DIVERSE BEAM SEARCH: DECODING DIVERSE SOLUTIONS FROM NEURAL SEQUENCE MODELS

Ashwin K Vijayakumar[1], **Michael Cogswell**[1], **Ramprasaath 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

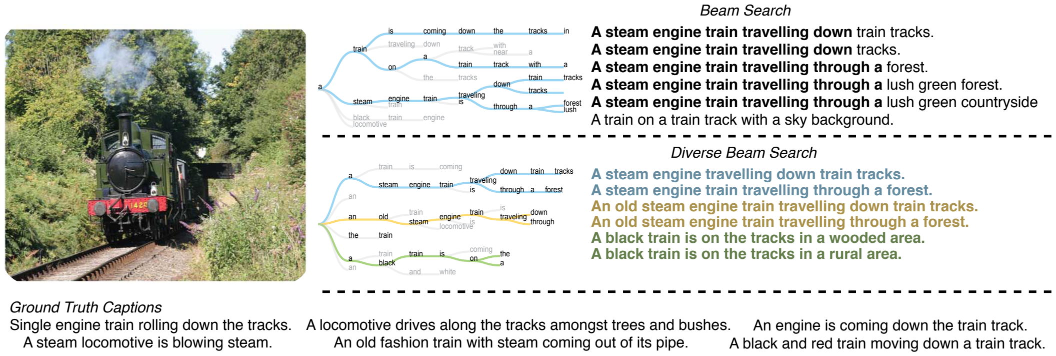
Neural sequence models are widely used to model time-series data. Equally ubiquitous is the usage of beam search (BS) as an approximate inference algorithm to decode output sequences from these models. BS explores the search space in a greedy left-right fashion retaining only the top B candidates. This tends to result 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 a diversity-augmented objective. We observe that our method not only improved diversity but also finds better top 1 solutions by controlling for the exploration and exploitation of the search space. Moreover, these gains are achieved with minimal computational or memory overhead compared to beam search. To demonstrate the broad applicability of our method, we present results on image captioning, machine translation, conversation and visual question generation using both standard quantitative metrics and qualitative human studies. We find 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 including 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 the 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.

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.,



Beam Search

A steam engine train travelling down train tracks.
A steam engine train travelling down tracks.
A steam engine train travelling through a forest.
A steam engine train travelling through a lush green forest.
A steam engine train travelling through a lush green countryside
A train on a train track with a sky background.

Diverse Beam Search

A steam engine travelling down train tracks.
A steam engine train travelling through a forest.
An old steam engine train travelling down train tracks.
An old steam engine train travelling through a forest.
A black train is on the tracks in a wooded area.
A black train is on the tracks in a rural area.

Ground Truth Captions

Single engine train rolling down the tracks. A locomotive drives along the tracks amongst trees and bushes. An engine is coming down the train track.
A steam locomotive is blowing steam. An old fashion train with steam coming out of its pipe. A black and red train moving down a train track.

Figure 1: Comparing image captioning outputs decoded by BS (top) and our method, Diverse Beam Search (middle) – 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 variability in human-generated ground truth captions (bottom).

2013; Li et al., 2015; Li & Jurafsky, 2016). Comparing the human (bottom) and BS (top) generated captions shown in Fig. 1 demonstrates this deficiency. 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 to *deliberately throttle BS to become a poorer optimization algorithm* by using reduced beam widths (Vinyals et al., 2015; Karpathy & Fei-Fei, 2015; Ferraro et al., 2016). This treatment of an optimization algorithm as a hyperparameter 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 (Kannan et al., 2016).

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 set of diverse sequences that can be used as an *alternative* to BS. At a high level, DBS decodes diverse lists by dividing the given 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).

Our primary technical contribution is Diverse Beam Search, a doubly greedy approximate inference algorithm to decode diverse sequences from neural sequence models. We report results on image captioning, machine translation, conversations and visual question generation to demonstrate the broad applicability of DBS. Results show that DBS produces consistent improvements on both task-specific oracle and other diversity-related metrics while maintaining run-time and memory requirements similar to BS. We also evaluate human preferences between image captions generated by BS or DBS. Further experiments show that DBS is robust over a wide range of its parameter values and is capable of encoding various notions of diversity through different forms of the diversty term.

Overall, our algorithm is simple to implement and consistently outperforms BS in a wide range of domains without sacrificing efficiency. Our implementation is publicly available at `https://github.com/ashwinkalyan/dbs`. Additionally, we provide an interactive demonstration of DBS for image captioning at `http://dbs.cloudcv.org`.

2 PRELIMINARIES: DECODING RNNS WITH BEAM SEARCH

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

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 $\theta(y_t) = \log \Pr(y_t | y_{t-1}, \ldots, y_1, \mathbf{x})$. To avoid notational clutter, we index $\theta(\cdot)$ with a single variable y_t, but it should be clear that it depends on all previous outputs, $\mathbf{y}_{[t-1]}$. We write the $\log$ probability of a partial solution (*i.e.* the sum of $\log$ probabilities of all tokens decoded so far) 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 a T-order Markov chain with nodes corresponding to output tokens at each time step. Not only does the size of the largest factor in such a graph grow as $|\mathcal{V}|^T$, but computing these factors also requires repetitively evaluating the sequence model. Thus, approximate algorithms are employed and the most prevalent method is beam search (BS).

Beam search is a heuristic search algorithm which stores the top B highest scoring partial 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 $Y_{[t-1]} = \{\mathbf{y}_{1,[t-1]}, \ldots, \mathbf{y}_{B,[t-1]}\}$. 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 retains the B highest scoring extensions. More formally, at each step the beams are updated as

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

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 complete beams according to their $\log$ 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 often only minor perturbations of a single sequence (and typically only towards the end of the sequences).

