This thesis set out to study the extent to which automatic image description systems are able to generate human-like descriptions. This question was split into three separate research questions:

1. How can we characterize human image descriptions? Specifically, what does the image description process look like, what do people choose to describe, to what extent do they differ in how they describe the same images, and how objective are their descriptions?
2. How can we characterize automatic image descriptions? Specifically, what does the image description process look like, how accurate are the automatically generated descriptions, and are they as diverse as human-generated descriptions?
3. Should we even want to mimic humans in all respects? Specifically, are all examples in current image description datasets suitable to be generated by automatic image description systems? If not, what kinds of examples should we avoid?

The first question aims to understand what human image descriptions look like, so as to see what kind of descriptions current data-driven systems aspire to produce. The second question aims to understand where stand in the development of systems producing human-like descriptions. The third, over-arching question, is meant to reflect on the differences between humans and machines. Is it wise to copy all human image description behavior?

8.1 What have we learned?

This thesis is split up into two parts. The first part focused on image description from a human perspective, where we looked at how humans describe images, and what the implications are of this for automatic image description. The second part of this thesis looked at image description from a machine perspective, assessing the state of current automatic image description systems. This section provides a summary of what we have learned from these two parts, followed by a reflection on (un)desirable image description behavior.

8.1.1 Image description from a human perspective

In the first part of this thesis, we have seen that there are three main properties of human image descriptions that have implications for automatic image description systems: (1) they are subjective, (2) they require reasoning, (3) they are task-dependent. We will now discuss these properties in turn.

Human descriptions are subjective

The canonical image description task is not deterministic; when you present the same image to five different participants, chances are that you will end up with five different descriptions. Assuming that this variation is not completely random, we have to conclude that the image descriptions are subjective. In other words, they depend on the participants’ interpretation of
the task itself, their interpretation of the images, and their personal thoughts, feelings, and associations with the images. Chapters 2 and 3 have shown several ways in which human image descriptions for the same image differ from each other:

1. They may present the same facts from a different perspective.
2. They may mention (or omit) different parts of the same image.
3. They may make reference to the same objects at different levels of granularity. That is: they may be more or less specific in the terms (and modifiers) that they use. The specificity of a description may depend on the background knowledge of the speaker and of the perceived background knowledge of the hearer.
4. They may rely on different interpretations of the same image. Actions in particular are underspecified in still images, because by definition photographs (presented in isolation) do not show any movement. For example, the difference between throwing and catching a ball may not be apparent from a picture of someone throwing a ball with two hands.
5. They may rely on different inferences based on the content of the image and the knowledge and beliefs of the participants.

This variation is not necessarily a bad thing. We are still in the early stages of image description research, and the diversity found in current image description datasets allows us to reflect on the question of what image descriptions should look like in the first place.

**Human descriptions require reasoning and world knowledge**

The subjectivity of the descriptions already hints at the idea that image description requires reasoning and world knowledge. After all: the descriptions depend on how different participants interpret the task and the images presented to them. We have seen further evidence that participants actively reason about the images in Chapters 2, 3, and 4.

Chapter 2 presented our basic findings for the English descriptions in the Flickr30K and MS COCO corpora. We found that crowd-workers often go beyond the contents of the images, and add their own inferences to their descriptions (e.g. about the goals, activities, ethnicity, or occupation of the people in the images). These unwarranted inferences are unexpected, because participants were instructed to not make any unfounded assumptions. Furthermore, the use of negations and adjectives also shows how crowd-workers compare the images to their past experiences and mark aspects of the images that are unusual or that deviate from the norm. The use of negations also shows that participants are reasoning about what is happening outside the frame, and about what happened before and after the picture was taken.

Chapter 3 showed that our findings also hold for other languages, and provided additional evidence from the comparison of Dutch, English, and German descriptions that differences in world knowledge affect the specificity of the descriptions. For example, American crowd-workers were unable to identify a traditional Dutch street organ, whereas every Dutch crowd-worker used the same term (*draaiorgel*) to refer to the instrument.

