What’s in a Caption? Dataset-Specific Linguistic Diversity and Its Effect on Visual Description Models and Metrics

David M. Chan1, Austin Myers2, Sudheendra Vijayanarasimhan2, David A. Ross2
Bryan Seybold2, John F. Canny1

1University of California, Berkeley 2Google Research
{davidchan, canny}@berkeley.edu
{aom, svnaras, dross, seybold}@google.com

Abstract

While there have been significant gains in the field of automated video description, the generalization performance of automated description models to novel domains remains a major barrier to using these systems in the real world. Most visual description methods are known to capture and exploit patterns in the training data leading to evaluation metric increases, but what are those patterns? In this work, we examine several popular visual description datasets, and capture, analyze, and understand the dataset-specific linguistic patterns that models exploit but do not generalize to new domains. At the token level, sample level, and dataset level, we find that caption diversity is a major driving factor behind the generation of generic and uninformative captions. We further show that state-of-the-art models even outperform held-out ground truth captions on modern metrics, and that this effect is an artifact of linguistic diversity in datasets. Understanding this linguistic diversity is key to building strong captioning models, we recommend several methods and approaches for maintaining diversity in the collection of new data, and dealing with the consequences of limited diversity when using current models and metrics.

1. Introduction

Automated visual description is an emergent field in computer vision, aiming to generate natural language descriptions of visual information. With various applications including digital accessibility [47] and video summarization [48] as well as indexing and search [24], methods for visual description have the potential to impact the daily lives of billions of users. Recent improvements such as vision and language pre-training [49], compositional and graph methods [9,27], and non-autoregressive training [21] have driven metric performance on standard benchmarks such as MSR-VTT [43] and MS-COCO [20] to new heights.

Unfortunately, despite recent improvements in model architectures [21,27], metrics [15,41], and datasets [25,42], automated visual description has been plagued by issues of poor generalization and description quality [1,33,36,45]. Models consistently perform poorly on novel data, generate nonsense descriptions, or produce descriptions that are too vague to be of use to visually impaired users [22]. It remains an open question in visual description to understand the source of these generalization issues.

This paper is motivated by both the fact that often state-of-the-art methods outperform leave-one-out experiments...
with ground truth sample data (explored in 3) as well as results demonstrating poor cross-dataset generalization in video captioning from Smeaton et al. [33] and Yang et al. [46]. We find that one major issue in current datasets—description linguistic diversity—explains a great deal about model evaluations. Our work, consisting of analyses on several popular visual description datasets, contains several primary contributions:

1. We demonstrate that a lack of linguistic diversity at a token and n-gram level can bias models to generate descriptions lacking in semantic detail (section 4).

2. We show that diversity among ground truths for a single visual context presents a catch-22: low within-sample linguistic diversity leads to generic captions, as information is repetitive; on the other hand, high within-sample diversity leads to a breakdown of single-sample metrics, causing inconsistencies in model evaluation and inaccurate understanding of model performance (section 5).

3. We detail how a lack of semantic diversity at the dataset level can encourage models to generate generic descriptions through classification, instead of learning to understand and relay visual phenomena at various levels of detail (section 6).

4. We discuss our findings demonstrating the need for future research in visual description datasets, methods, and metrics, present recommendations on possible solutions to current linguistic diversity, and introduce a new toolkit for dataset evaluation and split generation focused on linguistic diversity (section 7).

2. Experimental Design

In this work, we explore the field of visual description data through the lens of some of the most popular visual description datasets. While there are a large number of visual description datasets to choose from, we decided to focus on some of the most common datasets for video description, and an additional dataset for image description: 1 MSR-VTT [43], VATEX [42], MSVD [8] and MS-COCO [20] (for full details, see the supplementary materials).

