

This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version. The content of this paper is identical to the content of the officially published ACCV 2024 LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura

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

Abstract. In this work, we address the challenge of developing automatic evaluation metrics for image captioning, with a particular focus on robustness against hallucinations. Existing metrics are often inadequate for handling hallucinations, primarily due to their limited ability to compare candidate captions with multifaceted reference captions. To address this shortcoming, we propose DENEB, a novel supervised automatic evaluation metric specifically robust against hallucinations. DENEB incorporates the Sim-Vec Transformer, a mechanism that processes multiple references simultaneously, thereby efficiently capturing the similarity between an image, a candidate caption, and reference captions. To train DENEB, we construct the diverse and balanced Nebula dataset comprising 32,978 images, paired with human judgments provided by 805 annotators. We demonstrated that DENEB achieves state-of-the-art performance among existing LLM-free metrics on the FOIL, Composite, Flickr8K-Expert, Flickr8K-CF, Nebula, and PASCAL-50S datasets, validating its effectiveness and robustness against hallucinations. Project page at https://deneb-project-page-nc03k.kinsta.page/.

Keywords: vision and language \cdot hall ucination \cdot image captioning \cdot metrics

1 Introduction

Image captioning has been extensively researched and applied in various social applications, such as the assistance of visually impaired individuals, the analysis of medical images, and the generation of explanations in robotics [2, 4, 13, 14, 19, 33, 36]. In scenarios where 'AI safety' is paramount, generating appropriate and reliable captions is crucial to avoiding the misrepresentation of the content of images. In particular, erroneous captions that include words not depicted in the image, commonly referred to as 'hallucinations', are a prevalent issue in image captioning [39, 42]. However, a significant issue exists in current image captioning research: models failing to address hallucinations are often wrongfully overrated, as existing evaluation metrics predominantly focus only on correlation with human judgment, and overlook critical flaws. This misalignment is especially problematic in the context of social applications where reliability is



Fig. 1: Overview of DENEB. Our metric is designed to effectively evaluate hallucinated captions, which is crucial in scenarios where 'AI safety' is paramount. Unlike existing metrics such as CIDEr [46] and Polos [48], which often fail to distinguish between correct and hallucinated captions, DENEB demonstrates improved robustness by assigning lower scores to hallucinated captions than correct captions.

the issue of hallucinations. In fact, some studies have demonstrated that while most *data-driven* metrics [15,20,22,51,52] correlate well with human judgments, they are less effective in addressing hallucinations [15,40,46,48].

Hallucination is a prevalent issue in text generation models, including Large Language Models (LLMs) [6,21,25,31,32,42,45]. Despite their rapid advancement and adoption in various societal applications, LLMs often generate hallucinated text. This suggests that most LLMs are internally unable to evaluate hallucinations and that assessing hallucinations is a particularly challenging task. The FOIL benchmark [42] has been introduced to evaluate the robustness of metrics against hallucinations in image captioning. Several studies [15, 20, 40, 48] have revealed that most metrics still fall short of human performance on this benchmark, despite their strong correlation with human judgments.

Automatic evaluation metrics for image captioning can be broadly classified into four categories: classic metrics, reference-free metrics, pseudo-multifaceted metrics, and multifaceted metrics. Classic metrics [3, 5, 26, 35, 46, 47] are traditional approaches primarily based on n-grams and/or scene graphs. Despite their widespread use, they often provide low correlation with human judgments, leading to the development of data-driven metrics such as reference-free and pseudomultifaceted metrics. Reference-free metrics [15,23,40], which use images and do not rely on reference captions, correlate better with human judgments than classic metrics. However, they face challenges in robustness against hallucinations. This issue mainly stems from their heavy dependence on the alignment between image and language features, which has shown to be insufficiently effective in accurately grounding specific local areas within images. In fact, some studies have indicated that even the classic metric CIDEr [46] outperforms CLIP-S in terms of robustness against hallucinations [15]. In contrast to *classic* and *reference-free* metrics, *pseudo-multifaceted* metrics [15, 20, 40, 48] utilize both image and text encoders to process an image and multiple reference sentences. These metrics outperform *reference-free* metrics in terms of correlation with human evaluations but have significant issues in effectively leveraging multiple reference sentences. In general, they compute evaluation scores for each reference sentence independently, resulting in the virtual processing of only a single reference sentence and leading to suboptimal performance.

Handling multiple references becomes particularly important in the automatic evaluation of image captioning. Consider a scenario where an image depicts both a person and a dog, with two corresponding references. The first reference describes only the person, while the second focuses solely on the dog. Previous works, by utilizing an aggregation mechanism, virtually processes only a single reference. Therefore, they fail to evaluate a candidate that properly describes both subjects, such as "a person walking a dog." This underscores the necessity of handling multiple references.

Therefore, we propose DENEB¹, a supervised automatic evaluation metric specifically robust against hallucinations. Fig. 1 shows an overview of DENEB. Unlike *pseudo-multifaceted* metrics, DENEB is a *multifaceted* metric and thus can effectively compare multifaceted descriptions of an image against a candidate caption. This capability is achieved by the Sim-Vec Transformer, a novel architecture designed for the efficient processing of multiple reference captions.

Furthermore, for training our proposed metric, we construct the Nebula dataset by extending the Polaris dataset, increasing the visual diversity by a factor of three. Our dataset offers more effective data for training supervised metrics, particularly those capable of handling a wide range of cases, including hallucinated captions.

The main contributions of this study are summarized as follows:

- 1. We introduce the Sim-Vec Transformer, which handles the similarity between an image, a candidate caption, and reference captions.
- 2. We propose Sim-Vec Extraction (SVE), which utilizes a Hadamard product and element-wise differences to extract features beneficial for evaluation.
- 3. We construct the diverse and balanced Nebula dataset comprising 32,978 images, paired with human judgments provided by 805 annotators.
- 4. We achieve state-of-the-art performance among LLM-free metrics on FOIL, Composite, Flickr8K-Expert, Flickr8K-CF, PASCAL-50S, and Nebula.

