Bright as the Sun: In-depth Analysis of Imagination-driven Image Captioning

(Supplementary Material)

Huyen Thi Thanh Tran[1] and Takayuki Okatani[1,2]

[1] RIKEN Center for AIP
[2] Graduate School of Information Sciences, Tohoku University
{tran,okatani}@vision.is.tohoku.ac.jp

In this supplementary material, we first analyze the accuracy of CScorer considering the impact of training loss function. Then, we investigate the effectiveness of using CScorer in the proposed model. Next, we provide additional discussions on our dataset. Finally, we present additional results of the image captioning models and the failure cases of the proposed model.

1 More Analysis of CScorer

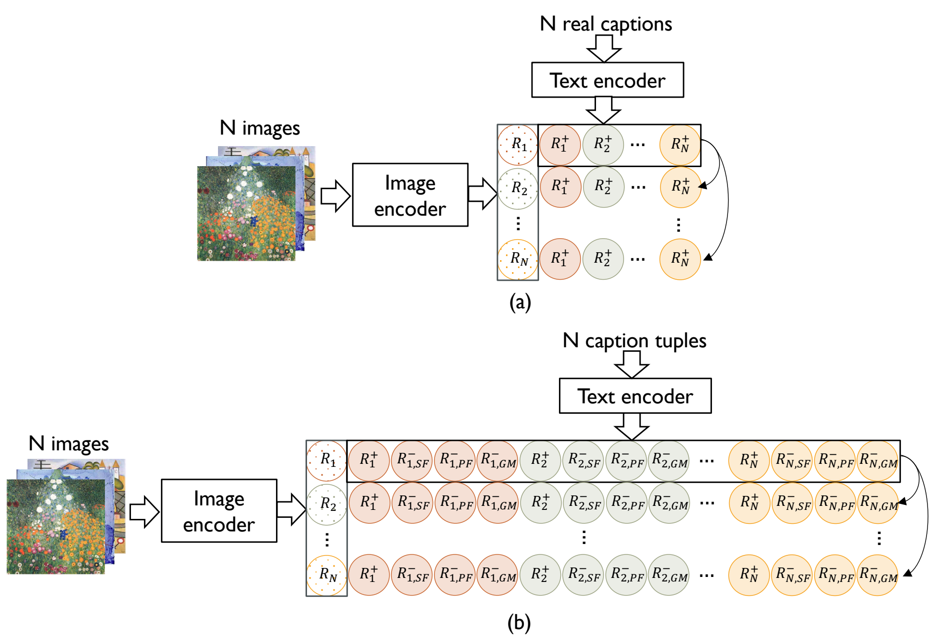


Fig. 1: Illustration of batch construction process for training CScorer: (a): Symmetric architecture [2], (b): Asymmetric architecture.

Table 1: Accuracy of CScorer using different loss functions.

Error type	Symmetry	Assymmetry							
		$o\mathcal{L}_{is}^{\mathcal{CE}}$	$o\mathcal{L}_{ip}^{\mathcal{CE}}$	$o\mathcal{L}_{g}^{\mathcal{CE}}$	$o\mathcal{L}_{s}^{\mathcal{CE}}$	$w.o\mathcal{L}_{is}^{\mathcal{CE}}$	$w.o\mathcal{L}_{ip}^{\mathcal{CE}}$	$w.o\mathcal{L}_{g}^{\mathcal{CE}}$	***Ours***
SF	0.831	0.912	0.872	0.517	0.884	0.891	0.896	0.896	**0.899**
PF	**0.794**	0.744	0.789	0.562	0.742	0.738	0.727	0.771	0.748
GM	0.608	0.574	0.645	**0.925**	0.869	0.875	0.869	0.839	0.867
Average	0.744	0.743	0.769	0.668	0.832	0.835	0.830	0.835	**0.838**

Figure 1 illustrates a comparison of the symmetric and asymmetric architectures, which can be used in the batch construction process for training CScorer. For the symmetric architecture, each input image I_i is paired with one real caption C_i^+. Assume that the batch size is N, there are $N \times N$ possible (image, caption) pairs passed through CScorer at one update [2]. Meanwhile, for the asymmetric architecture, each image I_i has four corresponding captions: one real caption C_i^+ and three fake captions, denoted as $C_{i,SF}^-$, $C_{i,PF}^-$, $C_{i,GM}^-$, leading to $N \times 4N$ possible (image, caption) pairs in one batch.

