Commonsense Learning: An Indispensable Path towards Human-centric Multimedia

Bin Huang*
huangb19@mails.tsinghua.edu.cn
Tsinghua University

Siao Tang*
tsa18@mails.tsinghua.edu.cn
Tsinghua University

Guangyao Shen
thusgy2012@gmail.com
Tsinghua University

Guohao Li
ligh16@mails.tsinghua.edu.cn
Tsinghua University

Xin Wang†
xin_wang@tsinghua.edu.cn
Tsinghua University

Wenwu Zhu†
wwzhu@tsinghua.edu.cn
Tsinghua University

ABSTRACT

Learning commonsense knowledge and conducting commonsense reasoning are basic human ability to make presumptions about the type and essence of ordinary situation in daily life, which serve as very important goals in human-centric Artificial Intelligence (AI). With the increasing number of media types and quantities provided by various Internet services, commonsense learning and reasoning with no doubt are playing key roles in making progresses for human-centric multimedia analysis. Therefore, this paper first introduces the basic concept of commonsense knowledge and commonsense reasoning, then summarizes commonsense resources and benchmarks, gives an overview on recent commonsense learning and reasoning methods, and discusses several popular applications of commonsense knowledge in real-world scenarios. This work distinguishes itself from existing literature that merely pays attention to natural language processing in focusing more on multimedia which include both natural language processing and computer vision. Furthermore, we also present our insights and thinking on future research directions for commonsense.

CCS CONCEPTS

• **Computing methodologies → Knowledge representation and reasoning**.

KEYWORDS

commonsense knowledge, reasoning

ACM Reference Format:
Bin Huang, Siao Tang, Guangyao Shen, Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Commonsense Learning: An Indispensable Path towards Human-centric Multimedia. In *1st International Workshop on Human-centric Multimedia Analysis (HuMA'20), October 12, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3422852.3423484

1 INTRODUCTION

Commonsense knowledge is known to be shared knowledge across a particular cultural group of people such that every individual in this group is expected to know or assume the amount of knowledge. Commonsense knowledge contains background information about the spatial, physical, social and temporal properties of entities, events and circumstances etc., so that it plays an important role in guiding our daily life. For example, humans will think it stupid to slam an egg on a stone in order to break the stone, because it is a commonsense to us that eggs are much more fragile than stones. Therefore, it is essential for AI systems to learn and utilize commonsense knowledge so that machines can better understand human world and act in a more human-like manner. However, learning commonsense knowledge and utilizing commonsense for reasoning have been very challenging in the research community, evidenced by the facts that existing machine learning algorithms perform poorly compared with human in many tasks requiring commonsense knowledge.

Human can conduct commonsense reasoning through utilizing commonsense knowledge, which serves as the core to perception, understanding and decision making. Figure 1 presents daily situations of human world, where commonsense knowledge is essential. As such, the evolution of Artificial Intelligence can seldom be successful without the involvement of commonsense reasoning. Although researchers from various communities ranging from natural language processing (NLP) to computer vision (CV) and robotic are putting more and more attentions on investigating commonsense reasoning, endowing machines with the ability of leveraging commonsense for reasoning is considered to be a bottleneck of current Artificial General Intelligence (AGI) [17].

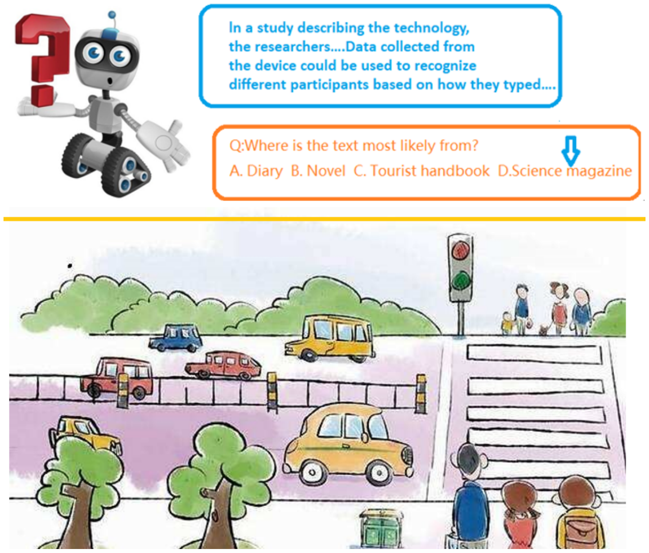
Existing works on commonsense knowledge mainly focus on two aspects, i.e., acquiring commonsense knowledge and evaluating commonsense learning or reasoning through various benchmarks.

Acquiring commonsense knowledge. On the one hand, researchers extract commonsense knowledge from some large-scale unstructured resources such as Wikipedia (the largest and most popular electronic encyclopedia around the world consisting of over 47 million entries). On the other hand, in order to collect more high-quality and well-organized commonsense knowledge, researchers also try to build large-scale commonsense knowledge bases including CYC [32], NELL [44], ConceptNet [61], WebChild [67] and ATOMIC [56] etc.

Evaluating commonsense learning or reasoning. Researchers have created many benchmarks to test algorithms' ability to learn

*Both authors contributed equally to this research.
†Corresponding authors.

HuMA'20, October 12, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8151-2/20/10…$15.00
https://doi.org/10.1145/3422852.3423484



In a study describing the technology, the researchers....Data collected from the device could be used to recognize different participants based on how they typed....

Q:Where is the text most likely from?

A. Diary B. Novel C. Tourist handbook D.Science magazine

Figure 1: The first example in the figure describes a reading comprehension task. We humans can easily choose the answer with background commonsense, but it's challenging for machines. Another example is about a traffic situation which requires traffic commonsense knowledge to make correct actions.

commonsense knowledge and conduct commonsense reasoning. Existing benchmarks can be approximately divided into several groups [58], i) social commonsense: SOCIAL IQA [57] and VCR [76] etc., ii) physical commonsense: PHYSICAl IQA [5] and SWAG [77] etc., iii) temporal commonsense: MCTACO [82], iiii) commonsense reading comprehension: COSMOS QA [26]. Some other benchmarks may involve more than one type of commonsense knowledge, for example, COMMONSENSEQA [65] takes both social and physical commonsense into consideration.

