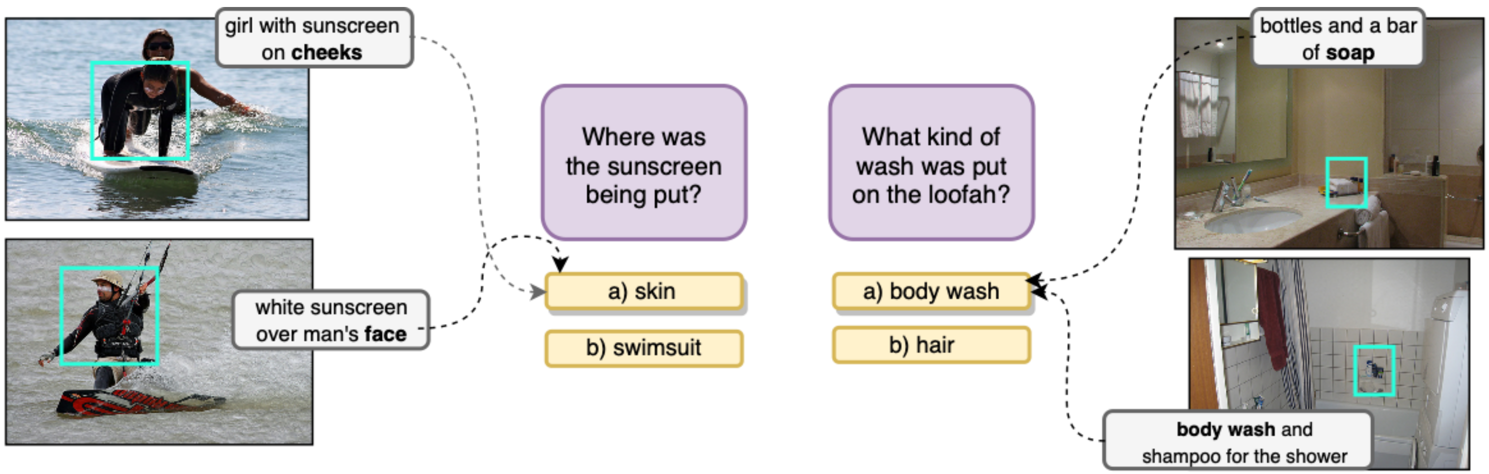
Image region descriptions improve BERT's SOTA performance

Data	MCScript			MCScript2.0			
Model	Text	Commonsense	Total	Text	Commonsense	Text/Commonsense	Total
Previous SOTA models							
TriAN (Wang et al., 2018)	-	-	85.27	-	-	-	-
HFL-RC (Chen et al., 2018)	-	-	86.46	-	-	-	-
Fine-tuned BERT							
vanilla-BERT	88.27	82.72	86.68	**86.02**	75.16	75.71	79.75
Visually Enhanced BERT	**88.37**	**83.70**	**87.03**	85.31	**77.74**	**76.77**	**80.79**

There is a positive effect of region descriptions on commonsense questions

Case study

- Are image descriptions helping BERT choose the correct answer?



Two sample questions answered incorrectly by BERT and two sample images with the regions retrieved by our framework. Using the region descriptions, Visually Enhanced BERT was able to select the correct answer.

Conclusion

- Dense image descriptions are an **alternative source of commonsense knowledge**.
- We extrinsically evaluated **Visual Genome** on a commonsense reading comprehension task.
- We show how a **pre-trained BERT** fine-tuned in the aforementioned task benefits from the **image descriptions** retrieved from our framework.

Future work

- Further investigate BERT’s knowledge
  - Does BERT-large contain more commonsense knowledge than BERT-base?
  - If so, is this knowledge the same as the one contained in dense image descriptions?

- Written input annotations vs. visual input annotations
  - Is the knowledge contained in dense image descriptions similar or complementary to that of ConceptNet?

- 日本語の質問を歓迎します！