A systematic comparison of commonsense knowledge usages between natural language processing (NLP) and computer vision (CV)

Adrian Kuiper
Supervisor(s): Gaole He, Jie Yang, Ujwal Gadijaru, Graduation Committee
EEMCS, Delft University of Technology, The Netherlands

24-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Commonsense knowledge is the key of human intelligence in generalizing their knowledge to deal with complex tasks. Over the past years, a lot of research has been done in both natural language processing (NLP) and computer vision (CV) on leveraging commonsense knowledge to improve AI models. However, no systematic comparisons of existing work have been made between the two domains. Therefore this survey aims to provide an overview of how commonsense knowledge is used within NLP and CV and how research varies between these two domains and what future challenges it may hold. An observation made from this survey is that leveraging commonsense is more difficult in CV than NLP, as commonsense is mostly incorporated textually and datasets need to be filtered to make them more relevant for visual commonsense. We hope to promote further research and create a better understanding of commonsense knowledge and its applications with this survey.

1 Introduction

Commonsense knowledge is an important topic within artificial intelligence (AI), especially in the areas of natural language processing and computer vision. Commonsense knowledge is knowledge that everybody is expected to know. It can normally be left out when conveying a message. Take an example “Lemons are sour”, people are expected to know this and therefore it can be left out when talking about lemons. This makes it difficult for AI-based models to possess commonsense knowledge, as it is not always available in the data [Ilievski et al., 2021]. However, AI models can benefit a lot from commonsense knowledge, like allowing them to reason about certain situations and explain how they come to an answer [Lin et al., 2019], improving machine translation [Vilares et al., 2018] or even supporting robot manipulation and navigation [Zhu et al., 2019].

In recent years a lot of research on commonsense knowledge is done, resulting in the creation of commonsense benchmarks like CommonsenseQA [Talmor et al., 2018] and VCR [Zellers et al., 2019]. These benchmarks are commonsense datasets where state-of-the-art language models like BERT [Devlin et al., 2018] are tested and trained on, to test their accuracy on particular commonsense tasks. Besides data sets, other benchmarks consist of models like TopicKA [Wu et al., 2020] or R2C[Zellers et al., 2019], which are developed for all kinds of commonsense tasks like Dialogue or VCR . However, although there is a lot of existing work in commonsense, no comparisons are made yet between the two domains.

In this survey, the main research question is “How does research related to commonsense knowledge vary across natural language processing (NLP) and computer vision (CV)”. As these are two major fields in artificial intelligence, more research is done on commonsense in these areas. However, right now no prominent surveys to evaluate and compare the existing NLP/CV methods exist, which serves as motivation for this survey. To answer the main research question, we divided it into two subquestions: (1) What kind of different NLP/CV tasks incorporate commonsense, and (2) what are the specific usages of commonsense knowledge within these different tasks?

Therefore the goal of this survey is to give an overview of the different NLP and CV tasks that incorporate commonsense and how this commonsense is specifically used within these tasks (Figure 1 for a high-level overview). We see that when comparing NLP to CV, there exists a lot more work in NLP, because commonsense is easier to incorporate textually and that incorporating commonsense in a lot of CV tasks goes in combination with NLP. In this survey, an analysis is made to compare the difference of each usage within NLP or CV to finally promote further research of commonsense knowledge with AI.

The structure of the paper is as follows. First, the methodology will show how and what kind of papers were collected and organized and how the research was done. Secondly, there will be a section for preliminary knowledge, explaining the different NLP/CV tasks. After this the survey will go more into detail about the usages of commonsense knowledge and their existing work; This survey will end off with a discussion and a final conclusion;

2 Methodology

In this section, a summary of the gathering process, the organization of research and the procedure of the survey will be provided.

2.1 Gathering papers

At the start of the survey, papers needed to be gathered related to the research question. Therefore the research question was split into multiple sub-questions to get a clearer idea of what kind of papers were relevant. For gathering papers, important questions were: “What kind of different NLP/CV tasks exist that incorporate commonsense?”, and “How does existing work of commonsense look like in these tasks?”. With these questions in mind, other surveys of NLP and CV were looked up to get a better overview of the tasks within these fields. Queries were then sent on specific keywords, for example, NLP tasks combined with commonsense, towards databases like Google Scholar and Scopus. Examples of keywords were *commonsense, natural language processing, computer vision, NLP tasks (§3.1), CV tasks (§3.2), commonsense benchmarks (§4)* and their abbreviations. See an example of the two main queries to start with below.

```
(commonsense* OR common sense*) AND (NLP OR
Natural Language Processing OR Natural Language)
```

```
(commonsense* OR common sense*) AND (CV OR
Computer Vision OR computer vision)
```

After finding papers from this query, we can move on to more advanced queries and search for specific tasks and models. Based on the number of citations, publication year, and source, papers were selected to be read. By reading the

