Submitted 5/2019; published 12/2019

Linguistic Variation and Anomalies in Comparisons of Human and Machine-Generated Image Captions

Minyue Dai MDAI@SMITH.EDU
Sandra Grandic SGRANDIC@SMITH.EDU
Jamie C. Macbeth JMACBETH@SMITH.EDU
Department of Computer Science, Smith College, Northampton, MA 01063 USA

Abstract

Describing the content of a visual image is a fundamental ability of human vision and language systems. Over the past several years, researchers have published on major improvements on image captioning, largely due to the development of deep learning systems trained on large data sets of images and human-written captions. However, these systems have major limitations, and their development has been narrowly focused on improving scores on relatively simple "bag-of-words" metrics. Very little work has examined the overall complex patterns of the language produced by image-captioning systems and how it compares to captions written by humans. In this paper, we closely examine patterns in machine-generated captions and characterize how conventional metrics are inconsistent at penalizing them for nonhuman-like erroneous output. We also hypothesize that the complexity of a visual scene should be reflected in the linguistic variety of the captions and, in testing this hypothesis, we find that human-generated captions have a dramatically greater degree of lexical, syntactic, and semantic variation. These results have important implications for the design of performance metrics, gauging what deep learning captioning systems really understand in images, and the importance of the task of image captioning for cognitive systems research.

1. Introduction

Composing natural language descriptions of visual and perceptual scenes is a key capability of human and machine cognition. Over the past several years, researchers have published major improvements in AI systems that generate natural language captions corresponding to visual images and scenes. Largely these advances correspond to the recent advent of deep learning and the development of large data sets of image files and human-written captions corresponding to them.

However, related research has raised concerns about the limitations of deep learning, demonstrating that deep networks for computer vision and natural language processing may not be performing these tasks as well as the highly publicized successes would lead us to believe (Nguyen et al., 2015; Jia & Liang, 2017). It appears that, for nearly all such models, researchers have found systematic methods for crafting evaluation data sets that are dramatic examples of these models going wrong, demonstrating how they are not performing these tasks in ways similar to humans. Image-captioning systems are no exception to this phenomenon, as adversarial examples have been developed for them as well (Chen et al., 2018).

The quantitative measures for evaluating these captioning systems compare a machine-generated caption with several human captions for the same image using "bag-of-words" or "bag-of-n-grams" metrics. While these metrics are used to gauge advances in the field, their simplicity appears to be the cause of peculiar claims of machine captioning systems "beating" humans at the task or passing the Turing Test (Vinyals et al., 2017), and they may also be partly to blame for the confusion about model performance. The development of modern-day image-captioning systems has been narrowly focused on improving aggregate scores on these relatively simple metrics on the individual images. Very little work has disregarded these metrics to examine the overall vision and language behavior patterns of image-captioning systems in comparison to those of humans.

The organization of light, intensities, colors, edges, textures, and other "low-level" features into perceived objects, figures, and backgrounds is fundamental to the visual cognitive processing of images in a variety of vision theories (Wagemans et al., 2012; Ullman, 1989; Marr, 1982). The main claim of this paper is that these recognition and identification processes are but precursors to the much richer processes of image understanding, which likely incorporate not only memories of prior viewings of similar objects, but also experiences with object motion, depth perception, and other senses. We present evidence to support the claim by differentiating the image-processing behavior of a learned neural network, which typifies classification and recognition, from that of humans, which are presumed to be in-depth understanding systems. We perform this comparison by examining the content of image caption texts produced by each, which act as indicators of the cognitive processing invoked by the visual stimulus.

Recent studies of the natural language models embedded in learned image-captioning systems indicate that they may only be recognizing individual objects and stringing together separate descriptions of them without ever understanding how those objects interact with each other and their environment in visual scenes (Heuer et al., 2016). We theorize that, unlike image classifiers, humans use complex knowledge structures to understand the physical and social interactions that are the subjects of images (Minsky, 1975; Schank & Abelson, 1977). These may be analogous to structures that humans use to understand and reason about natural language stories about similar subjects, and they may find their expression in descriptions that humans write about the images (Schank & Leake, 1989).

To differentiate humans from machine generation, we perform a close examination of a corpus of captions generated by a publicly-available deep network model trained on a widely-used image captioning data set and compare them to aggregate samples of human-generated captions on the same images. We characterize patterns in the quality and style of image captions that are obviously erroneous and reflect strong differences between the model and human intelligence. We hypothesize that the complexity of human understanding of the visual scene should be reflected in the complexity and variety of the linguistic responses to the images, and that this complexity may not be reflected in the conventional scores and metrics. To test this hypothesis, we also compare the linguistic variety of machine and human captions in our corpus in a number of ways, and find that a sample of human-generated captions has dramatically greater lexical, syntactic and semantic diversity.

These results have important implications for broader questions on the appropriateness of conventional performance metrics for image-captioning systems. They also address the degree to which recent captioning systems actually understand what is going on in images, such as the interactions between people and objects that appear in them, and whether image captioning should continue to be an important task for cognitive systems research on the vision and language functions of human intelligence.

2. Background

Humans have an ability to describe the content of an image, a task that has proved difficult to replicate in computer models. An image captioning model usually uses advanced techniques from areas of both computer vision and natural language processing to recognize objects and caption the relationships between them.

Traditional image-captioning systems (for a review, see Hossain et al., 2018) required handcrafted vision and language features that are task specific and constrained to the domain. In comparison, deep networks are claimed to automatically learn image and language features through large data sets. Such image-captioning systems typically use convolutional neural networks (CNNs) in image encoding to extract visual features and long short-term memory (LSTM) variants of recurrent neural networks (RNNs) to generate sentences describing the images (LeCun et al., 1998; Hochreiter & Schmidhuber, 1997; Hossain et al., 2018). These systems generate a full sentence conditioned on a given image. In each step, they predict one word w_t from the previous output $(w_0, w_1, ..., w_{t-1})$ and the visual information representation I.

