DIVERSE BEAM SEARCH: DECODING DIVERSE SOLUTIONS FROM NEURAL SEQUENCE MODELS

Ashwin K Vijayakumar[1], **Michael Cogswell**[1], **Ramprasath R. Selvaraju**[1], **Qing Sun**[1]
Stefan Lee[1], **David Crandall**[2] **& Dhruv Batra**[1]
{ashwinkv,cogswell,ram21,sunqing,steflee}@vt.edu
djcran@indiana.edu, dbatra@vt.edu

[1] Department of Electrical and Computer Engineering,
Virginia Tech, Blacksburg, VA, USA

[2] School of Informatics and Computing
Indiana University, Bloomington, IN, USA

ABSTRACT

Neural sequence models are widely used to model time-series data. Equally ubiquitous is the usage of beam search (BS) as an approximate inference algorithm to decode output sequences from these models. BS explores the search space in a greedy left-right fashion retaining only the top-B candidates – resulting in sequences that differ only slightly from each other. Producing lists of nearly identical sequences is not only computationally wasteful but also typically fails to capture the inherent ambiguity of complex AI tasks. To overcome this problem, we propose *Diverse Beam Search* (DBS), an alternative to BS that decodes a list of diverse outputs by optimizing for a diversity-augmented objective. We observe that our method finds better top-1 solutions by controlling for the exploration and exploitation of the search space – implying that DBS is a *better search algorithm*. Moreover, these gains are achieved with minimal computational or memory overhead as compared to beam search. To demonstrate the broad applicability of our method, we present results on image captioning, machine translation and visual question generation using both standard quantitative metrics and qualitative human studies. Further, we study the role of diversity for image-grounded language generation tasks as the complexity of the image changes. We observe that our method consistently outperforms BS and previously proposed techniques for diverse decoding from neural sequence models.

1 INTRODUCTION

In the last few years, Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs) or more generally, neural sequence models have become the standard choice for modeling time-series data for a wide range of applications such as speech recognition (Graves et al., 2013), machine translation (Bahdanau et al., 2014), conversation modeling (Vinyals & Le, 2015), image and video captioning (Vinyals et al., 2015; Venugopalan et al., 2015), and visual question answering (Antol et al., 2015). RNN based sequence generation architectures model the conditional probability, $\Pr(\mathbf{y}|\mathbf{x})$ of an output sequence $\mathbf{y} = (y_1, \ldots, y_T)$ given an input $\mathbf{x}$ (possibly also a sequence); where the output tokens y_t are from a finite vocabulary, $\mathcal{V}$.

Inference in RNNs. Maximum a Posteriori (MAP) inference for RNNs is te task of finding the most likely output sequence given the input. Since the number of possible sequences grows as $|\mathcal{V}|^T$, exact inference is NP-hard so approximate inference algorithms like Beam Search (BS) are commonly employed. BS is a heuristic graph-search algorithm that maintains the B top-scoring partial sequences expanded in a greedy left-to-right fashion. Fig. 1 shows a sample BS search tree.

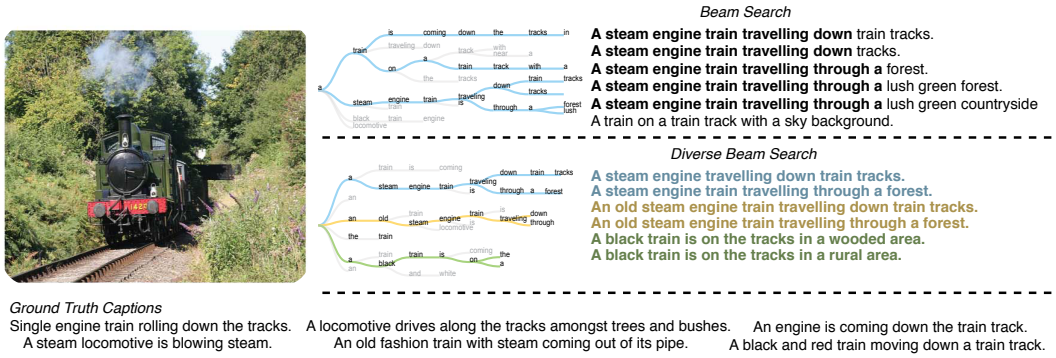


Figure 1: Comparing image captioning outputs decoded by BS and our method, Diverse Beam Search (DBS) – we notice that BS captions are near-duplicates with similar shared paths in the search tree and minor variations in the end. In contrast, DBS captions are significantly diverse and similar to the inter-human variability in describing images.

Lack of Diversity in BS. Despite the widespread usage of BS, it has long been understood that solutions decoded by BS are generic and lacking in diversity (Finkel et al., 2006; Gimpel et al., 2013; Li et al., 2015; Li & Jurafsky, 2016). To illustrate this, a comparison of captions provided by humans (bottom) and BS (topmost) are shown in Fig. 1. While this behavior of BS is disadvantageous for many reasons, we highlight the three most crucial ones here:

i) The production of near-identical beams make BS a computationally wasteful algorithm, with essentially the same computation being repeated for no significant gain in performance.

ii) Due to *loss-evaluation mismatch i.e.* improvements in posterior-probabilities not necessarily corresponding to improvements in task-specific metrics, it is common practice (Vinyals et al., 2015; Karpathy & Fei-Fei, 2015; Ferraro et al., 2016) to *deliberately throttle BS to become a poorer optimization algorithm* by using reduced beam widths. This treatment of an optimization algorithm as a hyper-parameter is not only intellectually dissatisfying but also has a significant practical side-effect – it leads to the decoding of largely bland, generic, and "safe" outputs, *e.g.* always saying "I don't know" in conversation models (Corrado, 2015).

iii) Most importantly, lack of diversity in the decoded solutions is fundamentally crippling in AI problems with *significant ambiguity – e.g.* there are multiple ways of describing an image or responding in a conversation that are "correct" and it is important to capture this ambiguity by finding several diverse plausible hypotheses.

Overview and Contributions. To address these shortcomings, we propose *Diverse Beam Search (DBS)* – a general framework to decode a list of diverse sequences that can be used as an *alternative* to BS. At a high level, DBS decodes diverse lists by dividing the beam budget into groups and enforcing diversity between groups of beams. Drawing from recent work in the probabilistic graphical models literature on Diverse M-Best (`DivMBest`) MAP inference (Batra et al., 2012; Prasad et al., 2014; Kirillov et al., 2015), we optimize an objective that consists of two terms – the sequence likelihood under the model and a dissimilarity term that encourages beams across groups to differ. This diversity-augmented model score is optimized in a *doubly greedy* manner – greedily optimizing along both time (like BS) and groups (like DivMBest).

