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

Ukyo Honda[1,2] Taro Watanabe[3] Yuji Matsumoto[2]
[1]CyberAgent, Inc. [2]RIKEN [3]Nara Institute of Science and Technology
honda_ukyo@cyberagent.co.jp taro@is.naist.jp yuji.matsumoto@riken.jp

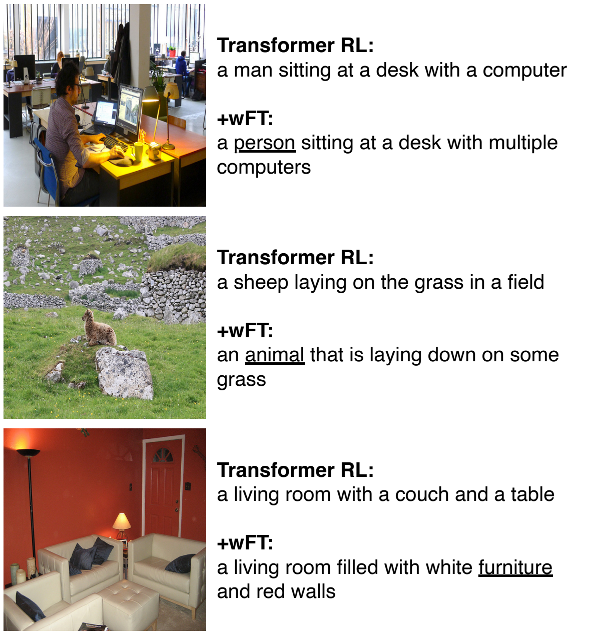

Transformer RL:
a man sitting at a desk with a computer
+wFT:
a person sitting at a desk with multiple computers
Transformer RL:
a sheep laying on the grass in a field
+wFT:
an animal that is laying down on some grass
Transformer RL:
a living room with a couch and a table
+wFT:
a living room filled with white furniture and red walls

Figure 1. Examples of the limitation of our methods. All the examples are from the MS COCO validation set. The underlined words are relatively low-frequency hypernyms.

1. Limitations and Ethical Considerations

Our experiments were limited to the MS COCO 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. To compensate for the limitation, we have demonstrated the effectiveness of our methods with the multiple baseline models.

Our current methods have a limitation in that they cannot select discriminative ones among low-frequency words. Although discriminative in general, low-frequency words do not always describe more specific information than others. Figure 1 shows the examples. Our model output relatively low-frequency hypernyms such as *person*, *animal*, and *furniture* instead of the more frequent but more specific hyponyms: *man*, *sheep*, and *couch*. Utilizing thesauruses like WordNet [4] will be a promising approach to reduce those relatively low-frequency hypernyms from outputs.

The dataset contains social biases, and captioning models have the risk of amplifying those biases [24, 25, 6]. Our methods are also not free from the risk, as they are not designed to reduce those social biases from existing models.

2. Further Output Examples

Figure 2 shows 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 observe that these blue words express 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 discriminativeness.

3. Peaky Distributions in Other Models

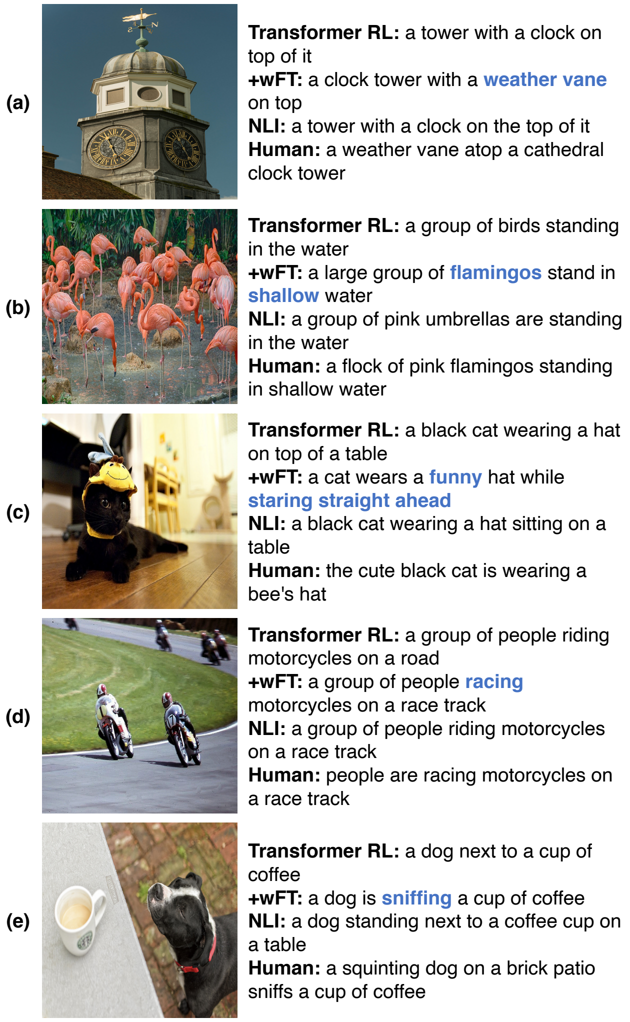
Figure 3 shows the results of the relative frequency of the words sampled for the training images by the LSTM-based models: Att2in [18] and UpDown [1]. Similar to the Transformer model, the sequences sampled with the LSTM-based RL models are clearly limited to high-frequency words, forming the peaky distributions.

