Switching to Discriminative Image Captioning
by Relieving a Bottleneck of Reinforcement Learning

Ukyo Honda¹,² Taro Watanabe³ Yuji Matsumoto²
¹CyberAgent, Inc. ²RIKEN ³Nara Institute of Science and Technology
honda_ukyo@cyberagent.co.jp taro@is.naist.jp yuji.matsumoto@riken.jp

Abstract

Discriminativeness is a desirable feature of image captions: captions should describe the characteristic details of input images. However, recent high-performing captioning models, which are trained with reinforcement learning (RL), tend to generate overly generic captions despite their high performance in various other criteria. First, we investigate the cause of the unexpectedly low discriminativeness and show that RL has a deeply rooted side effect of limiting the output words to high-frequency words. The limited vocabulary is a severe bottleneck for discriminativeness as it is difficult for a model to describe the details beyond its vocabulary. Then, based on this identification of the bottleneck, we drastically recast discriminative image captioning as a much simpler task of encouraging low-frequency word generation. Hinted by long-tail classification and debiasing methods, we propose methods that easily switch off-the-shelf RL models to discriminativeness-aware models with only a single-epoch fine-tuning on the part of the parameters. Extensive experiments demonstrate that our methods significantly enhance the discriminativeness of off-the-shelf RL models and even outperform previous discriminativeness-aware methods with much smaller computational costs. Detailed analysis and human evaluation also verify that our methods boost the discriminativeness without sacrificing the overall quality of captions.¹

1. Introduction

Image captioning plays a fundamental role at the intersection of computer vision and natural language processing by converting the information in images into natural language descriptions. Generated captions can be used in various downstream tasks: aiding visually impaired users [19], visual question answering on images and videos [16, 31], visual dialogue [68], and news generation [79].

¹The code will be made available at https://github.com/ukyohn/switch_disc_caption.git

Figure 1. Caption examples in the MS COCO validation set. Transformer RL is a Transformer captioning model trained with RL and wFT is our fine-tuning method. Transformer RL generates exactly the same caption for the four images. The underlined words indicate the characteristic information that are not mentioned by Transformer RL, and the blue words are those that have never appeared in the outputs of the model. See Appendix 2 for more examples.

For those downstream tasks, captions should be discriminative: captions should describe the characteristic and important details of the input images [51]. However, current captioning models tend to generate overly generic captions [12, 11, 64, 66]. In particular, models trained with the standard reinforcement learning (RL) [50], which is the de facto standard training method in current image captioning [55], unexpectedly perform poorly in discriminativeness despite the significant advantages in various other criteria [39, 62]. For example, a high-performing Transformer [57] captioning model trained with RL generates exactly the same caption for the four different images shown in Figure 1, ignoring the other salient details of each image.

To address the problem of overly generic captions, studies have been intensely conducted on discriminative image captioning, which is also called distinctive image captioning or descriptive image captioning. Previous research has created new RL rewards regarding discriminativeness or new model architectures to enhance discriminativeness. These approaches improved the discriminativeness; however, their models come with additional computations, re-
quire retraining from scratch, and do not shed light on the cause of existing models’ low discriminativeness.

Instead of creating or paying those computational costs, we first analyze the cause of the unexpectedly low discriminativeness of off-the-shelf RL models, i.e., pre-trained, existing RL models, to explore ways to improve their discriminativeness. **Our first contribution is the identification of a deeply rooted side effect in RL that limits output words to high-frequency words.** The limited vocabulary is a severe bottleneck for discriminativeness as it is difficult for a model to describe the details beyond its vocabulary.

Given this identification of the bottleneck, now we can directly address the bottleneck by simply encouraging the generation of low-frequency words. This task relaxation allows us to introduce long-tail classification and debiasing methods to discriminative image captioning for the first time. **Our second contribution is our effective and efficient methods that switch any off-the-shelf RL models to discriminativeness-aware models with only a single-epoch fine-tuning on the part of the parameters.** Unlike previous approaches, our methods do not require any discriminativeness rewards, new model architectures, or retraining from scratch.

