Long-Tail Classification for Distinctive Image Captioning: A Simple yet Effective Remedy for Side Effects of Reinforcement Learning

Anonymous ACL submission

Abstract

Distinctiveness is a desirable feature of image captions. Captions should cover the characteristic details of input images. However, recent high-performing captioning models that are trained with reinforcement learning (RL) tend to generate overly generic captions despite their high performance in various other criteria. Interestingly, it has also been reported that their outputs are composed of a limited number of common words and rarely contain tail-class words, *i.e.*, low-frequency words in the training corpus. Vocabulary size is closely related to distinctiveness as it is difficult for a model to describe details beyond its vocabulary. Based on this insight, we hypothesize that the limited vocabulary of RL models is the major factor limiting their distinctiveness. We recast distinctive image captioning as a simpler task of long-tail classification to increase the vocabulary and then propose lightweight fine-tuning methods to encourage tail-class word generation. The experimental results demonstrate that our methods significantly enhance the distinctiveness of existing RL models as well as their vocabulary size, without sacrificing quality. Our methods also outperform previous distinctiveness-aware methods with a small computational cost of minor modifications to pre-trained RL models.[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, such as aiding visually impaired users (Gurari et al., 2020), visual question answering on images and videos (Fisch et al., 2020; Kim et al., 2020), visual dialogue (White et al., 2021), and news generation (Zhang et al., 2021b).

[1]The code will be made available on our website.

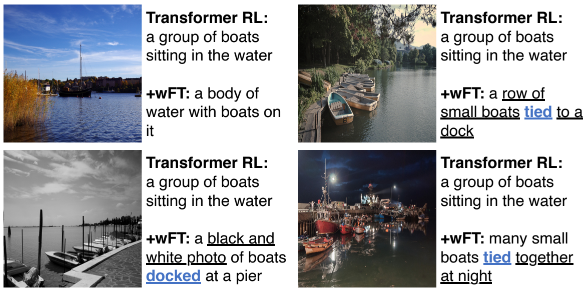

Figure 1: Caption examples in the MS COCO validation set. **Transformer RL** is a Transformer captioning model trained with RL and **+wFT** is our fine-tuning method. Transformer RL generates exactly the same caption for the four images. The underlined words indicate the characteristic information that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

For those downstream tasks, the generated captions should be **distinctive**: captions should cover the characteristic and important details of the input images. However, current captioning models tend to generate overly generic captions (Dai and Lin, 2017; Dai et al., 2017; Wang and Chan, 2019; Wang et al., 2020c). For example, a high-performing captioning model based on Transformer (Vaswani et al., 2017) 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, some studies have been conducted on **distinctive image captioning**, which is also called descriptive image captioning or discriminative image captioning. Previous research has created new rewards regarding distinctiveness or new model architectures to enhance distinctiveness. These approaches improved the performance with regard to distinctiveness and other evaluation metrics; however, their models come with additional computations and require training from scratch.

Instead of creating or paying those computational costs, we first analyze the cause of the current overly generic captions to explore ways to improve the distinctiveness of *pre-trained*, *existing models*. In particular, we focus on high-performing captioning models that are trained with the standard **reinforcement learning (RL)** (Rennie et al., 2017), which is the *de facto* standard training method in current image captioning (Stefanini et al., 2021). Those models have greater room to improve distinctiveness as they unexpectedly perform poor in distinctiveness despite the significant advantages in various other criteria (Liu et al., 2019; Wang et al., 2020a). Interestingly, some previous studies have reported that RL decreased the vocabulary size of output captions (Wang and Chan, 2019; Liu et al., 2019; Wang et al., 2020a). Vocabulary size is closely related to distinctiveness as it is difficult for a model to describe details beyond its vocabulary. Based on this insight, we hypothesize that the limited vocabulary of RL models is the major factor limiting their distinctiveness.

To directly increase the vocabulary of RL models, we recast distinctive image captioning as a simpler task of **long-tail classification**. Unlike previous approaches, our methods do not require any distinctiveness reward, new model architecture, or training from scratch. Our methods focus on generating **tail-class words**, *i.e.*, low-frequency words in the training corpus. Owing to their simplicity, our methods can be realized by single-epoch fine-tuning of pre-trained, existing RL models.

The experimental results confirm our hypothesis by revealing that our methods significantly boost both vocabulary size and distinctiveness from existing RL models. We also demonstrate that our methods outperform previous distinctiveness-aware methods with a small computational cost of minor modifications to pre-trained RL models.

2 RL Model Distinctiveness and Limited Vocabulary

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 (Stefanini et al., 2021). However, it does not improve distinctiveness and may even decrease it (Liu et al., 2019; Wang et al., 2020a). In this section, we examine the cause of overly generic captions generated by RL models and hypothesize that their limited vocabulary hinders their distinctiveness.

2.1 RL in Image Captioning

We provide a brief overview of the standard RL algorithm used in image captioning. It was proposed by Ranzato et al. (2015) and refined by Rennie et al. (2017). Their goal was to directly optimize non-differentiable test-time metrics by minimizing the negative expected reward:

$$\mathcal{L}_{\mathrm{RL}}(\theta) = -\mathbb{E}_{w^s \sim p_\theta(w^s|I)}[r(w^s)], \qquad (1)$$

where $w^s = (w_1^s, ..., w_T^s)$ is a sequence sampled from a policy p_θ, I is the input image, and $r(\cdot)$ is a reward function that returns a reward for w^s. To compute the gradient of $\mathcal{L}(\theta)$, Ranzato et al. (2015) applied the REINFORCE algorithm (Williams, 1992) to text generation. In their algorithm, the model parameters θ are updated with the following gradient:

$$\nabla_\theta \mathcal{L}_{\mathrm{RL}}(\theta) \approx -(r(w^s) - b)\nabla_\theta \log p_\theta(w^s \mid I). \qquad (2)$$

Here, b is a baseline reward that reduces the variance in the gradient. Typically, the reward function $r(\cdot)$ is CIDEr (Vedantam et al., 2015), and the baseline reward b is a reward for a sequence sampled with greedy decoding (Rennie et al., 2017).

2.2 RL Results in Limited Vocabulary

Despite its effectiveness, RL has been found to decrease distinctiveness and the number of unique n-grams in output captions (Liu et al., 2019; Wang et al., 2020a). As the relation between RL and those negative effects is not obvious, it was just considered a curious case.

However, recent research on the weaknesses of RL has revealed the relation between RL and a limited vocabulary. Recently, Choshen et al. (2020) and Kiegeland and Kreutzer (2021) empirically showed that RL makes the output distribution peaky. As shown in Section 2.1, RL samples sequences from policy p_θ. Typically, policy p_θ is computed using a captioning model pre-trained with the **Cross-Entropy (CE)** loss on ground-truth captions. However, text-generation models in general are known to output skewed distributions. Specifically, the distributions tend to be skewed towards **head-class words**, *i.e.*, high-frequency words in the training corpus (Nguyen and Chiang, 2018; Raunak et al., 2020; Demeter et al., 2020; Holtzman et al., 2020). Thus, RL can sample and reward head-class words but cannot sample or reward tail-class words during training.

