Validity-Based Sampling and Smoothing Methods for Multiple Reference Image Captioning

Shunta Nagasawa[1] **Yotaro Watanabe**[2] **Hitoshi Iyatomi**[1]
[1]Department of Applied Informatics, Graduate School of Science and Engineering,
Hosei University, Tokyo, Japan
[2]PKSHA Technology Inc, Tokyo, Japan
{shunta.nagasawa@stu., iyatomi@}hosei.ac.jp
y_watanabe@pkshatech.com

Abstract

In image captioning, multiple captions are often provided as ground truths, since a valid caption is not always uniquely determined. Conventional methods randomly select a single caption and treat it as correct, but there have been few effective training methods that utilize multiple given captions. In this paper, we propose two training techniques for making effective use of multiple reference captions: 1) validity-based caption sampling (VBCS), which prioritizes the use of captions that are estimated to be highly valid during training, and 2) weighted caption smoothing (WCS), which applies smoothing only to the relevant words the reference caption to reflect multiple reference captions simultaneously. Experiments show that our proposed methods improve CIDEr by 2.6 points and BLEU4 by 0.9 points from baseline on the MSCOCO dataset.

1 Introduction

Image captioning is a very challenging task that requires recognizing and understanding the objects in the image and then verbalizing the recognition results using natural language. This task is expected to have a wide range of practical applications, including use in text-based image retrieval systems and providing assistance for the visually impaired (Lin et al., 2014; Gurari et al., 2020). With the development of the field of deep learning, research in the area has primarily focused on the end-to-end method, which consists of an encoder that extracts information from images and a decoder that generates a description from the extracted information (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015; Xu et al., 2015; Lu et al., 2017). For example, some of the recent models use pretrained object detection models (Ren et al., 2015; Liu et al., 2016; Anderson et al., 2018) and self-attention mechanisms (Huang et al., 2019; Cornia et al., 2020) for encoders or decoders.

Image captioning is often a multi-reference task where multiple reference captions are used for training. MSCOCO (Lin et al., 2014), one of the most famous datasets of image captions, has about five reference captions for each image. Some of these reference captions are subject to uncertainty due to speculation, and may differ in subject matter and wording. Such label variance may affect the training of the model and the evaluation of the generated captions. In typical training for conventional models, one caption is randomly selected by uniform sampling at each training epoch, which means the validity and variance of reference captions are not considered. In addition, reference captions that were not selected in the training epoch are treated as incorrect. To address this problem, Yi et al. (2020) proposed a new metric that correlates well with human evaluation by taking into account the variance of captions. However, to the best of our knowledge, appropriate training methods that consider such variation in captions have not yet been sufficiently studied.

In this paper, we propose a simple and effective training method that uses multiple reference captions to generate appropriate captions. The proposed training method consists of two techniques: validity-based caption sampling (VBCS), which selects highly valid reference captions, and weighted caption smoothing (WCS), which reflects multiple reference captions simultaneously in training. We define that a highly valid caption has common phrases among reference captions. In VBCS, the validity score for each reference caption is estimated based on similarities among the reference captions. When training the model, the training captions to be used in each epoch are sampled, one per image, according to this score. In addition, WCS improves the generality of the model by applying soft labels only for highly relevant words based on their validity scores. By effectively utilizing multiple captions, the proposed method improves

Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pages 36–41

CIDEr by 2.6 points and BLEU4 by 0.9 points in the evaluation experiments using the MSCOCO dataset. Main contributions of this paper include:

- Validity-based caption sampling (VBCS) allows us to prioritize captions that are considered to be highly valid.
- Weighted caption smoothing (WCS) allows multiple reference captions to be reflected in training simultaneously.
- The proposed VBCS and WCS are architecture-independent and highly versatile for image captioning and can be applied to other NLP multi-reference tasks.

2 Related Work

2.1 Selection of Training Data

Preparing highly reliable training data is important, however open datasets often contain incorrectly labeled or mislabeled samples. In a typical supervised task, one training label is assigned to each piece of training data. In this common setting, several methods have been proposed to improve the performance of the model by selecting suitable data for training from a large amount of labeled data (Reed et al., 2014; Northcutt et al., 2021).