Extensive experiments demonstrate that increasing low-frequency words in outputs significantly boosts discriminativeness from off-the-shelf RL models and even outperforms previous discriminativeness-aware models with much smaller computational costs. These results verify that the limited vocabulary of RL models has been the major cause of their low discriminativeness. Detailed analysis and human evaluation also show that our methods enhance the discriminativeness without sacrificing the overall quality. We believe that our novel findings on the cause of low discriminativeness and the practical solutions to it will significantly impact future research on discriminative image captioning.

## 2. Discriminativeness and a Bottleneck of RL

Currently, RL is the de facto standard training method for models used in image captioning because it significantly improves the performance in various evaluation metrics [55]. However, it does not improve discriminativeness and may even decrease it [39, 62]. In this section, we examine the cause of the unexpectedly low discriminativeness.

### 2.1. RL in Image Captioning

We provide a brief overview of the standard RL algorithm used in image captioning. It was proposed by [48] and refined by [50]. Their goal was to directly optimize non-differentiable test-time metrics by minimizing the negative expected reward:

$$
\mathcal{L}_{RL}(\theta) = -\mathbb{E}_{w^* \sim p_\theta(w^* | I)}[r(w^*)],
$$

where $w^* = (w_1^*, ..., w_T^*)$ is a sequence sampled from a policy $p_\theta$, $I$ is an input image, and $r(\cdot)$ is a reward function. To compute the gradient of $\mathcal{L}(\theta)$, [48] applied the REINFORCE algorithm [69] to text generation. The algorithm approximates the gradient as follows:

$$
\nabla_\theta \mathcal{L}_{RL}(\theta) \approx -(r(w^*) - b) \nabla_\theta \log p_\theta(w^* | I).
$$

Here, $b$ is a baseline reward that reduces the variance in the gradient. Typically, the reward function $r(\cdot)$ is CIDEr [59], and the baseline reward $b$ is a reward for a sequence sampled with greedy decoding [50].

### 2.2. RL Limits Vocabulary

Despite its effectiveness, RL has been found not to improve discriminativeness and somehow decrease the number of unique n-grams in output captions [39, 62]. As the relation between RL and these two negative effects is not obvious, it has been just considered a curious case.

We elucidate for the first time the relation between RL and limited vocabulary by combining two recent findings.

(1) RL has been shown to make the output distribution peaky [8, 30]. RL samples sequences from policy $p_\theta$ (See Eq. (1)). Typically, $p_\theta$ is initialized with a text-generation model pre-trained with the Cross-Entropy (CE) loss on ground-truth text. In text generation, however, the initialized $p_\theta$ outputs peaky distributions, and thus, RL samples and rewards the words at the peak only, shaping more peaky distributions [8]. Then, where does $p_\theta$ tend to be peaky?
Text-generation models have been theoretically and empirically shown to output distributions peaky at high-frequency words in the training corpus \cite{46, 49, 13, 24}. These two findings conclude that RL shifts the probability mass from low-frequency words to high-frequency words by only sampling and rewarding the latter.

Figure 2 confirms the above by plotting the relative frequency of the words sampled for the training images. The words are sorted by their frequency in ground-truth captions and divided into 200 bins. Compared to the ground-truth captions and the sequences sampled with a CE model, the sequences sampled with an RL model are clearly limited to the high-frequency words, forming a peaky distribution\(^2\).

### 2.3. Vocabulary Limits Discriminativeness

Neural captioning models typically generate captions using sequential vocabulary-size classification \cite{61}. However, the actual vocabulary a model can generate is much smaller than the entire vocabulary as the output distribution is highly skewed towards high-frequency words. If the actual vocabulary cannot cover the details of an image, the model is forced to avoid those details and output only the information that high-frequency words can describe. For example, the blue words in Figure 1 are not in the actual vocabulary of the RL model; these words have never been generated during evaluation. As a result, the RL model had to ignore the characteristic relations tied and docked and ended up describing exactly the same for all four images.

