Measuring the Diversity of Automatic Image Descriptions

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

Desmond Elliott
University of Edinburgh
d.elliott@ed.ac.uk

Piek Vossen
Vrije Universiteit Amsterdam
piek.vossen@vu.nl

Abstract

Automatic image description systems typically produce generic sentences that only make use of a small subset of the vocabulary available to them. In this paper, we consider the production of generic descriptions as a lack of diversity in the output, which we quantify using established metrics and two new metrics that frame image description as a word recall task. This framing allows us to evaluate system performance on the head of the vocabulary, as well as on the long tail, where system performance degrades. We use these metrics to examine the diversity of the sentences generated by nine state-of-the-art systems on the MS COCO data set. We find that the systems trained with maximum likelihood objectives produce less diverse output than those trained with additional adversarial objectives. However, the adversarially-trained models only produce more types from the head of the vocabulary and not the tail. Besides vocabulary-based methods, we also look at the compositional capacity of the systems, specifically their ability to create compound nouns and prepositional phrases of different lengths. We conclude that there is still much room for improvement, and offer a toolkit to measure progress towards the goal of generating more diverse image descriptions.

1 Introduction

Automatic image description is a challenging task because natural language and the visual world both have an unbounded range of variation (Bernardi et al., 2016). Computational image description models are trained to generalize over datasets of images with multiple human descriptions, however, much of the variation present in these descriptions is lost in a trained model. Dai et al. (2017) note that the descriptions generated by recurrent neural networks using a maximum-likelihood objective are "overly rigid and lacking in variability." This rigidity and lack of variability in the output of state-of-the-art models is unfortunate because human descriptions are the exact opposite of this: Devlin et al. (2015) found that humans typically produce unique descriptions, i.e. only 4.8% of the human-described evaluation data in the MS COCO data set (Lin et al., 2014) also occur in the training data.

The lack of variability in machine-generated text is not limited to automatic image description; it is a general problem in natural language generation. Simply put: automatically generated text quickly becomes boring or repetitive. Recent efforts to address this problem include using maximum mutual information as an objective function, rather than the likelihood of the output, to improve the variablity of a neural conversation model (Li et al., 2016). Castro Ferreira et al. (2016) focused on the deterministic nature of NLG systems, in the sense that they repeatedly use the same referential forms to refer to the same entity in longer stretches of text. They addressed this problem by explicitly training their model to mimic human variability for referring expression generation.

In the image description literature, there have been two recent approaches to generating diverse outputs: (i) learn different description distributions simultaneously to generate multiple different descriptions for the same image (Wang et al., 2016); and (ii) augmenting a model with an additional (conditional) Generative Adversarial Network objective (Goodfellow et al., ; Mirza and Osindero, 2014, GAN)

Proceedings of the 27th International Conference on Computational Linguistics, pages 1730–1741
Santa Fe, New Mexico, USA, August 20-26, 2018.

to generate more natural and diverse descriptions. In this setting, the caption generator tries to fool a discriminator that is trying to distinguish human image descriptions from machine-generated ones (Dai et al., 2017; Shetty et al., 2017). From these papers, two definitions of diversity emerge:

Local diversity: The ability to generate many different descriptions for the same image.
Global diversity: The ability to use (many different combinations of) many different words.

The former is *local* because it can be evaluated for individual images. The latter is *global*, because it is a property at the corpus level. This paper focuses on global diversity, which means that we will study whether systems are able to produce as many different words and phrases as humans do in their descriptions of images. We know that word frequencies follow a *Zipfian* (or *power law*) distribution (Zipf, 1949; Van Heuven et al., 2014; Corral et al., 2015), which means that a small subset of the vocabulary accounts for the largest part of the data. Natural language processing systems trained on corpus data are sensitive to this, and tend to overfit on the head of the distribution (e.g. Postma et al., 2016). We will show that this also holds for the output of image description systems: all systems considered in this paper mainly use the top 20% most frequent words.

In this paper, we consider the following question: **How can we measure the diversity of the output generated by an image description model?** There is currently a lack of consensus about how to measure the diversity of model output but the metrics used thus-far fall into four broad areas:

(i) Modified[1] type-token ratio: the number of distinct unigrams or bigrams, divided by the total number of generated words (Li et al., 2016; Shetty et al., 2017).
(ii) mBLEU: compute the average BLEU score (Papineni et al., 2002) between each description and the other descriptions generated for the same image. This metric can only be used to evaluate models that produce multiple descriptions per image (Wang et al., 2016; Shetty et al., 2017).
(iii) Model-internal: a Generative Adversarial evaluator network that judges whether descriptions are more natural-sounding and semantically relevant than human descriptions (Dai et al., 2017); and
(iv) vocabulary size and the proportion of uniquely generated sentences (Shetty et al., 2017).

