Visual Choice of Plausible Alternatives: An Evaluation of Image-based Commonsense Causal Reasoning

Jinyoung Yeo POSTECH
Gyeongbok Lee* Yonsei University
Gengyu Wang* Yonsei Unversity
Seungtaek Choi Yonsei Unversity
Hyunsouk Cho POSTECH
Reinald Kim Amplayo Yonsei University
Seung-won Hwang Yonsei University
{jinyeo, prory}@postech.edu {alias_n, posuer, hist0613, rktamplayo, seungwonh}@yonsei.ac.kr

Abstract

This paper proposes the task of Visual COPA (VCOPA). Given a premise image and two alternative images, the task is to identify the more plausible alternative with their commonsense causal context. The VCOPA task is designed as its desirable machine system needs a more detailed understanding of the image, commonsense knowledge, and complex causal reasoning than state-of-the-art AI techniques. For that, we generate an evaluation dataset containing 380 VCOPA questions and over 1K images with various topics, which is amenable to automatic evaluation, and present the performance of baseline reasoning approaches as initial benchmarks for future systems.

Keywords: Commonsense knowledge, Causality, Reasoning, Evaluation, Image understanding

1. Introduction

Commonsense causal reasoning is one of the fundamental research problems in Knowledge Representation & Reasoning (KR) domain. It aims at understanding the general causal dependency between common events or actions. Recent efforts for such understanding are focused on measuring the plausibility of one event statistically leading to another, and in particular, are competing on an evaluation set called *Choice of Plausible Alternatives* (COPA) (Roemmele et al., 2011), which is to select the more plausible alternative as a cause (or effect) of the premise as:

Example 1 *Premise:* A janitor is cleaning the floor. *What is cause?*
Alternative 1: There is a broken cup on the floor.
Alternative 2: There is a cup of coffee on the table.

For the purpose of this reasoning, the state-of-the-art, called CausalNet (Luo et al., 2016), harvests causality scores of cause-effect term pairs, *e.g.*, ('broken','clean'), by mining their causal patterns, *e.g.*, "If...broken ..., then... clean...", from an extremely large text corpus (10TB). As a result, it achieved a remarkable accuracy (70.2%) from COPA.

In the real world, the desirable reasoning ability should be required not only in KR domain but also in Computer Vision (CV) domain. Toward the optimal goal of human-level intelligence, CV researchers have actively studied on Visual/Video Question Answering (VQA) (Antol et al., 2015; Ye et al., 2017; Zhao et al., 2017), which is to understand textual questions and images and give correct textual answers by machine. However, such QA tasks are at an early stage of "reasoning", being limited to object-level reasoning, *e.g.*, "What is the man holding in his hand?".

Beyond object-level, the *event-level* visual reasoning is at intersection between the top of KR and CV capabilities, as the boundaries between the two domains are crumbling down due to the huge success of neural-based image-to-text or text-to-image converting techniques (Karpathy et al., 2014; Jiang et al., 2017; Vinyals et al., 2017). We argue these two domains are in complementary nature (Aditya, 2017; Aditya et al., 2015). For example, unlike existing learning of end-to-end signal matching (*e.g.*, image-to-object) in CV, commonsense and background knowledge in KR can help rectifying noise in visual inference. Also, unlike existing language-specific reasoning in KR, visual detection and captioning can help realizing more general reasoning scenarios.

Figure 1: An example of VCOPA question. If a premise image is effect, plausible alternatives image should be cause, and vice versa. Red mark indicates correct answers.

In this paper, we thus propose a new reasoning task and its evaluation dataset, called Visual COPA (VCOPA) as a variant of COPA, which covers the visual questions for commonsense causal reasoning. Specifically, Figure 1 illustrates an example of VCOPA question, which is converted from a textual question on COPA (Example 1) into its corresponding visual question with three images. Similar to COPA, given a premise image, the goal of the VCOPA task is to identify a more plausible alternative image. As a public dataset including over 1K images, VCOPA is amenable to automatic quantitative evaluation, making it possible to effectively track progress on this task.