Based on the observations, now we can hypothesize that the unexpectedly low discriminativeness of RL models has been rooted in the limited vocabulary. This identification of the bottleneck is a key contribution as it allows us to address the low discriminativeness directly at the root.

### 3. Methods to Relieve the Bottleneck

We have shown that RL results in the limited vocabulary as it steals the probability mass from low-frequency words. Thus, increasing those low-frequency words is the easy yet critical solution to the bottleneck. One way to achieve this is to jointly optimize both the RL loss and the CE loss on ground-truth captions so that the low-frequency words in ground-truth captions would be more likely to be sampled during RL training \cite{64}. However, this approach still relies on the sampling from a skewed policy and requires retraining from scratch.

To increase the actual vocabulary more effectively and efficiently, we refine the mapping from encoded features to low-frequency words. This refinement can be applied to any RL models and can be achieved by modifying only the mapping function parameters with a single-epoch fine-tuning.

\(^2\)Although Figure 2 shows only the results obtained with the Transformer captioning model, we also confirmed that other models output peaky distributions \cite{50, 2}. See Appendix 3 for the details.

### 3.1. Simple Fine-Tuning (sFT)

The first method is a simple fine-tuning (sFT). It is based on a decoupled two-stage training \cite{27}, which is a current strong baseline model for long-tail classification \cite{56, 44, 65}. \cite{27} decoupled the learning procedure into representation learning and classification, and then found that classification, i.e., the mapping from representations to label distributions, is critical for long-tail classification. They decoupled the classification model \(f_\theta(\cdot)\) into an encoder \(g_\theta(\cdot)\) and a classifier consisting of weight and bias parameters: \(f_\theta(x) = W g_\theta(x) + b\). Representation learning is the first stage of training, where they trained the entire classification model \(f_\theta(\cdot)\) on a full training dataset. The second stage is classification, where they fixed the encoder parameters \(\theta_e\) and adjusted only the classifier parameters. For the second-stage adjustment, they applied class-balanced sampling to encourage learning on low-frequency labels.

Following \cite{27}, we decouple a captioning model into an encoder and a classifier. In image captioning, the first-stage training of \cite{27} corresponds to RL training on the full training dataset. The second-stage training corresponds to adjusting the classifier parameters on the vocabulary-balanced sequences. However, sampling from the skewed policy of text-generation models cannot provide sequences containing low-frequency words (Section 2.2). Thus, we use ground-truth captions as relatively vocabulary-balanced samples. sFT simply fine-tunes the classifier parameters of a pre-trained RL captioning model by minimizing the CE loss on ground-truth captions:

\[
\mathcal{L}_{CE}(\hat{\theta}) = -\frac{1}{T} \sum_{t=1}^{T} \log p_\hat{\theta}(w^g_t | w^g_{<t}, I),
\]

where \(w^g = (w^g_1, ..., w^g_T)\) is a ground-truth caption of image \(I\) and \(\hat{\theta}\) denotes the model parameters \(\theta\) that are initialized with RL training. Let the softmax function \(\phi(\cdot)\) be

\[
\phi_{\omega_{\gamma}, \beta}(z) = \frac{\exp(\beta z_{\omega_i})}{\sum_{w_i \in W} \exp(\beta z_{w_i})},
\]

where \(z_{w_i}\) indicates the element of a vector \(z \in \mathbb{R}^{|W|}\) at the index of a word \(w_i \in W\). \(W\) is the entire vocabulary. \(\beta\) is an inverse-temperature hyperparameter that controls the steepness of the softmax distribution. Then, the conditional probability \(p_\theta(w^g_t | w^g_{<t}, I)\) is computed as follows:

\[
p_\theta(w^g_t | w^g_{<t}, I) = \phi_{\omega_{\gamma}, \beta}(s_\theta(w^g_t, I)),
\]

\[
s_\theta(w^g_t, I) = W^T g_\theta(w^g_{<t}, I) + b,
\]

where \(W \in \mathbb{R}^{d \times |W|}\) and \(b \in \mathbb{R}^{|W|}\), \(d\) is the dimension of the hidden states of an encoder \(g_\theta(\cdot)\). We use LSTM \cite{23} or Transformer \cite{57} for \(g_\theta(\cdot)\). During fine-tuning, only the classifier parameters \{\(W, b\)\} \(\in \hat{\theta}\) are updated with the gradients \(\nabla_W \mathcal{L}_{CE}(\hat{\theta})\) and \(\nabla_b \mathcal{L}_{CE}(\hat{\theta})\), respectively.
3.2. Weighted Fine-Tuning (wFT)

