TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 00 (2021) 1–20

ISSN: 1756-8765 online
DOI: 10.1111/tops.12584

This article is part of the topic "Cognition-Inspired Artificial Intelligence," Daniel N. Cassenti, Vladislav D. Veksler and Frank E. Ritter (Topic Editors).

Knowledge Gaps: A Challenge for Agent-Based Automatic Task Completion

Goonmeet Bajaj,[a] Sean Current,[a] Daniel Schmidt,[b] Bortik Bandyopadhyay,[a] Christopher W. Myers,[b] Srinivasan Parthasarathy[a]

[a]*Department of Computer Science and Engineering, The Ohio State University*
[b]*Air Force Research Laboratory*

Received 8 January 2021; received in revised form 11 October 2021; accepted 11 October 2021

Abstract

The study of human cognition and the study of artificial intelligence (AI) have a symbiotic relationship, with advancements in one field often informing or creating new work in the other. Human cognition has many capabilities modern AI systems cannot compete with. One such capability is the *detection*, *identification*, and *resolution* of knowledge gaps (KGs). Using these capabilities as inspiration, we examine how to incorporate *detection*, *identification*, and *resolution* of KGs in artificial agents. We present a paradigm that enables research on the understanding of KGs for visual-linguistic communication. We leverage and enhance and existing KG taxonomy to identify possible KGs that can occur for visual question answer (VQA) tasks and use these findings to develop a classifier to identify questions that could be engineered to contain specific KG types for other VQA datasets. Additionally, we examine the performance of different VQA models through the lens of KGs.

Keywords: Artificial Intelligence; Cognitive Science; Computer Science; Computer Vision; Intelligent agents; Computational Cognitive Modeling; Neural Networks

This work was done when author Daniel Schmidt was at AFRL.

Bortik Bandyopadhyay is currently employed at Apple. This work was done when Bortik Bandyopadhyay was a student at The Ohio State University.

Correspondence should be sent to Goonmeet Bajaj, Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave, Columbus, OH 43210, USA. E-mail: bajaj.32@osu.edu

1. Introduction

The formal studies of human cognition and artificial intelligence (AI) are deeply entwined. Through research on understanding human and animal learning, Rescorla and Wagner (1972) developed a mathematical formula that has become the bedrock for reinforcement learning algorithms in machine learning (Sutton & Barto, 1998). Similarly, groundbreaking research on biological neural networks and their capacity to compute (cf., McClelland & Rumelhart, 1981; McCulloch & Pitts, 1943) has provided the foundation for machines to classify patterns and select actions within highly complex games (cf., Silver et al., 2017). Indeed, it is difficult to overstate the impact cognitive science research has had on AI.

Not only have the studies of human cognition and the processes governing human behavior informed AI, but research within AI has helped facilitate theories of human cognition. For example, theories that human cognition is optimal given the constraints on the cognitive system, or *boundedly optimal*, are often evaluated through the use of machine learning techniques that can provide provably optimal policies given human cognitive constraints (Acharya, Chen, Myers, Lewis, & Howes, 2017; Lieder & Griffiths, 2020). The derived optimal policies are then compared against human behavior to determine if humans approximate the optimal policies.

Continuing in this vein, research has focused on the development of machines that can be taught new information, either through interaction or instruction (Gluck & Laird, 2018; Kupitz et al., 2021). In both cases, new knowledge required for completing a specified task is provided to the intelligent system, agent, or model. Such approaches leverage discoveries and capabilities from AI and the cognitive sciences with demonstrable successes (Eberhart et al., 2020; Kirk, Mininger, & Laird, 2016; Li et al., 2019; Salvucci, 2021). Unfortunately, omitting a step required to complete a task, leaving a step to be inferred, or insufficient prior knowledge may lead to a knowledge gap (KG), which we define as the deficiency in knowledge that prevents an intelligent system from completing the newly instructed task.

Research and development toward intelligent systems (human and machine, alike) capable of determining that knowledge is insufficient to achieve its specified goals and rectifying the insufficiency would mutually benefit AI and cognitive science and require discoveries and capabilities from each. To this end, research has been initiated to study and develop intelligent systems capable of detecting insufficient knowledge to achieve its goals and rectifying the insufficiency. We broadly refer to these capabilities as KG processes.

When faced with KGs, intelligent systems (again, humans and machines alike) must be capable of (a) detecting the existence of a gap, (b) identifying the KG type detected, and (c) resolving the gap through some available method. The cognitive mechanisms supporting human capabilities to detect and resolve KGs could be researched, understood, codified, and leveraged to establish mechanisms within artificially intelligent systems and models of human cognition to promote increased flexibility and robustness (Walsh, Einstein, & Gluck, 2013) when faced with imperfect or incomplete information (see Fig. 1). To this end, models of cognitive capacities and processes supporting the detection, identification, and resolution of KGs require further research and understanding.

In this work, we propose a research agenda focused on the abilities of humans and machines alike to detect and resolve gaps in knowledge. The research agenda involves five elements

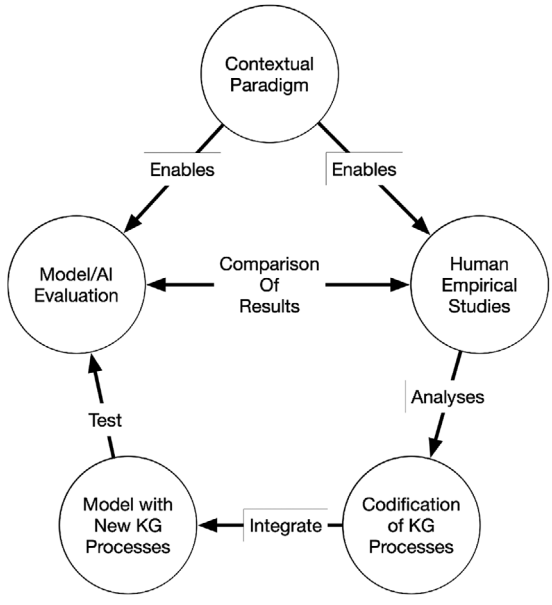


Fig. 1. Areas of research activity on detection and resolution process of knowledge gaps.

of investigation (not necessarily in order of importance nor are they necessary steps): (a) empirical studies with human participants to investigate when KG detection occurs and how gaps are resolved, (b) the codification of the identified KG detection and resolution processes, (c) the integration of the codified processes into AI systems and cognitive models, and (d) evaluating AI and cognitive models with integrated KG processes. The fifth element, (e), and the remaining focus of the paper, is the development of a paradigm that enables the rigorous evaluation of KGs in human empirical studies as well as evaluating KG processes integrated with cognitive models and other AI systems.