In addition to this lack of consensus about which metrics should be used to measure diversity, it is not known how state-of-the-art systems differ in terms of output diversity because it has not been standard practice to report this type of performance statistics. In this paper, we present an overview of metrics to assess the diversity of automatically generated English image descriptions, and compare them using nine state-of-the-art image description systems (Section 2). Besides covering existing metrics, like TTR and average sentence length, we also propose two word recall metrics that provide more information about the output vocabulary (Section 3). We also look at the compositional capacity of the different systems, by examining how many different compound nouns and prepositional phrases they can produce. We use these metrics to analyse how image description systems differ from human descriptions (Section 4). It is not our goal to evaluate the quality of the descriptions, though future research may find that more diverse descriptions are also more attractive for human readers (Section 5.2).

The main finding of our analysis is that recent GAN-based systems (Dai et al., 2017; Shetty et al., 2017), designed to produce more human-like image descriptions, do indeed produce more diverse output than the other MLE-based systems, but this increased diversity still mostly comes from the head of the vocabulary. In order to support future analyses, we release a toolkit to assess the output of any system and to compare the results with existing approaches.[2]

2 Existing metrics

This section discusses six general metrics to measure output diversity at the word level, along with a method to visually inspect the differences between systems. All of these methods require tokenized image descriptions – we use SpaCy 2.0.4 for this purpose and lowercase all of the tokens. The validation data is different from the system output, in that it consists of 5 reference descriptions per image, while

[1]This is similar to the type-token ratio (TTR; number of types divided by number of tokens), except that it is customary to compute TTR over a fixed number of tokens, as TTR decreases with corpus size (Youmans, 1990).

[2]Toolkit: `https://github.com/evanmiltenburg/MeasureDiversity`

Type	System	BLEU	Meteor	ASL	SDSL	Types	TTR_1	TTR_2	%Novel	Cov	Loc_5
MLE	Liu et al. 2017	32.3	25.8	10.3	1.32	598	0.17	0.38	50.1	0.05	0.70
	Mun et al. 2017	32.6	25.7	9.4	1.12	1009	0.16	0.38	50.0	0.08	0.78
	Shetty et al. 2016	31.9	25.2	9.0	1.03	1112	0.15	0.34	43.0	0.08	0.74
	Tavakoli et al. 2017	28.7	23.5	9.2	1.03	917	0.15	0.33	38.8	0.07	0.66
	Vinyals et al. 2017	32.1	25.7	10.1	1.28	953	0.21	0.43	90.5	0.07	0.69
	Wu et al. 2016	31.0	25.0	9.1	1.03	849	0.14	0.32	44.5	0.06	0.72
	Zhou et al. 2017	30.0	24.8	9.3	1.20	1334	0.22	0.51	60.1	0.10	0.80
GAN	Dai et al. 2017	20.7	22.4	9.8	1.63	1922	0.23	0.55	87.7	0.15	0.76
	Shetty et al. 2017	–	23.6	9.4	1.31	2611	0.24	0.54	80.5	0.20	0.71
	Validation data	–	–	11.3	2.61	9200	0.32	0.72	95.3	–	–

Table 1: System results: BLEU and Meteor scores; average sentence length; standard deviation of sentence length; mean-segmented type-token ratio (TTR); bigram TTR; percentage novel descriptions; coverage; and local recall with importance class 5. BLEU/Meteor scores are originally reported values, except for (Dai et al., 2017) and (Vinyals et al., 2017), which we computed on the validation set.

the systems only produce one description per image. Hence, for the validation data, we compute each score 5 times – once per reference description – and report the average.

1. The **average sentence length (ASL)** corresponds to the mean number of tokens per sentence.
2. The **standard deviation of the sentence length (SDSL)** is a measure of how much systems vary in their description lengths.
3. The **number of types** measures the number of unique word types in the output vocabulary.
4. The **mean segmented type-token ratio (TTR_1)** is the average number of types per 1000 tokens (Johnson, 1944). It is not affected by sentence length because it is computed for a fixed number of tokens. It is more difficult to artificially increase than the number of types because it is an average.
5. The **bigram TTR (TTR_2)** is the average number of bigram types per 1000 bigram tokens. This is based on Li et al.'s (2016) diversity metric (looking at bigram diversity), and the MSTTR metric (using a fixed size, averaging over multiple samples) so that it is not biased by description length.
6. The **percentage novel descriptions (%Novel)** refers to the generated descriptions that do not occur in the training data. Note that there may be duplicates among the novel descriptions.

