WEmbSim: A Simple yet Effective Metric for Image Captioning

Naeha Sharif
The University of Western Australia
Perth, Australia
naeha.sharif@research.uwa.edu.au

Lyndon White
Invenia Labs
Cambridge, UK
lyndon.white@invenialabs.co.uk

Mohammed Bennamoun
The University of Western Australia
Perth, Australia
mohammed.bennamoun@uwa.edu.au

Wei Liu
The University of Western Australia
Perth, Australia
wei.liu@uwa.edu.au

Syed Afaq Ali Shah
Murdoch University
Perth, Australia
Afaq.Shah@murdoch.edu.au

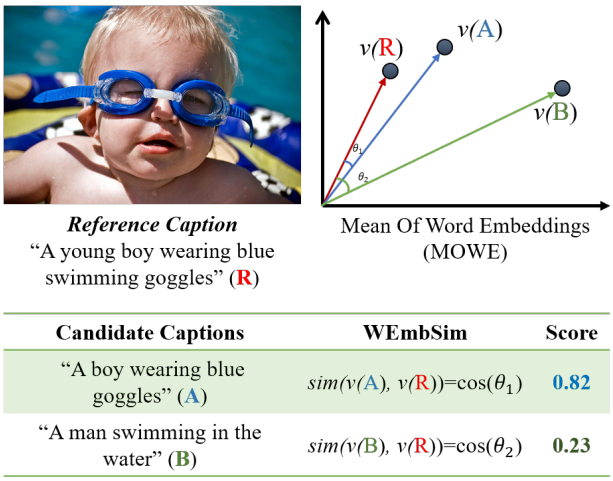
***Abstract*—The area of automatic image caption evaluation is still undergoing intensive research to address the needs of generating captions which can meet adequacy and fluency requirements. Based on our past attempts at developing highly sophisticated learning-based metrics, we have discovered that a simple cosine similarity measure using the Mean of Word Embeddings (MOWE) of captions can actually achieve a surprisingly high performance on unsupervised caption evaluation. This inspires our proposed work on an effective metric WEmbSim, which beats complex measures such as SPICE, CIDEr and WMD at system-level correlation with human judgments. Moreover, it also achieves the best accuracy at matching human consensus scores for caption pairs, against commonly used unsupervised methods. Therefore, we believe that WEmbSim sets a new baseline for any complex metric to be justified.**

***Index Terms*—Image Captioning, Automatic Evaluation Metric, Word Embeddings**

I. INTRODUCTION

Automatic caption evaluation is a challenging task which seeks to score machine generated captions on their quality. Due to a surge of interest in image captioning over the past few years [1], the development of effective evaluation metrics has become increasingly pressing. Human assessments are considered the gold standard for the evaluation task. However, obtaining human scores for captions is time-consuming and resource intensive.

Automatic evaluation metrics in contrast are more efficient and cost-effective. However, to serve as a replacement for humans, automatic measures are desired to generate human-like assessments such that their evaluation scores are consistent with human judgments. A number of sophisticated evaluation metrics have been proposed for captioning, amongst which some are borrowed from relevant domains, such as Machine Translation and others are developed specifically for captioning [2], [3]. While captioning specific metrics such as SPICE [19] and CIDEr [18] have shown improved performance over the ones adopted from other domains, there is still no consensus on a single metric that gives near human performance.



Candidate Captions	WEmbSim	Score
"A boy wearing blue goggles" (A)	$sim(v(A), v(R))=\cos(\theta_1)$	0.82
"A man swimming in the water" (B)	$sim(v(B), v(R))=\cos(\theta_2)$	0.23

Fig. 1. An overview of the proposed evaluation metric. The candidate and reference captions corresponding to an image are embedded into a Mean of Word Embeddings (MOWE) space. The evaluation score of a candidate caption is the cosine similarity between its MOWE and that of the reference caption.

A general inclination of the research community towards sophisticated and intricately crafted measures tends to overlook the efficacy and usefulness of simple methods. To avoid this trap of 'complex is better', our research is inspired by the principle of Parsimony (Occam's razor) [4]. This principle is interpreted as 'a simple solution is better than a complex one'. Thus, if two methods give near equal performance, then one should prefer the simpler one. Our motivation in this work is to propose a simple, intuitive, yet effective metric, that can serve as a strong baseline for caption evaluation, and also be used as a good feature for supervised measures.

