# Deneb: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning (Supplementary Material)

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura

Keio University, Japan
{k2matsuda0, yuiga, komei.sugiura}@keio.jp

## A Additional Related Work

The FOIL dataset, specifically designed to assess metric robustness against hallucination, is derived from the COCO dataset and comprises approximately 200,000 image-caption pairs with a mix of correct and hallucinated captions. Flickr8K-Expert [7], Flickr8K-CF, Composite [1], and Polaris [13] each consist of human Likert-scale judgments at the level of each image-caption pair. Specifically, Flickr 8K-Expert comprises 17,000 human judgments across 5,664 images, with captions rated on a four-point scale. Flickr8K-CF offers 145,000 human judgments collected from CrowdFlower, covering over 48,000 image-caption pairs. The Composite dataset, sourcing image-caption pairs from COCO [10], Flickr8K [7], and Flickr30K [17], contains approximately 12,000 human judgments. The Polaris dataset is tailored for the training and evaluation of metrics and consists of a diverse collection of captions, sourced from ten modern image captioning models and accompanied by approximately 130,000 human judgments.

## B Implementation Details

We divided the Nebula dataset into training, validation, and test sets, containing 26,382, 3,298, and 3,298 samples, respectively. We used the training set to train our model, the validation set for hyperparameter tuning, and the test set for evaluating the model's performance.

Table A shows the experimental settings of the proposed metric. We employed early stopping for model training with a focus on Kendall's $\tau$. This process entailed monitoring Kendall's $\tau$ on the validation set at each epoch. The training process was halted if Kendall's $\tau$ on the validation set did not show any improvement for a single epoch. Subsequently, we evaluated the model's performance using the test set.

Table A: The experimental settings for the proposed metric.

| | |
|---|---|
| Optimizer | Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) |
| Learning rate | $5.0 \quad 10^{-6}$ |
| Batch size | 16 |
| Loss function | Huber loss ($\delta = 0.5$) |

Our model had approximately 133 million trainable parameters. We trained our model on a Tesla A100 GPU and measured the inference time on the GeForce RTX 3090 with 24 GB of memory and an Intel Core i9 12900K with 64 GB of memory. The training phase was completed in approximately 2 hours, and the inference time per sample on GeForce RTX 3090 was approximately 22 ms.

## C Nebula Dataset

To construct the Nebula dataset, we employed ten standard image captioning models. These models include: SAT [16], $M^2$-Transformer [4], VinVL [18], GRIT [11], $\text{BLIP}_{\text{base}}$, $\text{BLIP}_{\text{large}}$ [9], GIT [14], OFA [15], BLIP-2 $_{\text{an}}$, and $\text{BLIP-2}_{\text{opt}}$ [8]. Here, $\text{BLIP}_{\text{base}}$ and $\text{BLIP}_{\text{large}}$ represent variants of BLIP that utilizes ViT-B and ViT-L [5], respectively. BLIP-2 $_{\text{an}}$ and $\text{BLIP-2}_{\text{opt}}$ are variants of BLIP-2 that employs Flan-T5 [3] and OPT [19] as their large language models, respectively.

The images in the Nebula dataset were sourced from the validation sets of the MS-COCO [10] and nocaps [2] datasets. MS-COCO was selected as it is a standard dataset for image captioning, whereas nocaps was selected for its greater diversity of classes compared to MS-COCO. The validation sets of MS-COCO and nocaps were chosen to avoid potential data leakage that could occur when using their training sets, particularly in terms of evaluating an image captioning model trained on MS-COCO. Furthermore, their test sets were not used because they lacked the reference captions necessary for multifaceted metrics.

We instructed the annotators to assess the quality of the captions from the perspectives of fluency, relevance, and descriptiveness. For fluency, they assessed the grammatical correctness of captions, deducting points for each grammatical error. For relevance, they evaluated whether the caption was closely related to the image and deducted points for irrelevant words. For descriptiveness, they assessed how comprehensively and accurately the caption describes the content of the image.