2 Related Work

Image captioning has been widely applied in various areas of society, including the assistance of visually impaired individuals [2, 10, 13, 14], medical image

¹ In the field of general text generation metrics, names of celestial bodies are often used as names for metrics, such as METEOR [5] and COMET [38]. Following this convention, we have named our metric and dataset after a celestial body.

analysis [4, 36], and the generation of explanations in robotics [19, 33]. Several comprehensive surveys [13, 17, 44] have provided an exhaustive overview of representative models, standard datasets, and metrics, including those specifically designed for evaluating hallucinations, such as CHAIR [39]. Notably, [13] offers a detailed summary of how different approaches, including convolutional methods, attention mechanisms, and generative adversarial networks, contribute to reducing hallucinations in image captioning.

Hallucinations. Hallucinations pose significant challenges not only in LLMs but also in image captioning models [6, 21, 25, 31, 32, 42, 45]. In image captioning, hallucinations are instances in which models generate captions with words not corresponding to any of the elements in the input image. As noted in [13], this issue is particularly critical in social applications where the correctness of captions is prioritized over content coverage. Such hallucinations substantially affect the reliability of these models.

A notable metric developed to evaluate this issue is CHAIR [39], which assesses hallucinations by calculating the proportion of generated words that accurately reflect objects in the input image. However, CHAIR's approach is rule-based and operates within a closed vocabulary, limiting its applicability and generalizability. The development of CHAIR underscores the need for more comprehensive benchmarks to evaluate the robustness against hallucinations.

In response to the challenge of hallucinations, Shekhar et al. [42] introduced the FOIL benchmark to evaluate the robustness of metrics against hallucinations. The FOIL dataset, derived from the COCO dataset [27], consists of approximately 200,000 image-caption pairs, featuring a mix of both correct and hallucinated captions. Hallucinated captions are generated by altering a single word in a correct caption to create a similar but inaccurate version — for instance, replacing "motorcycle" with "bicycle". Moreover, Hessel et al. [15] presented a widely-accepted methodology for using the FOIL dataset to evaluate metrics. As detailed in [15], this methodology entails the following steps: For each of the 32K test images in the dataset, a pair comprising a hallucinated caption and its correct counterpart is sampled. The robustness of each evaluation metric against hallucinations is then evaluated based on its capability to consistently assign higher scores to the correct captions than to their hallucinated versions.

Image Captioning Metrics. Standard metrics for image captioning include BLEU [35], METEOR [5], ROUGE [26], CIDEr [34,46], and SPICE [3,47]. These classic metrics, primarily based on *n*-grams and/or scene graphs, have been extensively used but often exhibit low correlation with human judgments. This discrepancy has led recent studies to shift their focus toward data-driven metrics, such as BERTScore [51], CLIP-S [15], MID [20], and Polos [48]. BERTScore and MoverScore, for instance, leverage a pre-trained BERT encoder to compare word token embeddings in both candidate captions and references. However, their lack of image incorporation can limit their effectiveness in evaluating image captioning models. Several metrics adopt strategies that leverage both visual and language embeddings from pre-trained vision-and-language models, such as CLIP [37], ViL-BERT [29], and UNITER [8]. A notable example is CLIP-S [15], which employs an unsupervised approach to evaluate captions by measuring their similarity to embeddings generated by CLIP encoders. The distinct feature of CLIP-S is its ability to evaluate captions in contexts both with and without reference images. In [40], the authors introduced PAC-S, a variant of CLIP-S. In this variant, the authors generated text-image pairs using image generators and the original CLIP model was fine-tuned with these generated pairs.

Supervised metrics, which are trained based on human judgments, have been studied in text generation tasks, including machine translation, text summarization, and text simplification [30, 38, 41, 43, 49]. In the field of image captioning, however, the development of such metrics is still in its infancy. UMIC is an example of a supervised metric in this field and employs the fine-tuning of UNITER through a ranking-based approach. While UMIC [23] has shown superior performance to other metrics, [48] pointed out that ranking models such as UMIC have shortcomings, such as varying focal points in captions and subjective variations in expression. A notable instance of a different approach from the ranking model is Polos, a supervised metric inspired by COMET [38]. BLEURT [41], and RUSE [43]. The authors of [48] introduced the M²LHF framework, which was used to develop a metric utilizing human judgments. Based on the M²LHF, Polos achieves state-of-the-art performance in various image captioning benchmarks. However, as outlined in Section 1, Polos falls into the category of *pseudo-multifaceted* metrics, facing challenges in effectively managing multiple references.

Datasets and Benchmarks. Standard datasets for evaluating image captioning metrics include FOIL [42], Flickr8K-Expert [16], Flickr8K-CF, Composite [1], Polaris [48], and PASCAL-50S [3]. In response to concerns about the lack of diversity in the Flickr8K-Expert, Flickr8K-CF, and Composite datasets, Wada et al. [48] introduced the Polaris dataset, which is currently the largest in the field. However, as we argue later, Polaris faces the challenge of a limited variety of images compared with the total number of samples, leading to concerns about imbalance in the dataset.

3 Methods

In this study, we propose a novel automatic evaluation metric DENEB, which is specifically robust to hallucinations. Fig. 2 provides an overview of the proposed metric DENEB. Unlike *pseudo-multifaceted* metrics, DENEB is a *multifaceted* metric and can therefore effectively compare multifaceted descriptions of an image with a candidate caption. This capability is achieved by the Sim-Vec Transformer, a novel architecture designed for the efficient processing of multiple reference captions.