To improve the performance of machine learning algorithms in various benchmarks, a lot of strategies incorporating commonsense knowledge into reasoning tasks have been proposed. In particular, self-supervised learning [20, 23] learns commonsense from large-scale unstructured corpus through different pre-training tasks. Relational reasoning [35] infers the relationships between different entities from the topology of knowledge graph that contains a wealth of commonsense knowledge. Other methods [39] combine self-supervised learning and relational reasoning together, aiming to keep the advantages of both strategies. These strategies have wide applications in various machine learning tasks such as visual segmentation, video description and recommender systems etc.

To summarize, in this paper we introduce the basic concept of commonsense knowledge and commonsense reasoning, present main commonsense resources and benchmarks, survey recent advances on commonsense learning and reasoning, and discuss several practical applications with commonsense knowledge. Different from existing literature with similar topic which focuses mainly on natural language processing, this work starts from the view related to the multimedia community, focusing on both natural language processing and computer vision. Last but not least, we further share our insights on future research directions for commonsense.

2 COMMONSENSE RESOURCES

On the road towards Artificial General Intelligence (AGI), one of the main bottlenecks is that machines lack the reasoning ability with commonsense knowledge [17]. To tackle this challenge and endow machines with the ability to learn and exploit commonsense knowledge, researchers built plentiful commonsense resources in the past decades. During the time when the Internet was thriving, people created huge amounts of raw data on the web that contains commonsense knowledge implicitly. We refer to this kind of commonsense resources as unstructured commonsense corpus. Based on that, researchers further generalize and sort out these resources to form well-organized, easy-to-use commonsense knowledge bases.

2.1 Unstructured Commonsense Corpus

Since electronic records have been available, a variety of unstructured resources that implicitly contain commonsense knowledge have emerged, such as English Wikipedia, BookCorpus [83], etc. English Wikipedia is the English-language version of the freely edited online encyclopedia Wikipedia. It is founded on 15 January 2001 and now consists of more than 6,136,733 articles, covering almost all aspects of human knowledge. BookCorpus is a large-scale text corpus which consists of 11,038 unpublished books from 16 different genres and 984 million words. These resources of book can provide very rich, descriptive text that convey high-quality semantics. And there are also some multi-modal corpora which also include images or videos. For example, Coco [13] is a data contains about 200000 images, each of which usually has five textual descriptions. ImageNet [18] is a dataset built upon the backbone of the WordNet [43] structure. The dataset originally aims to populate the majority of the 80,000 synsets of WordNet with an average of 500-1000 clean and full resolution images. And so far, it includes over 14 million images and over 21000 synsets indexed. The DeepMind Kinetics human action video dataset is a video dataset which describes 400 human action classes with over 400 video clips for each action[28].

Unstructured commonsense corpus includes all the texts, images, videos and other media which reflect commonsense in the human world. However, these unstructured corpus are difficult to be directly incorporated into downstream reasoning tasks. A mainstream approach that performs well on capturing the implicit commonsense knowledge is to pre-train neural models on the large-scale unstructured textual corpus and adapt to downstream tasks through fine-tuning. Recent methods based on pre-trained models such as BERT [19], GPT [51] and ELMo [49] have obtained remarkable results in many commonsense reasoning NLP tasks. And other models(VL-BERT [62],VisualBERT [34],etc.) trained on multimodal corpus obtain state-of-the-art results in some multi-modality tasks. Another popular approach is to explicitly extract commonsense knowledge from the unstructured corpus and organize them in the form of structured knowledge base, as shown in the following subsection.

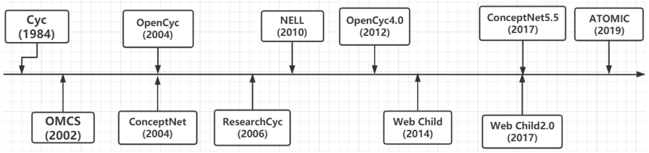


Figure 2: The development timeline of the several commonsense knowledge bases [58].

2.2 Well-organized Commonsense Knowledge Base

Researchers have been devoted to building structured large-scale commonsense knowledge bases, aiming to provide practical and high-quality commonsense knowledge for various downstream applications. We show a development timeline(Figure 2) of several popular knowledge bases and present a detailed description for each of them, respectively.

2.2.1 Cyc. Cyc [32] is one of the pioneers of commonsense knowledge base. Cyc expresses commonsense in a way of LISP-style logic [31]. It presents ontological relationships between objects, using so-called CycL language. Specifically, the concept names in Cyc are CycL terms or constants(individuals, Collections, Functions and Truth Functions) and facts about concepts are asserted using certain CycL sentences. OpenCyc is an open source version of Cyc knowledge base, providing API and downloadable dataset for developers and users. ResearchCyc [53] is a free version of Cyc released to the research community. In addition to containing the categorical information in OpenCyc, a significant amount of semantic knowledge was added to ResearhCyc.

2.2.2 ConceptNet. ConceptNet [61] is a frequently used commonsense knowledge base, which represents daily words and phrases in the form of graph, with a mass of concept nodes connected by kinds of relations. ConceptNet is originally a representation for the knowledge collected by Open Mind Common Sense (OMCS), which leverages a interactive website to acquire free text commonsense assertions from online visitors. The latest version, ConceptNet 5.5, consists of over 21 million edges and over 8 million nodes in more than 85 languages. ConceptNet5.5 incorporates knowledge from other crowd-sourced resources, especially data mined from Wiktionary and Wikipedia. Meanwhile, it links to other knowledge graphs like WordNet [43]and Freebase [6].

