Cross-validating Image Description Datasets and Evaluation Metrics

Josiah Wang and Robert Gaizauskas
Department of Computer Science
University of Sheffield, UK
{j.k.wang, r.gaizauskas}@sheffield.ac.uk

Abstract

The task of automatically generating sentential descriptions of image content has become increasingly popular in recent years, resulting in the development of large-scale image description datasets and the proposal of various metrics for evaluating image description generation systems. However, not much work has been done to analyse and understand both datasets and the metrics. In this paper, we propose using a leave-one-out cross validation (LOOCV) process as a means to analyse multiply annotated, human-authored image description datasets and the various evaluation metrics, i.e. evaluating one image description against other human-authored descriptions of the same image. Such an evaluation process affords various insights into the image description datasets and evaluation metrics, such as the variations of image descriptions within and across datasets and also what the metrics capture. We compute and analyse (i) human upper-bound performance; (ii) ranked correlation between metric pairs across datasets; (iii) lower-bound performance by comparing a set of descriptions describing one image to another sentence *not* describing that image. Interesting observations are made about the evaluation metrics and image description datasets, and we conclude that such cross-validation methods are extremely useful for assessing and gaining insights into image description datasets and evaluation metrics for image descriptions.

Keywords: Image Description, Evaluation Metrics, Vision and Language

1. Introduction

The ability to describe the content of an image is essential for various tasks such as image indexing and retrieval, and organising or browsing large image collections. With advances in visual object category recognition (e.g. being able to recognise and localise a *car* in an image), recent years have seen an upsurge of interest in moving beyond annotating images with isolated keywords to automatically generating sentence-level, natural language descriptions of image content (*Three boys playing with a ball in the park*).

To facilitate progress and benchmarking on such automatic image description generation tasks, various image datasets with textual descriptions have been developed and various metrics proposed to evaluate such systems. The textual descriptions in such datasets are distinct from generic image captions (e.g. on Flickr or the Web, or in news articles) in that they aim to describe the literal visual content (actors, attributes, activities, scenes, etc.) of the corresponding image (*A black dog jumping to catch a green ball in the field*), and exclude non-literal, subjective opinions or semantic information that requires external knowledge that cannot be determined from viewing the image alone (*My adorable two-year-old poodle named Bobby loves playing with his favourite toy in the sun, pictured here in Hyde Park in London* or *Happiness is a day out with your pet*).

The image datasets provide textual descriptions written by multiple human annotators per image, and are often used during evaluation as gold-standard *reference* descriptions against a system generated *candidate* description. However, little work has been done to analyse or evaluate the gold-standard descriptions against themselves, i.e. evaluating a human-authored description against other human-authored descriptions of the same image. Besides answering the obvious question of what the human upper-bound in the image description generation task is, performing such an evaluation can also provide further insights into (i) the **evaluation metrics** used for evaluation; (ii) the similarities and differences within and across a variety of **datasets**.

In this paper, we carry out this analysis. Throughout the paper, we use the term **leave-one-out cross-validation (LOOCV)** to refer to this method of evaluation[1]. Evaluation is performed on a variety of image description datasets that are currently available, using several commonly-used metrics for image description generation tasks. More specifically, we compute and analyse (i) human *upper-bounds* for major evaluation metrics on a range of image description datasets; (ii) *lower-bounds* by examining the performance when comparing a set of descriptions describing one image to another sentence *not* describing the particular image; these might describe a different image or could simply be random sentences. The lower-bounds will be useful for investigating how textual descriptions vary within and across datasets (and how well the metrics capture this). We hypothesise that by comparing multiply-annotated, human-authored descriptions in the proposed manner, we can discover subtle differences and biases within and across evaluation metrics and image description datasets. To our knowledge, this is the most extensive evaluation of image description datasets and evaluation metrics carried out to date using LOOCV over gold standard image descriptions.

Overview. The paper is structured as follows. We first present a review and discussion of existing image description datasets (section 2.) and automatic evaluation metrics that have been adopted for the image description generation task (section 3.). In section 4., we present an upper-bound evaluation of how well humans perform in the image description generation task when judged against other

[1] We use the term **cross-validation** loosely: later in the paper we explore replacing the description that has been 'left out' with a sentence not from the original set, stretching what is generally meant by "cross-validation". The term **jackknifing** has also been used previously, but again it too does not exactly fit what we do.

humans, on the different datasets using the various evaluation metrics described. This gives insight into the similarities and differences between the various datasets and metrics. We also compute the ranked correlations between pairs of metrics. Section 5. presents a lower-bound evaluation, i.e. the scores obtained if one evaluates a set of descriptions for one image against an unrelated sentence (descriptions of another image, random words, etc.). This allows us to investigate the variation in human-authored descriptions within and across datasets. Finally, section 6. offers conclusions.