The proposed metric distinguishes itself from existing metrics in three key aspects as follows: First, in contrast to *classic* metrics [5, 26, 35, 46], DENEB is

6 K. Matsuda et al.



Fig. 2: The architecture of DENEB. CLIP and RoBERTa are employed to extract embeddings from an image, a candidate, and references. These embeddings are then processed concurrently by the Sim-Vec Transformer, comprising two modules: the Sim-Vec Extraction and transformer. Sim-Vec Extraction (SVE) utilizes a Hadamard product and element-wise differences to extract features, capturing the similarity among $\boldsymbol{x}_{img}, \boldsymbol{x}_{cand}, \text{ and } \{\boldsymbol{x}_{ref}^{(i)}\}_{i=1}^{N}$.

a data-driven automatic evaluation metric that employs a supervised approach. Second, unlike *reference-free* metrics [15, 23, 40], DENEB utilizes reference captions, enhancing its robustness against hallucinations. Finally, whereas *pseudomultifaceted* metrics typically employ aggregate functions [15, 40, 48], DENEB more effectively addresses the challenge of hallucination by handling multiple references simultaneously through the Sim-Vec Transformer.

The Sim-Vec Transformer handles the similarity between an image, a candidate caption, and reference captions. This module can be broadly applied to automatic evaluation metrics that deal with image features, including CLIP-S [15] and PAC-S [40].

3.1 Feature Extraction

Given an image \boldsymbol{x}_{img} , a candidate \boldsymbol{x}_{cand} , and N references $\{\boldsymbol{x}_{ref}^{(i)}\}_{i=1}^N$, the automatic evaluation metrics for image captioning models output a score \hat{y} that captures the appropriateness of \boldsymbol{x}_{cand} in the context of \boldsymbol{x}_{img} and $\{\boldsymbol{x}_{ref}^{(i)}\}_{i=1}^N$. First, the input x to our metric is defined as follows:

$$\boldsymbol{x} = \left\{ \boldsymbol{x}_{\text{img}}, \left\{ \boldsymbol{x}_{\text{ref}}^{(i)} \right\}_{i=1}^{N}, \boldsymbol{x}_{\text{cand}} \right\},$$
(1)

where $\boldsymbol{x}_{img} \in \mathbb{R}^{3 \times H \times W}$, $\{\boldsymbol{x}_{ref}^{(i)}\}_{i=1}^{N} \in \{0, 1\}^{N \times V \times L}$, and $\boldsymbol{x}_{cand} \in \{0, 1\}^{V \times L}$ represent an image, N references, and a candidate, respectively. Furthermore, H, W, N, V, L denote the height and width of the image, the number of references, the vocabulary size, and the number of tokens, respectively.

In this study, we extract features from \boldsymbol{x} using CLIP [37] and RoBERTa [28]. We utilize the CLIP encoder (CLIP ViT-B/32) pre-trained with [40] and RoBERTa-base pre-trained with SimCSE [12]. Specifically, we use CLIP's text encoder to obtain text embeddings, $\{\boldsymbol{r}_{\text{clip}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N \times d_{\text{clip}}}$ and $\boldsymbol{c}_{\text{clip}} \in \mathbb{R}^{d_{\text{clip}}}$, from

 $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N}$ and $\boldsymbol{x}_{\mathrm{cand}}$, respectively, where d_{clip} denotes the output dimension of CLIP. Concurrently, using the CLIP image encoder, we obtain image features $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{clip}}}$ from $\boldsymbol{x}_{\mathrm{img}}$. Additionally, we employ RoBERTa to obtain text embeddings $\{\boldsymbol{r}_{\mathrm{rb}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N \times d_{\mathrm{rb}}}$ and $\boldsymbol{c}_{\mathrm{rb}} \in \mathbb{R}^{d_{\mathrm{rb}}}$ from $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N}$ and $\boldsymbol{x}_{\mathrm{cand}}$, respectively, where d_{rb} denotes the output dimension of RoBERTa and $\{\boldsymbol{r}_{\mathrm{rb}}^{(i)}\}_{i=1}^{N}$ and $\boldsymbol{c}_{\mathrm{rb}}$ are obtained from the [CLS] token of the inputs. In this study, both the CLIP and RoBERTa encoders are frozen.

3.2 Sim-Vec Transformer

Sim-Vec Extraction. In the context of metrics for image captioning, accurately capturing the similarity between the image, the candidate, and the multiple references is crucial. Several studies have incorporated mechanisms to extract these similarities, thereby achieving high performance [30, 38, 43, 48]. According to the observations in [30, 38, 43, 48], the exclusion of the candidate and the reference leads to a decline in performance. This suggests that these metrics do not fully leverage the extracted vector-form similarities, partially because the latter part of them, specifically responsible for processing vector-form similarities, is just a poorly performing MLP. Considering these mechanisms effectively extract similarities [30, 38, 43, 48], the inclusion of raw features of candidates and references could even hinder effective learning from vector-form similarities. Therefore, in this study, we introduce the Sim-Vec Extraction (SVE) module by extending them to enhance the focus on effectively leveraging these extracted similarities. Our transformer-based metric has greater potential in capturing vector-form similarities more effectively than the classic MLP approach.