2.2.3 NELL. Commonsense knowledge can change over time with the development of human society, as a result, it's crucial to update knowledge base. The Never-Ending Language Learner(NELL) [44], a never-ending learning engine, can automatically learn knowledge from web contents. Since January 2010, it has been reading the web and learning twenty-four hours a day. NELL has built a knowledge base with over 80 million confidence-weighted beliefs. All belief triples, such as "play(MapleLeafs,hockey)", own an associated confidence.

2.2.4 WebChild 2.0. WebChild 2.0 [67] is a large, clean, and semantically organized commonsense knowledge base, with over 2 million disambiguated concepts and activities, connected by over 18 million assertions. It automatically extracts commonsense from Web content (such as Google's large Web N-Gram) and other text resources. The knowledge base consists of fine-grained commonsense properties, connecting noun senses with adjective senses by a variety of relations (19 types) such as "hasSize", "hasAbility", etc.

2.2.5 ATOMIC. ATOMIC [56] is an atlas of machine commonsense, which focuses on inferential if-then knowledge rather than taxonomic knowledge. It describes everyday events in natural language, in the form of nine typed if-then relations. The commonsense knowledge graph consists of over 300k events associated with 877k inferential relations. ATOMIC applies a crowd-sourced method to collect free-form text annotations by asking online visitors to write answers to a question about a specific event.

Overall, these knowledge bases provide explicit, clean and well-organized commonsense knowledge in general, but they differ with each other in many aspects, such as the way of construction, the knowledge formats and scopes, etc. As for the way of knowledge base construction, Cyc leverages experts to add commonsense knowledge, while the others, ConceptNet, NELL, WebChild and ATMOIC extract knowledge from non-experts knowledge written in natural language. Table 1 present comparisons of knowledge representation format.

3 BENCHMARKS

Since the early 2000s, especially recent years, researchers have been eagerly devoted to creating benchmarks for commonsense reasoning. An excellent benchmark can adequately examine machine's ability to learn commonsense knowledge and reason with commonsense already mastered. Most benchmarks focus on natural language processing, but recent years visual benchmark such as VCR emerged, prompting the development of multi-modal models. We will give an overview(Table 2) of some existing benchmarks, and then describe several latest benchmarks in detail. Figure 4 presents some concrete examples of each benchmark. We can see it is effortless for humans to solve these reasoning problems, however, it is still challenging for machines. For example, as shown in Figure 4, the COMMONSENSEQA question says "Where can I stand on a river to see water falling without getting wet?". We humans can easily choose the correct answer - "bridge", because we know bridge is over the water thus we will not get wet. But the question is difficult for machines without commonsense knowledge.

According to the type of commonsense knowledge required for reasoning, benchmarks can be approximately classified into four categories: Physical [5, 77, 78, 80], Social [27, 45, 46, 54, 55, 57, 76], Temporal [82] and Reading Comprehension [26, 29, 79], as shown in Figure 3. In addition, some benchmarks begin to concern about hybrid commonsense [4, 65].

3.1 Social Commonsense

Social commonsense is the basic knowledge about social situations such as interpersonal interaction. Social commonsense is critical for humans' ability to reason about mental states of others and their likely actions [22], and it is this ability that enables us to navigate various of social situations [2].

Table 1: Comparisons of knowledge representation format.

	Representation format	Instance
Cyc	symbolic logic	#$capitalCity #$France #$Paris
ConceptNet	triple	IsA(cook,person)
NELL	triple	play(MapleLeafs,hockey)
WebChild 2.0	triple	<car, faster than, bike>
ATOMIC	If-then assertion	If X repels Y's attack,then Y feels ashamed...

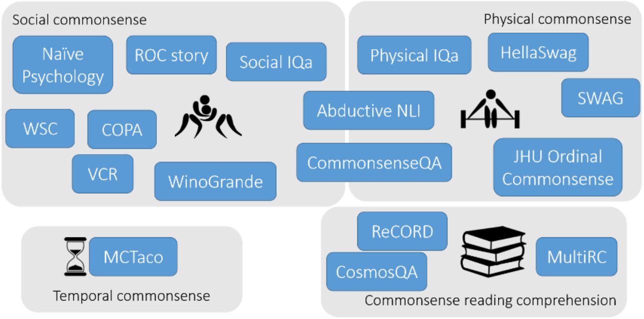


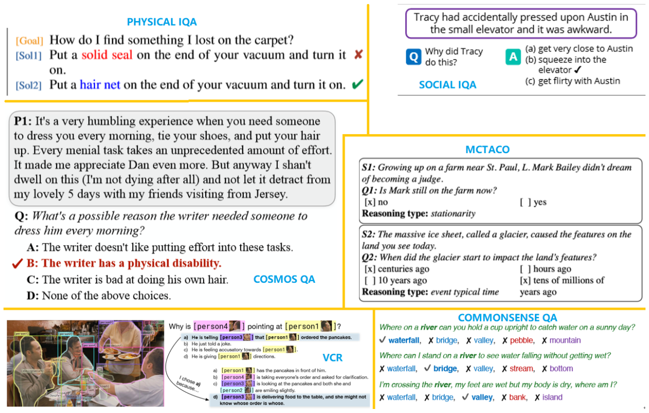
Figure 3: A classification of benchmarks according to the type of commonsense [58].

3.1.1 SOCIAL IQA. SOCIAL IQA [57] is the first large-scale benchmark which concentrates on commonsense reasoning about social situations. It consists of 38,000 multiple choice questions. SOCIAL IQA leverages a crowd-sourced method to collect social commonsense questions along with corresponding both correct and wrong answers. Researchers further take SOCIAL IQA as a resource of commonsense knowledge for transfer learning, which have proved effective in some other benchmarks.