To summarize, our primary technical contribution is Diverse Beam Search, a doubly greedy approximate inference algorithm for decoding diverse sequences. To demonstrate its broad applicability, we report results on two image-grounded language generation tasks, captioning and question generation and on machine translation. Our method consistently outperforms BS while being comparable in terms of both run-time and memory requirements. We find that DBS results in improvements on both oracle task-specific and diversity-related metrics against baselines. Further, we notice that these gains are more pronounced as the image becomes more complex consisting of multiple objects and interactions. We also conduct human studies to evaluate the role of diversity in human preferences between BS and DBS for image captions. We also analyze the parameters of DBS and show they are robust over a wide range of values. Finally, we also show that our method

is general enough to incorporate various forms for the dissimilarity term. Our implementation is available at https://github.com/ashwinkalyan/dbs. Also, a demo of DBS on image-captioning is available at dbs.cloudcv.org.

2 Preliminaries: Decoding RNNs with Beam Search

We begin with a refresher on BS, before describing our extension, Diverse Beam Search. For notational convenience, let $[n]$ denote the set of natural numbers from 1 to n and let $\mathbf{v}_{[n]} = [v_1, v_2, \dots v_n]$ index the first n elements of a vector $\mathbf{v} \in \mathbb{R}^m$, where $n \leq m$.

The Decoding Problem. RNNs are trained to estimate the likelihood of sequences of tokens from a finite dictionary $\mathcal{V}$ given an input $\mathbf{x}$. The RNN updates its internal state and estimates the conditional probability distribution over the next output given the input and all previous output tokens. We denote the logarithm of this conditional probability distribution over all tokens at time t as $\theta(y_t) = \log \Pr(y_t | y_{t-1}, \dots, y_1, \mathbf{x})$. To simplify notation, we index $\theta(\cdot)$ with a single variable y_t; but it should be clear that it depends on the previous outputs, $\mathbf{y}_{[t-1]}$ from the context. The log-probability of a partial solution (*i.e.* the sum of log-probabilities of all previous tokens decoded) can now be written as $\Theta(\mathbf{y}_{[t]}) = \sum_{\tau \in [t]} \theta(y_\tau)$. The decoding problem is then the task of finding a sequence $\mathbf{y}$ that maximizes $\Theta(\mathbf{y})$.

As each output is conditioned on all the previous outputs, decoding the optimal length-T sequence in this setting can be viewed as MAP inference on T-order Markov chain with the T nodes corresponding to output tokens. Not only does the size of the largest factor in such a graph grow as $|\mathcal{V}|^T$, but also requires wastefully forwarding of the RNN repeatedly to compute entries in the factors. Thus, approximate algorithms are employed.

Beam Search. The most prevalent method for approximate decoding is BS, which stores the top-B highly scoring candidates at each time step; where B is known as the *beam width*. Let us denote the set of B solutions held by BS at the start of time t as $Y_{[t-1]} = \{\mathbf{y}_{1,[t-1]}, \dots, \mathbf{y}_{B,[t-1]}\}$. At each time step, BS considers all possible single token extensions of these beams given by the set $\mathcal{Y}_t = Y_{[t-1]} \times \mathcal{V}$ and selects the B most likely extensions. More formally, at each step,

$$Y_{[t]} = \underset{\mathbf{y}_{1,[t]}, \dots, \mathbf{y}_{B,[t]} \in \mathcal{Y}_t}{\operatorname{argmax}} \sum_{b \in [B]} \Theta(\mathbf{y}_{b,[t]}) \quad s.t. \;\; \mathbf{y}_{i,[t]} \neq \mathbf{y}_{j,[t]} \tag{1}$$

The above objective can be trivially maximized by sorting all $B \times |\mathcal{V}|$ members of $\mathcal{Y}_t$ by their log-probabilities and selecting the top-B. This process is repeated until time T and the most likely sequence is selected by ranking the B beams based on log-probabilities.

While this method allows for multiple sequences to be explored in parallel, most completions tend to stem from a single highly valued beam – resulting in outputs that are typically only minor perturbations of a single sequence.

3 Diverse Beam Search: Formulation and Algorithm

To overcome this shortcoming, we consider augmenting the objective in Eq. 1 with a dissimilarity term $\Delta(Y_{[t]})$ that measures the diversity between candidate sequences. Jointly optimizing for all B candidates at each time step is intractable as the number of possible solutions grows with $|\mathcal{V}|^B$ (which can easily reach 10^{60} for typical language modeling settings). To avoid this joint optimization problem, we divide the beam budget B into G groups and greedily optimize each group using beam search while holding previous groups fixed. This doubly greedy approximation along both time and across groups turns $\Delta(Y_{[t]})$ into a function of only the current group's possible extensions. We detail the specifics of our approach in this section.

Diverse Beam Search. Let $Y_{[t]}$, the set of all B beams at time t be partitioned into G non-empty, disjoint subsets $Y^g_{[t]}$, $g \in [G]$. Without loss of generality, consider an equal partition such that each group contains $B' = B/G$ groups. Beam search can be applied to each group to produce B solutions; however, each group would produce identical outputs.