Chapter 4 further supported our claim that image description requires reasoning, by providing real-time evidence of participants reasoning about the images. By eliciting spoken image descriptions together with eye-tracking data, we were able to see what participants were looking at as they were describing the images. Participants seem to be actively predicting what the images were about as they are describing the images. Furthermore, their self-corrections indicate that they reason about (1) the appropriate level of specificity for their descriptions, and (2) whether their descriptions might be ambiguous.
Having shown that the descriptions in current image description datasets are the result of a higher-level reasoning process (rather than a one-to-one mapping of visual features to text), it seems clear that if we want automatic image description systems to be able to produce human-like descriptions, then they should also be able to perform this kind of reasoning.

**Human descriptions are task-dependent**

Chapter 5 considered the effect of the format of the image description task on the resulting descriptions. The chapter argues that the canonical image description task has just one out of many possible formats, and provides an overview of all the different parameters that one might manipulate to influence the outcome. Focusing on spoken versus written descriptions, we have found that speakers are more likely to ‘show themselves’ in their descriptions than writers. For example, they seem to use more *consciousness-of-projection* terms, indicating how certain they are about their observations. Future research should investigate whether users appreciate the spoken style more (or less) than the written style.

Given that differences in the image description task lead to different descriptions, we may ask ourselves whether the canonical format actually provides the best set-up for the task. This is important, because with the use of image description corpora for training automatic image description systems, we are implicitly telling models that this is what image descriptions should look like.

**Towards an understanding of the human image description process**

In his *Tractatus*, the philosopher Ludwig Wittgenstein noted that, though incorrect, his propositions were useful to gain a deeper understanding of the relation between language and reality. After gaining this newfound understanding, we can abandon the propositions and move on. Or in Wittgenstein’s words:

6.54 - My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them –as steps– to climb beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright.

(Wittgenstein, 1921/1961)

This idea has come to be known as *Wittgenstein’s Ladder* (although others have used this metaphor before him, see Gakis 2010). Datasets such as Flickr30K and MS COCO are similar: they are useful for us to gain a better understanding of how people describe images, but, having reached this level of understanding, it is clear that we need more controlled data. For example, it would be useful to specify the goal of the task, so that participants know how their descriptions will be used. This would enable them to adjust their descriptions accordingly, which would reduce variation in the descriptions. We have discussed other factors influencing the descriptions in Section 5.3. In a more controlled experiment, we could start to systematically manipulate these factors to see how they influence the image description process. Furthermore, it would be useful to retain participant IDs, so that it is possible to study individual variation in image description.\(^1\)

If we want to make the goal of the image description task more explicit, then more work is also needed to explore different applications of image description technology. We will

\(^1\)This data is available for the Dutch image description data collected for this thesis, as well as the German part of the Multi30K corpus (Elliott et al., 2016). It may also be available for other image description corpora, but this metadata is typically not listed with the publication of the data.
discuss this in more depth in Section 8.2, but for now it is important to recognize that different applications may also have different requirements regarding the form and content of the descriptions. This in turn means that we would need different image description corpora to study how images should be described for a particular task, in a particular domain. We may then find that the cognitive requirements for producing suitable image descriptions may differ between tasks and domains.

8.1.2 Image description from a machine perspective

The second part of this thesis focused on image description from a machine perspective. We have identified three main properties of current approaches. They (1) are inherently limited, (2) produce flawed descriptions, and (3) produce generic descriptions. We will now discuss these properties in turn.

Current approaches are inherently limited

Chapter 6 presented an overview of current image description technology, introducing different kinds of neural networks. But given that their goal is to produce human-like image descriptions, we have to ask ourselves: are they up to the task? As I have argued above in Section 8.1.1, it is clear that human image descriptions are subjective (depending on the participants’ interpretation of the task itself, their interpretation of the images, and their personal thoughts, feelings, and associations with the images), require world knowledge, and are highly contextual. However, looking at the general architectures that are used for automatic image description systems, it is clear that they assume a simple one-to-one mapping from images to text. There are (typically) no components that use external resources to reason about the images. Thus there is a clear contrast between what humans do, and what automatic image description systems are designed to do. As noted in the introduction of this thesis (§1.6), there are two possible ways to resolve this issue: either we should (1) build more advanced image description systems, or we should (2) change the (currently implicit) goal of trying to match human descriptions as closely as possible, and formulate a more restrictive standard for what image descriptions should look like.