Contributions from a research agenda focused on KG processes are fourfold. First, we present the concept of KGs with their hypothesized associated cognitive capacities (see Table 1). Second, a paradigm for evaluating the KG processes (i.e., detection, identification, and resolution) of humans, cognitive models, and other AI systems is presented. To this end, we leverage the *visual question answering* (VQA) task as our evaluation paradigm, as it sits at an intersection of three components of AI and human cognition: *language*, *vision*, and *reasoning*. Third, a simple classifier for identifying VQA questions that can be manipulated into specific KG types for VQA is developed and presented. Fourth, we evaluate existing VQA machine learning algorithms across different types of KGs.

2. Knowledge gaps and related processes

It is an unreasonable and impractical expectation for a person, or any other intelligent system, to have complete knowledge of everything. Rather, knowledge is accumulated over an

Table 1
Taxonomic mapping of gaps, cognitive capacities, and brief descriptions. KGs in bold are used for our KG-VQA Paradigm. A machine readable version of the taxonomy is available at https://github.com/schmidtDTN/General-Taxonomy-Graph.git

Gap Type	Related Cognitive Capacities	Brief Knowledge Gap Description
Lexical	Language, Reasoning, Decision-Making	A term outside of an agent's vocabulary is received.
Word-Sense	Language, Reasoning	A word with multiple meanings cannot be disambiguated.
Activity	Perception	The specific action upon an object or person cannot be recognized.
Material	Perception	The composition of an object is not clear (subtype of Attribute Gap).
State	Perception, Categorization	The specific condition of an object or person cannot be understood (subtype of Attribute Gap).
Location	Spatial, Perception, Attention	The physical location of a setting cannot be determined accurately.
Reasoning	Reasoning, Memory	There is insufficient knowledge to draw conclusions from a set of premises.
Attribute	Perception	The feature(s) of an item or object are unknown.
Sentiment	Language, Emotion	The sentiment of a statement cannot be interpreted.
Target	Language, Reasoning, Perception	The target of a phrase or instruction is unclear.
Context	Spatial, Perception, Memory	Seemingly disjoint information is provided.
Direction	Spatial, Perception, Attention	An agent is unsure of its position in space.
Size	Spatial, Perception, Attention	The size of an object or distance cannot be determined (subtype of Attribute Gap).
Entity Resolution	Perception, Attention	The presence of an object cannot be determined.
Explanatory	Reasoning, Categorization	An agent can perform accurately but not explain how or why.
Memory	Memory, Attention	Relevant information is inaccessible through forgetting. (Levine, 1983)

individual's lifetime of experience that can then be applied to tasks (Spelke, 2017). Additionally, knowledge required for completing novel tasks can be explicitly provided through instruction. Instructions often build on previously acquired knowledge to situate novel information and steps required for the completion of novel tasks. However, prior knowledge and new knowledge from instructions may be incomplete, producing a deficiency in an intelligent system's knowledge (declarative or procedural) required for completing a task or goal, resulting in one or more KGs.

Unfortunately, there is relatively limited published research focused specifically on KG processes. To overcome KGs, they must first be detected. Indeed, how often, and under what

conditions, humans are capable of detecting KGs is poorly understood at this point. Nonetheless, once detected, KG resolution is similar to human problem-solving, where problem-solving very generally requires the selection of an applicable method based on the task representation (Newell & Simon, 1972; Reed & Vallacher, 2020). The detection and resolution of KGs are related to problem-solving in important ways. First, KGs can occur from incomplete relevant semantic information associated with the problem (i.e., incorrect information or information gaps in the task representation) or the absence of actions available to apply to the problem. Second, like problem-solving, selecting method(s) to resolve KG(s) will depend on the representation of the task. Third, having resolved a KG, the newly acquired information/actions must be incorporated into the existing knowledge base for immediate use, future use on a similar task, and continuous learning across diverse tasks.

Indeed, the *impasse* problem-solving mechanism within the Soar cognitive architecture (Laird, 2012), based on human problem-solving research, provides one example of a KG resolution process. Within the Soar architecture, when the system is in a state where a decision between two or more operators cannot be applied, an impasse is encountered and the system establishes a subcontext in an attempt to identify knowledge that can overcome the previous impasse. Unfortunately, in instances where the required knowledge remains unavailable, a KG persists and failures may occur, impeding the system's ability of completing the task.

Considerable research has been conducted on whether missing information is critical for decision-making and when one should move on from collecting more information to making a decision. For example, a gap in knowledge could lead to equivalent or greater success than if the gap were filled, as in the case of *satisficing* (Simon, 1957) and heuristic-driven behavior and decision-making (Gigerenzer & Todd, 1999; Marewski & Schooler, 2011). Further, determining when to continue acquiring information versus proceeding on with the task or making a decision, or *information foraging* (Pirolli & Card, 1999) has demonstrated that humans are good at determining when to move on to a new source for information. Indeed, such a balance between seeking out knowledge to fill gaps and acting on the available knowledge must be balanced in the KG resolution process to prevent an agent from continuously searching for knowledge.

Others have theorized that humans should provide the KG capabilities for machines. Chandrasekaran, Yadav, Chattopadhyay, Prabhu, and Parikh (2017) argue that humans should develop a theory of a machine's "mind" to increase the effectiveness of human-AI teaming. They found that with only few examples, lay people can be trained to better understand predicted responses and future failures of a complex AI system. Similarly, Nushi, Kamar, and Horvitz (2018) propose a set of hybrid human–machine methods and tools for describing and explaining system failures. Their methods use both human and system-generated observations to summarize conditions of system malfunction with respect to the input content and system architecture. The methods are designed to predict the probability of failure given the input. Similarly, to address the issue of computer vision systems failing abruptly without warning or explanation, Zhang, Wang, Farhadi, Hebert, & Parikh (2014) explored an approach of evaluating the input itself. The authors found that the ability to detect the likelihood of correctly classifying an input prior to processing the input completely may provide an initial mechanism for detecting when and identifying how image processing may fail.