All of these datasets collect multiple ground truth descriptions per visual context, and the ground truth descriptions that they do collect are generated by human annotators (via Amazon Mechanical Turk for these datasets). Unfortunately, very large benchmark datasets such as Conceptual Captions [30] and HowTo-100M [24] often contain only a single description per image/video, of questionable quality as the datasets are not annotated by hand. While datasets like S-MiT [25] contain human-annotated ground truths, they post-process spoken language with automated speech recognition tools, making the dataset difficult to analyze from an n-gram metric angle. Both ActivityNet Captions [17] and YouCook [54] are dense video description datasets that contain high-quality descriptions, however only contain a single ground truth per video.

Given the datasets, we will contextualize our experiments through the lens of several standard metrics for visual description. The BLEU (or BLEU@N) [26] score is a measure of n-gram precision, the ROUGE-L [19] score is a measure of longest common sub-sequence recall, the METEOR [3] score is a F1-oriented alignment-based metric, and the CIDEr [40] score is a TF-IDF weighted similarity metric. For more details of the individual metrics, see Aafaq et al. [1]. Recently, metrics which focus more on including visual content directly such as TIGER [15] and FAIR [41] have shown improvements in human-judgement correlation and scores such as CLIP-score [28], BERT-score [50], and SMURF [13] have been shown to closer approximate semantic content. While improving the metrics is an extremely important area of research, we also believe that analyzing both why current metrics are failing and what patterns models exploit to optimize these metrics, can give essential insight into model improvements.

We selected a set of recent works from the field as representing the state of the art. For visual description on MSR-VTT and MSVD, we refer to SemSynAN [27], a recent work that uses semantic embeddings based on POS tagging to achieve strong results. SemSynAN was not evaluated on the VATEX dataset, so for VATEX, we refer to the performance of MGRMP (Motion Guided Region Message Passing) [9], a recent method for visual description which leverages message passing between object regions. For MS-COCO, we refer specifically to Vin-VL [49], a method that uses object-level attention and vision and language pre-training for visual description.

3. How Can Models Outperform Humans?

Recently, there has been a strong contrast between the metrics-based evaluation of methods for generating visual descriptions on data sets and whether those methods generalize to real-world use cases [36]. The goal of our analysis in this paper is to understand some of the core reasons why models are failing to generalize and to make recommendations for the future design of datasets, models, and metrics, in an attempt to avoid further generalization shortcomings. A core indicator of the difficulty of using standard metrics to improve generalization is that the “leave-one-out” performance of the ground truths for each dataset is typically poor. Because we investigate datasets that have more than a single ground truth sample per visual context, we can measure the metric scores between a randomly sampled ground
Within-sample diversity ranges between 11% and 35%, which is the number of unique tokens making up 90% of the tokens in the dataset. In this section, we analyze the diversity of visual descriptions in datasets. In addition to reporting the number of unique tokens, we hypothesize that this metric alone does not tell the full story of the diversity of the dataset. Many datasets use "vocabulary size", the number of unique tokens, as a proxy for the diversity of the dataset, however, we find that alone, the number of unique tokens comprising 90% of the total tokens lacks n-gram diversity, leading to domination of common n-grams. The number of unique n-grams does not allow for strong performance. One of the major issues in overall dataset diversity is a tendency for language models to accentuate a lack of n-gram diversity, leading to domination of common n-grams over visually likely n-grams [14]. A standard metric reported by Wang et al. [42] in VATEX is the number of unique n-grams in the dataset, however, we find that alone, the number of unique n-grams does not allow for strong comparison between datasets, both because the number is not normalized, and the number of n-grams says little about the overall distribution of those n-grams.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Unique</th>
<th>WS-Unique</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVD</td>
<td>9455</td>
<td>11.8%</td>
<td>944</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>22780</td>
<td>21.55%</td>
<td>1636</td>
</tr>
<tr>
<td>VATEX</td>
<td>31364</td>
<td>24.87%</td>
<td>1363</td>
</tr>
<tr>
<td>MS-COCO</td>
<td>35341</td>
<td>33.76%</td>
<td>824</td>
</tr>
</tbody>
</table>

Table 1. Vocabulary metrics for each of the datasets. Unique: The number of unique tokens. WS-Unique: Average percentage of tokens that are unique within a sample (image/video). Head: The number of unique tokens comprising 90% of the total tokens.
Further, this means that the number of captions models will be able to generate is restricted to $V^{ED}$, where $V$ is the size of the vocab, a notably smaller number than expected with large vocab sizes, and long captions. We believe that this is one of the reasons that non-auto-regressive approaches such as those in Liu et al. [21] and Yang et al. [44] are able to perform so well on these datasets: they can focus on the visual information, and don’t have to worry about the syntactic structure as it is similar for all descriptions.