4. Libraries for Evaluation

We used the following libraries for evaluation with all the hyperparameters set to the default values.

CIDEr, SPICE, CLIPS, and RefCLIPS https://github.com/jmhessel/pycocoevalcap

BERTS+ https://github.com/ck0123/improved-bertscore-for-image-captioning-eval



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

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

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

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

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 2. 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.

uation
TIGEr https://github.com/SeleenaJM/CapEval
R@K https://github.com/fartashf/vsepp; following [12], we used a publicly available model, coco_vse++_resnet_restval_finetune.

5. Best Hyperparameters

We searched for the best hyperparameters for the learning rate from $\{1\text{e-}3, 1\text{e-}4, 1\text{e-}5, 1\text{e-}6\}$, and the inverse-temperature hyperparameter β' of Eq. (7) from $\{0.1, 1\}$. The best learning rate was 1e-5 for Transformer models and 1e-4 for the other models (Att2in and UpDown). The best β' was 0.1 for wFT with p_θ decoding and 1 for wFT with BP decoding. Note that sFT does not use β'.

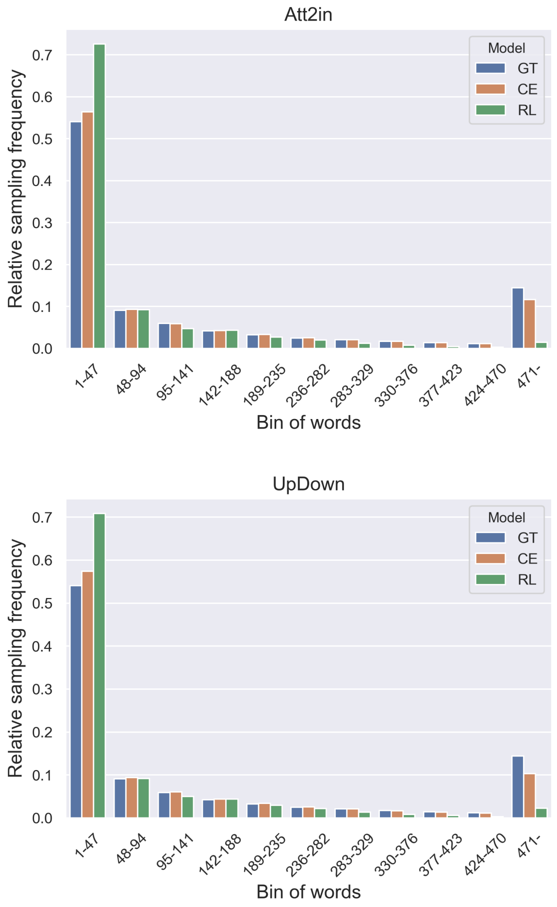


Figure 3. Relative frequency of the words in the sequences sampled for the training images. Five sequences were sampled for each image. The words (9,486 unique words excluding an out-of-vocabulary token $\langle \text{unk} \rangle$) are sorted by their frequency in ground-truth captions and divided into 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.

The best learning rate was the same in CE-based models (Joint CE and Only CE): 1e-5 for Transformer and 1e-4 for the others. The best $\lambda \in \{0.2, 0.5, 0.8\}$ for Joint CE was 0.8 for Transformer and 0.2 for the others.