In this paper, we present WEmbSim to evaluate machine generated captions. The key idea is to embed the captions into a semantic embedding space and then to assess the similarity between caption embeddings. There are various ways to embed sentences in a vector space. Sent2vec [5] is one such

example, which requires a pretrained model to generate sentence embeddings. In this paper, we adopt the simplest sentence embedding model to represent captions i.e., the representation of captions is obtained through the mean of word embeddings (MOWE). By word embeddings, we mean a family of learned distributed vector representations based on sound language models, such as Word2vec [6] and ELMo [7].

The strength of Word Embeddings in terms capturing semantics has been extensively verified in literature [8], [9]. Moreover, Sums of Word Embeddings (SOWE) have been successfully used as representations in a number of areas [10], [11]. We show that using the cosine distance between the MOWE of the candidate and reference caption, gives a score that is correlated to human evaluations on the captions quality. To the best of our knowledge, this is the first work which evaluates the performance of such metric for image captioning, using various pre-trained word embeddings across a number of benchmark datasets.

For our experiments we make use of pre-trained word embeddings from Word2vec [6], GloVE [12], FastText [13] and ELMo [7]. We do not use the recently proposed language representation model BERT [14] as it is primarily a fine-tuning based model for applying pre-trained language representations to downstream tasks. Our focus is towards using feature-based strategies for language representations such as ELMo and Word2vec, to avoid the overhead of including the Transformer-based architecture in our caption evaluation task.

Our contributions in this work are as follows:

- We propose a simple yet effective unsupervised metric WEmbSim for automatic caption evaluation.
- To demonstrate the strong performance of WEmbSim compared to rival methods, we evaluate its system-level correlation and accuracy in terms of matching human consensus scores for caption pairs.
- We analyze the robustness of our proposed metric to various sentence perturbations.
- We also examine its usefulness and complementarity to the existing measures.

II. RELATED LITERATURE

- **Captioning Vs. Machine Translation:**
  Caption generation has similarities to other natural language generation tasks, in particular machine translation and abstractive summarisation. Image captioning can be thought of as translating images into text. However, not all the contents of an image should be described in the caption: it is expected that the background for example would not be described except if it were particularly relevant to the main action occurring in the image.
  Given the similarities to machine translation and abstract summarisation, it is natural to expect that metrics for those tasks would be useful for the evaluation of captions. However, there are some important differences in the generated text. *First*, captions are normally short, single sentences; rather than paragraphs of complex sentences. *Second*, the possible text of captions is restricted by the domain of things that people photograph. While almost any idea can be put in text and translated; there are a wide range of subjects from intangible concepts to non-visually-interesting events, that will never be photographed, and thus never need captions generated for them.
- **Automatic Evaluation Metrics:**
  A number of automatic evaluation metrics have been proposed for captioning, which can be categorized into *supervised* [15], [16], [17] and *unsupervised* [18], [19], [20] methods. The work presented here, falls in the latter category.

  - **Textual Similarity-based Metrics:** Majority of existing caption evaluation metrics capture the linguistic correspondence between the candidate and the reference captions to generate an assessment score. Metrics such as BLEU [21], METEOR [22] and CIDEr [18] measure the n-gram overlap between the candidate and reference captions for evaluation. Such n-gram based metrics rely on word overlaps, therefore, their performance is dependent on the representativeness of the set of reference captions. However, exploiting linguistic information beyond the lexical level, i.e. using pre-trained embeddings helps remedy this issue by extending the reference lexicon. Thus, we believe that metrics which are based on distributed sentence representations are more promising for caption evaluation.
    SPICE [19], which is a sophisticated semantic metric, estimates the caption quality using a graph-based representation of captions. It relies on semantic parsers and completely ignores the syntactic quality of candidate captions. However, the competitive performance of SPICE shows that measuring the similarities between the semantic representation of captions is useful for caption evaluation. WMD [23], which is an embedding-based measure, was originally proposed to measure document similarity. It tends to overly penalize candidate captions which are shorter in length compared to the reference captions [3]. Moreover, it is also computationally expensive compared to WEmbSim (Sec. III-C).
  - **Visual Content-based Metrics:** TIGER [24] and VIFIDEL [25] are recently proposed metrics which use image content for caption evaluation. TIGEr uses a sophisticated pipeline to project the image (source) and text (target) into a common semantic space for quality assessment. VIFIDEL is an extension of WMD and uses word embedding's as a bridge for matching content in images and textual descriptions. The performance of TIGEr and VIFIDEL is highly dependent on correct visual detections. VIFIDEL prefers systems that just spit out correct object detections. Thus blurring the line between a list of object detections and captions.
  - **Textual Vs. Visual Content-based Metrics:** Major-