Current approaches produce flawed descriptions

Following an overview of the general architecture of automatic image description systems, Chapter 6 provided an error analysis for one specific model: Xu et al.’s (2015) attention-based architecture, trained for the Flickr30K dataset. Error analyses for image descriptions are subjective by nature, because classifying the type of error means that the annotator has to reason about what the model is supposed to say. Nevertheless, error analysis is useful to get a general sense of a model’s strengths and weaknesses.

Our results indicate that about 80 percent of the generated descriptions contains at least one error. Most of the errors fall in the generally unrelated category, which means that the description does not seem to have any relation to the image. After this category, most errors fall into the categories color of clothing (e.g. green shirt instead of red shirt), activity (walking instead of running), type of clothing (shirt instead of coat), and gender (man instead of woman). Furthermore, many of the errors made by the system are unlikely to be made by humans. For example, Figure 6.12 in Chapter 6 shows a little girl in a pink dress holding a large ball in her hands. This image is described by the system as A little boy in a white shirt playing soccer. Descriptions like these show us that we are still far away from human-level
automatic image descriptions. Despite the fact that we only looked at the performance of one model, we expect that other image description systems with similar architectures will also make errors like these. The distribution of errors will probably differ, but there is no fundamental reason to expect that another model will not produce any mistakes regarding color of clothing, for example. What is needed, is some way to ensure the visual fidelity of the descriptions (cf. Madhyastha et al. 2018), so that the automatically generated descriptions will not only be similar to the human-generated descriptions, but also correspond to the contents of the image.

**Current approaches produce generic descriptions**

Having looked at the content of the automatically generated descriptions, Chapter 7 examined the diversity of the output of 9 different automatic image description systems. We asked to what extent these systems display the same amount of variation as the human-generated descriptions, and whether these systems were able to use particular labels that all human annotators agreed on. In both of these areas, we found that there seems to be much room for improvement. Automatic image description systems tend to only use a small portion of the vocabulary that is available from the training data. Furthermore, if human annotators all agree that a particular term should be used in the description of an image, systems only use that term in 80% of the cases. Finally, image description systems seem to lag behind humans in terms of compositionality; they use fewer kinds of compound nouns, and fewer kinds of prepositional phrases. This may indicate that automatic image description systems are less expressive than humans. At the same time, we shouldn’t necessarily take humans as the standard to aspire to. In some cases, it might actually be beneficial for a system to produce relatively predictable descriptions, with only a limited vocabulary. More research is needed to establish when to use a more diverse vocabulary, and when generic descriptions would suffice. Either way, Chapter 7 provides a first step towards a better operationalization of diversity in image descriptions.

**8.1.3 How human-like should automatic image descriptions be?**

The third sub-question is difficult to answer, because it is not clear what it means for a description to be human-like. As noted above, the descriptions in the Flickr30K and MS COCO datasets are very diverse, and different annotators have different ideas about what an image description should contain. For the sake of simplicity, let us say that a system is fully human-like if it is able to produce any of the different kinds of descriptions that we see in existing image description datasets. Based on the above, there are three answers to the third sub-question:

**Computability.** Some kinds of descriptions are easier to produce than others. For example, a description like ‘A man in a red shirt is walking down the street’ is relatively straightforward, compared to descriptions containing negations, or interpretations of how people might be feeling in a particular situation. The latter require much more reasoning and background knowledge (e.g. about how different experiences may affect someone’s mood). It may not be feasible for current systems to produce these more advanced kinds of descriptions. We will also discuss this in Section 8.3.

**Systematicity and predictability.** The wide range of variation displayed by human image descriptions also makes the descriptions themselves unpredictable. This thesis posits that the amount of variation may (at least in part) be due to the fact that participants of the image description task differ in their understanding of what the task is about. With a clearer problem
definition (stating what the image descriptions should be used for), we might be able to establish some standards of what a proper image description should look like. (See §8.2 below for further discussion.) Following these standards, we should see less variation in the descriptions, which should also make the output of image description systems more predictable (and easier to evaluate). This predictability may help users understand when a system generates a particular kind of description, which also helps them make inferences about what likely isn’t in the image (because otherwise the system would have told them; cf. Grice 1975).