6. The Number of Parameters

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 parameters θ' are not included because they are not trainable and fixed through the entire training and evaluation; rather, the actual trainable parameters are decreased to the classifier parameters in our models. Visual Paraphrase has double decoders of Att2in; thus, it increases

	Epoch	Batch	Hour/Epoch	Total Hour
Att2in RL	20	10	0.68	13.54
+ sFT	1	10	**0.08**	**0.08**
+ wFT	1	10	0.12	0.12
CIDErBtw	50	10	0.70	35.11
NLI	50	16	0.87	43.55
Joint CE	20	10	1.15	22.97
UpDown RL	20	10	0.71	14.16
+ sFT	1	10	**0.09**	**0.09**
+ wFT	1	10	0.14	0.14
CIDErBtw	50	10	0.76	38.09
NLI	50	16	0.87	43.74
Joint CE	20	10	1.08	21.67
Transformer RL	25	10	3.23	80.66
+ sFT	1	10	**0.11**	**0.11**
+ wFT	1	10	0.18	0.18
CIDErBtw	25	10	3.27	81.76
NLI	25	16	2.74	68.54
Joint CE	25	10	4.06	101.43

Table 1. Time to train discriminativeness-aware captioning models. Note that we excluded the time for initialization before RL because there is not much difference among the methods. Results for the baseline RL models are shown in gray text because we did not train these models but used publicly-available pre-trained models.

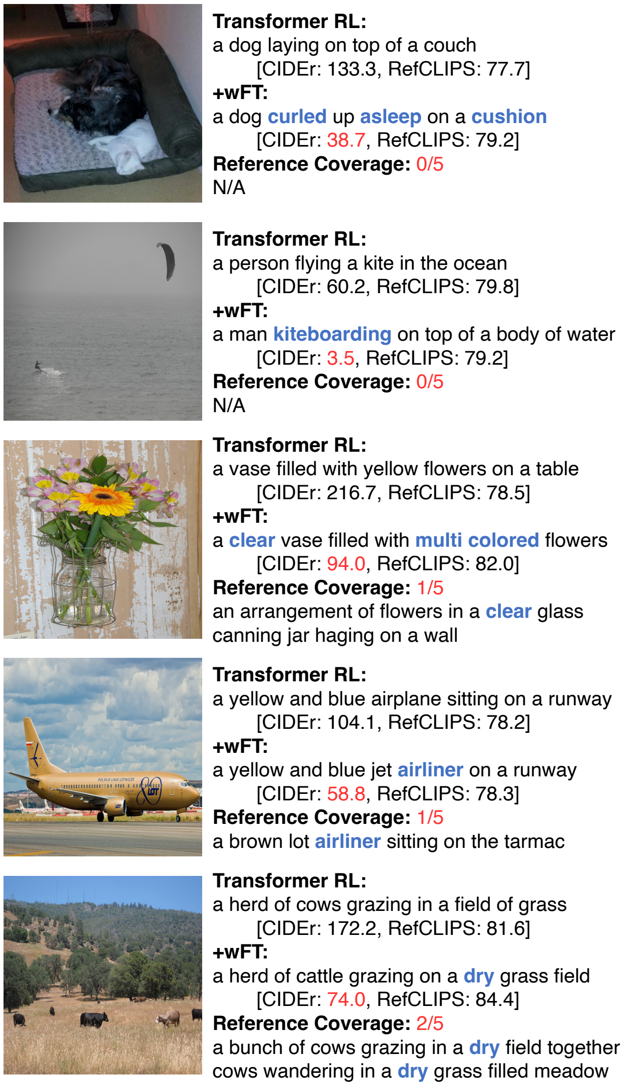
the number of trainable parameters and requires training of the specialized model from scratch.

7. Comparison of Computational Cost

Table 1 shows the time to train discriminativeness-aware captioning models. We used a single GPU of 16 GB memory for all training. Clearly, our methods require far less time for training. This is because our methods do not require retraining from scratch but only require a single-epoch fine-tuning to publicly-available pre-trained RL models.

8. Qualitative Analysis of Underrated Captions

Figure 4 shows caption examples, automatic evaluation scores, and reference captions. Clearly, our wFT model correctly described all five images with diverse vocabulary. However, the CIDEr scores for our captions were considerably lower than those for the baseline model captions. The cause of this underrating is the small coverage of the reference captions: the reference captions rarely include the low-frequency words colored in blue due to their low frequency. Conventional exact-matching metrics such as CIDEr cannot evaluate those correct-but-OOR words by the definition of exact-matching. In contrast, RefCLIPS, the state-of-the-art soft-matching metric, can consider the information not covered by reference captions by incorporating image features. Figure 4 shows that RefCLIPS evaluated the correct-but-OOR words more correctly and gave more plausible scores to our captions. These examples further support our conclusion 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.



