

Tell as You Imagine: Sentence Imageability-aware Image Captioning

Kazuki Umemura¹, Marc A. Kastner^{2,1}, Ichiro Ide¹,
Yasutomo Kawanishi¹, Takatsugu Hirayama¹,
Keisuke Doman^{3,1}, Daisuke Deguchi¹, and Hiroshi Murase¹

¹ Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan

`umemurak@murase.is.i.nagoya-u.ac.jp`

`{ide, kawanishi, murase}@i.nagoya-u.ac.jp`

`{takatsugu.hirayama, ddeguchi}@nagoya-u.jp`

² National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, 101-8430 Japan

`mkastner@nii.ac.jp`

³ Chukyo University, 101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan

`kdoman@sist.chukyo-u.ac.jp`

Abstract. Image captioning as a multimedia task is advancing in terms of performance in generating captions for general purposes. However, it remains difficult to tailor generated captions to different applications. In this paper, we propose a sentence imageability-aware image captioning method to generate captions tailoring to various applications. Sentence imageability describes how easily the caption can be mentally imagined. This concept is applied to the captioning model to obtain a better understanding of the perception of a generated caption. First, we extend an existing image caption dataset by augmenting its captions' diversity. Then, a sentence imageability score for each augmented caption is calculated. A modified image captioning model is trained using this extended dataset to generate captions tailoring to a specified imageability score. Experiments showed promising results in generating imageability-aware captions. Especially, results from a subjective experiment showed that the perception of the generated captions correlates with the specified score.

Keywords: Vision and language · Image captioning · Psycholinguistics.

1 Introduction

In recent years, image captioning that automatically generates image descriptions is advancing. State-of-the-art image captioning methods [14, 28, 29] commonly perform at a visually descriptive level for general purposes, but do not consider the perception of the generated captions. Because of this, it is difficult to tailor the captions to different applications.

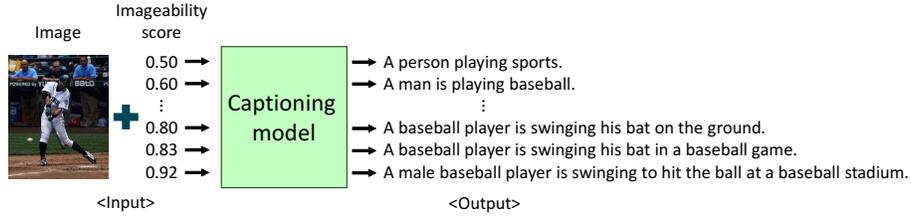


Fig. 1: Example of the proposed imageability-aware captioning. For an input image, the model generates diverse captions with different degrees of visual descriptiveness (i.e. imageability) tailoring to different target applications.

In the context of, e.g., news articles, the visual contents of an image are obvious and not needed to be captioned in detail. Accordingly, a caption should focus on additional information or context, rather than a pure visual description. News article captions also usually include many proper nouns [2]. If proper nouns in a caption are converted into a more abstract word (e.g., replacing *Donald Trump* with *A person*), the description becomes less detailed. Furthermore, such image captions usually include few adjectives [2]. Thus, captions in news articles are visually less descriptive. In contrast, captions targeting at visually impaired people would need a higher degree of visual description to be useful. Similarly, an image description used for image retrieval systems relies on a close connection between the visual contents of an image and the resulting caption.

In this research, we aim to generate captions with different levels of visual descriptiveness for such different applications. For this, we introduce the concept of “sentence imageability.” The concept of “imageability” originates from Psycholinguistics [19] and describes how easy it is to mentally imagine the meaning or the content of a word. Extending this idea to a sentence allows us to evaluate the visual descriptiveness of a caption. The proposed method generates diverse captions for an image corresponding to a given imageability score as shown in Fig. 1. Each caption is generated so that it contains a different degree of visual information, making them easier or harder to mentally imagine. This intrinsically tailors them to different target applications.

For this, we first augment the image captions in an existing dataset by replacing the words in them. Next, we propose a method to calculate the sentence imageability score based on word-level imageability scores. Then, we modify an existing image captioning model [29] to generate diverse captions according to sentence imageability scores.

The main contributions of this work are as follows:

- Proposal of a novel captioning method that generates captions tailoring to different applications by incorporating the concept of imageability from Psycholinguistics.
- Evaluation of the generated captions in a crowd-sourced fashion, and showing their imageability scores correlate to the mental image of users.