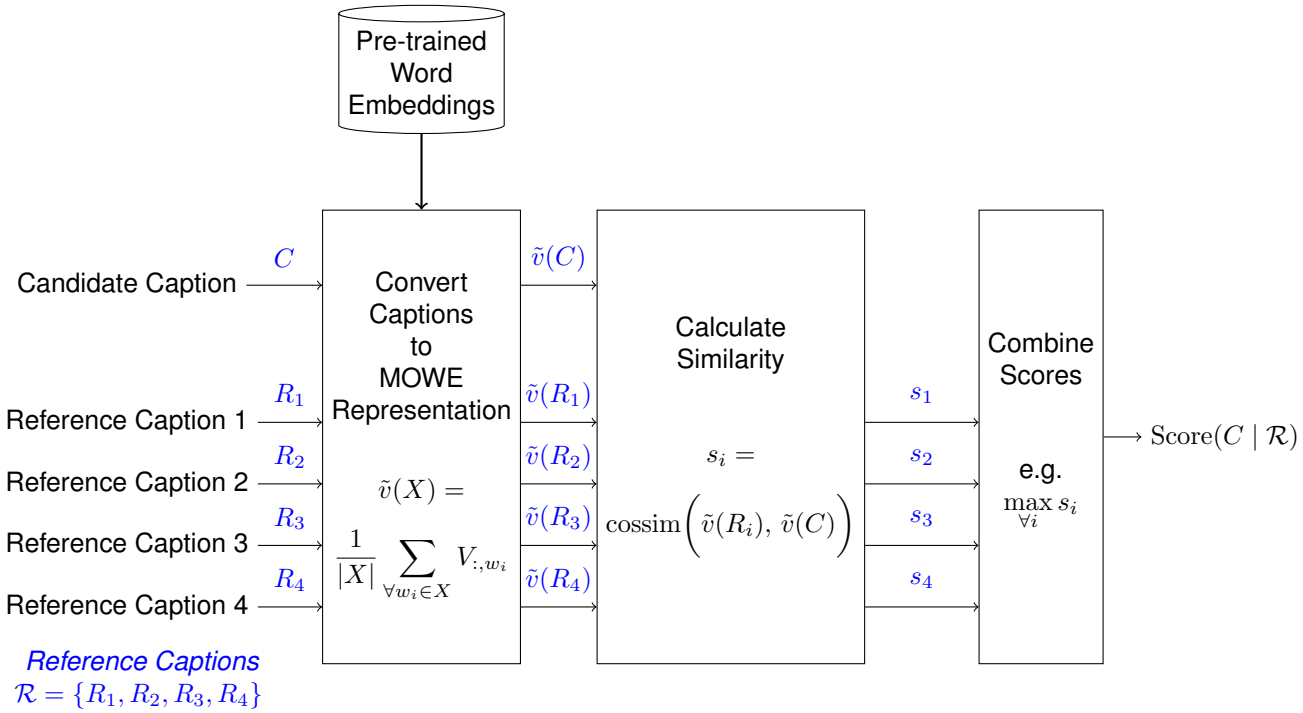


Fig. 2. Block diagram showing how WEmbSim is used to score one candidate caption (C) against four reference captions R_1, R_2, R_3, R_4.

ity of the caption evaluation measures rely on textual similarity to assess the caption quality, presuming that the reference caption sufficiently reflects the visual information. Thus, the quality of references play a crucial role in the case of text-similarity based measures. However, similar is the case with visual evaluation methods such as VIFIDEL which rely on visual detectors to accurately detect the objects. Moreover, comparing visual (source) and textual (target) information is not a straightforward task, and greatly adds to the complexity.

- **Supervised Metrics:** Supervised metrics such as NNEval [15] and LCEval [16] combine existing metrics into a single unified measure, which has shown improvement in performance. However, the drawbacks of learned metrics are their high complexity and subjectivity to the training examples.