In the multi-reference task, on the other hand, we expect to improve the performance by selecting appropriate labels from among them in the training. The choice can be deterministic, choosing the best one, or probability-based, depending on the characteristics of the data, such as likelihood (Hastings, 1970; Casella and George, 1992). The latter can be taken as a sampling problem. The proposed method prioritizes the sampling of highly valid captions to reduce the influence of less valid captions (i.e., noisy samples) and improves the performance.

2.2 Soft Label

Label smoothing (LS) (Pereyra et al., 2017) is a widely used soft labeling technique that prevents overfitting by creating soft supervised labels (i.e., adding a uniform distribution to each class of training labels). The introduction of LS has also been reported to improve the performance in language generation tasks, such as machine translation (Vaswani et al., 2017) and image captioning (Huang et al., 2019). Although the LS may contribute to the diversity of generated words, it treats all words in the vocabulary equally without taking into account their relevance to the image. Our WCS further improves the performance by constructing a novel soft label from multiple reference captions given to the image. Our soft label focuses on only relevant words among the reference captions based on the validity score.

3 Methodology

3.1 Validity-Based Caption Sampling (VBCS)

The proposed VBCS can take into account the validity and variance of reference captions. We define that a high validity caption has common phrases among reference captions, and assign a validity score to each reference caption. Let $R^{(i)} = \{\mathrm{ref}_1^{(i)}, \mathrm{ref}_2^{(i)}, \cdots, \mathrm{ref}_{K^{(i)}}^{(i)}\}$ be the reference caption set for image $I^{(i)} (i = 1, 2, \cdots, N)$. $K^{(i)}$ is the number of reference captions for image $I^{(i)}$. The similarity $s_j^{(i)}$ of the reference caption $\mathrm{ref}_j^{(i)}$ to other captions for image $I^{(i)}$ is calculated as follows:

$$\mathrm{s}_j^{(i)} = \frac{1}{K^{(i)} - 1} \sum_{\substack{k=1 \cdots K^{(i)}, \\ k \neq j}} f_{\mathrm{metric}}(\mathrm{ref}_j^{(i)}, \mathrm{ref}_k^{(i)}), \quad (1)$$

where f_{metric} is a metric of the similarity of the reference caption. Possible metrics that use word n-grams or longest match sequence include BLEU (Papineni et al., 2002), ROUGE-L (Lin et al., 2014), and CIDEr (Vedantam et al., 2015). Finally, the sampling probability $p_j^{(i)}$ for the j-th reference caption of image $I^{(i)}$ is calculated as follows:

$$p_j^{(i)} = \frac{\exp(\mathrm{s}_j^{(i)})}{\sum_{k=1}^{K^{(i)}} \exp(\mathrm{s}_k^{(i)})}. \quad (2)$$

This probability represents the validity of the reference caption and is referred to as the validity score in this paper. This allows us to prioritize training captions that have a high degree of similarity to other reference captions and are considered to be highly valid.

3.2 Weighted Caption Smoothing (WCS)

The proposed WCS solves the problem that unselected captions are treated as incorrect by introducing a soft label. Our soft label generated by WCS consists of only the words in each position of multiple reference captions, weighted by the validity score obtained by VBCS. This technique can reflect multiple captions in the training simultaneously.

	Evaluation Metric					
	B@1	B@4	M	R	C	S
Anderson et al. (2018)†	76.0 ±0.2	34.9 ±0.1	27.3 ±0.1	56.2 ±0.1	111.7 ±0.0	20.5 ±0.1
+ **LS**	76.1 ±0.1	35.2 ±0.2	27.4 ±0.0	56.3 ±0.1	112.8 ±0.3	20.6 ±0.2
+ **VBCS (ours)**	76.2 ±0.1	35.2 ±0.1	27.4 ±0.1	56.4 ±0.1	113.1 ±0.5	20.7 ±0.1
+ **WCS (ours)**	76.9 ±0.3	35.7 ±0.2	27.4 ±0.1	56.6 ±0.1	113.7 ±0.7	20.7 ±0.1
+ **VBCS + WCS (ours)**	**77.2 ±0.1**	**35.8 ±0.1**	**27.5 ±0.1**	**56.7 ±0.1**	**114.3 ±0.3**	**20.8 ±0.1**

Table 1: Summary of image captioning performance for MSCOCO test data, where B@N, M, R, C, and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr, and SPICE scores, respectively. For a robust evaluation, we run each method five times with different seeds. († are not the values given in the original paper, but the result of our best efforts to reimplement them.)