The SVE module employs a Hadamard product and element-wise differences to extract features that capture the similarity from the following vectors:

$$\boldsymbol{s}_{\rm in} = \left\{ \boldsymbol{c}_{\rm clip}, \left\{ \boldsymbol{r}_{\rm clip}^{(i)} \right\}_{i=1}^{N}, \boldsymbol{c}_{\rm rb}, \left\{ \boldsymbol{r}_{\rm rb}^{(i)} \right\}_{i=1}^{N}, \boldsymbol{v} \right\}.$$
(2)

We compute h_{clip} , h_{rb} , d_{clip} , d_{rb} as follows:

$$\boldsymbol{h}_{\text{clip}} = \left\{ \boldsymbol{c}_{\text{clip}} \odot \boldsymbol{v}, \left\{ \boldsymbol{c}_{\text{clip}} \odot \boldsymbol{r}_{\text{clip}}^{(i)} \right\}_{i=1}^{N} \right\},$$
(3)

$$\boldsymbol{d}_{\text{clip}} = \left\{ |\boldsymbol{c}_{\text{clip}} - \boldsymbol{v}|, \left\{ |\boldsymbol{c}_{\text{clip}} - \boldsymbol{r}_{\text{clip}}^{(i)}| \right\}_{i=1}^{N} \right\},$$
(4)

$$\boldsymbol{h}_{\rm rb} = \left\{ \boldsymbol{c}_{\rm rb} \odot \boldsymbol{r}_{\rm rb}^{(i)} \right\}_{i=1}^{N}, \tag{5}$$

$$\boldsymbol{d}_{\rm rb} = \left\{ |\boldsymbol{c}_{\rm rb} - \boldsymbol{r}_{\rm rb}^{(i)}| \right\}_{i=1}^{N}.$$
 (6)

It is important to note that, unlike RUSE [43] and COMET [38], we do not combine the candidate features c and the reference features r here to allow our model to focus more on vector-form similarity. Subsequently, we obtain g_{inter} by combining h_{clip} , d_{clip} , h_{rb} , and d_{rb} .

$$\boldsymbol{g}_{\text{inter}} = \{\boldsymbol{h}_{\text{clip}}, \boldsymbol{d}_{\text{clip}}, \boldsymbol{h}_{\text{rb}}, \boldsymbol{d}_{\text{rb}}\}.$$
 (7)

8 K. Matsuda et al.

Non-aggregate mechanism. Previous studies [15, 40, 48] exhibit a fundamental limitation in the handling of multiple references. Specifically, they calculate evaluation scores independently for each of the N references, so that the multiple references are not fully utilized. Generally, CLIP-S [15], PAC-S [40], and Polos [48] can be formulated as follows:

$$\hat{y} = \operatorname{Aggregate}_{i} f(\boldsymbol{x}^{(i)}), \tag{8}$$

where $\boldsymbol{x}^{(i)}$ denotes the *i*th reference included in $\boldsymbol{x}^{(i)} = \{\boldsymbol{x}_{img}, \boldsymbol{x}_{ref}^{(i)}, \boldsymbol{x}_{cand}\}$ and Aggregate (·) represents any mapping $f : \mathbb{R}^N \to \mathbb{R}$, such as the Max and Mean functions. In these methods, the core of the metric $f(\cdot)$ takes only a single reference, $\boldsymbol{x}_{ref}^{(i)}$, as input, implying the ineffective processing of multiple references. Moreover, the aggregate function cannot be optimized since it does not have trainable parameters. Therefore, it can be concluded that these methods do not fully utilize multiple references in the automatic evaluation of image captioning.

To overcome this limitation, DENEB employs an *N*-layer transformer to effectively handle multiple references. Initially, we obtain g by concatenating the [CLS] token $g_{\text{[CLS]}}$ to g_{inter} obtained from the SVE as follows:

$$\boldsymbol{g} = \{\boldsymbol{g}_{[\text{CLS}]}, \boldsymbol{g}_{\text{inter}}\}.$$
(9)

Here, g_{inter} is defined as the feature extracted from the SVE. Subsequently, we calculate h_{N} using an N-layer transformer encoder:

$$\boldsymbol{h}_{\mathrm{N}} = \mathrm{TransformerEncoder}\left(\boldsymbol{g}\right).$$
 (10)

In this study, we set N to 3 for achieving a balance between computational efficiency and the model's ability to handle multiple references effectively. Subsequently, we input the [CLS] token $\boldsymbol{h}_{[CLS]}$ from \boldsymbol{h}_{N} into an MLP and then apply a sigmoid function to compute the final evaluation score \hat{y} .

We employ the Huber loss as our loss function due to its robustness against outliers. The Huber loss is defined using \hat{y} , the human judgments y, and the hyperparameter δ as follows:

$$\mathcal{L} = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{if } |\hat{y} - y| < \delta, \\ \delta \cdot (|\hat{y} - y| - \frac{1}{2} \cdot \delta) & \text{otherwise.} \end{cases}$$
(11)

In this study, we set δ to 0.5.

4 Experiments and Results

4.1 Experimental Setup

Nebula dataset. The development of supervised metrics for image captioning requires a large-scale, diverse dataset, but there are limited available datasets for effectively training these models. Moreover, the largest dataset in this field [48]

has a notable issue: a significant imbalance, characterized by a limited variety of images relative to the number of captions. Specifically, there is a discrepancy where the number of images is approximately only one-tenth of the total number of samples. This imbalance could potentially lead to suboptimal evaluation of various types of images.

To mitigate this imbalance, we constructed the Nebula dataset² by extending the Polaris dataset to have approximately three times the number of images. Following the standard procedure, human judgments were adopted on a five-point scale to assess the appropriateness of a candidate for a given image and references. The annotation process was carried out through a crowdsourcing service. Following previous studies [23,47,48], we instructed the annotators to assess the quality of the candidates from the perspectives of fluency, relevance, and descriptiveness. To ensure the reliability of our data, we excluded data from evaluators who exhibited suspicious behavior, such as extremely short response times or consistently providing identical values. In addition, the human judgments, given on a five-point scale, were normalized to the range [0, 1]. For a comprehensive description of the dataset, please refer to the Appendix.

4.2 Comparison to State-of-the-Art

We evaluated the automatic evaluation metrics based on the accuracy they achieved when applied to the FOIL benchmark and PASCAL-50S, as well as their correlation coefficients when applied to Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula. Given the importance of efficiency in practical applications, we also conducted experiments to measure the inference times of these metrics. This enables a comprehensive evaluation of both the performance effectiveness and practical usability of these metrics.