Transformer RL:
a dog laying on top of a couch
[CIDEr: 133.3, RefCLIPS: 77.7]
+wFT:
a dog **curled** up **asleep** on a **cushion**
[CIDEr: 38.7, RefCLIPS: 79.2]
Reference Coverage: 0/5
N/A

Transformer RL:
a person flying a kite in the ocean
[CIDEr: 60.2, RefCLIPS: 79.8]
+wFT:
a man **kiteboarding** on top of a body of water
[CIDEr: 3.5, RefCLIPS: 79.2]
Reference Coverage: 0/5
N/A

Transformer RL:
a vase filled with yellow flowers on a table
[CIDEr: 216.7, RefCLIPS: 78.5]
+wFT:
a **clear** vase filled with **multi colored** flowers
[CIDEr: 94.0, RefCLIPS: 82.0]
Reference Coverage: 1/5
an arrangement of flowers in a **clear** glass canning jar haging on a wall

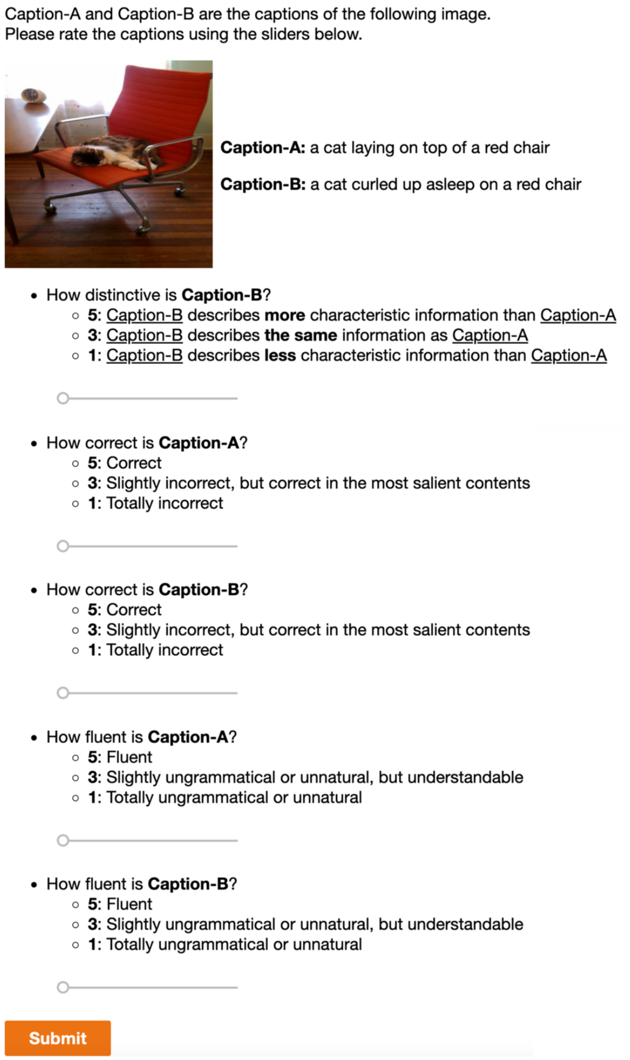
Transformer RL:
a yellow and blue airplane sitting on a runway
[CIDEr: 104.1, RefCLIPS: 78.2]
+wFT:
a yellow and blue jet **airliner** on a runway
[CIDEr: 58.8, RefCLIPS: 78.3]
Reference Coverage: 1/5
a brown lot **airliner** sitting on the tarmac

Transformer RL:
a herd of cows grazing in a field of grass
[CIDEr: 172.2, RefCLIPS: 81.6]
+wFT:
a herd of cattle grazing on a **dry** grass field
[CIDEr: 74.0, RefCLIPS: 84.4]
Reference Coverage: 2/5
a bunch of cows grazing in a **dry** field together
cows wandering in a **dry** grass filled meadow

Figure 4. 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). *Reference Coverage* shows the number of reference captions (out of five) that cover at least one of the blue words.

9. Details of Human Evaluation

We show our AMT interface in Figure 5. 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



Caption-A and Caption-B are the captions of the following image.
Please rate the captions using the sliders below.

Caption-A: a cat laying on top of a red chair

Caption-B: a cat curled up asleep on a red chair

- How distinctive is **Caption-B**?
    - **5:** Caption-B describes **more** characteristic information than Caption-A
    - **3:** Caption-B describes **the same** information as Caption-A
    - **1:** Caption-B describes **less** characteristic information than Caption-A

○

- How correct is **Caption-A**?
    - **5:** Correct
    - **3:** Slightly incorrect, but correct in the most salient contents
    - **1:** Totally incorrect

○

- How correct is **Caption-B**?
    - **5:** Correct
    - **3:** Slightly incorrect, but correct in the most salient contents
    - **1:** Totally incorrect

○

- How fluent is **Caption-A**?
    - **5:** Fluent
    - **3:** Slightly ungrammatical or unnatural, but understandable
    - **1:** Totally ungrammatical or unnatural

○

- How fluent is **Caption-B**?
    - **5:** Fluent
    - **3:** Slightly ungrammatical or unnatural, but understandable
    - **1:** Totally ungrammatical or unnatural

○

Figure 5. A screenshot of our AMT interface.

higher than the minimum wage in the U.S., $7.25 per hour.