Ground-truth captions contain more low-frequency words than sampled sequences, but some low-frequency words are still difficult to learn because of their low frequency. Our second method is weighted fine-tuning (wFT), which further pursues vocabulary balance by rebalancing the loss of high-frequency words and low-frequency words in ground-truth captions.

To rebalance the loss, we exploit the frequency bias of RL models: RL models overly assign the probability to high-frequency words but not to low-frequency words. Given the properties of the frequency bias, fine-tuning for discriminativeness should focus more on the words that an RL model is not confident of but should be avoided on the words that an RL model is confident of. wFT incorporates these heuristics by modifying the probability $p_{θ}$ of $L_{CE}$ to the bias product (BP) \[9, 20, 22\] probability, $p_{θ,θ'}$:

\[
\begin{align*}
p_{θ,θ'}(w_{t}^{θ} | w_{<t}^{θ}, I) &= \phi_{w_{t,1}} \left[ \log \phi_{θ}(s_{θ}(w_{t}^{θ}, I)) + \log ϕ_{θ'}(s_{θ'}(w_{t}^{θ'}, I)) \right], \\
p_{θ}(· | w_{<t}^{θ}, I) &= p_{θ'}(· | w_{<t}^{θ'}, I)
\end{align*}
\]

where $ϕ_{θ}(z) ∈ \mathbb{R}^{|W|}$. By inserting $p_{θ,θ'}$ into $L_{CE}$, we define the objective function of wFT as follows:

\[
L_{BP}(θ) = -\frac{1}{T} \sum_{t=1}^{T} \log p_{θ,θ'}(w_{t}^{θ} | w_{<t}^{θ}, I).
\]

Similar to sFT, the parameters θ and θ' are initialized with the same RL model to be $θ$ and $θ'$. The difference is that, although the classifier parameters of $θ$ are updated, all the parameters of $θ'$ are fixed during fine-tuning\(^5\). Figure 3 shows the change in the BP loss compared to the CE loss. The BP severely suppresses the loss when the frequency-biased policy $p_{θ'}$ is confident, and largely increases the loss when $p_{θ'}$ is not confident. In this way, the BP allows models to unlearn the frequency bias learned with RL. As with sFT, only the classifier parameters $\{W, b\} ∈ θ$ are updated with the gradients $\nabla_{W}L_{BP}(θ)$ and $\nabla_{b}L_{BP}(θ)$, respectively.

The previous BP methods used the probability $p_{θ}$ during evaluation to avoid incorporating the bias of $P_{θ'}$ into the predictions \[9, 20\]. Although it worked well in their classification tasks, we found this train–test gap makes the decoding unstable in text generation. To mitigate the train–test gap, we use two variants of decoding: (1) decode with $p_{θ}$ but use a small $β'$ for $p_{θ'}$ during training to ease the gap between $P_{θ}$ and $P_{θ,θ'}$, or (2) use $p_{θ,θ'}$ during both training and decoding (BP decoding) as $P_{θ,θ'}$ itself is already less biased than $P_{θ'}$.

4. Experiments

4.1. Setup

Dataset and Metrics. We used the MS COCO captioning dataset\(^4\) [38, 6] with Karpathy splitting [28]. After preprocessing, the entire vocabulary size $|W|$ was 9,487\(^5\).