3.1.2 VCR. Visual understanding is a significant challenge for AI systems, which goes beyond simple recognition to advanced cognition-level understanding. For human beings, with a rapid glance at a visual scene, we not only can effortlessly recognize the people and objects in the scene, but also can infer the people's goals and mental states, which is not visually obvious. VCR [76] is a large-scale dataset about Visual Commonsense Reasoning, consisting of 290k multiple choice question-answer problems derived from 110k movie scenes. The problem not only requires machine to correctly answer a challenge question about an image, but also asks for a rationale which can justify the answer.

3.2 Physical Commonsense

Physical commonsense consists of physical properties of everyday objects such as shape and material, and it also consists of knowledge about affordances and manipulation of these objects. For example, PHYSICAL IQA [5] is a commonsense reasoning benchmark concerning physical aspect of everyday events. The objective is to test whether AI systems can tackle physical commonsense questions without experiencing the real physical world. The dataset consists of over 16,000 question-answer pairs for training, about 2,000 pairs for development and 3,000 pairs for test.



PHYSICAL IQA

[Goal] How do I find something I lost on the carpet?
[Sol1] Put a solid seal on the end of your vacuum and turn it on. ✗
[Sol2] Put a hair net on the end of your vacuum and turn it on. ✓

Tracy had accidentally pressed upon Austin in the small elevator and it was awkward.

Q Why did Tracy do this?

A (a) get very close to Austin
(b) squeeze into the elevator ✓
(c) get flirty with Austin

SOCIAL IQA

P1: It's a very humbling experience when you need someone to dress you every morning, tie your shoes, and put your hair up. Every menial task takes an unprecedented amount of effort. It made me appreciate Dan even more. But anyway I shan't dwell on this (I'm not dying after all) and not let it detract from my lovely 5 days with my friends visiting from Jersey.

Q: *What's a possible reason the writer needed someone to dress him every morning?*

A: The writer doesn't like putting effort into these tasks.
✓ **B: The writer has a physical disability.**
C: The writer is bad at doing his own hair.
D: None of the above choices.

COSMOS QA

MCTACO

S1: *Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.*
Q1: *Is Mark still on the farm now?*
[x] no [ ] yes
Reasoning type: *stationarity*

S2: *The massive ice sheet, called a glacier, caused the features on the land you see today.*
Q2: *When did the glacier start to impact the land's features?*
[x] centuries ago [ ] hours ago
[ ] 10 years ago [x] tens of millions of years ago
Reasoning type: *event typical time*

Why is [person4] pointing at [person1]?
a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

VCR

COMMONSENSE QA

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*
✓ **waterfall**, ✗ bridge, ✗ valley, ✗ pebble, ✗ mountain

*Where can I stand on a **river** to see water falling without getting wet?*
✗ waterfall, ✓ **bridge**, ✗ valley, ✗ stream, ✗ bottom

*I'm crossing the **river**, my feet are wet but my body is dry, where am I?*
✗ waterfall, ✗ bridge, ✓ **valley**, ✗ bank, ✗ island

Figure 4: Illustrations of the samples of several benchmarks [5, 26, 57, 65, 76, 82].

3.3 Temporal Commonsense

Temporal commonsense consists of knowledge about temporal phenomena, for example: how long an event takes? how often an event occurs? Temporal commonsense that people rarely express obviously is crucial for understanding daily events. MCTACO [82](multiple choice temporal commonsense) is an important benchmark which focuses on temporal commonsense reasoning. It consists of 1,893 questions and 13,225 question-answer pairs. Founders clarify all the questions into five temporal reasoning types: duration, temporal ordering, typical time, frequency and stationarity. MCTACO is constructed via crowd-sourcing on Amazon Mechanical Turk to collect questions and corresponding answers and distractors, with elaborate guidelines to guarantee high quality.

3.4 Commonsense Reading Comprehension

Reading comprehension requires the ability to understand clues explicitly stated in text and to read between the lines with background knowledge. For example, COSMOS QA [26] (Commonsense Machine Comprehension) is a dataset consisting of 35,588 reading comprehension problems which require background commonsense about the causes and effects of events. It also uses a crowd-sourcing way (Amazon Mechanical Turk) to collect questions and answers.

3.5 Hybrid Commonsense

Many benchmarks refer to more than one type of commonsense, such as COMMONSENSEQA [65], which concerns both social and physical commonsense knowledge. It is a challenging benchmark in the form of question answering, which consists of 12,247 questions.

Its construction process first extracted a subgraph from ConceptNet centered on a single concept, and then asked crowd-workers to create corresponding natural language questions and answers according to the subgraph.

4 INCORPORATING COMMONSENSE KNOWLEDGE IN MULTIMEDIA

With the aforementioned resources and benchmarks, researchers attempts to integrate human-centric commonsense knowledge into ML systems for deeper understanding of the society and the world. There are two mainstream methods for incorporating commonsense knowledge: one is self-supervised learning for acquiring commonsense knowledge from unstructured corpus, and the other is relational reasoning on structured commonsense knowledge bases. Furthermore, combining the above two ideas, some hybrid methods have achieved promising results in many benchmarks recently. We list the overall methods in Figure 5

4.1 Self-supervised Learning on Unstructured Corpus

For the implicit commonsense knowledge hidden in the large-scale unstructured multimedia corpus, manually annotating and extracting knowledge would consume unacceptable efforts. Therefore, researchers tend to self-supervised learning, which leverages the information of the corpus itself to construct pseudo tags for training. Self-supervised learning could provide better representations and transfer the commonsense knowledge learned from pre-trained tasks to many downstream tasks.

4.1.1 Pre-training from Language Corpus. At present, commonsense knowledge is mainly represented as texts. The most common self-supervised pre-training task for texts is called language model (LM), where some tokens are expected to be recovered from their contexts. Early pre-trained language model can be traced back to word2vec [41, 42], which converts every word into an unique embedding. It has two training paradigms CBOW and Skip-gram as simple version of LM: CBOW method predicts current token from its context tokens, and Skip-gram method predicts context tokens from the current token. Here, the context tokens are taken from a fixed-length sliding window. To better handle polysemy and consider long-term language dependency, ELMo [49] predicts current token from all history tokens with a 2-layer bidirectional LSTM, while GPT [51] replace the LSTMs with stacked transformers [68] and achieves superior performances.