2.1 Systems

For any analysis of output diversity, it is essential to have the generated descriptions. Unfortunately, this data is generally not available for most published systems. We contacted the authors of papers that appeared in relevant conferences and journals between 2016–2017[3], and received nine responses with descriptions generated for the MS COCO validation set. All these systems are listed in Table 1. With the exception of the two GAN-based systems (Dai et al., 2017; Shetty et al., 2017), the other systems are based on a conditioned recurrent neural network, trained using a Maximum Likelihood (MLE) objective.

2.2 Results

Table 1 presents the results for the metrics discussed above. We discuss each of them in turn.

Average sentence length. We observe that all models produce shorter sentences than humans, on average, perhaps also conveying less information. It also means that the BLEU *brevity penalty* (Papineni et al., 2002) and Meteor *length penalty* (Denkowski and Lavie, 2014) are affecting the metric scores. However, *producing shorter sentences* does not necessarily mean *producing worse descriptions*.

Standard deviation of sentence length. We observe that the GAN-based systems vary more than most other systems, but the systems by Liu et al. (2017) and Vinyals et al. (2017) have more variation than other MLE-based systems. Humans vary much more than any model in the length of their descriptions.

Number of types. The model by Liu et al. (2017) produces the fewest distinct word types (598), which severely limits the output diversity of the system. The two GAN-based models produce the most

[3]We surveyed AAAI, ACL, BMVC, COLING, CVPR, EACL, EMNLP, ICCV, ICLR, ICPR, IJCAI, NAACL, and NIPS.

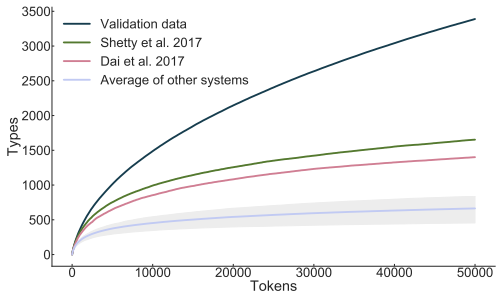

Figure 1: Type-token curves for nine systems. The validation data grows much faster than any of the systems and the GAN-based systems clearly outperform the other systems (shaded, with a line plotting the average performance).

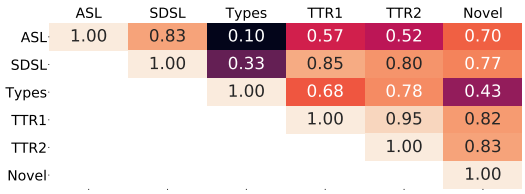

Figure 2: Absolute Spearman correlation between the different diversity metrics, computed over the results for the 9 different systems

distinct word types: 1,922 and 2,611. This is still much lower than the human type count, which averages at 9,200. The total number of types in the validation set is much higher, at 17,557.

TTR$_{\{1,2\}}$ We find that the GAN-based models again outperform the rest. But in terms of variation, there is still much room for improvement before they reach human parity.

Percentage novel descriptions. We find that the model by Vinyals et al. (2017) outperforms the rest (90.5% novel), with the GAN-based systems following close behind at 87.7% and 80.5% novel. The remainder of the systems reproduce a sentence from the training data approximately 50% of the time.

We visualize the differences between the systems using a **type-token curve (TTC)**, which shows how the number of types develops as one reads more output tokens (Youmans, 1990). This curve was originally proposed to compare different texts, which means that sentence order is fixed. With automatic image description, we do not have this constraint. Rather than taking a single sample, and reading the image descriptions in a single order, we randomized the order of the descriptions ten times, and computed the average TTC for the validation data for each system. Figure 1 shows the type-token curves for the validation data and all systems. We observe that the TTC for the validation data develops much more rapidly than that of the systems. Moreover, we can clearly see how the two GAN-based systems stand out from the others in producing more diverse output.

We now inspect how strongly the different existing metrics correlate with each other. Figure 2 shows the correlation matrix between the different general metrics for measuring diversity. We observe that TTR_1 and TTR_2 are almost perfectly correlated. We conclude from this that a single type-token ratio measure is enough to capture differences between systems in their use of different types. The number of novel descriptions is strongly correlated with the type-token ratio. An intuitive explanation for this is that whenever a model produces more varied output, it is also more likely to produce novel output. In this light, it is interesting to observe the lower correlation between the number of types and the percentage of novel sentences. An explanation for this may be that producing more different types in total does not necessarily mean more diverse output. A system has to *consistently* produce more different types to have an impact.