In Sec. 2 we briefly discuss the related work on Psycholinguistics and image captioning. Next, Sec. 3 introduces the proposed method on image captioning considering sentence imageability. We evaluate the proposed method through experiments in Sec. 4 and conclude the paper in Sec. 5.

2 Related Work

We will briefly introduce related work regarding psycholinguistic word ratings and image captioning.

Psycholinguistics: In 1968, Paivio et al. [19] first proposed the concept of imageability which describes the ease or difficulty with which “words arouse a sensory experience”, commonly represented as a word rating on the Lickert scale. Existing dictionaries [6, 23, 24] used in Psycholinguistics are typically created through labor-intensive experiments, often resulting in rather small corpora. For that reason, researchers have been working towards the estimation of imageability or concreteness using text and image data-mining techniques [10, 13, 17]. Imageability and similar word ratings have been used in multimodal applications like improving the understanding of text-image relationships [30].

Image Captioning: Image captioning is receiving great attention lately thanks to the advances in both computer vision and natural language processing. State-of-the-art models [14, 29] commonly take an attention guided encoder-decoder strategy, in which visual information is extracted from images by deep CNNs and then natural language descriptions are generated with RNNs.

In recent research, the goal to generate captions considering sentimental information, which not only contain the visual description of an image but also tailor to specific styles and sentiments, are receiving an increasing attention. Chen et al. [4] and Guo et al. [8] proposed methods to generate captions for combinations of four kinds of stylized captions: humorous, romantic, positive, and negative styles. Mathews et al. [16] considered the semantics and style of captions separately in order to change the style of captions into, e.g., story-like sentences. Most recently, Shuster et al. [25] proposed a method to better engage image captions to humans by incorporating controllable style and personality traits, such as sweet, dramatic, anxious, and so on. While these works aim to control caption styles, some other works [3, 5] aim to adjust the contents of the generated captions. However, although they focus on sentence variety, they do not consider the perception or imageability of the output.

Some methods [2, 20] targeting news images have been proposed. Similarly, we expect to be able to generate better image captions for news articles by generating captions with low imageability scores. Furthermore, the proposed method can not only generate captions similar to news image captions when targeting lower imageability scores, but also generate captions for other purposes.

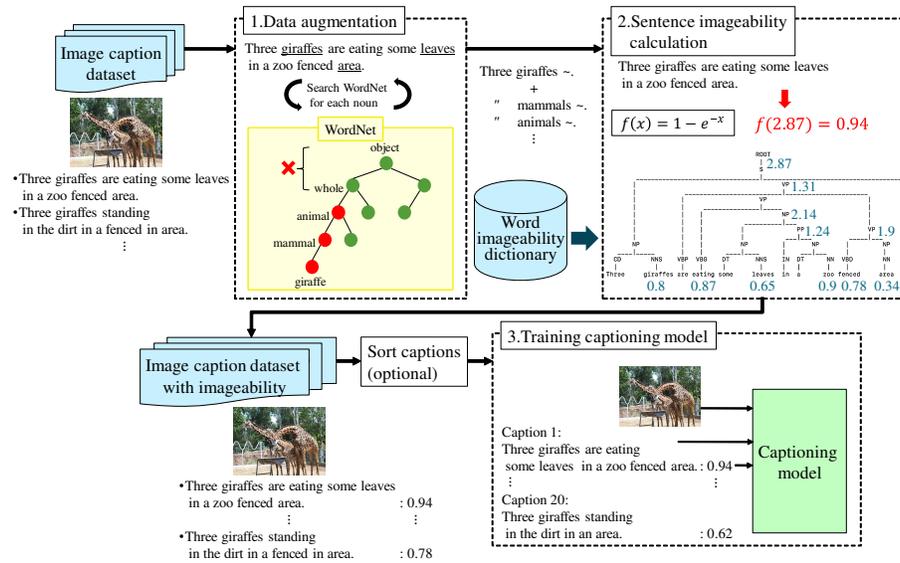


Fig. 2: Training process of the proposed image captioning model.