While LM is a default choice, researchers begin to design various pre-training tasks for self-supervised learning, which opens new doors for language understanding from large-scale corpus. BERT [19], a landmark pre-trained model, designs the tasks of masked language model (MLM) and next sentence prediction (NSP). While MLM randomly mask some tokens and expect the model to recover them, NSP predicts whether a sentence is the next sentence of the input sentence in the raw corpus. The last two years have witnessed many pre-training tasks, such as full sentence training [37], sentence order prediction (SOP) [30], and permuted language modeling (PLM) [73], which aims to improve the performance of pre-trained models and better obtain implicit knowledge.

Once pre-trained, the models can be seen as implicit knowledge library which could also provides commonsense knowledge. As most commonsense benchmarks are formulated as question answering tasks, pre-trained models should accept QA pairs as inputs. For example, when BERT deals with a question with four alternative answers, four inputs will be constructed by concatenating the question and every candidate answer. For each input, A [CLS] symbol is inserted before the text, a [SEP] symbol is used to separate question and answer. So the final input form for BERT is "[CLS] question [SEP] answer [SEP]", and the output vector corresponding to the symbol is used as the semantic representation of the whole text, which can be used to predict the correct answer. When applied to different downstream tasks with different forms of input and output, researchers usually add adaptation layers into the model, and fine-tune it layer by layer to avoid catastrophic forgetting. Another idea is to unify all downstream tasks into text-to-text format, which is adopted by T5 [52] and GPT-3 [8]. The difference is that GPT-3 is very hard to do fine-tune due to the huge amount of parameters (more than 170 billion) . Therefore, it directly takes the training data and test data of downstream tasks as the input of the model, and also achieves good results in many tasks, showing the powerful ability of large-scale pre-training model.

Generally speaking, pre-training on larger corpus would achieve better performance and incorporate more commonsense. ELMo uses 1B Word Benchmark [11] for pre-training, which contains about 1 billion words. GPT uses Bookcorpus for pre-training, with more than 10000 books and more than 70 million sentences. BERT uses Bookcorpus [84] and English Wikipedia with a size of about 13GB, while XLNet uses more than 100GB of corpus. T5 uses Colossal Clean Crawled Corpus captured from the Common Crawl website with a size of 750GB. GPT-3 uses a larger data set for training, and the whole English Wikipedia only accounts for 0.6% of its training data. These unstructured but large-scale text corpus contains various commonsense and reflects human intelligence, which deserve more efforts for further navigation.

4.1.2 Pre-training from Multimodal Corpus. While the commonsense knowledge is mainly represented as texts, the multimodal corpus, commonly visual-language corpus, is larger and more general in multimedia. While they are more difficult in processing, visual-language corpus provides more choices for pre-training tasks. For example, in addition to the aforementioned MLM, researchers designs the tasks of masked region feature regression, masked region feature classification, and sentence-image alignment. The first two tasks cover the feature vectors of region randomly, but masked region feature regression expect the model to recover them directly, while masked region feature classification expect the model to predict the labels after the vectors pass through R-CNN[24]. Sentence-image alignment randomly replaces image or text for the input image-text pair, and finally predicts whether there is a corresponding relationship between the image and text, which could be seen as a binary classification task.

Unlike uni-modal pre-training from text corpus, visual-language pre-training must be able to handle visual representations, which are usually the region of interest (RoI) in the images. Another key challenge is to fuse visual and text information in visual-language pre-training. There are two mainstream fusion methods. The first

Table 2: An overview of the commonsense benchmarks in recent years.

	Year	Size	Type(and modality)	Metric	State-of-the-art	Human
COMMONSENSEQA [65]	2019	12,247 questions	social and physical(text)	accuracy	79.1	88.9
SOCIAL IQA [57]	2019	38,000 questions	social(text)	accuracy	83.2	88.1
PHYSICAL IQA [5]	2020	over 21,000 QA pairs[1]	physical(text)	accuracy	90.1	94.9
MCTACO [82]	2019	1,893 questions	temporal(text)	exact match	59.4	75.8
COSMOS QA [26]	2019	35,588 questions	reading comprehension(text)	accuracy	91.8	94.0
VCR [76]	2019	290k questions	social(text+vision)	accuracy	63.0(Q->AR)	85.0(Q->AR)

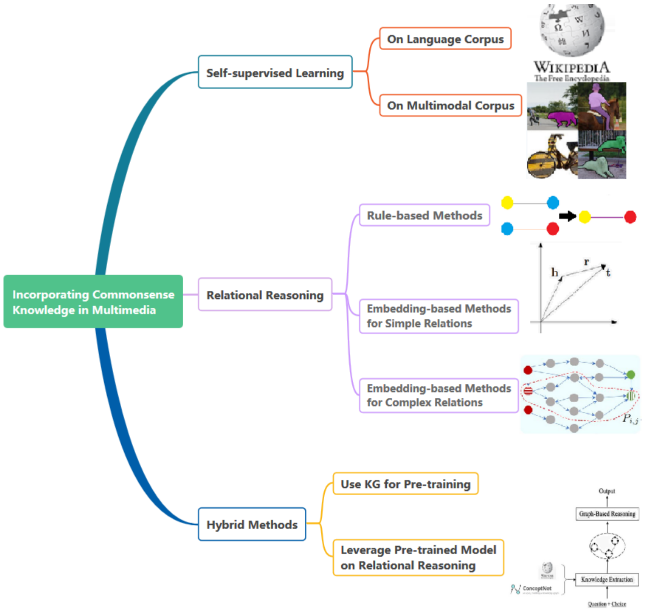


Figure 5: Methods of incorporating commonsense knowledge in multimedia. Image taken from [7, 13, 35, 39]

one includes VisualBERT [34], Unicoder-VL [33], VL-BERT [62], B2T2 [1], UNITER [14], VideoBERT [64], *etc.*, They combine caption tokens with visual feature embedding as input to transformer, aligning and fusing the image and language information at the beginning. The second is that the image and text are encoded independently by the encoder of transformer, and then fused by a co-attention mechanism module. The representative models are VilBERT [38], LXMERT [66], CBT [63], *etc.*.