5. Within Sample Diversity

While we have seen that token-level diversity is important for the generation of high quality captions, we also want to understand how within-sample diversity (i.e. diversity within a collection of ground truths for a single visual context) impacts the performance of visual description models. To define how much within sample diversity there is in a dataset, there are several methods that we can use. One metric, common to many papers, is an analysis of how many captions in each sample are novel. VATEX (100%) and MS-COCO (99.9%) have high caption novelty, while MSR-VTT (92.66%) and MSVD (85.3%) contain somewhat less exact novelty. Further, we could look at within-sample token diversity (shown in Table 1), which suggests that within a sample, diversity is actually relatively high, with 11% to 33% of tokens being unique within a sample. Further, the within sample verb (15% to 56%) and noun (13% to 35%) uniqueness is relatively high as well, suggesting that individually, captions discuss unique parts of a visual context (Full results are given in the supplementary materials). This is demonstrated qualitatively in Figure 4.

The issue with these measures of novelty is that they account only for novelty at the caption or token level by exact matching, but do not directly target the semantic novelty of the captions. In order to look closer at within-sample diversity, we compute the pairwise semantic distance between each description and all other unique descriptions in the sample using the cosine distance between MP-Net embeddings [34] trained for sentence similarity. Figure 2 shows the minimum of the inter-sample cosine distances, a metric we call sample redundancy. Notably, almost 10% of the samples in MSVD have a very close semantic match, suggesting that MSVD has more semantically redundant information than other description datasets.

Sample redundancy is both a blessing and a curse. Datasets that have very high sample redundancy will tend to have high performance on leave-one-out ground truth metrics, as most of the ground truth captions will share large amounts of information. This means that pair-wise metrics such as those in Liu et al. [21] and Yang et al. [44] are able to perform so well on these datasets: they can focus on the visual information, and don’t have to worry about the syntactic structure as it is similar for all descriptions.

### Table 2. Effective vocab size (EVS), number of tokens per caption (TPC) and Effective Decision (ED@N). The EVS-N is the percentage of n-grams that do not act like 1-grams in the dataset. A large EVS-n means that language is more diverse, while a small EVS-n means that there are very few combinations of possible n-grams. The ED@N is the expected number of decision that a model has to make when generating captions of length N. WT-103 is WikiText-103 [23], a common natural language dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>TPC</th>
<th>EVS-2</th>
<th>EVS-3</th>
<th>EVS-4</th>
<th>ED@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVD</td>
<td>7.03</td>
<td>47.83%</td>
<td>25.29%</td>
<td>14.67%</td>
<td>2.90</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>9.32</td>
<td>52.96%</td>
<td>26.44%</td>
<td>13.68%</td>
<td>2.88</td>
</tr>
<tr>
<td>VATEX</td>
<td>15.29</td>
<td>54.84%</td>
<td>32.60%</td>
<td>18.86%</td>
<td>3.38</td>
</tr>
<tr>
<td>MS-COCO</td>
<td>11.33</td>
<td>53.91%</td>
<td>32.59%</td>
<td>20.56%</td>
<td>3.51</td>
</tr>
<tr>
<td>WT-103</td>
<td>87.04</td>
<td>95.19%</td>
<td>34.49%</td>
<td>17.81%</td>
<td>3.72</td>
</tr>
</tbody>
</table>

