Seeing the world through text: Evaluating image descriptions for commonsense reasoning in machine reading comprehension

Diana Galvan-Sosa[1], **Jun Suzuki**[1,2], **Kyosuke Nishida**[3],
Koji Matsuda[2,1], **Kentaro Inui**[1,2]
[1]Tohoku University, [2]RIKEN, [3]NTT Media Intelligence Laboratories
{dianags, jun.suzuki, matsuda}@ecei.tohoku.ac.jp,
kyosuke.nishida.rx@hco.ntt.co.jp,
inui@ecei.tohoku.ac.jp

Abstract

Despite recent achievements in natural language understanding, reasoning over commonsense knowledge still represents a big challenge to AI systems. As the name suggests, common sense is related to perception and as such, humans derive it from experience rather than from literary education. Recent works in the NLP and the computer vision field have made the effort of making such knowledge explicit using written language and visual inputs, respectively. Our premise is that the latter source fits better with the characteristics of commonsense acquisition. In this work, we explore to what extent the descriptions of real-world scenes are sufficient to learn common sense about different daily situations, drawing upon visual information to answer script knowledge questions.

1 Introduction

The recent advances achieved by large neural language models (LMs), such as BERT (Devlin et al., 2018), in natural language understanding tasks like question answering (Rajpurkar et al., 2016) and machine reading comprehension (Lai et al., 2017) are, beyond any doubt, one of the most important accomplishments of modern natural language processing (NLP). These advances suggest that a LM can match a human's stack of knowledge by training on a large text corpora like Wikipedia. Consequently, it has been assumed that through this method, LMs can also acquire some degree of *commonsense* knowledge. It is difficult to find a unique definition, but we can think of *common sense* as something we expect other people to know and regard as obvious (Minsky, 2007). However, when communicating, people tend not to provide information which is obvious or extraneous (as cited in Gordon and Van Durme (2013)). If common sense is something obvious, and therefore less likely to be reported, what LMs can learn from text is already being limited. Liu and Singh (2004) and more recently Rashkin et al. (2018) and Sap et al. (2019) have tried to alleviate this problem by collecting crowdsourced annotations of commonsense knowledge around frequent phrasal events (e.g., PERSONX EATS PASTA FOR DINNER, PERSONX MAKES PERSONY'S COFFEE) extracted from stories and books. From our perspective, the main limitation of this approach is that even if we ask annotators to make explicit information that they will usually omit for being too obvious, the set of commonsense facts about the human world is too large to be listed. Then, what other options are there?

As the name suggests, common sense[1] is related to *perception*, which the Oxford English Dictionary defines as the ability of becoming aware of something through our senses: SIGHT *(e.g., the sky is blue)*, HEARING *(e.g., a dog barks)*, SMELL *(e.g., trash stinks)*, TASTE *(e.g., strawberries are sweet)*, and TOUCH *(e.g., fire is hot)*. Among those, vision (i.e., sight) is one of the primary modalities for humans to learn and reason about the world (Sadeghi et al., 2015). Therefore, we hypothesize that annotations of visual input, like images, are an option to learn about the world without actually experiencing it. This paper explores to what extent the textual descriptions of images about real-world scenes are sufficient to learn common sense about different human daily situations. To this end, we use a large-scale image dataset as knowledge base to improve the performance of a pre-trained LM on a commonsense machine reading comprehension task. We find that by using image descriptions, the model is able to answer some

[1]Latin *sensus* (perception, capability of feeling, ability to percieve)

Proceedings of the Second Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 23–29
Barcelona, Spain (Online), December 13, 2020

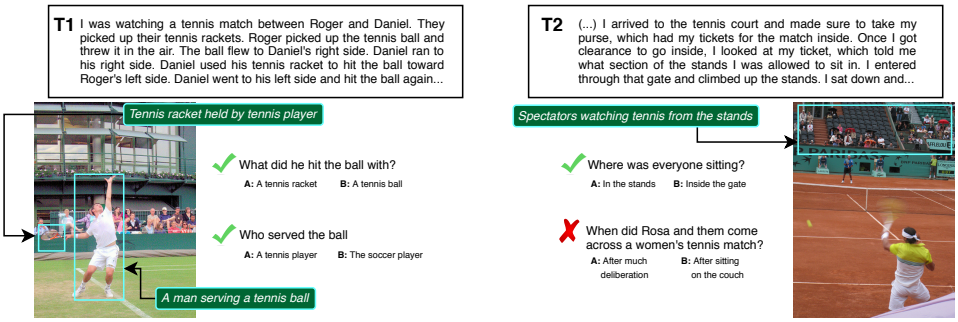

Figure 1: Example of three selected and one removed commonsense questions from two MCScript2.0 instances.

questions about common properties and locations of objects that it previously answered incorrectly. The ultimate goal of our work is to discover an alternative to the expensive (in terms of time) and limited (in terms of coverage) crowdsourced-commonsense acquisition approach.