Unlike BS, we optimize a modified version of the objective of eq. 1 which adds a dissimilarity term $\Delta(\mathbf{y}_{[t]}, Y^g_{[t]})$, measuring the dissimilarity of a sequence $\mathbf{y}_{[t]}$ against a group $Y^g_{[t]}$. While $\Delta(\cdot, \cdot)$ can

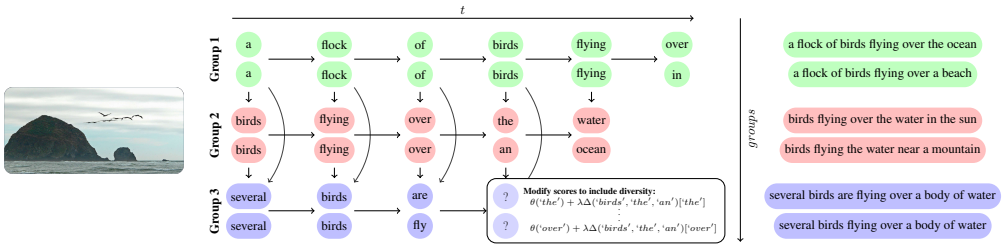


Figure 2: Diverse beam search operates left-to-right through time and top to bottom through groups. Diversity between groups is combined with joint log-probabilities, allowing continuations to be found efficiently. The resulting outputs are more diverse than for standard approaches.

take various forms, for simplicity we define one broad class that decomposes across beams within each group as:

$$\Delta(\mathbf{y}_{[t]}, Y^g_{[t]}) = \sum_{b=1}^{B'} \delta\left(\mathbf{y}_{[t]}, \mathbf{y}^g_{b,[t]}\right) \tag{2}$$

where $\delta(\cdot,\cdot)$ is a measure of sequence dissimilarity – *e.g.* a negative cost for each co-occurring n-gram in two sentences or distance between distributed sentence representations. The exact form of the sequence-level dissimilarity term can vary and we discuss some choices in Section 5.1.

As we optimize each group with the previous groups fixed, extending group g at time t amounts to a standard BS using dissimilarity augmented log-probabilities and can be written as:

$$Y^g_{[t]} = \underset{\mathbf{y}^g_{1,[t]},\ldots,\mathbf{y}^g_{B',[t]} \in \mathcal{Y}^g_t}{\operatorname{argmax}} \sum_{b\in[B']} \Theta(\mathbf{y}^g_{b,[t]}) + \sum_{h=1}^{g-1} \lambda_g \Delta\left(\mathbf{y}^g_{b,[t]}, Y^h_{[t]}\right) \quad s.t.\ \mathbf{y}^g_{i,[t]} \neq \mathbf{y}^g_{j,[t]},\ \lambda_g \geq 0 \tag{3}$$

This approach, which we call Diverse Beam Search (DBS) is detailed in Algorithm 1. An example run of DBS is shown in Figure 2 for decoding image-captions. In the example, B=6 and G=3 and so, each group performs a smaller, diversity-augmented BS of size 2. In the snapshot shown, group 3 is being stepped forward and the diversity augmented score of all words in the dictionary is computed conditioned on previous groups. The score of all words are adjusted by their similarity to previously chosen words – 'birds', 'the' and 'an' (Algorithm 1, Line 5). The optimal continuations are then found by standard BS (Algorithm 1, Line 6).

Algorithm 1: Diverse Beam Search

1 Perform a diverse beam search with G groups using a beam width of B
2 **for** $t = 1, \ldots T$ **do**
  `// perform one step of beam search for first group without diversity`
3 $Y^1_{[t]} \leftarrow \operatorname{argmax}_{(\mathbf{y}^1_{1,[t]},\ldots,\mathbf{y}^1_{B',[t]})} \sum_{b\in[B']} \Theta(\mathbf{y}^1_{b,[t]})$
4 **for** $g = 2, \ldots G$ **do**
    `// augment log-probabilities with diversity penalty`
5 $\Theta(\mathbf{y}^g_{b,[t]}) \leftarrow \Theta(\mathbf{y}^g_{b,[t]}) + \sum_h \lambda_g \Delta(\mathbf{y}^g_{b,[t]}, Y^h_{[t]}) \quad b \in [B'], \mathbf{y}^g_{b,[t]} \in \mathcal{Y}^g$ and $\lambda_g > 0$
    `// perform one step of beam search for the group`
6 $Y^g_{[t]} \leftarrow \operatorname{argmax}_{(\mathbf{y}^g_{1,[t]},\ldots,\mathbf{y}^g_{B',[t]})} \sum_{b\in[B']} \Theta(\mathbf{y}^g_{b,[t]})$
7 Return set of B solutions, $Y_{[T]} = \bigcup_{g=1}^{G} Y^g_{[T]}$

There are a number of advantages worth noting about our approach. By encouraging diversity between beams at each step (rather than just between highest ranked solutions like in Gimpel et al. (2013), our approach rewards each group for spending its beam budget to explore different parts of the output space rather than repeatedly chasing sub-optimal beams from prior groups. Furthermore, the staggered group structure enables each group beam search to be performed in parallel with a time offset. This parallel algorithm completes in $T + G$ time steps compared to $T \times G$ running time for a black-box approach of Gimpel et al. (2013).

In summary, DBS is a task agnostic, doubly greedy algorithm that incorporates diversity in beam search with little memory or computational overhead. Moreover, as the first group is not conditioned on other groups during optimization, our method is guaranteed to be at least as good as a beam search of size B/G.

4 Related Work

Diverse M-Best Lists. The task of generating diverse structured outputs from probabilistic models has been studied extensively (Park & Ramanan, 2011; Batra et al., 2012; Kirillov et al., 2015; Prasad et al., 2014). Batra et al. (2012) formalized this task for Markov Random Fields as the `DivMBest` problem and presented a greedy approach which solves for outputs iteratively, conditioning on previous solutions to induce diversity. Kirillov et al. (2015) show how these solutions can be found jointly for certain kinds of energy functions. The techniques developed by Kirillov are not directly applicable to decoding from RNNs, which do not satisfy the assumptions made.

Most related to our proposed approach is that of Gimpel et al. (2013) who apply the `DivMBest` approach to machine translation using beam search as a black-box inference algorithm. To obtain diverse solutions, beam searches of arbitrary size are sequentially performed while retaining the top-scoring candidate and using it to update the diversity term. This approach is extremely wasteful because in each iteration only one solution returned by beam search is kept. Consequently, the iterative method is time consuming and is poorly suited for batch processing or producing a large number of solutions. Our algorithm avoids these shortcomings by integrating diversity within BS such that *no* beams are discarded. By running multiple beam searches *in parallel* and at staggered time offsets, we obtain large time savings making our method comparable to classical BS. One potential advantage over our method is that more complex diversity measures at the sentence-level can be incorporated. However, as observed empirically by us and Li et al. (2015), initial words tend to significantly impact the diversity of the resultant sequences – suggesting that later words may not contribute significantly to diverse inference.

Diverse Decoding for RNNs. Some efforts have been made to produce diverse decodings from recurrent models for conversation modeling and machine translation.