In the evaluation, the captions were decoded using a beam search of size 5 and evaluated using various evaluation metrics. Specifically, we used CIDEr [59] SPICE [1], Improved BERTScore (BERTS+) [74], TIGER [25], CLIPScore (CLIPS), and RefCLIPScore (RefCLIPS) [21]. Note that the correlation with human judgments increases in the above order, with RefCLIPS indicating the state-of-the-art correlation [21, 29]. Following the previous studies \[39, 62, 54\], we evaluated discriminativeness with R@K scores: the percentage of captions with which a pre-trained image–text retrieval model \[15\] could correctly retrieve the original images from the entire validation/test images within the rank of $K ∈ \{1, 5, 10\}$. A higher R@K indicates

\(^{5}\)\(^{5}\)\(^{5}\) Each split of training/validation/test contained 113,287/5,000/5,000 images, and each image had around five ground-truth captions.

\(^{5}\) The words that occur less than five times in the training captions were converted to \{\text{unk}\} token.
that the model generates more discriminative captions with characteristic information of images. Evaluation was conducted in a single run for each model. See Appendix 4 for the libraries and settings we used for these evaluations.

Comparison Models. Following [62], we used Att2in [50], UpDown [2], and Transformer [57] as the baseline models. The models were pre-trained with the standard RL [50] and are publicly available. In addition to the baseline models, we compared our models with state-of-the-art discriminativeness-aware models: CIDErBtw [62], NLI [54], DiscCap [41], and Visual Paraphrase [39]. The first three created new discriminativeness rewards to be optimized with RL. Visual Paraphrase introduced a new model architecture to paraphrase simpler captions to more complex captions. See Section 5 for more details of these models. As we mentioned in the beginning of Section 3, the CE loss on ground-truth captions can be utilized in a different way from our methods. We report the results of jointly optimizing the RL loss and CE loss (Joint CE) [62, 54, 14]). It optimizes $\mathcal{L}_{\text{Joint}}(\theta) = \lambda \mathcal{L}_{\text{RL}}(\theta) + (1 - \lambda) \mathcal{L}_{\text{CE}}(\theta)$ during RL training. We also tested Only CE, which sets $\lambda = 0$ to solely optimize the CE loss, as the baseline without RL. See Appendices 12 and 13 for more comparisons [76, 36, 7].

Hyperparameters. Our models used the same hyperparameters as the baseline models, except for the epoch size, learning rate, and $\beta'$ in Eq. (7). We set the epoch size for fine-tuning to 1 and searched for the best learning rate from $\{1e-3, 1e-4, 1e-5, 1e-6\}$. For BP in Eq. (7), we set $\beta = 1$ and searched for the best $\beta'$ from $\{0.1, 1\}$. As with our models, we set all hyperparameters of the CE-based models to the same as the baseline models except for the $\lambda \in \{0, 0.2, 0.5, 0.8\}$. We disabled scheduled sampling [4] for our fine-tuning and the CE loss to separate them from the RL loss strictly. We took the best hyperparameters according to the R@1 scores in the validation set. Note that we used different hyperparameters for the wFT with different decoding methods (See Section 3.2). Appendix 5 shows the best hyperparameters. We followed the previous work for the hyperparameters of the other models.

All the models except Visual Paraphrase had the same size of trainable parameters as their baselines. See Appendix 6 for the exact number of parameters. Our fine-tuning was completed in around 10 minutes using a single GPU of 16 GB memory. See Appendix 7 for the exact time for training and comparison with other methods.

### 4.2. Comparison with Baseline Models and Discriminativeness-Aware Models

Table 1 shows the results compared to those obtained with the baseline models and state-of-the-art discriminativeness-aware models.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Standard Evaluation</th>
<th>Discriminativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique-I</td>
<td>Unique-S</td>
<td>Length</td>
</tr>
<tr>
<td>Att2in RL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ sFT</td>
<td>1,197</td>
<td>21.5</td>
</tr>
<tr>
<td>+ wFT</td>
<td>1,102</td>
<td>21.5</td>
</tr>
<tr>
<td>+ wFT (BP decoding)</td>
<td>470</td>
<td>21.5</td>
</tr>
<tr>
<td>CIDErBtw</td>
<td>NLI</td>
<td></td>
</tr>
<tr>
<td>DiscCap$^*$</td>
<td>Joint CE</td>
<td></td>
</tr>
<tr>
<td>Only CE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Paraphrase</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparison of baseline models, our models, and state-of-the-art discriminativeness-aware models. Automatic evaluation results on the MS COCO test set. Unique-I and Unique-S indicate the number of unique unigrams and sentences, respectively. Length is the average length of the output captions. Scores with | were reported in [39]. Other scores were reproduced by us.