**Truthfulness and fairness.** For an image description system to be usable, it should provide reliable descriptions, that treat all subjects fairly and without prejudice. We have seen in Chapters 2 and 3 that participants of the image description task don’t restrict themselves to the contents of the images, but often speculate about what is happening in the image, what caused the events in the image, and what is likely to happen. In their speculations, people often resort to stereotypes. Furthermore, people display biases in the way that they mark people and situations that are different from what they perceive to be the default. It would not be advisable for systems to display the same behavior, because of the potential for this behavior to be harmful or offensive (next to the fact that speculations and generalizations are simply not always true).

### 8.2 Application: supporting blind and visually impaired people

What could automatic image description systems be used for? This section will discuss one of the most important applications for automatic image description technology: supporting blind and visually impaired users in their interaction with the world around them. As I have argued in this thesis, current image description datasets display an overwhelming amount of variation, with many different ways to describe the same image. This section argues that we need to talk to potential users of image description technology to understand what is the best way to describe any particular image (§8.2.1), provides an overview of existing research using image description technology to help blind and visually impaired users (§8.2.2), and discusses possible next steps (§8.2.3).

#### 8.2.1 Developing sign-language gloves: A cautionary tale

In developing any application, it is important to keep the end users in mind, and to try and understand their needs. One of the best examples of what not to do is the development of sign-language gloves. In an article titled *Why sign-language gloves don’t help deaf people*, Michael Erard (2017) describes how different groups of researchers developed high-tech gloves for deaf people to wear, so that their gestures could automatically be translated into spoken English.\(^2\) The main problem with these kinds of gloves is that they misconstrue the problem. Many sign-language gloves only focus on what the hands do (e.g. finger-spelling). But sign-language also uses arm-gestures, facial expressions, and lip movements, which are not captured by the gloves. Thus, the gloves cannot possibly translate the entire message. Furthermore, there is no way for hearing people to respond, so the conversation remains one-way traffic. The moral of the story is that, in developing assistive technology, we should always involve the potential users themselves. Ideally, they should be consulted from the beginning, so that the research does not start out on the wrong foot, and our solutions are actually useful in practice.

---

\(^2\) Also see the open letter by Forshay et al. (2016).
8.2.2 Existing research on supporting the blind

The automatic image description literature regularly refers to the potential of this technology to help blind or visually impaired people, but we are still in the early stages of establishing what these people actually want or need, in terms of image descriptions. Existing research can be categorized as follows:

**Alt-text**

Petrie et al. (2005) provide an overview of existing guidelines for *alt-text*: `alternative text` to be displayed instead of images for visually impaired users browsing the web using a screen reader. The authors also describe the results of a series of interviews with visually impaired users, asking them how images on the web should be described. Petrie et al.’s (2005) conclusion is that descriptions are very context-dependent, but the following elements should usually be included:

1. Objects, buildings, and people in the image.
2. Activities taking place in the image.
3. The use of color.
4. The purpose of the image.
5. Emotion and atmosphere.
6. The location of the depicted events or activities.

Since this study predates most of today’s social media outlets, or at least their widespread use, it does not tell us which properties of images are important in the context of social media. Furthermore, these guidelines are also not informative about life outside the web; how should real-life situations be described?

**Automatic image description**

Gella and Mitchell (2016) contrast the capabilities of automatic image description systems with the needs or blind or visually impaired people. They note that current automatic image description systems mostly focus on objects, attributes, and actions. Talking to blind or visually impaired people, however, Gella and Mitchell found that users would also like to have a description of the emotion and atmosphere, and whether the image is humorous or not (which perhaps coincides with what Petrie et al. call the purpose of the image). Furthermore, they would like to see descriptions for different types of domains: personal, news, and social media images.

Studies about automatic image description for social media images have been carried out by MacLeod et al. (2017); Zhao et al. (2017b), and Wu et al. (2017b). MacLeod et al. (2017) carried out a user study with automatically generated descriptions for images from Twitter. They provided blind or visually impaired people with actual tweets, that were enriched with

---


4 I will ignore related areas, such as object detection, depth estimation, (micro-)navigation, text extraction and text summarization. See Weiss et al. 2018 for a short survey.

5 Facebook was introduced to college students in 2005, and Twitter was launched in 2006; see Boyd and Ellison 2007 for a timeline.