3 DIVERSE BEAM SEARCH: FORMULATION AND ALGORITHM

To overcome this, we augment the objective in Eq. 1 with a dissimilarity term $\Delta(Y_{[t]})$ that measures the diversity between candidate sequences, assigning a penalty $\Delta(Y_{[t]})[c]$ to each possible sequence completion $c \in \mathcal{V}$. Jointly optimizing this augmented objective for all B candidates at each time step is intractable as the number of possible solutions grows with $|\mathcal{V}|^B$ (easily 10^{60} for typical language modeling settings). To avoid this, we opt for a greedy procedure that divides the beam budget B into G groups and promotes diversity between these groups. The approximation is doubly greedy – across both time and groups – so $\Delta(Y_{[t]})$ is constant with respect to other groups and we can sequentially optimize each group using regular BS. We now explain the specifics of our approach.

Diverse Beam Search. As joint optimization is intractable, we form G smaller groups of beams and optimize them sequentially. Consider a *partition* of the set of beams $Y_{[t]}$ into G smaller sets $Y^g_{[t]}, g{\in}[G]$ of $B' = B/G$ beams each (we pick G to divide B). In the example shown in Fig. 2, $B = 6$ beams are divided into $G = 3$ differently colored groups containing $B' = 2$ beams each.

Considering diversity only between groups, reduces the search space at each time step; however, inference remains intractable. To enforce diversity efficiently, we consider a greedy strategy that steps each group forward in time sequentially while considering the others fixed. Each group can then evaluate the diversity term with respect to the fixed extensions of previous groups, returning the search space to $B' \times |\mathcal{V}|$. In the snapshot shown in Fig. 2, the third group is being stepped forward at time step $t = 4$ and the previous groups have already been completed. With this staggered beam-front, the diversity term of the third group can be computed using these completions. Here we use

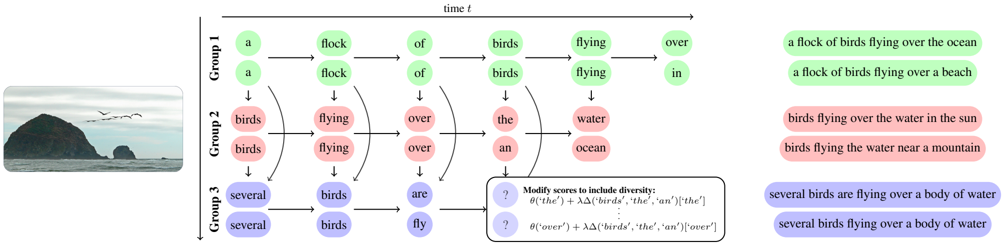

time t
Group 1
Group 2
Group 3
a
flock
of
birds
flying
over
a
flock
of
birds
flying
in
birds
flying
over
the
water
birds
flying
over
an
ocean
several
birds
are
several
birds
fly
?
?
Modify scores to include diversity:
$\theta(\text{'the'}) + \lambda\Delta(\text{'birds'}, \text{'the'}, \text{'an'})[\text{'the'}]$
$\vdots$
$\theta(\text{'over'}) + \lambda\Delta(\text{'birds'}, \text{'the'}, \text{'an'})[\text{'over'}]$
a flock of birds flying over the ocean
a flock of birds flying over a beach
birds flying over the water in the sun
birds flying the water near a mountain
several birds are flying over a body of water
several birds flying over a body of water

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.

hamming diversity, which adds diversity penalty -1 for each appearance of a possible extension word at the same time step in a previous group – 'birds', 'the', and 'an' in the example – and 0 to all other possible completions. We discuss other forms for the diversity function 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:

$$Y^g_{[t]} = \underset{\mathbf{y}^g_{1,[t]},\ldots,\mathbf{y}^g_{B',[t]} \in \mathcal{Y}^g_t}{\text{argmax}} \quad \sum_{b\in[B']} \Theta\left(\mathbf{y}^g_{b,[t]}\right) + \lambda\Delta\left(\bigcup_{h=1}^{g-1} Y^h_{[t]}\right)[y^g_{b,t}], \qquad (2)$$
$$s.t.\ \lambda \geq 0,\ \mathbf{y}^g_{i,[t]} \neq \mathbf{y}^g_{j,[t]} \forall i \neq j$$

where λ is scalar controlling the strength of the diversity term. The full procedure to obtain diverse sequences using our method, Diverse Beam Search (DBS), is presented in Algorithm 1. It consists of two main steps for each group at each time step –

1) augmenting the $\log$ probabilities of each possible extension with the diversity term computed from previously advanced groups (Algorithm 1, Line 5) and,

2) running one step of a smaller BS with B' beams using the augmented $\log$ probabilities to extend the current group (Algorithm 1, Line 6).

Note that the first group ($g = 1$) is not 'conditioned' on other groups during optimization, so our method is guaranteed to perform at least as well as a beam search of size B'.