As a baseline, we leverage CausalNet by reducing image-based questions to text-based question with the state-of-the-art neural image captioning technique (Vinyals et al., 2017). Although this approach cannot achieve significantly better performance when compared to the random baseline

*Authors in alphabetical order with equal contribution

of 50% accuracy, it can be a starting point on the task for the multiple communities such as KR and CV. As this guidance, VCOPA poses a rich set of challenges, many of which have been viewed as the holy grail of automatic image understanding and causal reasoning in general. However, it includes several components that the KR and CV communities have made significant progress on during the past few decades. Thus, we provide an attractive list of solution techniques accessible enough for the communities to start making progress on the VCOPA task.

2. VCOPA Dataset Collection

This section presents the Visual Choice of Plausible Alternatives (VCOPA) dataset by describing our process of collecting image (and text) questions.

2.1. Image Collection

The VCOPA task consists of 380 questions of commonsense causality with 1,140 images. The image question set was created using a specific collecting methodology that ensured breadth of topics on images and quality of the questions. We now explain the details.

Similar to COPA, the first major concern of the collecting methodology is the breadth of the image question set. Our approach is to identify question topics as inspired by COPA questions, which is already validated with a high degree of breadth, and then apply these topics to collecting premise and alternative images through our own creativity. This approach helps balance the generative and analytic aspects of this task, ensuring that the skewed topic interests of the image collectors are not over-represented in the question set, but still allowing for the creative design solutions that each of these questions required.

More specifically, as shown in Example 1 and Figure 1, we first try to directly convert the textual question to its semantically equivalent image question. As a result, we made 224 image questions, which can be compared with their corresponding 224 COPA textual questions. The rest of the 776 COPA questions is difficult to convert to images. Below is an example:

Example 2 *Premise:* The engine of the airplane was faulty. *What is effect?*
Alternative 1: The airplane crashed.
Alternative 2: The pilot made an error.

In the above example, although *Alternative 1* can be easily visualized to an image, *Alternative 2* is hard to visualize using a single image. In this case, we leverage our own creativity to make a new question, taking into account the topic inspired by COPA as possible as we can. To illustrate, Figure 2 recasts Example 2 into a different problem while capturing ideas about 'airplane'. This process made 276 image questions, which is not completely matched with COPA textual questions.

A challenging part of designing VCOPA questions is to establish the incorrect alternative for each question. This image is intended to be similar in form to the correct alternative image, and somewhat related to the premise image, but with no obvious causal connection, especially by eight

Figure 2: VCOPA question recasted from Example 2

main challenges, as we discuss later in Section 4. This design is intended to ensure that answering these questions requires both computer vision techniques and commonsense knowledge harvesting, and cannot be easily answered when using an individual technique. As a result, we made 500 questions, each of which has one premise image and two alternative images.

The second major concern of the collecting methodology is the quality of questions, that is, the strong agreement among human raters who were asked to answer each question. To validate the set, we enlisted the help of 10 volunteers, each of which validates the overall set of 500 questions. Agreement between raters was high (Cohen's K=0.942). In all, at least one volunteer answered 120 questions differently than was intended by the collector of the question. We strictly remove these 120 questions from the set, then the final set contained 380 questions, each of which has the perfect agreement among raters. Especially, 148 questions in this question set are interchangeable with COPA questions with the same meaning.

The order of the question set collected by the image collectors is randomized and the position of the correct alternative image is also randomized, ensuring that a random baseline would answer exactly 50% of the questions correctly.

2.2. Text Annotation

The VCOPA task focuses on causal reasoning with only images (and their metadata automatically generated by machine, *e.g.*, image captions) as input. Despite this fact, we aim at supporting other research scenarios, for example, mutlimodal questioning with image and human-generated text, *i.e.*, the combination of Figure 1 and Example 1. For this purpose, while the selected 148 questions are matched with their corresponding COPA questions, we generate the COPA-like text questions for the rest of questions. For example, the VCOPA question in Figure 2 is matched with a COPA-like question as follows:

Example 3 *Premise:* A plane is landing at the airport. *What is effect?*
Alternative 1: The plane is burning on the ground.
Alternative 2: The ship is burning on the ocean.