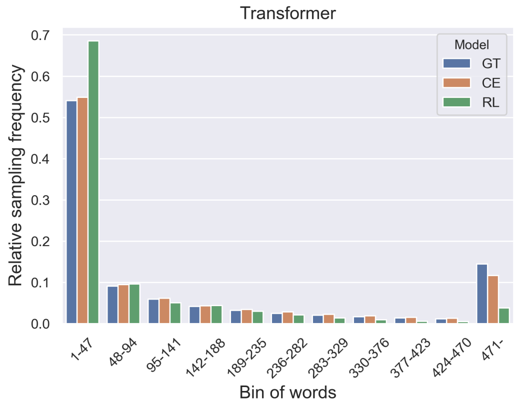

Figure 2: Relative frequency of the words in the sequences sampled for the training images. Five sequences were sampled for each image. The words (9,486 unique words excluding an out-of-vocabulary token $\langle \mathrm{unk} \rangle$) are sorted by their frequency in ground-truth captions and divided into 200 bins. We show the first 10 bins and the sum of the rest. GT is the ground-truth caption of the training images, CE is the output of a captioning model trained with the CE loss, and RL is the output of a captioning model trained with RL. Here, we used the Transformer captioning model.

This imbalance results in shifts of the probability mass from tail-class words to head-class words, further limiting the vocabulary to head-class words.

Figure 2 confirms this phenomenon in image captioning 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 sequences sampled with a CE model, the sequences sampled with an RL model are clearly limited to the head-class words, forming a peaky distribution[2].

2.3 Limited Vocabulary Results in Overly Generic Captions

Standard encoder–decoder captioning models generate captions using sequential vocabulary-size classification. However, the actual vocabulary a model can generate is much smaller than the entire vocabulary as the output distribution is highly skewed towards head-class 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 can be described by head-class words. For example, the blue words in Figure 1 are not in the actual vocabulary of the RL model; these words have never been generated by the RL model. As a result, the RL model had to ignore the characteristic relations *tied* and *docked* and ended up describing exactly the same information for all four images.

Based on the above observations, we hypothesize that the limited vocabulary of RL models hinders their distinctiveness. To directly address this limitation, we recast distinctive image captioning as a simpler task of increasing the actual vocabulary.

3 Long-Tail Classification to Remedy the Side Effects of RL

RL results in the limited vocabulary because it steals the probability mass from tail-class words of ground-truth captions. Thus, those tail-class words are the key to addressing the limitation. Wang and Chan (2019) jointly optimized both the RL loss and the CE loss on ground-truth captions so that the tail-class words in ground-truth captions would be more likely to be sampled during RL training. However, this approach still relies on the sampling from a skewed policy and requires training from scratch.

To increase the actual vocabulary more effectively and efficiently, we propose two fine-tuning methods based on long-tail classification. Our methods are designed to directly encourage the generation of tail-class words with only single-epoch fine-tuning on pre-trained, existing RL models.

3.1 Simple Fine-Tuning

The first method is a **simple fine-tuning (sFT)** method for ground-truth captions. It is based on a decoupled two-stage training (Kang et al., 2020), which is a current strong baseline model for long-tail classification (Tang et al., 2020; Menon et al., 2020; Wang et al., 2020b). Kang et al. (2020) decoupled the learning procedure into representation learning and classification, and then found that classification is critical for long-tail classification. They decoupled the classification model $f_\theta(\cdot)$ into an encoder $g_{\theta_e}(\cdot)$ and a classifier consisting of weight and bias parameters: $f_\theta(x) = \boldsymbol{W}^\top g_{\theta_e}(x) + \boldsymbol{b}$. In the first stage of training, they trained the entire classification model $f_\theta(\cdot)$ on a full training dataset. In the second stage, they fixed the encoder parameters θ_e and adjusted only the classifier parame-

[2]Although Figure 2 shows only the results obtained with the Transformer captioning model, we also confirmed that other models output peaky distributions (Rennie et al., 2017; Anderson et al., 2018). See Appendix A for the details.

ters. For the second-stage adjustment, they applied class-balanced sampling to encourage learning on tail-class labels.

Following Kang et al. (2020), we decouple a captioning model into an encoder and a classifier. In image captioning, the first-stage training of Kang et al. (2020) corresponds to RL training on the full training dataset. Likewise, 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 tail-class 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:

$$\mathcal{L}_{\mathrm{CE}}(\hat{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\theta}}(w_t^g \mid w_{<t}^g, I), \quad (3)$$

where $w^g = (w_1^g, ..., w_T^g)$ is a ground-truth caption of image I, and the model parameters $\hat{\theta}$ are initialized with RL training. The conditional probability $p_\theta(w_t^g \mid w_{<t}^g, I)$ is computed using the following softmax function:

$$p_\theta(w_t^g \mid w_{<t}^g, I) = \frac{\exp(\beta \boldsymbol{z}_{w_t^g})}{\sum_{w_i \in \mathcal{W}} \exp(\beta \boldsymbol{z}_{w_i})}, \quad (4)$$

$$\boldsymbol{z} = \boldsymbol{W}^\top g_{\theta_e}(w_{<t}^g, I) + \boldsymbol{b}, \quad (5)$$

where $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{W}|}$, $\boldsymbol{W} \in \mathbb{R}^{d \times |\mathcal{W}|}$, and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{W}|}$. $\mathcal{W}$ is the entire vocabulary, and d is the dimension of the hidden states of an encoder $g_{\theta_e}(\cdot)$. $\boldsymbol{z}_{w_i}$ indicates the element of $\boldsymbol{z}$ at the index of a word $w_i \in \mathcal{W}$. β is an inverse-temperature hyperparameter that controls the steepness of the softmax distribution. We use LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Vaswani et al., 2017) for $g_{\theta_e}(\cdot)$. During fine-tuning, only the classifier parameters $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\theta}$ are updated with the gradients $\nabla_{\boldsymbol{W}} \mathcal{L}_{\mathrm{CE}}(\hat{\theta})$ and $\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathrm{CE}}(\hat{\theta})$, respectively.

3.2 Weighted Fine-Tuning

Ground-truth captions contain more tail-class words than sampled sequences, but some tail-class 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 for head-class words and tail-class words in ground-truth captions.