2. Image description datasets

In this section, we provide a review of existing image datasets that are coupled with multiple human-authored descriptions of image content. As mentioned, noisy, large-scale datasets with user-generated captions exist for news images (Berg et al., 2004; Feng and Lapata, 2008) and Flickr (Ordonez et al., 2011; Chen et al., 2015; Thomee et al., 2015). However, in this paper, we are mainly interested in literal descriptions of what is depicted in the image, rather than non-literal or non-visual descriptions that require significant inference from additional knowledge about the image context. As such, we only explore image datasets that are annotated with multiple, sentential descriptions of the visually observable content of the corresponding image. The requirement for multiple descriptions per image also rules out the IAPR TC-12 dataset (Grubinger et al., 2006) which contains only one English description per image.[2] Eight datasets meet the above criteria:

1. UIUC PASCAL Sentence Dataset **(PASCAL1K)** (Farhadi et al., 2010) contains 1,000 real-world images and five crowd-sourced descriptions per image. The images are taken from the PASCAL Visual Object Classes (VOC) 2008 Challenge (Everingham et al., 2015) (which in turn are sourced from Flickr), and are thus biased towards 20 selected object categories (aeroplane, bird, chair, etc.). The descriptions are authored by Amazon's Mechanical Turk (AMT) workers based in the US.

2. The Visual and Linguistic Treebank Dataset (**VLT2K**) (Elliott and Keller, 2013) comprises 2,424 images of various human actions (e.g. person using computer, riding a horse or a bicycle), along with three crowd-sourced descriptions per image. The images are again taken from the PASCAL VOC challenge, specifically the 2011 Action Classification Taster Competition to recognise 10 action classes (jumping, playing instrument, etc.). The descriptions are produced by AMT workers, and are generally made up of two sentences: the first sentence describes the main action in the image ("A band is playing on stage."), and the second covers other background objects ("They are in a white tent."). For this paper, we retain only the *first* sentence of each description.[3]

3. **Abstract Scenes Dataset** (Zitnick and Parikh, 2013) consists of scenes illustrated from clip art and crowd-sourced descriptions for scenes. It is aimed at exploring image description generation without the complexities of visual recognition, as clip art instances can act as gold standard visual annotations. The dataset contains 10,020 images with six AMT crowd-sourced descriptions each (2 sets of 3 descriptions, each description per set describes different aspects of the image). Some of the images are also *semantically* similar since they originate from the same seed description (1,002 seed descriptions used to generate 10 images each), and as such we expect this dataset to contain many semantically similar descriptions.

4. **Flickr30k** (Young et al., 2014), an extension of the Flickr8k (Rashtchian et al., 2010) dataset, contains over 30,000 Flickr images with five AMT crowd-sourced descriptions each. The original Flickr8k dataset is the successor of PASCAL1K ((1) above), and later extended as the Flickr30k dataset. Images are collected directly from Flickr, and depict various actions, events and human activities.

5. **MS COCO** (Microsoft Common Objects in Context) (Lin et al., 2014) contains approximately 80,000 training images and 40,000 validation images with at least five AMT crowd-sourced descriptions per image. Like previous datasets, the images are sourced from Flickr. The emphasis on this dataset is to gather large numbers of images for a small set of 80 categories. As such, images and the descriptions may be biased towards these categories.

6. The **ImageCLEF2015** development set from the Scalable Image Annotation, Localization and Sentence Generation task (Gilbert et al., 2015) of the ImageCLEF2015 (Villegas et al., 2015) challenge consists of 2,000 *web* images with 5 to 51 descriptions per image (with a mean of 9.5 descriptions). The descriptions are crowd-sourced using CrowdFlower. Unlike previous image description datasets, the images are obtained from a large set of generic web pages gathered by the challenge organisers, making it a highly varied dataset, albeit still constrained by the 251 object categories defined for the challenge.

7. **Pascal50S** (Vedantam et al., 2015) is an extension of Pascal1K ((1) above) with 50 descriptions per image. This dataset is used primarily for improving the reliability of evaluation given the significantly larger number of reference image descriptions per image.

8. **Abstract50S** (Vedantam et al., 2015) is an extension of the Abstract Scenes Dataset ((3) above), also with 50 descriptions per image.

Parallel to our work, Bernardi et al. (2016) summarised most of these datasets, among others, and reviewed different approaches to image description generation. Ferraro et

[2] While some images have multiple descriptions, they each describe different aspects of the image.

[3] We have experimented using both sentences as the description, and predictably found the overall scores to be low. This is further compounded by the fact that only two reference descriptions remain after leaving one out.

al. (2015) also summarised a subset of these datasets, along with other vision and language related datasets. They also analysed the datasets based on different criteria, such as abstract:concrete word ratios, syntactic complexity and perplexity. Most related to our work is their proposed measure of pairwise perplexity across different datasets to predict the words in a test set given a language model trained on another dataset. Our proposed LOOCV method can also achieve this, albeit in a different manner, and additionally allows us to evaluate the metrics as well as the datasets.