3 Image description as word recall

Image description can be simplified to a word recall problem, where the goal is simply to produce a bag of words that should overlap with the reference data. By ignoring sentence structure, we can focus on the richness of the vocabulary, and study system performance for different classes of words. We distinguish between global recall, looking at the corpus as a whole, and local recall, looking at the corpus image-by-image. We also introduce ranking measures based on these concepts.

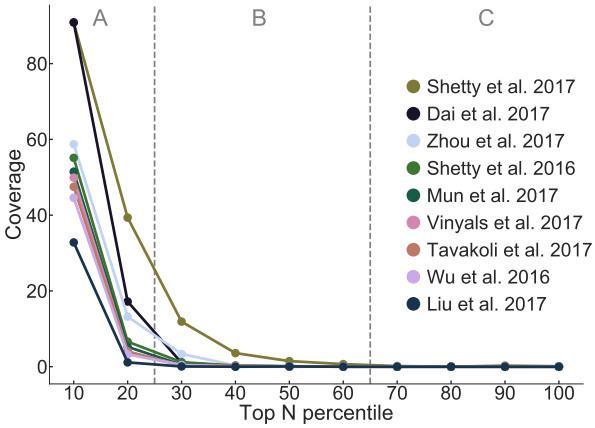

Figure 4: Coverage (equation 3) for different subsets of the learnable words. Recall for all systems is best for the top 10% most frequent words, but immediately drops for the next 10% of most frequent words.

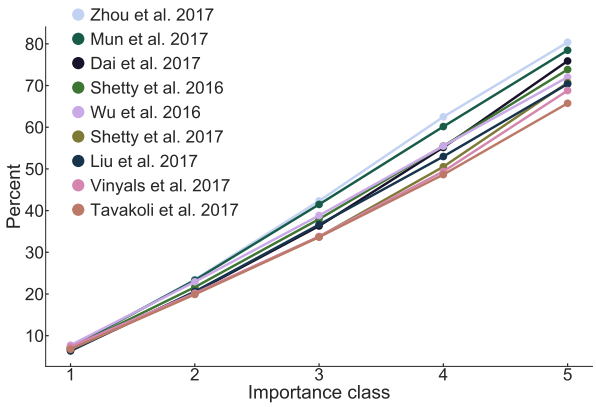

Figure 5: Local recall scores for all systems for each word importance class. Systems have low recall for words that occur only once in the reference descriptions, but their recall grows to 65-80% when all five references mention the same word.

3.1 Global recall

We formally define the *global recall* metrics in Figure 3. The sets TRAIN, EVAL, GEN correspond to the words that are in the training set, evaluation set, or those generated by the model. Any word type that is both in TRAIN and EVAL is *learnable* from the training data (Eq. 1).[4] Recalled words are those that are both learnable and generated by the model (Eq. 2). We quantify the success of a system as the percentage of learnable words it can recall, i.e. coverage (Eq. 3). Since the set of learnable word types is a subset of the word types in EVAL (this follows from (Eq. 1)), systems that are trained on the training data alone cannot recall *all* word types in EVAL. We define this limit in (Eq. 4). Intuitively, a model that has a higher coverage (Eq. 3) can recall more types from the learnable set (Eq. 1), therefore the model is producing a more globally diverse output.

Using the coverage metric to evaluate the nine systems, we find that the GAN-based systems of Shetty et al. (2017) and Dai et al. (2017) once again achieve the highest scores, achieving 15-20% coverage. This still leaves much room for improvement. We further explore coverage for 10 different subsets of the learnable word types, ranging from the 10% most to the 10% least frequent types in the validation data (based on the counts in the validation set).

$$\text{Learnable} = \text{TRAIN} \cap \text{EVAL} \quad (1)$$

$$\text{Recalled} = \text{GEN} \cap \text{Learnable} \quad (2)$$

$$\text{Coverage} = \frac{|\text{Recalled}|}{|\text{Learnable}|} \quad (3)$$

$$\text{Limit} = \frac{|\text{Learnable}|}{|\text{Eval}|} \quad (4)$$

Figure 3: Definitions for global recall.

Figure 4 shows the results. We see that the two GAN-based systems achieve almost 90% coverage of the most frequent types, but this score quickly degrades. Other systems only achieve about 60% coverage of the head, and degrade even more quickly than the GAN-based systems. Furthermore, we observe that the GAN-based systems only achieve better coverage than the other systems on the head of the distribution. Dai et al.'s system is only better for the 0–20% most frequent terms (part A), and Shetty et al.'s (2017) system still shows higher coverage than the others up to the 60% mark (part B), but there is no difference for the rest of the lexicon (part C). We emphasize that, for global recall, a system only has to use a type *once* for it to be counted. The Limit for the MS COCO validation set is 0.75. This means that the other 25% (4356 words) in the validation set cannot be learned on the basis of the training set.