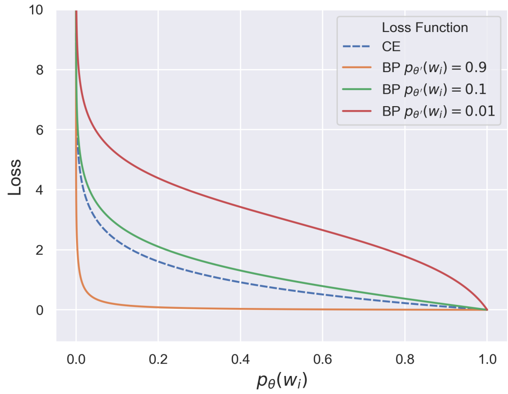

Figure 3: Visualization of the CE loss $-\log p_\theta(w_i)$ and BP loss $-\log p_{\theta,\theta'}(w_i)$. To compute the BP loss, we need the entire distribution of $\{p_\theta(w_i)\}_{w_i \in \mathcal{W}}$ and $\{p_{\theta'}(w_i)\}_{w_i \in \mathcal{W}}$. Here, we set the index i to 1 and assigned $\frac{1}{5}(1 - p_\theta(w_1))$ to the words of the next five indices, $w_2, ..., w_6$. This is because we observed that the five most probable words occupied 99% of the probability in the output distribution of the RL models. We assumed that the five most probable words were the same between p_θ and $p_{\theta'}$ as the parameters were initialized with the same RL model. Thus, we assigned $\frac{1}{5}(1 - p_{\theta'}(w_1))$ to the words of the next five indices, $w_2, ..., w_6$, likewise p_θ. Here, β was set to 1.

To rebalance the loss, we exploit the head-class bias of RL models: RL models overly assign probability to head-class words, but not to tail-class words. Based on the head-class bias, a ground-truth word that an RL model is confident of should be a head-class word that the model is refrained from further learning, whereas a ground-truth word that an RL model is not confident of should be a tail-class word for the model to learn intensely. wFT incorporates these heuristics by modifying the probability p_θ of $\mathcal{L}_{\mathrm{CE}}$ to the probability of the **bias product (BP)** (Clark et al., 2019; He et al., 2019), $p_{\theta,\theta'}$, as follows:

$$p_{\theta,\theta'}(w_t^g \mid w_{<t}^g, I) = \frac{\exp(s_\theta^t(w_t^g) + s_{\theta'}^t(w_t^g))}{\sum_{w_i \in \mathcal{W}} \exp(s_\theta^t(w_i) + s_{\theta'}^t(w_i))}, \quad (6)$$

where

$$s_\theta^t(w_i) = \log p_\theta(w_i \mid w_{<t}^g, I), \quad (7)$$

$$s_{\theta'}^t(w_i) = \log p_{\theta'}(w_i \mid w_{<t}^g, I). \quad (8)$$

By inserting $p_{\theta,\theta'}$ into $\mathcal{L}_{\mathrm{CE}}$, we define the objective of wFT as follows:

$$\mathcal{L}_{\mathrm{BP}}(\hat{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\theta},\hat{\theta}'}(w_t^g \mid w_{<t}^g, I). \quad (9)$$

Similar to sFT, both parameters $\hat{\theta}$ and $\hat{\theta}'$ are initialized with a captioning model pre-trained with RL. The difference is that, although the classifier parameters of $\hat{\theta}$ are updated, all the parameters of $\hat{\theta}'$ are fixed during fine-tuning. Figure 3 shows the change in the BP loss compared to the CE loss. The BP severely suppresses the loss when the head-class-biased policy $p_{\theta'}$ is confident, and largely increases the loss when $p_{\theta'}$ is not confident. In this way, the BP allows models to unlearn the head-class bias learned with RL. As with sFT, only the classifier parameters $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\theta}$ are updated with the gradients $\nabla_{\boldsymbol{W}} \mathcal{L}_{\mathrm{BP}}(\hat{\theta})$ and $\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathrm{BP}}(\hat{\theta})$, respectively.

We follow Clark et al. (2019) and He et al. (2019) at the evaluation stage, too. We use the probability p_θ with updated parameters rather than the BP probability $p_{\theta,\theta'}$ to avoid incorporating the head-class bias of $p_{\theta'}$ into the predictions.

4 Experiments

4.1 Setup

Dataset and Metrics. We used the MS COCO captioning dataset[3] (Lin et al., 2014; Chen et al., 2015) with Karpathy splitting (Karpathy and Fei-Fei, 2015). After preprocessing, the entire vocabulary size $|\mathcal{W}|$ was 9,487[4]. In the evaluation, the captions were decoded using a beam search of size 5 and evaluated using various evaluation metrics[5]: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and RefCLIP (Hessel et al., 2021). Following the previous studies (Liu et al., 2019; Wang et al., 2020a; Shi et al., 2021b), we evaluated distinctiveness with **R@K** scores: the percentage of captions with which an image–text retrieval model[6] (Faghri et al., 2018) could correctly retrieve the original images from the entire validation/test images within the rank of $\mathrm{K} \in \{1, 5, 10\}$. A higher R@K indicates that the model captures more characteristic information of images and generates more distinctive captions. Evaluation was conducted in a single run for each model.

Comparison Models. Following Wang et al. (2020a), we used Att2in (Rennie et al., 2017), UpDown (Anderson et al., 2018), and Transformer (Vaswani et al., 2017) as the baseline models. The models were pre-trained with the standard RL (Rennie et al., 2017) and are publicly available[7]. In addition to the baseline models, we compared our models with state-of-the-art distinctiveness-aware models: **CIDErBtw** (Wang et al., 2020a), **NLI** (Shi et al., 2021b), **DiscCap** (Luo et al., 2018), and **Visual Paraphrase** (Liu et al., 2019). The first three created new distinctiveness rewards to be optimized with RL. Visual Paraphrase introduced a new model architecture to paraphrase simpler captions to more complex captions. 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** (Wang and Chan, 2019; Edunov et al., 2018)), and also those of solely optimizing the CE loss (**Only CE**) as the baseline without using RL[8].

Hyperparameters. Our models used the same hyperparameters as the baseline models, except for the epoch size, learning rate, and β in Eq. 4. 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}. We set β to 1 for p_θ and searched for the best β of the fixed policy $p_{\theta'}$ from $\{0.1, 1\}$. The best hyperparameters were chosen according to the R@1 scores in the validation set. See Appendix B for the best hyperparameters.

All the models except Visual Paraphrase had the same parameter size as their baseline models[9].

[3]This dataset was intended for image captioning, which is consistent with our use. The dataset was created with the instruction to anonymize people's proper names (Chen et al., 2015). The dataset was licensed under CC BY 4.0. Each split of training/validation/test set contained 113,287/5,000/5,000 images, and each image had five ground-truth captions.

[4]The words that occur less than five times in the training captions were converted to $\langle$unk$\rangle$ token.

[5]We used the following library, and all the hyperparameters were set to the default values: `https://github.com/jmhessel/pycocoevalcap` (All Rights Reserved)

[6]Following Liu et al. (2019), we used a pre-trained model, `coco_vse++_resnet_restval_finetune`, which is available at `https://github.com/fartashf/vsepp` (Apache License, Version 2.0)

[7]`https://github.com/ruotianluo/self-critical.pytorch` (MIT License): `{Att2in, UpDown, Transformer}+self_critical` models.