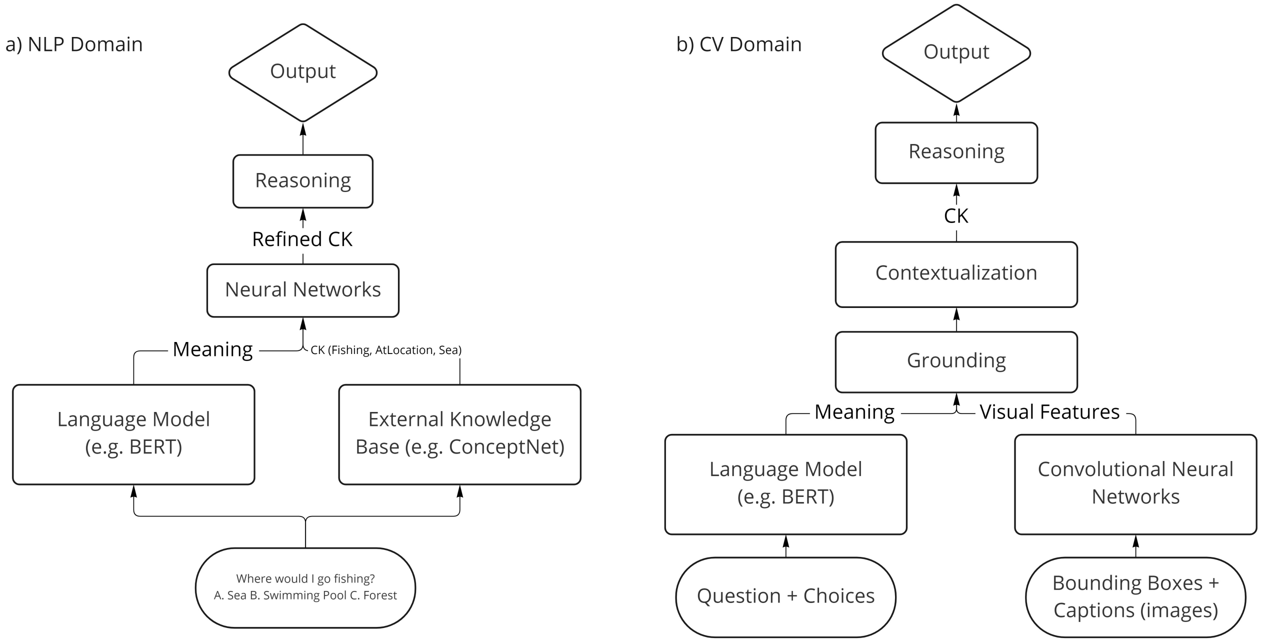

a) NLP Domain
Output
Reasoning
Refined CK
Neural Networks
Meaning
CK (Fishing, AtLocation, Sea)
Language Model (e.g. BERT)
External Knowledge Base (e.g. ConceptNet)
Where would I go fishing?
A. Sea B. Swimming Pool C. Forest
b) CV Domain
Output
Reasoning
CK
Contextualization
Grounding
Meaning
Visual Features
Language Model (e.g. BERT)
Convolutional Neural Networks
Question + Choices
Bounding Boxes + Captions (images)

Figure 1: a) High-level overview of an example of how commonsense knowledge can be incorporated within NLP domain. The question and answers are put through the language model and relational concepts (commonsense triples) are found within an external knowledge base. These results are then combined and refined by neural networks to reason over it and eventually predict the outcome. b) High-level overview of commonsense in CV domain performing VCR task [Zellers et al., 2019]. Text and images are first grounded together, then relations between them are contextualized and finally, this is used to reason over the query, response, and images.

abstract, introduction, and conclusion, papers were finally deemed to be relevant.

2.2 Organization

When organizing the collected papers, initially, Google Docs was used and papers were divided into sections for NLP, CV, and Surveys. After this, papers were read more in detail and put in Mendeley Reference Manager. Mendeley is a tool to organize papers and there is the possibility to highlight, annotate and make remarks on specific parts of the paper. Another feature is tagging, where you can manually put multiple tags on a paper, which will then be categorized and can be sorted on their tags later on. In Mendeley important parts were highlighted and papers were tagged, indicating (1) What NLP/CV task is used, (2) How users involve explicit commonsense knowledge in this research, and (3) What implications/observations are provided. With these steps, the organization of the research was concluded.

2.3 Survey procedure

To answer the research question, the focus of this survey came to lie on the specific usages of commonsense instead of on the different NLP/CV tasks that incorporated commonsense. The survey summarized the existing work of the different usages and analyzed how these usages were used within NLP and CV in general, instead of categorizing them per task. From this analysis, a discussion was formed and a conclusion was drawn.