In the training phase of model development, the training data is provided as an image with part of a human annotated caption, such as "A cat is eating", and the model learns to predict the correct next word, e.g., "food" through an optimization process. The actual output of the model is a vector where each entry is the probability of a particular word in the dictionary being the next word. In the testing phase, the model generates a caption based on an image as a prediction of the first word, the second word, and so on. This means that the previous prediction may bias the future output. The model generates a full sentence iteratively by generating the highest-probability word in the vector at each step. The full caption is generated and typically compared with one or more ground truth human-generated captions and scored using common metrics (described below).

Microsoft Common Objects in Context (MSCOCO, Lin et al., 2014) is a large-scale image data set used in various computer vision tasks, such as object detection, segmentation, and caption generation. Its first version, commonly known as MSCOCO2014, consists of 82,783 training images, 40,504 validation images, and 40,775 test images. The training and validation images are annotated with five ground-truth captions written by humans which have been collected via crowd-sourcing, and are provided for researchers to train and validate machine-learning systems for image captioning. The COCO evaluation server allows researchers to upload generated captions for the test images and make direct comparisons with competing captioning systems. The server calculates and posts the scores of competing models on the test data set using a set of standardized evaluation metrics while keeping the ground truth captions of the test data set private and hidden from researchers. MSCOCO2014 has been a long-standing image captioning challenge task since the Large-Scale Scene Understanding workshop (Yu et al., 2015).

MSCOCO also provides an evaluation API with implementations of metrics (Chen et al., 2015) for evaluating machine-generated captions against the validation data, including BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Although these have become standard metrics, there is scant evidence that they are correlated with human judgment. Researchers who created the "Show and Tell" system that tied for first place in the MSCOCO 2015 image captioning challenge (Vinyals et al., 2017) found that their model did exceptionally well on standard metrics, but fared poorly when human raters evaluated the captions. Although Vinyals et al. (2017) were "hoping for more discussion and research to arise regarding the choice of metric", very little can be found in the literature. More recently, however, Krause et al. (2017) used statistical diversity of part of speech tags to compare paragraph-length descriptions of images to sentence-length captions.

3. Methodology

To learn more about actual image captions, metrics commonly used to evaluate image-captioning systems, and the linguistic diversity of machine-generated captions, we ran a state-of-the-art, readily available deep learning-based image-captioning system on the MSCOCO data set. Hardware support, training processes, hyperparameter settings, and network design all affect the performance of deep learning-based models. Publications on these models often do not provide enough of these details, making it difficult for other researchers to fully reproduce the results, even when the authors provide source code for the model. Also, researchers usually implement pre-processing and post-processing algorithms separate from the deep learning architecture to boost performance; these are frequently not published or even mentioned.

Ultimately we chose a system called NOC, the Novel Object Captioner (Venugopalan et al., 2017), to generate image captions for our study. This consists of a CNN-based image recognition model and an LSTM-based language model, and, as such, it is an excellent example of the standard architecture used in deep learning captioning systems. Also, NOC has achieved state-of-the-art performance on the MSCOCO task fairly recently, and, most importantly, unlike many of the available models, a version of NOC that has been pre-trained on the MSCOCO2014 data was available online,[1] letting us avoid issues with training and hyperparameter tuning conflicts.

We ran the NOC model and generated captions for all images in the MSCOCO validation data set, and used the MSCOCO evaluation API to calculate the scores of these captions on standard metrics. The METEOR score of 0.212 that we obtained is nearly identical to that published in Venugopalan et al. (2017), confirming that the model was performing as expected. Additionally, the scores on ROUGE-L and CIDEr stood at 0.436 and 0.487 respectively, while the scores on SPICE and BLEU-2 were 0.148 and 0.507.

To show that the NOC model is state of the art, we sought to compare its metric scores to the scores of other models on the MSCOCO captioning task leaderboard.[2] The scores are based on the test data set, and we are unable to calculate metrics for the NOC model on the test data because this would require the ground-truth captions for the test data, which are kept hidden. We compared the

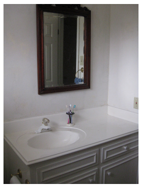
[1] https://github.com/vsubhashini/noc

[2] http://cocodataset.org/#captions-leaderboard

Table 1. An example of a “stutter” caption. All metrics give unexpectedly high scores except for SPICE.

Image file: COCO_val2014_000000300216.jpg

Human-generated ground-truth captions:
1. A clean white bathroom with a simple mirror above the vanity.
2. A bathroom with a white sink and mirror.
3. Bathroom sink with toothbrushes in a stand and a mirror.
4. A white bathroom a sink and a brown mirror.
5. A bathroom sink underneath a mirror and tooth brushes.



Machine caption: **A bathroom with a sink, mirror and a mirror.**

Metric	BLEU-2	METEOR	ROUGE-L	CIDEr	SPICE
Score	0.82	0.373	0.832	1.678	0
Percentile	98.49	94.85	98.89	95.92	9.02

leaderboard scores on test data to our scores on the validation data instead and found that the scores achieved by our model were not the best, but they were at least comparable to the leaderboard scores. Overall, there has been little improvement in the performance of models posted to the leaderboard for this task since 2017, the year that the NOC model was first published. Due to this and the fact that the underlying “vision model-language model” structure is widely used in neural network based image-captioning systems (Hossain et al., 2018), NOC is a suitable representative of these systems.

4. Nonhuman-like Anomalies in Machine-Generated Captions

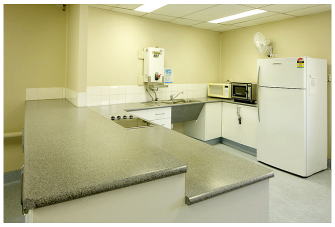
There are many ways to classify errors and anomalies in image captions. Minor misidentifications of objects and locations, as well as inaccurate descriptions of activities, occur in both human and machine-generated captions. But some machine-generated captions have clearly nonhuman-like characteristics in terms of their structure and content. A close examination of just a few captions generated by the learned model confirmed previous observations of nonhuman-like anomalies and discoveries of new ones.

In this section we provide samples of these machine-generated captions, the images that generated them, and the corresponding human-generated ground-truth captions. These are accompanied by scores on metrics provided by the MSCOCO evaluation tools. In addition to illustrating the metrics’ performance, this demonstrates the discrepancies between human judgment and automatic evaluation tools that potentially mislead the development of image-captioning systems. The samples are representative since these types of errors are common in machine-generated captions and designers of future metrics should consider these inconsistencies.