10. Comparison with Other Long-Tail Classification Methods

We adapted the long-tail classification method of [9] to relieve the bottleneck 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 [17].

[17] explored ways to employ long-tail classification methods for machine translation. Their first method was τ-normalization (τ-norm), which directly adopted the method of [9]. Based on an observation that the norm of classifier parameters correlates with the frequency of the classes, they normalized the classifier weight $\boldsymbol{W}$ as follows:

$$\widetilde{\boldsymbol{W}}_{w_i} = \frac{\boldsymbol{W}_{w_i}}{\|\boldsymbol{W}_{w_i}\|^{\tau}}, \tag{1}$$

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 [17] were Focal loss (FL) and Anti-Focal loss (AFL). AFL is a variant of FL [11], 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:

$$\mathcal{L}_{\mathrm{FL}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}(1 - p_\theta^t)^\gamma \log p_\theta^t, \tag{2}$$

$$\mathcal{L}_{\mathrm{AFL}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}(1 + \alpha p_\theta^t)^\gamma \log p_\theta^t, \tag{3}$$

where γ and α are hyperparamters that control the degree of the reweighting. Other work also explored ways to employ long-tail classification methods for text generation, but those approaches are categorized as either τ-norm [14] or variants of FL [5, 8, 22], 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 $\boldsymbol{b}$[1] in addition to the weight term $\boldsymbol{W}$ as we found it performed better than normalizing the weight term only. We applied FL and AFL as the alternative weighting to BP for a fair comparison with our methods. That is, we fine-tuned the classifier parameters by optimizing $\mathcal{L}_{\mathrm{FL}}(\hat{\theta})$ or $\mathcal{L}_{\mathrm{AFL}}(\hat{\theta})$, where $\hat{\theta}$ were initialized with the pre-trained RL models. We used the best hyperparameters reported in [17]: $\tau = 0.2$, $\gamma = 1$, 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 for these hyperparameters as our models: we set the epoch size for fine-tuning to 1 and searched for the best learning rates from $\{$1e-3, 1e-4, 1e-5, 1e-6$\}$. We selected the best learning rate according to the R@1 scores in the validation set. The best learning rate was 1e-4 for Att2in RL + FL/AFL, 1e-4 for UpDown RL + FL/AFL, and 1e-5 for Transformer RL + FL/AFL. Note that we did not explore the learning rate for τ-norm because it does not require training.

In open-ended text generation tasks, *e.g.*, story generation and text generation after prompts, stochastic sampling methods are used instead of beam search to increase the diversity in output text [7, 2, 13]. Although image captioning does not fall in the category of open-ended text generation

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

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS+	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Att2in RL	445	2,524	9.3	**117.4**	**20.5**	43.6	73.9	73.0	79.7	16.3	41.9	57.2
+ sFT	880	3,156	9.0	115.4	20.4	**43.9**	74.3	73.7	**80.3**	20.1	48.0	62.8
+ wFT	**1,197**	**3,732**	8.9	104.3	19.5	43.1	74.2	73.9	80.2	20.6	49.7	64.5
+ wFT (BP decoding)	1,102	3,615	9.4	109.3	20.1	43.7	**74.4**	**74.0**	80.2	**21.1**	**50.5**	**64.8**
+ τ-norm	437	2,414	9.1	117.3	20.4	43.5	73.8	72.9	79.7	15.4	40.7	55.8
+ FL	903	3,217	9.0	114.8	20.4	43.8	74.3	73.7	**80.3**	20.1	48.1	63.2
+ AFL	886	3,116	9.0	115.3	20.4	43.8	74.3	73.7	**80.3**	19.7	47.6	62.7
+ Nucleus sampling	475	2,726	9.3	116.5	20.3	43.5	73.9	72.9	79.7	16.5	41.9	57.1
UpDown RL	577	3,103	9.5	**122.7**	**21.5**	**44.2**	74.6	74.0	80.5	21.1	49.9	64.6
+ sFT	1,190	3,788	9.2	115.9	21.0	**44.2**	**74.9**	74.8	**80.9**	25.0	56.8	71.2
+ wFT	**1,479**	**4,268**	9.1	101.8	19.5	43.1	74.6	74.9	80.7	26.0	57.6	72.2
+ wFT (BP decoding)	1,275	4,177	9.6	110.0	20.6	44.1	**74.9**	**75.0**	80.8	**26.7**	**58.7**	**72.4**
+ τ-norm	576	2,967	9.3	122.6	21.3	**44.2**	74.4	73.8	80.5	19.6	48.1	63.4
+ FL	1,201	3,830	9.2	114.9	20.9	44.1	**74.9**	74.7	**80.9**	25.2	57.0	70.9
+ AFL	1,171	3,760	9.2	116.4	20.9	**44.2**	**74.9**	74.7	**80.9**	24.9	56.6	70.7
+ Nucleus sampling	592	3,339	9.5	120.7	21.3	**44.2**	74.6	73.9	80.4	20.9	49.7	64.4
Transformer RL	753	3,433	9.2	**127.7**	**22.5**	**45.1**	75.0	75.0	81.3	26.6	56.2	70.5
+ sFT	1,458	3,959	9.1	118.7	21.7	44.8	**75.2**	75.6	81.5	30.6	62.3	75.7
+ wFT	1,776	4,274	9.1	103.1	20.0	43.3	74.8	75.8	81.2	32.5	64.5	77.1
+ wFT (BP decoding)	**1,964**	**4,373**	9.4	107.2	21.1	44.2	**75.2**	**76.1**	81.5	**33.5**	**65.9**	**78.2**
+ τ-norm	1,027	3,483	9.2	124.4	22.1	44.9	74.8	74.9	81.2	26.1	55.8	69.7
+ FL	1,523	4,018	9.1	116.5	21.4	44.6	**75.2**	75.7	81.5	31.2	63.1	76.3
+ AFL	1,402	3,908	9.1	120.5	21.9	44.8	**75.2**	75.6	**81.6**	30.0	62.1	75.9
+ Nucleus sampling	1,053	3,751	9.3	123.7	22.0	44.8	74.9	75.0	81.2	26.9	55.8	70.4