2 Related work

Knowledge extraction. Previous works have already recognized the rich content of computer vision datasets and investigated its benefits for commonsense knowledge extraction. For instance, Yatskar et al. (2016) and Mukuze et al. (2018) derived 16K commonsense relations and 2,000 verb/location pairs (e.g., *holds*(dining-table, cutlery), *eat*/*restaurant*) from the annotations included in the Microsoft Common Objects in Context dataset (Lin et al., 2014) (MS-COCO). However, they only focused on physical commonsense. A more recent trend is to query LMs for commonsense facts. While a robust LM like BERT has shown a strong performance retrieving commonsense knowledge at a similar level to factual knowledge (Petroni et al., 2019), this seems to happen only when that knowledge is explicitly written down (Forbes et al., 2019).

Machine reading comprehension (MRC). MRC has long been the preferred task to evaluate a machine's understanding of language through questions about a given text. The current most challenging datasets such as Visual Question Answering (Goyal et al., 2017), NarrativeQA (Kočiskỳ et al., 2018), MCScript (Ostermann et al., 2018; Ostermann et al., 2019), CommonsenseQA (Talmor et al., 2018), Visual Commonsense Reasoning (Zellers et al., 2019) and CosmosQA (Huang et al., 2019) were designed to be solvable only by using both context (written or visual) and background knowledge. In all of these datasets, no system has been able to reach the upper bound set by humans. This emphasizes the need to find appropriate sources for systems to equal human knowledge.

Our work lies in the intersection of these two directions. We aim to use computer vision datasets for broad commonsense knowledge acquisition. As a first step, we explore whether visual text from images provides the implicit knowledge needed to answer questions about an MRC text. Ours is an ongoing attempt to emulate the success of multi-modal information in VQA and VCR on a MRC task.

3 Approach

We evaluate image descriptions through a MRC task for which commonsense knowledge is required, and assume that answering a question incorrectly means the reader lacks such knowledge. Most of what humans consider obvious about the world is learned from experience, and we believe there is a fair amount of them written down in an image's description. We will test this idea by using image descriptions as external knowledge. Out of the different types of common sense, the text passages in the selected MRC dataset focus on *script knowledge* (Schank and Abelson, 2013), which covers everyday *scenarios* like BRUSHING TEETH, as well as the *participants* (persons and objects) and the *events* that take place during them. Since scenarios represent activities that we do on a regular basis, we expect to find images of it. Ideally, for each passage, we would automatically query an image dataset to retrieve descriptions related to what the passage is about. Retrieval is a key step in our approach and for the time

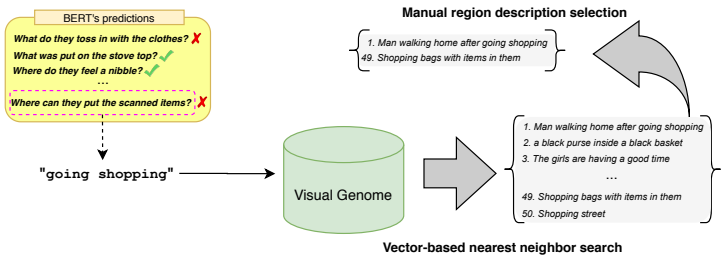

Figure 2: Retrieval process for one of the questions BERT answered incorrectly. Identifying the GOING SHOPPING scenario, querying Visual Genome and selecting the most related region descriptions to the scenario was manually done.

being, such process was done manually so we can focus on the image's description content rather than in the retrieval process itself.

There is a considerable number of crowdsourced image datasets whose image descriptions are available, which means they can be collected (and extended, if needed) for a reasonable cost. The motivation behind our approach is that once such descriptions are proven to contain useful commonsense knowledge that it is not easily obtained from text data, one can think of extending the description collection.

4 Experiments

4.1 Data

Image dataset. Visual Genome (Krishna et al., 2017) is a large-scale collection of non-iconic, real-world images with dense captions for multiple objects and regions in a single image. Each of the 108K images in the dataset has an average of 50 region descriptions of 1 to 16 words. To use this dataset as a knowledge base, we first used BERT-sentence embeddings (Reimers and Gurevych, 2019) to embedded all of the region descriptions and then created a semantic search index using FAISS (Johnson et al., 2017). When querying the index, we retrieved the top 50 results.