Instead of only looking at the number of n-grams, in order to measure the amount of n-gram diversity that is introduced into a dataset, we introduce the N-Gram Effective Vocab Size metric (EVS-N), which measures the percentage of n-grams that do not act like 1-grams in practice. Formally, EVS-N is the percentage of tokens for which an N-gram language model has zero conditional variance (i.e. the percentage of tokens for which an n-gram language model does not assign 100% probability to a single next token). This metric can be thought of as a language-generation complexity metric — a higher EVS means that it will be more difficult for a model to memorize captions, while a low EVS suggests that models need only determine the first few words in order to generate a high-quality caption. Table 2 shows EVS-N performance, and a shocking result. The EVS-2 is approximately 50% for all datasets, suggesting that in the majority of cases, the model is able to make only one decision to generate two tokens, contrasting with WikiText-103 [23], where the EVS-2 is 95.19%.

In addition to just understanding the EVS, we can combine the EVS scores with the average number of tokens in a dataset to compute the average number of “decisions” that a model has to make during generation. The ED@N, or expected number of decisions made in a description of length N is also given in Table 2. Formally, the ED@N is the expected number of tokens in a description of length N for which an n-gram language model of the dataset has non-zero variance conditioned on the sentence so far. Surprisingly, most of the datasets have very similar ED scores (despite their differing average token lengths), and the number is low: only 3-3.5 decisions have to be made on average to get the desired caption. This low number has major implications in the quality of the captions: the fewer the number of decisions that need to be made at training, the less diverse the captions will be during test time, and the less likely models trained on the low-ED data will be able to generalize to fine-grained differences between samples. Further, this means that the number of captions models will
Figure 2. Histogram of within-sample minimum distances under the MP-Net [34] BERT-style embeddings. MSVD and MSR-VTT both have a high number of descriptions which have 0 within-sample minimum distance, while MS-COCO and VATEX have a higher within-sample diversity.

Figure 3. Plot showing the relationship between semantic variance and the performance of leave-one-out ground truth estimates of human performance on the BLEU@4 metrics. As we increase semantic variance, the average minimum distance between ground truth samples increases, and metric performance falls.

supplementary materials for a figure). Because of this increase in distance, the leave-one-out performance of ground truths decreases, as shown in Figure 3, leading to a breakdown of the n-gram metrics (and all metrics that rely on a single-sample pairwise comparison to the set of ground truths). This effect is what causes SOTA models to outperform leave-one-out samples as demonstrated in section 3. While ideally, metrics should be independent of the variance in the ground truth data, for the datasets we analyze in the paper it is clear the sample variance is sufficient that this is not the case. Interestingly, the leave-one-out fall-off occurs at different rates for the different datasets, suggesting that some datasets are more-redundant to semantic variance than others: while we hypothesize that this is due to the choice of tokens and distribution of semantic structure, it is interesting future work to confirm this hypothesis.

Why are SOTA models immune to the effects of sample variance? It’s important to note that when evaluating models, we only look at a single sample from the model distribution. We hypothesize that instead of attempting to approximate the full distribution of captions, models are picking up on trends between samples in the data, such as a wealth of descriptions that contain simple semantic structures (as described in section 4) or individually strong training descriptions (which we will discuss in section 6) which allow the model to reduce the effective variance of the ground truth dataset during the evaluation phase by ignoring most of the ground truth captions, and only focusing on a specific subset of descriptions. While these trends are likely model-specific, we believe it is important future work to quantify and understand the kinds of descriptions that models learn to approximate, and more closely monitor the effects of over-fitting to a small subset of captions to reduce the effects of ground-truth sample variance.

The effect of reducing semantic variance appears in practice via a training trick exploited by both Perez et al. [27] and Liu et al. [21] who find that decreasing the number of reference captions during training leads to improved evaluation performance on n-gram metrics. By artificially restricting the semantic variance of the training dataset, models are able to over-fit to a smaller subset of semantically redundant captions, and exploit current pairwise metrics.