6 This question also gives rise to an ethical dilemma: determining what to describe also means that you are effectively withholding information about other parts of the image. How should we handle this responsibility? We leave this question for future research.
automatically generated descriptions. Their first experiment was a think-aloud study, where users were asked to describe their experiences with the automatically generated descriptions. The authors note that users generally trusted the descriptions (without double-checking the information), despite the fact that they were often wrong. Moreover, in cases where the descriptions did not line up with the content of the Tweet, the users tried to provide explanations for why the Tweet-caption combination could still be coherent, rather than dismissing the captions for being implausible. MacLeod et al. (2017) note that this bears some risk for users of automatic image description software, because they may wrongly act upon misleading descriptions. Thus it is important to clearly communicate the accuracy of automatically generated image descriptions to the users. In a follow-up experiment, the authors looked at different ways to communicate (un)certainty about descriptions that are (in)congruent with the images they are associated with. They found that negatively framed descriptions encourage users to remain skeptical about the descriptions in situations where the system is uncertain. Examples of negative frames are: I have absolutely no idea but my best guess is . . . ; I am not completely sure, but I think it’s . . . This works better than positive framing (e.g. I’m only sort of confident, but . . . ; I’m pretty sure it’s . . .), where users are more likely to accept the descriptions as valid.

Zhao et al. (2017a) interviewed 12 visually impaired participants to understand their experiences with photo sharing on Facebook. The authors developed an automatic image description system to aid visually impaired users of the mobile Facebook application. Afterwards, they evaluated this application using a seven-day diary study with six visually impaired users. Based on the 12 interviews, the authors identified three aspects that users would like to know before uploading an image to Facebook:

1. Key visual elements: main landmarks and objects depicted in the image.
2. People: the identities and relative location of the people in the image.
3. Photo quality: technical (focus, lighting), composition (e.g. no people cut off), and subject behavior (e.g. smiling, no eyes closed).

The diary study indicated that users found the application helpful, but they were unsure about the reliability of the descriptions. Having used the application they also had further requests to improve the descriptions. They should provide information about:

4. The kind and color of different objects, especially for common objects like flowers. For example, ‘flowers’ could be specified to ‘yellow tulips’.
5. Non-salient items, especially those that may help distinguish multiple similar images.\(^7\)
6. The luminance and the level of blurriness (some blurriness may be acceptable).

Wu et al. (2017b) present another user evaluation for Facebook’s Automatic Alt Text (AAT) functionality. Their participants noted two further improvements that they would like to see:

7. The ability to extract and recognize text.
8. More detailed descriptions of people, “including their identity, age, gender, clothing, action, and emotional state.”

Finally, Zhao et al. (2017b) also found that their participants were re-appropriating the app to organize their photo collections. This also shows that there is room for the development of

---
\(^7\) The idea to automatically produce pragmatically informative descriptions that distinguish an image from similar images, is explored by Andreas and Klein (2016) and Cohn-Gordon et al. (2018).
personal photo organization applications, which may have different requirements than social media image descriptions.

**Visual Question Answering and the VizWiz grand challenge**

Following initial work on Visual Question Answering (Antol et al., 2015; Goyal et al., 2017), where computers are asked to answer different questions about a set of images, Gurari et al. (2018) presented the VizWiz grand challenge. The VizWiz dataset consists of 31,000 questions from blind people, about pictures they took themselves. This dataset represents a real-life application (VizWiz; Bigham et al. 2010), which blind people use to answer everyday questions, such as: what type of soup is this? or what temperature is the oven set to? Because the pictures are taken by blind users (who cannot see the screen), the images are often of low quality, and the questions are spoken rather than written. The VizWiz grand challenge consists of two subtasks: 1. predicting the answer to a visual question; and 2. predicting the answerability of a visual question.

The VizWiz grand challenge is a great addition to the existing multimodal Natural Language Processing and Computer Vision tasks, because it confronts us with the noise and uncertainty of real-life data. Moreover, the dataset itself is a very rich source of information about the domains that blind and visually impaired people are interested in. For example, we may use the subjects of the questions and images to understand what kind of information should be highlighted in automatic image descriptions.

**8.2.3 Future research supporting blind and visually impaired people**

Summarizing the above, there is a growing list of aspects that are generally important for automatic image description systems describe. But it is still unclear:

1. Which of those aspects are relevant to mention, given a particular image and context.
2. How specific the description of those aspects should be.
3. What is the best way to phrase the descriptions.