Table 2. 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.

as input images tightly scope the correctness of captions, we additionally test whether the randomness in stochastic sampling can increase the output vocabulary. We used Nucleus sampling [7] with a hyperparameter $p = 0.95$, which is the best hyperparameter reported [7, 13].

Table 2 shows the results. τ-norm and Nucleus sampling showed the similar results. Both methods slightly increased the output vocabulary but the performance generally remained the same as the baseline models. These results indicate that the output vocabulary cannot be significantly increased while maintaining the relative probability of words: Nucleus sampling samples according to the original output distributions and τ-norm changes the distribution only by the difference in the norm, basically flattening the distribution. In contrast, FL and AFL drastically change the relative probability of words by refining the mapping from encoded features to low-frequency words, as with sFT and wFT. They successfully increased the vocabulary size and discriminativeness. However, the gains were smaller than those of wFT.

To analyze the cause of the difference between FL, AFL, and the BP loss (wFT), we visualized the losses in Figure 6. 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 frequency-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 wFT's larger vocabulary size and higher discriminativeness.

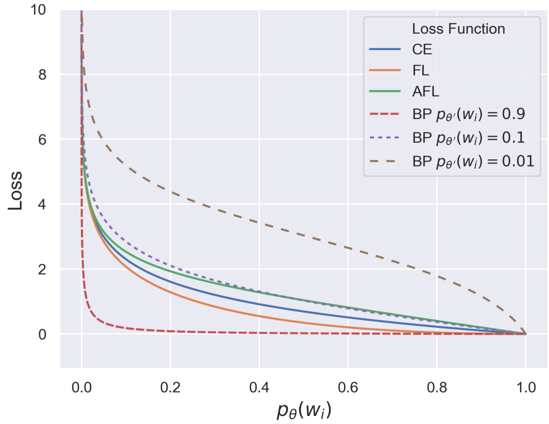


Figure 6. Visualization of the losses: CE $-\log p_\theta(w_i)$, BP $-\log p_{\theta,\theta'}(w_i)$, FL $(1 - p_\theta(w_i))^\gamma \log p_\theta(w_i)$, and AFL $(1 + \alpha p_\theta(w_i))^\gamma \log p_\theta(w_i)$. We set $\beta = \beta' = 1$, $\gamma = 1$, and $\alpha = 1$.

11. Validation Performance for Reproduction

Table 3 shows the performance of our models on the MS COCO validation set. We report these results for the future reproduction of our experiments. The code will be made available at https://github.com/ukyh/switch_disc_caption.git.