Table 2. An example of a “never-ending” caption in which all metrics except for SPICE give low scores.

Image file: COCO_val2014_000000452695.jpg

Human-generated ground-truth captions:
1. A view of kitchen with two microwaves in it.
2. A kitchen with counters, a sink, a refrigerator and a microwave in it.
3. Modern kitchen facility in open space with large counters.
4. A kitchen that has a icebox and some microwaves in it.
5. A very clean and empty kitchen and counter area.



Machine caption: **A white sink sitting under a mirror in a bathroom . . . . a very clean and ready to ready to be a sink in a mirror . . . . a** . . .

Metric	BLEU-2	METEOR	ROUGE-L	CIDEr	SPICE
Score	0.245	0.172	0.335	0.014	0.636
Percentile	28.51	32.12	25.41	13.65	99.99

4.1 “Stutter” Captions

Sometimes the image-captioning system demonstrates that it successfully detects objects, but it creates a sentence with unusual repeated phrases describing a particular object in the image. In the example shown in Table 1, the model recognizes that the image was taken in a bathroom, but the caption states that there are two mirrors in the image, which is only partially accurate given that there is only one mirror. We found instances of these “stutter” captions easily in a cursory examination of machine-generated captions, confirming the acknowledgment of their existence by the creators of the NOC model (Venugopalan et al., 2017). However, while stuttering does occur in human speech and writing, we found no examples of this behavior in a similar search over the human captions in the data set.

For the stutter caption in the table, BLEU, METEOR, and ROUGE give scores that seem very high given this anomaly, whereas SPICE penalizes it far too much. We speculate that this is due to the fact that SPICE parses the caption to build a semantic graph of the sentence, and for some reason this process failed on this caption. For the other metrics, it appears that the extra mention of the mirror only amounts to a tiny penalty because all of them primarily count overlapping short phrases. We conclude that these metrics generally have difficulty penalizing captions that human readers can easily recognize are not accurate.

4.2 “Never-Ending” Captions

Table 2 illustrates another anomaly that is prevalent in our machine-generated captions, which we have named the “never-ending” caption. This caption and others like it contain long sequences of periods and repeated words and phrases. Also, all of these captions are 50 tokens long, which is the maximum length of a caption produced by the NOC model using the default settings the authors provide in the source code. When we analyzed the lengths of all of the captions generated by our model, we found that 10,486 (25.89%) of them are at the maximum length.

To confirm that this behavior represented a proper functioning of the model, we examined a set of captions that have been made publicly available by the NOC model authors.[3] We searched the 20,252 captions in this data set for two or more consecutive periods and found that 26.3% of the captions (5,335 out of 20,252) had this characteristic, nearly identical to the percentage of our captions. However, visually scanning the captions in the data set provided by the authors, we realized that this issue might not have been noticed previously because the maximum caption length appeared to be set at 25 words.

Researchers commonly use post-processing techniques to make captions more presentable before metric evaluation. These never-ending captions could be fixed with a post-processing algorithm that truncates the caption at the first period. However, the caption might be incomplete at the first period and cutting it there runs the risk of removing salient information. We also found instances of anomalous maximum length captions where this would not work because they contained no periods at all (e.g., “Parking meter parking meter parking meter . . . ”). The appearance of consecutive periods in the data set provided by the NOC authors indicates that no post-processing is being applied except for a truncation of the captions at 25 words.

The never-ending caption scores fairly poorly in a majority of the standard evaluation metrics; this is because if the machine-generated caption (the candidate) is very long and contains many short phrases that do not exist in the ground-truth human caption (the reference), metrics that contain a precision score component will penalize these as false positives. Even if a phrase that appears repeatedly in the candidate appears in the reference, additional instances of it in the candidate will be penalized as noise. On the other hand, the SPICE metric gives an extremely high score in spite of the fact that it was designed to be a better metric for evaluating image-captioning systems. We speculate that this issue is also due to the fact that SPICE parses the caption to build a semantic graph, and the process may not penalize sequences of periods or ungrammatical repeated phrases in the sentence structure.