To train CScorer, we use the asymmetric architecture. Our objective is to maximize the scores of N pairs of images and their real captions while minimizing the scores of the other pairs. To do that, we train three MLP modules of CScorer (i.e., $\text{MLP}_s, \text{MLP}_p, \text{MLP}_g$) using four component losses, namely $\mathcal{L}_{is}^{\mathcal{CE}}$, $\mathcal{L}_{ip}^{\mathcal{CE}}$, $\mathcal{L}_{g}^{\mathcal{BCE}}$, and $\mathcal{L}_{s}^{\mathcal{CE}}$. To force these modules to learn error types, each of the first three losses is computed using one of the three error types as follows.

$$\mathcal{L}_{is}^{\mathcal{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(R_i^\top \text{MLP}_s(R_i^+))}{\sum_{j=1}^{N}(\exp(R_i^\top \text{MLP}_s(R_j^+)) + \exp(R_i^\top \text{MLP}_s(R_{j,SF}^-)))}\right), \quad (1)$$

$$\mathcal{L}_{ip}^{\mathcal{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(R_i^\top \text{MLP}_p(R_i^+))}{\sum_{j=1}^{N}(\exp(R_i^\top \text{MLP}_p(R_j^+)) + \exp(R_i^\top \text{MLP}_p(R_{j,PF}^-)))}\right), \quad (2)$$

$$\mathcal{L}_{g}^{\mathcal{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{1+\exp(-\text{MLP}_g(R_{i,GM}^-))}\right), \quad (3)$$

$$\text{MLP}_k(R) = W_k^1 \text{GELU}(W_k^0 R), \quad (4)$$

where $k \in \{s, p, g\}$; $W_s^0 \in \mathbb{R}^{4d\times d}$, $W_s^1 \in \mathbb{R}^{d\times 4d}$, $W_p^0 \in \mathbb{R}^{4d\times d}$, $W_p^1 \in \mathbb{R}^{d\times 4d}$, $W_g^0 \in \mathbb{R}^{4d\times d}$, and $W_g^1 \in \mathbb{R}^{1\times 4d}$ are learnable parameters. Similar to [2], we swap R_i and $\text{MLP}(R_i^+)$ to maximize the learning effectiveness.

The end-to-end loss $\mathcal{L}_{s}^{\mathcal{CE}}$ is calculated by

$$\mathcal{L}_{s}^{\mathcal{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(f_s(R_i, R_i^+))}{\sum_{j=1}^{N}(\exp(f_s(R_i, R_j^+)) + \sum_{e\in E}\exp(f_s(R_i, R_{j,e}^-)))}\right), \quad (5)$$

$$f_s(R, R^\pm) = (R^\top \text{MLP}_s(R^\pm) + R^\top \text{MLP}_p(R^\pm))/2 - \gamma \text{MLP}_g(R^\pm), \quad (6)$$

Table 2: Performance of the proposed model with and without using CScorer.

Models	**B1**	**B2**	**B3**	**B4**	**M**	**R**
Random	59.7	35.4	19.3	9.9	14.7	32.2
Ours	**61.6**	**37.9**	**22.3**	**13.4**	**15.8**	**33.2**

where $E = \{SF, PF, GM\}$.

Finally, the overall loss function is

$$\mathcal{L} = \mathcal{L}_{is}^{\mathcal{CE}} + \alpha\mathcal{L}_{ip}^{\mathcal{CE}} + \beta\mathcal{L}_{s}^{\mathcal{CE}} + \mu\mathcal{L}_{g}^{\mathcal{BCE}}, \tag{7}$$

where α, β, μ are trade-off parameters of the component losses.