3 Image Captioning Considering Imageability

In this section, the proposed method is introduced in three steps: (1) Generation of an augmented image caption dataset with a higher variety of captions and different imageability scores for each. (2) Calculation of a sentence imageability score based on word-level imageability scores. (3) Incorporation of a sentence imageability vector into an image captioning model. The training process of the proposed image captioning model is illustrated in Fig. 2.

3.1 Data Augmentation

Existing datasets [12, 22] for image captioning commonly provide multiple annotated captions for each image. For example, MSCOCO [12] comes with five distinct captions for each image, differently describing the same contents. However, since our work requires captions with different degrees of visual descriptiveness, the variety of these captions is not sufficient; We require a variety based on different imageability scores, e.g., *human*, *male person*, and *teenage boy* all describe the same object, but result in different degrees of mental clarity if used in a caption. Thus, we augment the sentence variety of existing image-text pairs by replacing words in the original captions. All nouns in a caption are replaced with a selection of hypernyms using the WordNet [18] hierarchy. At most five closest hypernyms are used in order to avoid the word replacements getting too abstract or unrelated to the word from the original caption. Similarly, we do

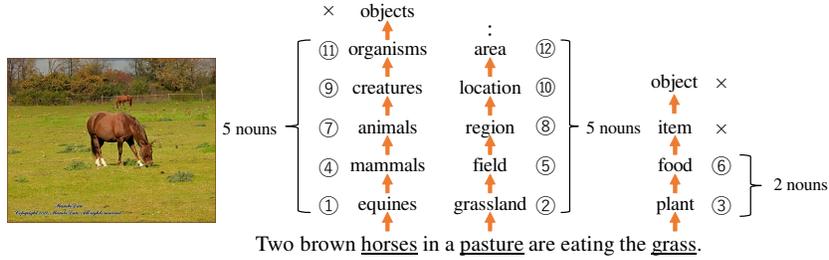


Fig. 3: Example of data augmentation. Each noun is replaced by a selection of differently abstract words, incorporating the WordNet hierarchy. This ensures a sentence variety with very different degrees of visual descriptiveness. The number next to each hypernym indicates the order of the replacement.

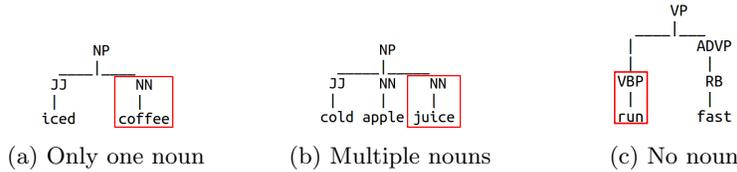


Fig. 4: Three relationship patterns for deciding the most significant words. The square indicates the selected word for each pattern.

not replace with words too close to the root of the WordNet hierarchy as they get too abstract. By this method, the dataset is augmented to contain a large variety of similar sentences with differently abstract word choices. An example of this method is illustrated in Fig. 3. If there are multiple nouns, the order of replacement occurs as exemplified in the figure. In the experiments, we use this order for sampling a subset of captions in case the augmentation generates too many of them.

3.2 Sentence Imageability Calculation

A sentence imageability score is calculated for each caption in the augmented dataset. Although the concept of word imageability has been part of works such as [19], there are few works that aim to determine the imageability of sentences. To be able to rate the descriptiveness of a caption, we propose to use a sentence imageability score of a caption. As there is no existing method for this, we introduce a way to calculate the sentence imageability score of a caption by using the imageability scores of words composing it. For this method, we assume that the imageability score of a word increases when being modified by other words in the same-level. For example, *coffee* gets less ambiguous when modified by the word *iced* before it (see Fig. 4a).

First, the tree structure of a given sentence is parsed using StanfordCore NLP [15]. We then take a bottom-up approach to calculate the imageability of sub-parts of the sentence based on word imageability scores, normalized to $[0, 1]$. Along the tree structure, the nodes in the same-level are weighted based on selecting the most significant word (Fig. 4). The significant word is selected according to three relationship patterns: (a) When there is only a single noun at the same depth, it is selected as the significant word for weighting. (b) When there is more than one noun at the same depth, the last one is selected as the significant word. (c) When there is no noun at the same depth, the first word of the sequence is selected as the significant word. Note that stop words and numerals are ignored.