6https://github.com/ruotianluo/self-critical.py
torch: {Att2in, UpDown, Transformer}+self_critical
wFT) successfully increase the actual vocabulary size: both of them considerably increased Unique-I compared to all the baseline models. wFT increased the vocabulary more than sFT, indicating that rebalancing the loss further encouraged low-frequency word generation. The increased vocabulary resulted in more specific captions to each image: Unique-S also increased significantly. Consistent with previous studies [64, 39, 62], the models trained with the CE loss (Joint CE and Only CE) achieved the larger vocabulary than the baseline RL models. However, the improvement of our methods was even larger than these CE-based models. Despite the significant increase in the vocabulary size, our method kept the captions concise: the average sentence length was close to those of the baseline models.

**Discriminativeness.** Our goal is to enhance the discriminativeness of RL models by addressing their limited vocabulary. As expected, our methods successfully improved the discriminativeness: the R@K scores of our models were considerably higher than those of the baselines. Corresponding to the better improvement in vocabulary size, wFT increased discriminativeness more than sFT. **These results confirm our hypothesis that the limited vocabulary of RL models has been a major bottleneck for discriminativeness.**

Among the Att2In-based models, Visual Paraphrase achieved the highest discriminativeness. However, this model is not directly comparable to the others because it increases the trainable parameters for its specialized model architecture. Moreover, its improvement in discriminativeness was achieved at the expense of conciseness, which is another desirable property for discriminative image captions [51]: its sentence length was substantially longer than the other models. DiscCap performed comparably with our models, but its reward requires high computational costs. CIDErBtw and NLI proposed more lightweight rewards to be applicable to larger models, but they still need retraining from scratch. Among the larger models (UpDown and Transformer), our models achieved the highest discriminativeness despite the small computational cost.

**Standard Evaluation.** As our methods increase low-frequency words in outputs, the outputs are likely to include the words that are out-of-references (OOR). That is, low-frequency words may not be covered by reference captions regardless of their correctness due to the low frequency. These low-frequency OOR words unfairly decrease scores in conventional evaluation metrics because those metrics count exact matches in the surface form of text.\(^7\)

To fairly evaluate the OOR words, recent metric research has focused on soft matching metrics [26, 21]. Soft-matching metrics can evaluate the semantic similarity between target captions and reference captions beyond the surface form of text by utilizing pre-trained language models (PLMs) [74, 77] or pre-trained cross-modal models (PCMs) [25, 33, 21]. Their correlation with human judgments is significantly higher than that of exact-matching metrics in both precision and recall [29]. In particular, PCM-based metrics, which can utilize image features in addition to reference captions, have substantially enhanced the evaluation performance and have achieved the state-of-the-art correlation with human judgments [21, 29].

Given the above advantages, we employed soft-matching metrics in addition to conventional exact-matching metrics. Not surprisingly, our models decreased the scores in the exact-matching metrics (CIDEr and SPICE). However, our models scored comparably with the baselines in the PLM-based metric (BERTS+) and rather outperformed them in the state-of-the-art PCM-based metrics (TIGEr, CLIPS, and RefCLIPS). The higher performance in the superior soft-matching metrics indicates that our methods do not degrade the overall quality of captions. To further validate the overall quality of our output captions, the following Section 4.3 analyzes the cause of this performance gap in more detail.