It is likely that there are different types of KGs, and that not all KGs are resolved using the same methods. For example, a KG due to an unknown lexical term could be resolved using a dictionary, WordNet (Miller, 1995), or a thesaurus. However, a KG associated with the spatial arrangement of information in a scene may require visuospatial processes to resolve. Determining which method to apply to resolve a detected KG can be informed from the current task context (*a la* problem-solving) as well as from the type of gap and its associated cognitive capacities. To begin addressing the detection and resolution of KGs, a set of potential KG types and their associated cognitive capacities is proposed to provide intelligent systems with clues on how the gaps can be resolved as well. A taxonomy of KGs and their associated capacities can also help researchers to focus on particular gaps or be informed on what types of gaps might arise as they research and develop particular cognitive capacities. See Bajaj et al. (2020) and Schmidt (2020) for an initial taxonomic representation of KGs. In Table 1, we expand our understanding of this initial representation by providing a mapping between the gaps that have been identified (Bajaj et al., 2020) and the cognitive processes or capacities we hypothesize to be involved in their resolution.

Table 1 is incomplete, and will likely change as research progresses. For example, certain use cases of the taxonomy may include relationships between gap types and cognitive processes that are not related in the current taxonomy; in other cases, additional domain-specific gaps may be identified as subtypes of existing gap types or even as novel gap types.

The relationship between KGs and cognitive capacities presented in Table 1 is of benefit to the greater research community by providing a starting point from which to conduct further, targeted research into specific KG types, whether by identifying and incorporating new forms of gaps or new connections between gaps and cognitive processes. This mapping between KGs and cognitive capabilities was derived from prior work associated with developing large-scale cognitive models required to operate with human team members over extended periods of time, specifically, errors that arose within the models through the lack of language and task knowledge (Myers et al., 2019). Additionally, this taxonomic mapping can be applied in situations where knowledge integrity is essential, granting insight into what kinds of failures from an absence of knowledge could arise given the capacities of the agent and enabling improvements with this targeted knowledge.

3. A paradigm for evaluating knowledge gap detection, identification, and resolution

A paradigm is required for thoroughly and objectively evaluating the performance of humans, cognitive models, and AI systems on their abilities to detect KGs, identify the KG type, and resolve the identified KG. Such a paradigm for rigorously evaluating KG processes should (a) enable a researcher to identify and focus on specific KGs of interest, (b) provide the same set of stimuli to humans, cognitive models, and AI systems for evaluation and direct comparisons, and (c) be sufficiently complex yet tractable for intelligent systems to perform and require the use of KG detection, identification, and resolution processes. Related to such functionalities within a paradigm is the ability to control the distribution of stimuli given different KG types, as it enables researchers to better control human experiments using

the paradigm as well as understand how the training data affect the capabilities learned by AI and machine learning systems. Based on these desired requirements, the VQA paradigm was selected, as it sits at the intersection of three areas of cognitive science and AI: *language*, *vision*, and *reasoning*, thus making it an ideal task. In the following sections, we provide steps to transitioning the VQA paradigm from a machine learning task to a paradigm for rigorously evaluating KG processes. We refer to this paradigm as KG-VGA.

3.1. VQA paradigm

Within the VQA task, images and related information (e.g., scene graphs) are provided to algorithms to evaluate their ability to answer questions about image scenes. More formally, given an image and question pair, (I, Q), VQA systems are trained to provide an answer, A. Answers provided by algorithms performing VQA tasks are in the form of natural language responses or the selection of an option among alternatives. VQA is an active area of AI research with multiple ongoing visual-question challenges (e.g., VQA, TextVQA + TextCaps, VizWiz, and GQA) and different datasets (e.g., VQA 2.0 [Antol et al., 2015a], VCR,[1] GQA, OK-VQA,[2] KBVQA,[3] FVQA,[4] ConVQA,[5] and VQA Introspect[6]) that are used for measuring progress for VQA algorithms. VQA datasets contain questions about spatial information of objects, object attributes, or general scene understanding questions (see Fig. 2). Recent VQA datasets have focused increasingly on questions that rely on external knowledge or commonsense knowledge to reduce system dependency on memorized linguistic features and enhance visual understanding. Consequently, we can assume that all of the information needed to answer a question may not be readily available in the image or the agent's knowledge repository for these datasets. However, it is still unclear which knowledge capabilities are learned by or are built into existing VQA algorithms.

Traditional machine learning and deep-learning VQA algorithms methods consist of three primary steps (Kafle & Kanan, 2017): (a) extraction of image features (image featurization), (b) extraction of question features (question featurization), and (c) designing and implementing an algorithm that combines these features to produce an answer (see Fig. 3). However, the more recent transformer-based models that use unsupervised self-training are now considered as state-of-the-art methods for many computer vision tasks (Lu, Batra, Parikh, & Lee, 2019). The visual transformer architectures use a self-attention mechanism to learn the relationships between elements of a sequence. Self-attention transformer models consist of a two-stage training pipeline. The first stage (i.e., pre-training) is an unsupervised task that uses a large-scale dataset to learn initial weights. The second stage fine-tunes the pre-trained weights for a downstream task. However, neither traditional nor transformer-based VQA models have integrated KG processes. To the best of our knowledge, we are the first to study how VQA questions can produce KGs within intelligent systems to understand their knowledge capabilities.

We use our understanding of KGs in the VQA paradigm to examine the characteristics of VQA dataset(s) and the strength and limitations of current VQA models through the lens of KGs. The benefit of this is twofold: (a) We can assess the different types of KGs that can occur in VQA questions. (b) We can investigate how the model architecture affects which

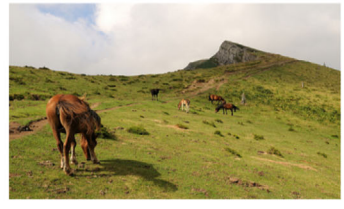
(a) **Q**: Are there any large mouse pads?
KG: Attribute, Entity Resolution, Size
Q: What kind of device is to the left of the cup on the right?
KG: Direction

(b) **Q**: Which are less healthy, the brownies or the cherries?
KG: Reasoning

(c) **Q**: What is the device that the happy man is holding?
KG: Sentiment

(d) **Q**: Is the large propeller blue and still?
KG: Attribute, Size, State

(e) **Q**: Which material makes up the white sink, porcelain or chrome?
KG: Material, Attribute



(f) **Q**: Where is the horse that looks white and brown walking?
KG: Attribute, Location

Fig. 2. Sample images and questions from the GQA dataset with knowledge gap tags.