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS+	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Att2in RL	435	2,583	9.3	**116.5**	**20.3**	43.6	N/A	73.1	79.8	16.2	42.5	57.0
+ sFT	874	3,189	9.0	113.7	20.1	**43.7**	N/A	73.9	**80.3**	19.2	47.9	62.9
+ wFT	**1,196**	**3,792**	9.0	104.8	19.3	43.2	N/A	**74.2**	**80.3**	19.6	50.4	64.6
+ wFT (BP decoding)	1,105	3,633	9.4	108.6	20.0	**43.7**	N/A	74.1	**80.3**	**20.6**	**50.6**	**64.9**
UpDown RL	563	3,161	9.5	**122.3**	**21.3**	**44.2**	N/A	74.2	80.6	20.6	50.2	65.7
+ sFT	1,222	3,805	9.2	115.3	20.7	44.1	N/A	74.9	**80.9**	24.6	56.2	70.9
+ wFT	**1,502**	**4,301**	9.1	100.4	19.2	43.0	N/A	75.0	80.7	26.1	57.4	71.4
+ wFT (BP decoding)	1,278	4,226	9.6	108.9	20.5	43.9	N/A	**75.1**	**80.9**	**26.4**	**58.6**	**73.6**
Transformer RL	713	3,432	9.2	**126.4**	**22.1**	**45.0**	N/A	75.0	81.2	25.4	56.3	69.8
+ sFT	1,496	3,953	9.1	118.4	21.4	44.6	N/A	75.7	**81.5**	30.2	62.7	75.8
+ wFT	1,836	4,268	9.1	102.2	19.8	43.2	N/A	75.9	81.3	32.2	64.3	76.8
+ wFT (BP decoding)	**2,004**	**4,392**	9.4	105.6	20.6	43.9	N/A	**76.1**	81.4	**32.8**	**66.1**	**79.0**

Table 3. 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. TIGEr scores are N/A as the TIGEr evaluation tool currently does not support evaluation on the MS COCO validation set.

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS+	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
VinVL RL	1,126	4,298	10.0	**140.9**	**25.2**	**46.1**	**75.7**	77.6	**83.3**	36.1	68.5	80.2
+ sFT	1,834	4,649	10.0	126.0	23.8	45.5	75.6	**78.2**	**83.3**	39.2	**72.1**	83.8
+ wFT	**1,852**	4,652	10.0	124.9	23.7	45.5	75.6	**78.2**	**83.3**	39.2	72.0	83.9
+ wFT (BP decoding)	1,734	**4,717**	9.8	122.4	23.5	45.2	**75.7**	**78.2**	**83.3**	**39.6**	**72.1**	**84.6**

Table 4. Test on the more recent captioning model. 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.

12. Effectiveness on More Recent Models

To further demonstrate the effectiveness of our methods, we tested our fine-tuning methods on a more recent captioning model, **VinVL** [23, 10]. VinVL boosts its performance through large-scale cross-modal pre-training. The significant performance improvements have made VinVL a popular captioning model and one of the most advanced captioning models available today [20, 21, 15].

We used the best-performing pre-trained model as our baseline: `coco_captioning_large_scst` model that is publicly available at `https://github.com/microsoft/Oscar/blob/master/VinVL_MODEL_ZOO.md#Image-Captioning-on-COCO`. Note that this model was trained with the standard RL [18].

As in the previous experiments, we applied our fine-tuning methods for one epoch only. We searched for the best learning rates for fine-tuning from $\{1\text{e-}5, 1\text{e-}6\}$, and the inverse-temperature hyperparameter β' of Eq. (7) from $\{0.01, 0.1, 1\}$. Other hyperparameters were set to the same as the baseline model. The best learning rate was 1e-5. The best β' was 0.01 for wFT with p_θ decoding and 1 for wFT with BP decoding. Note that sFT does not use β'.

Table 4 shows similar results as Table 1 *in the main paper*. Our methods significantly increased the vocabulary size from the baseline and accordingly enhanced the discriminativeness. The standard evaluation metrics also showed the same tendency. Although our models scored lower than the baseline in the conventional exact-matching metrics (CIDEr and SPICE), the gap became smaller in the more advanced soft-matching metrics (BERTS+ and TIGEr). In the state-of-the-art soft-matching metrics (CLIPS and RefCLIPS), our models achieved the same or even higher scores than the baseline. These results show that our methods are also effective on the more recent model. Moreover, these results further validate that our methods can switch any off-the-shelf RL models to discriminativeness-aware models while maintaining the overall quality of captions.

13. Comparison and Combination with More Recent Discriminativeness-Aware Models

Contemporaneous to our work, [3] showed that maximizing CLIPS-based reward enhanced discriminativeness significantly. In this section, we clarify the advantages of our methods over the CLIPS-based RL by comparing and combining our methods with it.