NLP	CV
Question Answering (QA)	Scene Graph Generation (SGG)
Machine reading Comprehension (MRC)	Visual Question Answering (VQA)
Generating NLE	Visual Commonsense Reasoning (VCR)
Dialogue	Vision-Language Navigation (VLN)
Natural Language Inference (NLI)	
Machine Translation (MT)	

Table 1: Overview of the different NLP and CV tasks covered by this survey and their abbreviations.

3 Preliminary Knowledge

In the preliminary knowledge, different NLP and CV tasks are explained to get a better picture of how usages of commonsense knowledge can be connected with these tasks (Table 1). A thing to notice is that these tasks do not necessarily cover every task leveraging commonsense knowledge, but are deemed the most relevant by this survey.

3.1 NLP tasks

Natural Language Processing is all about the relation of human language with computers. Commonsense knowledge in this area is very important, as text can be ambiguous [Davis and Marcus, 2015]. Therefore real-life knowledge is needed to let AI models understand ambiguity and other data where commonsense is required. The following part aims to give an overview of the different NLP tasks that incorporate commonsense knowledge.

Question Answering
Question Answering (QA) is one of the major tasks in NLP and can benefit a lot from commonsense. In QA human-posed questions are automatically answered by the model. As humans mostly use their commonsense knowledge to answer questions, the QA task comes back often when commonsense is researched.

Machine Reading Comprehension
The goal of Machine Reading Comprehension (MRC) is for machines to learn how to read and understand human languages [Zhang et al., 2020]. To test the MRC ability of machines, researchers often utilize cloze-style questions [Mostafazadeh et al., 2017]. These questions contain a placeholder, where the machine should choose which word or sentence is the most suitable.

Generating NLE
Natural Language Explanations (NLE) are an important part of NLP. Generating NLEs can help us gain a better understanding of the predictions made by a neural model. An example of this is users getting an insight into why something is recommended to them [Chen et al., 2021].

Dialogue
Dialogue is a fundamental part of natural language. Nowadays dialogue systems like Alexia, Siri, online website assistants, and many more exist. These systems can support us by generating the right response. Think about examples such as automated customer service or simply finding the right directions.

Natural Language Inference
One of the main tasks in NLP is Natural Language Inference (NLI). The goal is to decide, given a premise, if a hypothesis is true, false or undetermined [Ruder, 2019].

Machine Translation
In natural language processing machine translation (MT) is an important task that translates natural languages using computers. [Tan et al., 2020]

3.2 CV tasks

Computer Vision is a field that focuses on identifying and understanding images and videos. Humans can automatically reason about objects in images, like when knives are sticking on the kitchen wall, humans can deduce that it is probably due to a magnet. For AI models in CV, this is more difficult to deduce. Therefore commonsense knowledge is needed in CV. In the following paragraphs, different CV tasks that incorporate commonsense are introduced.

Scene Graph Generation
The goal of Scene Graph Generation is to gain a higher-level understanding of visual scenes and have the ability to reason about it [Chang et al., 2022]. This is done by labeling objects and relations in the images and mapping them all together towards a structured scene graph.

Visual Question Answering
As the name says, Visual Question Answering (VQA) is the same as Question Answering, only images are now included. Given an image and a question, an answer needs to be generated. Therefore this task is also actually a combination between NLP and CV.

Visual Commonsense Reasoning
Visual Commonsense Reasoning (VCR) is one of the more prominent tasks incorporating commonsense. It is a task like VQA, however, after answering a question, a model must also provide a rationale as to why they chose the answer [Zellers et al., 2019].

Vision-Language Navigation
Vision-Language Navigation (VLN) is a CV task, where agents need to navigate through their surroundings by following human language instructions [Wu et al., 2021]. Commonsense is needed here, as agents should not walk over a table for example but around it.

4 Usages of commonsense knowledge

In this section, the focus is on the different usages of commonsense knowledge in NLP and CV (Figure 2). We are going to talk about the terminology of the usages and what existing work looks like. The terminology is made by taking the usages literally out of the papers. Usages like pre-training and reasoning are the collective name of different usages that essentially fall under the same category. Other usages like answering questions and response generation are divided based on the difference in methods for these usages. Following that, we will talk about the difference between the usages themselves, when comparing their research in NLP to CV.

4.1 Answering a question

One of the main usages of commonsense knowledge in AI models is using this knowledge to answer questions. Researchers came up with many benchmarks to improve question answering. These benchmarks consist of frameworks and datasets.