Thus, we are stuck in a catch-22 when it comes to adding more captions per sample. If we increase the number of captions, we decrease our metrics’ ability to accurately discern caption quality, however if we reduce the number of captions, we can improve the accuracy of current metrics, and obtain models that achieve higher metric scores, at the cost of bland and generic captions.

6. Dataset Level Diversity

Not only do sample level diversity and within-sample diversity have important impacts on models and metrics, but dataset-level conceptual diversity matters as well. A common criticism of captioning models is that they are not generative, but instead, reproduce captions from the training set based on a set of global criteria. In general, we hypothesize that a lack of diversity in the dataset, both in the lack of overall visual concept diversity, and the exact distribution of that diversity in the dataset itself leaves models vulnerable to choosing classification over generation. We further hypothesize that a lack of conceptual diversity leads models to produce a few generic captions based on high-level visual features, instead of generating semantically detailed captions. In order to support this hypothesis, we attempt to answer two questions: “how much performance can we achieve with classification alone?” and “how much does the explicit selection of visual samples encourage models to-
Figure 4. A qualitative example from MSR-VTT demonstrating several diversity effects. The blue description is a description with the minimum distance from the sentence embedding mean, while the red description maximizes the mean BLEU@4 score to all other captions in the sample. Notably, both captions are much more generic than the other captions in the data, a trend which is consistent across all samples. We can see that the variance within this sample is high, however the tokens themselves are similar (annotators select similar tokens for the same sample). Captions are ordered from top to bottom by similarity to the mean caption embedding (See section 5).

6.1. How many captions make up a dataset?

One interesting question to ask is, how many captions do you reasonably need to use in order to solve a dataset to a particular score? This metric is a reasonable proxy for concept-level diversity, and can more globally measure the performance of a model. To answer this question, we used a greedy approximation algorithm for optimal set cover to approximate the minimum number of captions from the training set that need to be chosen for MSR-VTT and MSVD in order to achieve a particular BLEU@4 score on the validation set. We don’t compute this number for VATEX/MSCOCO or metrics beyond BLEU due to the computational cost of computing a full matrix of caption distances. Figure 5 demonstrates the results of this experiment. We can see here that to achieve SOTA BLEU@4 performance, we need only to select optimally from a set of 43 captions in the case of MSVD, and 156 captions in the case of MSR-VTT. Even further, it’s interesting to see that with only 58 captions in MSVD and 289 captions in MSR-VTT, we can achieve almost optimal BLEU scores.

This particular result, combined with the fact that models only need to make a few token-level decisions when generating language (See subsection 4.2) appears to be a real cause for models producing generic captions. Not only do models not have to make many decisions, but overall, they don’t have to select from many visual concepts either.

6.2. Does the feature set matter?

Caption models are limited not only by a classification effect but also by the concept-level diversity of the feature extractors that they use. When models rely on particular feature extraction methods, we expect pre-initialized features to bias models towards classification over generation, particularly classification among the concepts present in the pre-training data. Recently, Srinivasan et al. [35] showed that these biases can compound - so it seems natural to ask the question: how much do we expect biases in our datasets to compound with feature extractor bias?

In order to measure how much particular datasets are biased towards particular feature extractors, we compute a concept-level “overlap” between several popular feature datasets [7, 12, 20, 53], and the visual description datasets. Table 3 demonstrates the percentage of samples in the visual description datasets which contain at least one description that has a sub-string matching a label from the pre-training dataset.

```
<table>
<thead>
<tr>
<th>Dataset</th>
<th>GT</th>
<th>ImageNet</th>
<th>Kinetics</th>
<th>COCO</th>
<th>Places</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSVD</td>
<td>0.453</td>
<td>0.652</td>
<td>0.442</td>
<td>0.634</td>
<td>0.470</td>
</tr>
<tr>
<td>MSR-VTT</td>
<td>0.210</td>
<td>0.678</td>
<td>0.467</td>
<td>0.650</td>
<td>0.521</td>
</tr>
<tr>
<td>VATEX</td>
<td>0.234</td>
<td>0.576</td>
<td>0.460</td>
<td>0.547</td>
<td>0.485</td>
</tr>
<tr>
<td>COCO</td>
<td>0.152</td>
<td>0.680</td>
<td>0.515</td>
<td>0.704</td>
<td>0.292</td>
</tr>
</tbody>
</table>
```