The Nebula dataset comprises 32,978 images and 32,978 human judgments collected from 805 annotators, and contains approximately three times more images than the Polaris dataset. The total number of references is 183,472, with a vocabulary size of 32,870, a total word count of 1,945,956, and an average sentence length of 10.61 words. The total number of candidates is 32,978, with a vocabulary size of 3,695, a total word count of 288,922, and an average sentence length of 8.76 words. All sentences are in English.

Table B: Categorization of failed samples.

| Error Type | #Error |
|---|---|
| Focus Area Discrepancy | 40 |
| Caption Accuracy Deficiency | 28 |
| Caption Detail Insufficiency | 16 |
| Grammatical Error | 8 |
| Annotation Error | 4 |
| Others | 4 |

## D Additional Qualitative Analysis

Figs. A and B present additional qualitative results from the Nebula and FOIL datasets, respectively. In our analysis, we compared the performance of the Deneb model with three representative metrics: CIDEr [12] (*classic*), CLIP-S [6] (*reference-free*), and Polos [13] (*pseudo-multifaceted*). Specifically, these methods have a tendency to overestimate the quality of instances where $x_{\text{cand}}$ were inappropriate but contained words related to the image. This discrepancy primarily stems from their limited capability to effectively compare candidates with multifaceted references and their significant reliance on the alignment of image and language features. In contrast, Deneb consistently assigned low evaluation scores to such captions and aligned closely with human judgments. These results show its effectiveness and robustness against hallucinations.

## E Error Analysis

To investigate the limitations of the proposed method, we analyzed the worst 100 samples with the largest absolute differences between $\hat{y}$ and $y$. We defined samples that satisfy $|y - \hat{y}| > 0.25$ as failure cases. Within the test set of Nebula dataset, a total of 503 samples were identified as failure cases.

Table B categorizes the failure cases. The causes of failure can be grouped into six main categories:

Focus Area Discrepancy: This category pertains to samples where our metric incorrectly scores captions that focus on different areas than the references.

Caption Accuracy Deficiency: This category refers to samples where our metric inappropriately scores captions with incorrect expressions.

Caption Detail Insufficiency: This category pertains to samples where our metric outputs inappropriate scores when the candidate lacks details.

Grammatical Error: This category refers to samples where our metric outputs inappropriate scores for captions containing grammatical errors.

Annotation Error: This category includes samples where the human judgment was inappropriate.

Others: This category covers other types of errors that do not fall into the aforementioned categories

Table B shows that the main bottleneck was errors due to differences in areas of focus. In samples corresponding to these errors, $\boldsymbol{x}_{\mathsf{cand}}$ describes elements absent in $\boldsymbol{x}_{\mathsf{ref}}^{(1)}$, suggesting that the proposed metric may not effectively capture the relationship between $\boldsymbol{x}_{\mathsf{cand}}$ and the local region of $\boldsymbol{x}_{\mathsf{img}}$. Consequently, the introduction of a mechanism to extract the relationship between features in local image regions and language features, as suggested in [20], is anticipated to offer a possible solution. In future work, we plan to extend Deneb by introducing a mechanism to extract the relationship between features in local image regions and language features, as suggested in [20].

In future work, we plan to extend Deneb by introducing a mechanism to extract the relationship between features in local image regions and language features, as suggested in [20].

## References

1. Aditya, S., Yang, Y., Baral, C., Fermüller, C., Aloimonos, Y.: From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. arXiv preprint arXiv:1511.03292 (2015)
2. Agrawal, H., Desai, K., et al.: nocaps: Novel Object Captioning at Scale. In: ICCV. pp. 8948–8957 (2019)
3. Chung, H., Hou, L., Longpre, S., Zoph, B., Tay, Y., et al.: Scaling Instruction-finetuned Language Models. arXiv preprint arXiv:2210.11416 (2022)
4. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-Memory Transformer for Image Captioning. In: CVPR. pp. 10578–10587 (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)
6. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP. pp. 7514–7528 (2021)
7. Hodosh, M., et al.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. JAIR 47, 853–899 (2013)
8. Li, J., Li, D., et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: ICML (2023)
9. Li, J., et al.: BLIP: Bootstrapping Language-image Pre-training for Unified Vision-language Understanding and Generation. In: ICML. pp. 12888–12900 (2022)
10. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., et al.: Microsoft COCO: Common Objects in Context. In: ECCV. pp. 740–755 (2014)
11. Suganuma, M., Okatani, T., et al.: GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features. In: ECCV. pp. 167–184 (2022)
12. Vedantam, R., Zitnick, L., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. In: CVPR. pp. 4566–4575 (2015)
13. Wada, Y., Kanta, K., Daichi, S., Sugiura, K.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In: CVPR (2024)
14. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., et al.: GIT: A Generative Image-to-text Transformer for Vision and Language. TMLR (2022)

15. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-sequence Learning Framework. In: ICML. pp. 23318–23340 (2022)
16. Xu, K., Ba, J., Kiros, R., et al.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: ICML. pp. 2048–2057 (2015)
17. Young, P., et al.: From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. TACL 2, 67–78 (2014)
18. Zhang, P., Li, X., Hu, X., et al.: VinVL: Revisiting Visual Representations in Vision-language Models. In: CVPR. pp. 5579–5588 (2021)
19. Zhang, S., Roller, S., Goyal, N., Artetxe, M., et al.: OPT: Open Pre-trained Transformer Language Models. arXiv preprint arXiv:2205.01068 (2022)
20. Zhong, Y., Yang, J., Zhang, P., Li, C., et al.: RegionCLIP: Region-based Language-image Pretraining. In: CVPR. pp. 16793–16803 (2022)

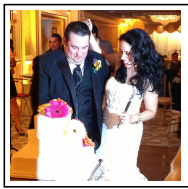
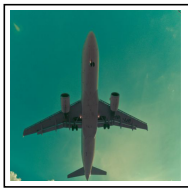
| | | | |
|---|---|---|---|
|  | $x_{\text{ref}}^{(1)}$: "The bride and groom are cutting the cake."<br>$x_{\text{cand}}$: "a bride and groom cutting their wedding cake."<br>**Human: 1.0** | CIDEr CLIP-S Polos<br>**2.14 0.59 0.83** | Deneb<br>**0.87** |
|  | $x_{\text{ref}}^{(1)}$: "a pedestrian walk sign among these billboards signs"<br>$x_{\text{cand}}$: "a pedestrian crossing sign on a pole in front of a building"<br>**Human: 1.0** | CIDEr CLIP-S Polos<br>**0.62 0.57 0.62** | Deneb<br>**0.79** |
|  | $x_{\text{ref}}^{(1)}$: "A stop sign on the corner of the street"<br>$x_{\text{cand}}$: "the building is orange"<br>**Human: 0.75** | CIDEr CLIP-S Polos<br>**0.00 0.43 0.30** | Deneb<br>**0.54** |
|  | $x_{\text{ref}}^{(1)}$: "A half eaten sandwich sitting on a wrapper."<br>$x_{\text{cand}}$: "a man in a plaid shirt eating a sandwich"<br>**Human: 0.0** | CIDEr CLIP-S Polos<br>**0.46 0.43 0.49** | Deneb<br>**0.07** |
|  | $x_{\text{ref}}^{(1)}$: "A large airplane flies overhead through the sky."<br>$x_{\text{cand}}$: "a woman flying a kite in a clear blue sky"<br>**Human: 0.0** | CIDEr CLIP-S Polos<br>**0.51 0.44 0.38** | Deneb<br>**0.02** |
|  | $x_{\text{ref}}^{(1)}$: "person running along the beach flying a kite"<br>$x_{\text{cand}}$: "a little boy playing with a tennis racket on the beach"<br>**Human: 0.0** | CIDEr CLIP-S Polos<br>**0.35 0.55 0.44** | Deneb<br>**0.09** |

Fig. A: Additional qualitative examples from the Nebula dataset. Existing metrics, such as CIDEr [12], CLIP-S [6], and Polos [13] do not closely align with human evaluations. Specifically, these methods have a tendency to overestimate the quality of instances where $\boldsymbol{x}_{\text{cand}}$ are inappropriate but contain words related to the image. In contrast, Deneb appropriately assigns lower scores to these instances, thereby demonstrating a more accurate reflection of their quality.