Graph-based Models: A textual inference framework like KagNet [Lin et al., 2019], is used to incorporate commonsense knowledge for question answering. This framework measures the plausibility of each answer by first grounding question-answer sets as a schema graph extracted from external knowledge graphs like ConceptNet [Speer et al., 2016]. Following this, the framework uses graph convolutional networks (GCNs) [Kipf and Welling, 2016] to encode the schema graphs and find relational paths between questions and answers. With these relational paths, the plausibility of each answer can finally be measured. Another approach is graph-based reasoning [Lv et al., 2020]. This approach consists of two graph-based modules. One contextual representation learning module to find the relative position between words and an inference model using GCNs, like KagNet. The previous two models mainly focused on solely textual question answering, while there is also an approach

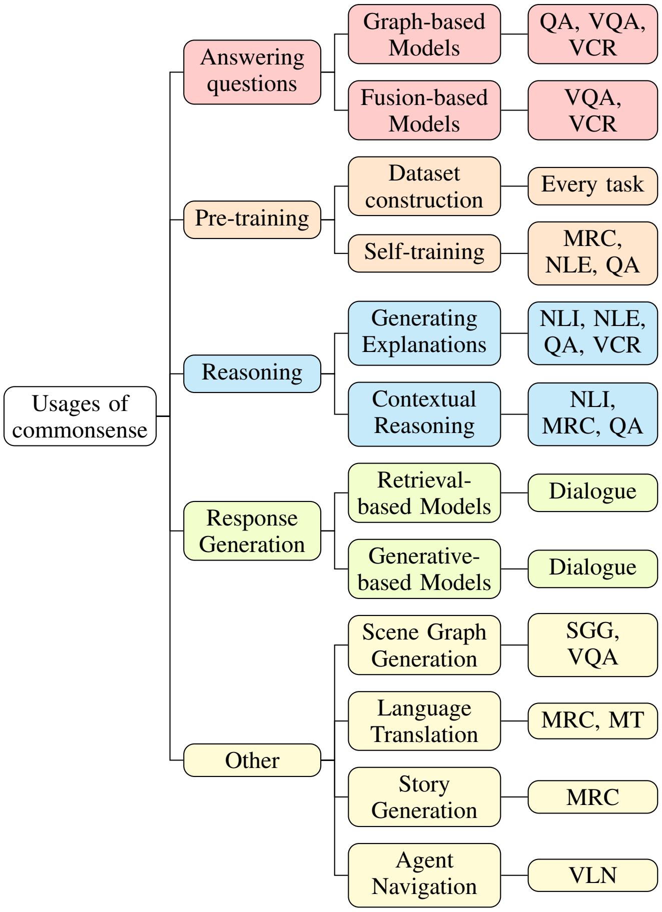


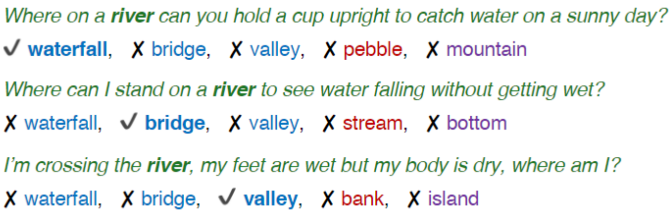
Figure 2: The main usages of commonsense and their subcategories together with the corresponding NLP/CV tasks that are performed with these usages.

that focuses on VQA (§3). This approach utilizes scene description graphs (SDGs) [Aditya et al., 2018] to answer questions related to images. These SDGs are obtained by a deep learning-based perception model that combines the images with commonsense extracted from a knowledge base. After this with the detailed information in the SDGs, questions can be answered. However, in practice, SDGs are not always as informative compared to other existing methods.

Fusion-based Models: There exist other frameworks that use a different method than graphs to improve question answering with commonsense knowledge specifically for computer vision. "Bounding Boxes in Text Transformer" (B2T2) [Alberti et al., 2019] is a model that classifies answers by combining natural language and vision and is developed for Visual Question Answering (§3). B2T2 is an early fusion architecture where text, image, and bounding boxes are embedded at the same level as word tokens, before the classification of the answers. This model showed better performance on the VCR task (§3) than the "Dual Encoder" [Alberti et al., 2019], another architecture that does not make use of bounding boxes. R2C [Zellers et al., 2019] is another model that can perform the VQA task, which will be mentioned more in the reasoning paragraph.

Datasets: Datasets have also been developed for improving question answering. Language models like BERT [Devlin et al., 2018] are then pre-trained on these datasets to perform question answering with commonsense. A typical dataset is CommonsenseQA [Talmor et al., 2018]. This is a dataset containing commonsense-required questions, where there are multiple answers related to the concept and a few are not. Due to having a few distractors, researchers can test the commonsense in a model (Figure 3). Other examples of datasets that are developed for question answering are PIQA [Bisk et al., 2020], a dataset that focuses on physical commonsense, and CosmosQA [Huang et al., 2019], a dataset focused on answering questions through machine reading comprehension.

Analysis: What we observe from existing work is that for NLP-based models, graph-based methods are more dominant, when researching the QA task with commonsense knowledge. For CV-based models, transformers and classifiers prove to be effective when performing the VQA task, while graph-based methods are less explored on CV tasks as they show worse performance. However, due to recent graph-based methods working well for NLP models, future research will try to build further on these methods and incorporate them for CV tasks as mentioned in KagNet [Lin et al., 2019].