Assuming the imageability score of the most significant word increases by other words modifying it, the imageability score of a sub-tree is calculated with

$$I = x_s \prod_{i=1(i \neq s)}^n (2 - e^{-x_i}), \quad (1)$$

where x_i ($i = 1, \dots, n | i \neq s$) is the score of each word and x_s is the score of the significant word as described above. The resulting score I is used recursively in a bottom-up manner in the parent node. Additionally, when a sub-tree forms a coordinate conjunction, the score I is calculated as the sum of the scores of their nodes. When there is only one node at the same depth of the tree, its score is directly transferred to its parent node. Finally, when reaching the root node of the parsing tree, the score is normalized to the scale of $[0, 1]$ by applying Equation 2, which represents the imageability score of the sentence.

$$f(x) = 1 - e^{-x} \quad (2)$$

3.3 Image Captioning

Based on a state-of-the-art captioning model incorporating attention [29], to consider the sentence imageability score of a caption, imageability feature vectors are added. The sentence imageability score for each caption is converted into the same dimensionality as the image feature and caption feature vectors to form an imageability feature vector \mathbf{A} . Given a ground-truth caption $\{\mathbf{t}_0, \mathbf{t}_1, \dots, \mathbf{t}_N\}$, image features \mathbf{I}_f extracted from a pre-trained ResNet network, and an imageability feature \mathbf{A} , a caption $c_i := \{\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_N\}$ is generated, where \mathbf{w}_i is a word vector for the i -th word as follows:

$$\begin{aligned} \mathbf{x}_t &= W_e \mathbf{w}_{t-1}, \quad t \in \{1, 2, \dots, N\}, \\ \mathbf{I}_t &= \text{Att}(\mathbf{h}_{t-1}, \mathbf{I}_f), \\ \mathbf{h}_t &= \text{LSTM}(\text{concat}(\mathbf{x}_t, \mathbf{I}_t, \mathbf{A}), \mathbf{h}_{t-1}), \\ \mathbf{w}_t &= \text{softmax}(W_l \mathbf{h}_t), \end{aligned} \quad (3)$$

where W_e, W_l are learnable parameters, \mathbf{h}_{t-1} is a hidden state of the previous iteration in LSTM. The hidden state \mathbf{h}_{t-1} and the image feature \mathbf{I}_f are input to

the Attention Network (Att) and yield the attention weighted vector \mathbf{I}_t of the image. Next, the attention weighted vector \mathbf{I}_t , the embedded word vector \mathbf{x}_t , and the imageability vector \mathbf{A} are concatenated. Lastly, the model is trained using the concatenated vector by optimizing Equation 4, minimizing the cross-entropy loss between the ground-truth word \mathbf{t}_i and the generated word \mathbf{w}_i .

$$L = - \sum_{i=0}^N q(\mathbf{t}_i) \log p(\mathbf{w}_i), \quad (4)$$

where p and q are the occurrence probabilities of \mathbf{w}_i and \mathbf{t}_i , respectively.

After training, we generate n caption candidates controlled by parameters W_e and W_l , and select the best caption with the smallest Mean Squared Error (MSE) regarding the sentence imageability scores calculated by the method in Sec. 3.2. For generating the image caption, we use the Beam Search which is a search algorithm that stores the top- b sequences (b : beam size) at each time step.

4 Evaluation

We evaluated the proposed method by conducting three experiments. First, the imageability scores of generated captions are evaluated. Second, the generated captions are evaluated by existing image captioning metrics. Lastly, the generated captions are evaluated through a crowd-sourced subjective experiment regarding their actual perception of visual descriptiveness.

4.1 Environment

Word Imageability Dataset: We use the word imageability dictionaries by Scott et al. [24] and Ljubešić et al. [13]. As the latter was predicted through datamining, the former is preferred if a word exists in both of them. Note that captions with a noun not available in the dictionaries are excluded from the datasets.

Image Caption Dataset: We use MSCOCO [12] as the base dataset, which we augment based on the proposed method. Two sampling methods are tested: (1) Sampling based on the order of caption augmentation. It reflects both the order of the nouns in a sentence, and then their respective WordNet hierarchy (*Without Sorting*). (2) Sampling after sorting captions by the calculated sentence imageability scores (*With Sorting*). Here, the augmented captions are sorted by their sentence imageability scores and the captions with the highest and the lowest imageability scores are selected in turn. By this, the model will be trained towards a higher diversity of imageability scores. Note that images with too few captions are excluded from the dataset.