### 4.3. Analysis of the Performance Gap

**Properties of OOR Words.** The critical difference between the conventional exact-matching metrics and the recent soft-matching metrics is the (in)ability to evaluate OOR words\(^8\). Based on the difference, we hypothesize that the performance gap is caused by a difference in the properties of OOR words. We analyzed the OOR words of our models, comparing with those of RL baselines and Only CE, which scores similarly to our models in exact-matching metrics but decreases soft-matching scores in contrast to our models. Table 2 shows the number of OOR words and their average frequency rank. The frequency rank refers to the order of words when sorted by their frequency in training captions; the most frequent word ranks 1st, and the value of rank increases as the frequency decreases. Although our models and Only CE output the similar number of OOR words, the significant difference in the frequency rank indicates that the properties of our OOR words are different from those of Only CE; that is, the OOR words of our models consist of much more low-frequency words than those of Only CE. Low-frequency words are likely to be OOR by the nature of their frequency, regardless of their correctness.

The soft-matching metrics could tell this difference and scored our models higher than Only CE models and even higher than baseline RL models. Especially, this trend was more clear in the state-of-the-art PCM-based metrics

---

\(^7\)Some metrics use stemming, lemmatization, and/or WordNet synsets to evaluate synonyms but their coverage is limited.

\(^8\)Note that this difference does not mean that exact-matching metrics represent precision, and soft-matching metrics represent recall. Exact-matching metrics cannot represent precision because the reference captions do not cover all correct descriptions. That is, exact-matching metrics can only represent the flawed precision with false negatives. Actually, exact-matching metrics correlate with human judgments worse than soft-matching metrics not only in recall but also in precision [29].
Table 2. Comparison of OOR words and the resulting difference in exact-matching and soft-matching metrics. We report the results on the MS COCO test set. A higher value in Rank indicates a lower frequency rank of the OOR words. We also report the rate of repetition.

<table>
<thead>
<tr>
<th>Repetition</th>
<th>Discriminativeness</th>
<th>Correctness</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep (%) ↓</td>
<td>Transformer RL</td>
<td>3.00</td>
<td>4.42</td>
</tr>
<tr>
<td>+ wFT</td>
<td>3.34**</td>
<td>4.45</td>
<td>4.84</td>
</tr>
<tr>
<td>NLI</td>
<td>3.18**</td>
<td>4.54</td>
<td>4.76</td>
</tr>
</tbody>
</table>

Table 3. Human evaluation results on the subset of the MS COCO test set. The discriminativeness score of Transformer RL was fixed at 3.00 because we set it as the baseline. "*"/"** indicates that a score is statistically significantly different from that of the baseline model (t-test with p < 0.05/0.01); one-sample t-test for discriminativeness and independent two-sample t-test for the other criteria.

Conclusion. From the above results, we conclude that the lower exact-matching scores of our models are caused by the nature of low-frequency words and the deficiency of exact-matching metrics, not by the degeneration of our models. The results of the human evaluation in the following Section 4.4 further support this conclusion.

4.4. Human Evaluation

As discussed in Sections 4.2 and 4.3, automatic evaluation of our models has difficulty due to the OOR words caused by the low frequency. To further validate the performance of our models, we conducted human evaluations using Amazon Mechanical Turk (AMT) on three criteria: discriminativeness, correctness, and fluency. Correctness and fluency are absolute scores: we instructed workers to give a maximum score 5 to the captions that did not contain incorrect information (grammatical or unnatural expressions) in terms of correctness (fluency). In contrast, discriminativeness is designed as a relative score because it is difficult to set an absolute standard for discriminativeness; unlike correctness or fluency, we cannot define the perfectly discriminative captions. Following [62], we instructed the workers to determine the discriminativeness of a caption by comparing the caption with that of a baseline model.

We evaluated the Transformer-based models, which performed the best in the automatic evaluation. Although wFT with BP decoding performed better, here we picked up wFT with $p_0$ decoding to set the total number of parameters for decoding strictly the same across the models.
[62], we randomly selected 50 images from the MS COCO test set and assigned five workers to each image. See Appendix 9 for more details on the AMT instruction. Table 3 shows the results. wFT, which had the highest R@K scores, also achieved the highest discriminativeness here. wFT achieved the same or higher correctness and fluency than the baseline model, in contrast to the exact-matching scores in Table 2. These results are consistent with the results of the state-of-the-art soft-matching metrics, confirming again that our methods do not degrade the quality of captions.