*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 3: In this figure we see a few examples of the dataset commonsenseQA [Talmor et al., 2018]. Here we see a commonsense-required question, with a few answers. The answers a,b,and c are related to the concept, while the answers d and e serve as distractors, answers that do not have any relation with the concept, to make the task more difficult.

4.2 Pre-training

Pre-training is something that happens in every NLP or CV task. Models are pre-trained to perform a specific task like question answering on a dataset. There are different types of pre-training; Datasets are constructed for pre-training containing examples of commonsense knowledge from external knowledge bases like ConceptNet [Speer et al., 2016]; While there are also models that use methods to train themselves.

Dataset Construction Methods: There are several methods that exist to construct datasets for pre-training; A common way to construct datasets is by using crowd sourcing methods. Crowd-sourcing workers are asked to author commonsense examples for the dataset. Examples of datasets constructed by crowd-sourcing are CommonsenseQA (§4.1), SocialIQA [Sap et al., 2019], a dataset for social interactions and CoS-E [Rajani et al., 2019]. Next to crowd-sourcing different methods like AMS [Ye et al., 2019] exist (Figure 4). This method constructs a multiple-choice (MC) dataset by first aligning sentences with triples from ConceptNet, then masking the alignments as questions and finally selecting some distractors for the questions. By doing this, a MC dataset is constructed to pre-train language models for improving commonsense abilities, while their language representation skills are not sacrificed for this improvement. Another method is Knowledge-based Commonsense Generation (KCG) [Xing et al., 2021]. This method is used to pre-train models to perform the VCG task [Park et al., 2020]. KCG makes use of generated commonsense from COMET [Bosselut et al., 2019], a model pre-trained on commonsense knowledge graphs (CSKGs) [Ilievski et al., 2020]. However COMET only uses textual information and no visual information. Therefore a self-training data filtering method is applied to filter out examples that resemble the VCG dataset to finally construct the pre-training dataset. An example of another filtering method is used in the VCR dataset [Zellers et al., 2019]. VCR uses an *interestingness* filter, interestingness meaning complex examples that humans can solve without additional context.

Self-training: Although KCG in the previous paragraph showed a form of self-training, there are also other forms of self-training. An example of this is self-talk [Shwartz et al., 2020]. Given some context, this model generates clarification questions and answers to incorporate as additional context. In this way, the model trains itself to better understand and answer questions and therefore does not need to depend on external knowledge or supervision.

Analysis: In both NLP and CV models, datasets are constructed to pre-train them on specific tasks. And the methods for creating these datasets, like crowdsourcing and using transformers mostly overlap. However, there are occasions that CV models require filtering on their pre-training dataset to get relevant examples. This is because commonsense is often incorporated textually and therefore examples need to be selected to also be relevant to the images.

4.3 Reasoning

Reasoning is maybe the most important usage of commonsense knowledge. Every commonsense task essentially comes back to reasoning. Models must reason to get a better understanding of context and to answer questions correctly where commonsense is required. However, when answering questions we cannot say for sure that models contain commonsense, as they can also choose an answer randomly or for the wrong reason. Therefore by adding a reasoning task, it is possible to measure commonsense within a model with more certainty.

Datasets, frameworks, and models like R2C [Zellers et al., 2019] are made to incorporate commonsense for reasoning. R2C for example is a model specifically developed for the VCR task (§3). This model first grounds the images and queries and then contextualizes them together to finally reason over the relationships between image, query, and response, where the reasoning is used as additional context to come to a prediction. Next to this, there are other usages of commonsense within the reasoning area.

(1) A triple from ConceptNet
(population, AtLocation, city)
(2) **Align** with the English Wikipedia dataset to obtain a sentence containing "population" and "city"
The largest **city** by **population** is Birmingham, which has long been the most industrialized city.
(3) **Mask** "city" with a special token "[QW]"
The largest **[QW]** by **population** is Birmingham, which has long been the most industrialized city?
4) **Select** distractors by searching (population, AtLocation, $*$) in ConceptNet
(population, AtLocation, Michigan) (population, AtLocation, Petrie dish) (population, AtLocation, area with people inhabiting) (population, AtLocation, country)
5) Generate a multi-choice question answering sample
question: The largest **[QW]** by **population** is Birmingham, which has long been the most industrialized city? **candidates**: *city*, Michigan, Petrie dish, area with people inhabiting, country

Figure 4: An example is taken from the dataset construction method AMS [Ye et al., 2019]. First commonsense triples extracted from ConceptNet [Speer et al., 2016] are aligned with sentences. After this the aligned concepts are masked and distractors (Figure 3) are selected to finally generate an MC question answering sample. By using this method, language models preserve their language representation skills, while improving their commonsense abilities.