[8]We optimized $\mathcal{L}_{\mathrm{Joint}}(\theta) = \lambda \mathcal{L}_{\mathrm{RL}}(\theta) + (1 - \lambda)\mathcal{L}_{\mathrm{CE}}(\theta)$ during RL training. We explored $\lambda \in \{0.2, 0.5, 0.8\}$. $\lambda = 0.8$ for Transformer and $\lambda = 0.2$ for the others achieved the best R@1 scores in the validation set. $\lambda = 0$ for Only CE. As with our models, all hyperparameters were set to the same as the baseline models except for the λ and scheduled sampling (Bengio et al., 2015). We disabled scheduled sampling for the CE loss to strictly separate it from the RL loss.

[9]The exact number of parameters was 14,451,985 for Att2in, 52,125,025 for UpDown, and 57,474,832 for Transformer. Note that the fixed parameters θ' were not included because they were neither trained nor used in the prediction. Visual Paraphrase has double decoders of Att2in.

		Vocabulary			*Standard Evaluation*						*Distinctiveness*		
		Unique-1	Unique-S	Length	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	RefCLIP	R@1	R@5	R@10
Att2in	**Att2in RL**	445	2,524	9.3	35.3	27.1	56.7	117.4	20.5	79.7	16.3	41.9	57.2
	+ sFT (Ours)	880	3,156	9.0	35.6	27.0	56.5	115.4	20.4	**80.3**	20.1	48.0	62.8
	+ wFT (Ours)	**1,091**	**3,749**	9.0	32.6	26.4	54.9	108.6	19.9	**80.3**	**21.7**	**50.8**	65.2
	CIDErBtw	470	2,630	9.3	35.7	27.2	56.9	**119.0**	20.7	79.8	17.2	44.1	58.7
	NLI	465	2,626	9.2	35.7	27.2	**57.0**	**119.0**	20.6	79.9	17.6	44.4	59.8
	DiscCap†		3,093	9.3	**36.1**			114.2	**21.0**		21.6	50.3	**65.4**
	Joint CE	700	2,907	9.1	36.0	**27.3**	56.4	111.7	19.9	80.0	19.1	46.7	61.5
	Only CE	689	2,845	9.2	35.7	27.1	56.1	110.7	20.1	79.9	19.0	46.6	61.1
	Visual Paraphrase†		4,576	12.9	27.1			86.9	21.1		26.3	57.2	70.8
UpDown	**UpDown RL**	577	3,103	9.5	**36.7**	27.9	57.6	**122.7**	**21.5**	80.5	21.1	49.9	64.6
	+ sFT (Ours)	1,190	3,788	9.2	35.7	27.5	56.5	115.9	21.0	**80.9**	25.0	56.8	71.2
	+ wFT (Ours)	**1,227**	**4,263**	9.3	32.0	26.5	54.3	107.9	20.4	**80.9**	**25.5**	**58.0**	**72.6**
	CIDErBtw	582	3,108	9.4	**36.7**	**28.0**	**57.7**	122.4	21.4	80.7	21.9	50.9	65.9
	NLI	575	3,144	9.4	**36.7**	**28.0**	**57.7**	122.4	21.4	80.6	21.5	50.7	65.6
	Joint CE	857	3,120	9.4	35.4	27.6	56.0	111.8	20.5	80.2	21.8	51.2	65.2
	Only CE	878	3,126	9.4	34.2	27.3	55.5	109.2	20.1	80.0	21.8	49.9	64.5
Transformer	**Transformer RL**	753	3,433	9.2	**39.0**	28.7	**58.7**	127.7	22.5	81.3	26.6	56.2	70.5
	+ sFT (Ours)	1,458	3,959	9.1	36.9	28.2	57.2	118.7	21.7	**81.5**	30.6	62.3	75.7
	+ wFT (Ours)	**1,776**	**4,274**	9.1	31.3	26.2	53.0	103.1	20.0	81.2	**32.5**	**64.5**	**77.1**
	CIDErBtw	837	3,609	9.5	38.6	28.8	58.6	128.2	22.6	81.2	27.7	57.6	71.6
	NLI	876	3,744	9.5	38.9	28.9	58.5	**129.1**	**23.0**	**81.5**	29.8	59.9	73.4
	Joint CE	1,083	3,491	9.3	38.6	**29.0**	58.3	123.8	21.9	81.2	27.3	57.2	70.8
	Only CE	935	3,599	9.4	35.0	27.7	56.0	112.2	20.8	80.9	26.5	55.8	69.7

Table 1: Comparison with the baseline models and state-of-the-art distinctiveness-aware models. Automatic evaluation results on the MS COCO test set. *Unique-1* 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 by Liu et al. (2019). Other scores were reproduced by us. The results of our models are colored in gray.

Our fine-tuning was completed in approximately 10 minutes for each model using a single GPU of 16 GB memory.

4.2 Comparison with Baseline Models and Distinctiveness-Aware Models

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

Vocabulary. First, we observed that our methods (sFT and wFT) successfully increased the actual vocabulary size: both of them considerably increased Unique-1 compared to all the baseline models. wFT increased the vocabulary more than sFT, indicating that rebalancing the loss further encouraged tail-class word generation. The increased vocabulary resulted in the captions more specific to each image: Unique-S also increased significantly. Consistent with previous studies (Wang and Chan, 2019; Liu et al., 2019; Wang et al., 2020a), the models trained with the CE loss (Joint CE and Only CE) achieved the larger vocabulary than the baseline RL models. The improvement of our methods were even larger than these CE models. Despite the significant increase in the vocabulary size, our method kept the captions concise: the average sentence length was similar to that of the baseline models.

Distinctiveness. Our goal was to address the limited vocabulary of RL models in order to increase their distinctiveness. As expected, our methods increased distinctiveness compared to the baseline models: the R@K scores of our models were considerably higher than those of all the baseline models. Corresponding to the better improvement in vocabulary size, wFT increased distinctiveness more than sFT. These results confirm our hypothesis that the major bottleneck for distinctiveness is the limited vocabulary of RL models.

Among the Att2in-based models, Visual Paraphrase achieved the highest distinctiveness. However, this model is not directly comparable because it increases the parameters for its specialized model architecture. 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 training from scratch. Among the larger models (UpDown and Transformer), our models achieved the highest distinctiveness despite the small computational cost.