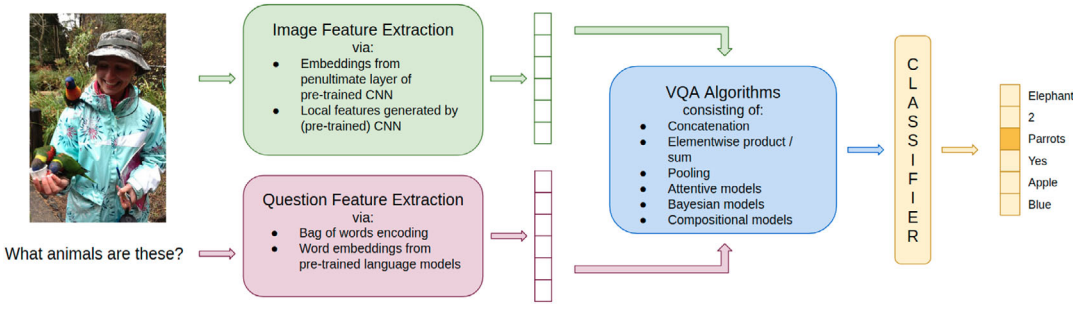


Fig. 3. The traditional visual question answering framework and pipeline (adapted from Kafle & Kanan, 2017).

types of questions an agent is able answer, thus removing possible KGs. The current VQA agents we examine do not have KG processes built into them; therefore, we only consider the number of correct and incorrect responses to evaluate their performance. Particular KGs can be introduced for VQA tests by altering the information available to an agent or the agent itself. Specifically, for the VQA paradigm, a KG can be introduced by removing or modifying important information from the input image or question. By altering the question, we can change the specific tasks the VQA agent has to complete to arrive at the correct answer,

resulting in the introduction of new KGs to the problem domain. In this work, we study two VQA datasets: GQA (Hudson & Manning, 2019) and VQA 2.0 (Antol et al., 2015a). We use the GQA dataset to formulate our problem of identifying possible KGs in the VQA domain and to enhance the existing KG taxonomy. We use the VQA 2.0 dataset to test our ability to automatically identify possible KGs when rich image and question metadata are not available.

3.2. GQA dataset

The GQA dataset (Hudson & Manning, 2019) consists of 22 million questions about various day-to-day images. There are 148,855 images in the dataset, 1878 possible answers, and a vocabulary of 3097 words. Questions in the dataset require multiple reasoning skills, including spatial understanding and multistep inferencing. The dataset is balanced by controlling the answer distribution for various collections of questions to limit educated guesses using language and world priors. The dataset contains images, scene graphs, questions, and object and spatial features. Scene graphs are machine readable image representations that can provide the semantic input used to train VQA models but are not sufficient to answer many VQA questions. For each image, the scene graph contains the image's objects, object attributes, and relations among other objects (see Fig. 4). Questions in the dataset are annotated with different characteristics (see Fig. 5). Section 3.4 contains an overview of how the characteristics outlined in color in Fig. 5 are used to identify potential KGs for questions. We work with the balanced training and validation dataset splits and direct the readers to the official GQA website for more information. Additionally, Fig. 2 contains sample image and question pairs from the GQA dataset and KGs assigned by our system.

3.3. VQA 2.0 dataset

The VQA 2.0 dataset (Antol et al., 2015b) is another popular dataset for VQA. The dataset contains open-ended questions about images that require an understanding of vision, language, and commonsense knowledge to answer. There are 204,721 images, 1,105,904 questions, and 11,059,040 ground truth answers. Questions in the VQA 2.0 dataset are not annotated with the same metadata as the questions in the GQA dataset. Additionally, images in VQA 2.0 are not accompanied with scene graphs (see the VQA 2.0 website https://visualqa.org for additional details and examples).

3.4. Analysis of knowledge gaps in the GQA Dataset

We refine the general-purpose KG taxonomy presented in Section 2 using the GQA dataset for the VQA paradigm (see bolded KGs in Table 1) to understand the types of questions in the GQA dataset and the cognitive capacities required to answer the questions. We first create a manual mapping to identify KGs for questions in the GQA dataset. Questions in the GQA dataset are annotated with rich metadata that also contain information about the image (see Fig. 5). In particular, we use the *global group*, *detailed type*, and *semantic filters* annotations associated with each question to assign KGs (see Fig. 5 for an example). The *global group* tag

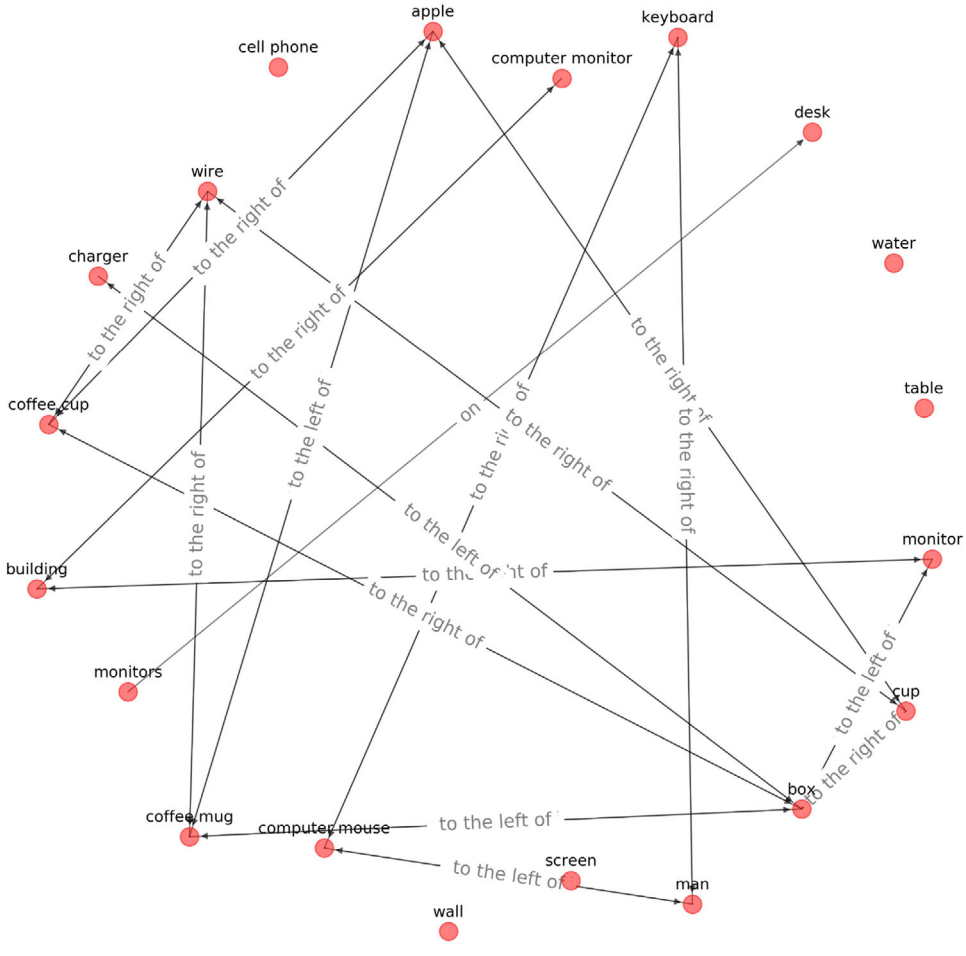


Fig. 4. Partial-sample scene graph (object attributes are not visualized).

