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

David M. Chan[1], Austin Myers[2], Sudheendra Vijayanarasimhan[2], David A. Ross[2]
Bryan Seybold[2], John F. Canny[1]
[1]University of California, Berkeley [2]Google 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.

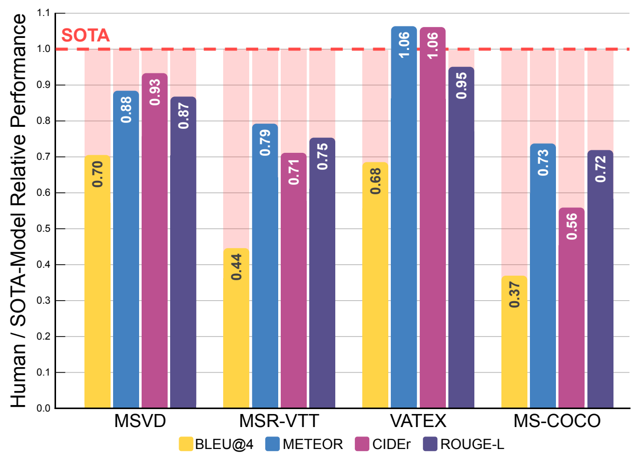


Figure 1. Captions generated by state-of-the-art (SOTA) models outperform held-out ground truth captions written by humans on common visual description datasets and metrics. Despite being far from human-level, SOTA models appear to outperform humans on most datasets and metrics, with the exception of VATEX, a relatively new dataset (and not even on all metrics). This discrepancy begs the question, "What causes these effects?" and "Are these effects indicative of a more serious issue with visual description datasets or model evaluation methods?" The figure above shows metric performance normalized to a recent SOTA model across several visual description datasets.

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 FAIEr [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 pretraining 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

[1] As described in section 7, we make the tools available for this analysis public, so any additional datasets can be analyzed.

Dataset	Unique	WS-Unique	Head
MSVD	9455	11.8%	944
MSR-VTT	22780	21.55%	1636
VATEX	31364	24.87%	1363
MS-COCO	35341	33.76%	824

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.

truth, and the remaining ground truths for that visual context. When averaged over many trials, this stochastic approach generates an estimate of human performance on the dataset (see the supplementary materials for details).

Our results are summarized in Figure 1. We can see that SOTA methods significantly outperform this estimate of human performance on the MSVD, MSR-VTT, and MS-COCO datasets. This result is not only counter-intuitive, but detrimental to progress in the field of video description, as it draws into question the usefulness of standard metrics as an indicator of model performance and generalization. These results motivate questions of understanding: "Why, and how, do models exploit the current metrics to achieve strong performance?" and "How can we limit the the exploitation of N-Gram centric metrics". The goal of the next several sections is to explore these questions through the lens of data diversity. Through analysis of single-token, n-gram, within-sample, and cross-sample diversities, we demonstrate how linguistic patterns affect models and metrics and explore how we can mitigate these effects.

4. Single Sample Diversity

To understand and analyze the impact of caption diversity on both model and metric quality, we need to first understand the diversity of the dataset itself. Many datasets use "vocabulary size", the number of unique tokens (usually words), as a proxy for the diversity of the dataset, however we hypothesize that this metric alone does not tell the full story of token level diversity in visual description datasets. In this section, we analyze the diversity of visual description datasets bottom-up, starting from tokens and working our way up to measures of n-gram complexity.

4.1. Token-Level Diversity

Table 1 provides a token-level analysis of each of the datasets. In addition to reporting the number of unique tokens in the dataset, we also introduce three new measures of diversity: Within-Sample uniqueness, which measures the percentage of tokens that are unique within a particular image/video; and "Vocab-Head", which measures the number of tokens making up 90% of the tokens in the dataset. Within-sample diversity ranges between 11% and 35%, suggesting that within samples, the descriptions are relatively varied. We discuss the impact of within-sample diversity in section 5. As expected, a small fraction of tokens represent 90% of the occurrence in most of the datasets. In MS-COCO, 2% of the tokens represent 90% of the occurrences, while at the other extreme 10% of the tokens are required for MSVD. This begs the question: how does the effective vocab size impact performance?

To validate how effective vocab size impacts performance, we used the same setup as in section 3 to compute the performance of the ground truths, however, replaced tokens in the tail of the token distribution with unique "UNK" tokens. Performance dropped significantly in all cases, with the most dramatic drop for MSVD (drop of 63.87%) and the least for MS-COCO (drop of 51.23%). MSR-VTT experienced a decrease of 58.66% and VATEX experienced a decrease of 56.20%). Counter-intuitively, the longer the tail, the less performance decreased. This result, confirmed in classification by Tang *et al*. [37], implies that models which generate from a limited vocabulary are advantaged (in terms of n-gram performance) when the head is relatively small, leading to undesirable generation behavior.