Generating Explanations: One usage of commonsense knowledge in reasoning is generating explanations that can be incorporated as additional context. CAGE [Rajani et al., 2019] is a framework developed for generating explanations on the CommonsenseQA dataset. CAGE consists of 2 phases; In the first phase, models are trained to generate NLE (§3) with examples from the CoS-E [Rajani et al., 2019] dataset, a dataset containing questions and answers from CommonsenseQA and their corresponding explanation for the correct answer. In the second phase, explanations that are generated by the pre-trained model, are provided to another model to help them with their predictions. A more recent framework is RExC [Majumder et al., 2021]. RExC uses self-extracted rationales to incorporate commonsense knowledge and generate fitting NLEs that help in prediction making. These rationales are extracted from the input and used in queries to knowledge modules. This framework works on NLP tasks, like ComVE [Wang et al., 2020], a commonsense validation and explanation task, and CoS-E as well as on CV tasks like VCR .

Contextualized Reasoning: Contextualized reasoning means when given a context like a paragraph, reason about the causality of events or facts about people, etc. Essentially it requires machine reading comprehension (§3). Datasets and models have been constructed to support commonsense in reading comprehension. Examples of datasets are ReCoRD [Zhang et al., 2018] and CosmosQA [Huang et al., 2019]. ReCoRD is a dataset containing passages from news articles and therefore reduces the bias. A cloze-style query [Mostafazadeh et al., 2017] that is supported by the passage is posed with an X in it. And the goal is to reason what best fits the X. CosmosQA is a dataset that contains examples that are less "exact", meaning less focus on facts and literal understanding and instead focusing on people's everyday situations, asking what-if questions. Therefore to answer these questions, reasoning ability is required and with these datasets, existing language models are trained and commonsense is incorporated.

Analysis: When looking at models and frameworks, the more recent work for generating additional contexts like explanations and reasons does well on both NLP and CV tasks. However, when trying to reason without factual context, research for now is still focused on textual areas, with datasets constructed for improving machine reading comprehension.

4.4 Response generation

Response Generation is seen in dialogue systems and to create human-like conversations, it can benefit a lot from commonsense knowledge. There are two types of conversational models where commonsense is used differently, retrieval-based and generative-based models.

Retrieval-based Models: In retrieval-based models, commonsense knowledge is used to select a suitable response from a predefined repository. An example of this is using commonsense knowledge in the form of external memory to augment dialogue systems [Young et al., 2017]. In this work, a match score is defined to see how well commonsense assertions go with their responses and only the highest score is taken into account when selecting an appropriate response. This is done by using a Tri-LSTM encoder to encode commonsense assertions, messages, and responses.

Generative-based Models: In generative-based models, commonsense knowledge is used to generate new conversations. A commonsense knowledge-aware conversational model (CCM) is an example of a generative-based model [Zhou et al., 2018]. This model uses static and dynamic graph attention mechanisms to understand the posts and generate better responses. These are mechanisms that allow the model to focus on an input one at a time. Another example is TopicKA [Wu et al., 2020], a model developed for open-domain dialogue. TopicKA generates responses that are conditioned on the query and a topic fact, which are commonsense facts related to the query and the response.

Analysis: Current existing work on response generation incorporated with commonsense knowledge only focuses on responses based on natural language dialogues. Dialogue itself is an NLP task, although response generation can also be in the form of responding to or with an image or video. However, currently, there is no research available on response generation in CV. Future research may be generating responses based on actions seen in visuals.

4.5 Other usages

In the previous sections, four main usages of commonsense knowledge were mentioned together with their existing work. However, there are also other usages of commonsense knowledge that are less prominently researched. These usages will be mentioned short below.

Scene Graph Generation: Scene graph generation is one of the tasks within CV where commonsense is researched. However, it is not always clear whether the perception or commonsense in scene graph generation is correct or not. Therefore GLAT [Zareian et al., 2020] was proposed, a method for generating plausible scene graphs. First commonsense is acquired from annotated scene graphs and then both perception and commonsense models are fused to support each other to correct obvious mistakes and finally generate plausible scene graphs.

Language Translation: Machine translation is an area that can benefit a lot from commonsense. For example, when translating fast food, you could also get the translation of quick food. Although this translation is correct, the meaning is different. Knowledge bases like BabelSenticNet [Vilares et al., 2018] and test suites like [He et al., 2020] are developed to incorporate and promote future research of commonsense in machine translation.

Next to machine translation, other forms of translation like translating videos into captions exist. An example model of this is Video2Commonsense [Fang et al., 2020]. In this model first global representations of a video are encoded and then decoded by a transformer that generates commonsense caption and finally a cross-modal self-attention model is used to capitalize on joint visual-text embeddings.

Story Generation: Although different forms of language generation like response generation and explanation generation were mentioned earlier in this survey, another less researched form is story generation. To improve story generation with commonsense, a model was developed with an incremental encoding scheme and a multi-source attention mechanism [Guan et al., 2018]. Context clues are built up by incremental reading and commonsense is gathered from knowledge graphs through the mechanism.