Standard Evaluation. Our models degraded the performance in the text-based evaluation metrics (BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE), but they rather outperformed the baseline

	Distinctiveness	Correctness	Fluency
Transformer RL	3.00	4.42	4.83
+ wFT (Ours)	**3.34****	4.45	**4.84**
NLI	3.18**	**4.54**	4.76

Table 2: Human evaluation results on the subset of the MS COCO test set. The distinctiveness score of Transformer RL was fixed at 3.00 because we set it as the baseline model. ** indicates a model scored higher than the baseline model with statistical significance (t-test with $p < 0.01$): one-sample t-test for distinctiveness and independent two-sample t-test for the other criteria. The results of our model are colored in gray.

models in the text-and-image-based evaluation metric (RefCLIP). Because RefCLIP correlates with human evaluation better than text-based evaluation metrics (Hessel et al., 2021; Kasai et al., 2021), the high performance in RefCLIP demonstrates that our methods do not sacrifice the quality of captions. We found that our models tended to generate tail-class words that were correct but not covered by ground-truth captions, which unfairly lowered the scores in the text-based evaluation metrics. Appendix C shows those underrated captions.

4.3 Human Evaluation

We conducted human evaluations using Amazon Mechanical Turk (AMT) on three criteria: distinctiveness, correctness, and fluency. Correctness and fluency are absolute scores: we instructed workers to give a maximum score 5 to captions that *did not contain* incorrect information (ungrammatical or unnatural expressions) in terms of correctness (fluency). In contrast, distinctiveness is designed as a relative score because it is difficult to set an absolute standard for distinctiveness; unlike correctness or fluency, we cannot perfectly define distinctive captions across images. Following Wang et al. (2020a), we instructed the workers to determine the distinctiveness of a caption by comparing the caption with that of a baseline model[10].

We evaluated the Transformer-based models, which performed the best in the automatic evaluation. We randomly selected 50 images from the MS COCO test set and assigned five workers to each image. See Appendix D for more details on the AMT instruction. Table 2 shows the results. wFT, which had the highest R@K scores, also achieved the highest distinctiveness here. wFT did not achieve the highest correctness or fluency but achieved the same or higher correctness and fluency than the baseline model. This is consistent with the scores of RefCLIP, confirming again that our methods do not degrade the quality of captions.

[10]If a target caption describes the same information as a baseline caption, the workers give the target caption a score of 3; if the target caption describes more (less) characteristic information than the baseline caption, the workers give the target caption a score of 4 or 5 (1 or 2).

4.4 Qualitative Analysis

Figure 1 shows the caption examples in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model. The number of blue words indicates that our model successfully increased the vocabulary of the baseline model. Furthermore, we observed that these blue words were utilized in the description of the characteristic information of the images. See Appendix E for more examples.

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 (Vinyals et al., 2015), attention mechanism (Xu et al., 2015), RL training (Ranzato et al., 2015; Rennie et al., 2017), attention over bounding box features (Anderson et al., 2018), large-scale pre-training (Li et al., 2020b), and large-scale captioning datasets (Young et al., 2014; Lin et al., 2014; Chen et al., 2015; Krishna et al., 2017; Sharma et al., 2018). Despite these advancements, current captioning models generate overly generic captions (Dai and Lin, 2017; Dai et al., 2017; Wang and Chan, 2019; Wang et al., 2020c).

Distinctive Image Captioning has been explored to generate more informative captions. Sadovnik et al. (2012) were the first to study it. They considered the more concise and more informative captions as those that describe information distinctive from *distractor images*, *i.e.*, images similar to an input image. Andreas and Klein (2016) proposed neural listener and speaker models that cooperate to generate distinctive captions for abstract scenes. Monroe et al. (2017) adapted the models to single-colored images. Vedantam et al. (2017) and Cohn-Gordon et al. (2018) extended the domain to real images and improved inference efficiency. Recently, Wang et al. (2021) proposed a memory attention network to describe objects that are unique among distractor images.

These approaches require selecting distractor images for inference. Luo et al. (2018) and Liu et al. (2018) proposed the methods that do not require this step. Their models learn to generate distinctive captions by optimizing the R@K scores for sampled captions using RL (Rennie et al., 2017). The R@K scores are computed with a pre-trained image–text retrieval model (Faghri et al., 2018) over images in a mini batch. Vered et al. (2019) 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, Wang et al. (2020a) and Shi et al. (2021b) achieved state-of-the-art distinctiveness 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 (Wang et al., 2020a) or their entailment scores against other ground-truth captions (Shi et al., 2021b). Another approach exploited unrelated captions as negative examples and trained caption generators with contrastive learning (Dai and Lin, 2017) or GAN (Dai and Lin, 2017; Goodfellow et al., 2014).

Liu et al. (2019) and Wu et al. (2021) are related to our work in that they exploited low-frequency n-grams to enhance distinctiveness. Liu et al. (2019) divided ground-truth captions into two subsets according to n-gram TF-IDF scores and proposed a new model architecture to paraphrase low TF-IDF captions into high TF-IDF ones. Wu et al. (2021) 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 distinctiveness 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 (Liu et al., 2019), additional computational cost of rewards (Wu et al., 2021), or training of a model from scratch.

Diverse Image Captioning is the task of generating a set of diverse captions for a given image (Wang et al., 2016). Diverse image captioning is aimed at enumerating various pieces of information with a set of captions, whereas distinctive 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 tail-class words, such as ground-truth captions (Wang and Chan, 2019; Luo and Shakhnarovich, 2020) or captions sampled from CE models (Shi et al., 2021a). 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 training of a model from scratch.

Long-Tail Classification has been studied extensively in various tasks as label imbalance is prevalent across datasets (Zhang et al., 2021a; Li et al., 2020a). In text-generation tasks, label imbalance exists in the frequency of words. Previous approaches have addressed this imbalance by normalizing classifier weights (Nguyen and Chiang, 2018; Raunak et al., 2020) or using variants of Focal loss (Raunak et al., 2020; Gu et al., 2020; Jiang et al., 2019; Wu et al., 2020; Lin et al., 2017). In contrast to these approaches, we adopted long-tail classification to mitigate the side effects of RL in the context of distinctive image captioning. We also tried these approaches and found that our methods performed the best. See Appendix F for the details.

6 Limitations and Risks

Our experiments were limited to the MS COCO captioning dataset, which is the standard dataset for image captioning. The images belong to the general domain (real images of common objects) and the captions are in English only. The dataset contains social biases and captioning models have the risk of amplifying those biases (Zhao et al., 2021, 2017; Hendricks et al., 2018). Our methods are not free from the risk too as they are not designed to reduce those biases from existing models.

7 Conclusion

In this study, we have investigated the problem of overly generic captions of RL captioning models with the hypothesis that their limited vocabulary is the major hindrance to distinctiveness. We recast distinctive image captioning as a simpler task of long-tail classification to increase the vocabulary and then propose lightweight fine-tuning methods to encourage tail-class word generation. The experimental results confirm our hypothesis by demonstrating that our methods significantly enhance the distinctiveness of existing RL models as well as their vocabulary size. Our methods also outperform previous distinctiveness-aware methods with a small computational cost of minor modifications to pre-trained RL models.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2020. On the weaknesses of reinforcement learning for neural machine translation. In *ICLR*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP-IJCNLP*.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *NAACL-HLT*.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *NeurIPS*.