Similarly to prior work, we employ Karpathy-splits [9] resulting in 113,287 images for training and 5,000 images each for validation and testing. After excluding images with an insufficient number of captions, 109,114 images for training, 4,819 images for validation, and 4,795 images for testing were left.

Table 1: Results of imageability analysis.

Captioning method	Sampling method	Caption Variety (\uparrow)	Imageability Range (\uparrow)	Average MSE (\downarrow)			Average RMSE (\downarrow)		
				Low	Mid	High	Low	Mid	High
Proposed	W/o Sorting	4.68	0.083	0.405	0.118	0.011	0.632	0.334	0.098
	With Sorting	4.63	0.182	0.338	0.089	0.014	0.573	0.276	0.107
Baseline	W/o Sorting	3.50	0.070	0.434	0.131	0.015	0.655	0.354	0.117
	With Sorting	3.26	0.164	0.378	0.103	0.022	0.607	0.300	0.142

Captioning Methods: Using the proposed captioning model, captions for nine levels of imageability in the range of $[0.1, 0.9]$ are generated. We regard the ranges of $[0.1, 0.3]$ as *Low*, $[0.4, 0.6]$ as *Mid*, and $[0.7, 0.9]$ as *High*. For the Beam Search, the beam size is set to $b = 5$. The feature vector has a dimensionality of 512. The proposed method first generates n caption candidates. Next, a candidate with the smallest MSE between the input and the predicted imageability score is chosen. For comparison, a baseline method is prepared, which is a simplified version where a single caption is generated instead of n candidates. We set $n = b$.

4.2 Analysis on the Sentence Imageability Scores

For analyzing the sentence imageability scores, we evaluate the number of unique captions (Caption Variety), the range of imageability scores ($\max - \min$; Imageability Range), as well as the MSE and Root Mean Squared Error (RMSE). The error is calculated between the input and the imageability scores of the generated caption calculated by the method introduced in Sec. 3.2.

The results are shown in Table 1. For all metrics, the results of the proposed method are better than the baseline method. The proposed method generates captions with both a large Caption Variety and a wider Imageability Range. Since the captions generated by Beam Search are all different, there is no identical caption candidate. We select the caption which has the closest imageability score with that of the target, i.e., having the smallest error between the predicted imageability score of the generated caption and the target score. The captions generated like this usually have larger caption variety than the baseline method. Similarly, the proposed method shows smaller MSE and RMSE. For the sampling method, the Imageability Range is wider, and there is smaller error in the low- and the mid-range imageability captions for the *With Sorting* sampling method. On the other hand, the error for high-range imageability captions is smaller for the *Without Sorting* sampling method. As the training data consists of a larger number of long, high-imageability captions, the model favors generating long captions. Due to this, the average error on mid- and high-range imageability scores is lower.

An example of the output of the proposed method is shown in Table 2. While each caption targets a different imageability score, the *Without Sorting* method outputs identical captions. In contrast, the *With Sorting* method produces results resembling the target imageability score.

Table 2: Example of the output of the proposed method corresponding to a given imageability score. The upper half of the generated captions was sampled using the *Without Sorting* method, while the lower half was sampled using the *With Sorting* method in the training phase. For comparison, the caption at the bottom row is generated by a state-of-the-art captioning method [29] not considering imageability.

Image	Imageability	Generated captions
	0.6	A cat laying on top of a device keyboard.
	0.7	A cat laying on top of a device keyboard.
	0.8	A cat laying on top of a device keyboard.
	0.9	A cat laying on top of a device keyboard.
	0.6	A placental is laying on a keyboard on a desk.
	0.7	A vertebrate is laying on a keyboard on a desk.
	0.8	A feline is laying on a keyboard on a desk.
	0.9	A cat is laying on a computer keyboard.
	—	A black and white cat laying next to a keyboard.

Table 3: Results of image captioning metrics.