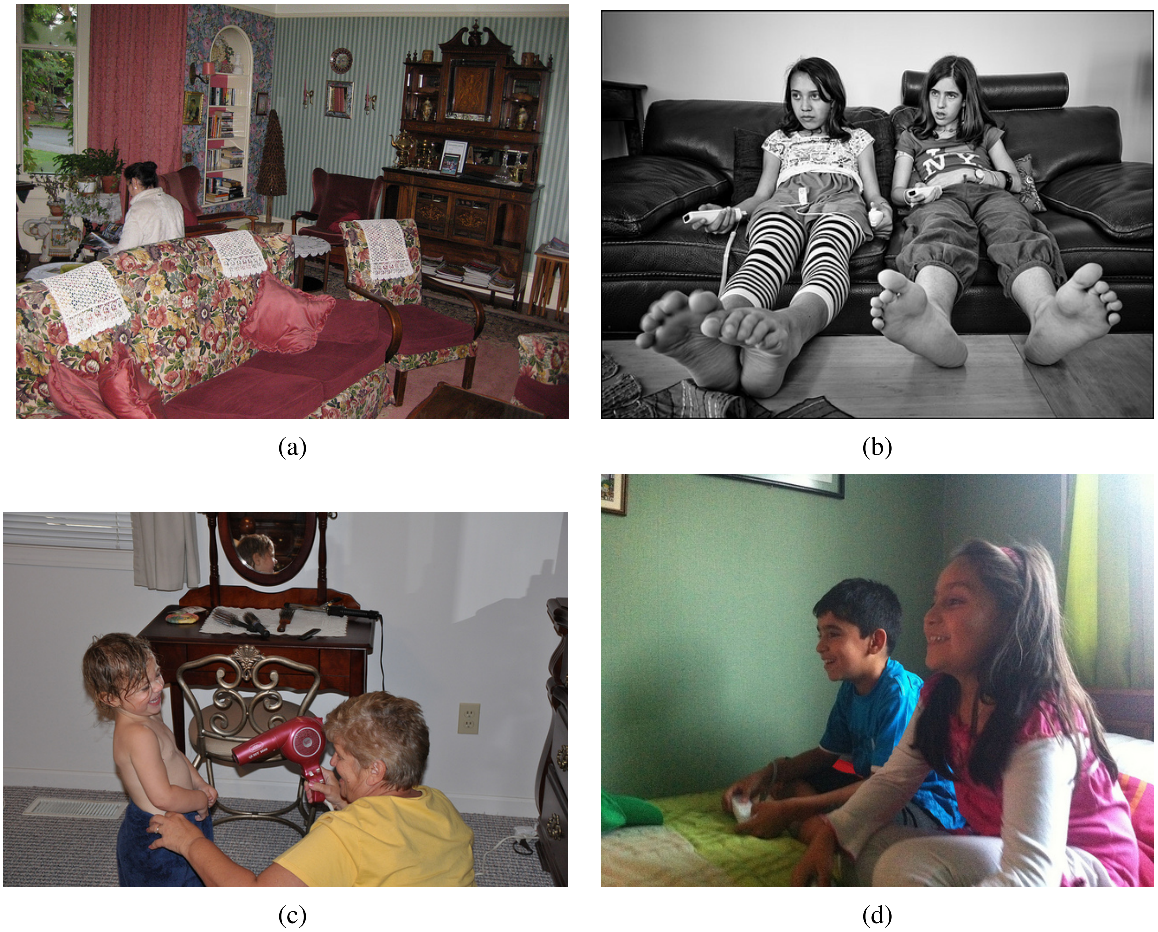
5. Linguistic Variation

We also worked to determine whether captions produced by deep learning were human-like and displayed the natural variety, diversity, and originality of words and structures that are present in natural language. Tables 1 and 2 show examples of this kind of variety in the five ground truth captions written by humans to describe the MSCOCO images.

5.1 Tests for Originality

Examining the captions, we noticed that the model frequently produced the same caption for many different images and that these repeated captions were a good match for some of those images but not others. For example, the model captioned 29 separate images with the identical caption “A

[3] https://vsubhashini.github.io/noc.html



(a)

(b)

(c)

(d)

Figure 1. Four images captioned as “A woman sitting on a bed with a teddy bear”. The model captioned 29 separate images with this identical caption, which exists as a ground-truth caption in the training set, but the model seems to use it incorrectly on many of the images.

woman sitting on a bed with a teddy bear”, but the model seems to use it incorrectly on many of them. Four instances are shown in Figure 1. This caption also appears 11 times as a ground-truth caption in the training set, but there are more than ten times as many captions in the training data as in our collection of machine captions. This example raises a question: is the system really generating unique captions based on the objects it finds in the images or is it merely “remembering” and reproducing a ground truth caption verbatim based on a few features in the image?

The first way that we assessed the originality in the machine-generated captions was by determining how many were exact duplicates of captions in the training data. We found that, out of the 40,504 captions generated from the MSCOCO validation data, there were exact matches for 4,988 (12.3%) in the 413,915 ground-truth captions in the 82,783 training cases.

We were also concerned about whether the system generally tends to reproduce the same captions over and over again, regardless of whether they are identical to ground-truth captions or not, so

Table 3. Texts and frequencies for the ten highest frequency machine captions.

Caption	Frequency
“A bathroom with a toilet and a sink.”	110
“A man riding a wave on top of a surfboard.”	109
“A man riding a skateboard down a street.”	108
“A man riding a surfboard on top of a wave.”	96
“A man riding skis down a snow covered slope.”	93
“A plate of food with a fork and a plate of food.”	77
“A clock tower with a clock on top of it.”	67
“A man riding a skateboard down a ramp.”	59
“A woman sitting at a table with a plate of food.”	56
“A baseball player swinging a bat at a ball.”	55

we calculated the frequency of each machine-generated caption in the whole corpus. Of the 40,504 captions we generated, 25,631 (63%) are unique captions that appear only once, while the remaining 14,873 (37%) appear as duplicates of another 3,423 captions. Some 563 of these repeated captions appear more than five times and 227 of them appear over ten times.

Table 3 lists the ten highest-frequency machine captions. We notice that these captions all have a similar structure, such as “an object is doing something on/in some place”. Moreover, half of them are about similar objects and actions, such as “riding a skateboard”. We hypothesize that there may be systematic biases in the model, a prevalence of these kinds of images in the data set, or both. While these high-frequency captions may match the image content (which we did not check) they generally show a lack of variety in terms of structure and content.

5.2 Vocabulary Size

To give a sense of the lexical variation present in the machine-generated captions, we compared the size of the vocabulary of the captions generated by our model with that of human-generated ground-truth captions (a method also employed by Krause et al., 2017). We calculated the number of unique words and tokens in our 40,504 machine-generated captions. We also calculated unique words and tokens in the human-generated captions in both the 82,783 training data items and the 40,504 validation data items. Since five human-written sentences are provided for each image in the training and validation sets, we randomly selected one of them from each image to ensure a fair comparison between the humans and the model. We did not perform stemming, lemmatization, or spelling correction on any of the captions beforehand.

We found that the machine-generated captions contain 2,297 unique words, which is about 24% of the vocabulary of 9,466 unique words in the 20% sample of the human-generated ground truth captions in the validation data; these 2,297 words are also only 9% of the 24,781 unique words in our 20% training sample. This shows that the image-captioning model is limited in its use of

Table 4. Examples of part-of-speech tags (b), and dependency parses (c) of a 15-word machine-generated caption and a 16-word human-written ground-truth caption describing similar scenes (a). Both are from our data set. However, the machine-generated caption shown was changed slightly from "snow covered" to "snow-covered" so that it would parse correctly in this example.

	Machine-generated Caption	Human-generated Caption
(a)	"A large wing view of a airplane sitting on top of a snow-covered mountain."	"Several airplanes can be seen at the airport, but there is also snow on the ground."
(b)	A/DT large/JJ wing/NN view/NN of/IN a/DT airplane/NN sitting/VBG on/IN top/NN of/IN a/DT snow-covered/JJ mountain/NN ./.	Several/JJ airplanes/NNS can/MD be/VB seen/VBN at/IN the/DT airport/NN ,/, but/CC there/EX is/VBZ also/RB snow/NN on/IN the/DT ground/NN ./.