3. Evaluation metrics

Several automatic metrics have been proposed for evaluating image description generation systems. We review and compare an array of evaluation measures (and their variants) that have been proposed or adopted for the task:

1. **BLEU** (Papineni et al., 2002) is a *precision*-based metric adopted from the machine translation community. It measures the number of n-grams in a candidate sentence also appearing in at least one reference sentence, with the count clipped to avoid positive terms being over-repeated in the candidate sentence. BLEU-n is the geometric mean between the modified precision (p_n) for each n, multiplied by the brevity penalty (BP) to penalise short sentences:

$$BLEU_N = BP \times \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big) \quad (1)$$

   where w_n is usually set to $\frac{1}{N}$, and N being the maximum n-gram order of $BLEU_N$.

   While BLEU was originally devised as a corpus-level metric, it has also been used to measure sentence-level performance, with various smoothing techniques proposed to address the issue of n-gram sparseness at sentence level, especially for higher order n-grams (Lin and Och, 2004; Gao and He, 2013; Chen and Cherry, 2014). For this paper, we evaluate sentence-level **BLEU-1**, **BLEU-2**, **BLEU-3** and **BLEU-4** with smoothing.[4] Our implementation is different from the BLEU evaluation used in previous image description generation work; in their case, no smoothing is performed and the brevity penalty is always set to 1. Although the latter may have been useful in earlier, more constrained work ("a cow" is considered a good sentence), recent progress in techniques and the availability of larger datasets have shifted the focus of the task to "describing the image as a human would". As such we argue that precision alone is insufficient, and that recall should now be factored in as part of the evaluation process.

2. **ROUGE** (Lin, 2004) is a *recall*-based metric used to evaluate automatic summarisation systems. In its original formulation, **ROUGE-N** computes the n-gram recall between a candidate summary and a set of reference summaries. Its variants, such as **ROUGE-L** and **ROUGE-S**, are *f-measure*-based metrics. ROUGE-L considers the longest common subsequence between two summaries, while ROUGE-S uses skip-bigram co-occurrences as statistics for measuring the similarity between two summaries, allowing for gaps between the two terms of a bigram. **ROUGE-W** is a variant of ROUGE-L, and awards higher scores to contiguous n-grams over skip-grams. **ROUGE-SU** is an extension of ROUGE-S which also captures unigram co-occurrences in addition to skip bigrams. In this paper, we evaluate – using the official `rouge-1.5.5.pl` script – the following variants: ROUGE-1[5], ROUGE-L, ROUGE-W1.2, and ROUGE-SU4.

3. **Meteor** (Denkowski and Lavie, 2014), again adopted from machine translation, is an *f-measure*-based measure that finds the optimal alignment of chunks of matched text, incorporating semantic knowledge by allowing terms to be matched to stemmed words, synonyms and paraphrases. Content and function word matches can be assigned different weights, and each type of matcher (exact, stemmed, synonym, paraphrase) is also weighted individually. Word ordering is accounted for by encouraging fewer matched chunks, indicating less fragmentation. Meteor matches a candidate text to each reference one-to-one, and takes the *maximum* score out of all references as the final score. We use the official version 1.5 of Meteor for this paper, with the default recommended parameters for English.

4. **CIDEr** (Vedantam et al., 2015) is a measure developed specifically for evaluating image descriptions by consensus. The measure computes the cosine similarity (per n-gram length, n) between a candidate and reference description, each represented as TF-IDF weighted bag of n-grams. The scores are averaged over all reference descriptions belonging to the same image, and further averaged across n. A variant of the measure, **CIDEr-D** has also been proposed to prevent gaming issues with the metric. Note that in the official version of CIDEr/CIDEr-D, the score is arbitrarily multiplied by a factor of 10 so that the scores do not appear too discouragingly low. The theoretical possible range is thus between 0.0 and 10.0,where the scores of state-of-the-art image description generation systems are often between 0.8-1.0[6]. In this paper,

[4] Our implementation of smoothing is based on the official `mteval-v13a.pl` script, which assigns a geometric sequence $1/2^k$ to n-grams with zero matches – please refer to Chen and Cherry (2014) for the algorithm (Smoothing Technique 3). We found an oddity in the script's handling of short hypotheses (length $l < n$), where division-by-zero cases are assigned $\log p_{>l} = 0$ (e.g. a hypothesis of length 2 will result in $p_3 = \frac{0}{0}$ for trigram matches, and $\log p_3$ is set to 0 in this case). Eq (1) will inflate the scores of such cases (because $\exp(0)$=1), making the BLEU-3 score undesirably higher than the BLEU-2 in this example. Our implementation modifies this by setting the denominator of such candidates to 1 and performing smoothing as described.

[5] For ROUGE-1, we consider the f-measure variant with equally weighted precision and recall. We have tested the recall version of ROUGE-1 and found the scores to be just slightly higher than the f-measure. The general trend however is similar.