Captioning method	Sampling method	BLEU-4 (\uparrow)			CIDEr (\uparrow)			ROUGE (\uparrow)			METEOR (\uparrow)			SPICE (\uparrow)		
		Low	Mid	High	Low	Mid	High	Low	Mid	High	Low	Mid	High	Low	Mid	High
Proposed	W/o Sorting	0.27	0.27	0.26	0.68	0.68	0.68	0.50	0.50	0.50	0.23	0.24	0.24	0.09	0.09	0.09
	With Sorting	0.25	0.27	0.26	0.59	0.50	0.64	0.49	0.50	0.50	0.23	0.23	0.24	0.09	0.09	0.09
Baseline	W/o Sorting	0.28	0.28	0.28	0.71	0.71	0.70	0.51	0.51	0.51	0.24	0.24	0.24	0.09	0.09	0.09
	With Sorting	0.25	0.27	0.28	0.61	0.65	0.65	0.49	0.51	0.51	0.23	0.24	0.24	0.09	0.09	0.09
Comp. [29]	—	0.30			0.91			0.52			0.25			0.18		

4.3 Evaluation of Image Captioning Results

We evaluate the proposed method in the general-purpose image captioning framework. For this, we look at the accuracy of the generated captions through standard metrics for image captioning evaluation, namely BLEU [21], CIDEr [27], ROUGE [11], METEOR [7], and SPICE [1]. For training this model five captions per image are used, and one caption per image is generated for testing.

The results are shown in Table 3. For comparison, results of a state-of-the-art captioning model which does not consider imageability [29] is shown, which is slightly better than the proposed method. This is because the proposed method focuses on caption diversification. The existing image captioning metrics evaluate the textual similarity to the ground truth, mainly evaluating the linguistic accuracy of the captions. In contrast, the proposed method aims for linguistic diverseness of each caption, intentionally generating different wordings for each caption. Thus, this will naturally decrease the textual similarity, as the generated captions will use different wordings than the ground truth. Following, we aim for a higher diversity of captions (discussed in Sec. 4.2) while maintaining a reasonable captioning quality (discussed here). Therefore, we regard these metrics as

Table 4: Results of subjective evaluation.

(a) Correlation		(b) Examples whose correlation failed	
ρ	#Images		
1.0	62 (31%)		An organism holding a banana in his hands.
0.5	86 (43%)		An organism holding a banana in his hand.
-0.5	42 (21%)		An equipment holding a banana in his hand.
-1.0	10 (5%)		A structure with a toilet and a sink.
			An area with a toilet and a sink.
			A white toilet sitting in a bathroom next to a structure.

not necessarily feasible to evaluate the proposed method. However, the results show similar performance with the general-purpose image captioning method while still considering an additional factor; the imageability of a caption.

4.4 Subjective Evaluation

In order to evaluate the actual perception of the generated captions, we evaluate three captions each for 200 randomly chosen images in a crowd-sourced subjective experiment on the Amazon Mechanical Turk platform⁴. As the majority of generated imageability scores is above 0.5, we focus on the upper half range of imageability scores. To compare how differently generated captions are perceived, we thus uniformly sample three generated captions, resulting in imageability scores of 0.5, 0.7, and 0.9. For each caption pair, we ask 15 English-speaking participants from the US to judge which of the presented two captions has a higher sentence imageability. Note that we don’t present the participants the image itself, but rather let them judge the imageability solely based on the textual contents of a caption. Based on the judgments, we rank the three captions for each image using Thurstone’s paired comparison method [26].

We calculate the Spearman’s rank correlation ρ between the target imageability scores and the actually perceived order obtained by asking participants of the crowd-sourced survey. The average correlation for all images was 0.37, which confirms that the perceived imageability of captions matches relationship between the target values to some extent. To further understand the results, we look into the distribution of the correlation shown in Table 4a. We found that, the number of “correctly” selected responses in line with the imageability scores was very high (approx. 65.8%). However, there are a few outliers with strong negative correlations. These results bring down the overall average performance for all images. Table 4b shows outliers whose captions have strong negative correlations. We found that in these cases, the generated captions were not describing the image contents correctly. In other cases, similar captions seem to have prevented the participants to decide which one had higher imageability.

⁴ <https://www.mturk.com/>

5 Conclusion

In this paper we proposed and evaluated an adaptive image captioning method considering imageability. By this method, we aim to control the degree of visual descriptiveness of a caption. For future work, we expect to generate captions with larger variety in terms of imageability. For that, we will try to augment captions in terms of their length as training data.