The image description literature has generally avoided these issues by delegating them to the crowd-workers annotating the images. A technical solution is still far on the horizon, because formulating a suitable description, mentioning the relevant aspects of an image, at the right level of specificity is still too difficult for current technology. (The next section discusses third-wave approaches that should be able to provide satisfying descriptions to users.) An alternative would be to take a Q&A-style approach (similar to Visual Dialog; Das et al. 2017), where the system would generate a ‘basic description’ and the user can ask for specific details. The basic description would then serve as a starting point for the conversation. Whatever approach we end up taking, we should always keep the end users in mind. By involving them in the process, we can establish clear guidelines to develop image description solutions that actually address the needs of blind and visually impaired people. These guidelines in turn allow us to develop evaluation metrics that show our progress in generating suitable descriptions.

**8.3 Automatic image description in the context of Artificial Intelligence**

The work in this thesis can be seen as part of the more general area of Artificial Intelligence. This section aims to present a short overview of the recent progress in this field.
8.3.1 Three waves of AI

In a recent DARPA\(^8\) video, Launchbury (2017) describes the development of Artificial Intelligence (AI) as coming in three waves:

1. **Handcrafted knowledge**: this first wave of development involves experts translating knowledge from a particular domain into formal rules for computers to follow. This works very well for narrow domains, where the computer can take a set of basic facts and reason through their implications. The downside, according to Launchbury is that rule-based systems are less suited to learn from experience, to abstract away from specific problems and apply their knowledge in a different domain. Furthermore, they are not able to perceive the outside world and see what’s going on. In Launchbury’s (2017) words, they “stumble when it comes to the natural world.”

2. **Statistical learning**: this second wave of development focuses on the ability to extract knowledge from data. AI systems in this second wave are much better at perceiving the world and learning from data to adapt to new situations. At the same time, these systems are limited in terms of logical reasoning. Launchbury (2017) summarizes the strengths and weaknesses of second-wave AI systems by saying that they have “nuanced capabilities to classify data and to predict the consequences of data, but they don’t really have any ability to understand the context in which they’re taking place and they have minimal capability to reason.” Hence, DARPA is foreseeing a third wave:

3. **Contextual adaptation**: Launchbury (2017) describes this future wave as one where “the systems themselves over time will build underlying explanatory models that allow them to characterize real-world phenomena.” An important feature of these systems is the ability to properly *explain* their decisions.\(^9\) Furthermore, third-wave systems should be able to learn from only a handful of examples, rather than the thousands of training examples required for current statistical learning systems.

The automatic image description systems demonstrated in this thesis are clearly part of the second wave of AI; current systems mostly aim to generate ‘the most probable description’ given an image, without developing an explanatory model that could tell us *why* an image should be described in a particular way. Current systems are also unable to adapt to the context in which they are providing their descriptions. Chapters 6 and 7, then, are an exploration of second wave systems and the limits of this kind of technology.

8.3.2 Requirements

How do we move from second to third-wave AI? In a recent paper, Lake et al. (2017) present an overview of the requirements for “building machines that learn and think like people.” They broadly categorize these requirements into three sets of ingredients:

1. **“Start-up software”** This first set of ingredients corresponds to cognitive capabilities that children have from an early age:

---

\(^8\)DARPA is the Defence Advanced Research Projects Agency, which funds scientific research in the United States.

\(^9\)The ability to explain decisions is not just nice to have. Not only do explanations signal a deeper understanding of the problem that the system is built to solve, they also satisfy public demands for transparency in automated decision making systems. See Goodman and Flaxman 2017; Selbst and Powles 2017 for a discussion of the European ‘right to an explanation’ and Lipton 2016 for a discussion of what it means for a machine learning model to be interpretable.
Intuitive physics: infants have a basic understanding of how the physical world works, and they know, for example, which kinds of movements are possible and impossible. They can use (and improve) this understanding with every new task they learn.

Intuitive psychology: infants can attribute mental states (goals, beliefs, desires, intentions, knowledge) to other people, which helps them reason about other people’s behavior. In turn, this helps them infer other properties about the world (e.g. which objects are good and which are bad).