[4] We ignore zero-shot learning approaches that could learn to describe images using words outside the training data.

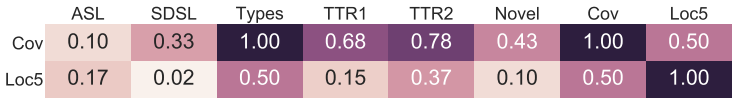

Figure 6: Spearman correlations between our coverage and local recall metric and the existing metrics.

3.2 Local recall

Local recall considers each image in the evaluation data as a separate word recall problem. We define the local target set as the union of the descriptions (sets of words, D) for an image I_i (Eq. 5). The goal is to recall the content words that are important to describe the image. We used SpaCy 2.0.4 to tag the descriptions and we only use adjectives, verbs, nouns, and adverbs as content words for the analysis.

Recalled words are those that are generated for a specific image I_i and occur in the local target set (Eq. 6). We define the importance of a word w for an image I in terms of the number of descriptions D that the word w occurs in (Eq. 7), resulting in a value between 1 and N (here N=5, as there are 5 descriptions per image). We use the importance metric to measure how well a system recalls the essential (with a score of 5) or the majority (3 or higher) words.

$$\text{Local}_i = \bigcup_{D_j \in I_i} \{w : w \in D_j\} \tag{5}$$

$$\text{Recalled}_i = \text{Gen}_i \cap \text{Local}_i \tag{6}$$

$$\text{Importance}(w, I) = |\{D : w \in D \wedge D \in I\}| \tag{7}$$

$$\text{Local recall score}_k = \frac{1}{|\text{Val}|} \sum_{I \in \text{Val}} \frac{|\{w : w \in \text{Recalled}_i \wedge \text{Importance}(w, I_i) = k\}|}{|\{w : w \in \text{Local}_i \wedge \text{Importance}(w, I_i) = k\}|} \tag{8}$$

The *local recall score* for words of *k* importance is computed by dividing the total number of recalled words with an importance of *k* by the total number of words with an importance of *k* (Eq. 8). Figure 5 shows the scores for all 9 systems. All models achieve local recall scores between 65% and 80% for types that are mentioned in all five references. This time, the GAN-based models do not outperform the rest, although they still have recalls around 75%. Although local recall is not strictly about diversity in output vocabulary, it does test each system's ability to use the right words at the right time (even if those words are rare).

Figure 6 shows the correlations between coverage (Eq. 3) and the local recall metric with the existing measures of diversity that were discussed earlier. We find that coverage and the number of types are perfectly correlated. Future work may find that these two measures do not always correlate perfectly, since coverage is based on the word types in the evaluation set. If future systems start producing more word types that are not in the evaluation set, we would see a divergence between coverage and number of types. Local recall (Loc_5 in the table), does not correlate as strongly with the other metrics.

3.3 Global ranking of omitted words

Instead of using local and global recall to produce scores summarizing model performance, we can use these metrics to construct a ranking of words typically produced by a model, or that a model typically fails to produce. We refer to ranking on the basis of global recall as *global ranking*. The most straightforward way to use global ranking is to construct a frequency table for all words in the evaluation set that are not recalled by a model. This gives us a list of the *most common omissions* for that model. Table 2 presents the 15 most frequent words that *all* systems failed to produce. The first ranking is based on the frequency in the training set; the second ranking on the basis of the validation set frequency. The advantage of the former is that we see which words are omitted even though there is sufficient evidence. The advantage of the latter is that we see which words are omitted, even though there are sufficient contexts in which those words could have been used.

Two types that immediately stand out are *'s* and *n't*. One possible reason that both these types were never produced by any system is that they are (cognitively) complex. The possessive *'s* indicates abstract

Global ranking	
Train	Validation
's	's
elderly	elderly
toast	toast
we	we
whole	thrown
laughing	whole
displays	ham
meadow	located
located	driver
ham	mat
nicely	n't
n't	heading
almost	displays
more	amongst
picking	simple

Local ranking		
Absolute	Relative	Relative$_{10}$
man	pillow	door
dog	kitten	paper
woman	flag	van
people	turkey	pink
cat	milk	head
umbrella	ice	doll
dogs	chips	hair
sign	rainbow	pool
pizza	potatoes	fork
ball	map	tray
cake	eggs	carrot
bear	cream	girls
bed	butter	apple
table	strawberries	women
elephant	pregnant	rice

Table 2: Global and local rankings of omitted words. These rankings show the most frequent words that *are not* produced at all (Global ranking), or that are most commonly omitted by the 9 image description systems (Local ranking).

relations between animate entities and objects that vary from scene to scene, making it difficult to learn how to use this type on the basis of visual information alone. The use of negations like *n't* typically requires the speaker to reason about whether or not an image conforms with their expectations (van Miltenburg et al., 2016). Another difficult case is *thrown*, which refers to a throwing action taking place before the picture was taken. Completing the top-3 in both rankings are *elderly* (253 occurrences in the training data, 140 in the validation data) and *toast* (237 and 124). These are less complex than the examples mentioned above, and could be determined on the basis of visual information alone. Further research is needed to determine why these words could not be produced by any system.