Specifically, our soft label $\tilde{y}_t^{(i)}$ used for predicting the t-th word of the image $I^{(i)}$ obtained by WCS is defined with two terms $y_{j,t}^{(i)}$ and $\hat{y}_t^{(i)}$:

$$\tilde{y}_t^{(i)} = (1-\alpha) y_{j,t}^{(i)} + \alpha \hat{y}_t^{(i)}, \tag{3}$$

where $y_{j,t}^{(i)}$ is the one-hot representation for the t-th word of the j-th reference caption selected by VBCS and α is hyperparameter that adjusts the smoothing. $\hat{y}_t^{(i)}$ is the weighted sum of the t-th word one-hot representation of multiple reference captions by the validity score and is obtained by:

$$\hat{y}_t^{(i)} = \sum_{k=1}^{K^{(i)}} p_k^{(i)} y_{k,t}^{(i)}. \tag{4}$$

Here, the length of each reference caption is padded or cropped according to the length of $y_j^{(i)}$.

The main difference between WCS and LS is the number of words to be smoothed. In our WCS, smoothing is not done uniformly for all words, but only for words that are in the same position in the assigned reference caption, weighted individually according to their validity score (i.e., words that are highly relevant).

4 Experiment

4.1 Dataset

We used the MSCOCO 2014 caption dataset (Lin et al., 2014), which contains 123,287 images labeled with five captions each. The "Karpathy" data split (Karpathy and Fei-Fei, 2015) was used for performance comparisons, and 5,000 images were used for validation, 5,000 images for testing, and the rest for training. As for pre-processing, we converted all sentences to lower case and dropped any words that occurred less than five times. To evaluate caption quality, we used several standard metrics, such as BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016).

4.2 Models

For our evaluation, we used the Up-Down (Anderson et al., 2018) model as a baseline, which has a typical structure in the field of image captioning and has been reported to be highly accurate. We compared the following training methods: +**LS** with its uniform smoothing for all words; +**VBCS**, which prioritizes highly valid reference captions for training based on the validity score; +**WCS** with smoothing for highly relevant words based on the validity score; and +**VBCS**+**WCS**, which is our proposed method. To ensure robust evaluation, we ran each method five times with different seeds.

4.3 Implementation Details

In the Up-Down model, we used the Faster-RCNN model (Ren et al., 2015), which was pre-trained with ImageNet (Deng et al., 2009) and Visual Genome (Krishna et al., 2017), as a content vector generator. We used beam search when generating captions, and set the beam size to 5. In this study, we decided to select CIDEr for f_{metric}, as it is the most widely used in image captioning and is capable of focusing on the importance of caption phrases. In Section 5.2, we will discuss the effectiveness of other metrics for f_{metric}. The hyperparameter of LS was set to 0.2 according to Huang et al. (2019). This corresponds to α when $\hat{y}_t{}^{(i)}$ is regarded as a soft label equal to all words in Eq 3. In WCS, α was set to 0.2 for comparison.

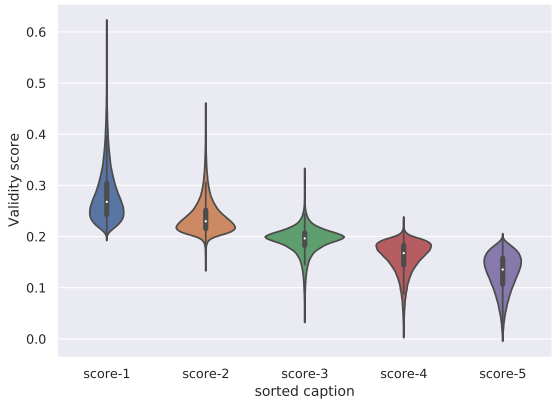

Figure 1: Distribution of the sorted validity scores in descending order.