(c)	det(view-4, A-1)
	amod(view-4, large-2)
	compound(view-4, wing-3)
	root(ROOT-0, view-4)
	case(airplane-7, of-5)
	det(airplane-7, a-6)
	nmod(view-4, airplane-7)
	acl(view-4, sitting-8)
	case(top-10, on-9)
	nmod(sitting-8, top-10)
	case(mountain-14, of-11)
	det(mountain-14, a-12)
	amod(mountain-14, snow-covered-13)
	nmod(top-10, mountain-14)
		amod(airplanes-2, Several-1)
		nsubjpass(seen-5, airplanes-2)
		aux(seen-5, can-3)
		auxpass(seen-5, be-4)
		root(ROOT-0, seen-5)
		case(airport-8, at-6)
		det(airport-8, the-7)
		nmod(seen-5, airport-8)
		cc(seen-5, but-10)
		expl(is-12, there-11)
		conj(seen-5, is-12)
		advmod(is-12, also-13)
		nsubj(is-12, snow-14)
		case(ground-17, on-15)
		det(ground-17, the-16)
		nmod(snow-14, ground-17)

words in comparison to human annotators, and this is another respect in which the model lacks the linguistic variety of human-generated captions.

5.3 Syntactic and Semantic Variation

In order to investigate the syntactic and semantic structures of machine and human captions, we used the Stanford CoreNLP parsing system (Manning et al., 2014) to parse the 40,504 machine captions and the 20% sample of human captions from the validation data set for comparison. Two basic parsers in the Stanford CoreNLP suite of tools were used for evaluation: the part of speech parser and the dependency parser.

The part of speech parser tags each word with one of several dozen well-known grammatical categories, including noun, verb, adjective and so on. The dependency parser employs phrase-structure rules and detects dependency relationships between pairs of words in a sentence (De Marneffe et al., 2014). Dependency parsing systems are designed as an alternative to traditional phrase structure representations to provide output that is more easily accessible to non-linguists focused on extracting information, textual relations, and graph representations of text. We propose that variety in the

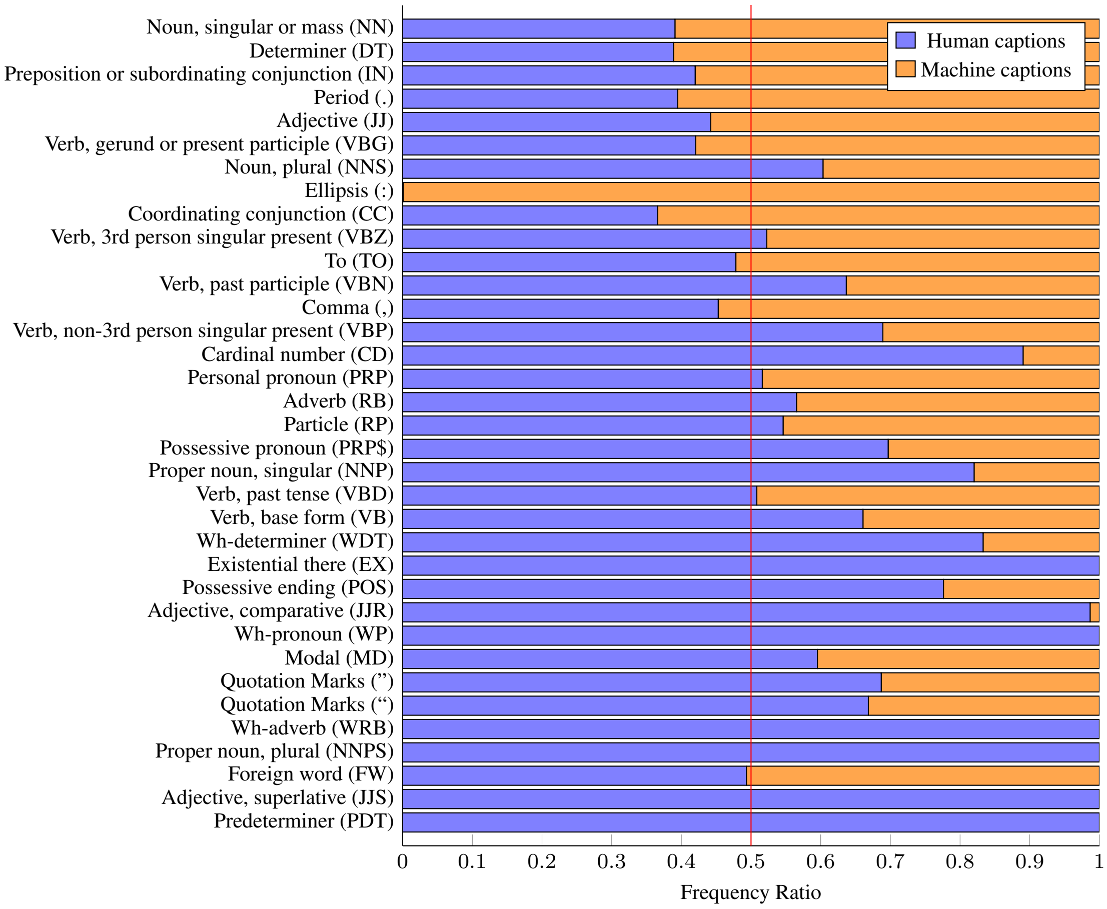


Figure 2. A comparison of the sentence structures of machine captions and human annotated captions via the frequency ratios of part of speech tags of each in our data set. The plot is ordered by the total count of each tag, with most common tags appearing at the top. Only tags occurring more than ten times are shown.

part-of-speech and dependency categories is a signal of the overall complexity of the syntactic and semantic structures of the captions from different sources. Table 4 shows examples of tagging and dependency parses of human and machine captions from our data set.