- **Sums of Skip-gram-like Embeddings:**
The general idea of using sums of skip-gram-like embeddings to represent the meaning of phrases was presented in [26]. FastText, GloVE and Word2vec are all skipgram-like, in that they have a linear relationship between the input embedding and the log of the probability of co-occurring words. Due to this linear-log relationship, a sum of embeddings has an approximately linear-log relationship with the probability of words co-occurring with all the words in the sum (i.e., the intersection of the individual concurrences). This is similar to what would be predicted by a skip-gram-like formulation if the phrase being embedded had been treated as a single token. Thus by the distributional hypothesis ('Words that occur in the same context tend to have similar meaning' [27]), such a sum of words is expected to characterize the sentence meaning.
Authors in [11] considered a similar approach to measure the sentence similarity as part of their method for extractive summarisation. They found that using sums of embeddings, rather than more complex methods to generate sentence representations gave a superior performance. Authors in [28] and [29] found that the idea of using simple mean of embeddings shows strong performances in capturing sentence meaning.

- **Cosine Similarity:**
The use of the cosine similarity as a measure of correspondence between natural language strings, dates back to the work of [30] on latent semantic indexing (LSI). This approach was used in information retrieval, to find the most similar documents, based on LSI matrix. LSI vectors are dimensionally reduced word-document occurrence matrices, which often use a weighting that is proportional to with point-wise mutual information (e.g., tdf-if). Skip-gram-like word embeddings are effectively dimensionally reduced word-word occurrence matrices, using a weighting that is proportional to the point-wise mutual information [31].

The efficacy of sums of embeddings as sentence representations has been advocated in a number of research areas [10], [11]; and the cosine similarity has been very commonly used as a measure [42], [32]. Despite the success of simple mean of embeddings-based measures in other related domains,

their effectiveness for image caption evaluation has never been analyzed. To the best of our knowledge, this is the first work which evaluates the effectiveness of such a measure for captioning, across a number of benchmark datasets.

An unpublished concurrent work [33] uses pre-trained BERT embeddings and token-level computations to evaluate the similarity between two captions. In contrast, our metric is based on sentence-level representations, and is comparatively simpler. We focus on using feature-based strategies for language representations such as ELMo and Word2vec to avoid the overhead of including the Transformer based encoder architecture, which is used by BERTScore. Moreover, what we are proposing is a framework, which is not confined to a particular embedding technique.

III. THE PROPOSED METRIC

WEmbSim is a cosine similarity based measure which uses the mean of word embeddings for caption evaluation. WEmbSim is defined mathematically as follows. For a set of reference captions $\mathcal{R}$, and a candidate caption to be evaluated $C = [w_1, w_2, \ldots, w_n]$, we define the function $\tilde{v}(\cdot)$ which maps a caption to a vector via the mean of word embeddings representation using the embedding matrix V.

$$\tilde{v}(C) = \frac{1}{n} \sum_{\forall w_i \in C} V_{:,w_i} \tag{1}$$

Using this, and the cosine similarity function:

$$\operatorname{cossim}\left(\tilde{a},\, \tilde{b}\right) = \frac{|\tilde{a} \cdot \tilde{b}|}{|\tilde{a}||\tilde{b}|} \tag{2}$$

We define the scoring function as:

$$\operatorname{Score}(C \mid \mathcal{R}) = \underset{\forall R_i \in \mathcal{R}}{\operatorname{Combine}} \operatorname{cossim}\left(\tilde{v}(C),\, \tilde{v}(R_i)\right) \tag{3}$$

The scoring function is defined in terms of the *combining rule*. The combining rule specifies how to combine the scores for multiple references. The similarity is defined between two captions, and as in caption evaluation usually multiple references are provided, thus the scores must be combined. It is standard in other similar methods, such as METEOR [22] to use *max* as the combining function [22]. This corresponds to being generous with the score i.e., it is acceptable to be at least similar to one of the reference captions.

The alternative approach is to use *min* which requires the candidate to be similar to all of the reference captions since it reports the worst score. In contrast, using the *mean* strikes a balance between the two. We investigated all three i.e., *max*, *mean* and *min*, however, for brevity, we report the results for using the *mean* combining function for WEmbSim, as *mean* shows consistently better performance than either *min* and *max*.