In this context, our work is closely related to Li & Jurafsky (2016), who propose a BS diversification heuristic to overcome the shortcomings of Gimpel et al. (2013). This discourages sequences from sharing common roots, implicitly resulting in diverse lists. Introducing diversity through a modified objective as in DBS rather than a heuristic provides easier generalization to incorporate different notions of diversity and control for the exploration-exploitation trade-off as detailed in Section 5.1. Furthermore, we find that DBS outperforms this method.

Through a novel decoding objective that maximizes mutual information between inputs and predicted outputs, Li et al. (2015) penalize decoding generic, input independent sequences. This is achieved by training an additional target language model. Although this work and DBS share the same goals (producing diverse decodings), the techniques developed are disjoint and complementary – Li et al. (2015) develops a new model (RNN translation model with an RNN target language model), while DBS is a modified *inference* algorithm that can be applied to *any* model where BS is applicable. Combination of these complementary techniques is left as interesting future work.

5 Experiments

We first explain the baselines and evaluation metrics used in this paper. Next, we proceed to the analysis of the effects of DBS parameters. Further, we report results on image-captioning, machine translation and visual question generation. In the context of image-grounded language generation tasks, we additionally study the role of diversity with varying *complexity* of the image. Although results are reported on these tasks, it should be noted that DBS is a task-agnostic algorithm that can replace BS to decode diverse solutions.

Baselines. We compare with beam search and the following existing methods:

- Li & Jurafsky (2016): This work modifies BS by introducing an intra-sibling rank. For each partial solution, the set of $|\mathcal{V}|$ continuations are sorted and assigned intra-sibling ranks $k \in [L]$ in

order of decreasing log-probabilities, $\theta_t(y_t)$. The log-probability of an extenstion is then reduced in proportion to its rank, and continuations are re-sorted under these modified log-probabilities to select the top-B *diverse* beam extensions.

- Li et al. (2015): These models are decoded using a modified objective, $P(\mathbf{y}|x) - \lambda U(\mathbf{y})$, where $U(\mathbf{y})$ is an unconditioned target sequence model. This additional term penalizes generic input independent decoding.

Both works use secondary mechanisms such as *re-rankers* to pick a single solution from the generated lists. As we are interested in evaluating the quality of the generated lists and in isolating the gains due to diverse decoding, we do not implement any re-rankers. Instead, we simply sort the list based on log-probability. We compare to our own implementations of these methods as none are publicly available.

Evaluation Metrics. We evaluate the performance of the generated lists using the following two metrics that quantify complementary details:

- *Oracle Accuracy*: Oracle or top-k accuracy for some task-specific metric like BLEU is the maximum value of the metric over a list of k potential solutions. It is an upper bound on the potential impact diversity plays in finding relevant solutions.

- *Diversity Statistics*: We count the number of distinct n-grams present in the list of generated outputs. Similar to Li et al. (2015), we divide these counts by the total number of words generated to bias against long sentences.

Simultaneous improvements in both metrics indicate that output lists have increased diversity without sacrificing fluency and correctness with respect to target tasks. Human preference studies which compare image captions produced by DBS and BS also compare these methods. Finally, We discuss the role of diversity by relating it to intrinsic details contained in images.

5.1 Sensitivity Analysis and Effect of Diversity Functions

In this section, we study the impact of the number of groups, the strength of diversity penalty, and various forms of diversity functions for language models. Further discussion and experimental details are included in the supplementary materials.

Number of Groups (G). Setting $G{=}B$ allows for the maximum exploration of the space, while setting $G{=}1$ reduces our method to BS, resulting in increased exploitation of the search-space around the 1-best decoding. Thus, increasing the number of groups enables us to explore various modes of the model. Empirically, we find that maximum exploration correlates with improved oracle accuracy and hence use $G{=}B$ to report results unless mentioned otherwise.

Diversity Strength (λ). The diversity strength λ specifies the trade-off between the joint log-probability and the diversity terms. As expected, we find that a higher value of λ produces a more diverse list; however, excessively high values of λ can overpower model probability and result in grammatically incorrect outputs. We set λ by performing a grid search on the validation set for all experiments. We find a wide range of λ values (0.2 to 0.8) work well for most tasks and datasets.

Choice of Diversity Function (δ). As mentioned in 3, the sequence level dissimilarity term $\delta(\cdot, \cdot)$ can be design to satisfy different design choices. We discuss some of these below:

- *Hamming Diversity.* This form penalizes the selection of tokens used in previous groups proportional to the number of times it was selected before.
- *Cumulative Diversity.* Once two sequences have diverged sufficiently, it seems unnecessary and perhaps harmful to restrict that they cannot use the same words at the same time. To encode this 'backing-off' of the diversity penalty we introduce cumulative diversity which keeps a count of identical words used at every time step, indicative of overall dissimilarity. Specifically, $\delta(\mathbf{y}_{[t]}, \mathbf{y}^g_{b,[t]}) = \exp\{-(\sum_{\tau \in t} \sum_{b \in B'} I[y^h_{b,\tau} \neq y^g_{b,\tau}])/\Gamma\}$ where Γ is a temperature parameter controlling the strength of the cumulative diversity term and $I[\cdot]$ is the indicator function.
- *n-gram Diversity.* The current group is penalized for producing the same n-grams as previous groups, regardless of alignment in time – similar to Gimpel et al. (2013). This is proportional to the number of times each n-gram in a candidate occurred in previous groups. Unlike hamming

diversity, n-grams capture higher order structures in the sequences.

- *Neural-embedding Diversity.* While all the previous diversity functions discussed above perform exact matches, neural embeddings such as word2vec (Mikolov et al., 2013) can penalize semantically similar words like synonyms. This is incorporated in each of the previous diversity functions by replacing the hamming similarity with a soft version obtained by computing the cosine similarity between word2vec representations. When using with n-gram diversity, the representation of the n-gram is obtained by summing the vectors of the constituent words.

Each of these various forms encode different notions of diversity. Hamming diversity ensures different words are used at different times, but can be circumvented by small changes in sequence alignment. While n-gram diversity captures higher order statistics, it ignores sentence alignment. Neural-embedding based encodings can be seen as a semantic blurring of either the hamming or n-gram metrics, with word2vec representation similarity propagating diversity penalties not only to exact matches but also to close synonyms. We find that using any of the above functions help outperform BS in the tasks we examine; hamming diversity achieves the best oracle performance despite its simplicity. A comparison of the performance of these functions for image-captioning is provided in the supplementary.