Reading comprehension dataset. MCScript2.0 is a dataset with stories about 200 everyday scenarios. Each instance has a text passage paired with a set of questions, which in turn have two answer candidates (one correct and one incorrect). In total, MCScript2.0 has 19,821 questions, out of which 9,935 are commonsense questions that require script knowledge. We split the dataset into train, dev and test sets as in (Ostermann et al., 2019). The train set is used as it is. However, for evaluation, we worked with a subset of 56 and 81 questions from the original dev and test sets, respectively (more details of this in the next section). The subsets include instances with passages about 15 out of the 200 scenarios. For each instance, we took all of its commonsense questions and further selected those in which the necessary commonsense knowledge might be present in one (or more) image descriptions. An example is shown in Figure 1.

4.2 Models

BERT (Baseline). Fine-tuned BERT (base-uncased) on MCScript2.0 using three different random seeds (See Appendix A).

Visually Enhanced BERT. As introduced in Section 3, we hypothesize there is commonsense knowledge present in image descriptions. This model aims to improve on the baseline by using region descriptions from Visual Genome to answer those questions were BERT was wrong. We will refer to these questions as the *unanswerable questions* set. All of them were manually inspected to identify the scenario they are about. As shown in Figure 2, the scenario name is used to query our Visual Genome index. If the results do not contain information about the scenario's events or participants, we refined the query using keywords from the question (e.g., querying "going fishing" returns no results mentioning "rod", a new query would be "going fishing rod"). To be careful not to exceed BERT's sequence length, we selected a maximum of 6 region descriptions from the results and concatenated them at the beginning

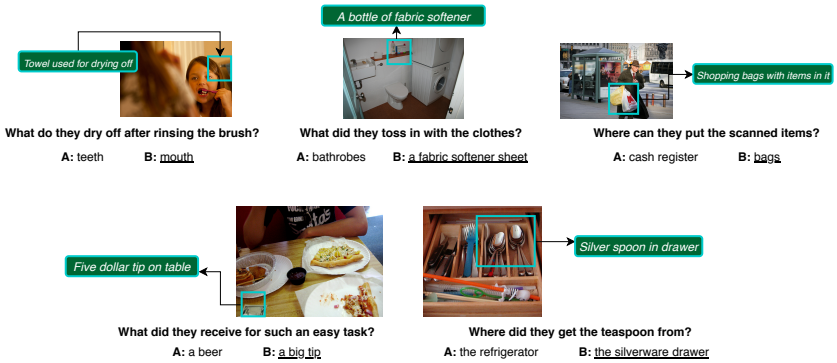

Figure 3: Examples of questions from the unanswerable set and one of the manually selected region descriptions from Visual Genome.

of the given question's text passage. Finally, we fine-tuned the model just as we did with the baseline model.

The whole retrieval process was done manually, which did not represent much of a problem for the dev and test subsets. However, it would be time-consuming to follow this approach with the train set. We fine-tuned on the complete train data, but we limited the use of image descriptions to 225 train questions that were selected in the same way as the dev and test subsets.

5 Results

For most of the questions in the unanswerable set, we did find related region descriptions. Figure 3 shows some of the images retrieved and the regions that matched what the question is asking. Besides its size, one of the main advantages of Visual Genome annotations is that they cover several regions that compose the scene in an image. Thanks to this, we were able to find region descriptions that not only mention an object (e.g., *a towel, scissors, a dollar-bill*), but also add a description of how the object can be used (e.g., *towel used for drying off, scissors for cutting string*) or what does it represent (e.g., *five dollar tip on table*). This suggests that our hypothesis mentioned in Section 1 about annotations of visual input might be correct. As shown in Table 1, region descriptions helped BERT to achieve a better

Model	**Dev** Commonsense	**Test** Commonsense
fine-tuned BERT (base-uncased)	.780	.732
Visually Enhanced BERT	.857	.749

Table 1: Accuracy of BERT baseline and our manually visually enhanced BERT in both MCScript2.0 development and test sets. The results come from three different random seeds.

accuracy. If our hypothesis is true, the improvement should come from correctly answering questions from the unanswerable set. This was true for those related to affordances.[2] Some examples of questions that became answerable for Visually Enhanced BERT are *What did they toss in with the clothes?*, and *What do they cut out the pieces with?*. Another type of question BERT initially had problems answering required commonsense knowledge about an object's location. Some examples of those questions are *Where did they get the teaspoon from?* (Answer: the silverware drawer) and *Where did they get the paper plate from?* (Answer: the kitchen). Our results suggest that region descriptions were more beneficial to these type of questions, since they were no longer unanswerable for Visually Enhanced BERT. However, there were cases in which we could not see an improvement. Questions like *What did they receive for such an easy task?* (Answer: big tip) and *What does a list keep them on?* (Answer: budget) do require commonsense knowledge about the SERVING A DRINK and GOING SHOPPING scenarios, but the

[2] An object's properties that show the possible actions users can make with it.

concept that needs to be understood is too abstract. Even though we found region descriptions that match the correct answer candidate (e.g., *Five dollar tip on table. Tip on the table.*), these type of questions remained unanswerable for Visually Enhanced BERT. See Appendix B for more examples.