To investigate the roles of the losses, we train CScorer using more seven difference cases of loss functions, denoted as $o\mathcal{L}_{is}^{\mathcal{CE}}$, $o\mathcal{L}_{ip}^{\mathcal{CE}}$, $o\mathcal{L}_{g}^{\mathcal{BCE}}$, $o\mathcal{L}_{s}^{\mathcal{CE}}$, $w.o\mathcal{L}_{is}^{\mathcal{CE}}$, $w.o\mathcal{L}_{ip}^{\mathcal{CE}}$, $w.o\mathcal{L}_{g}^{\mathcal{BCE}}$. We use only one of the four component losses in the first four cases. In the remaining cases, we remove one of the three losses $\mathcal{L}_{is}^{\mathcal{CE}}$, $\mathcal{L}_{ip}^{\mathcal{CE}}$, and $\mathcal{L}_{g}^{\mathcal{BCE}}$. Also, we compare the accuracy of the symmetric and asymmetric architectures. Table 1 shows the obtained results. It can be seen that the proposed scorer has the highest average accuracy. Compared to using all the four losses, the accuracy slightly reduces when using only $\mathcal{L}_{s}^{\mathcal{CE}}$, and significantly decreases in the other cases of $o\mathcal{L}_{is}^{\mathcal{CE}}$, $o\mathcal{L}_{ip}^{\mathcal{CE}}$, and $o\mathcal{L}_{g}^{\mathcal{BCE}}$. This indicates the essential role of the end-to-end loss $\mathcal{L}_{s}^{\mathcal{CE}}$. Also, the results show the benefit of adding $\mathcal{L}_{is}^{\mathcal{CE}}$, $\mathcal{L}_{ip}^{\mathcal{CE}}$, and $\mathcal{L}_{g}^{\mathcal{BCE}}$ to the loss function. The relative importance of the losses can be expressed as $\mathcal{L}_{s}^{\mathcal{CE}} > \mathcal{L}_{ip}^{\mathcal{CE}} > \mathcal{L}_{is}^{\mathcal{CE}} \approx \mathcal{L}_{g}^{\mathcal{BCE}}$. In comparison between the two architectures, the average accuracy of the asymmetric architecture is considerably higher (i.e., 0.838 vs. 0.744), demonstrating its effectiveness on imagination-driven caption evaluation.

2 Effectiveness of CScorer in Proposed Model

To investigate the effectiveness of using CScorer, we evaluate the performance of the model when randomly selecting between the literal and imagination-driven captions as the output caption, denoted as *Random*. Table 2 shows the obtained results. It can be seen that using CScorer indeed boosts the performance of the model. This suggests that it is beneficial to use CScorer to make decision in selecting which to generate between literal and imagination-driven captions.

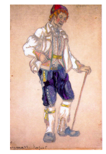
3 Discussions on Our Dataset

3.1 Diversity of Imagination-driven Descriptions

Figure 2 shows examples of human-generated imagination-driven descriptions collected from the ArtEmis dataset [1]. Given an image, annotators create remarkably diverse imagination-driven descriptions. For the image of Fig. 2(a),

- "***The vertical lines*** look like prison bars locking up someone"
- "***It*** reminds me of a baby lying down in a crib"
- "***The red figures make it*** look like they're dead bodies"
- "***This*** looks like a flag for a brand new country"
- "**I love the colors right here. The blue mixed in with the red and white,** reminds me of a buffalo bills football flag. It is a sweet look"
- "***It*** looks like what someone on ayahuasca would see while staring at the American flag"
- "***The red splashes of paint*** reminds me of pools of blood in a prison, which is frightening"

(a)



- "**This man in blue pants and bright red hat is amusing because his cheeks and jowels** remind me of a puppet..."
- "**The man's red cheeks and ginger fair makes him** seem like a comical clown for me."
- "**This** looks like a pirate with his cane. I think the colors and his face are amusing, especially because of his red cheeks."
- "**He** reminds me of a jester or some kind of eccentric entertainer who would be fun to watch."
- "**Something seems wrong,** as though the comic figure is about to pull a blade out of that cane."
- "**This man** looks as though he is disabled and unable to walk without a cane."
- "**The almost cherubic shape of the subject's face is comical to me, with rosy cheeks and wide eyes that also** remind me of medieval art."
- "**The mans ginger hair** reminds me of carrots I had for supper."