5.2 Estimating Image Complexity

Diversity in the output space is often dependent on the input. For example, "complex" scenes consisting of various objects and interactions tend to be described in multiple ways as compared to "simple" images that tend to focus on one specific object. We study this by inspecting the gains due to DBS with varying complexity of images. One notion of image complexity is studied by Ionescu *et al*. Ionescu et al. (2016), defining a "difficulty score" as the human response time for solving a visual search task for images in PASCAL-50S Vedantam et al. (2015). Using the data from Ionescu et al. (2016), we train a Support Vector Regressor on ResNet (He et al., 2016) features to predict this difficulty score. This model achieves a 0.41 correlation with the ground truth (comparable to the best model of Ionescu et al. (2016) at 0.47). To evaluate the relationship between image complexity and performance gains from diverse decoding, we use this trained predictor to estimate a difficulty score s for each image in the COCO Lin et al. (2014) dataset. We compute the mean ($\mu = 3.3$) and standard deviation ($\sigma = 0.61$) and divide the images into three bins, `Simple` ($s \leq \mu - \sigma$), `Average` ($\mu - \sigma > s < \mu + \sigma$), and `Complex` ($s \geq \mu + \sigma$) consisting of 745, 3416 and 839 images respectively. Figure 3 shows some sample `Simple`, `Average`, and `Complex` images from the PASCAL-50S dataset. While simple images like close-up pictures of cats may only be described in a handful of ways by human captioners (first column), complex images with multiple objects and interactions will be described in many different ways depending on what is the focus of the captioner (last column). In the subsequent experiments on image-grounded language generation tasks, we show that improvements from DBS are greater for more complex images.

5.3 Image Captioning

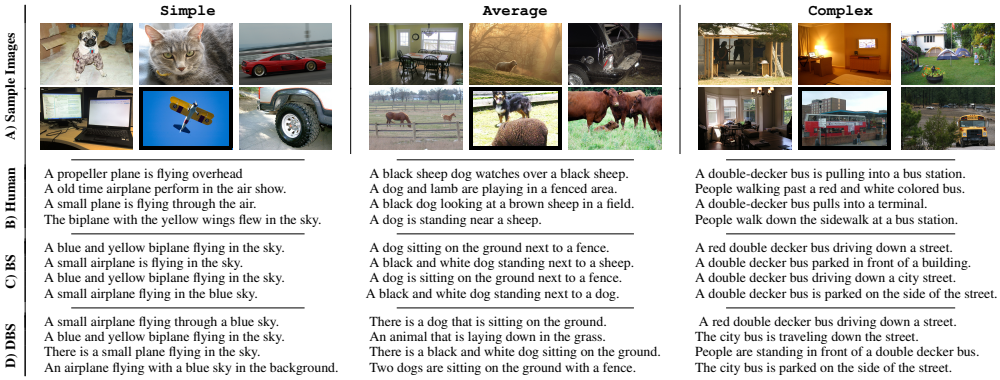
Dataset and Models. We evaluate on two datasets – COCO (Lin et al., 2014) and PASCAL-50S (Vedantam et al., 2015). We use the public splits as in Karpathy & Fei-Fei (2015) for COCO. PASCAL-50S is used only for testing save 200 validation images used to tune hyperparameters. We train a captioning model (Vinyals et al., 2015) using the `neuraltalk2`[1] code repository.

Results. As it can be observed from Table 1 (Top), DBS outperforms both BS and Li & Jurafsky (2016) on both COCO and PASCAL-50S datasets. We observe that gains on PASCAL-50S are more pronounced (7.24% and 9.60% Oracle@20 improvements against BS and Li & Jurafsky (2016)) than COCO. This suggests diverse predictions are especially advantageous when there is a mismatch between training and testing sets making DBS a better inference strategy in real-world applications.

Table 1 (Top) also shows the number of distinct n-grams produced by different techniques. Our method produces significantly more distinct n-grams (almost 300% increase in the number of 4-grams produced) as compared to BS. We also note that our method tends to produce slightly longer captions compared to beam search on average. Moreover, on the PASCAL-50S test split we observe that DBS finds more likely top-1 solutions on average – DBS obtains a maximum log-probability of -

[1] https://github.com/karpathy/neuraltalk2

Figure 3: **A)** Sample PASCAL-50S images of different difficulty. Simple images are often close-ups of single objects while complex images involve multiple objects in a wider view. **B)** Random human captions for the black-bordered images. Complex images have more varied captions than simpler images. **C**) which are not captured well by beam search compared to **D)** DBS.



6.53 as against -6.91 got by BS of same beam width. While the performance of DBS is guaranteed to be better than a BS of size $^{B}/_{G}$, this experimental evidence suggests that using DBS as a replacement to BS leads to better or at least comparable performance.

Results by Image Complexity. From Table 1, we can see that as the complexity of images increases DBS outperforms standard beam search (difference shown in parentheses) and other baselines by larger margins for all values of k. For example, at Oracle Spice@20, DBS achieves significant improvements over BS of 0.67, 0.91, and 1.13 for `Simple`, `Average`, and `Complex` images respectively. While DBS improves over BS in all settings, complex images benefit even more from diversity-inducing inference than simple images.

Human Preference by Difficulty. To further establish the effectiveness of our method, we evaluate human preference between captions decoded using DBS and BS. In this forced-choice test, DBS captions were preferred over BS 60% of the time by human annotators. Further, they were preferred about 50%, 69% and 83% of the times for `Simple`, `Average` and `Difficult` images respectively. Furthermore, we observe a positive correlation ($\rho = 0.73$) between difficulty scores and humans preferring DBS to BS. Further details about this experiment are provided in the supplement.