We adopted BLEU [35], ROUGE [26], METEOR [5], CIDEr [46], and SPICE [3] as they are standard metrics for image captioning tasks. Additionally, we included MoverScore [52], BERTScore [51], BARTScore [50], TIGEr [18], LEIC [9], ViLBERTScore [22], UMIC [23], MID [20], CLIP-S [15], PAC-S [40], Polos [48], CLAIR [7] and FLEUR³ [24] as baseline metrics since they are representative metrics for image captioning. It is important to note that among these, CLAIR and FLEUR are LLM-based metrics, while the others are LLM-free.

Hallucination and Likert judgments. We conducted comparative experiments against baseline models across the FOIL dataset to investigate the robustness against hallucinations. Additionally, we performed quantitative comparisons on various datasets, including Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula. Table 1 shows the performance of the proposed and baseline metrics on FOIL. Following previous studies [15, 40, 48], our experiments were conducted in settings where either one or four reference captions were provided.

 $^{^{2}}$ The Nebula dataset can be downloaded from this link.

³ Note that FLEUR and RefFLEUR was released after the submission of our manuscript. At the time of our submission, DENEB exhibited the best performance.

10 K. Matsuda et al.

Table 1: A quantitative comparison with baseline metrics. Boldface indicates the best values, and <u>underlining</u> indicates the second best values. A "-" indicates either non-executable code or unavailable data. Note that, for a fair comparison, we have also included DENEB using OpenCLIP ViT-L/32 backbone as the encoder.

	FOIL 1-ref	FOIL 4-ref	Composite	Flickr8K Expert	Flickr8K CF	Nebula	Inference time
	Accura	acy [%]	Kendall's τ				[ms]
LLM-free metrics							
BLEU [35]	66.5	82.6	30.6	30.8	16.4	40.4	0.13
ROUGE [26]	71.7	79.3	32.4	32.3	19.9	42.6	0.51
CIDEr [46]	82.5	90.6	37.7	43.9	24.6	48.1	0.40
METEOR [5]	78.8	82.6	38.9	41.8	22.2	46.8	0.14
SPICE [3]	75.5	86.1	40.3	44.9	24.4	44.0	17
UMIC [23]	_	-	56.1	46.8	30.1	_	_
CLIP-S [15]	87.2	87.2	53.8	51.2	34.4	46.9	1.7
PAC-S [40] (ViT-B/32)	89.9	89.9	55.7	54.3	36.0	47.2	2.5
PAC-S [40] (ViT-L/14)	93.1	93.1	54.0	55.7	37.0	47.9	12
SPARCS [11]	_	_	43.1	48.1	10.4	46.5	1.5
BERTScore [51]	88.6	92.1	30.1	46.7	22.8	47.0	7.6
BARTScore [50]	85.3	91.1	43.5	37.8	24.3	43.8	130
ViLBERTScore [22]	_	_	52.4	50.1	_	_	_
MID [20]	90.5	90.5	55.7	54.9	37.3	51.3	52
RefCLIP-S [15]	91.0	92.6	55.4	53.0	36.4	46.9	2.8
RefPAC-S [40] (ViT-B/32)	93.7	94.9	57.3	55.9	37.6	50.6	4.5
RefPAC-S [40] (ViT-L/14)	94.4	94.9	56.0	56.4	37.8	50.4	17
Polos [48]	93.2	95.1	<u>57.6</u>	56.4	37.8	53.9	6.9
DEVER (V/T B (89)	95.1	96.1	57.9	56.5	38.0	54.1	7.3
DENEB (VIT-B/32)	(+0.7)	(+1.0)	(+0.3)	(+0.1)	(+0.2)	(+0.2)	
	95.4	96.5	58.2	56.8	38.3	54.3	22
DENEB (V11-L/14)	(+1.0)	(+1.4)	(+0.6)	(+0.4)	(+0.5)	(+0.4)	
LLM-based metrics	-						
CLAIR [7]	81.4	83.4	55.0	44.6	34.4	52.7	1600
FLEUR 24	96.8	96.8	<u>63.5</u>	53.0	38.6	-	700
RefFLEUR [24]	97.3	98.4	64.2	51.9	38.8	-	760

In both single-reference (1-ref) and four-reference (4-ref) settings, our metric achieved state-of-the-art (SOTA) performance, with scores of 95.4% and 96.4%, respectively, marking improvements of 1.7 and 1.4 points over existing LLM-free metrics. This suggests that our metric is more effective in addressing hallucinations compared to existing metrics.

Table 1 also shows the quantitative comparison results with baseline metrics for the Composite, Flickr8K, Flickr8K-CF, and Nebula datasets. Consistent with previous studies [3, 15, 20, 48], we employed τ_b (Kendall-B) for Flickr8K-CF and τ_c (Kendall-C) for the other datasets. It should be noted that the Kendall-C evaluation values for the CLAIR metric were not provided in [7]; thus, for a fair comparison, we reproduced CLAIR and reevaluated it using Kendall-C. This illustrates that our proposed metric DENEB achieved SOTA results with scores of 58.2, 56.8, 38.3, and 54.3 on the Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula datasets, respectively. In particular, DENEB outperformed the existing SOTA LLM-free metrics by margins of 0.6, 0.4, 0.5, and 0.4 on the Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula datasets, respectively.