Following Wang *et al*. [42], we analyze the datasets at the level of the parts of speech in the dataset (See the supplementary materials for details). VATEX has more than 2 verbs per caption on average (by design, see [42]) while the other datasets have at most 1.3 verbs. While VATEX is the most linguistically complex, the distribution has significantly different base statistics, likely explaining poor cross-dataset generalization to VATEX from MSR-VTT and MSVD trained models. MSR-VTT is the most diverse from an object perspective (1512 nouns representing 90% of the noun mass), which lends additional support to the observations by Zhang *et al*. [51], who find that a strong object detector and good object features are necessary for strong MSR-VTT performance. Notably, MS-COCO has a very high within-sample noun diversity, suggesting that many of the captions in MS-COCO focus on different objects in each sample, and supporting hypotheses introduced in Anderson *et al*. [2] based on multiple-object attention for this dataset.

4.2. N-Gram-Level Diversity

From tokens, we can move on to exploring how the tokens fit together. 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.

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

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.

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 the 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 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 the standard n-gram metrics will often perform well, as any generated sample should also lie close to at least one ground truth sample. Unfortunately, as we increase the number of diverse ground truths (increase the sample variance), the minimum distance between samples increases (See the

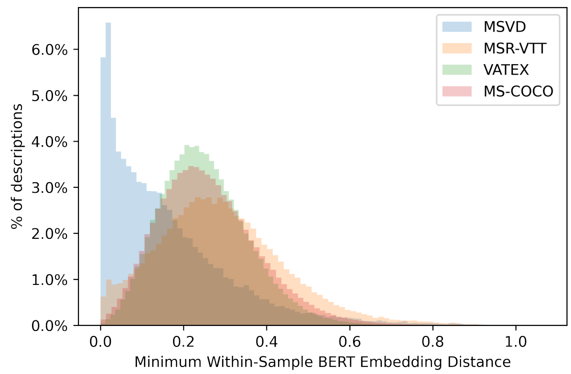


Figure 2. Histogram of within-sample minimum distances under the MP-Net [34] BERT-style embeddings. MSVD and MSR-VTT both have a high number of descriptions which have 0 within-sample minimum distance, while MS-COCO and VATEX have a higher within-sample diversity.

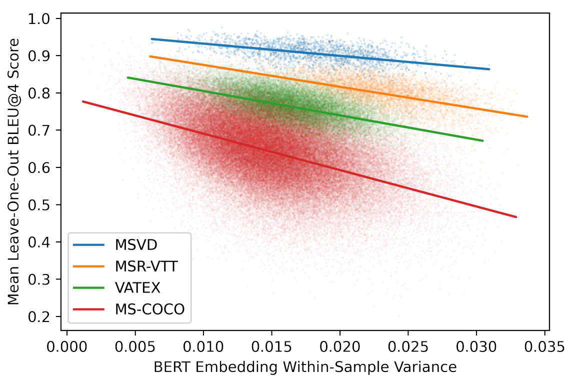


Figure 3. Plot showing the relationship between semantic variance and the performance of leave-one-out ground truth estimates of human performance on the BLEU@4 metrics. As we increase semantic variance, the average minimum distance between ground truth samples increases, and metric performance falls.

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-

Dataset: MSR-VTT (Video 6641)
Inter-Sample BERT-Embedding Distance [Min, Avg., Max]: [0.41,0.58,0.66]
Mean Leave One Out BLEU@4 Score: 0.4028

- there is a woman is talking about a simple recipe (BLEU@4: 0.920)
- there is a woman is serving a dish on the table
- a plate of food is displayed with two chopsticks next to it
- a woman is making a singapore style noodle dish in a pot
- a noodle dish with sprouts and shrimp sits on a countertop
- a pair of black chopsticks with decorative bands rests on a bowl of noodles with shrimp white bean sprouts and leafy green garnish
- some singapore noodles of egg is kept on the plate and the stick is placed on it
- there is a woman is talking about noodles
- the narrator tells about making singapore style fried noodles with prawns in a prawn-based broth
- noodles kept on the plate along with the sticks to eat it
- a woman describing a noodle dish in a white bowl (BLEU@4: 0.978)
- a noodles is been placed in the plate and a girl is describing about it
- delicious noodles are in a bowl with chopsticks
- a lady is explaining how to make a singapore dish which consists of fried noodles and prawns
- a woman describes an elegant recipe consisting of fried noodles with prawns in a prawn-based broth garnished with fresh lime wedges
- a bowl of vegtables prawns and noodles is shown chop sticks rest on the edge of the bowl while a woman's voice talks about the dish
- a lady is presenting fried noodles with prawns in a prawn broth
- there is a bowl on the table with a singapore style dish of fried noodles with prawn in a prawn broth

O2NA Baseline: there is a man is talking about a dish (BLEU@4: 0.467)

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).

wards classification over generation?"

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. While exact overlap from labels to descriptions may exclude some cases (for example the label "playing baseball" does not overlap with any description which has

Dataset	ImageNet	Kinetics	COCO	Places
MSVD	98.27%	38.88%	89.03%	55.68%
MSR-VTT	68.88%	23.51%	59.82%	46.44%
VATEX	98.60%	40.12%	76.86%	60.55%
MS-COCO	93.22%	8.83%	91.70%	60.49%

Table 3. 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.