3.4 Local ranking of omitted words

We refer to rankings produced on the basis of local recall as *local ranking*. With local ranking, we can look at the words that models failed to produce most often. For reasons of space, we will only look at the words with importance class $k = 5$. Table 2 presents three local rankings:

1. An *absolute* ranking, where we look at the aggregate number of times each word was missed by the models under investigation.
2. A *relative* ranking, where we look at the rate at which each word was missed (Eq. 9). In the case of a tie, the most frequent word 'wins', so that words with the largest impact on model performance are ranked higher.

$$\text{MissRatio}(w) = \frac{\text{missed}(w)}{\text{missed}(w) + \text{recalled}(w)} \tag{9}$$

3. A relative ranking with an *occurrence threshold*, where each word with importance class $k = 5$ has to occur at least $n = 10$ times for each system. This eliminates words from the ranking that occur only a few times, but that are missed by all systems (so MissRatio(w) = 1).

All three rankings provide a starting point to explore system performance. For example, in the first ranking, we observe that some of the most common terms in the MS COCO dataset overall (*man* and *woman*) are often missed by image description systems, when all annotators *do* use those terms. Since these words are ranked high, they have a big impact on the quality of the descriptions. For reasons of space, we cannot discuss this example in depth, but a natural next step would be to look at example descriptions where systems fail to produce *man* or *woman* and identify potential causes of this behavior (e.g. an inability to determine people's gender using only visual information).

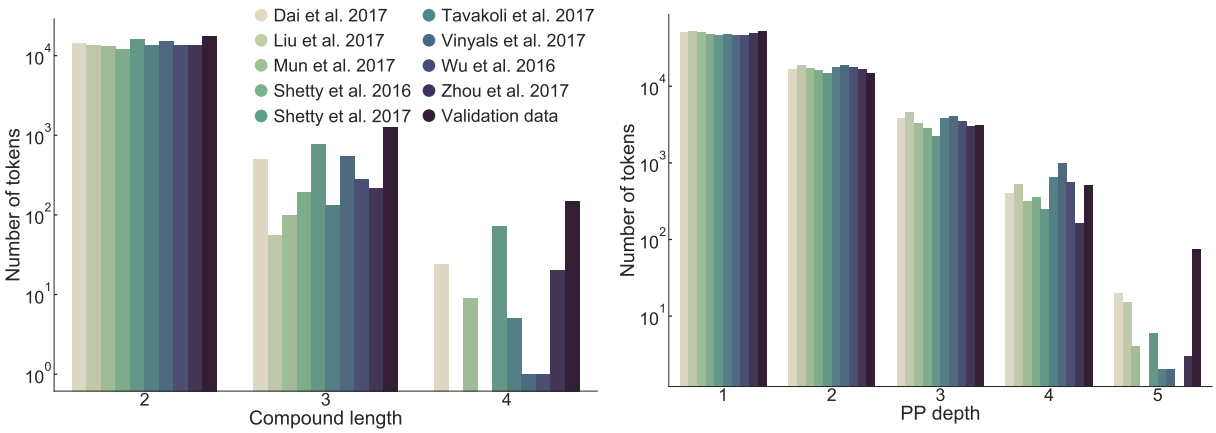

Figure 7: Histograms showing the number of tokens with compound length 2, 3, and 4, and the number of tokens with PP-depth 1–5, for all 9 systems and the MS COCO validation data. The validation data and the two GAN-based systems (Dai et al., 2017; Shetty et al., 2017) clearly have more compound tokens than the other systems. We do not observe this difference with PP depth.

4 Compound nouns and prepositional phrases

Beyond the word level, we can look at how words are combined to form new phrases (i.e. *compositionality* (Szabó, 2017)). We detect compound nouns using a part-of-speech tagger (SpaCy 2.0.4), assuming that any sequence of nouns is a nominal compound. We also compute the *compound ratio*: the average number of compounds per description. Figure 7 and Table 3 show the results. We observe that the validation data has a larger number of compound nouns, resulting in a higher compound ratio. When we separate the compounds by length, we see that humans produce most compounds in any category, and the GAN-based systems (Dai et al., 2017; Shetty et al., 2017) produce more compounds of length 3 and 4 than the other systems. The system by Vinyals et al. (2017) also stands out in this regard. Finally, we see that the GAN-based systems produce more compound types of length 2 than any other system, but there is still a big gap between the GAN-based systems and human performance.