	HC	\mathbf{HI}	$\mathbf{H}\mathbf{M}$	$\mathbf{M}\mathbf{M}$	Mean		
		Accuracy [%]					
LLM-free metrics							
BLEU [35]	60.4	90.6	84.9	54.7	72.7		
METEOR [5]	63.8	97.7	93.7	65.4	80.2		
ROUGE [26]	63.7	95.3	92.3	61.2	78.1		
SPICE [3]	63.6	96.3	86.7	68.3	78.7		
CIDEr [46]	65.1	98.1	90.5	64.8	79.6		
CLIP-S [15]	56.5	99.3	96.4	70.4	80.7		
PAC-S [40] (ViT-B/32)	60.6	99.3	96.9	72.9	82.4		
PAC-S [40] (ViT-L/14)	59.6	99.7	96.9	75.2	82.9		
UMIC [23]	66.1	99.8	98.1	76.2	85.1		
BERTScore [51]	65.4	98.1	96.4	60.3	80.1		
MoverScore [52]	65.1	97.1	93.2	65.6	80.3		
TIGEr [18]	56.0	99.8	92.8	74.2	80.7		
MID [20]	67.0	99.7	97.4	76.8	85.2		
RefCLIP-S [15]	64.5	99.6	95.4	72.8	83.1		
RefPAC-S $[40]$ (ViT-B/32)	67.7	99.6	96.0	75.6	84.7		
RefPAC-S [40] (ViT-L/14)	65.0	99.8	97.3	76.1	84.6		
Polos [48]	<u>70.0</u>	99.6	97.4	79.0	86.5		
DENER (VET B /22)	74.4	99.8	97.3	76.5	87.0		
DENEB (VII- $\mathbf{D}/32$)	(+4.4)				(+1.3)		
	76.1	99.7	97.4	77.9	87.8		
DENEB (VII-L/14)	(+6.1)				(+0.5)		
LLM-based metrics							
CLAIR-E [7]	57.7	99.8	94.6	75.6	81.9		
FLEUR [24]	61.3	99.7	97.6	74.2	83.2		
RefFLEUR [24]	68.0	99.8	98.0	76.1	85.5		

 Table 2: A quantitative comparison on PASCAL-50S. DENEB achieved the best or second-best performance across all categories.

Inference time. In the context of practical applications, it is essential to consider not only the performance of metrics but also their inference times. Table 1 shows the inference times per sample, measured on a system equipped with a GeForce RTX 3090 and an Intel Core i9-10900KF. The inference times for recent LLM-free metrics, RefCLIP-S and RefPAC-S, were 2.8 ms and 4.5 ms, respectively. Similarly, DENEB demonstrated an inference time of 7.3 ms, suggesting that the proposed metric is comparable in terms of inference speed. In contrast, LLM-based metrics such as CLAIR and FLEUR exhibited significantly longer inference times of 1600 ms and 700 ms respectively, which are at least 95 times slower than DENEB. These results indicated the limitations of LLM-based metrics, rendering them impractical for real-time applications. Considering that standard evaluation datasets sometimes contain more than 40,0000 samples [16], it is crucial that the evaluations are completed within a practical timeframe.

Pairwise ranking on Pascal-50S. PASCAL-50S consists of 1,000 images each with 50 reference captions, and presents the task of identifying which caption out of a given pair has received the majority vote from human judg-

				-
(a)		$m{x}_{ m ref}^{(1)}$: "A kitchen with vending machines and a black and white checkered floor" $m{x}_{ m cand}$: "black and white checkered floor" Human: 0.5	RefCLIP-S Polos 0.719 0.816	Deneb 0.634
(b)	SAN GARLOS Contractions CONTRACTOR	$m{x}_{ref}^{(1)}$: "Three traffic signs arranged on a sign post." $m{x}_{cand}$: ""a man sitting on a stop sign on a street corner" Human: 0.0	RefCLIP-S Polos 0.442 0.392	Deneb 0.023
(c)		$\boldsymbol{x}_{ref}^{(1)}$: "A small cat standing on a wood railing next to an arrange- ment in a vase." \boldsymbol{x}_{cand} : "a cat on a fence" Human: 0.25	RefCLIP-S Polos 0.591 0.703	Deneb 0.691

Fig. 3: Qualitative results on the Nebula dataset. Panels (a) and (b) illustrate successful cases, and panel (c) depicts a failure case.

ments. Specifically, this task focuses on pairwise preference judgments between two captions in the following categories: pairs of HC (human correct) captions, HI pairs (both human-written, with one being incorrect), HM pairs (one from a human and the other machine-generated), and MM pairs (both generated by machines). Table 2 presents the performance of our proposed metric with baselines in PASCAL-50S. As indicated in Table 2, DENEB achieved SOTA performance with scores of 76.1% in HC, 77.9% in MM, and Mean of 87.8%, outperforming existing metrics by margins of 6.1, 0.6, and 1.3 points, respectively.

4.3 Qualitative Analysis

Fig. 3-(a) shows a successful sample in the Nebula dataset. In this sample, \boldsymbol{x}_{cand} is "black and white checkered floor," and $\boldsymbol{x}_{ref}^{(1)}$ is "A kitchen with vending machines and a black and white checkered floor". The human judgment in this sample was rated 0.5 because \boldsymbol{x}_{cand} was partially correct. While both Polos [48] and RefCLIP-S [15] incorrectly evaluated it as 0.816 and 0.719, respectively, DENEB evaluated it closer to \boldsymbol{y} with a score of 0.634. Fig. 3-(b) shows another successful sample in the Nebula dataset. In this sample, \boldsymbol{x}_{cand} is "a man sitting on a stop sign on a street corner," and $\boldsymbol{x}_{ref}^{(1)}$ is "Three traffic signs arranged on a sign post." The human judgment for this sample was 0.0 due to the presence of hallucinations ("a man sitting on") in \boldsymbol{x}_{cand} . Polos [48] and RefCLIP-S incorrectly evaluated this sample with scores of 0.392 and 0.442, respectively, whereas DENEB evaluated it more accurately with a value of 0.023. These results demon-

strate that our proposed metric DENEB is not only aligned more closely with human judgments but also shows robustness against hallucinations.