When generating text questions, we basically follow the authoring guidance in (Roemmele et al., 2011). For this task, six volunteers, mutually exclusive to the original image collectors, are asked to generate the texts, to exclude the subjectivity of intending causality context. Other than the supplementary files, the VCOPA dataset including visual and textual questions is accessible in our github site[1].

[1] https://github.com/antest1/VCOPA-Dataset

3. VCOPA Task Analysis

3.1. Reasoning Baselines

The VCOPA evaluation is designed so that a random baseline system, where one of the two alternative images is randomly chosen for each question, would perform at exactly 50%. In addition, we adopt a somewhat stronger baseline based on the state-of-the-art causal reasoning system in language domain, and investigate its reasoning performance. While we do not expect these baselines to be competitive with future sophisticated approaches in the CV and KR domains, successful systems must demonstrate improvements over these baseline results.

Our baseline approaches explore the simple idea that visual causality can be converted into textual causality by automatic image captioning techniques, as the causal inference achieved a high performance on COPA task. Accordingly, one would expect that causality between words in the captioning sentences captures image causality as well. Let L_p, L_{a_1}, and L_{a_2} be each automatic captioning sentence of p, a_1, and a_2, by a state-of-the-art system (Vinyals et al., 2017). Then, the more plausible alternative a^* can be identified as:

$$\begin{aligned} a^* &= \underset{a\in\{a_1,a_2\}}{\text{argmax}}\ \text{plausibility}(p, a) \\ &\approx \underset{a\in\{a_1,a_2\}}{\text{argmax}}\ \text{plausibility}(L_p, L_a) \\ &= \underset{a\in\{a_1,a_2\}}{\text{argmax}}\ \frac{1}{|L_p|+|L_a|}\sum_{t_i\in L_p}\sum_{t_j\in L_a} CS(t_i,t_j) \end{aligned} \tag{1}$$

where $CS(t_i, t_j)$ is a causality score between a cause term t_i and an effect term t_j extracted from CausalNet.

CausalNet (Luo et al., 2016) is a weighted and directed graph $G(L, E, W)$ with nodes (lemmatized English terms) $L = \{t_1, t_2, ...\}$ and edges (causal relations) E. The edge weights are captured by the function $W : E \rightarrow [0, 1]$. The weight $w_{i,j}$ associated with an edge (t_i, t_j) represents the causality score, denoted as $CS(t_i, t_j)$, of a cause t_i and an effect t_j. Causality scores depend on the number of occurrences that two terms t_i and t_j are in linguistic patterns known as causal cues (Chang and Choi, 2004) identifying precise cause/effect roles, *e.g.*, "If...t_i ..., then... t_j..." and "... t_j..., because...t_i ...". That is, as more occurrences of (t_i, t_j) in causal cues, its causality score is higher as:

$$W : w(t_i, t_j) = CS(t_i, t_j) \propto freq(t_i, t_j) \tag{2}$$

where $freq(t_i, t_j)$ is the frequency of observing the causal pair (t_i, t_j) from an English corpus. We omit the details of the list of causal cues and Eq. 2 and refer the readers to (Luo et al., 2016).

Despite building on a rather simple and shallow text analysis, by leveraging the scale and richness from a extremely large (10TB) text corpus, CausalNet achieves the state-of-the-arts accuracy on COPA tasks. The corpus contains 1.6B web pages, which result in 64,436 nodes in CausalNet.

3.2. Reasoning Results

Table 1 shows the reasoning performance on COPA and VCOPA evaluation. As reported in (Luo et al., 2016), although CausalNet formally achieved 70.2% accuracy on COPA evaluation, its accuracy on the overlapped set between COPA and VCOPA is 67.9%. Compared to this accuracy, replacing the manual annotation by the automatic caption gains much lower accuracy 52.7%.