[6] https://competitions.codalab.org/competitions/3221

we report the *raw* CIDEr/CIDEr-D scores before this multiplication process. We also compute IDF scores independently per dataset.[7]

Elliott and Keller (2014) evaluated how well the first three metrics correspond to human judgements. This is done by asking human annotators to score the output sentences from one of the systems of Hodosh et al. (2013) on a scale of 1-4, and computing the correlation between the human scores and the scores from the same system for each metric. They found that Meteor correlates best with human judgements, followed by ROUGE-SU4 and BLEU-4 (with smoothing). Vedantam et al. (2015) also compared the four metrics in terms of how well they correlate with human judgement on a *consensus* task, i.e. which is more similar to sentence A? Sentence B or sentence C? They found CIDEr captured human consensus best.

Other metrics have also been proposed to evaluate the content of image descriptions, for example using semantic tuples (Ellebracht et al., 2015) and concentrating only on the content selection phase (Wang and Gaizauskas, 2015). These however require additional annotations.

4. Human upper-bound evaluation

Given the multiply annotated datasets, we first evaluate human performance on each using LOOCV, i.e. by withholding one human-authored image description as a candidate description, and evaluating it against all remaining descriptions for the same image, repeating the process in turn for each description of the image. The final score is produced either by micro-averaging the scores across all descriptions in the dataset or by macro-averaging the scores across all images (average scores per image, and average the mean scores). Other statistics such as standard deviation, median, minimum and maximum scores are also computed. We evaluate the images descriptions for eight datasets (section 2.), using the various metrics described in section 3. In section 4.2., we further measure the ranking correlation between metrics. As it is impractical to present all these numbers in this paper, we provide all computed statistics online[8], and instead concentrate here on highlighting and discussing interesting observations.

As a preprocessing step, all image descriptions are stripped of punctuation, case-normalised and tokenised (words separated by hyphens are always tokenised).

4.1. Human upper-bound results

We discuss the human upper-bound evaluation results by metric, interleaved with dataset-specific observations where relevant.

[7]We have experimented concatenating all datasets to compute a common DF statistic, which results in significant bias towards the two large datasets (MS COCO and Flickr30k). We did not observe any notable differences except for the Abstract Scenes Dataset and Abstract50S, which ended up with much higher CIDEr scores. We presume that this is mainly due to the words 'mike' and 'jenny', common in these datasets, being assigned inflated IDF weights as they do not occur often (if ever) in the dominant MS COCO and Flickr30k datasets.

[8]http://visualsense.github.io/loocv/

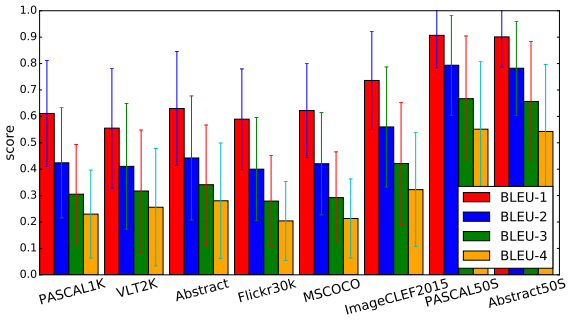

Figure 1: Upper-bound, micro-averaged BLEU scores for the eight datasets (best viewed in colour).

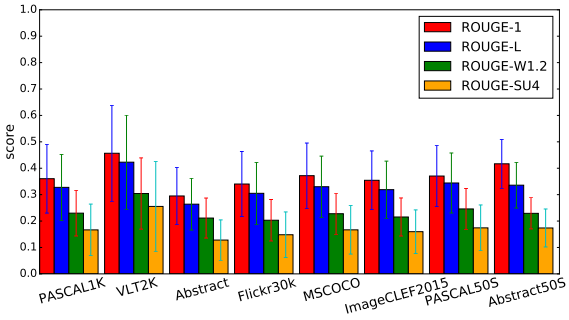

Figure 2: Upper-bound, micro-averaged ROUGE scores for the eight datasets.

BLEU. Figure 1 shows the absolute BLEU scores for the dataset. We found the overall BLEU-1 scores to be high (0.56-0.91), and as expected are lower for increased n-gram length. We also noticed that BLEU is sensitive to the number of references per image. Datasets with many descriptions per image (ImageCLEF2015, PASCAL50S, Abstract50S) produced higher scores, while VLT2K (two references per image) yielded lower scores. In fact, the median BLEU score for PASCAL50S is actually 1.00. The reason for this observation is that BLEU measures the overlap between a candidate and the *union* of n-grams in the corresponding references, thus increasing the number of references increases the chances of overlap. We further verified this by repeating the experiments, but sampling only 5 descriptions per image for ImageCLEF2015, PASCAL50S and Abstract50S. The BLEU scores on these reduced datasets are now in the same range as the other datasets. As such, we do not recommend BLEU for datasets with too many reference descriptions because of the higher likelihood of spurious matching.