provides the question with an answer type, such as "place" (Location) or "texture" (Material), while the *detailed type* tags the question with a specific reasoning step required for the answer, such as "existObj" (Entity Resolution) or "comparativeChoose" (Reasoning). The *semantic filter* details a specific attribute that should be used to differentiate between subjects in the image, such as "color" (Attribute), "height" (Size), or "facial expression" (Sentiment). This mapping allows for questions to be tagged with multiple KGs. The detailed mapping for question metadata fields to KGs and details about the KG identification method can be found in Bajaj et al. (2020) and a sample mapping is presented in Table 2.

After tagging questions with possible KGs, we observe a skew in the distribution of the number of questions per KG category in the GQA dataset. Fig. 6 displays the number of training questions per KG, respectively. The blue bar indicates the total number of questions tagged with a particular KG. The orange bar indicates the number of unique question

```
{'semantic':
     [{'operation': 'select', 'dependencies': [],
             'argument': 'mouse pad (4117948)'},
     {'operation': 'filter size', 'dependencies': [0],
             'argument': 'large'},
     {'operation': 'exist', 'dependencies': [1],
             'argument': '?'}],
     'entailed': ['14583226', '14583227',
             '14583228', '14583233', '14583232'],
     'equivalent': ['14583231'],
     'question': 'Are there any large mouse pads?',
     'imageId': '2328415', 'isBalanced': True,
     'groups': {'global': None,
             'local': '04-mouse pad_large'},
     'answer': 'yes',
     'semanticStr': 'select: mouse pad
             (4117948)->filter size:
          large [0]->exist: ? [1]',
     'annotations': {'answer': {},
             'question': {'4:6': '4117948'},
     'fullAnswer': {'5:7': '4117948'}},
     'types': {'detailed': 'existAttr',
             'semantic': 'obj',
          'structural': 'verify'},
     'fullAnswer': 'Yes, there is a large mouse pad.'}
```


Fig. 5. Metadata for the question associated with the Image 2a in Fig. 2.

Table 2
GQA metadata tags to KG mapping

Knowledge Gaps	Detailed Types	Global Group Label	Semantic Filters
Activity	activity, activityWho, activityChoose	activity, sportActivity	choose activity, choose sportActivity, activity, sportActivity
Location	place, placeVerify, placeVerifyC, placeChoose, locationVerifyC, locationVerify	place, room, nature environment, urban environment, road	location, place, room

texts tagged with a particular KG. The GQA dataset primarily contains questions that inquire about spatial relations (Direction Gap), object attributes (Attribute, Size, or Material Gaps), or the existence of objects (Entity Resolution Gaps). We calculate the coefficient of determination and Kullback–Leibler divergence (Kullback & Leibler, 1951) ($r^2 = 0.999$, $KL = 0.00$) between the training and validation datasets to ensure that the test set approximates the training set distribution for the number of questions per KG type. The observed skew in the distribution of image-question stimuli across KGs is important for researchers to be aware of as they may impact what is learned through training or result in an undesired disproportionate amount of KG types provided to human participants. If necessary, the skew can be alleviated by generating new KG-specific questions or using stratified sampling. In the next section, we present a method to identify KGs for the VQA 2.0 dataset. Identifying KGs for questions

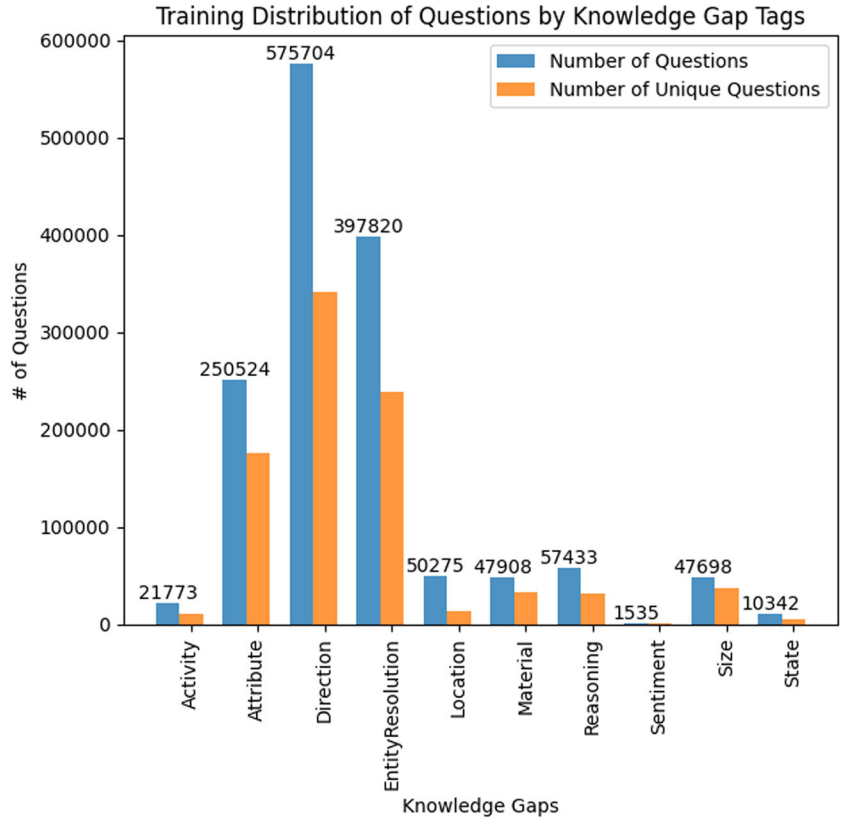


Fig. 6. Skew in the distribution of training questions per KG.

in different VQA datasets helps us understand how the training data impact what cognitive capacities are built into or learned by AI models.