Pre-training of other modalities is similar to visual-language pre-training. For example, SpeechBERT [16] in audio-language pre-training, which is a cross-modal transformer-based pre-trained language model, encodes audio and text with a single transformer.

Unstructured multimodal corpus is usually some datasets collected for other visual-language research areas. VisualBERT is pre-trained on COCO [13], which contains about 200000 images with five descriptions per image. ViLBERT, B2T2, Unicoder-VL are pre-trained on Conceptual Captions [59], which contains 3 million images with descriptions. VL-BERT not only uses Conceptual Captions for image related pre-training, but also uses BooksCorpus and English Wikipedia for language related pre-training. In video-language pre-training, both VideoBERT and CBT are pre-trained on Cooking312k [64], which is a “cooking” related video set with 312k videos and a total of 23186 hours.

4.2 Relational Reasoning on Structured Corpus

As mentioned in Sec. 2.2, the well organized commonsense knowledge bases are highly structured mainly in the form of knowledge graph. Knowledge graph contains a large number of commonsense knowledge in the form of (h, r, t), which is a triplet of head, relation and tail. Here, head and tail denotes two entities. In contrast to the pre-training methods with implicit commonsense knowledge, relational reasoning on structured corpus can make explicit use of the commonsense and get better interpretability.

4.2.1 *Rule-based Methods.*

Rule-based relational reasoning is to infer new relationships from existing relationships according to reasoning rules. In the past, reasoning rules were mainly constructed manually, which could not cover all reasoning rules. Currently, the researchers can automatically extract inference rules from the knowledge graph so as to carry out logical reasoning. Typical methods include inductive logic program (ILP) [50] and association rule mining (ARM) [21]. While rule-based method will provide a variety of reasoning rules automatically, which could be used for commonsense relational reasoning on the knowledge graphs, it would fail on the rules that have not been seen.

4.2.2 *Embedding-based Methods for Simple Relations.*

Embedding-based methods maps entities and relationships into low-dimensional space, then a scoring function is defined to measure the rationality of the facts. To model simple relations in commonsense, embedding models can be divided into two types with regard to different scoring functions: translation distance model and semantic matching model. Translation distance models like TransE [7], TransH [69], and KG2E [25] use the distance-based score functions to measure the rationality of a commonsense fact by the distance between two entities. Semantic matching models like RESCAL [48], DistMult [72], and HolE [47] use similarity-based score functions to measure the credibility of facts by matching the underlying semantics of entities and the relationships contained in the vector space representation. After obtaining the low-dimensional distributed embedding of knowledge graph, some simple relationships can be inferred by the embedding. Taking TransE [7] as an example, it takes advantage of the translation invariance of the word vector space and considers that the relationship vector carries the potential features of head

[1] A question combined with its arbitrary candidate answer can be called a Question-answer(QA) pairs.

entity to tail entity. TransE regards the relation vector r as the translation between the head entity vector h and the tail entity vector t, that is, $h + r = t$. For example, $Beijing + TheCapitalOf = China$. Therefore, the distance between $t - h$ and r can be used to estimate the possibility of the relationship r between h and t. For example, when the model find that $China - Beijing \approx America - Washington$, it can infer that Washington is the capital of America.

In summary, these embedding-based methods will map the structure of a graph to a low dimensional space, where the relationships between vectors are predicted. Pure embedding-based methods have performed well in some tasks, but fails in multi-hop commonsense reasoning.

4.2.3 Embedding-based Methods for Complex Relations. When the relationship between two entities is complex, relational reasoning may have to go through many other entities on the graph. Therefore, it is hard to capture the relationship between two entities without the information of their neighbors and other entities on the route between them. For example, DeepPath [71] applies reinforcement learning to find the best path to link two entities. KagNet [35] looks for paths with lengths shorter than 4 between two entities to construct a schema graph. The graph is encoded with GCN and each path is encoded with LSTM. The hierarchical path based attention mechanism is used to select the path that has greater impact on the QA problem for reasoning. They all solved the multi-hop reasoning problem very well.

4.3 Hybrid Methods

While self-supervised learning methods are pre-trained on unstructured corpus to provides implicit commonsense knowledge, relational reasoning methods works explicitly on structured commonsense knowledge bases. To combine the advantage of both sides, researchers seek to propose some hybrid methods for more effective commonsense learning and reasoning.

Lv et al. [39] leverage pre-trained XLNet on graph reasoning. They automatically extract evidence from heterogeneous knowledge sources, that is, from both ConceptNet and Wikipedia articles. Based on XLNet, the distance between words is refined by the structure of graph. After obtaining the contextual word representation of each word, they further use the structural information of the graph to make inference at the graph structure level.

Some studies use knowledge graphs for pre-training. KG-BERT [74] modifies the input of the BERT to make it suitable for the triplets in knowledge graphs. ERNIE-THU [81] first identifies the named entities in the text, and then matches the mentioned entities with those in the knowledge map. The structures of knowledge graphs are encoded by knowledge embedding algorithm, and multi-information entities are embedded as the input of ERNIE. However, it does not use the relational information in the knowledge graph. K-BERT [36] injects the triplet information into the sentence to form a sentence tree with rich background knowledge, and it uses visible matrix to introduce the structure information of tree into the model.

As a conclusion, self-supervised learning methods have a wider source of corpus, and do not need manual annotation. It can complete a variety of downstream commonsense reasoning tasks through pre-training and fine-tuning, but the knowledge are provided implicitly in the pre-trained models. Relational reasoning methods take advantage of the valuable structural relationships between entities and makes explicit reasoning, which is better in accuracy and interpretability, but they are restricted by the size of available commonsense knowledge graphs. As a promising research direction, hybrid methods that combines self-supervised learning and relational learning are expected to make use of both large-scale corpus and explicit reasoning procedure.