Acknowledgment

Parts of this research were supported by JSPS KAKENHI 16H02846 and MSR-CORE16 program.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic propositional image caption evaluation. In: Proc. 2016 European Conf. Comput. Vis., LNCS. vol. 9909, pp. 382–398 (2016)
2. Biten, A.F., Gomez, L., Rusinol, M., Karatzas, D.: Good news, everyone! Context driven entity-aware captioning for news images. In: Proc. 2019 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 12466–12475 (2019)
3. Chen, S., Jin, Q., Wang, P., Wu, Q.: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: Proc. 2020 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 9962–9971 (2020)
4. Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., Luo, J.: “Factual” or “Emotional”: Stylized image captioning with adaptive learning and attention. In: Proc. 2018 European Conf. Comput. Vis., LNCS. vol. 11214, pp. 527–543 (2018)
5. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: A framework for generating controllable and grounded captions. In: Proc. 2019 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8307–8316 (2019)
6. Cortese, M.J., Fugett, A.: Imageability ratings for 3,000 monosyllabic words. *Behav. Res. Methods Instrum. Comput.* **36**(3), 384–387 (2004)
7. Denkowski, M., Lavie, A.: METEOR universal: Language specific translation evaluation for any target language. In: Proc. 2014 EACL Workshop on Statistical Machine Translation. pp. 376–380 (2014)
8. Guo, L., Liu, J., Yao, P., Li, J., Lu, H.: MSCap: Multi-style image captioning with unpaired stylized text. In: Proc. 2019 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 4204–4213 (2019)
9. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3128–3137 (2015)
10. Kastner, M.A., Ide, I., Nack, F., Kawanishi, Y., Hirayama, T., Deguchi, D., Murase, H.: Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimed. Tools Appl.* **79**(25), 18167–18199 (2020)
11. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Proc. 2004 ACL Workshop on Text Summarization Branches Out. pp. 74–81 (2004)

12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO Common Objects in Context. In: Proc. 2014 European Conf. Comput. Vis., LNCS. vol. 8693, pp. 740–755 (2014)
13. Ljubešić, N., Fišer, D., Peti-Stantić, A.: Predicting concreteness and imageability of words within and across languages via word embeddings. In: Proc. 3rd Workshop on Representation Learning for NLP. pp. 217–222 (2018)
14. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proc. 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 7219–7228 (2018)
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proc. 52nd Annual Meeting Assoc. Comput. Linguist.: System Demonstrations. pp. 55–60 (2014)
16. Mathews, A., Xie, L., He, X.: Semstyle: Learning to generate stylised image captions using unaligned text. In: Proc. 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8591–8600 (2018)
17. Matsuhira, C., Kastner, M.A., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., Deguchi, D., Murase, H.: Imageability estimation using visual and language features. In: Proc. 2020 ACM Int. Conf. Multimed. Retr. pp. 306–310 (2020). <https://doi.org/10.1145/3372278.3390731>
18. Miller, G.A.: WordNet: A lexical database for English. *Comm. ACM* **38**(11), 39–41 (1995)
19. Paivio, A., Yuille, J.C., Madigan, S.A.: Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psycho.* **76**(1), 1–25 (1968)
20. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proc. 2020 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 10971–10980 (2020)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting Assoc. Comput. Linguist. pp. 311–318 (2002)
22. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. 15th IEEE Int. Conf. Comput. Vis. pp. 2641–2649 (2015)
23. Reilly, J., Kean, J.: Formal distinctiveness of high-and low-imageability nouns: Analyses and theoretical implications. *Cogn. Sci.* **31**(1), 157–168 (2007)
24. Scott, G.G., Keitel, A., Becirspahic, M., Yao, B., Sereno, S.C.: The Glasgow Norms: Ratings of 5,500 Words on Nine Scales. Springer (2018)
25. Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proc. 2019 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 12516–12526 (2019)
26. Thurstone, L.L.: The method of paired comparisons for social values. *J Abnorm Psychol.* **21**(4), 384–400 (1927)
27. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 4566–4575 (2015)
28. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3156–3164 (2015)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proc. 32nd Int. Conf. Mach. Learn. pp. 2048–2057 (2015)

30. Zhang, M., Hwa, R., Kovashka, A.: Equal but not the same: Understanding the implicit relationship between persuasive images and text. In: Proc. 2018 Br. Mach. Vis. Conf. No. 8. pp. 1–14 (2018)