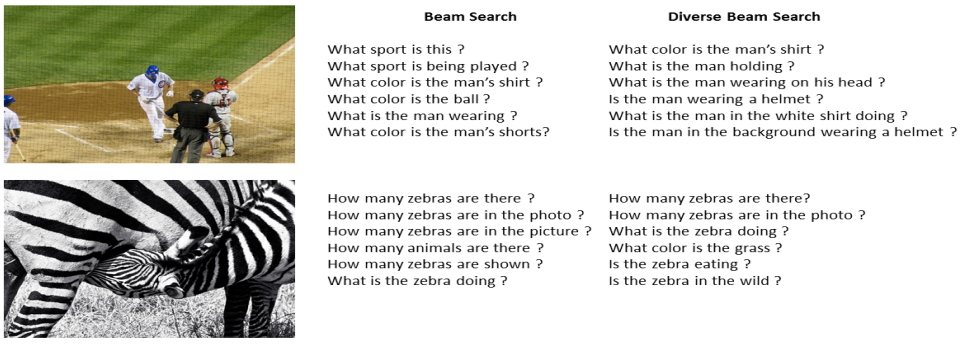


Figure 4: Qualitative results on Visual Question Generation. It can be noted that DBS generates questions that are non-generic which belong to different question types

5.4 Visual Question Generation

We also report results on Visual Question Generation (VQG) (Mostafazadeh et al., 2016), where a model is trained to produce questions *about an image*. Generating visually focused questions requires reasoning about multiple problems that are central to vision – *e.g.*, object attributes, relationships between objects, and natural language. Similar to captioning, there are many sensible questions for a given image.

Table 1: **Top:** Oracle SPICE@k and distinct n-grams on the COCO image captioning task at $B = 20$. While we report SPICE, we observe similar trends in other metrics (reported in the supplement). **Bottom:** Breakdown of results by difficulty class, highlighting the relative improvement over BS.

	Method	SPICE	Oracle SPICE@k			Distinct n-Grams			
			@5	@10	@20	n = 1	2	3	4
COCO	BS	16.27	22.96	25.14	27.34	0.40	1.51	3.25	5.67
	Li & Jurafsky (2016)	16.35	22.71	25.23	27.59	0.54	2.40	5.69	8.94
	DBS	**16.783**	**23.08**	**26.08**	**28.09**	**0.56**	**2.96**	**7.38**	**13.44**
	Li et al. (2015)	16.74	23.27	26.10	27.94	0.42	1.37	3.46	6.10
PASCAL-50S	BS	4.93	7.04	7.94	8.74	0.12	0.57	1.35	2.50
	Li & Jurafsky (2016)	5.08	7.24	8.09	8.91	0.15	0.97	2.43	5.31
	DBS	**5.357**	**7.357**	**8.269**	**9.293**	**0.18**	**1.26**	**3.67**	**7.33**
	Li et al. (2015)	5.12	7.17	8.16	8.56	0.13	1.15	3.58	8.42

	Method	SPICE	Oracle SPICE@k (Gain over BS)		
			@5	@10	@20
Simple	BS	17.28 (0)	24.32 (0)	26.73 (0)	28.7 (0)
	Li & Jurafsky (2016)	17.12 (-0.16)	24.17 (-0.15)	26.64 (-0.09)	29.28 (0.58)
	DBS	**17.42 (0.14)**	**24.44 (0.12)**	**26.92 (0.19)**	**29.37 (0.67)**
	Li et al. (2015)	17.38 (0.1)	24.48 (0.16)	26.82 (0.09)	29.21 (0.51)
Average	BS	15.95 (0)	22.51 (0)	24.8 (0)	26.55 (0)
	Li & Jurafsky (2016)	16.19 (0.24)	22.59 (0.08)	24.98 (0.18)	27.23 (0.68)
	DBS	**16.28 (0.33)**	**22.65 (0.14)**	**25.08 (0.28)**	**27.46 (0.91)**
	Li et al. (2015)	16.22 (0.27)	22.61 (0.1)	25.01 (0.21)	27.12 (0.57)
Complex	BS	16.39 (0)	22.62 (0)	24.91 (0)	27.23 (0)
	Li & Jurafsky (2016)	16.55 (0.16)	22.55 (-0.07)	25.18 (0.27)	27.57 (0.34)
	DBS	**16.75 (0.36)**	**22.81 (0.19)**	**25.25 (0.34)**	**28.36 (1.13)**
	Li et al. (2015)	16.69 (0.3)	22.69 (0.07)	25.16 (0.25)	27.94 (0.71)

The VQG dataset (Mostafazadeh et al., 2016) consists of 5 human-generated questions per image for 5000 images from COCO (Lin et al., 2014). We use a model similar to the one used for captioning, except that it is now trained to output questions rather than captions. Similar to previous results, using beam search to sample outputs results in similarly worded question while DBS decoded questions ask about multiple details of the image (see Fig. 4).

We show quantitative evaluations in Table 2 for the VQG dataset as a whole and when partitioned by image difficulty. We find DBS significantly outperforms the baseline methods on this task both on standard metrics (SPICE) and measure of diversity. We also observe that gap between DBS and the baseline methods is more pronounced than in the captioning task and attribute this to the increased variety of possible visually grounded questions compared to captions which often describe only a few major salient objects. The general trend that more complex images benefit more from diverse decoding also persists in this setting.

5.5 Machine Translation

Dataset and Models. We use the English-French parallel data from the *europarl* corpus as the training set. We report results on *news-test-2013* and *news-test-2014* and use the *newstest2012* to tune DBS parameters. We train a encoder-decoder architecture as proposed in Bahdanau et al. (2014) using the `dl4mt-tutorial`[2] code repository. The encoder consists of a bi-directional recurrent network (Gated Recurrent Unit) with attention. We use sentence level BLEU scores (Papineni et al., 2002) to compute oracle metrics and report distinct n-grams similar to image-captioning. From Table 3, we see that DBS consistently outperforms standard baselines with respect to both metrics.

[2] https://github.com/nyu-dl/dl4mt-tutorial

Table 2: **Top:** Oracle SPICE@k and distinct n-grams on the VQG task at $B = 20$. **Bottom:** Results by difficulty class, highlighting the relative improvement over BS.