5 Results and Discussion

5.1 Quantitative Analysis

Table 1 demonstrates the performance of our proposed method with other comparable models. With the introduction of efficient caption sampling, our VBCS improved performance in all metrics against the baseline. In particular, the CIDEr score improved by 1.4 points. This confirmed that sampling using the validity scores contributes to improving the score for each metric. Figure 1 shows the distribution of the validity scores in descending order using the violin plot. Since the validity of each reference caption is different, the distribution from the validity score is very different from the commonly used uniform distribution.

Our WCS outperformed LS on all metrics and was 0.5 and 0.9 points higher on BLEU4 and CIDEr, respectively. Since WCS smooths only a limited number of relevant words, we believe that it can learn more efficiently than LS, which smooths all words uniformly. The proposed techniques (+ VBCS + WCS) scored the highest on all the metrics. The improvements in BLEU4, ROUGE-L, and CIDEr, which are based on n-grams and longest matching sequence are particularly clear.

5.2 Effect of Hyperparameters

In this section, we investigate the impact of hyperparameters in our proposed methods.

Effect of f_{metric} for Validation Data Table 2 demonstrates the performance with the validation data, where BLEU4, ROUGE-L, and CIDEr were applied to f_{metric}. Regardless of the choice of f_{metric}, the proposed method produces results equal to or better than baseline. These results indicate that CIDEr is superior to the others and can capture more important phrases than other metrics.

	Evaluation Metric					
f_{metric}	B@1	B@4	M	R	C	S
baseline	75.8	34.7	27.2	56.1	109.4	20.1
BLEU4	**77.0**	**35.8**	27.2	**56.6**	111.4	20.3
ROUGE-L	76.6	35.2	27.2	56.5	110.7	20.4
CIDEr	76.7	35.4	**27.4**	**56.6**	**112.0**	**20.5**

Table 2: Comparison of scores for validation data under different f_{metric} choices in VBCS.

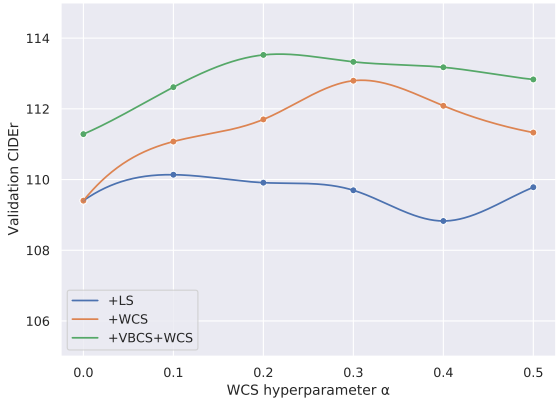

Figure 2: The effect of α, a smoothing hyperparameter of WCS for validation data. The proposed method achieves higher performance than LS with any α.

Effect of Hyperparameter in WCS Figure 2 demonstrates the effect of the hyperparameter α on the validation data in WCS. Our proposed +VBCS+WCS with $\alpha = 0.2$ performed the best. Since WCS applies to smooth to a limited number of words, it results in higher scores than those of LS with any α.

6 Conclusion and Future Works

In this paper, we proposed two novel techniques called VBCS and WCS that effectively utilize multiple references in image captioning tasks, and demonstrated their advantages. The former determines a sampling probability (i.e., validity score), for each caption based on similarities among the reference captions. The latter simultaneously reflects multiple reference captions in the training. In the future, we would like to consider the grammar in WCS and, extend the proposed method to be adaptable to reinforcement learning.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceeding of the European Conference on Computer Vision*, pages 382–398.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

George Casella and Edward I George. 1992. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Proceedings of the IEEE European Conference on Computer Vision*, pages 417–434.

Wilfred Keith Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

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 *Proceedings of the European Conference on Computer Vision*, pages 740–755.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*, pages 21–37.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–383.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *CoRR preprint arXiv:1911.00068v3*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR preprint arXiv:1701.06548*.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *CoRR preprint arXiv:1412.6596*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

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 *International Conference on Machine Learning*, pages 2048–2057.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994.