3.5. Knowledge gap identification for VQA 2.0 questions

Questions in the VQA 2.0 dataset are not annotated with the same metadata as the questions in the GQA dataset; thus, we cannot directly use our GQA mapping approach to tag questions with KGs. Instead, we developed a classifier and apply self-training to tag questions with possible KGs that can occur when an agent tries to answer a question. We fine-tuned bidirectional encoder representations from transformer (BERT) (Devlin, Chang, Lee, & Toutanova, 2019) on the GQA dataset and adopted questions from the VQA 2.0 into the training data using self-training. BERT is a language model that is bidirectionally trained over a large text corpus. The transformer-based model takes as input a sentence and outputs context-aware vectors for each word of the sentence. Fine-tuned transformer models have achieved state-of-the-art results in many downstream natural language processing tasks such as question answering, named entity recognition, and language understanding (see Devlin et al., 2019, for further details

Table 3
Fleiss' Kappa per knowledge gap

Knowledge Gaps	Fleiss' Kappa
Activity	0.608
Attribute	0.944
Direction	0.905
Entity Resolution	0.940
Location	0.886
Material	0.854
Reasoning	0.903
Sentiment	0.958
Size	0.937
State	0.904

of BERT or the tutorial blog, Khalid, 2019). We train our classifier using a max of 10,000 questions from each KG (to reduce the KG-type imbalance issue) and 40,000 questions from the VQA 2.0 dataset. We fine-tune the Sequence Classification BERT model for two epochs. To adapt questions from the VQA 2.0 dataset, we evaluate the model after every 600 batches to determine which questions to include as a part of the training data. Questions in the VQA 2.0 that are assigned KG(s) with at least 95% probability are adapted into the training data. We conducted several experiments by varying the number of training questions, questions to adapt from the VQA 2.0 dataset, and the probability of adaption threshold. Empirically, we find that as we increase the number of questions to adapt from the VQA 2.0 dataset, we see a decrease in the performance of the KG classifier. We evaluate our KG classifier on a validation dataset of 685 questions from the VQA 2.0 dataset with KGs manually identified for these questions by three human judges. Table 3 presents the Fleiss's Kappa (Fleiss & Cohen, 1973) per KG to show a measure of the agreement between the three human raters for 500 questions.

Our current experiments indicate that automatic KG identification without metadata is a challenging task for machine learning models, particularly when trained on and applied to different datasets (see Table 4). Consequently, we plan to improve our KG classifier to use it to identify KGs for other VQA datasets as part of future research initiatives.

3.6. Performance of VQA models

A key component of our paradigm is to evaluate the performance of AI agents through the lens of KGs. We examine the performance of the vision-and-language BERT (ViLBERT) (Lu et al., 2019), MCAN (Yu, Yu, Cui, Tao, & Tian, 2019), and MAC model (Hudson & Manning, 2018) for the GQA dataset. We refrain from reporting the performance of these models on the VQA 2.0 dataset because of the acknowledged limitations of our KG classifier, see Table 4. The MAC model uses a unique memory, attention, and composition (MAC) recurrent cell in a neural network architecture, which is composed of a control state and

Table 4
KG classifier results for VQA 2.0 dataset. Precision is the ratio of true positives identified by the model to the total number of positive samples identified. Recall is the ratio of true positives identified to the total number of positive samples in the data. F1-score is the harmonic mean of precision and recall

Knowledge Gaps	Precision	Recall	F1-Score
No Gap	0.0	0.0	0.0
Activity	0.557	0.542	0.50
Attribute	0.356	0.656	0.462
Direction	0.75	0.29	0.419
Entity Resolution	0.952	0.079	0.146
Location	0.396	0.588	0.473
Material	0.333	0.5	0.4
Reasoning	0.5	0.021	0.04
Sentiment	1.0	0.261	0.414
Size	0.097	0.929	0.176
State	1.0	0.029	0.056
Micro Avg	0.323	0.255	0.285
Macro Avg	0.54	0.354	0.285
Weighted Avg	0.644	0.255	0.244
Samples Avg	0.315	0.26	0.273

a memory state. The control state performs reasoning-related tasks, while the memory state reads, extracts, and integrates specific information from the image, guided by the control state. We use the code and data provided by the author to train a model for this work. The modular coattention network (MCAN) model takes advantage of region-based convolutional neural networks (R-CNNs) and long–short-term memory recurrent units (LSTMs) to perform feature extraction on the input image and question, which are passed as input to deep coattention mechanisms that select important information from the question–image pair. Traditional attention mechanisms attend to a single input source, usually a text segment, but the adapted coattention units can relay information between the question and image to inform information extraction. The MCAN model then uses feedforward networks on the attention-processed input to select an answer to the question. We use the MCAN Large pretrained model and code provided by OpenVQA (footnote: https://openvqa.readthedocs.io/en/latest/) to evaluate the MCAN model on the GQA Testdev Dataset. Finally, the more recent ViLBERT model uses coattention-based transformers, which expands on the previously described BERT architecture (see Section 3.5) to include visual image information. Similar to the MCAN model, ViLBERT uses coattention mechanisms to extract important information from question–image pairs as it processes data. Traditional BERT models operate on text segments, while the modified ViLBERT architecture jointly operates on text segments and image regions. We use the code and data provided by the original author to evaluate the pretrained multitask model for the GQA task (i.e., Task 15).

There are 943,000 training questions and 132,062 validation questions in the GQA dataset. Table 5 presents the accuracy of each model on the GQA Testdev questions with identified KGs using our KG to question mapping (see Table 2). Experiment results indicate that the

Table 5
Accuracy of VQA systems with respect to KGs for the GQA dataset based on author dataset splits

Knowledge Gaps	ViLBERT	MAC Network	MCAN
Activity	0.449	0.372	0.426
Attribute	0.672	0.615	0.649
Direction	0.551	0.505	0.534
Entity Resolution	0.506	0.473	0.481
Location	0.443	0.411	0.417
Material	0.663	0.607	0.644
Reasoning	0.672	0.618	0.642
Sentiment	0.600	0.627	0.587
Size	0.655	0.591	0.643
State	0.681	0.569	0.644

transformer-based model outperforms the two other VQA models. Additionally, all models have the highest accuracy for the entity resolution gap. This is partly because all questions tagged with entity resolution gaps are binary questions (e.g., yes/no). Moreover, we can see that all models do not perform similarly for all KGs; in particular, the MCAN model performs significantly worse on activity and location gaps compared to the MCAN model's performance on the other gaps. For location, this discrepancy is likely due to the homogeneity of questions associated with the gap, which has a low number of unique questions; if the MCAN or MAC model is unable to properly leverage image information to answer the question, the model will likely converge on the most probable answer based on language or world prior to respond with, which may not be correct. For questions associated with activity gaps, attention to the surroundings of the subject in the image is important to discern the motion and position of the subject. From our understanding, the deeply interconnected multihead attention mechanisms of ViLBERT exceed the attention mechanisms found in the MCAN and MAC models; thus, the drop in the performance for activity-related questions for ViLBERT is not as substantial as the drops in the performance observed in the MAC or MCAN models. Additionally, the transformer-based ViLBERT model is trained on multiple computer vision tasks, giving it the benefit of contextualized pre-training.