Another noteworthy point about all the datasets is that the maximum BLEU scores (all variants, micro-averaged) are 1.00 across all datasets. This means that for each dataset, there is at least one description that has all its 4-grams matched with the union of all 4-grams in the corresponding references.

ROUGE. Like BLEU, the unigram-based ROUGE-1 resulted in higher absolute scores than its skip-gram counterparts (Figure 2). ROUGE-W1.2's absolute scores are lower than ROUGE-L as the measure penalises non-contiguous

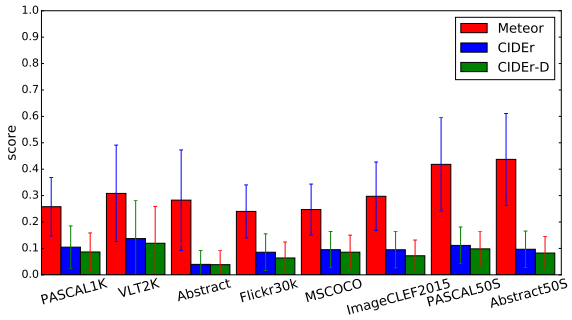

Figure 3: Upper-bound, micro-averaged Meteor, CIDEr and CIDEr-D scores for the eight datasets.

common subsequences. In contrast to BLEU, ROUGE is not sensitive to number of descriptions per image, as it performs averaging over all reference descriptions. This can be seen from the more uniform scores across datasets in Figure 2. With the scores being less dependent on reference size, we observe several outliers. VLT2K yields higher ROUGE scores than other datasets, most likely because the first sentence describes the main action, making the problem more constrained, and has better a chance of word or skip-gram overlaps. In fact, there are images with all three reference sentences being identical (six such cases). The standard deviation on this dataset, however, is also higher. This suggests that there are also many main actions in the images that can be described differently by different annotators. The other outlier is the Abstract dataset, with the lowest ROUGE-1, ROUGE-L and ROUGE-SU4 scores. This confirms the fact that the agreement between descriptions of the same image is quite low, i.e. many descriptions of the same image describe something different about the image. This can also be confirmed by comparing Abstract against Abstract50S, where the score range of Abstract50S is more similar to that of other datasets as it does not have this constraint of defining different aspects of the same image. Therefore, ROUGE manages to highlight this fact by averaging recall across *all* reference descriptions, compared to BLEU which computing precision against matching segments in *any* reference descriptions.

Meteor. We can observe from Figure 3 that Meteor is also sensitive to the number of reference descriptions per image, albeit not to the extent of BLEU. This is because Meteor takes the best matching reference (maximum score) as the final score, thus increasing the likelihood for a good match with more references. Contrast this to BLEU which sums co-occurrences of each n-gram from *any* reference (i.e. matching n-grams do not have to be in the same single reference 'robI've changed the wording here to make it clearer – is this correct?), which explains why BLEU is even more dependent on the number of reference descriptions. Again, like BLEU, the maximum Meteor score across all datasets is 1.0 (exact match). This suggests that there is at least one image with at least one pair of identical descriptions; we have manually verified this fact. Thus, Meteor is the only metric covered in this paper that can capture such cases. Again, VLT2K shows high scores despite having only two reference sentences. Interestingly, Abstract actually showed the highest standard deviation with Meteor compared to other datasets (and other metrics). We attempt to draw some inferences from the second order macro-average statistics, more specifically by computing the standard deviation across descriptions per image, and computing the mean and standard deviation of these standard deviations. The value for these are quite high, demonstrating that (i) there is on average high variation within descriptions of the same image; (ii) some descriptions of the same image have similar pairs, while others are quite dissimilar (the standard deviation themselves vary).

CIDEr. Compared to other metrics, the absolute *raw* CIDEr scores are much lower. This is due to CIDEr being designed to evaluate descriptions by consensus, i.e. a good candidate should agree with the majority of reference descriptions. To achieve a CIDEr score of 1.00 requires the candidate description and *all* reference descriptions to be exact matches. Our informal experiments showed that changing just a few words to *one* reference description can greatly affect the score. CIDEr is, however, very intuitive and reinforces our previous observations. The Abstract Scenes Dataset achieved extremely low CIDEr scores ($<$0.04) compared to other datasets ($\approx$0.09-0.11), showing low consensus among the descriptions. Again, VLT2K showed high consensus (0.14), with a high standard deviation (0.13), and a maximum score of 1.00 (all three descriptions are exact matches). CIDEr-D showed similar results.

Other observations. Across all metrics, MS COCO consistently achieved slightly higher scores than Flickr30k, suggesting that MS COCO might be more homogeneous than Flickr30k. PASCAL1K and MS COCO seems to be about equally homogeneous. The homogeneity of ImageCLEF2015 is hard to assess as it varies across metrics, but in general seems to lie between Flickr30k and MS COCO.