5 APPLICATIONS

As incorporating commonsense knowledge into current machine learning systems can provide more powerful reasoning ability and interpretability, many downstream multimedia applications, including recommender systems, visual understanding, and robotics, *etc.* have been benefited from deep understanding of commonsense knowledge.

5.1 Recommender System

Recommender systems recommend products to users according to their preferences. Usually, there are many items and many users in the environment, and the system will give users a clear reason to recommend, which requires the support of commonsense knowledge.

Catherine and Cohen [9] proposed to use a series of manual rules for reasoning and recommendation. Ma et al. [40] proposed a new joint learning framework, which combines the induction of interpretable rules in knowledge graphs with the construction of rule-guided neural recommendation model. The framework encourages the two modules to complement each other when generating valid and interpretable recommendations. Xian et al. [70] proposed a symbolic reasoning method called NSER, which first interprets the user behavior in a coarse-grained way, and then generates a more fine-grained explanation based on the inferential path of the knowledge graph. It has achieved good results in four evaluation metrics: normalized discounted cumulative gain, recall, hit rate, and precision.

5.2 Visual Understanding

Visual understanding is an important part of computer vision area. Commonsense knowledge could benefit many downstream visual understanding tasks such as image retrieval, scene graph generation, emotion reasoning [60], and so on.

Nowadays, the search of documents still relies mainly on text information and fails to make good use of images in documents. Chowdhury et al. [15] proposed Know2Look, which integrates visual commonsense knowledge to get more accurate picture description. They can use text and image to retrieve documents at the same time, and get better results than text-only methods.

Scene graph generation is to create a graph to represent the relationship between different objects in the scene. However, sometimes the complex environment will lead to the scene graph violating the laws of the real world, it can be corrected by the commonsense. Chen et al. [12] uses statistical methods to learn commonsense from training data, but it is limited to the frequency of scene relations appearing in data. Zareian et al. [75] extended the transformer model to incorporate the structure of the scene graph and trained a global-local attention transformer, which can automatically acquire the visual knowledge such as affront and intuitive physics, and can be applied to any scene graph generation model.

5.3 Robotics

Robot is a kind of intelligent machine which can work autonomously, which has the basic characteristics of perception, decision-making and execution. To work in a complex real environment, robots need to have some commonsense.

For example, when you assign a robot to perform a task, the robot needs to select the appropriate tool in the environment. Bansal et al. [3] propose a neural model ToolNet, which use graph neural network to encode the current environmental state and goal-conditional spatial attention to predict the most appropriate tool. When you need a robot to find something, it also requires commonsense to guess where it is in the room. Chaplot et al. [10] propose a modular system called *Goal-Oriented Semantic Exploration* , which learns the semantic prior knowledge of relative arrangement of objects in the scene and uses them to explore effectively.

6 FUTURE DIRECTION

6.1 Definition of Commonsense

The definition of commonsense knowledge shown in Sec. 1 is popular in the domain of machine learning. However, there are some other kinds of definitions. For example, in the philosophical sense, the beliefs and thoughts commonly possessed by rational and normal persons also belong to commonsense knowledge. To make machine commonsense reasoning more human-like and human-centric, commonsense knowledge should also include simple modes of reasoning shared among most people such as syllogism and analogical reasoning. A syllogism is an argument with three parts: the major premise, the minor premise and the conclusion. For example, when the major premise is "All people will die" and the minor premise is "Socrates is a man", humans can easily conclude that "Socrates will die" even though they don't know the concept of syllogism. Analogical reasoning is the process of inferring one object's attributes by comparing it to another similar object. For example, hitting your head with a stone will make you feel painful, so we can effortlessly infer that hitting with a watermelon is the same, because watermelon is also hard. Maybe existing commonsense reasoning methods can solve syllogism and analogy problems by leveraging semantic and syntactic knowledge, but we believe it's advantageous to directly incorporate these thinking commonsense into downstream tasks.

6.2 Commonsense in Multimedia

Today, we live in a multimedia world constructed by a large number of multimodal contents (text, image, video, audio, sensor data, *etc.*), which are highly relevant in specific events and applications. At present, multimodal commonsense knowledge are mostly implicit captured in pre-trained models on unstructured corpus, as shown in Sec. 4.1.2. However, the multimedia community has not yet seen available large-scale structured multi-modal knowledge graph with explicit commonsense knowledge, which is challenging not only in the data collection cost but also in the definition of the structure of *multimodal* commonsense knowledge. In the future, researchers can explore the construction of multimodal knowledge graph and use it for relational reasoning, or combine it with pre-training to give full play to their advantages.

6.3 Explicit Commonsense Reasoning

One of the core advantages of incorporating commonsense knowledge into the model is stronger interpretability, which could be user-friendly in many downstream tasks. However, this aspect is not good enough at present. Although self-supervised learning obtain commonsense knowledge from large-scale corpus, the huge amount of parameters and the black-box architecture make researchers unknown about what it has learned and how to explicitly apply it to downstream tasks. In the domain of relational learning from structured corpus, most embedding-based methods would lost a lot of accurate and structured semantic information and weaken the unique advantages of knowledge graph. Machine learning based on statistics can only be weak artificial intelligence, which is good at finding correlation but weak in logical reasoning. Therefore, on the way towards artificial general intelligence, it is worth exploring to make explicit commonsense reasoning with high interpretability.