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     $Y^1_{[t]} \leftarrow \text{argmax}_{(\mathbf{y}^1_{1,[t]},\ldots,\mathbf{y}^1_{B',[t]})} \sum_{b\in[B']} \Theta(\mathbf{y}^1_{b,[t]})$
4     for $g = 2, \ldots G$ do
          // augment log probabilities with diversity penalty
5         $\Theta(\mathbf{y}^g_{b,[t]}) \leftarrow \Theta(\mathbf{y}^g_{b,[t]}) + \lambda\Delta(\bigcup_{h=1}^{g-1} Y^h_{[t]})[y^g_{b,t}] \quad b \in [B'], \mathbf{y}^g_{b,[t]} \in \mathcal{Y}^g_t$ and $\lambda > 0$
          // perform one step of beam search for the group
6         $Y^g_{[t]} \leftarrow \text{argmax}_{(\mathbf{y}^g_{1,[t]},\ldots,\mathbf{y}^g_{B',[t]})} \sum_{b\in[B']} \Theta(\mathbf{y}^g_{b,[t]}) \quad$ s.t. $\mathbf{y}_{i,[t]} \neq \mathbf{y}_{j,[t]} \ \forall i \neq j$
7 Return set of B solutions, $Y_{[T]} = \bigcup_{g=1}^{G} Y^g_{[T]}$
```

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 (non-greedily) 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 the work of Gimpel et al. (2013), who applied DivMBest to machine translation using beam search as a black-box inference algorithm. Specifically, in this approach, DivMBest knows nothing about the inner-workings of BS and simply makes B sequential calls to BS to generate B diverse solutions. This approach is *extremely* wasteful because BS is called B times, run from scratch every time, and even though each call to BS produces B solutions, only one solution is kept by DivMBest. In contrast, DBS 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 *a single run* of classical BS. One potential disadvantage of our method w.r.t. Gimpel et al. (2013) is that sentence-level diversity metrics cannot be incorporated in DBS since no group is complete when diversity is encouraged. However, as observed empirically by us and Li et al. (2015), initial words tend to disproportionally impact the diversity of the resultant sequences – suggesting that later words may not be important for diverse inference.

Diverse Decoding for RNNs. Efforts have been made by Li et al. (2015) and Li & Jurafsky (2016) to produce diverse decodings from recurrent models for conversation modeling and machine translation. Both of these works propose new heuristics for creating diverse M-Best lists and employ mutual information to re-rank lists of sequences. The latter achieves a goal separate from ours, which is simply to re-rank diverse lists.

Li & Jurafsky (2016) proposes a BS diversification heuristic that discourages beams from sharing common roots, implicitly resulting in diverse lists. Introducing diversity through a modified objective (as in DBS) rather than via a procedural heuristic provides easier generalization to incorporate different notions of diversity and control the exploration-exploitation trade-off as detailed in Section 5.1. Furthermore, we find that DBS outperforms the method of Li & Jurafsky (2016).

Li et al. (2015) introduced a novel decoding objective that maximizes mutual information between inputs and predicted outputs to penalize generic sequences. This operates on a principle orthogonal and complementary to DBS and Li & Jurafsky (2016). It works by penalizing utterances that are generally more frequent (diversity independent of input) rather than penalizing utterances that are similar to other utterances produced for the same input (diversity conditioned on input). Furthermore, the input-independent approach *requires training a new language model* for the target language while DBS just requires a diversity function Δ. Combination of these complementary techniques is left as interesting future work.

In other recent work, Wu et al. (2016) modify the beam search objective by introducing length-normalization to favor longer sequences and a coverage penalty that favors sequences that account for the complete input sequence. While the coverage term does not generalize to all neural sequence models, the length-normalization term can be implemented by modifying the joint-log-probability of each sequence. Although the goal of this method is not to produce diverse lists and hence not directly comparable, it is a complementary technique that can be used in conjunction with our diverse decoding method.

5 Experiments

In this section, we evaluate our approach on image captioning, machine translation, conversation and visual question generation tasks to demonstrate both its effectiveness against baselines and its general applicability to any inference currently supported by beam search. We also analyze the effects of DBS parameters, explore human preferences for diversity, and discuss diversity's importance in explaining complex images. We first explain the baselines and evaluations used in this paper.