The pre-trained models of [3] are publicly available at `https://github.com/j-min/CLIP-Caption-Reward`. We used the transformer model trained with the standard CIDEr reward (**Transformer* RL (CIDEr)**; `clipRN50_cider`) and the one trained with the reward proposed by [3] (**Transformer* RL (CLIPS + Grammar)**; `clipRN50_clips_grammar`)[2]. The proposed reward is computed by the weighted sum of CLIPS and grammaticality scores.

[2] Note that `clipRN50` does not mean that the model used the CLIPS-based reward. It denotes that the model used CLIP [16] as the image encoder, unlike the other models tested in this paper.

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS+	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Transformer* RL (CIDEr)	691	3,650	9.5	**126.0**	**22.8**	**45.2**	74.6	75.8	81.6	27.1	57.2	70.6
+ sFT	1,265	4,071	9.1	122.9	22.2	**45.2**	**74.8**	76.4	**82.0**	31.4	62.0	75.0
+ wFT	**1,546**	4,337	9.0	111.3	21.0	44.2	74.5	76.5	81.8	31.6	63.3	75.7
+ wFT (BP decoding)	1,543	**4,471**	9.5	112.3	21.7	44.9	**74.8**	**76.9**	81.9	**34.0**	**65.4**	**78.4**
Transformer* RL (CLIPS + Grammar)	952	4,847	13.0	74.1	19.8	43.6	**75.0**	**79.2**	81.2	44.2	77.0	86.9
+ sFT	969	4,848	12.8	76.4	20.1	43.8	**75.0**	**79.2**	81.2	44.6	**77.3**	87.0
+ wFT	969	4,847	12.9	76.4	20.1	43.8	**75.0**	**79.2**	81.2	44.8	77.2	**87.1**
+ wFT (BP decoding)	**1,001**	**4,853**	12.2	**82.5**	**20.6**	**44.1**	**75.0**	**79.2**	**81.3**	**45.5**	77.2	**87.1**

Table 5. Test on the more recent discriminativeness-aware model. Transformer* used a different image encoder than the other transformer models tested in this paper. 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.

As in the previous experiments, we applied our fine-tuning methods for one epoch only. We searched for the best learning rates for fine-tuning from $\{1\text{e-}5, 1\text{e-}6, 1\text{e-}7\}$, and the inverse-temperature hyperparameter β' of Eq. (7) from $\{0.01, 0.1, 1\}$. Other hyperparameters were set to the same as the baseline model. The best learning rate for Transformer* RL (CIDEr) was 1e-5; the best β' was 0.1 for wFT with p_θ decoding and 1 for wFT with BP decoding. The best learning rates for Transformer* RL (CLIPS + Grammar) were 1e-6 for wFT with BP decoding and 1e-7 for the others; the best β' was 1 for wFT with both decoding methods. Note that sFT does not use β'.

Table 5 shows the results. Similar to the previous results, our methods significantly enhanced the vocabulary size and discriminativeness from the RL models while maintaining or even increasing the scores in the state-of-the-art soft-matching metrics. The CLIPS + Grammar reward also achieved the high discriminativeness compared with the standard CIDEr reward.

However, the improvement of the CLIPS-based RL came at the expense of the *conciseness* and overall quality of captions in contrast to our methods: compared to Transformer* RL (CIDEr), Transformer* RL (CLIPS + Grammar) significantly increased the sentence length and decreased scores in the standard evaluation metrics, including the state-of-the-art metric, RefCLIPS. Although increasing the sentence length is one way to describe images in detail, concise description is more desirable to convey the most characteristic information clearly and efficiently [19].

These results indicate that our methods and the CLIPS-based RL increased discriminativeness by different factors: more specific vocabulary and longer descriptions, respectively. In other words, the contribution of our methods is orthogonal to that of the CLIPS-based RL. To utilize the strength of each, we applied our methods to the CLIPS-based RL model. Although the CLIPS-based RL achieved the high discriminativeness and relatively large vocabulary size due to the longer sentences, our methods further enhanced the discriminativeness and vocabulary size. Surprisingly, our methods also improved the standard evaluation scores, including exact matching scores. This result suggests that our fine-tuning with ground-truth captions restored the overall quality of captions, which was degraded by over-optimization for CLIPS.

Another critical advantage of our methods is computational efficiency. Training of CLIPS-based RL took *one day using eight GPUs* [3], while ours only took *40 minutes using a single GPU*.

The above results conclude that our methods are orthogonal to the more recent discriminative image captioning method and have important advantages in conciseness and efficiency.

References

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

[2] Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *ICLR*, 2021.

[3] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022.

[4] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

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

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

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

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

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

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

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

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

[13] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Typical decoding for natural language generation. *arXiv preprint arXiv:2202.00666*, 2022.

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

[15] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*, 2022.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

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

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

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

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

[21] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

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

[23] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

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

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