Fig. 3-(c) illustrates a failure sample in the Nebula dataset. In this sample, \boldsymbol{x}_{cand} is described as "a cat on a fence," while $\boldsymbol{x}_{ref}^{(1)}$ is "A small cat standing on a wood railing next to an arrangement in a vase." The human judgment for this sample was 0.25 due to the lack of sufficient detail. However, DENEB inappropriately evaluated it with a score of 0.691. Polos [48], RefPAC-S [40], and RefCLIP-S [15] also output significantly erroneous evaluation scores of 0.703, 0.847, and 0.591, respectively. These consistent errors across metrics utilizing image features suggest that the failure of the proposed metric is likely due to an overestimation of the quality of \boldsymbol{x}_{cand} , which only describes objects that are prominently displayed in the image.

4.4 Ablation Studies

We conducted ablation studies to investigate the contribution of each module. Table 3 presents the quantitative results of the ablation studies.

Sim-Vec Transformer ablation. We replaced the Sim-Vec Transformer with MLP to investigate the contribution of the Sim-Vec Transformer. Table 3 shows that the accuracy for 1-ref and 4-ref on FOIL in Metric (i) were 76.2% and 76.5%, respectively, representing a decrease of 19.2 and 20.0 points compared to Metric (iv). Table 3 also shows that the correlation coefficients for Composite, Flickr8K-Expert,Flickr8K-CF, and Nebula in Metric (i) were 37.7, 40.1, 25.1, and 48.1, respectively. In comparison to Metric (iv), these values have decreased significantly by 20.7, 16.1, 13.2, and 6.2 points. These results demonstrate that the SVE Transformer contributes to both performance and robustness against hallucinations.

SVE ablation. To analyze the contribution of the SVE, we simply used Eq. (2) as input instead of Eqs. (3) to (6). Table 3 shows that the accuracy for 1-ref and 4-ref on FOIL in Metric (ii) were 84.3% and 89.3%, respectively, representing a decrease of 11.1 and 7.2 points compared to Metric (iv). Table 3 also shows that the correlation coefficients for Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula in Metric (ii) were 35.8, 49.6, 24.7, and 45.2, respectively. These values were lower by 22.6, 6.6, 13.6, and 9.1 points than Metric (iv). These results indicate that the SVE is effective in extracting significant features for automatic evaluation and enhancing robustness against hallucinations.

Multifaceted references ablation. We did not originally employ Aggregate functions in DENEB to handle multifaceted references, as detailed in Section 3. To explore their effect on performance, we conducted an ablation study by adding Aggregate functions. Table 3 shows that the accuracy for 1-ref and 4-ref on FOIL in Metric (iii) were 94.4% and 96.1%, respectively, representing a decrease of 1.0 and 0.4 points compared to Metric (iv). Table 3 also demonstrates that the correlation coefficients for Composite, Flickr8K-Expert, Flickr8K-CF, and Nebula

14 K. Matsuda et al.

Table 3: Quantitative results of the ablation studies. The incorporation of the SVE Transformer, as indicated for Metrics (i) and (ii), emerges as the most significant factor in enhancing performance.

	Sim-Vec Transformer	SVE	Multifaceted references	FOIL 1-ref	FOIL 4-ref	Flickr8K CF	Flickr8K Expert	Composite	Nebula
			Accura	acy [%]		Kend	all's τ		
(i)				76.2	76.5	25.1	40.1	37.7	48.1
(ii)	\checkmark		\checkmark	84.3	89.3	24.7	49.6	35.8	45.2
(iii)	\checkmark	\checkmark		94.4	96.1	37.2	55.7	57.4	53.2
(iv)	\checkmark	\checkmark	\checkmark	95.4	96.5	38.3	56.2	58.4	54.3

in Metric (iii) were 57.4, 55.7, 37.2, and 53.2, respectively. In comparison to Metric (iv), these values were lower by 1.0, 0.5, 1.1, and 1.1 points. These results indicate that the inclusion of a mechanism for handling multifaceted references contributed to both performance and robustness against hallucinations.

5 Limitations and Discussion

While our metric has clearly been shown to provide impressive robustness against hallucinations, it is not without its limitations. The primary limitation is the occurrence of errors stemming from differences in the areas of focus. Moreover, the metric tends to erroneously overevaluate candidates that only describe prominent objects in an image while neglecting finer details. These limitations indicate that the proposed metric may not effectively capture the relationship between x_{cand} and the local regions of x_{img} . For further error analysis, see Appendix.

6 Conclusions

In this study, we have proposed the novel automatic evaluation metric DENEB, which is specifically robust to hallucinations. Our key contributions are as follows: (1) We have introduced the Sim-Vec Transformer, which handles the similarity between an image, a candidate caption, and reference captions. (2) We proposed the Sim-Vec Extraction (SVE), which utilizes a Hadamard product and element-wise difference to extract features beneficial for automatic evaluation. (3) We constructed the diverse and balanced Nebula dataset comprising 32,978 images, paired with human judgments provided by 805 annotators. (4)We achieved state-of-the-art performance on FOIL, Composite, Flickr8K-Expert, Flickr8K-CF, PASCAL-50S, and the Nebula dataset.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number 23H03478, JST CREST, and NEDO.