Baselines & Metrics. Apart from classical beam search, we compare DBS with the diverse decoding method proposed in Li & Jurafsky (2016). We also compare against two other complementary decoding techniques proposed in Li et al. (2015) and Wu et al. (2016). Note that these two techniques are not directly comparable with DBS since the goal is not to produce diverse lists. We now provide a brief description of the comparisons mentioned:

- Li & Jurafsky (2016): modify BS by introducing an intra-sibling rank. For each partial solution, the set of $|\mathcal{V}|$ beam extensions are sorted and assigned intra-sibling ranks $k \in [|\mathcal{V}|]$ in order of decreasing log probabilities, $\theta_t(y_t)$. The log probability of an extension 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): train an additional unconditioned target sequence model $U(\mathbf{y})$ and perform BS decoding on an augmented objective $P(\mathbf{y}|x) - \lambda U(\mathbf{y})$, penalizing input-independent decodings.
- Wu et al. (2016) modify the beam-search objective by introducing length-normalization that favors longer sequences. The joint log-probability of completed sequences is divided by a factor, $(5 + |\mathbf{y}|)^\alpha / (5+1)^\alpha$, where $\alpha \in [0, 1]$.

We compare to our own implementations of these methods as none are publicly available. Both Li & Jurafsky (2016) and Li et al. (2015) develop and use re-rankers to pick a single solution from the generated lists. Since 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, simply sorting by $\log$-probability.

We evaluate the performance of the generated lists using the following two metrics:

- *Oracle Accuracy*: Oracle or top k accuracy w.r.t. some task-specific metric, such as BLEU (Papineni et al., 2002) or SPICE (Anderson et al., 2016), is the maximum value of the metric achieved over a list of k potential solutions. Oracle accuracy is an upper bound on the performance of any re-ranking strategy and thus measures the maximum potential of a set of outputs.
- *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 sequences have increased diversity without sacrificing fluency and correctness with respect to target tasks.

5.1 Sensitivity Analysis and Effect of Diversity Functions

Here we discuss the impact of the number of groups, strength of diversity , and various forms of diversity for language models. Note that the parameters of DBS (and other baselines) were tuned on a held-out validation set for each experiment. The supplement provides further discussion and experimental details.

Number of Groups (G). Setting $G{=}B$ allows for the maximum exploration of the search space, while setting $G{=}1$ reduces DBS to BS, resulting in increased exploitation of the search-space around the 1-best decoding. Empirically, we find that maximum exploration correlates with improved oracle accuracy and hence use $G{=}B$ to report results unless mentioned otherwise. See the supplement for a comparison and more details.

Diversity Strength (λ). The diversity strength λ specifies the trade-off between the model score and diversity terms. As expected, we find that a higher value of λ produces a more diverse list; however, very large values of λ can overpower model score and result in grammatically incorrect outputs. We set λ via grid search over a range of values to maximize oracle accuracies achieved on the validation set. 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). In Section 3, we defined $\Delta(\cdot)$ as a function over a set of partial solutions that outputs a vector of dissimilarity scores for all possible beam completions. Assuming that each of the previous groups influences the completion of the current group independently, we can simplify $\Delta(\bigcup_{h=1}^{g-1} Y_{[t]}^h)$ as the sum of each group's contributions as $\sum_{h=1}^{g-1} \Delta(Y_{[t]}^h)$. In Section 3, we illustrated a simple hamming diversity of this form that penalizes selection of tokens proportionally to the number of time it was used in previous groups. However, this factorized diversity term can take various forms in our model – with hamming diversity being the simplest. For language models, we study the effect of using cumulative (i.e. considering all past time steps), n-gram and neural embedding based diversity functions. Each of these forms encode differing notions of diversity and result in DBS outperforming BS. We find simple hamming distance to be effective and report results based on this diversity measure unless otherwise specified. More details about these forms of the diversity term are provided in the supplementary.