We detect prepositional phrases (PPs), such as *in the kitchen*, using SpaCy's part-of-speech tagger and dependency parser. First, we identify each preposition in the description (e.g. *in, with, on*). Then we inspect the subtree headed by those prepositions. For each of those subtrees, we count their depth in terms of PP embeddings, e.g. *on top of a pan on a table* (1) has a depth of 3.

(1) [**on** top [**of** a pan [**on** a table]]]

We also compute the *preposition ratio*, which is the average number of prepositions per description. Table 3 shows the results. We do not see a big difference between the validation data and the systems. The only difference is that humans produce more types of PPs with depth 1: twice as many as the System by Dai et al. (2017). We conclude that image description systems still have much to gain in terms of compositionality. For further discussion of this topic, also see recent work by Lake and Baroni (2017).

	Compound stats		PP stats	
	Ratio	Types-2	Ratio	Types-1
Liu et al. 2017	0.33	122	1.86	1145
Mun et al. 2017	0.33	300	1.74	2423
Shetty et al. 2016	0.30	319	1.65	2426
Tavakoli et al. 2017	0.33	259	1.72	1888
Vinyals et al. 2017	0.39	275	1.74	1678
Wu et al. 2016	0.34	237	1.69	1732
Zhou et al. 2017	0.34	472	1.71	3451
Dai et al. 2017	0.37	2576	1.78	11709
Shetty et al. 2017	0.42	1446	1.58	8439
Validation data	0.47	6089	1.74	22237

Table 3: Statistics for nominal compounds and prepositional phrases. Compound ratio corresponds to the number of compounds per description. Types-2 refers to the number of compound types of length 2. Preposition ratio corresponds to the number of prepositional phrases per description. Types-1 refers to the number of PP types of depth 1.

5 Discussion and Future Research

5.1 Other metrics

In addition to the metrics proposed in this paper, there are other options that could be explored in future work. For reasons of space, we were also not able to cover metrics based on the **frequency distribution** of words in the training and validation data. We already mentioned Shetty et al.'s (2017) use of frequency ratios in the introduction. Their approach could be extended (perhaps also using log-likelihood (Rayson and Garside, 2000)) to produce a ranking of words that are over- or underused by a particular system. Overused words could be further analyzed by computing a '**local precision**' metric, measuring how often a generated word is also used in at least one reference description. Ferraro et al. (2015) present other metrics in their survey of datasets for vision and language research, including:

Yngve and Frazier measurements of syntactic complexity (Yngve, 1960; Frazier, 1985). Ferraro et al. (2015) found that the MS COCO and Flickr30K datasets have the most complex sentences, compared to other vision & language datasets. It is still an open question whether machine-generated descriptions are of equal complexity and, if not, what are the differences.

Abstract-to-concrete ratio The authors also compare the proportion of abstract words that each corpus contains. They count abstract words by using a list of abstract words compiled in earlier work. In the literature, there are two definitions of abstractness and concreteness. Concrete words are either said to be (1) more closely tied to perception, or (2) more specific (Spreen and Schulz, 1966; Theijssen et al., 2011). It is unclear which is meant by Ferraro et al., but it would be interesting to see whether machine-generated descriptions are more closely tied to perception than human descriptions, who also speculate about the context of the images (van Miltenburg, 2016).

Part-of-speech distribution Ferraro et al. (2015) compared the distribution of *nouns, verbs, adjectives*, and *other* parts of speech. Our work on detecting prepositional phrases and compound nouns (Section 4) suggests that differences in the distribution of parts of speech between human- and machine-generated descriptions could be an interesting avenue to explore.

Besides the measures discussed above, it may also be interesting to study some types of linguistic phenomena in more detail. For example, van Miltenburg et al. (2016) provide a thorough overview of the uses of *negations* in human-generated image descriptions. Even though this is a low-frequent (or long-tail) phenomenon, studying a subset of the image descriptions informs us about the human image description process, and the *cognitive requirements* to produce a description containing a negation. It remains to be seen whether image description systems could produce similar descriptions.

5.2 Limitations and human validation

Earlier work has shown that automated evaluation metrics do not correlate well with human judgments (Elliott and Keller, 2014; Kilickaya et al., 2017). For this reason, we should not blindly trust evaluation metrics in their assessment of system performance. Still, this paper only includes automatic, intrinsic metrics. This is by design: we want to gain insight into the descriptions, not to evaluate their quality.