Agent Navigation: For VLN (§3), the AuxRN [Zhu et al., 2019] framework was developed to improve navigation learning for agents. This framework consists of 4 phases: Explaining the previous actions, evaluating the navigation progress, predicting the next move, and aligning vision and language encoding. Using commonsense, these tasks can improve a lot.

5 Discussion

In the discussion section, different observations are highlighted and interesting statements made from these observations are discussed.

Reused methods: Commonsense knowledge is used for a variety of tasks within both natural language processing and computer vision. For a lot of commonsense usages, existing work is seen in both NLP and CV fields. One thing observed from reading papers was that earlier papers on NLP-based commonsense methods like KagNet [Lin et al., 2019] mentioned that future directions regarding their research included using their work in a visual context. From this, we can reason that when research is proven to be useful, there will also be an attempt to integrate it into CV tasks. This is also a commonly seen phenomenon in research, that when methods work in particular fields, they are also applied to other fields and then put off as new work.

Amount existing work: Another observation made when looking only at the amount of existing work in both NLP and CV, was that more research existed for NLP tasks. This is an interesting topic, as both fields are important within artificial intelligence. Although it may be because of what was discussed in the previous paragraph, there could be multiple reasons for this. One reason could be that it is a lot easier to incorporate commonsense in NLP models. This is because commonsense is mostly incorporated textually and aligning the encodings for both text and images together is more difficult. Another observation to support this is that there are hardly any usages of commonsense, solely for CV models and if there are, they still make use of textual commonsense, like VLN [Zhu et al., 2019] or [Zareian et al., 2020]. Finally, images and videos on their own contain a very rich amount of information, therefore commonsense knowledge may not be needed to reason about or caption everything that is happening in a visual.

Bias: Furthermore a case to consider is existing bias in commonsense datasets. Models are pre-trained on datasets, however, a lot of datasets are constructed by crowd-sourcing. This means that there could be a possibility that examples contain bias from crowd-workers. Although commonsense is regarded as knowledge that every human should know, it can still differ in some cases due to for example cultural differences or language use. Commonsense in America can not always be the same as commonsense in Asian countries. This may be a point to take into account when constructing such datasets.

Future research domains: In the future there could be several extra challenges when researching commonsense knowledge. Currently, existing research is more focused on natural language processing and computer vision, however, there are many more areas where commonsense can be applied. Just like humans, when commonsense is fully applied in AI, it can use commonsense for example to recognize audio fragments, or maybe in the far future robots can even recognize objects by touch using commonsense. For audio there already exists some research like [Zhang et al., 2021], although not much. It is an interesting challenge for the future to get AI using commonsense just like humans do, in a way that when AI is developed to use the six senses, commonsense can be involved.

6 Conclusions and Future Work

This survey aims to give an overview of what commonsense knowledge is, and how commonsense is used within NLP and CV. Furthermore, this survey analyzes the differences in commonsense usages in both fields. From this survey, the main findings were, that there are four main usages of commonsense knowledge: Answering questions, pre-training models, reasoning, and response generation. These are the usages of commonsense that we found had the most research about themselves. In this research, we see that most commonsense usages have existing work in both NLP and CV fields, however, the implementation of these usages has some differences in both fields. Specifically for pre-training, often when creating datasets for CV tasks the datasets will have an extra filtering process to make examples more relevant for visuals. For other usages, implementations have great similarities in both fields, although there are other usages than the four main ones, which only happen in one field. The main difference between these implementations is that research in NLP is mostly more advanced than in CV, as it is easier to incorporate commonsense textually. However, we see research on CV catching up while building on research already done in NLP. We hope that this survey will provide other researchers with a better understanding of existing work in NLP and CV and hope that this will support future research of commonsense and its applications.

7 Responsible Research

As this research solely contains a survey on commonsense and its existing work, only a literature review is needed and no experiments were carried out outside of the review. However, the existing works mentioned in the survey contain some ethical aspects in their research. When creating datasets for pre-training, it must be made sure that the examples do not contain any bias. Furthermore, every paper that is used in this survey is cited at least once, to make clear what the source is. And finally, this literature review is easily reproducible, as all the steps are described in the methodology section (§2).

References

[Aditya et al., 2018] Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., and Fermüller, C. (2018). Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45.

[Alberti et al., 2019] Alberti, C., Ling, J., Collins, M., and Reitter, D. (2019). Fusion of detected objects in text for visual question answering.

[Bisk et al., 2020] Bisk, Y., Zellers, R., bras, R. L., Gao, J., and Choi, Y. (2020). Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

[Bosselut et al., 2019] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). Comet: Commonsense transformers for automatic knowledge graph construction.

[Chang et al., 2022] Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., and Hauptmann, A. G. (2022). A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

[Chen et al., 2021] Chen, H., Chen, X., Shi, S., and Zhang, Y. (2021). Generate natural language explanations for recommendation.