5.2 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 (with 200 held out images used to tune hyperparameters). We train a captioning model (Vinyals et al., 2015) using the `neuraltalk2`[1] code repository.

Results. Table 1 shows Oracle (top k) SPICE for different values of k. DBS consistently outperforms BS and Li & Jurafsky (2016) on both datasets. We observe that gains on PASCAL-50S are more pronounced (7.14% and 4.65% SPICE@20 improvements over BS and Li & Jurafsky (2016)) than COCO. This suggests diverse predictions are especially advantageous when there is a mismatch between training and testing sets, implying DBS may be better suited for real-world applications.

Table 1 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 on average. Moreover, on the PASCAL-50S test split we observe that DBS finds more likely top-1 solutions on average – DBS obtains an average maximum log probability of -6.53 opposed to -6.91 found by BS of the same beam width. This empirical evidence suggests that using DBS as a replacement to BS may lead to lower inference approximation error.

Table 1: Oracle accuracy and distinct n-grams on COCO and PASCAL-50S datasets for image captioning at $B = 20$. While we report SPICE, we observe similar trends in other metrics (reported in supplement).

Dataset	Method	Oracle Accuracy (SPICE)				Diversity Statistics			
		@1	@5	@10	@20	distinct-1	distinct-2	distinct-3	distinct-4
PASCAL-50S	Beam Search	4.933	7.046	7.949	8.747	0.12	0.57	1.35	2.50
	Li & Jurafsky (2016)	5.083	7.248	8.096	8.917	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**
	Wu et al. (2016)	5.301	7.322	8.236	8.832	0.16	1.10	3.16	6.45
	Li et al. (2015)	5.129	7.175	8.168	8.560	0.13	1.15	3.58	8.42
COCO	Beam Search	16.278	22.962	25.145	27.343	0.40	1.51	3.25	5.67
	Li & Jurafsky (2016)	16.351	22.715	25.234	27.591	0.54	2.40	5.69	8.94
	DBS	**16.783**	**23.081**	**26.088**	**28.096**	**0.56**	**2.96**	**7.38**	**13.44**
	Wu et al. (2016)	16.642	22.643	25.437	27.783	0.54	2.42	6.01	7.08
	Li et al. (2015)	16.749	23.271	26.104	27.946	0.42	1.37	3.46	6.10

Human Studies. To evaluate human preference between captions generated by DBS and BS, we perform a human study via Amazon Mechanical Turk using all 1000 images of PASCAL-50S. For each image, both DBS and standard BS captions are shown to 5 different users. They are then asked – *"Which of the two robots understands the image better?"* In this forced-choice test, DBS captions were preferred over BS 60% of the time by human annotators.

Is diversity *always* needed? While these results show that diverse outputs are important for systems that interact with users, is diversity *always* beneficial? While images with many objects (*e.g.*, a park or a living room) can be described in multiple ways, the same is not true when there are few objects (*e.g.*, a close up of a cat or a selfie). This notion is studied by Ionescu et al. (2016), which defines a "difficulty score": the human response time for solving a visual search task. On the PASCAL-50S dataset, we observe a positive correlation ($\rho = 0.73$) between difficulty scores and humans preferring DBS to BS. Moreover, while DBS is generally preferred by humans for 'difficult' images, both are about equally preferred on 'easier' images. Details are provided in the supplement.

5.3 Machine Translation

We use the WMT'14 dataset containing 4.5M sentences to train our machine translation models. We train stacking LSTM models as detailed in Luong et al. (2015), consisting of 4 layers and 1024-dimensional hidden states. While decoding sentences, we employ the same strategy to replace `UNK` tokens. We train our models using the publicly available `seq2seq-attn`[2] code repository. We report results on *news-test-2013* and *news-test-2014* and use the *news-test-2012* to tune the parameters of DBS. We use sentence level BLEU scores to compute oracle metrics and report distinct n-grams

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

[2] https://github.com/harvardnlp/seq2seq-attn

similar to image captioning. Results are shown in Table 2 and we again find that DBS consistently outperforms all baselines.

Table 2: Quantitative results on English-German translation on the newstest-2013 and newstest-2014 datasets combined (at $B = 20$).

Method	**Oracle Accuracy (BLEU-4)**				**Diversity Statistics**			
	@1	@5	@10	@20	distinct-1	distinct-2	distinct-3	distinct-4
Beam Search	20.5	22.4	23.8	24.2	0.04	0.75	2.10	3.23
Li & Jurafsky (2016)	20.7	22.6	24.0	24.3	0.04	0.81	2.92	4.61
DBS	**20.8**	**22.9**	**24.4**	**24.8**	**0.06**	**0.95**	**3.67**	**5.54**
Wu et al. (2016)	20.6	22.6	24.3	24.6	0.05	0.88	2.83	4.50
Li et al. (2015)	20.7	23.1	24.4	24.6	0.04	0.86	2.76	4.31

5.4 Dialog Generation