5.3.1 Part-of-Speech Tag Comparisons

We calculated counts of the appearance of each part of speech tag over all of the parses of the machine-generated and human-generated captions. Since the frequencies of different tags over the data set varied from over 100,000 for nouns (nn) and determiners (dt) to less than 100 for superlative adjectives (jjs), we calculate and plot the ratios of tag frequencies between the human and machine instead of raw frequencies. Figure 2 shows the frequency ratios of different tags arranged from highest to lowest frequency, denoted as blue and orange bars for human and machine captions, respectively; the topmost tag is most frequent, while the one on the bottom occurs least often.

We found that the machine captions have greater percentages for many of the more frequently occurring part-of-speech tag categories, with the exception of plural nouns (NNS). In both the example parses (Table 4) and the frequency ratio plot (Figure 2) the machine-generated captions had higher instances of parts of speech used to form simple sentences: singular or mass nouns (NN), determiners (DT), coordinating conjunctions (CC), prepositions/subordinating conjunctions (IN), adjectives (JJ), the word "to" (TO), third person singular present verbs (VBZ), and present participle ("-ing") verb forms (VBG). We found that machine-generated captions also had more commas and periods. All ellipsis tags (appearing as ":" in the tag set) came from machine captions; the parser recognizes the long sequences of periods in never-ending captions as concatenations of ellipses.

In contrast, human-generated captions have dramatically greater numbers of plural nouns (NNS), past participle verbs (VBN), past-tense verbs (VBD), non-third person singular present verbs (VBP), verb base forms (VB), third-person singular verbs (VBZ), modal verbs (MD), possessive and personal pronouns (PRP and PRP$), possessive endings (POS), comparative adjectives (JJR), singular proper nouns (NNP), and wh-determiners (WDT). They also had more particles (RP, e.g., "over", "down", "out"), adverbs (RB), and cardinal numbers (CD). These and elements like the existential *there* (EX) in Table 4 are used to form more diverse and complex sentences.

5.3.2 Dependency Parse Comparisons

We also obtained dependency parses of all of the human and machine-generated captions and calculated counts of the appearance of each dependency relationship. Again, since there was a huge range of frequencies of different tags over the data set, we calculate and plot the ratios of dependency relationship frequencies between the human and machine instead of raw frequencies. Figure 3 shows the frequency ratios of different dependency relationships arranged from highest to lowest total frequency; the topmost tag has the highest frequency while the tag on the bottom is lowest.

In the dependency parses,[4] again the machine-generated captions dominated relationships for simpler sentences: determiners (det), prepositions (case, nmod), adjectival modification (amod, acl), conjunctions (conj, cc), compound nouns (compound), and punctuation (punct). The dependency parse of the example machine-generated caption in Table 4 has seven kinds of dependency relationships, while the human-written caption has fifteen. As with the part of speech tags, the human-generated captions more frequently featured relationships of more complex sentences, such as clausal modifier relationships (xcomp, acl:relcl, advcl, ccomp, csubj), passive voice (auxpass, nsubjpass), negation (neg), and multi-word expressions (mwe), among others.

5.3.3 Permutation Tests

We also gathered quantitative statistical evidence that machine-generated captions have greater instances of higher-frequency tags and dependency relationships. We performed permutation tests to demonstrate that the difference between the average ratio of machine captions is statistically significant between high-frequency and low-frequency categories. To do this with the tags, we divided them in half according to frequency, one set being the high-frequency tags and the other the low-frequency tags. We define the means of the ratios of tag frequencies in these two sets as $\bar{r}_h$ and $\bar{r}_l$.