[Davis and Marcus, 2015] Davis, E. and Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58:92–103.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., Google, K. T., and Language, A. I. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding.

[Fang et al., 2020] Fang, Z., Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. (2020). Video2commonsense: Generating commonsense descriptions to enrich video captioning.

[Guan et al., 2018] Guan, J., Wang, Y., and Huang, M. (2018). Story ending generation with incremental encoding and commonsense knowledge.

[He et al., 2020] He, J., Wang, T., Xiong, D., and Liu, Q. (2020). The box is in the pen: Evaluating commonsense reasoning in neural machine translation. pages 3662–3672. Association for Computational Linguistics.

[Huang et al., 2019] Huang, L., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

[Ilievski et al., 2021] Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., and Szekely, P. (2021). Dimensions of commonsense knowledge.

[Ilievski et al., 2020] Ilievski, F., Szekely, P., and Zhang, B. (2020). Cskg: The commonsense knowledge graph.

[Kipf and Welling, 2016] Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks.

[Lin et al., 2019] Lin, B. Y., Chen, X., Chen, J., and Ren, X. (2019). Kagnet: Knowledge-aware graph networks for commonsense reasoning.

[Lv et al., 2020] Lv, S., Guo, D., Xu, J., Tang, D., Duan, N., Gong, M., Shou, L., Jiang, D., Cao, G., and Hu, S. (2020). Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8449–8456.

[Majumder et al., 2021] Majumder, B. P., Camburu, O.-M., Lukasiewicz, T., and McAuley, J. (2021). Rationale-inspired natural language explanations with commonsense.

[Mostafazadeh et al., 2017] Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. F. (2017). Lsdsem 2017 shared task: The story cloze test.

[Park et al., 2020] Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., and Choi, Y. (2020). Visualcomet: Reasoning about the dynamic context of a still image.

[Rajani et al., 2019] Rajani, N. F., McCann, B., Xiong, C., and Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning.

[Ruder, 2019] Ruder, S. (2019). NLP-progress/natural$_{l}anguage_{i}nference.md.$

[Sap et al., 2019] Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. (2019). Socialiqa: Commonsense reasoning about social interactions.

[Shwartz et al., 2020] Shwartz, V., West, P., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). Unsupervised commonsense question answering with self-talk.

[Speer et al., 2016] Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5: An open multilingual graph of general knowledge.

[Talmor et al., 2018] Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2018). Commonsenseqa: A question answering challenge targeting commonsense knowledge.

[Tan et al., 2020] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., and Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools.

[Vilares et al., 2018] Vilares, D., Peng, H., Satapathy, R., and Cambria, E. (2018). Babelsenticnet: A commonsense reasoning framework for multilingual sentiment analysis. pages 1292–1298. IEEE.

[Wang et al., 2020] Wang, C., Liang, S., Jin, Y., Wang, Y., Zhu, X., and Zhang, Y. (2020). Semeval-2020 task 4: Commonsense validation and explanation.

[Wu et al., 2020] Wu, S., Li, Y., Zhang, D., Zhou, Y., and Wu, Z. (2020). Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact.

[Wu et al., 2021] Wu, W., Chang, T., and Li, X. (2021). Vision-language navigation: A survey and taxonomy.

[Xing et al., 2021] Xing, Y., Shi, Z., Meng, Z., Lakemeyer, G., Ma, Y., and Wattenhofer, R. (2021). Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation.

[Ye et al., 2019] Ye, Z.-X., Chen, Q., Wang, W., and Ling, Z.-H. (2019). Align, mask and select: A simple method for incorporating commonsense knowledge into language representation models.

[Young et al., 2017] Young, T., Cambria, E., Chaturvedi, I., Huang, M., Zhou, H., and Biswas, S. (2017). Augmenting end-to-end dialog systems with commonsense knowledge.

[Zareian et al., 2020] Zareian, A., Wang, Z., You, H., and Chang, S.-F. (2020). Learning visual commonsense for robust scene graph generation.

[Zellers et al., 2019] Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. pages 6713–6724. IEEE.

[Zhang et al., 2018] Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., and Durme, B. V. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension.

[Zhang et al., 2020] Zhang, Z., Zhao, H., and Wang, R. (2020). Machine reading comprehension: The role of contextualized language models and beyond.

[Zhang et al., 2021] Zhang, Z., Zhou, Z., Tang, H., Li, G., Wu, M., and Zhu, K. Q. (2021). Enriching ontology with temporal commonsense for low-resource audio tagging. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3652–3656.

[Zhou et al., 2018] Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Commonsense knowledge aware conversation generation with graph attention.

[Zhu et al., 2019] Zhu, F., Zhu, Y., Chang, X., and Liang, X. (2019). Vision-language navigation with self-supervised auxiliary reasoning tasks.