2. Learning Lake et al. (2017) note that they “view learning as a form of model building, or explaining observed data through the construction of causal models of the world” (emphasis in original). These models of the world include the intuitive notions of physics and psychology that infants start out with, and that gradually improves as they learn. The authors argue that compositionality and learning-to-learn are essential ingredients to make rapid model learning possible.

Compositionality is the key to understand complex scenes or objects. Rather than treat each complex scene or object as completely new, we can begin to understand those scenes or objects by decomposing them into their primitive parts. This makes the reasoning process more efficient, and it improves generalization, because each encounter with a complex scene or object informs us about the properties of its more primitive parts (and vice versa), which we can use in the next situation where we encounter those parts again.

Causality means knowing or reasoning about how different situations come to be; providing an explanation. Lake et al. (2017) argue that people also understand scenes like the ones in the Flickr30K and MS COCO dataset by building causal models. Specifically: “human-level scene understanding involves composing a story that explains the perceptual observation, drawing upon and integrating the ingredients of intuitive physics, intuitive psychology, and compositionality.” In other words, understanding a scene requires us to identify the individual components and to be aware of what they might contribute to the scene (compositionality), it requires us to reason about the way that the objects in the scene are held together (intuitive physics), and it requires us to think about the goals and intentions of the people in the scene (intuitive psychology), in order to construct a coherent story about what is going on. Lake et al. (2017) note that causality might also help us understand the role of unfamiliar objects in a scene.

The errors that we have seen in Chapter 6 of this thesis are either foundational errors (the visual features being flat-out misleading), or they could be the result of missing one or more of the ingredients listed so far. Lake et al. (2017) note that image description systems often seem to get the key objects correct, but are unable to relate these objects to each other (and thus they do not build the right causal model –if they build causal models at all).

Learning-to-learn refers to the idea that previous learning experiences can make it easier to learn new tasks (Harlow, 1949). Lake et al. (2017) note that this is similar to transfer learning, multi-task learning or representation learning in the field of Machine Learning. The authors note that, while these concepts are already used, there is still room for improvement, because humans are still much more efficient at leveraging their past experiences to learn to perform new tasks. One way to improve learning-to-learn skills is to focus on the ingredients listed earlier.

3. Speed Rich and complex causal models that humans develop about the world (such as the ones that Lake et al. (2017) propose), are typically slow, as they may require multiple
reasoning steps to get to the answer. Lake et al. (2017) observe that this contrasts with speed of perception and thought. Somehow, the authors note, humans successfully combine rich models with efficient inference. Though Lake et al. (2017) do not make the connection, this is reminiscent of Kahneman’s (2011) theory of Thinking Fast and Slow. He argues that we have two modes of thought, which he refers to as System 1 and System 2. System 1-thinking is fast, instinctive, and emotional, while System 2-thinking is slower, deliberative, and logical. We may also interpret one of the main examples from Chapter 4 of this thesis in these terms. Figure 4.3 showed a picture of a restaurant with people sitting around a table. The participant describing this image immediately inferred from the setting that the group of people was eating. But as the participant continued to describe the image, they found that the group wasn’t in fact eating anything yet; they were still looking at their menus. In this example, the quick interpretation of the image would be a good example of System-1 thinking, which was later corrected as the participant collected more information and had more time to think about the image.

The requirements laid out by Lake et al. (2017) are based on findings from a wide range of disciplines. And these are just the cognitive requirements. If we want artificially intelligent systems to have any role in society, we also need to think about the ethical implications of developing such systems (e.g. Hovy and Spruit 2016; Friedman et al. 2013; IEEE 2018). In short: it is impossible to study AI in isolation. Developing AI means talking to many different groups of researchers. For this conversation to be successful, it is important to make our research as accessible as possible. The next paragraph highlights ways of doing so.

8.3.3 A way forward: more interaction with related fields

Epstein et al. (2018) discuss the rise of Artificial Intelligence as a field, and note that there are strong incentives to develop new systems that improve upon the state-of-the-art performance on particular tasks, but there is less emphasis on the study of those systems themselves. This gives rise to a knowledge gap in AI: development of AI systems moves faster than our understanding of them. What we need, Epstein et al. argue, is a centralized platform where researchers can upload their systems, and others can easily test them, without the need to install anything on their own computer. This would allow social scientists (and I would add: linguists and philosophers) to test the biases and competence of AI systems without requiring any technological knowledge.