Table 4. Performance on BLEU@4 score when using the best core-set ground truth from overlapping categories. Performance remains surprisingly high when using shared captions, implying that models are able to leverage template captions instead of scene understanding. GT: random within-sample leave-one-out ground truth performance.
only the word “baseball”), we found that fuzzy matching induced significant numbers of false-positives. This metric, thus, represents a lower-bound on the overlap (as can be seen in the case of MS-COCO, where only 91% of the descriptions contain an object from the official label set).

We can see that in datasets except for MSR-VTT, the dataset overlap with ImageNet is relatively high, likely leading to models which achieve performance based solely on the use of ImageNet features, as the classification effect detailed in both subsection 6.1 and subsection 4.2 can be exaggerated. Similarly, for datasets besides MSR-VTT, adding object detection features is likely to exaggerate the classification effect, as the model will be pre-disposed to split samples into object-category bins.

To explore exactly how much classification performance can be achieved splitting only along feature extractor boundaries, we generate sets of captions that match (using exact matching) a particular label in the feature extractor pre-training dataset. For each sample, we generate a hypothesis using a randomly sampled caption from the union of the matching concepts and compute the metric score of that hypothesis (See the supplementary materials for a detailed discussion). The results of this experiment are given in Table 4, and we can see that without sufficient conceptual diversity, models can achieve strong performance by segmenting samples among higher-order labels instead of leveraging visual understanding.

7. Recommendations & Limitations

Our aim in this work is to demonstrate that there are three unique levels of diversity that need to be maintained when collecting a dataset: Token-level diversity, within-sample diversity, and dataset conceptual diversity.

In section 4 we showed that a lack of token diversity can lead to simple captions from a core data level: few decisions need to be made to generate captions, and a large number of the tokens responsible for this generation are relatively common, opening the door for potential limits to model diversity. Token-level diversity is primarily controlled during the labeling phase of dataset collection, so we believe that both when researchers collect novel data, and when they are building splits for current datasets, they should focus on token diversity. Primarily, to encourage models to generate from a diverse set of captions, we recommend maximizing the ED@N score from section 4, along with increasing token EVS by improving the diversity of collected captions. Prompts encouraging crowd-source workers to include higher semantic detail and limits on sentence complexity (such as those introduced in VATEX [42] and Barbosa et al. [4]) could help prevent token-diversity effects from appearing in downstream models.

On the other hand, collecting too many ground truths, as discussed in section 5 presents a model training issue. Currently, models are trained to reduce semantic variance, which can lead to captions which are less complex than we expect. We believe that it is essential future research to explore how to account for the fact that variance in ground truth video descriptions is signal and not noise. Methods for managing multi-modal conditional distributions such as Slade et al. [32] or multi-label learning such as Tsoumakas et al. [39] may represent step towards such methods. Further, metrics that we use reinforce semantic variance effects by computing maximums with single samples. We believe that investigating metrics which focus on comparing multiple model samples to the full set of ground truth samples represents a possible solution. By forcing models to approximate the entire ground truth distribution we may avoid creating models which optimize away variance in the data. Finally in section 6, we discussed how a lack of diversity at a concept level can impact the performance of models. When metrics have fewer global concepts, or high overlap with feature extraction methods, they are more likely to trend towards classification over generation. In order to remedy this effect, we recommend the creation of datasets through sampling independent from the label sets of feature models. We additionally recommend that when creating training, validation, and testing splits in the dataset, the concept-level diversity is monitored to avoid introducing potential feature or concept biases with respect to popular feature extraction methods.