	Method	SPICE	Oracle SPICE@k			Distinct n-Grams			
			@5	@10	@20	n = 1	2	3	4
VQG	BS	15.17	21.96	23.16	26.74	0.31	1.36	3.15	5.23
	Li & Jurafsky (2016)	15.45	22.41	25.23	27.59	0.34	2.40	5.69	8.94
	DBS	**16.49**	**23.11**	**25.71**	**27.94**	**0.43**	**2.17**	**6.49**	**12.24**
	Li et al. (2015)	16.34	22.92	25.12	27.19	0.35	1.56	3.69	7.21

	Method	SPICE	Oracle SPICE@k (Gain over BS)		
			@5	@10	@20
Simple	BS	16.04 (0)	21.34 (0)	23.98 (0)	26.62 (0)
	Li & Jurafsky (2016)	16.12 (0.12)	21.65 (0.31)	24.64 (0.66)	26.68 (0.04)
	DBS	**16.42 (0.38)**	**22.44 (1.10)**	**24.71 (0.73)**	**26.73 (0.13)**
	Li et al. (2015)	16.18 (0.14)	22.18 (0.74)	24.16 (0.18)	26.23 (-0.39)
Average	BS	15.29 (0)	21.61 (0)	24.12 (0)	26.55 (0)
	Li & Jurafsky (2016)	16.20 (0.91)	21.90 (0.29)	25.61 (1.49)	27.41 (0.86)
	DBS	**16.63 (1.34)**	**22.81 (1.20)**	**24.68 (0.46)**	**27.10 (0.55)**
	Li et al. (2015)	16.07 (0.78)	22.12 (-0.49)	24.34 (0.22)	26.98 (0.43)
Complex	BS	15.78 (0)	22.41 (0)	24.48 (0)	26.87 (0)
	Li & Jurafsky (2016)	16.82 (1.04)	23.20 (0.79)	25.48 (1.00)	27.12 (0.25)
	DBS	**17.25 (1.47)**	**23.35(1.13)**	**26.19 (1.71)**	**28.01 (1.03)**
	Li et al. (2015)	17.10 (1.32)	23.31 (0.90)	26.01 (1.53)	27.92 (1.05)

Table 3: Quantitative results on En-Fr machine translation on the newstest-2013 dataset (at $B = 20$). Although we report BLEU-4 values, we find similar trends hold for lower BLEU metrics as well.

Method	Oracle Accuracy (BLEU-4)				Diversity Statistics			
	@1	@5	@10	@20	distinct-1	distinct-2	distinct-3	distinct-4
Beam Search	13.52	16.67	17.63	18.44	0.04	0.75	2.10	3.23
Li & Jurafsky (2016)	13.63	17.11	17.50	18.34	0.04	0.81	2.92	4.61
DBS	**13.69**	**17.51**	**17.80**	**18.77**	**0.06**	**0.95**	**3.67**	**5.54**
Li et al. (2015)	13.40	17.54	17.97	18.86	0.04	0.86	2.76	4.31

6 Conclusion

Beam search is the most commonly used approximate inference algorithm to decode sequences from RNNs; however, it suffers from a lack of diversity. Producing multiple highly similar and generic outputs is not only wasteful in terms of computation but also detrimental for tasks with inherent ambiguity like image captioning. In this work, we presented *Diverse Beam Search*, which describes beam search as an optimization problem and augments the objective with a diversity term. The result is a 'doubly greedy' approximate algorithm that produces diverse decodings while using about the same time and resources as beam search. Our method consistently outperforms beam search and other baselines across all our experiments without *extra computation* or *task-specific overhead*. Further, in the case of image-grounded language generation tasks, we find that DBS provides increased gains as the complexity of the images increases. DBS is *task-agnostic* and can be applied to any case where BS is used – making it applicable in multiple domains.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2425–2433, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Repre-*

sentations (ICLR), 2014.

Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

Greg Corrado. Computer, respond to this email. *Google Research Blog*, November 2015. URL https://research.googleblog.com/.

Francis Ferraro, Ishan Mostafazadeh, Nasrinand Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiadong He, Pushmeet Kohli, Dhruv Batra, and C Lawrence Zitnick. Visual storytelling. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 2016.

Jenny Rose Finkel, Christopher D Manning, and Andrew Y Ng. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 618–626, 2006.

K. Gimpel, D. Batra, C. Dyer, and G. Shakhnarovich. A systematic exploration of diversity in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. abs/1303.5778, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim Papadopoulos, and Vittorio Ferrari. How hard can it be? Estimating the difficulty of visual search in an image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Alexander Kirillov, Bogdan Savchynskyy, Dmitrij Schlesinger, Dmitry Vetrov, and Carsten Rother. Inferring m-best diverse labelings in a single one. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2016.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2002.

Dennis Park and Deva Ramanan. N-best maximal decoders for part models. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

Adarsh Prasad, Stefanie Jegelka, and Dhruv Batra. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4534–4542, 2015.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

APPENDIX

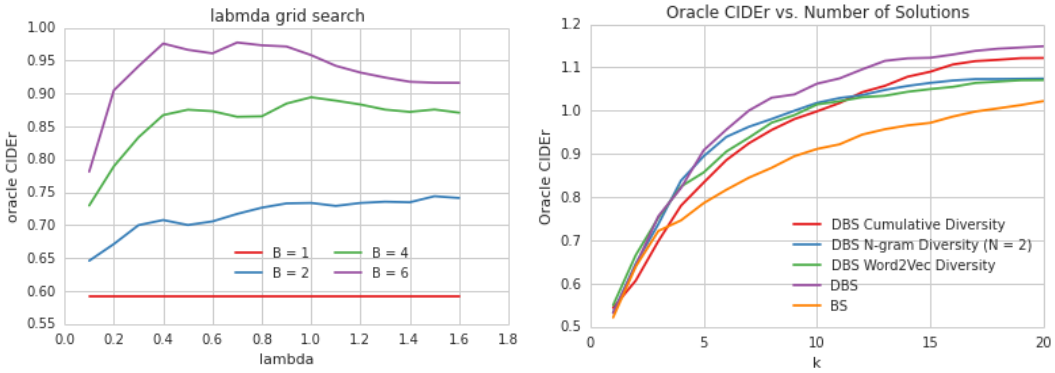
SENSIVITY STUDIES

Number of Groups. Fig. 5 presents snapshots of the transition from BS to DBS at $B = 6$ and $G = \{1, 3, 6\}$. As beam width moves from 1 to G, the exploration of the method increases resulting in more diverse lists.

Figure 5: Effect of increasing the number of groups G. The beams that belong to the same group are colored similarly. Recall that diversity is only enforced across groups such that $G = 1$ corresponds to classical BS.

Diversity Strength. As noted in Section 5.1, our method is robust to a wide range of values of the diversity strength (λ). Fig. 6a shows a grid search of λ for image-captioning on the PASCAL-50S dataset.

Choice of Diversity Function. Fig. 6b shows the oracle performace of various forms of the diversity function described in Section 5.1. We observe that hamming diversity surprisingly performs the best. Other forms perform comparably while outperforming BS.