(b)

Fig. 2: Examples of human-generated imagination-driven descriptions collected from the ArtEmis dataset [1]. The texts in bold letters indicate the "subjects" while the rests indicate the "predicates" of captions.

the first annotator pays attention to vertical lines, *a visual entity*, and imagines prison bars, *an imaginary entity*. Other annotators think about a baby lying down in a crib, dead bodies, flags, or pools of blood. For the image of Fig. 2(b), although only one man is depicted in the image, the annotators imagine many things such as a puppet, a clown, a pirate, a comic figure, a disabled person, and even medieval art or carrots. These examples demonstrate that imagination-driven descriptions of images are enormously diverse due to the variety of attentive visual entities and imaginary entities generated from boundless human imagination capability.

3.2 Imagination-driven Caption Tuples in *IdC-II*

Figure 3 shows four examples of caption tuples in *IdC-II*. In the first example, to generate a fake caption of SF type, the visual entity *a young girl flies a kite* in the real caption is replaced by a wrong entity *two birds eating on what*. The original imaginary entity *shaped like a butterfly* is substituted with *shaped like a pirate boat* to create the fake caption of PF type. *butterfly* is excluded from the real caption to cause an incomplete caption of GM type. Observing these fake captions, we can see that they are improper and dissimilar from human-generated descriptions of images.

3.3 Commonly Used Words

Figure 4 depicts the wordclouds of the commonly used words in the imagination driven captions extracted from the four source datasets. The word sizes are

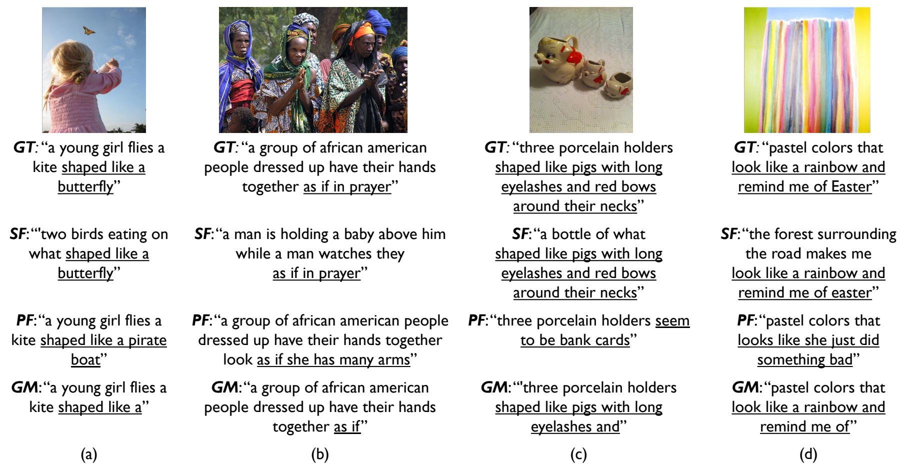


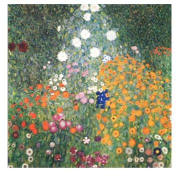
Fig. 3: Examples of caption tuples in *IdC-II*. *GT* denotes real captions. *SF*, *PF*, and *GM* denote fake captions corresponding to the three error types. The images and real captions are collected from (a) *MS COCO*, (b) *Flickr30K*, (c) *VizWiz*, and (d) *ArtEmis*.

Fig. 4: Wordclouds of commonly used words in imagination-driven captions extracted from (a) *MS COCO*, (b) *Flickr30K*, (c) *VizWiz*, and (d) *ArtEmis*.

linearly proportional to their frequencies. Since we used the list of keywords to detect imagination-driven captions, the words included in the keywords such as *look* and *seem* frequently appear in our dataset. Besides, for the captions collected from *MS COCO*, words related to objects or animals such as *cake*, *kite*, *bear*, and *dog* also have high occurrence frequencies. When using the source dataset of *Flickr30K*, a lot of captions include nouns referring to humans such as *man*, *people*, and *woman*.