[4] `http://universaldependencies.org/docs/en/dep/`

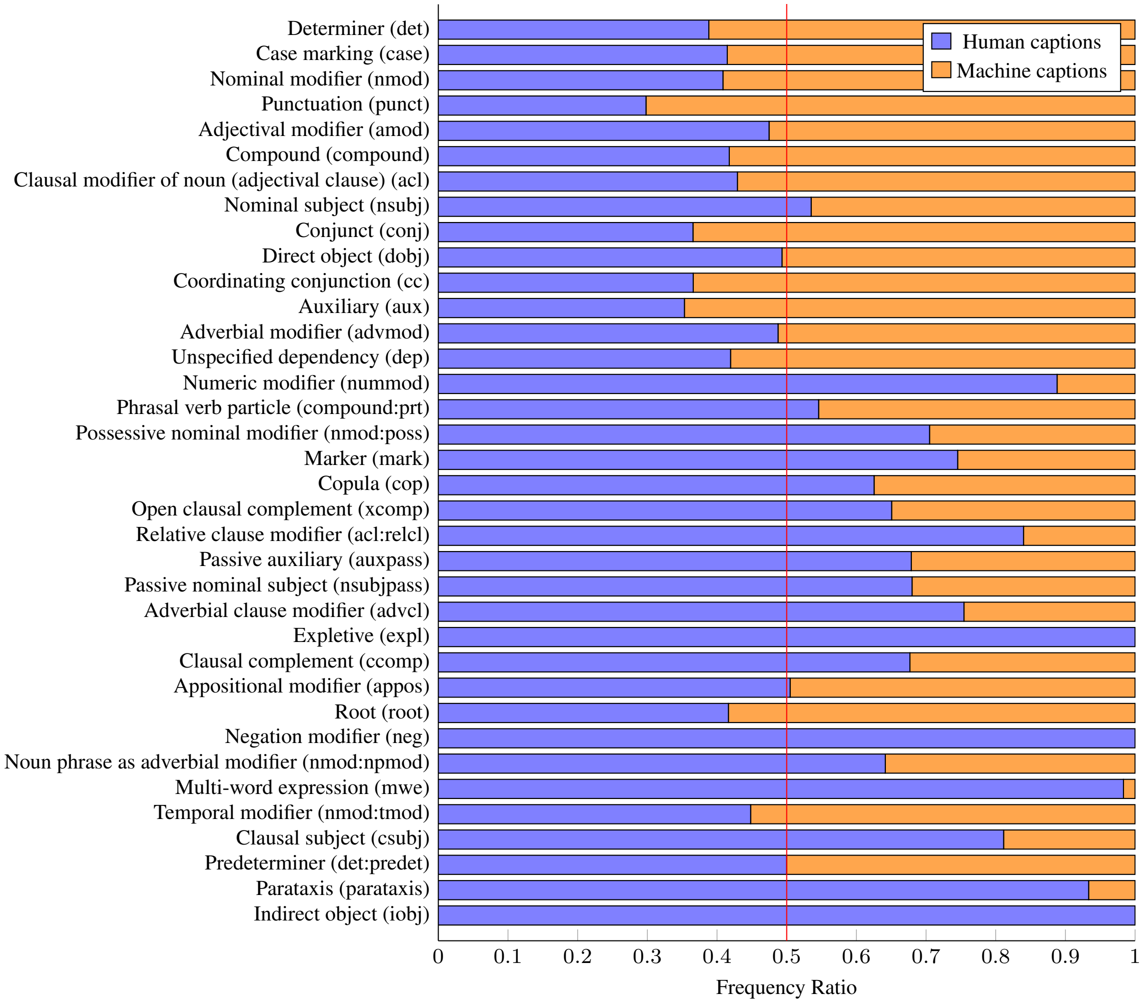


Figure 3. A comparison of the sentence structures of machine captions and human annotated captions via the frequency ratios of dependency relations in our data set. The plot is ordered by the total count of each dependency relation, with the more common relations appearing at the top. Only relations occurring more than ten times are shown.

The null and alternative hypotheses are

$$H_0 : \bar{r}_h - \bar{r}_l = 0$$

and

$$H_A : \bar{r}_h - \bar{r}_l > 0.$$

We did the same for dependency relationships. The permutation test statistics are shown in Table 5. We have enough statistical evidence to reject the null hypothesis H_0 and we can conclude that, on average, the machine captions have greater numbers of high-frequency tags and dependency relationships than low-frequency ones.

Table 5. Permutation test results for ratios of part of speech tags and dependency relationships.

Test	Part of Speech Tags	Dependency Relations
Number of Rounds	100,000	100,000
Observed Value of $\bar{r}_h - \bar{r}_l$	0.3836	0.2630
p-value	0	5×10^{-5}

6. Discussion

Our study provides insights into understanding the functioning of image-captioning models and areas for improving them. While we confirmed the existence of "stutter" captions, we were amazed that more than a quarter of the captions produced by our model were anomalous maximum-length "never-ending" captions, providing obvious differences between how the models and humans generate captions. Stutter captions may be occurring due to weaknesses in the recurrent architecture in the language model and in the word embedding representations. With never-ending captions, during pre-processing special "start" and "end" tokens are attached to the beginning and end of each caption in the training data, and the recurrent language model is supposed to learn to produce an "end" token to signal that the caption is complete. Apparently, in many instances, the model fails to do this and the system continues running the model until the maximum caption length is reached. Since most systems employ the same language models and pre-processing techniques as NOC, we suspect that this is a widespread issue.

However, the evidence we gathered from comparing the overall linguistic variation of machine- and human-generated captions is unique and strong in supporting our claim that human-like understanding of an image is a more involved cognitive process than simple classification or object recognition. The comparison of vocabulary size demonstrates that the model only "remembers" a small percentage of words in the training data—likely the most frequently-occurring ones. In addition to the roughly 25,000 word vocabulary in the training data, the NOC model word-embedding system uses an external corpus with six billion tokens and a vocabulary size of 72,700 words (Venugopalan et al., 2017). However, this additional vocabulary available at training does not appear to have a strong effect on the vocabulary of the generated captions. We hypothesize that the neural network training process inherently places greater weight on high-frequency phrases, syntactic patterns, and entire sentences. This could explain the high percentage of repeated captions and captions with exact matches in the training data.

When humans compose a caption for an image, they obviously use visual perception and linguistic knowledge. But humans do more than simply identify the objects in the image and their relative positions in a natural language description; they also apply complex knowledge structures about the physical and social situations in which the objects are involved, and how they interact, as in the Heuer et al. (2016) example of "woman and dog with frisbee on grass near fence" versus "a woman playing tug of war with a dog over a white frisbee." Observing human-generated captions in aggregate, the complexity of the language given in response to visual scenes reflects the sophistication of understanding a visual scene, the different aspects of scenes and different parts of images that people notice, their implications and consequences, and individual differences among them.

Our finding that the machine-generated captions tend not to have this linguistic complexity is evidence to the claim that these image-captioning systems are not learning these kinds of knowledge structures and are not truly understanding images in the same ways as people. This is consistent with related findings (Heuer et al., 2016) supporting the view that, by and large, these systems are simply recognizing objects and composing lists of simple noun phrases. At the same time, this is somewhat unsurprising, given that the models, the data sets used to train, validate, and test them, and the metrics used to judge their performance generally have not been designed with human-like understanding in mind.

7. Directions for Future Work

In this paper, we illustrated that neural network image-captioning models have several special challenges with respect to caption errors and syntactic structures, and, unfortunately, widely-used evaluation tools do not seem to detect these problems. Our study, while showing various ways in which image understanding is a much greater process than image object classification, provides important illustrations about how these machine-captioning systems actually work and the ways their behavior does not resemble human performance. In our corpus of machine-generated captions, we found a large number of examples of captions that are clearly erroneous, or appear repeatedly and often poorly describe the source image. The “stutter” and “never-ending” caption errors clearly reveal a deficiency in the language models used to handle this vision-to-language task. Our experiment also demonstrates that the models may converge towards simple sentences in the training data. This appears to be due to approaches to image captioning where it is posed primarily as a computer vision and classification problem, rather than a joint task between computer vision and natural language generation. We also measured and observed how machine-generated captions lack the linguistic and descriptive variety of human captions.

One broader question concerns the modern big-data paradigm of machine learning research and the possibility that systems in this paradigm only give good results because of the narrow scope of the tasks and the way the data sets are constructed. There may usually be a “good answer” or a “good response” to a stimulus in the validation or test data based on “shallow” patterns in the training data. If “shallow” metrics are what define a “good response”, then these systems can exploit simple patterns in data sets to achieve “good” performance without ever needing to perceive, think, reason, and respond at a human level. Resolving this question requires further investigation of the data sets and the metrics themselves.