A. Pre-trained Embeddings

We use off-the-shelf pre-trained embeddings, without any fine tuning for this task. The details of the embeddings we used in our investigations are given in Table I. For ELMo, we use the representations from the lowest layer of BiLM [7], as they show consistently better performance on our test datasets, compared to the other two higher layers or a linear combination of the three layers. As discussed in [7], the lower level LSTM states capture syntactical aspects of a sentence. For our captioning datasets, syntactic information proved to be more useful in differentiating between captions. We evaluate WEmbSim using different embedding to assess how various embeddings affect its performance.

B. Preprocessing

We perform preprocessing of the captions to convert them to a MOWE representation. We remove stop words, using the standard NLTK [35] stop-word list. We found in preliminary investigations that removing stop-words gives a small improvement across the board for all embeddings. We also remove any words that are outside the pre-trained vocabulary for the embeddings. We conducted some investigations into using charngram embeddings for out-of-vocabulary words; we however achieved only a very small improvement, as rare and stop words contribute little to the performance. Therefore, for simplicity we do not use these methods in our experiments.

C. Complexity Analysis

Our proposed metric WEmbSim, superficially sounds similar to the Word Mover's Distance (WMD) proposed by [23]. Both methods use word embeddings to compute the similarity of natural language. However, WMD is a much more complex method, as it attempts to attribute the distance at the word level. Whereas, our proposed method WEmbSim, just uses the mean of the word embeddings to embed sentences and measures the similarity between sentence embeddings. WMD has time complexity per pairwise evaluation $O(p^3 \log p)$, in contrast per pairwise evaluation time complexity of WEmbSim is $O(p^2)$, where p is the length of the captions. WEmbSim is much more similar to the lower bound that [23] places on WMD, which is referred to as the word centroid distance.

IV. EXPERIMENTS AND RESULTS

In this section we present the details of the four different experiments that we performed to validate the effectiveness of our proposed metric. Since our metric falls in the category of *unsupervised textual-comparison metrics*, we compare its performance to other commonly used image captioning metrics, which belong to the same category.

A. System-level Correlation

Strong correlation with human judgments is considered one of the most important qualities of an automatic evaluation metric [34]. We investigate the system-level correlation of our proposed metric against the human assessments. Since image captioning metrics are used to evaluate captioning systems, it is therefore important to compare metrics in terms of their system-level performance.

For this experiment we use the system-level manual evaluations of 12 teams that participated in MSCOCO Captioning Challenge and submitted their results on the validation set.

TABLE I
THE DETAILS OF THE PRE-TRAINED EMBEDDINGS USED IN OUR INVESTIGATIONS.

Name	Source	Dims	Corpus	Corpus Size	Vocabulary Size
GloVE 840B	[12]	300	Common Crawl	$8.4 \cdot 10^{11}$	$2 \cdot 10^6$
Word2vec	[6]	300	Google News (100B)	$1.0 \cdot 10^{11}$	$3 \cdot 10^6$
FastText	[13]	300	Wikipedia	$4.0 \cdot 10^9$	$3 \cdot 10^6$
ELMo	[7]	1,024	1B Word Benchmark	$1.0 \cdot 10^9$	$8 \cdot 10^5$

TABLE II
PEARSON'S p CORRELATION OF COMMONLY USED EVALUATION METRICS AGAINST 12 COMPETITION ENTRIES ON THE COCO VALIDATION SET [17]. OUR REPORTED SCORES WITH A * DIFFER FROM THE ONES REPORTED IN [17] M1: PERCENTAGE OF CAPTIONS THAT ARE EVALUATED AS BETTER OR EQUAL TO HUMAN CAPTION. M2: PERCENTAGE OF CAPTIONS THAT PASS THE TURING TEST. HIGHEST CORRELATION VALUES ARE SHOWN IN BOLDFACE.