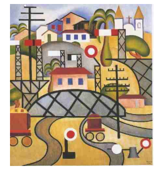
GT-1: "the person looks like a ghost or zombie from a horror film"
GT-2: "the man looks like he has blood on his face and it looks scary"
NN: "very strange how he or she looks all reddish brown in an otherwise sky colored blue and white world"
SAT: "the colors are very dark and gloomy"
$\mathcal{M}^2$: "the woman looks like she is in pain"
CLIPCap: "The colors are very dark and the face looks like it has been through a lot."
Oscar: "the colors are bright and cheerful, and the shapes are fun and whimsical. it's like a child's painting. it's fun to look at and try to"
OFA: "the blue and green colors make me feel sad"
1GEN: "the face looks like it is melting into the eyes"
LCGen (2GEN): "the colors are very sad and the face is sad"
IdCGen *(2GEN)*: "the man looks like a zombie is bleeding to be a lot of a scary"

(a)



GT-1: "it reminds me of an english garden full of flowers in bloom very cheerful"
GT-2: "portrait of a beautiful colorful field of flowers of all kinds"
NN: "this makes me feel relaxed and content because the shades of yellows browns and blue greens can be calming and peaceful"
SAT: "the bright colors of the flowers make me feel happy"
M2: "the bright colors of the flowers and the bright colors make me feel happy"
CLIPCap: "The flowers are so vibrant and colorful, it makes me feel happy."
Oscar: "the flowers are beautiful and the colors are bright and happy."
OFA: "the bright colors of the flowers make me feel happy"
1GEN: "the bright colors of the flowers are beautiful"
LCGen (2GEN): "the flowers are very colorful and the colors are beautiful"
IdCGen *(2GEN)*: "the bright colors and flowers remind me of a garden"

(b)



GT-1: "this piece reminds me of a childhood tv show with the bright colors and the rail carts and all the different traffic signs i love the block work to the city as if your expecting a cartoon train to come around the corner"
GT-2: "the suppleness of the shapes used for the objects reminds me of toy cars and those rugs that are made to look like maps"
NN: "the people look happy and are friendly with each other"
SAT: "the colors are bright and cheery"
$\mathcal{M}^2$: "the colors are bright and cheerful and the shapes are fun"
CLIPCap: "The colors are bright and fun."
Oscar: "the colors are bright and cheerful and the shapes are fun and whimsical."
OFA: "the bright colors of the buildings and the roads make me feel happy"
1GEN: "the colors are bright and playful"
LCGen (2GEN): "the colors are bright and the shapes are very fun"
IdCGen *(2GEN)*: "the bright colors and yellow colors make it look like a cartoon town"

(c)

Fig. 5: Examples of human-generated descriptions (*GT-1*, *GT-2*) and captions generated by the models

Since the *VizWiz* dataset consists of images taken by blind photographers in their daily lives, nouns of everyday objects such as *bottle* and *table* often appear in the extracted captions. In addition, it can be seen that annotators usually use adjectives to describe colors such as *white*, *black*, and *blue*. Since *ArtEmis* was created to research affective human experiences, the word of *feel* has an extremely high occurrence frequency in the captions extracted from this dataset.

4 Additional Results of Image Captioning Models

Examples of Generated Captions In Fig. 5, we show examples of human-generated imagination-driven descriptions and captions generated by all the considered models in our study. In addition, Fig. 6 depicts additional examples of captions generated by our model.

Failure Cases Based on the experimental results obtained in our study, the proposed model is found to generate imagination-driven captions closer to human-generated descriptions than the existing methods for standard image captioning. However, it still needs to be improved to generate diverse, precise, and comprehensive captions like humans do. Figure 7 shows some examples of failure cases of the proposed model. In the first example, the model detects a wrong visual entity (i.e., *the zebra*). In the second example, the model fails in forming an imaginary entity (i.e., *a fish*), which is different from those mentioned in the human-generated descriptions (i.e., *a letter c*, *an mri scan of a person's brain*). In the third example, the model makes a grammatical mistake: *a cherry blossoms*. In the last example, the model generates an incomplete caption, *the woman is playing piano and the piano*. These examples show the difficulties with generating precise and comprehensive captions like humans, especially when the images are artworks.