While you cannot evaluate a system using only automated metrics, they do tell us something about how a system behaves. Future researchers could try to improve the diversity metrics while maintaining or improving the quality of the descriptions (ideally measured by human judgments). At that point, we should determine if more diverse descriptions (as measured by the metrics covered in this paper) are perceived by humans as more interesting to read. One issue is that it is unclear how human judgments could be used to rate the diversity of the generated descriptions, because diversity is a *global* property of the data. In other words: you cannot judge the diversity of a single description, because that is not what diversity is about. You can only judge the diversity of a larger collection of descriptions. One way to do this might be to generate descriptions for sets of very similar images, and have participants rate the diversity of different batches of descriptions.

6 Conclusion

We explored several metrics to analyze the richness of computer-generated image descriptions, most of which focus on diversity at the word level. In our analysis of the output of nine state-of-the-art systems, we found that there are clear differences between human and system output: humans produce more word types; more different types when averaged over multiple 1000-token samples; more compound nouns per description; more long compound nouns; and more compound noun types than image description systems. Not all of these observations hold for prepositional phrases: humans *don't* produce more prepositional phrases per description, and neither do they produce more embedded prepositional phrases, however, they *do* produce a larger number of different prepositional phrases than the systems. At the sentence level, we found that humans produce longer descriptions, vary more in their description length, and produce more novel descriptions. We also found that GAN-based systems produce more diverse descriptions than MLE-based systems. However, we caution that the GAN-based systems are the only ones in our evaluation that are designed with diversity in mind. Further research is needed to find out what kind of approach is best for producing diverse descriptions.

We also proposed to frame image description as a word recall task to further explore the differences highlighted above. *Global recall* looks at the types from all the validation data that are learnable from the training data. *Local recall* measures whether systems are able to produce content words that are mentioned in n reference descriptions for a single image. These metrics show that there is plenty of room for improvement, both in terms of vocabulary size, as well as using the right words at the right time. One way to approach this challenge is by *ranking* terms that are often missed by a system, and looking for ways to learn when to use these words.

We provide all the code and data to to apply the metrics discussed in this paper and compare systems. We encourage readers to use this overview to start exploring the output of their own image description systems, but note that the metrics covered here are just the tip of the iceberg. As more researchers focus on producing more diverse descriptions, we will hopefully also develop a better understanding of what makes a description human-like. Formalizing these notions enables us to measure our progress towards richer and more diverse descriptions.

Acknowledgements

This research is funded through the 2013 NWO Spinoza prize, awarded to Piek Vossen. Desmond Elliott is supported by an Amazon Research Award.

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany.

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. 2015. Zipf's law for word frequencies: word forms versus lemmas in long texts. *PloS one*, 10(7):e0129031.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the 2017 International Conference on Computer Vision*, pages 2970–2979, Venice, Italy.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd*

Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 100–105, Beijing, China.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June. Association for Computational Linguistics.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *EMNLP*, pages 207–213, Lisbon, Portugal.

Lyn Frazier. 1985. Syntactic complexity. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 129–189. Cambridge University Press, Cambridge.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Montreal, Canada.

Wendell Johnson. 1944. I. a program of research. *Psychological Monographs*, 56(2):1.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain, April. Association for Computational Linguistics.

Brenden M Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *arXiv preprint arXiv:1711.00350*. Under review at ICLR 2018.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL:HLT*, pages 110–119, San Diego, California, June. ACL.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Chang Liu, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. 2017. Mat: A multimodal attentive translator for image captioning. In *IJCAI*, pages 4033–4039.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv:1411.1784*.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *AAAI Conference on Artificial Intelligence*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016. Addressing the mfs bias in wsd systems. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, pages 1–6.

Rakshith Shetty, Hamed R.-Tavakoli, and Jorma Laaksonen. 2016. Exploiting scene context for image captioning. In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*, iV&L-MM '16, pages 1–8, New York, NY, USA. ACM.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*, Oct.

Otfried Spreen and Rudolph W Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5):459–468.

Zoltán Gendler Szabó. 2017. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *CVPR*, pages 2487–2496.

DL Theijssen, H van Halteren, LWJ Boves, and NHJ Oostdijk. 2011. On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal*, 1:61–77, December.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany, August. ACL.

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, April.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *IJCAI*, pages 2957–2964.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Victor H Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

Gilbert Youmans. 1990. Measuring lexical style and competence: The type-token vocabulary curve. *Style*, pages 584–599.

George Kingsley Zipf. 1949. *Human behaviour and the principle of least effort: an introduction to human ecology.* Cambridge, MA: Addison-Wesley.