	M1		M2	
Metric	p	p-value	p	p-value
$BLEU_4$	-0.02	0.95	-0.01	0.99
$ROUGE_L$	0.09	0.77	0.10	0.75
METEOR	0.61	0.03	0.59	0.03
CIDEr	0.47*	0.13*	0.50*	0.10*
SPICE	0.76*	0.00*	0.75*	0.01*
WMD	0.68	0.02	0.71	0.01
$WEmbSim_E$	0.47	0.13	0.48	0.12
$WEmbSim_G$	0.80	0.00	**0.76**	0.00
$WEmbSim_W$	0.71	0.01	0.64	0.02
$WEmbSim_F$	**0.83**	0.00	**0.76**	0.01

TABLE III
PAIRWISE CLASSIFICATION ACCURACY ON TWO DIFFERENT TASKS REPORTED ON ABSTRACT-50S AND PASCAL-50S [18]. THE HIGHEST ACCURACY FOR EACH TASK IS SHOWN IN BOLD FACE.

Metric	ABSTRACT-50S		PASCAL-50S		Avg.
	HHC	HHI	HHC	HHI	
$BLEU_4$	51.5	81.0	53.7	93.2	69.9
$ROUGE_L$	52.0	82.0	56.5	95.3	71.5
METEOR	48.5	46.5	**61.1**	97.6	63.4
CIDEr	50.5	78.5	57.8	98.0	71.2
SPICE	56.0	86.5	58.0	96.7	74.3
WMD	50.0	93.5	56.2	98.4	74.5
$WEmbSim_E$	55.5	**94.5**	56.5	98.3	76.2
$WEmbSim_G$	56.0	89.5	58.0	97.6	75.3
$WEmbSim_W$	57.5	90.5	59.3	**98.9**	76.6
$WEmbSim_F$	**60.5**	90.0	60.9	98.1	**77.4**
# Instances	1000	1000	1000	1000	4000

Similar to [17], we assume that the manual assessments on the validation and test set are fairly similar. In Table II we report Pearson's p against M1 and M2 human assessments, which were used to rank the captioning models. Where, M1 is the 'percentage of captions that are evaluated as better or equal to human captions' and M2 is the 'percentage of captions that pass the Turing test'.

In Table II, we report four variants of our metric; $WEmbSim_E$, $WEmbSim_G$, $WEmbSim_F$ and $WEmbSim_W$, which use ELMo, GloVe, FastText and Word2vec respectively. Table II shows that our best performing metric $WEmbSim_F$ significantly outperforms all exsisting handcrafted metrics, achieving a correlation of 0.834 and 0.755 with human judgments M1 and M2 respectively.

B. Pairwise Classification Accuracy

We also evaluate the accuracy of our proposed and several other key metrics at matching human consensus score for candidate captions. For this experiment we use PASCAL-50S and ABSTRACT-50S [18] datasets, which contain human consensus judgements for pairs of captions. The human scores for the candidate captions were collected through AMT, where the workers were shown three captions, two candidates and one reference, and were asked to pick the candidate which is most similar to the reference caption. This choice was based solely on sentence similarity, as the workers were never shown the corresponding image. PASCAL-50S contains 50 whereas ABSTRACT-50S contains 48 reference captions per candidate pair.

For our experiments we focus on comparing two human written captions i.e., 1) Human Human Correct (HHC) and **2**) Human Human Incorrect (HHI). In HHC and HHI categories, both captions are human written, where in HHC both are correct and in HHI one candidate caption is incorrect. PASCAL-50S also contains caption pairs generated by machine models (MM category). However, we notice that majority of the machine generated captions are of inferior quality, as the image captioning systems used for generating them are retrieval-based, rule-based or template-based [18]. Captions generated by such models are not representative of the state-of-the-art. Therefore, we only used captions from HHC and HHI category, as they are comparatively closer in quality to the captions generated by state-of-the-art models. We compute the scores for the given candidate pairs using five randomly selected references, as done in literature [19]. Ideally, a metric should assign a higher score to the captions preferred by humans.

Table III shows the results of this experiment. In HHC, most of the metrics achieve a low accuracy since differentiating between good quality captions is a difficult task. However, $WEmbSim_F$ achieves the highest accuracy of 60.5 on ABSTRACT-50S and comes very close (60.9) to the best performing metric METEOR (61.1) on PASCAL-50S. All other variants of our proposed measure, show a higher overall accuracy compared to other handcrafted measures. The results in Table III highlight that WEmbSim, is a strong baseline at matching human consensus scores for captions which differ in quality.