David Demeter, Gregory Kimmel, and Doug Downey. 2020. Stolen probability: A structural weakness of neural language models. In *ACL*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *The Ninth Workshop on Statistical Machine Translation*.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *NAACL-HLT*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. 2020. Capwap: Captioning with a purpose. In *EMNLP*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *EMNLP*.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *ECCV*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *EMNLP-IJCNLP*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *WWW*.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *ICLR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. 2021. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940*.

Samuel Kiegeland and Julia Kreutzer. 2021. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *NAACL-HLT*.

Hyounghun Kim, Zineng Tang, and Mohit Bansal. 2020. Dense-caption matching and frame-selection gating for temporal localization in videoqa. In *ACL*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020a. Dice loss for data-imbalanced nlp tasks. In *ACL*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.

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

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019. Generating diverse and descriptive image captions using visual paraphrases. In *ICCV*.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

Ruotian Luo and Gregory Shakhnarovich. 2020. Analysis of diversity-accuracy tradeoff in image captioning. *arXiv preprint arXiv:2002.11848*.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. In *ICLR*.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *TACL*, 5:325–338.

Toan Q Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *NAACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. In *ICLR*.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Amir Sadovnik, Yi-I Chiu, Noah Snavely, Shimon Edelman, and Tsuhan Chen. 2012. Image description with a goal: Building efficient discriminating expressions for images. In *CVPR*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Jiahe Shi, Yali Li, and Shengjin Wang. 2021a. Partial off-policy learning: Balance accuracy and diversity for human-oriented image captioning. In *ICCV*.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021b. Enhancing descriptive image captioning with natural language inference. In *ACL*.

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

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *CVPR*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. 2019. Joint optimization for cooperative image captioning. In *ICCV*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020a. Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2021. Group-based distinctive image captioning with memory attention. In *ACM MM*.

Qingzhong Wang and Antoni B Chan. 2019. Describing like humans: on diversity in image captioning. In *CVPR*.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. 2020b. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*.

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. 2020c. Towards unique and informative captioning of images. In *ECCV*.

Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. 2016. Diverse image captioning via grouptalk. In *IJCAI*.

Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh, and Noah Goodman. 2021. Open-domain clarification question generation without question examples. In *EMNLP*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. 2021. Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia*, 23:2413–2427.

Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2020. Importance-aware learning for neural headline editing. In *AAAI*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2021a. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhongping Zhang, Yiwen Gu, and Bryan A Plummer. 2021b. Show and write: Entity-aware news generation with image information. *arXiv preprint arXiv:2112.05917*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *ICCV*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

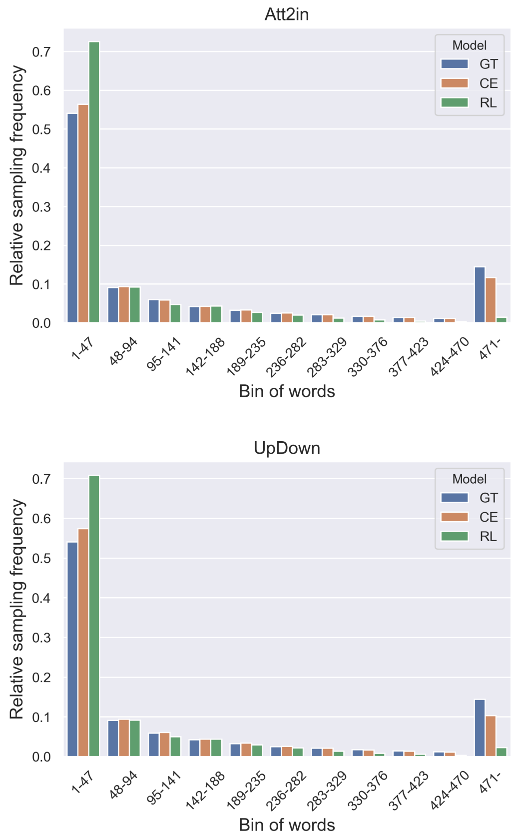

Figure 4: Relative frequency of the words in the sequences sampled for the training images. Five sequences were sampled for each image. The words (9,486 unique words excluding an out-of-vocabulary token ⟨unk⟩) are sorted by their frequency in ground-truth captions and divided into 200 bins. We show the first 10 bins and the sum of the rest. GT is the ground-truth caption of the training images, CE is the output of a captioning model trained with the CE loss, and RL is the output of a captioning model trained with RL.

Figure 5: Underrated captions in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model (Transformer RL). *Human* shows the full reference captions (ground-truth captions) of each image.

A Peaky Distributions in Other Models

Figure 4 shows the results of the plotting in Figure 2 for the LSTM-based models: Att2in (Rennie et al., 2017) and UpDown (Anderson et al., 2018). Similar to the Transformer model, the sequences sampled with the LSTM-based RL models are clearly limited to head-class words, forming the peaky distributions.

B Best Hyperparameters

As described in Section 4.1, we searched for the best hyperparameters for the learning rate (LR) from $\{1\text{e-}3, 1\text{e-}4, 1\text{e-}5, 1\text{e-}6\}$, and the inverse-temperature hyperparameter β of the fixed policy $p_{\theta'}$ from $\{0.1, 1\}$. Note that sFT does not use $p_{\theta'}$. The best hyperparameters were as follows.

Att2in RL + sFT: LR $=$ 1e-4,
Att2in RL + wFT: LR $=$ 1e-4, $\beta = 1$,
UpDown RL + sFT: LR $=$ 1e-4,
UpDown RL + wFT: LR $=$ 1e-4, $\beta = 1$,
Transformer RL + sFT: LR $=$ 1e-5,
Transformer RL + wFT: LR $=$ 1e-5, $\beta = 0.1$.

C Examples of Underrated Captions

Figure 5 shows caption examples, reference captions, and their automatic evaluation scores. It is clear that our +wFT model correctly described all three images with diverse vocabulary. However, the CIDEr scores were quite low compared with those of the baseline model, Transformer RL. The cause

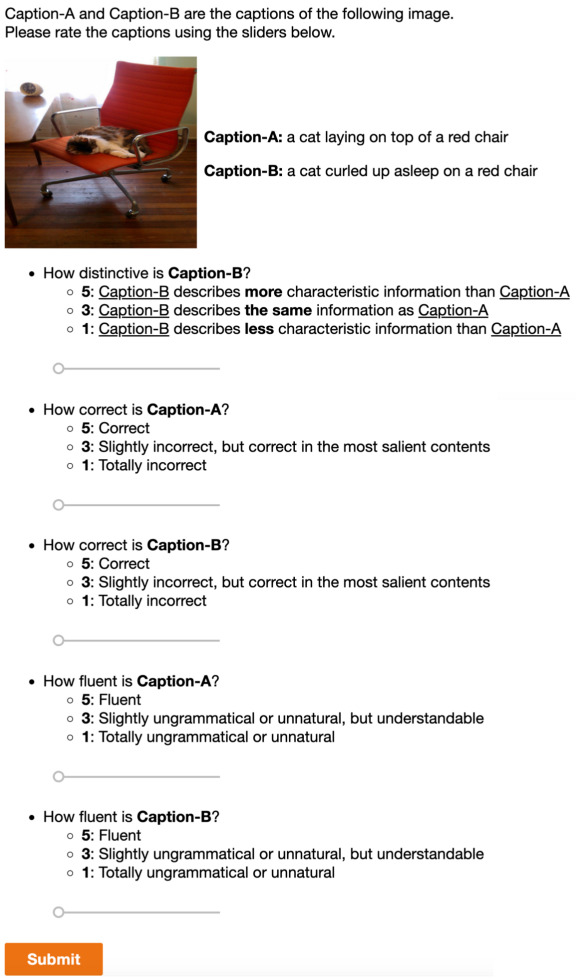

Figure 6: A screenshot of our AMT interface.