Table 1: Reasoning performance on VCOPA evaluation

Dataset	Method	Accuracy
COPA ∩ VCOPA (148 questions)	Automatic Caption	52.7
	Manual Annotation	67.9
Test Set (190 questions)	Automatic Caption	54.2
	Manual Annotation	56.3
Dev Set (190 questions)	Automatic Caption	53.2
	Manual Annotation	55.8
Test + Dev (380 questions)	Automatic Caption	53.7
	Manual Annotation	56.1

In the overall VCOPA dataset, we compare using machine-generated captions with using human-generated annotations again. As a result, although using the captioning technique achieves a better accuracy than the random baseline, we find that using automatic captions cannot outperform the manual annotations. Note that as the VCOPA questions are designed to infer visual causality not textual causality, the average accuracy of their manual annotation is lower than that in COPA while the accuracy of automatic caption is consistent among all data divisions. Despite this trend, manual annotation is consistently better than automatic caption.

These results also suggests the limitation of our baseline approach, that is, using only textual information cannot achieve the high accuracy in VCOPA evaluation regardless of automatic and manual texts. We pose the challenges in VCOPA in the next section.

4. VCOPA Challenges

Figure 3 illustrates the list of challenges and their future work in VCOPA, which we enumerate as follows:

(a) Visual disambiguation. VCOPA dataset contains many image triples that are visually ambiguous. For example, in Figure 3(a), the two alternative images are blurred, however *Alternative 1* is more plausible than *Alternative 2* because the former blur corresponds to 'smog' while the latter corresponds to 'fog'. Due to the unambiguous nature of the ImageNet dataset (Deng et al., 2009), current object recognition systems do not consider visual ambiguity. In (Gella et al., 2016), they used multimodal embeddings by leveraging captions to disambiguate the visual senses. However, this would require very informative captions in order to be effective.

(b) Temporal disambiguation. The dataset also includes ambiguity on the chronological sense. For example, in Figure 3(b), both alternatives are strongly correlated to the premise. *Alternative 1* seems to happen before the premise, while *Alternative 2* happens after the premise. However, the question is asking for the effect, which means the latter alternative is the correct answer. This entails that simple correlation is not enough to find the solution. One possible solution is to use visual storytelling machines (Ferraro et al.,

2016) which describes images in sequence in order to determine if the current sequence makes sense.

(c) Fine-grained object recognition. Causal reasoning requires object recognition in the fine-grained level. For example, in Figure 3(c), an object recognition system should be able to recognize that the car in the premise image is a police car and the yellow tape in the first alternative is a police line in order to do reasoning. There are few systems which are able to detect distinct features of fine-grained objects by picking deeper filters (Zhang et al., 2016) and by localizing parts of the image (Wei et al., 2016).

(d) Event recognition. A good system should also be able to recognize events based on the identified objects. For example, in Figure 3(d), all images contain two objects (*e.g.*, a ball and a player), however all three images correspond to different events (*e.g.*, the premise image is a scene of a player kicking a ball). Recent works on event grounding (Cho et al., 2016) might help in recognizing these events.

(e) Inter-event relationship. Relationship between the events found on the pairs of images should also be considered. For example, in Figure 3(e), there is a relationship between the 'cleaning' event in the premise and the 'broken glass' event in the first alternative, even though they do not share the same objects.

(f) Event-sentiment relationship. Knowing the sentiment shown in the image can also help in causal reasoning. For example, in Figure 3(f), since the sentiment of the premise leans to negative, the second alternative, which has the same sentiment, is more plausible. Image sentiment analysis (You et al., 2015) is a growing sub-field of computer vision, and we expect solutions for VCOPA to also employ techniques used here.

(g) Inter-sentiment relationship. The relationship between the sentiments of two images is also important. Moreover, detecting more fine-grained *emotions* (Abdul-Mageed and Ungar, 2017), such as 'excited' and 'relieved', can help in causal reasoning. For example, in Figure 3(g), although all the images are showing smiling faces, fine-grained emotion detection tells us that the second alternative is 'scary', and therefore is incorrect.