Original Caption	
'a single man standing above the crowd at a busy bar'	
Perturbed Captions	**Distraction Type**
a single man is wearing a life jacket in a boat'	*Share Person*
'a waitress is serving cola at a busy bar'	*Share Scene*
'a single man'	*Just Person*
'a busy bar'	*Just Scene*

Fig. 3. shows an image, corresponding correct caption, perturbed versions of the original caption and the perturbation types.

TABLE IV
THE DISTRACTION ANALYSIS ON BINARY FORCED-CHOICE TASKS. THE HIGHEST ACCURACY FOR EACH TASK IS SHOWN IN BOLD FACE.

Metric	Share Person	Share Scene	Just Person	Just Scene	Avg.
$BLEU_4$	83.5	82.4	54.9	67.7	72.1
$ROUGE_L$	86.8	86.8	83.4	94.1	87.8
METEOR	92.4	91.4	91.9	98.4	93.5
CIDEr	94.1	**93.1**	73.3	81.5	85.5
SPICE	88.5	88.8	78.1	92.0	86.9
WMD	**94.3**	92.5	86.3	96.9	92.5
$WEmbSim_E$	92.5	91.4	89.4	96.6	92.5
$WEmbSim_G$	91.8	90.2	95.1	99.6	94.2
$WEmbSim_W$	93.9	92.2	93.6	98.8	94.6
$WEmbSim_F$	92.5	91.2	**95.7**	**99.6**	**94.8**
# Instances	4596	2621	5811	2624	15,652

C. Distraction Analysis

We also analyse the robustness of our proposed metric against various sentence perturbations. For this experiment we use the dataset introduced in [43]. We report the performance on four different tasks namely **1**) Share Person (SP), **2**) Share Scene (SS), **3**) Just Person (JP) and **4**) Just Scene (JS). An example of each of the four tasks is shown in Figure 3. For the SP and SS tasks, the distractors contain the same person/scene as the correct caption, however, the remaining part of the sentence is different. The JS and JP distractors are just noun phrases and consist only of the scene/person of the correct caption.

A robust metric should assign a higher score to the correct caption as compared to its perturbed version. Table IV shows that $WEmbSim_F$ achieves the highest accuracy in JP and JS category, showing that it can identify when a complete sentence (unperturbed) is better than just a noun phrase (distractor). It can be noted, that other sophisticated metrics are comparatively not that robust to such perturbations. We also notice that the embedding-based measures such as WMD and the variants of our WEmbSim achieve a higher overall accuracy on the four tasks compared to the other measures.

Our metric performs no direct assessment of the word order and primarily focuses on the word content (semantics). Similar to SPICE, WEmbSim neglects fluency of captions as compared to other n-gram based measures. An obvious weakness of WEmbSim is that it can trivially be fooled if it is presented with a caption that has the right content, in a random order. For example, one could create an unordered list of keywords generated using object detection. This would be scored quite highly by WEmbSim. The metric is thus 'gameable'. However, if this were expected, such inputs could be readily detected and assigned a poor score either by using a general purpose language model, or by setting a minimal threshold on the BLEU score.

D. Is WEmbSim complementary to other commonly used captioning measures?

To elucidate the usefulness and complementarity of WEmbSim to the existing unsupervised metrics, we compute the correlation between our proposed measure and other existing metrics. For this experiment, we use the same dataset that we used for the system-level evaluation (Sect. IV-A for details). We report Spearman correlation coefficient to maintain consistency with the literature [2]. From Figure 4 we observe that the correlation between metrics that capture the similar linguistic properties is higher. For example lexical measures such as BLEU and $ROUGE_L$ correlate very strongly with each other compared to semantic measures such as SPICE and WMD. Correlation of METEOR is relatively lower to lexical measures compared to the semantic ones, since it captures the semantic information using synonym matching. Amongst all measures $WEmbSim_F$ shows a relatively lower correlation with the existing measures and even with the two other variants $WEmbSim_E$ and $WEmbSim_W$. This shows that $WEmbSim_F$ is complementary to the existing measures.

To further strengthen our claim, we linearly combine the scores of the commonly used measures and the variants of WEmbSim for 12 teams that participated in MSCOCO Captioning Challenge (Sec. 4.1 for details). We evaluate and report the Pearson correlation of the combined scores with the human judgements in Table V. It can be seen from Table V that combining $WEmbSim_F$ to the existing measures significantly improves their performance. This reflects the complementarity and effectiveness of our proposed measure.