7 CONCLUSION

This article presents a comprehensive survey on commonsense learning and reasoning – how the AI systems acquire commonsense knowledge and utilize commonsense for reasoning. In this survey, we first summarize the commonsense resources and evaluation benchmarks, then we review mainstream state-of-the-art methods for incorporating commonsense knowledge. We also discuss several popular applications with commonsense. Throughout this article, we focus more on multimedia commonsense in contrast with existing literature. In addition to the descriptive review, we discuss several promising directions for future research. In particular, we suggest incorporating other modes of commonsense knowledge, e.g. syllogism, into downstream tasks, as well as multimedia knowledge base construction and interpretable commonsense reasoning. We believe this article will help readers to build a big picture of commonsense learning and benefit the field for future research.

8 ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China Major Project (No. U1611461), National Key R&D Program of China under Grand No. 2018AAA0102000.

REFERENCES

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of Detected Objects in Text for Visual Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2131–2140.

[2] Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

[3] Rajas Bansal, Shreshth Tuli, Rohan Paul, et al. 2020. ToolNet: Using Commonsense Generalization for Predicting Tool Use for Robot Plan Synthesis. *arXiv preprint arXiv:2006.05478* (2020).

[4] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning.. In *ICLR*.

[5] Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language.. In *AAAI*. 7432–7439.

[6] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human

knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1247–1250.

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*. 2787–2795.

[8] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[9] Rose Catherine and William Cohen. 2016. Personalized Recommendations using Knowledge Graphs: A Probabilistic Logic Programming Approach. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 325–332.

[10] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. 2020. Object Goal Navigation using Goal-Oriented Semantic Exploration. *arXiv preprint arXiv:2007.00643* (2020).

[11] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[12] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6163–6171.

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740* (2019).

[15] Sreyasi Nag Chowdhury, Niket Tandon, and Gerhard Weikum. 2016. Know2Look: Commonsense Knowledge for Visual Search. *Proceedings of AKBC* (2016), 57–62.

[16] Yung-Sung Chuang, Chi-Liang Liu, and Hung-Yi Lee. 2019. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559* (2019).

[17] Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58, 9 (2015), 92–103.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[20] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*. 1422–1430.

[21] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. 2015. Fast rule mining in ontological knowledge bases with AMIE. *The VLDB Journal* 24, 6 (2015), 707–730.

[22] MY Ganaie and Hafiz Mudasir. 2015. A study of social intelligence & academic achievement of college students of district Srinagar, J&K, India. *Journal of American Science* 11, 3 (2015), 23–27.

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

[24] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[25] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 623–632.

[26] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2391–2401.

[27] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20, 8 (2016), 589–604.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[29] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 252–262.

[30] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

[31] Douglas B Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38, 11 (1995), 33–38.

[32] Douglas B Lenat, Mayank Prakash, and Mary Shepherd. 1985. CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI magazine* 6, 4 (1985), 65–65.

[33] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training.. In *AAAI*. 11336–11344.

[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[35] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2822–2832.

[36] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling Language Representation with Knowledge Graph.. In *AAAI*. 2901–2908.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[39] Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering.. In *AAAI*. 8449–8456.

[40] Weizhi Ma, Min Zhang, Yue Cao, Woojeong Jin, Chenyang Wang, Yiqun Liu, Shaoping Ma, and Xiang Ren. 2019. Jointly learning explainable rules for recommendation with knowledge graph. In *The World Wide Web Conference*. 1210–1221.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[43] George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

[44] T Mitchell, W Cohen, E Hruschka, P Talukdar, J Betteridge, A Carlson, B Dalvi, M Gardner, B Kisiel, J Krishnamurthy, et al. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2302–2310.

[45] Leora Morgenstern and Charles Ortiz. 2015. The winograd schema challenge: Evaluating progress in commonsense reasoning. In *Twenty-Seventh IAAI Conference*.

[46] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 839–849.

[47] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *AAAI*. 1955–1961.

[48] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 809–816.

[49] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*. 2227–2237.

[50] JR QUINLAN. 1990. Learning Logical Definitions from Relations. *Machine Learning* 5 (1990), 239–266.

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. [n.d.]. Improving Language Understanding by Generative Pre-Training. ([n. d.]).

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[53] Deepak Ramachandran, Pace Reagan, and Keith Goolsbey. 2005. First-orderized researchcyc: Expressivity and efficiency in a common-sense ontology. In *AAAI workshop on contexts and ontologies: theory, practice and applications*. 33–40.

[54] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

[55] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint*

arXiv:1907.10641 (2019).

[56] 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*, Vol. 33. 3027–3035.

[57] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4453–4463.

[58] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. ACL 2020 Commonsense Tutorial. *ACL2020 Tutorial* (2020). https://homes.cs.washington.edu/~msap/acl2020-commonsense/

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[60] Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. MEmoR: A Dataset for Multimodal Emotion Reasoning in Videos. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[61] Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*. Springer, 161–176.

[62] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[63] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. [n.d.]. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743* 3, 5 ([n. d.]).

[64] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. [n.d.]. VideoBERT: A Joint Model for Video and Language Representation Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 7463–7472.

[65] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*. 4149–4158.

[66] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5103–5114.

[67] Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*. 115–120.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[69] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes.. In *AAAI*, Vol. 14. 1112–1119.

[70] Yikun Xian, Zuohui Fu, Qiaoying Huang, Shan Muthukrishnan, and Yongfeng Zhang. 2020. Neural-Symbolic Reasoning over Knowledge Graph for Multi-stage Explainable Recommendation. *arXiv preprint arXiv:2007.13207* (2020).

[71] Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 564–573.

[72] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*. 1–13.

[73] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5753–5763.

[74] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193* (2019).

[75] Alireza Zareian, Haoxuan You, Zhecan Wang, and Shih-Fu Chang. 2020. Learning Visual Commonsense for Robust Scene Graph Generation. *arXiv preprint arXiv:2006.09623* (2020).

[76] 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*. 6720–6731.

[77] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 93–104.

[78] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4791–4800.

[79] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885* (2018).

[80] Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics* 5 (2017), 379–395.

[81] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1441–1451.

[82] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3354–3360.

[83] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *The IEEE International Conference on Computer Vision (ICCV)*.

[84] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.