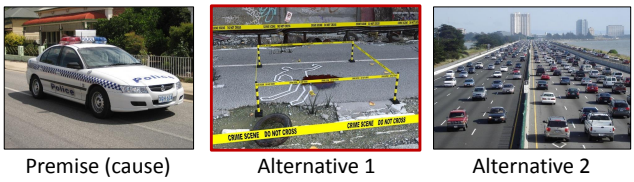
(h) Commonsense knowledge. Beyond scene understanding at object, event, and sentiment levels, commonsense causal reasoning inherently requires commonsense knowledge, which is a non-visual dimension, such as "people prefer white wine to red wine with seafood", which entails that *Alternative 2* is the answer in Figure 3(h).

5. Conclusion

We introduce the task of visual commonsense causal reasoning with VCOPA evaluation dataset. Given a premise image and two alternatives as cause or effect, the task is to provide a more plausible answer with causality context. We provide the VCOPA dataset containing 380 questions of diverse variety on domains with over 1K images. We believe VCOPA has the distinctive advantage of pushing the frontiers on "multi-discipline" problems, while being amenable to automatic evaluation. A promising future work is to automatically harvest the VCOPA image triples to construct a large-scale image database for neural-based visual commonsense causal reasoning.

(a) Visual disambiguation

(b) Temporal disambiguation

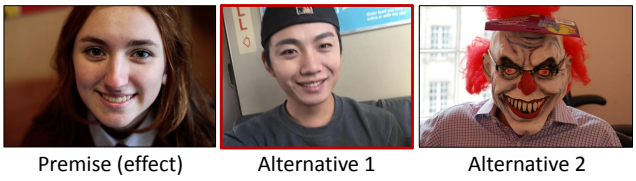


(c) Fine-grained object recognition

(d) Event recognition

(e) Inter-event relationship

(f) Event-sentiment relationship



(g) Inter-sentiment relationship

(h) Commonsense knowledge

Figure 3: Challenges in VCOPA

6. Acknowledgement

This work was supported by Microsoft Research, and Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01778,Development of Explainable Human-level Deep Machine Learning Inference Framework). S. Hwang is a corresponding author.

7. References

Abdul-Mageed, M. and Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *ACL*.

Aditya, S., Yang, Y., Baral, C., Fermuller, C., and Aloimonos, Y. (2015). Visual common-sense for scene understanding using perception, semantic parsing and reasoning. In *AAAI Symposium*.

Aditya, S. (2017). Explainable image understanding using vision and reasoning. In *AAAI*.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *ICCV*.

Chang, D.-S. and Choi, K.-S. (2004). Causal relation extraction using cue phrase and lexical pair probabilities. In *ICON*.

Cho, H., Yeo, J., and Hwang, S.-W. (2016). Event grounding from multimodal social network fusion. In *ICDM*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.

Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., et al. (2016). Visual storytelling. *arXiv preprint*.

Gella, S., Lapata, M., and Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint*.

Jiang, L., Kalantidis, Y., Cao, L., Farfade, S., Tang, J., and Hauptmann, A. G. (2017). Delving deep into personal photo and video search. In *WSDM*.

Karpathy, A., Joulin, A., and Li, F. F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*.

Luo, Z., Sha, Y., Zhu, K. Q., Hwang, S.-w., and Wang, Z. (2016). Commonsense causal reasoning between short texts. In *KR*.

Roemmele, M., Bejan, C. A., and Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Symposium*.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*.

Wei, X.-S., Xie, C.-W., and Wu, J. (2016). Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *arXiv preprint*.

Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., and Zhuang, Y. (2017). Video question answering via attribute-augmented attention network learning. *arXiv preprint*.

You, Q., Luo, J., Jin, H., and Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*.

Zhang, X., Xiong, H., Zhou, W., Lin, W., and Tian, Q. (2016). Picking deep filter responses for fine-grained image recognition. In *CVPR*.

Zhao, Z., Yang, Q., Cai, D., He, X., and Zhuang, Y. (2017). Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*.