Future work could employ crowdsourcing for a more thorough analysis of patterns in machine-generated captions that would gain greater insights into the inner workings of these models. Another possibility for future work is to run a number of other models to build a larger corpus of captions on which to perform further studies. Although the NOC model is representative of these systems, a larger study could reveal more patterns in machine-generated image captions. Better optimization strategies or ways of weighing visual and sentence representations may help reduce obvious anomalies in these captions. Also, more could be done to detect and penalize these kinds of problems in the model training process. The kinds of anomalies we found through cursory inspection can be easily detected, so post-processing may reduce their impact on the caption quality with respect to commonly-used simple metrics.

Due to the optimization limitations of neural networks, the training phase for deep network captioning systems requires caption-level comparisons, and all widely-used metrics, such as BLEU and SPICE, only consider captions for a single image. However, our analysis shows the noticeable corpus-level differences between machine and human captions, and novel training systems that reward corpus-level similarity will let researchers evaluate and improve their models from this perspective. The ideal metric can also include variance in vocabulary and sentence structure, in a way similar to our tests of linguistic variance. If the models are redesigned to take word and caption frequencies into consideration, they may avoid overfitting to simple sentence structures. To be specific, we might give a linguistic complexity weight to each image-caption training case that favors captions with lower-frequency words or more complicated syntactic structures during optimization. Another possible strategy for improvement is to train the language model with a more linguistically-diverse text corpus to give it a broader search space for semantics and vocabulary.

But the prevalence of these anomalies in spite of the huge amount of training data provided is an indication of the great limitations of incremental development of these kinds of models. Assuming the larger claim of this paper, cognitive systems researchers will probably be well served by the pursuit of novel image-captioning architectures that perform sophisticated and complex human-like understanding and language generation processes.

It is obvious that, in performing the captioning task, humans have access to more visual information, episodic and semantic memories, knowledge, and reasoning capability than that provided in data sets like MSCOCO. A better approach to building a cognitive system for image captioning based on machine learning could use sensory information in the form of active, non-canonical views and videos of similar objects, observations of objects with other senses, and interactions with them via actuators. These kinds of data could easily be provided by the sensors and actuators of cognitive robotic systems (Haber & Sammut, 2013) and tied to the language used to describe the activities in visual scenes (Scheutz et al., 2013).

However, if the generation of complex language by humans is driven by complex knowledge and reasoning, as we have hypothesized, these better language models will need to be integrated with complex commonsense knowledge structures (e.g., Gilpin et al., 2018). This will require subsystems that can reason about the objects, figures, and background information present in an image, and draw conclusions about the acts and events that are happening and their consequences as part of caption generation. To support this kind of reasoning, they will need access to structured knowledge about stereotypical social situations and about the goals and plans of actors, particularly when humans are in the images (Lenat, 1995; Speer & Havasi, 2013).

Finally, performance metrics are driving the development of these systems, and a more comprehensive analysis of the most commonly-used evaluation metrics, including their distinct properties and potential drawbacks, is currently underway. New metrics to assess image captions from other dimensions (e.g., level of detail or commonsense knowledge) that align with a more comprehensive human judgment of caption quality are worthy of invention. One of the most interesting characteristics we discovered about image captioning is the intrinsic structural variation of the language of acceptable captions for a particular image. While we did perform comparisons of structural variation of language in this paper, we are motivated to design quantitative metrics of variation to aid in these evaluations and comparisons, and diversity metrics used in ecology and other life sciences might be adapted to this purpose. The idea that human intelligence can seemingly generate an infinite variety of language behaviors from finite means formed an important part of the cognitive revolution. It is only natural that cognitive systems research should incorporate better measures and use them to raise expectations in assessing the human-like qualities of computational artifacts.

Acknowledgements

We thank the anonymous reviewers, the ACS Conference Program Chair and the Editor of this journal for their thoughtful comments, which shaped our revisions and led to significant improvements of this paper. We also thank Sambit Bhattacharya for discussions that inspired this work.

References

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. *Proceedings of the Fourteenth European Conference on Computer Vision* (pp. 382–398). Amsterdam: Springer.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Ann Arbor, MI: Association for Computational Linguistics.

Chen, H., Zhang, H., Chen, P.-Y., Yi, J., & Hsieh, C.-J. (2018). Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *Proceedings of the Fifty-Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 2587–2597). Melbourne, Australia: Association for Computational Linguistics.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 4585–4592). Reykjavik, Iceland: European Language Resources Association.

Gilpin, L. H., Macbeth, J. C., & Florentine, E. (2018). Monitoring scene understanders with conceptual primitive decomposition and commonsense knowledge. *Advances in Cognitive Systems*, *6*, 45–63.

Haber, A., & Sammut, C. (2013). A cognitive architecture for autonomous robots. *Advances in Cognitive Systems*, *2*, 257–275.

Heuer, H., Monz, C., & Smeulders, A. W. (2016). Generating captions without looking beyond objects. *arXiv:1610.03708*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hossain, M. Z., Sohel, F. A., Shiratuddin, M. F., & Laga, H. (2018). A comprehensive survey of deep learning for image captioning. *Computing Research Repository*, *abs/1810.04020*.

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2021–2031). Copenhagen: ACL.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. *Proceedings of The Thirtieth IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 317–325). Honolulu, HI: IEEE.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*, 33–38.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Proceedings of the Thirteenth European Conference on Computer Vision* (pp. 740–755). Zurich: Springer.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of Fifty-Second Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, MD: Association for Computational Linguistics.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 427–436). Boston, MA: IEEE.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C., & Leake, D. B. (1989). Creativity and learning in a case-based explainer. *Artificial Intelligence*, *40*, 353–385.

Scheutz, M., Harris, J., & Schermerhorn, P. (2013). Systematic integration of cognitive and robotic architectures. *Advances in Cognitive Systems*, *2*, 277–296.

Speer, R., & Havasi, C. (2013). ConceptNet 5: A large semantic network for relational knowledge. In I. Gurevych & J. Kim (Eds.), *The People's Web meets NLP*, 161–176. New York: Springer.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*, 193–254.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4566–4575). Boston, MA: IEEE.

Venugopalan, S., Hendricks, L. A., Rohrbach, M., Mooney, R., Darrell, T., & Saenko, K. (2017). Captioning images with diverse objects. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5753–5761). Honolulu, HI: The Computer Vision Foundation.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and Tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 652–663.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*, 1172–1217.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.