Dialog generation is a task that is inherently diverse as there are multiple valid ways to respond to a statement. We train a seq2seq model consisting of LSTMs as in Vinyals & Le (2015) on the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil & Lee, 2011) using the `neuralconvo`[3] repository. The training dataset consists of 222,182 conversational exchanges between movie characters. Since automatic evaluation of dialog generation responses is an open research problem with existing metrics being poorly correlated with human judgement (Liu et al., 2016), we show qualitative results to demonstrate the effectiveness of DBS. Table 3 compares BS and DBS at B=3.

Table 3: Qualitative examples of dialog generation comparing best-first search, BS and DBS

Prompt	**Beam Search**	**Diverse Beam Search**
Why do you want to be the president?	I don't know. I don't know... I don't know, but I am the best	I don't know. Because you know! Because, I have another life.
It's such a beautiful day!	I know! It's true. Yeah.	I know. Well... That's not true, honey.
What's your name?	Jacob. Berger, my man. Berger, Thomas.	Jacob. Berger, darling My mother used to hum that to me.

5.5 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 is interesting because it requires reasoning about multiple problems that are central to vision – *e.g.*, object attributes, relationships between objects, and natural language. Furthermore, many questions could make sense for one image, so it is important that lists of generated questions be diverse.

We use the VQA dataset (Antol et al., 2015) to train a model similar to image captioning architectures. Instead of captions, the training set now consists of 3 questions per image. Similar to previous results, using beam search to sample outputs results in similarly worded questions (see Fig. 3) and DBS brings out new details captured by the model. Counting the number of *types* of questions generated (as defined by Antol et al. (2015)) allows us to measure this diversity. We observe that the number of question types generated per image *increases* from 2.3 for BS to 3.7 for DBS (at $B = 6$).

6 Conclusion

Beam search is widely a used approximate inference algorithm for decoding sequences from neural sequence models; 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

[3] `https://github.com/macournoyer/neuralconvo`

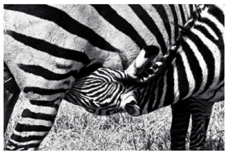
Input Image	Beam Search	Diverse Beam Search
	What sport is this? What sport is being played? What color is the man's shirt? What color is the ball? What is the man wearing? What color is the man's shorts?	What color is the man's shirt? What is the man holding? What is the man wearing on his head? Is the man wearing a helmet What is the man in the white shirt doing? Is the man in the background wearing a helmet?
	How many zebras are there? How many zebras are in the photo? How many zebras are in the picture? How many animals are there? How many zebras are shown? What is the zebra doing?	How many zebras are there? How many zebras are in the photo? What is the zebra doing? What color is the grass? Is the zebra eating? Is the zebra in the wild?

Figure 3: Qualitative results on Visual Question Generation. DBS generates questions that are non-generic and belong to different question types.

inherent ambiguity like many involving language. In this work, we modify Beam Search with a diversity-augmented sequence decoding objective to produce *Diverse Beam Search*. We develop a 'doubly greedy' approximate algorithm to minimize this objective and produce diverse sequence decodings. Our method consistently outperforms beam search and other baselines across all our experiments without *extra computation* or *task-specific overhead*. DBS is *task-agnostic* and can be applied to any case where BS is used, which we demonstrate in multiple domains. Our implementation available at https://github.com/ashwinkalyan/dbs.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 6

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. 1, 8

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

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. 2, 4

Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011. 8

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. 2

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. 1

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. 1, 5, 12

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. abs/1303.5778, 2013. 1

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. 7

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, et al. Smart reply: Automated reeponse suggestion for email. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016. 2

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. 2, 7

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. 2, 4

Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016. 2, 5, 6, 7, 8, 13, 14

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. 2, 5, 6, 7, 8, 13, 14

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. 7

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. 2016. URL http://arxiv.org/abs/1603.08023. 8

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015. 7

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. 12

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. 8

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. 6

Dennis Park and Deva Ramanan. N-best maximal decoders for part models. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011. 4

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. 2, 4

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. 7

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. 1

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015. 1, 8

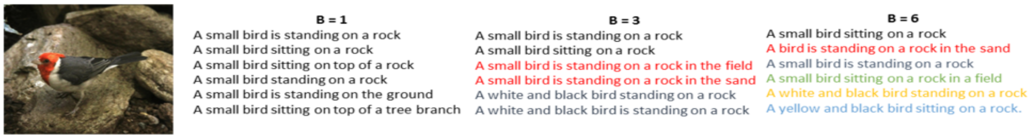
Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 7

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 5, 6, 7, 8, 13, 14

APPENDIX

SENSIVITY STUDIES

Number of Groups. Fig. 4 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.



B = 1	B = 3	B = 6
A small bird is standing on a rock	A small bird is standing on a rock	A small bird sitting on a rock
A small bird sitting on a rock	A small bird sitting on a rock	A bird is standing on a rock in the sand
A small bird sitting on top of a rock	A small bird is standing on a rock in the field	A small bird is standing on a rock
A small bird standing on a rock	A small bird is standing on a rock in the sand	A small bird sitting on a rock in a field
A small bird is standing on the ground	A white and black bird standing on a rock	A white and black bird standing on a rock
A small bird sitting on top of a tree branch	A white and black bird is standing on a rock	A yellow and black bird sitting on a rock.

Figure 4: 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. 5a shows a grid search of λ for image-captioning on the PASCAL-50S dataset.

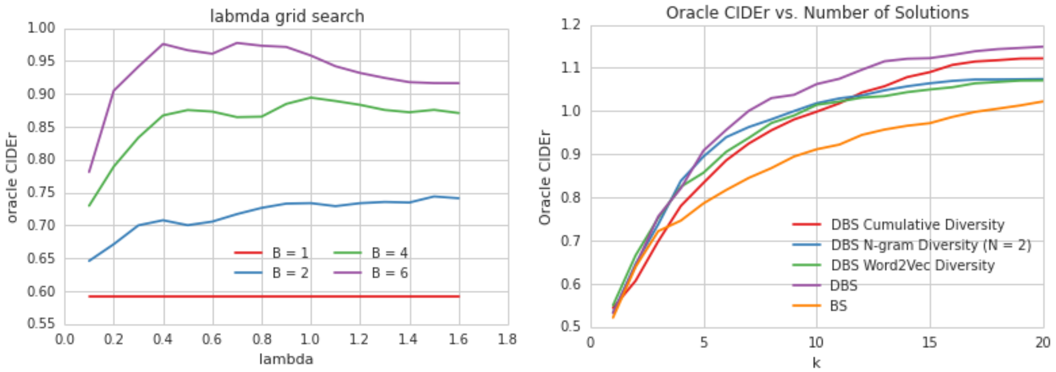
Choice of Diversity Function. The diversity function can take various forms ranging from simple hamming diversity to neural embedding based diversity. We discuss some forms for language modelling below:

- *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, $\Delta(Y^h_{[t]})[y^g_{[t]}] = \exp\{-(\sum_{\tau \in t} \sum_{b \in B'} I[y^h_{b,\tau} \neq y^g_{b,\tau}])/\Gamma\}$ 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. Fig. 5b shows the oracle performace of various forms of the diversity function described in Section 5.1. 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.

IMAGE CAPTIONING EVALUATION

While we report oracle SPICE values in the paper, our method consistently outperforms baselines and classical BS on other standard metrics such as CIDEr (Table 4), METEOR (Table 5) and ROUGE (Table 6). We provide these additional results in this section.



(a) Grid search of diversity strength parameter

(b) Effect of multiple forms for the diversity function

Figure 5: Fig. 5a 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. 5b 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.

Table 4: CIDEr Oracle accuracy on COCO and PASCAL-50S datasets for image captioning at $B = 20$.

Dataset	**Method**	**Oracle Accuracy (CIDEr)**			
		@1	@5	@10	@20
PASCAL-50S	Beam Search	53.79	83.94	96.70	107.63
	Li & Jurafsky (2016)	54.61	85.21	99.80	110.64
	DBS	**57.82**	**89.38**	**103.75**	**113.43**
	Wu et al. (2016)	47.77	72.12	84.64	105.66
	Li et al. (2015)	49.80	81.35	96.87	107.37
COCO	Beam Search	87.27	121.74	133.46	140.98
	Li & Jurafsky (2016)	**91.42**	111.33	116.94	119.14
	DBS	86.88	**123.38**	**135.68**	**142.88**
	Wu et al. (2016)	87.54	122.06	133.21	139.43
	Li et al. (2015)	88.18	124.20	138.65	150.06

Table 5: METEOR Oracle accuracy on COCO and PASCAL-50S datasets for image captioning at $B = 20$.

Dataset	**Method**	**Oracle Accuracy (METEOR)**			
		@1	@5	@10	@20
PASCAL-50S	Beam Search	12.24	16.74	19.14	21.22
	Li & Jurafsky (2016)	13.52	17.65	19.91	21.76
	DBS	**13.71**	**18.45**	**20.67**	**22.83**
	Wu et al. (2016)	13.34	17.20	18.98	21.13
	Li et al. (2015)	13.04	17.92	19.73	22.32
COCO	Beam Search	24.81	28.56	30.59	31.87
	Li & Jurafsky (2016)	24.88	29.10	31.44	33.56
	DBS	**25.04**	**29.67**	**33.25**	**35.42**
	Wu et al. (2016)	24.82	28.92	31.53	34.14
	Li et al. (2015)	24.93	30.11	32.34	34.88