The KG-VQA paradigm enables initial steps toward understanding VQA through the lens of KGs and cognitive capabilities. In our future work, we plan to generate questions to further test these data-hungry VQA models to better understand their sensitivity with respect to different dataset characteristics (e.g., number of questions, number of questions per KG, number of images, etc.). Further, the KG-VQA paradigm will be used to collect human data and cognitive models, and used to evaluate other AI systems and methodologies.

3.7. Challenges in the VQA domain

While developing the KG-VQA paradigm, several aspects of VQA required caution. First, some questions are ambiguous, even for humans (i.e., Word-Sense KG). A solution has not been incorporated to identify questions with Word-Sense KGs, further increasing the

complexity of the task. For example, *"Is the table clear?"* can be interpreted as asking about the table's opacity or asking if there are any objects on the table. These questions further add to the difficulty of VQA. Second, we note that there are inconsistencies in annotations across the GQA dataset. For the following questions, the *global group* label in the dataset is *color*:

- *Which is healthier, the orange or the muffins?*
- *Which is healthier, the candies or the orange?*

This annotation is incorrect, as the question is not about the color of an object, but is instead about the object *orange*. Additionally, for some questions, parts of the annotation are missing, for example, *"Is the large pot to the right or to the left of the jar in the middle of the picture?"* There is no mention of *size* or *large* in the question annotation.

Furthermore, VQA system answers are evaluated using exact string match, multiple choice, or by looking at the top k most likely responses. These metrics ignore any semantic similarity between the target answer and the answer suggested by the agent. This could misrepresent the ability of an agent to answer the question. For example, if an agent is posed with the question *"What is the girl wearing?,"* the answers *t-shirt* and *dress* are much more similar than the answers *t-shirt* and *utensil*, yet neither *dress* nor *utensil* matches the target string of *t-shirt*. For this question, the agent that answers *dress* might display the same performance as the agent that answers *utensil*, even though *dress* is a better answer. The Wu–Palmer similarity (WUPS) (Wu & Palmer, 1994) has also been used to evaluate VQA models (Malinowski & Fritz, 2014). WUPS tries to measure the difference between the semantic meaning of a predicted answer and the ground truth by finding the least common subsumer between two semantic senses (Kafle & Kanan, 2017). However, WUPS cannot handle pairs of words that are lexically similar and text phrases, which is a problem for VQA systems that accept open-text responses.

There are also social biases present in the VQA 2.0 dataset questions. For example, the question *"What are these Asians going to eat?"* is presented alongside an image of an Asian family playing a game, with no food present. The associated answers with this question are "Sushi," "Pizza," "Rice," and "Stew," along with generic answers of "Food" and "Nothing." By training a VQA agent to associate people of Asian descent with specific food types, social biases are strengthened and reinforced. Additionally, numerous questions in the VQA 2.0 dataset ask about the sexuality of a target with no pertinent context in the image; one such question asks *"Does this guy look gay?"* alongside a portrait of a man in business attire with a generic background and no other objects in frame. This forces the agent to perpetuate possibly harmful assumptions without reasonable evidence or context.

4. Discussion and future work

We introduce a paradigm for KG detection, identification, and resolution. Our work is far from complete but allows us to examine the VQA domain from a new lens. Our work is currently focused on KG identification using the GQA dataset, though we do consider classifying questions in different VQA datasets (i.e., VQA 2.0) to determine the types of KGs

questions are amenable to. However, our results in Table 4 indicate that this is a challenging task. As future work, we aim to improve our KG classifier and test VQA models on different datasets as well to analyze the performance across different KGs. Currently, we are only able to identify KGs for 59% (559,020 out of 943,000) questions in the GQA-balanced training dataset and 59% (78,343 out of 132,062) questions GQA-balanced validation dataset. Generalizing KGs for different VQA datasets is challenging without the additional metadata the GQA dataset provides for questions. Moreover, we identify KGs through textual questions; however, KGs potentially occur due to the absence of information in a question–image pair, not just the question. Our simplistic, textual model for KG identification could fail to identify KGs that arise due to information in the image. Additionally, we observe that VQA models are data hungry models, making it difficult to manipulate the distribution of questions per KG to examine the performance of these models without augmenting the dataset with new questions.

In this work, we focus on the identification of KGs. Orthogonally, this can be expanded to understand how to automatically detect KGs. Our future work on KG detection, identification, and resolution in the VQA domain will focus primarily on targeted question generation for the introduction of KGs as well as methods of resolving a KG when it is present. Possible methods of augmenting the VQA datasets to cover a set of questions amenable to specific KGs include question generation and question adaptation. For question generation, we introduce a new model that takes an image and target KG as input and produces a question–answer pair as output. The generative model will produce a stylized question amenable to the target KG using information that can be learned or collected from the image and associated object and spatial features. We propose a similar model for question adaptation, taking influence from StyleGAN (Karras, Laine, & Aila, 2019), to generate new questions from existing ones while adapting them to contain new possible KGs. These methods of KG question generation enable us to inject specific KGs into the by changing or introducing data.

To resolve KGs, we propose a reinforcement learning model with a set of atomic actions the agent can perform to help resolve the identified KG. Atomic actions are simple, discrete actions the agent can take to collect information relevant to the KG it is experiencing to help it in decision-making and reasoning. For example, if a VQA agent experiences a lexical gap due to not recognizing a particular word, it might query a lexical database such as WordNet (Miller, 1995) to find appropriate synonyms or antonyms to use in place of the original word. In another instance, the agent might even query the user for more information it can use to find the answer to the question. Additionally, we plan to explore how to evaluate detection, identification, and resolution capabilities in AI models. Broader directions of future research include: (a) examining formal processes to expand and refine on the current sent of specified KGs; (b) designing systematic approaches to identifying and resolving relevant KGs for a particular problem or set of intelligent agents; (c) exploring mechanisms to codify lessons learned from human empirical studies (see Fig. 1); and (d) developing new types of AI and cognitive models that inculcate KG processes.