5. Related Work

Image Captioning is the task of describing images in natural languages. The quality of captions has been remarkably improved by recent advances such as the encoder–decoder captioning model [61], attention mechanism [73], RL training [48, 50], attention over bounding box features [2], large-scale pre-training [36], and large-scale captioning datasets [75, 38, 6, 32, 52]. Despite these advancements, current captioning models generate overly generic captions [12, 11, 64, 66].

Discriminative Image Captioning has been explored to generate more informative captions. [51] was the first to study it. They defined the more informative captions as the captions that concisely describe the information discriminative from distractor images, i.e., images similar to an input image. [3] proposed neural listener and speaker models that cooperate to generate discriminative captions for abstract scenes. [45] adapted the models to single-colored images. [58, 10] extended the domain to real images and improved inference efficiency. [63] proposed a memory attention network to describe unique objects among distractor images. [43] introduced a dataset with harder distractor images.

These approaches require selecting distractor images for inference. [41] and [40] proposed the methods that do not require this step. Their models learn to generate discriminative captions by maximizing the R@K scores for sampled captions using RL [50]. The R@K scores are computed with a pre-trained image–text retrieval model [15] over images in a mini-batch. [60] proposed a method to jointly train the image–text retrieval model and captioning model. Despite their effectiveness, R@K scores are associated with high computational costs and require a large batch size. Recently, CIDErBtw [62] and NLI [54] achieved state-of-the-art discriminativeness with more lightweight rewards. They weighted the contribution of ground-truth captions for the CIDEr reward according to their differences from similar but different captions [62] or their entailment scores against other ground-truth captions [54]. Another approach exploited unrelated captions as negative examples and trained caption generators with contrastive learning [12] or GAN [12, 17].

Visual Paraphrase [39] and [70] are related to our work in that they exploited low-frequency n-grams to enhance discriminativeness. [39] divided ground-truth captions into two subsets according to n-gram TF-IDF scores and proposed a new model to paraphrase low TF-IDF captions into high TF-IDF ones. [70] proposed the use of n-gram TF-IDF scores as an additional reward to a variant of R@K reward.

Different from above approaches, our objective is set to remedy the low discriminativeness of existing RL models. Our models can be achieved with single-epoch fine-tuning of pre-trained RL models, without requiring either drastic changes in the model architecture [39], additional computational costs of rewards [70], or retraining from scratch.

Diverse Image Captioning is the task of generating a set of diverse captions for a given image [67]. Diverse image captioning is aimed at enumerating various pieces of information with a set of captions, whereas discriminative image captioning aims to concisely describe the most characteristic information with a single caption. Similar to this study, some studies utilized captions that contained more low-frequency words, such as ground-truth captions [64, 42] or captions sampled from CE models [53]. Their models learn to generate these captions in addition to the captions sampled from RL models. However, these approaches still rely on sampling from skewed policies and require retraining of a model from scratch.

Long-Tail Classification has been studied extensively in various tasks as label imbalance is prevalent across datasets [78, 35]. In text-generation tasks, label imbalance exists in the frequency of words. Previous approaches have addressed the imbalance by normalizing classifier weights [46, 49] or using variants of Focal loss [49, 18, 26, 72, 37]. In contrast to these approaches, we adapted long-tail classification to mitigate the side effects of RL in the context of discriminative image captioning. Appendix 10 shows that our methods outperformed these approaches.

6. Conclusion

We have investigated the cause of overly generic captions of RL models and found out that RL decreases the discriminativeness by limiting the output words to high-frequency words. We propose the lightweight fine-tuning methods to address the bottleneck directly and achieve significantly higher discriminativeness with only the slight modification on off-the-shelf RL models. Our identification of the bottleneck and practical solutions will significantly impact future research on discriminative image captioning.

As an additional practical advantage, our models can control the granularity of descriptions from coarse to fine by just switching the off-the-shelf/fine-tuned classifier parameters. In terms of broader impact, our methods can be easily applied to the RL models in other text generation tasks, such as machine translation [71], summarization [47], and dialogue generation [34] to enrich the output vocabulary.
References