Modified SPICE evaluation. To measure both the quality and the diversity of the generated captions, we compute SPICE-score by comparing the graph union of all the generated hypotheses with the ground truth scene graph. This measure rewards all the relevant relations decoded as against oracle accuracy that compares to relevant relations present only in the top-scoring caption. We observe that DBS outperforms both baselines under this measure with a score of 18.345 as against a score of 16.988 (beam search) and 17.452 (Li & Jurafsky, 2016).

Table 6: ROUGE Oracle accuracy on COCO and PASCAL-50S datasets for image captioning at $B = 20$.

Dataset	Method	Oracle Accuracy (ROUGE-L)			
		@1	@5	@10	@20
PASCAL-50S	Beam Search	45.23	56.12	59.61	62.04
	Li & Jurafsky (2016)	46.21	56.17	60.15	62.95
	DBS	**46.24**	**56.90**	**60.35**	**63.02**
	Wu et al. (2016)	43.73	52.29	56.49	61.65
	Li et al. (2015)	44.12	54.67	57.34	60.11
COCO	Beam Search	52.46	58.43	62.56	65.14
	Li & Jurafsky (2016)	52.87	59.89	63.45	65.42
	DBS	**53.04**	**60.89**	**64.24**	**67.72**
	Wu et al. (2016)	52.13	58.26	62.89	65.77
	Li et al. (2015)	53.10	59.32	63.04	66.19

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. 6. The lists were shuffled to guard the task from being gamed by a turker.

Table 7: Frequency table for image difficulty and human preference for DBS captions on PASCAL50S dataset

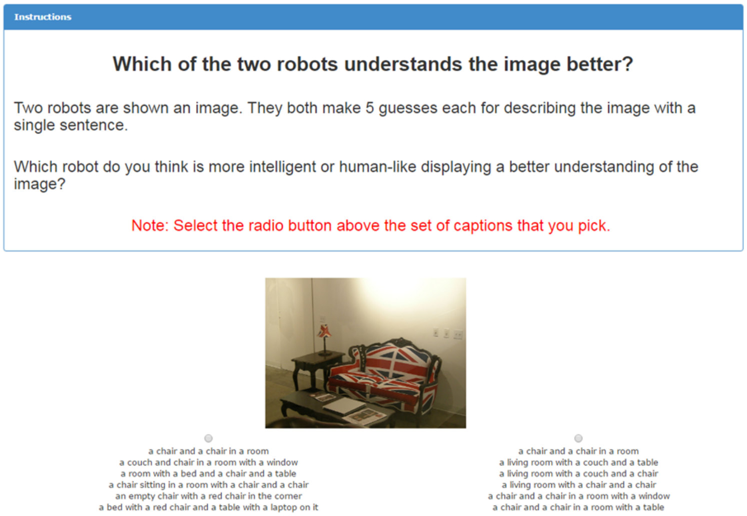
difficulty score bin range	# images	% images DBS was preffered
$\leq \mu - \sigma$	481	50.51%
$[\mu{-}\sigma, \mu{+}\sigma]$	409	69.92%
$\geq \mu + \sigma$	110	83.63%

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 7. 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. 7. More general qualitative examples are provided in Fig. 8.

Discussion

Are longer sentences better? Many recent works propose a scoring or a ranking objective that depends on the sequence length. These favor longer sequences, reasoning that they tend to have more details and resulting in improved accuracies. We measure the correlation between length of a sequence and its accuracy (here, SPICE) and observe insignificant correlation between SPICE and sequence length. On the PASCAL-50S dataset, we find that BS and DBS have are negatively correlated ($\rho = -0.003$ and $\rho = -0.015$ respectively), while (Li & Jurafsky, 2016) is correlated positively ($\rho = 0.002$). Length is not correlated with performance in this case.

Efficient utilization of beam budget. In this experiment, we emperically show that DBS makes efficient use of the beam budget in exploring the search space for better solutions. Fig. 9 shows the variation of oracle SPICE (@B) with the beam size. At really high beam widths, all decoding techniques achieve similar oracle accuracies. However, diverse decoding techniques like DBS achieve the same oracle at much lower beam widths. Hence, DBS not only produces sequence lists that are significantly different but also efficiently utilizes the beam budget to decode better solutions.



Instructions

Which of the two robots understands the image better?

Two robots are shown an image. They both make 5 guesses each for describing the image with a single sentence.

Which robot do you think is more intelligent or human-like displaying a better understanding of the image?

Note: Select the radio button above the set of captions that you pick.

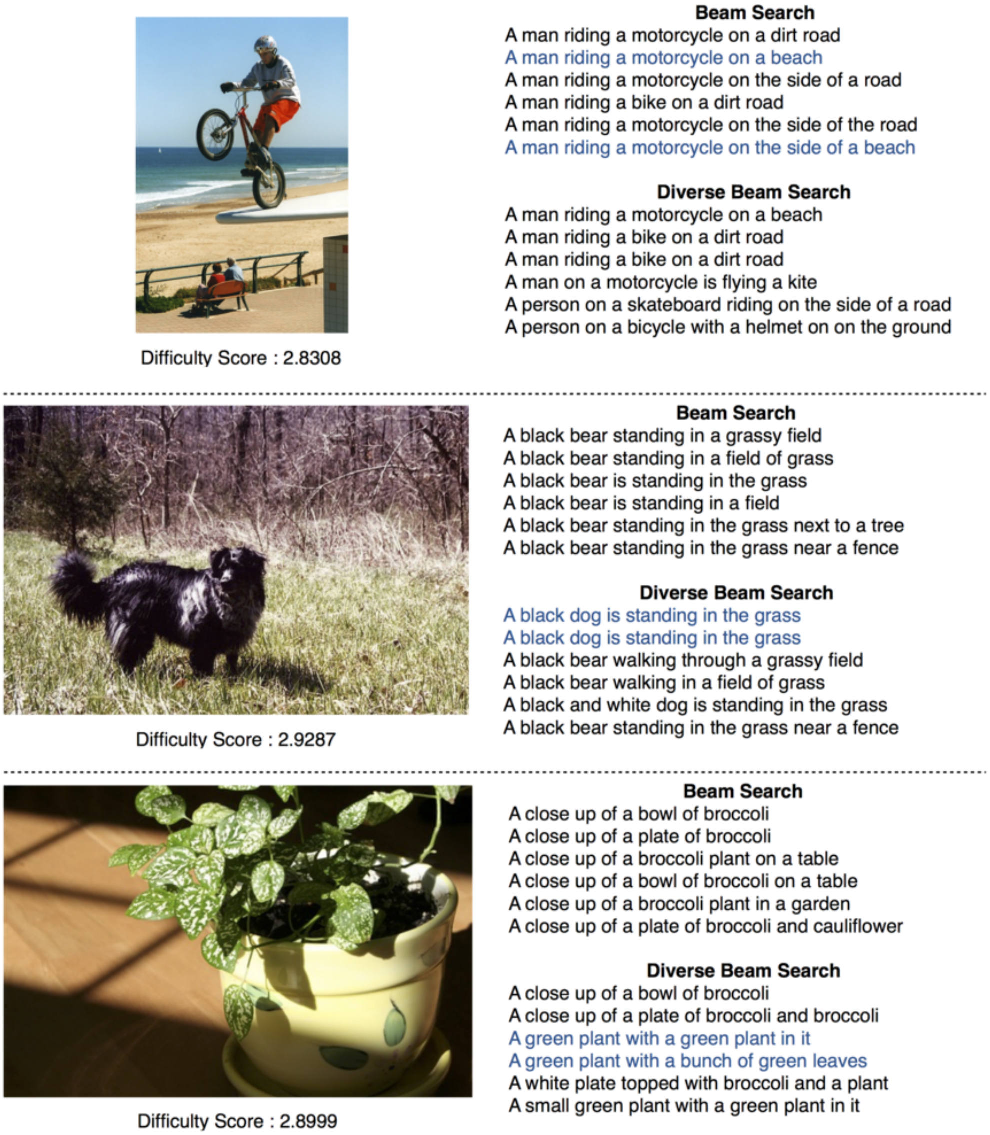
○

a chair and a chair in a room
a couch and chair in a room with a window
a room with a bed and a chair and a table
a chair sitting in a room with a chair and a chair
an empty chair with a red chair in the corner
a bed with a red chair and a table with a laptop on it

○

a chair and a chair in a room
a living room with a couch and a table
a living room with a couch and a chair
a living room with a chair and a chair
a chair and a chair in a room with a window
a chair and a chair in a room with a table

Figure 6: Screen-shot of the interface used to perform human studies



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 : 2.8308

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 : 2.9287

Beam Search

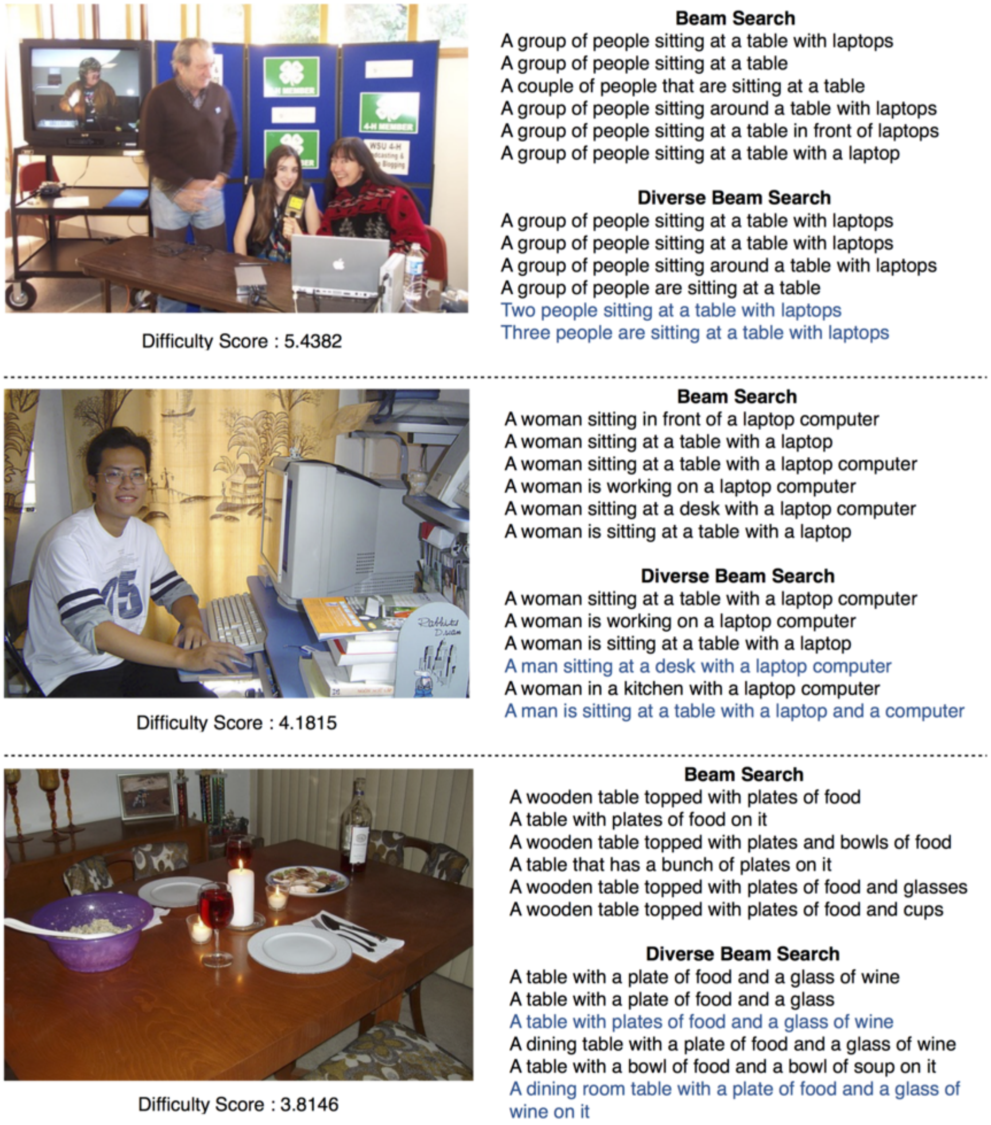
A close up of a bowl of broccoli
A close up of a plate of broccoli
A close up of a broccoli plant on a table
A close up of a bowl of broccoli on a table
A close up of a broccoli plant in a garden
A close up of a plate of broccoli and cauliflower

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

Difficulty Score : 2.8999

Figure 7: 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

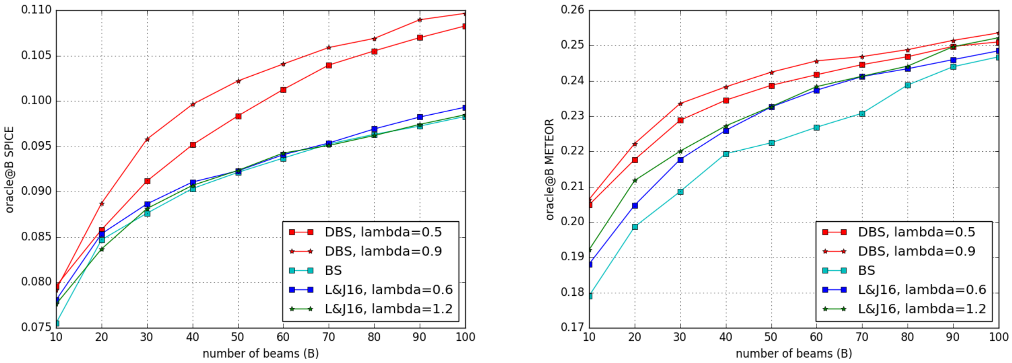
Beam Search

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



(a) Oracle SPICE (@B) vs B

(b) Oracle METEOR (@B) vs B

Figure 9: As the number of beams increases, all decoding methods tend to achieve about the same oracle accuracy. However, diverse decoding techniques like DBS utilize the beam budget efficiently achieving higher oracle accuracies at much lower beam budgets.