4.2. Ranking correlation between metrics

In this section, we investigate whether different metrics rank datasets similarly. This gives insight into which metric pairs correlate better with each other. Two setting are explored: (i) ranking per image (average the upper-bound scores across all descriptions for the same image); (ii) ranking per description (use the upper-bound score for each description directly, independent of the image). Kendall's τ ranking coefficient is used to measure the monotonic correlation between metric pairs. This measure is more intuitive and less sensitive to outliers compared to the more commonly used Spearman's ρ.[9] We used Kendall's τ-b variant which makes adjustments for ties. The p-values (two-tailed) are found to be extremely small (the largest being 9.9E-40). Again, we concentrate on highlighting interesting facts – detailed numbers are provided online.

Per image ranking correlation. All metric pairs across all datasets are found to be positively correlated (τ between 0.37-0.97), indicating that there is a monotonic relationship between the rankings of each metric pair. First, looking at the coefficient values among BLEU variants,

[9] We have computed Spearman's ρ and found the overall trend to be similar to Kendall's τ, but with higher values.

BLEU-2 and BLEU-3 show very strong correlation (0.83-0.88), BLEU-3 and BLEU-4 show even stronger correlations (0.89-0.93). BLEU-2 and BLEU-4 show strong correlation (0.74-0.82), but not as strong as BLEU-2/BLEU-4 and BLEU-3/BLEU-4. BLEU-1 shows much weaker correlation: 0.67-0.79 against BLEU-2, 0.57-0.70 (BLEU-3), and 0.51-0.66 (BLEU-4). BLEU-1 also seems to show much weaker correlation when compared against all other metrics (0.37-0.70), especially on the datasets with many descriptions (0.38-0.49 against PASCAL50S/Abstract50S). All ROUGE variants strongly correlate with each other, especially between ROUGE-L and ROUGE-W (0.87-0.93). All BLEU variants almost always correlate better with ROUGE-SU4 than with the other ROUGE variants. ROUGE also shows weak correlation with Meteor and CIDEr (0.39-0.68). ROUGE generally correlates slightly better with Meteor than with CIDEr (especially for Flickr30k, VLT2K and MS COCO), although the correlation is comparable for PASCAL50S.
The correlation between CIDEr and CIDEr-D is extremely strong (0.80-0.97), especially for Abstract, MS COCO, Abstract50S and PASCAL50S (0.90 or above). Meteor and CIDEr/CIDEr-D are moderately correlated (0.55-0.68).

Per description ranking correlation. Compared to per image ranking, the overall correlation for per description ranking is predictably slightly lower, except for Abstract which varies depending on the metric. The datasets with many descriptions show a larger drop in scores, because there can be more disagreements in rankings with a larger set of descriptions. There is quite a large drop in correlation scores for Meteor compared to per image ranking. This is likely because of the max function used by Meteor, making the variation in rankings larger; the mean aggregation for per image ranking appears to help alleviate this.

5. Lower-bound evaluation

In this section, we employ LOOCV to compute various lower-bounds for evaluation metrics across different image description datasets. More specifically, we investigate the performance of various metrics when comparing a set of reference descriptions describing one image to a candidate sentence that is *not* from the original set of descriptions for that particular image. The candidate might describe a different image from the same or different dataset, or could simply be random sentences from a generic corpus or even a random list of words. Lower-bound evaluation by LOOCV will be useful for investigating how image descriptions vary within and across dataset, and how well the metrics capture this.
To explore lower-bound evaluation, we perform LOOCV as in section 4., but substitute the candidate description that has been left out with one of the following candidate 'descriptions':

1. **Random Intra-Dataset:** A random description from a *different* image from the *same* dataset. More specifically, for each candidate description, a description from another image in the dataset is randomly selected. This allows us to investigate how descriptions of different images vary within the same dataset.

2. **Random Inter-Dataset:** A random description from a *different* dataset (selected at random). This allows us to explore how descriptions vary *across* datasets. Also, by comparing this to **Random Intra-Dataset**, we can establish how domain-specific a particular dataset is. To be precise, for each candidate description, one dataset (not the source dataset) is first randomly selected, and a description is randomly chosen from the selected dataset. To avoid bias from highly similar datasets, we rule out using descriptions from Pascal50S as candidates for Pascal1K (and vice versa); and similarly for Abstract and Abstract50S.

3. **Random Brown:** A random sentence from the Brown corpus. This is useful to ensure that the metrics are indeed measuring something more specific to image descriptions (content, structures, style, etc.) We retain only sentences with at least five word tokens and at least ten characters.

4. **Gibberish Dataset:** A randomly generated 'gibberish' n-word sentence, with each word drawn independently from the unigram distribution of the image descriptions of the *same* dataset. This explores how well a metric captures *structure*, as these 'sentences' are not grammatically well-formed. We experiment with $n = 10$ and $n =$ average number of words in the descriptions of the corresponding dataset.