In a classic reading comprehension task, word matching usually helps to find the correct answer. However, MCScript2.0 evaluates beyond mere understanding of the text and as such, it was designed to be robust against it. Out of the 56 questions in our dev set, we observed that the number of times a passage mentions the correct and the incorrect answer candidates is similar (42 and 36, respectively) and in either case this seemed to have influenced BERT's predictions. This stayed roughly the same after we appended the region descriptions.

6 Conclusion and Future Work

Pre-trained large LMs have significantly closed the gap between human and computer performance in a wide range of tasks, but the commonsense knowledge they capture is still limited. In this work, we presented a controlled experimental setup to explore the plausibility of acquiring commonsense knowledge from dense image descriptions. Our preliminary results on a commonsense-MRC task suggest that such descriptions contain simple but valuable information that humans naturally build through experiencing the world. In future work, our aim is to automate the retrieval process and explore better ways of using region descriptions than the presented approach of modifying BERT's input format.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR1513, Japan. We thank the anonymous reviewers for their insightful comments and suggestions.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Hugo Liu and Push Singh. 2004. Conceptnet: a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Marvin Minsky. 2007. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster.

Nelson Mukuze, Anna Rohrbach, Vera Demberg, and Bernt Schiele. 2018. A vision-grounded dataset for predicting typical locations for verbs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *arXiv preprint arXiv:1803.05223*.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. Mcscript2. 0: A machine comprehension corpus focused on script events and participants. *arXiv preprint arXiv:1905.09531*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Mark Yatskar, Vicente Ordonez, and Ali Farhadi. 2016. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

A Appendix A. Implementation

A.1 Input format

We fine-tuned a vanilla-BERT with the following input configuration: the question and one of its answer candidates are appended to segment one and the text passage is appended to segment two. Therefore, we have two inputs per instance. To help BERT differentiate between the question and answer-candidate tokens, we used a special separator token[3]. The maximum sequence length was set to 384. We trained the model up to 5 epochs with a learning rate (Adam) of 5e-5 and a training batch size of 8 using 3 different random seeds.

B Appendix B. Input and output examples

Figure 4 shows how we build the input representation of two MCScript2.0 questions. The question and answer candidate A in segment one, and the text passage in segment two. Similarly, there is a second input representation with the question and answer candidate B in segment one, and the text passage in segment two. BERT computes a softmax over the two choices to predict the correct answer candidate. Visually Enhanced BERT builds the input in a similar way. The difference is that the manually selected region descriptions are appended at the beginning of the text passage. The number of tokens in the text passage increases, but the input configuration remains the same.

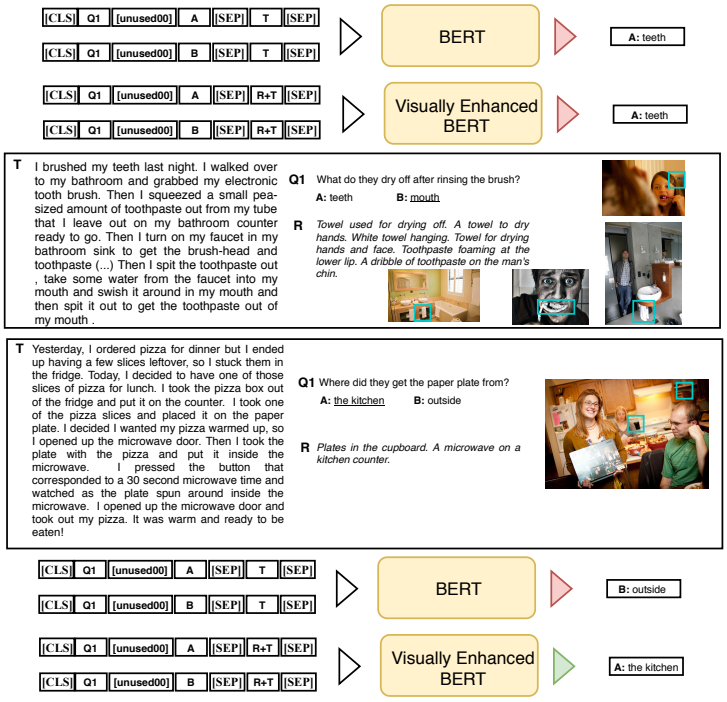

Figure 4: Two input/output examples. In the top example, region descriptions were not helpful to chose the correct answer candidate. In the bottom example, they were.

[3]We used '[unused00]' as the special separator token, which is included in BERT's vocabulary