References

- 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)
- Ahsan, H., Bhalla, N., Bhatt, D., Shah, K.: Multi-Modal Image Captioning for the Visually Impaired. In: NAACL. pp. 53–60 (2021)
- Anderson, P., Fernando, B., Johnson, M., et al.: SPICE: Semantic Propositional Image Caption Evaluation. In: ECCV. pp. 382–398 (2016)
- 4. Ayesha, H., et al.: Automatic Medical Image Interpretation: State of the Art and Future Directions. Pattern Recognition **114**, 107856 (2021)
- Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: ACL. pp. 65–72 (2005)
- Biten, A., Gómez, L., et al.: Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning. In: IEEE WACV. pp. 2473–2482 (2022)
- Chan, D., et al.: CLAIR: Evaluating Image Captions with Large Language Models. In: EMNLP. pp. 13638–13646 (2023)
- Chen, Y., Li, L., Yu, L., Kholy, A., Faisal, Gan, Z., Cheng, Y., et al.: UNITER: Universal Image-text Representation Learning. In: ECCV. pp. 104–120 (2020)
- Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.: Learning to Evaluate Image Captioning. In: CVPR. pp. 5804–5812 (2018)
- Dognin, P., et al.: Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. JAIR 73, 437–459 (2022)
- Feinglass, J., et al.: SMURF: SeMantic and linguistic UndeRstanding Fusion for Caption Evaluation via Typicality Analysis. In: IJCNLP. pp. 2250–2260 (2021)
- Gao, T., Yao, X., Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: EMNLP. pp. 6894–6910 (2021)
- Ghandi, T., Pourreza, H., Mahyar, H.: Deep Learning Approaches on Image Captioning: A Review. ACM Computing Surveys 56(3) (2023)
- Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning Images Taken by People Who Are Blind. In: ECCV. pp. 417–434 (2020)
- Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A Referencefree Evaluation Metric for Image Captioning. In: EMNLP. pp. 7514–7528 (2021)
- Hodosh, M., et al.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. JAIR 47, 853–899 (2013)
- Hossain, Z., Sohel, F., Shiratuddin, F., Laga, H.: A Comprehensive Survey of Deep Learning for Image Captioning. ACM 51(6) (2019)
- Jiang, M., Huang, Q., Zhang, L., Wang, X., et al.: TIGEr: Text-to-image Grounding For Image Caption Evaluation. In: EMNLP. pp. 2141–2152 (2019)
- Kambara, M., Sugiura, K.: Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks. In: IEEE ICIP. pp. 2601–2605 (2022)
- Kim, J., et al.: Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In: NeurIPS. pp. 35072–35086 (2022)
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., et al.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. IJCV 123, 32–73 (2017)
- Lee, H., Yoon, S., et al.: ViLBERTScore: Evaluating Image Caption Using Visionand-Language BERT. In: Eval4NLP Workshop. pp. 34–39 (2020)
- Lee, H., Yoon, S., et al.: UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In: ACL. pp. 220–226 (2021)

- 16 K. Matsuda et al.
- Lee, Y., Park, L., Kang, M.: FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In: ACL (2024)
- Li, Y., Du, Y., Zhou, K., Wang, J., et al.: Evaluating Object Hallucination in Large Vision-Language Models. In: EMNLP. pp. 292–305 (2023)
- Lin, C.: ROUGE: A Package For Automatic Evaluation Of Summaries. In: ACL. pp. 74–81 (2004)
- 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)
- Liu, Y., Ott, M., Goyal, N., Du, J., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)
- Lu, J., et al.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: NeurIPS. pp. 13–23 (2019)
- Maddela, M., Dou, Y., Heineman, D., Xu, W.: LENS: A learnable evaluation metric for text simplification. In: ACL. pp. 16383–16408 (Jul 2023)
- Manakul, P., Liusie, A., Gales, M.: SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In: EMNLP (2023)
- Mishra, A., Asai, A., Balachandran, V., Wang, Y., Neubig, G., et al.: Fine-grained hallucinations detections. arXiv preprint arXiv:2401.06855 (2024)
- 33. Ogura, T., Magassouba, A., Sugiura, K., Hirakawa, T., et al.: Alleviating the Burden of Labeling: Sentence Generation by Attention Branch Encoder-Decoder Network. IEEE RAL 5, 5945–5952 (2020)
- Oliveira, G., Colombini, E., Avila, S.: CIDEr-R: Robust Consensus-based Image Description Evaluation. In: W-NUT. pp. 351–360 (2021)
- Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: ACL. pp. 311–318 (2002)
- Pavlopoulos, J., Kougia, V., Androutsopoulos, I.: A Survey on Biomedical Image Captioning. In: SiVL. pp. 26–36 (2019)
- Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
- Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A Neural Framework for MT Evaluation. In: EMNLP. pp. 2685–2702 (2020)
- Rohrbach, A., Hendricks, L., Burns, K., Darrell, T., Saenko, K.: Object Hallucination in Image Captioning. In: EMNLP. pp. 4035–4045 (2018)
- 40. Sarto, S., Barraco, M., Cornia, M., et al.: Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In: CVPR. pp. 6914–6924 (2023)
- Sellam, T., Das, D., Parikh, A.: BLEURT: Learning Robust Metrics for Text Generation. In: ACL. pp. 7881–7892 (2020)
- Shekhar, R., Pezzelle, S., Klimovich, Y., et al.: FOIL it! Find One Mismatch Between Image and Language caption. In: ACL. pp. 255–265 (2017)
- Shimanaka, H., et al.: RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In: WMT. pp. 751–758 (2018)
- 44. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From Show to Tell: A Survey on Deep Learning-Based Image Captioning. IEEE TPAMI 45, 539–559 (2021)
- 45. Sun, Z., Shen, S., Cao, S., et al.: Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv preprint arxiv:2309.14525 (2023)
- Vedantam, R., Zitnick, L., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. In: CVPR. pp. 4566–4575 (2015)
- Wada, Y., Kaneda, K., Sugiura, K.: JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models. In: CoNLL. pp. 424– 435 (2023)

- 48. Wada, Y., Kanta, K., Daichi, S., Sugiura, K.: Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In: CVPR (2024)
- Xu, W., Qian, X., Wang, M., Li, L., et al.: SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes. In: ACL. pp. 5166–5183 (2023)
- 50. Yuan, W., Neubig, G., et al.: BARTScore: Evaluating Generated Text as Text Generation. In: NeurIPS. pp. 27263–27277 (2021)
- 51. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: ICLR (2020)
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S.: MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In: EMNLP. pp. 563–578 (2019)