5. **Gibberish Brown:** A randomly generated 'gibberish' n-word sentence, with each word drawn independently from the unigram distribution of the Brown corpus. This is useful to investigate how well a metric captures structure *and* content. We expect the scores for these to be the lowest out of all the substitute candidates above. Again, we experiment with a fixed $n = 10$ and a variable n per dataset (average length of descriptions of the dataset).

5.1. Results for lower-bound evaluation

Overall, the lower-bound scores are significantly lower than the upper-bound scores. As in section 4., we discuss the results by metric and, where relevant, highlight dataset-specific observations, providing detailed numbers online.

BLEU. The most notable observation is that the BLEU scores (all variants) for **Random Brown** are lower than for **Gibberish Brown**. This is mainly because BLEU is a *precision* measure that favours shorter sentences: the sentences in the Brown corpus can be quite long. This also explains why the scores are better for shorter sentences when we vary the n of the **Gibberish** candidate sets (scores are higher for the shorter of n=10 and n=average). Also, BLEU does not seem to be able to capture too well the subtle differences in structure between **Gibberish Dataset** and **Random Intra-Dataset**. This is slightly surprising for the higher order BLEU-4, which is expected to capture some structure; this is most likely because there is minimal overlap between high order n-grams in the first place. Another interesting observation is that for many datasets, **Random Inter-Dataset** actually achieved higher scores than **Random Intra-Dataset**. Upon investigation, this is again found

to be due to BLEU being biased towards shorter descriptions, rather than because of the correctness of the descriptions. However, candidates drawn from a dataset of image descriptions perform better than those from the Brown corpus. As such, we can conclude that BLEU captures word overlap well, although we could not gain much insight about the datasets themselves.

ROUGE. In contrast to BLEU, ROUGE is less sensitive to the candidate sentence length. Although we detected some slight bias towards longer sentences (e.g. **Gibberish** fared better with n=10 than n=6 for Abstract), the metric is balanced by both precision and recall, as evidenced from the comparable scores between **Random Brown** and **Gibberish Brown**. Like BLEU, the ROUGE scores are much higher for candidates drawn from dataset-specific corpora than for those using the Brown corpus, showing the importance of domain-specific tuning for generating image descriptions. **Random Intra-Dataset** yielded much higher ROUGE scores than **Random Inter-Dataset**, something not quite captured by BLEU. This suggests that besides domain-specific tuning, dataset-specific tuning is also important for these datasets, especially Abstract (ROUGE-1 score of 0.23 for Intra-Dataset vs. 0.09 for Inter-Dataset) which contains many dataset-specific vocabularies and structure/style not present in other datasets. We also found VLT2K to be quite dataset-specific (0.23 vs. 0.15), while ImageCLEF2015/MS COCO (0.17 vs. 0.14) are less so and PASCAL1K/PASCAL50S/Flickr30k even less again (less than 0.02 difference in score).

Another interesting observation is how the scores of **Random Intra-Dataset** for ROUGE-1, ROUGE-L and ROUGE-W1.2 are much closer to their upper-bound scores (ratio is about 1:2). This either shows that the datasets are pretty homogeneous, or that the ROUGE measure overrewards irrelevant terms. The Abstract Scenes Dataset, with a ROUGE-1 score of 0.23 (compared to its upper-bound of 0.30), and Abstract50S (0.28 vs. 0.40) show the smallest difference between the upper-bound and the lower-bound **Random Intra-Dataset** scores. The small difference in score suggests that the datasets are quite homogeneous – recall that the dataset contains images that are semantically similar. Thus, ROUGE successfully picks out this fact. The distances between the scores are much larger for ROUGE-SU4 (ratio about 1:3), although it still picks out the homogeneity of the two Abstract datasets very well. For this reason, ROUGE-SU4 may be the best measure for evaluating image descriptions among the ROUGE variants.

Meteor. The lower-bounds for Meteor also echo the same observations from ROUGE: (i) using dataset-specific corpora gives higher scores than using the Brown corpus; (ii) structured sentences are better than 'gibberish'. Like ROUGE, Meteor also captures dataset specificity, but interestingly shows an even larger difference in the Intra-Dataset vs Inter-Dataset scores for Abstract/Abstract50S (0.17/0.23 vs. 0.07/0.08 respectively) compared to the ratio of other datasets, and also showed that Flickr30k is more dataset-specific than previously suggested by ROUGE. Like ROUGE-SU4, Meteor also captures the homogeneity of the datasets (ratio is about 1:3 to 1:4, except the Abstract datasets with a ratio of about 1:1.6-1:1.9). In this case, we find that Meteor might be an even better measure than ROUGE-SU4, or otherwise comparable.