References

1. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11569–11579 (2021)
2. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)

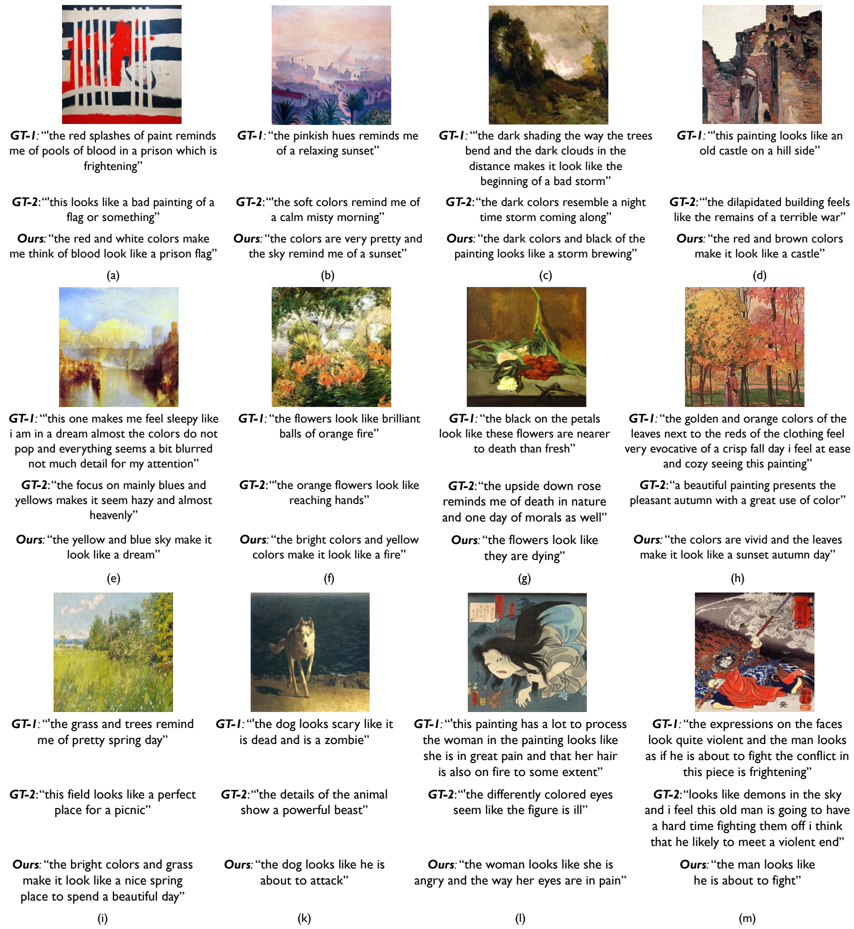


Fig. 6: Examples of human-generated descriptions (*GT-1* and *GT-2*) and captions generated by the proposed model.

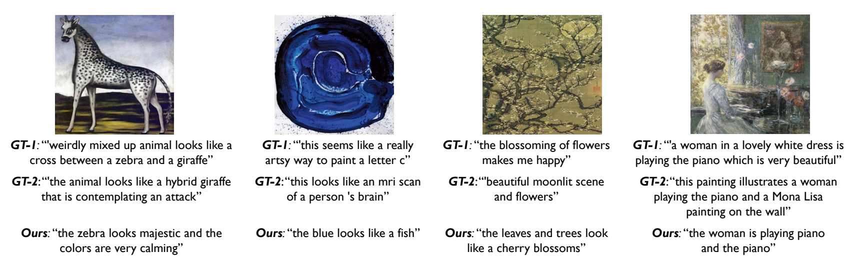


Fig. 7: Examples of failure cases of the proposed model. *GT-1* and *GT-2* denote human-generated descriptions, *Ours* denotes captions generated by our model.