Dataset	GT	ImageNet	Kinetics	COCO	Places
MSVD	0.453	0.652	0.442	0.634	0.470
MSR-VTT	0.210	0.678	0.467	0.650	0.521
VATEX	0.234	0.576	0.460	0.547	0.485
COCO	0.152	0.680	0.515	0.704	0.292

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 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.

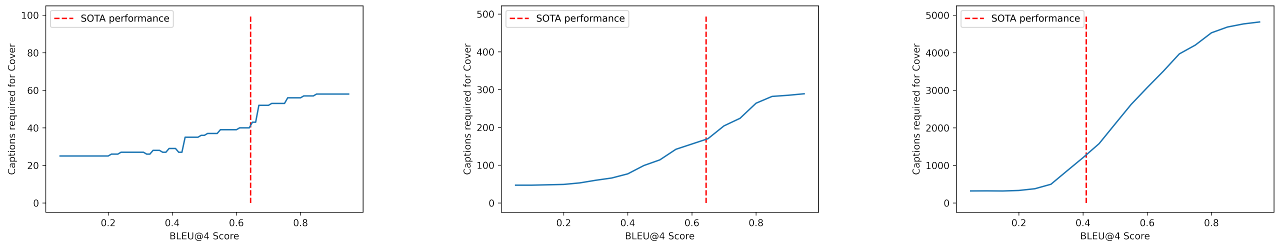
On the other hand, collecting too many ground truths, as discussed in section 5 presents a model training issue. Currently, models are trained to reduce semantic variance, which can lead to captions which are less complex than we expect. We believe that it is essential future research to explore how 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`



(a) MSVD Dataset (b) MSR-VTT Dataset (c) MS-COCO Dataset

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.

a caption. As far as we are aware, no such measure of the "usefulness" of a caption to a visually impaired user exists, that we can use to evaluate our current caption data. Figure 4 and the additional qualitative examples in the supplementary material) demonstrate some correlations between caption complexity, and the mean caption, however we believe that deeper analysis is necessary.

Our methods are also limited by the choice of metrics used in this work. Explorations of recent metrics such as FAIer [41] may indicate that they alleviate diversity effects by focusing on visual information over textual information, and leveraging pre-trained grounding models. While novel metrics may solve some of the problems, the training effects observed in section 5 remain common between all models, and the diversity in section 4 and section 6 are local to the datasets, and will remain regardless of the metric used.

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

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019. 1, 2

[2] Peter Anderson et al. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 2

[4] Natã M Barbosa and Monchu Chen. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. 7, 8

[5] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019. 8

[6] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:841–852, 2019. 8

[7] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6

[8] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[9] Shaoxiang Chen and Yu-Gang Jiang. Motion guided region message passing for video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1543–1552, 2021. 1, 2

[10] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *arXiv preprint arXiv:1912.05534*, 2019. 8

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to model and ignore dataset bias with mixed capacity ensembles. *arXiv preprint arXiv:2011.03856*, 2020. 8

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] Joshua Feinglass and Yezhou Yang. Smurf: Semantic and linguistic understanding fusion for caption evaluation via typicality analysis. *arXiv preprint arXiv:2106.01444*, 2021. 2

[14] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 3, 8

[15] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*, 2019. 1, 2

[16] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pages 1–5, 2020. 8

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 2

[18] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581, 2019. 8

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 6

[21] Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. O2na: An object-oriented non-autoregressive approach for controllable video captioning. *arXiv preprint arXiv:2108.02359*, 2021. 1, 4, 5

[22] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999, 2017. 1

[23] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. 4

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1, 2

[25] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. 1, 2

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 2

[27] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF

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

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2

[29] Deven Shah, H Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*, 2019. 8

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[31] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020. 8

[32] P Slade and Tamás D Gedeon. Bimodal distribution removal. In *International Workshop on Artificial Neural Networks*, pages 249–254. Springer, 1993. 7

[33] Alan F Smeaton, Yvette Graham, Kevin McGuinness, Noel E O'Connor, Seán Quinn, and Eric Arazo Sanchez. Exploring the impact of training data bias on automatic generation of video captions. In *International Conference on Multimedia Modeling*, pages 178–190. Springer, 2019. 1, 2, 8

[34] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020. 4, 5

[35] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666*, 2021. 6

[36] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*, 2021. 1, 2

[37] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020. 3

[38] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pages 633–645, 2021. 8

[39] Grigorios Tsoumakas and Min-Ling Zhang. Learning from multi-label data. 2009. 7

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[41] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059, 2021. 1, 2, 8

[42] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 1, 2, 3, 7

[43] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 2

[44] Bang Yang, Yuexian Zou, Fenglin Liu, and Can Zhang. Non-autoregressive coarse-to-fine video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3119–3127, 2021. 4

[45] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020. 1, 8

[46] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 8

[47] Ilmi Yoon, Umang Mathur, Brenna Gibson, Tirumalashetty Pooyan Fazli, and Joshua Miele. Video accessibility for the visually impaired. In *International Conference on Machine Learning AI for Social Good Workshop*, 2019. 1

[48] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782, 2016. 1

[49] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1, 2

[50] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 2

[51] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 3

[52] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. *arXiv preprint arXiv:2106.08503*, 2021. 8

[53] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[54] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2