Visual Description Toolkit: Alongside this work, we are releasing a new toolkit 2 for visual description dataset evaluation, which is designed to analyze the performance of models (or ground truths) across the axes explored in this work. We hope that by making tools for evaluating visual description datasets easily accessible, we can encourage the field to deeply investigate the sample diversity in their data and predictions. Further, as part of the analysis toolkit, we are also releasing a set of splits and of the validation and test data for the given datasets, designed to test the performance of models along several of the axes that we discuss in this work, including conceptual labels and caption length among others. We hope that such methods for evaluation can help uncover the deviations of the model from the ground truth data, and paint a more complete picture of our descriptive models beyond n-gram scores.

Limitations: While we have demonstrated how diversity at several levels directly impacts the performance of downstream models, we believe that additional research is required to further understand how the problem of visual description differs from classification and natural language processing. In section 5, we use several proxies for caption complexity, however it is not immediately clear that such proxies are good measures for the semantic complexity of

---

2 Toolkit available at https://github.com/CannyLab/vdtk
Figure 5. For several datasets, how many captions from the training dataset are required to achieve a particular BLEU@4 score on the test set. We can see that in the optimal case, only a few (58 for MSVD, 197 for MSR-VTT, 1578 for MS-COCO) captions are required to achieve SOTA performance on the dataset. Notably, MS-COCO uniquely requires a unique description for each image.

8. Background & Related Work

This is not the first work to analyze video description data from a dataset and metric perspective, however, we believe that it is the first to focus on how dataset diversity and metric choices directly affect caption generalization. Hendricks et al. [14], Bhargava et al. [5], Tang et al. [38] and Zhao et al. [52] have all demonstrated that visual description data is often biased with respect to protected attributes (such as race, gender or religion), and introduced new methods for handling specific biases - however, they do not discuss the impact of general biases on model performance. Both Smeaton et al. [33] and Yang et al. [46] demonstrate poor cross-dataset generalization in visual description, and demonstrate that the choice of dataset directly affects model generalization ability, as well as introduce additional model-centric methods for mitigating the impact of dataset effects. These works complement our own, and they support our core hypotheses that we discuss in section 7.

Outside of visual description, the evaluation of how linguistic data and metrics affects the performance of downstream vision and language models is prevalent. Cadene et al. [6] demonstrate unimodal language biases in visual question answering and Choi et al. [10] do the same for action recognition. While many papers [11, 16, 18, 29, 31, 45] make recommendations for reducing linguistic bias based on the modeling framework, these works do not focus on the quality of generation, and instead, focus on the equally important trend of models relying heavily on language priors to solve tasks. Barbosa et al. [4] introduce methods for dataset collection which attempt to reduce linguistic bias, which represents a great leap forward from standard Amazon Mechanical Turk (AMT) collection methods, but does not discuss how the diversity impacts the performance of downstream models beyond balancing language priors.

9. Conclusion

In this work we have taken a close look at linguistic diversity in common visual description datasets, and detailed how diversity can impact models and metrics. At the token level, we showed that a lack of diversity impacts the ability of metrics to assess the quality of captions, and the ability of models to generate diverse descriptions. At the sample level, we demonstrated that high within-sample diversity is both a blessing and a curse, leaving us with either a failure of metrics to correctly measure performance, or leaving us with correct metrics, but bland and generic captions. Finally, at the dataset level, we demonstrated that even when single sample and within-sample diversity is maintained, a lack of conceptual diversity at the dataset level can bias models towards visual classification over language generation, opening the door for models which can use a few, generic, samples to solve the visual description task instead of generating captions which are rich in semantics.

While this work demonstrates the potential pitfalls of a lack of diversity in visual description datasets, we believe that by introducing new tools for analysis, and additional recommendations for data collection and model evaluation, the field will be able to investigate the sources of poor model generalization more closely, and build models which are both robust to visual diversity and can generate diverse, high quality, and semantically meaningful captions.
References


Winter Conference on Applications of Computer Vision, pages 3039–3049, 2021. 1, 2, 5