5. Conclusion

Human cognition has many capabilities that modern AI systems lack, such as the detection, identification, and resolution of KGs. In this work, we present an evaluation paradigm for KG detection, identification, and resolution for the VQA task. We share a simple classifier for identifying VQA questions that can be manipulated into specific KG types (per our taxonomic representation) for VQA. Additionally, we examine VQA datasets and analyze VQA models in a novel fashion through the lens of the KG-VQA paradigm to benchmark performance for specific KGs. Finally, we consider extensions to this work with question generation for KG introduction and reinforcement learning agents for KG resolution.

Acknowledgments

We thank Dr. Wei-Lun Chao for suggestions on this work. The authors also acknowledge a seed grant from Air Force Research Laboratory and Ohio State University, and a cooperative AI Institute grant (AI-EDGE), from the National Science Foundation under CNS-2112471. All content represents the opinion of the authors, and is not necessarily endorsed by their institutions or sponsors.

Notes

1 https://visualcommonsense.com/download/
2 https://okvqa.allenai.org
3 https://bitbucket.org/sxjzwq1987/kb-vqa-dataset
4 https://www.dropbox.com/s/iyz6l7jhbt6jb7q/new_dataset_release.zip?dl=0
5 https://arijitray1993.github.io/ConVQA/
6 https://www.microsoft.com/en-us/research/project/vqa-introspect/
7 https://cs.stanford.edu/people/dorarad/gqa/about.html
8 https://cs.stanford.edu/people/dorarad/gqa/download.html
9 https://huggingface.co/transformers/model_doc/bert.html
10 https://github.com/stanfordnlp/mac-network/tree/gqa
11 https://github.com/MILVLG/mcan-vqa
12 https://github.com/facebookresearch/vilbert-multi-task

References

Acharya, A., Chen, X., Myers, C. W., Lewis, R. L., & Howes, A. (2017). Human visual search as a deep reinforcement learning solution to a POMDP. In *CogSci* (pp. 51–56).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015a). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425–2433).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015b). VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Bajaj, G., Bandyopadhyay, B., Schmidt, D., Maneriker, P., Myers, C., & Parthasarathy, S. (2020). Understanding knowledge gaps in visual question answering: Implications for gap identification and testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 386–387).

Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186).

Eberhart, A., Shimizu, C., Stevens, C., Hitzler, P., Myers, C. W., & Maruyama, B. (2020). A domain ontology for task instructions. In B. Villazón-Terrazas, F. Ortiz-Rodríguezm, S. M. Tiwari, & S. K. Shandilya (Eds.), *Knowledge Graphs and Semantic Web. Second Iberoamerican Conference and First Indo-American Conference, KGSWC 2020*, vol. *1232* (pp. 1–13). Mérida, Mexico: Communications in Computer and Information Science.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. USA: Oxford University Press.

Gluck, K. A., & Laird, J. E. (Eds.). (2018). *Interactive task learning: Humans, robots, and agents acquiring new tasks through natural interactions*. Strüngmann forum reports. The MIT Press.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kafle, K., & Kanan, C. (2017). Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, *163*, 3–20.

Karras, T., Laine, S. & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4401–4410).

Khalid, S. (2019). Bert explained: A complete guide with theory and tutorial. Retrieved from https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/

Kirk, J., Mininger, A., & Laird, J. (2016). A demonstration of interactive task learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 2016-Janua, 4248–4249.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals Of Mathematical Statistics*, *22*(1), 79–86.

Kupitz, C., Eberhart, A., Schmidt, D., Stevens, C., Shimizu, C., Hitzler, P., … Myers, C. W. (2021). Toward undifferentiated cognitive models. In *International Conference on Cognitive Modeling* (pp. 1–6). Virtual.

Laird, J. E. (2012). *The soar cognitive architecture*. Cambridge, MA: MIT Press.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361.

Li, T. J.-J., Radensky, M., Jia, J., Singarajah, K., Mitchell, T. M., & Myers, B. A. (2019). Interactive task and concept learning from natural language instructions and gui demonstrations. In *AAAI-20 Workshop on Intelligent Process Automation (IPA-20)*, New York, New York, USA.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. doi: https://doi.org/10.1017/S0140525X1900061X

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems* (pp. 13–23).

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*, *27*, 1682–1690.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. doi: https://doi.org/10.1037/0033-295X.88.5.375

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communication-ACM*, *38*(11), 39.

Myers, C. W., Ball, J., Cooke, N., Freiman, M., Caisse, M., Rodgers, S., … McNeese, N. (2019). Autonomous intelligent agents for team training. *IEEE Intelligent Systems*, *34*(2), 3–14. doi: https://doi.org/10.1109/MIS.2018.2886670

Newell, A., & Simon (1972). *Human problem solving*. NJ: Prentice Hall, Englewood Cliffs.

Nushi, B., Kamar, E., & Horvitz, E. (2018). Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Volume *6*.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675. doi: https://doi.org/10.1037//0033-295x.106.4.643

Reed, S. K., & Vallacher, R. R. (2020). A comparison of information processing and dynamical systems perspectives on problem solving. *Thinking and Reasoning*, *26*(2), 254–290.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Salvucci, D. D. (Wiley Online Library; 2021). Interactive grounding and inference in instruction following. *Topics in Cognitive Science*, to appear.

Schmidt, D. P. (2020). Identifying Knowledge Gaps Using a Graph-based Knowledge Representation (Master's Thesis, Wright State University).

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.

Spelke, E. S. (2017). Core knowledge, language, and number. *Language Learning and Development*, *13*(2), 147–170.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge MA: MIT Press.

Walsh, M. M., Einstein, E. H., & Gluck, K. A. (2013). A quantification of robustness. *Journal of Applied Research in Memory and Cognition*, *2*(3), 137–148. doi: https://doi.org/10.1016/j.jarmac.2013.07.002

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133–138).

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6281–6290).

Zhang, P., Wang, J., Farhadi, A., Hebert, M., & Parikh, D. (2014). Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3566–3573).