(a) **Transformer RL:** a tower with a clock on top of it
+wFT: a clock tower with a **weather vane** on top
NLI: a tower with a clock on the top of it
Human: a weather vane atop a cathedral clock tower

(b) **Transformer RL:** a group of birds standing in the water
+wFT: a large group of **flamingos** stand in **shallow** water
NLI: a group of pink umbrellas are standing in the water
Human: a flock of pink flamingos standing in shallow water

(c) **Transformer RL:** a black cat wearing a hat on top of a table
+wFT: a cat wears a **funny** hat while **staring straight ahead**
NLI: a black cat wearing a hat sitting on a table
Human: the cute black cat is wearing a bee's hat

(d) **Transformer RL:** a group of people riding motorcycles on a road
+wFT: a group of people **racing** motorcycles on a race track
NLI: a group of people riding motorcycles on a race track
Human: people are racing motorcycles on a race track

(e) **Transformer RL:** a dog next to a cup of coffee
+wFT: a dog is **sniffing** a cup of coffee
NLI: a dog standing next to a coffee cup on a table
Human: a squinting dog on a brick patio sniffs a cup of coffee

Figure 7: Caption examples in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model (Transformer RL). *Human* shows a ground-truth caption of each image.

of this underrating is the small coverage of the reference captions: the reference captions rarely contain the tail-class words colored in blue, probably due to their low frequency. Text-based evaluation metrics such as CIDEr cannot evaluate the expressions that are correct but not covered by reference captions. In contrast, RefCLIP incorporates image features and can consider information that is not covered by reference captions. We observed that the RefCLIP scores were more plausible in these examples.

D Details of Human Evaluation

We show our AMT interface in Figure 6. Each image was evaluated with the five questions in the discrete 5-point scale. We required workers to satisfy the following qualifications: being an AMT Master and living in the U.S. Workers were notified that this experiment was intended to evaluate caption quality. We paid \$0.1 for each image, and the median of the actual working time was 41 seconds per image. The hourly reward was estimated as \$8.78, which is higher than the minimum wage in the U.S., \$7.25 per hour.

E Detailed Qualitative Analysis

Figure 7 shows more caption examples in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model. We observed that these blue words expressed various types of characteristic information of the images. Here, *weather vane* and *flamingos* are characteristic objects of the images (a) and (b); *shallow*, *funny*, and *staring straight ahead* are characteristic attributes of the images (b) and (c); and *racing* and *sniffing* are characteristic relations in the images (d) and (e). These examples further support our hypothesis that the limited vocabulary of RL models hinders their distinctiveness.

	Vocabulary			*Standard Evaluation*						*Distinctiveness*		
	Unique-1	Unique-S	Length	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	RefCLIP	R@1	R@5	R@10
Att2in RL	445	2,524	9.3	35.3	**27.1**	**56.7**	**117.4**	**20.5**	79.7	16.3	41.9	57.2
+ sFT (Ours)	880	3,156	9.0	**35.6**	27.0	56.5	115.4	20.4	**80.3**	20.1	48.0	62.8
+ wFT (Ours)	**1,091**	**3,749**	9.0	32.6	26.4	54.9	108.6	19.9	**80.3**	**21.7**	**50.8**	**65.2**
+ τ-norm	437	2,414	9.1	35.4	27.0	**56.7**	117.3	20.4	79.7	15.4	40.7	55.8
+ FL	902	3,236	9.0	35.2	27.0	56.4	114.6	20.3	**80.3**	20.3	48.4	63.4
+ AFL	886	3,104	9.0	35.4	27.0	56.6	115.2	20.4	**80.3**	19.6	47.5	62.7
UpDown RL	577	3,103	9.5	36.7	**27.9**	57.6	**122.7**	**21.5**	80.5	21.1	49.9	64.6
+ sFT (Ours)	1,190	3,788	9.2	35.7	27.5	56.5	115.9	21.0	**80.9**	25.0	56.8	71.2
+ wFT (Ours)	**1,227**	**4,263**	9.3	32.0	26.5	54.3	107.9	20.4	**80.9**	**25.5**	**58.0**	**72.6**
+ τ-norm	576	2,967	9.3	**37.0**	27.7	**57.7**	122.6	21.3	80.5	19.6	48.1	63.4
+ FL	1,208	3,838	9.2	35.4	27.4	56.3	114.8	20.8	**80.9**	25.3	57.2	71.0
+ AFL	1,168	3,746	9.2	35.9	27.5	56.7	116.4	20.9	**80.9**	24.7	56.5	70.5
Transformer RL	753	3,433	9.2	**39.0**	**28.7**	**58.7**	**127.7**	**22.5**	81.3	26.6	56.2	70.5
+ sFT (Ours)	1,458	3,959	9.1	36.9	28.2	57.2	118.7	21.7	81.5	30.6	62.3	75.7
+ wFT (Ours)	**1,776**	**4,274**	9.1	31.3	26.2	53.0	103.1	20.0	81.2	**32.5**	**64.5**	**77.1**
+ τ-norm	1,027	3,483	9.2	38.5	28.4	58.3	124.4	22.1	81.2	26.1	55.8	69.7
+ FL	1,523	4,018	9.1	36.1	28.0	56.6	116.5	21.4	81.5	31.2	63.1	76.3
+ AFL	1,402	3,908	9.1	37.4	28.3	57.5	120.5	21.9	**81.6**	30.0	62.1	75.9

Table 3: Comparison with the other long-tail classification methods. Automatic evaluation results on the MS COCO test set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of output captions. The results of our models are colored in gray.

F Comparison with Other Long-Tail Classification Methods

We adapted the long-tail classification method of Kang et al. (2020) to remedy the side effects of RL and proposed sFT and wFT. Both methods were carefully designed for RL models, but these were not the only way to employ long-tail classification methods. In this section, we discuss the other possible adaptations based on Raunak et al. (2020).