Another solution to address the AI knowledge gap is to create shared events with researchers from other fields. One such example is the Workshop on Building Linguistically Generalizable Natural Language Processing Systems, which aims to bring together linguists and NLP researchers (Ettinger et al., 2017). The first edition of the workshop also featured a build it, break it-challenge, where there are two kinds of participants: the builders and the breakers. The former aim to build NLP systems that are robust to linguistic variation, while the latter aim to construct difficult test cases that might trip up the systems. The first edition saw four breaker teams submit test cases for the builders to evaluate their systems on. Those test cases focused on (morpho)syntactic, semantic, and pragmatic phenomena, as well as the ability to use world

10 This echoes the sentiment from Ali Rahimi and Ben Recht’s controversial acceptance speech for the NIPS 2017 ‘test of time’ award (Rahimi and Recht, 2017). They argued that AI and machine learning are still at a pre-scientific stage, similar to alchemy before it developed into physics and chemistry. We have some idea of what works and what doesn’t, but we don’t know why. Jordan (2018) uses the metaphor of civil engineering: “While the building blocks have begun to emerge, the principles for putting these blocks together have not yet emerged, and so the blocks are currently put together in ad-hoc ways.”
knowledge to reason about the examples. These test cases help us to better understand model performance in terms of phenomena that are well-studied in linguistics. In a way, the build it, break it-challenge is a real-life version of the platform that Epstein et al. (2018) propose. But the organization of shared workshops has the additional benefit of engaging with each other in person. At the same time, a permanent platform where researchers can continuously interact with existing systems allows for more experimentation.

In short: publishing papers and open-sourcing code and data is not enough. We need to think more about the accessibility of our work, and whether it is also feasible for non-technical researchers to study the fruits of AI and NLP research. Opening up the field to ‘outsiders’ may help us deepen our understanding of what AI and NLP are capable of.

8.4 Future research

There are different ways to make a contribution to NLP. David Marr (1982) posited that a full description of any cognitive system requires an explanation at three levels:

1. The computational level: what task is the system solving?
2. The algorithmic level: how does it actually solve the task?
3. The implementational level: how is this algorithm physically realized?

Most work in NLP seems to focus on the algorithmic level: assuming a well-defined task, can we find a better solution to that task? This thesis has mostly on the computational level, trying to give a better characterization of the task of automatic image description, through analyzing existing image description data. One of the main problems (or perhaps even the main problem) with image description research right now is that the task is not well-defined. What we need is a combination of:

1. **User studies** asking potential users of image description systems what the descriptions should look like. These studies should identify different classes of properties that image descriptions should have. We have already seen some of these kinds of studies in our discussion of image description for blind and visually impaired people (§8.2), but user studies shouldn’t be limited to this target group only. Others may also benefit from image description applications, e.g. users of voice assistants like Siri, Google Home, or Alexa.

2. **Metric development** where researchers determine for a given feature how to measure whether a particular system is able to competently produce descriptions with that feature. For example, whether a system is able to use negations in its image descriptions. Having more fine-grained test sets with targeted evaluation metrics hopefully allows for a ‘divide-and-conquer’ situation where different groups work towards solving different sub-problems of automatic image description.

3. **Feasibility studies** where we look at which features are feasible for an image description system to produce. These studies could either target a single feature, or see to what extent a particular is able to competently produce a wider array of different features. In these kinds of contexts, it is often proposed to develop a ‘summary score’ to see how well systems are doing overall. I would argue against this idea, because it is not clear what such summary statistics

11 A cognitive system could be defined here as ‘any information processing system,’ which could equally apply to both humans and machines trying to produce a description for a given image.

12 This has come to be known in Cognitive Science as the *Tri-Level Hypothesis* (Dawson, 1998).
mean. Is a system with an overall score of 0.8 better than one with an overall score of 0.75? That depends on how important you think the individual features are that make up the overall score, and this importance may differ from situation to situation.

As noted in the previous section, we cannot do this alone. It takes a concerted effort of researchers in NLP, human-computer-interaction, and linguistics to bring us towards a future where computers can finally handle all the pragmatic factors in automatic image description.