(a) Grid search of diversity strength parameter

(b) Effect of multiple forms for the diversity function

Figure 6: Fig. 6a shows the results of a grid search of the diversity strength (λ) parameter of DBS on the validation split of PASCAL 50S dataset. We observe that it is robust for a wide range of values. Fig. 6b compares the performance of multiple forms for the diversity function (Δ). While naïve diversity performs the best, other forms are comparable while being better than BS.

HUMAN STUDIES

For image-captioning, we conduct a human preference study between BS and DBS captions as explained in Section 5. A screen shot of the interface used to collect human preferences for captions generated using DBS and BS is presented in Fig. 7. The lists were shuffled to guard the task from being gamed by a turker.

difficulty score bin range	# images	% images DBS was preffered
$\leq \mu - \sigma$	481	50.51%
$[\mu - \sigma, \mu + \sigma]$	409	69.92%
$\geq \mu + \sigma$	110	83.63%

Table 4: Frequency table for image difficulty and human preference for DBS captions on PASCAL50S dataset

As mentioned in Section 5, we observe that *difficulty score* of an image and human preference for DBS captions are positively correlated. The dataset contains more images that are less difficulty and so, we analyze the correlation by dividing the data into three bins. For each bin, we report the % of images for which DBS captions were preferred after a majority vote (*i.e.* at least 3/5 turkers voted in favor of DBS) in Table 4. At low difficulty scores consisting mostly of iconic images – one might expect that BS would be preferred more often than chance. However, mismatch between the statistics of the training and testing data results in a better performance of DBS. Some examples for this case are provided in Fig. 8. More general qualitative examples are provided in Fig. 9.

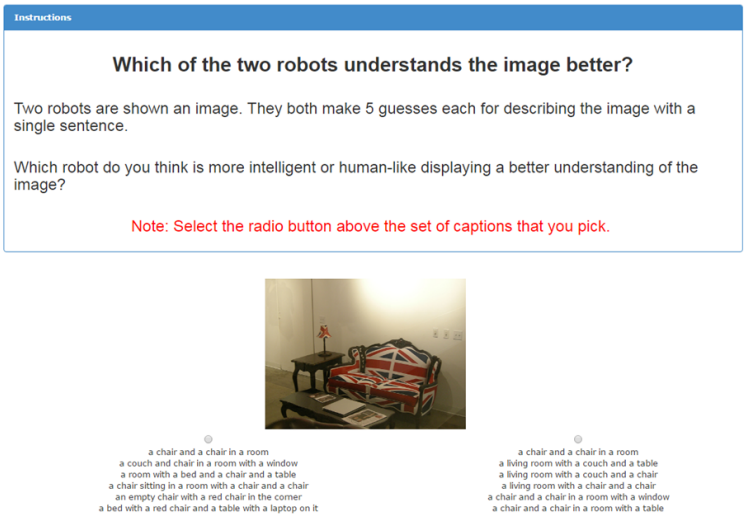


Figure 7: Screen-shot of the interface used to perform human studies

Difficulty Score : 2.8308

Beam Search
A man riding a motorcycle on a dirt road
A man riding a motorcycle on a beach
A man riding a motorcycle on the side of a road
A man riding a bike on a dirt road
A man riding a motorcycle on the side of the road
A man riding a motorcycle on the side of a beach

Diverse Beam Search
A man riding a motorcycle on a beach
A man riding a bike on a dirt road
A man riding a bike on a dirt road
A man on a motorcycle is flying a kite
A person on a skateboard riding on the side of a road
A person on a bicycle with a helmet on on the ground

Difficulty Score : 2.9287

Beam Search
A black bear standing in a grassy field
A black bear standing in a field of grass
A black bear is standing in the grass
A black bear is standing in a field
A black bear standing in the grass next to a tree
A black bear standing in the grass near a fence

Diverse Beam Search
A black dog is standing in the grass
A black dog is standing in the grass
A black bear walking through a grassy field
A black bear walking in a field of grass
A black and white dog is standing in the grass
A black bear standing in the grass near a fence

Difficulty Score : 2.8999

Beam Search
A close up of a bowl of broccoli
A close up of a plate of broccoli
A close up of a broccoli plant on a table
A close up of a bowl of broccoli on a table
A close up of a broccoli plant in a garden
A close up of a plate of broccoli and cauliflower

Diverse Beam Search
A close up of a bowl of broccoli
A close up of a plate of broccoli and broccoli
A green plant with a green plant in it
A green plant with a bunch of green leaves
A white plate topped with broccoli and a plant
A small green plant with a green plant in it

Figure 8: For images with low difficulty score, BS captions are preferred to DBS – as show in the first figure. However, we observe that DBS captions perform better when there is a mismatch between the statistics of the testing and training sets. Interesting captions are colored in blue for readability.

Difficulty Score : 5.4382

Beam Search
A group of people sitting at a table with laptops
A group of people sitting at a table
A couple of people that are sitting at a table
A group of people sitting around a table with laptops
A group of people sitting at a table in front of laptops
A group of people sitting at a table with a laptop

Diverse Beam Search
A group of people sitting at a table with laptops
A group of people sitting at a table with laptops
A group of people sitting around a table with laptops
A group of people are sitting at a table
Two people sitting at a table with laptops
Three people are sitting at a table with laptops

Difficulty Score : 4.1815

Beam Search
A woman sitting in front of a laptop computer
A woman sitting at a table with a laptop
A woman sitting at a table with a laptop computer
A woman is working on a laptop computer
A woman sitting at a desk with a laptop computer
A woman is sitting at a table with a laptop

Diverse Beam Search
A woman sitting at a table with a laptop computer
A woman is working on a laptop computer
A woman is sitting at a table with a laptop
A man sitting at a desk with a laptop computer
A woman in a kitchen with a laptop computer
A man is sitting at a table with a laptop and a computer

Difficulty Score : 3.8146

Beam Search
A wooden table topped with plates of food
A table with plates of food on it
A wooden table topped with plates and bowls of food
A table that has a bunch of plates on it
A wooden table topped with plates of food and glasses
A wooden table topped with plates of food and cups

Diverse Beam Search
A table with a plate of food and a glass of wine
A table with a plate of food and a glass
A table with plates of food and a glass of wine
A dining table with a plate of food and a glass of wine
A table with a bowl of food and a bowl of soup on it
A dining room table with a plate of food and a glass of wine on it

Figure 9: For images with a high difficulty score, captions produced by DBS are preferred to BS. Interesting captions are colored in blue for readability.