Raunak et al. (2020) explored ways to employ long-tail classification methods to machine translation. Their first method was τ-normalization (τ-norm), which directly adopted the method of Kang et al. (2020). To simply make the output distributions flatter, they normalized the classifier weight $\boldsymbol{W}$ as follows:

$$\widetilde{\boldsymbol{W}}_{w_i} = \frac{\boldsymbol{W}_{w_i}}{\|\boldsymbol{W}_{w_i}\|^{\tau}}, \qquad (10)$$

where $\boldsymbol{W}_{w_i} \in \mathbb{R}^d$ indicates a vector at the index of a word w_i and τ is a temperature hyperparameter that controls the degree of the normalization.

The other methods of Raunak et al. (2020) were Focal loss (FL) and Anti-Focal loss (AFL). AFL is a variant of FL (Lin et al., 2017), which was aimed at reweighting the loss according to the confidence of the model predictions. Let $p_\theta^t = p_\theta(w_t^g \mid w_{<t}^g, I)$. FL and AFL in image captioning are then written as follows:

$$\mathcal{L}_{\mathrm{FL}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}(1-p_\theta^t)^\gamma \log p_\theta^t, \qquad (11)$$

$$\mathcal{L}_{\mathrm{AFL}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}(1+\alpha p_\theta^t)^\gamma \log p_\theta^t, \qquad (12)$$

where γ and α are hyperparamters that control the degree of the reweighting. Other work also explored ways to employ long-tail classification methods to text generation, but those approaches can be categorized as either τ-norm (Nguyen and Chiang, 2018) or the variants of FL (Gu et al., 2020; Jiang et al., 2019; Wu et al., 2020), which we already explored above.

We compared our methods (sFT and wFT) with τ-norm, FL, and AFL. In our experiments, we normalized the bias term $\boldsymbol{b}$[11] in addition to the weight term $\boldsymbol{W}$ as we found it performed better than normalizing the weight term only. For a fair comparison with our methods, we applied FL and AFL at the fine-tuning of RL models. That is, we optimized $\mathcal{L}_{\mathrm{FL}}(\hat{\theta})$ and $\mathcal{L}_{\mathrm{AFL}}(\hat{\theta})$, where $\hat{\theta}$ were initialized with the pre-trained RL models. We used the best hyperparameters reported in Raunak et al. (2020): $\tau = 0.2$, $\gamma = 1$, and $\alpha = 1$. Similar to our models, other hyperparameters were set to the same values as the baseline models, except for the epoch size and learning rate. We explored the same values

[11] $\tilde{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|^{\tau}}$, where the value of the hyperparameter τ was set to the same as that of $\widetilde{\boldsymbol{W}}$.

	Vocabulary			*Standard Evaluation*						*Distinctiveness*		
	Unique-1	Unique-S	Length	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	RefCLIP	R@1	R@5	R@10
Att2in RL	435	2,583	9.3	**35.0**	**27.0**	**56.7**	**116.5**	**20.3**	79.8	16.2	42.5	57.0
+ sFT (Ours)	874	3,189	9.0	**35.0**	**27.0**	56.3	113.7	20.1	80.3	19.2	47.9	62.9
+ wFT (Ours)	**1,092**	**3,806**	9.0	32.5	26.4	54.8	107.2	19.7	**80.4**	**20.6**	**50.7**	**65.3**
UpDown RL	563	3,161	9.5	**36.7**	**27.9**	**57.7**	**122.3**	**21.3**	80.6	20.6	50.2	65.7
+ sFT (Ours)	1,222	3,805	9.2	35.4	27.4	56.7	115.3	20.7	80.9	24.6	56.2	70.9
+ wFT (Ours)	**1,230**	**4,311**	9.3	31.8	26.4	54.3	106.5	20.2	**81.0**	**25.9**	**58.2**	**73.6**
Transformer RL	713	3,432	9.2	**38.9**	**28.7**	**58.7**	**126.4**	**22.1**	81.2	25.4	56.3	69.8
+ sFT (Ours)	1,496	3,953	9.1	37.5	28.3	57.4	118.4	21.4	**81.5**	30.2	62.7	75.8
+ wFT (Ours)	**1,836**	**4,268**	9.1	31.1	26.3	53.3	102.2	19.8	81.3	**32.2**	**64.3**	**76.8**

Table 4: Automatic evaluation results on the MS COCO *validation* set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of output captions. The results of our models are colored in gray.

for these hyperparameters as our models: 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}. The best learning rates were chosen according to the R@1 scores in the validation set[12]. Note that we did not explore the learning rate for τ-norm because it does not require training.

Table 3 shows the results. In contrast to the results reported in machine translation (Raunak et al., 2020; Nguyen and Chiang, 2018), τ-norm models performed lower than the baseline models. These results indicate that simply flattening the output distributions does not work in image captioning. Although FL and AFL increased the vocabulary size and distinctiveness, the gains were smaller than those of wFT.

To analyze the cause of the difference between the FL, AFL, and the BP loss (wFT), we visualized the losses in Figure 8. FL suppresses the loss when a model is confident, whereas AFL increases the loss when a model is moderately confident. Compared with these losses, BP changes the loss more drastically. When the head-class-biased policy $p_{\theta'}$ is highly confident, BP strictly suppresses the loss to prevent further learning on that word; when $p_{\theta'}$ is not confident, BP highly increases the loss to encourage the learning on that word. This drastic rebalancing of the loss resulted in the larger vocabulary size and higher distinctiveness of wFT.

G Validation Performance for Reproduction

Table 4 shows the performance of our models on the MS COCO validation set. We report these results for the future reproduction of our experiments.

[12]The best learning rates were 1e-4 for Att2in RL + FL/AFL, 1e-4 for UpDown RL + FL/AFL, and 1e-5 for Transformer RL + FL/AFL.

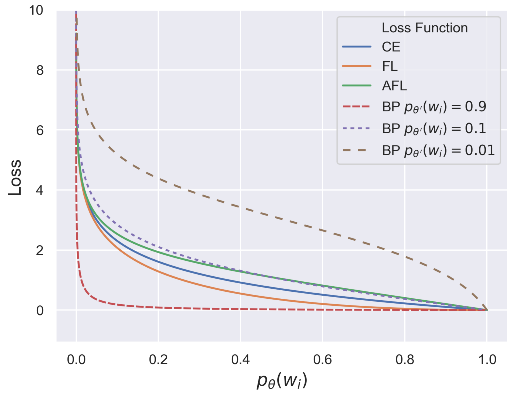

Figure 8: Visualization of the losses: CE $-\log p_\theta(w_i)$, BP $-\log p_{\theta,\theta'}(w_i)$, FL $(1 - p_\theta(w_i))^\gamma \log p_\theta(w_i)$, and AFL $(1 + \alpha p_\theta(w_i))^\gamma \log p_\theta(w_i)$. Here, we set $\beta = 1$, $\gamma = 1$, and $\alpha = 1$.

The code will be also available on our website for the reproduction.