CIDEr. Finally, CIDEr/CIDEr-D also demonstrates the same observations as ROUGE, giving higher scores to candidates from image description specific domains than from the Brown corpus. It also captures the fact that the two Abstract datasets are much more dataset-specific than the other datasets. What is most striking is that the differences in scores between the upper-bounds and **Random Intra-Dataset** lower-bounds are much larger compared to ROUGE and Meteor (e.g. 0.1372 vs. 0.0095 for VLT2K, 0.0949 vs. 0.0029 for ImageCLEF2015, and 0.0384 vs. 0.0066 for Abstract). Thus it captures the homogeneity of datasets while still keeping the lower-bound/upper-bound score differences large. We have not ascertained why, but we assume that this is because of the way CIDEr tries to ensure that the candidate must be similar to all reference descriptions (and by extension the majority). This is a very interesting observation, which makes CIDEr an extremely valuable metric for evaluating image descriptions.

6. Conclusion

We proposed using leave-one-out cross validation (LOOCV) to analyse and gain insights into multiply annotated, human-authored, image description datasets as well as commonly used metrics for evaluating image descriptions. We estimated a human upper-bound performance on the task with regards to each evaluation metric, for each dataset. The upper-bound performance scores provided us insights into both the datasets and the metrics. For example, BLEU is sensitive to the number of reference descriptions, Meteor was useful for discovering the fact that there is at least one image with at least one pair of identical descriptions, and ROUGE and CIDEr were useful for measuring agreement/consensus among reference descriptions. We also ranked the image descriptions per dataset by their scores, and analysed the rank correlation between difference metrics. We found BLEU to be weakly correlated with other metrics, and that ROUGE shows better correlation than BLEU to Meteor and CIDEr, which in turn are moderately correlated. We further estimated various lower-bounds for the evaluation metrics, again using LOOCV, to investigate how image descriptions vary within and across datasets and how well the evaluation metrics capture this. From the results, we concluded that learning to generate descriptions from image description specific datasets does yield better performance than from a generic corpus, and that using dataset-specific image descriptions further improves results, even more so with datasets like the Abstract Scenes Dataset, which has a unique vocabulary set and language structure.

Future work should consider characterising image description datasets and discovering which components of image descriptions matter more to each evaluation metric.

7. Acknowledgements

The authors acknowledge funding from the EU CHIST-ERA D2K 2011 Visual Sense (ViSen) project – UK EPSRC grant reference: EP/K019082/1.

8. Bibliographical References

Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. A. (2004). Who's in the picture? In *Neural Information Processing Systems*, pages 137–144.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 362–367.

Chen, J., Kuznetsova, P., Warren, D., and Choi, Y. (2015). Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proc. NAACL-HLT*, pages 504–514.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proc. of the EACL 2014 Workshop on Statistical Machine Translation*.

Ellebracht, L. D., Ramisa, A., Madhyastha, P. S., Cordero-Rama, J., Moreno-Noguer, F., and Quattoni, A. (2015). Semantic tuples for evaluation of image to sentence generation. In *Proc. of the 4th Workshop on Vision and Language*, pages 18–28.

Elliott, D. and Keller, F. (2013). Image description using visual dependency representations. In *Proc. Conf. on Empirical Methods in Natural Language Processing*.

Elliott, D. and Keller, F. (2014). Comparing automatic evaluation measures for image description. In *Proc. Association for Computational Linguistics*.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The Pascal Visual Object Classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every picture tells a story: Generating sentences for images. In *Proc. European Conf. on Computer Vision*.

Feng, Y. and Lapata, M. (2008). Automatic image annotation using auxiliary text information. In *Proc. ACL-08: HLT*, pages 272–280.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (2015). A survey of current datasets for vision and language research. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 207–213.

Gao, J. and He, X. (2013). Training MRF-based phrase translation models using gradient ascent. In *Proc. NAACL-HLT*, pages 450–459.

Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., and Mikolajczyk, K. (2015). Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. In *Language Resources and Evaluation*, pages 13–23.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899.

Lin, C.-Y. and Och, F. J. (2004). ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proc. International Conf. on Computational Linguistics*, pages 501–507.

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *CoRR*, abs/1405.0312.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proc. Association for Computational Linguistics*, pages 311–318.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L. (2015). The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*.

Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A., Bromuri, S., Amin, M. A., Mohammed, M. K., Acar, B., Uskudarli, S., Marvasti, N. B., Aldana, J. F., and del Mar Roldán García, M. (2015). General overview of ImageCLEF at the CLEF 2015 labs. In Josiane Mothe, et al., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 9283 of *Lecture Notes in Computer Science*, pages 444–461. Springer International Publishing.

Wang, J. and Gaizauskas, R. (2015). Generating image descriptions with gold standard visual inputs: Motivation, evaluation and baselines. In *Proc. 15th European Workshop on Natural Language Generation*, pages 117–126.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. of the Association for Computational Linguistics*, 2:67–78, February.

Zitnick, C. L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*.