V. DISCUSSION

The quality of the generated text can be rated on its *fluency* and *adequacy* [3]. *Fluency* is the quality of the text to flow-well, and have both natural, and grammatically correct word order. *Adequacy* is the quality of containing the correct and suitable information. A vast majority of the state-of-the-art

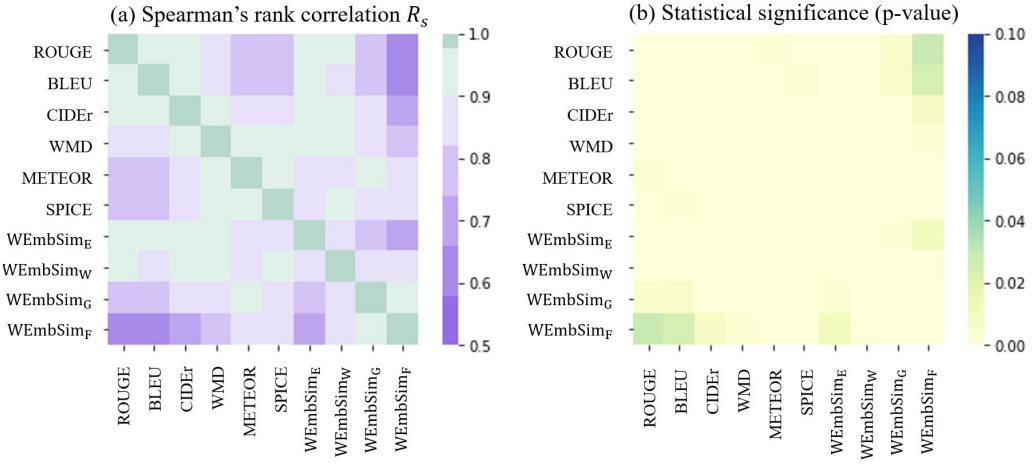


Fig. 4. Pairwise Spearman correlation of automatic metrics on MSCOCO captioning challenge validation set.

TABLE V
PEARSON CORRELATION BETWEEN STANDALONE/COMBINED METRICS AND HUMAN ASSESSMENTS FOR 12 TEAMS ON MSCOCO VALIDATION SET.

	Metric	Metric + WEmbSim$_E$	Metric + WEmbSim$_W$	Metric + WEmbSim$_G$	Metric + WEmbSim$_F$
BLEU$_4$	0.529	0.520	0.609	0.613	**0.652**
ROUGE$_L$	0.388	0.411	0.519	0.522	**0.567**
METEOR	0.744	0.658	0.748	0.778	**0.811**
WMD	0.682	0.618	0.712	0.736	**0.773**
CIDEr	0.465	0.466	0.500	0.500	**0.515**
SPICE	0.762	0.656	0.745	0.785	**0.813**

captioning models [1] are based on Recurrent Neural Network (RNN) [36] language models for text generation. RNN models perform exceedingly well at generating text with the correct order [36]. We thus can assume that the generated captions are usually fluent. Therefore, the meaningful composition of the words becomes a better discriminator in assessing the caption quality. Thus checking the word content takes priority over the word order.

Linear combination of word embeddings (mean,sum) have shown to be surprisingly powerful at the semantic tasks [37]. This is explained in large by their ability to capture how words are used [38], and a certain degree of lexical similarities, such as synonyms [39], [12]. If one has the correct word content, even a basic tri-gram language model is often able to produce the correct sentence, particularly for short sentences [40]. Captions are normally short, single sentences; rather than paragraphs of complex sentences. The simple linear combination of word embeddings has shown its ability in capturing the compositional semantics [41], which confirms our finding on its usefulness for caption evaluation in this work.

Our experiments reflect the significance of the choice of embeddings in our proposed model. FastText achieved the most promising results compared to other word embeddings. The difference in performance can be attributed to various factors such as the word semantics encoded by the pre-trained word embeddings, training criteria and training data. However, as the word embedding models will continue to evolve, the performance of our proposed metric can be improved even further by using stronger, effective and relevant embeddings. Note however, the purpose of this work is not to propose the best metric, but rather setting up a hard-to-beat simple baseline for more complex metrics to be justified.

VI. CONCLUSION

We have presented a new metric WEmbSim to automatically evaluate captions. WEmbSim$_F$, using FastText set of embeddings achieves the highest system-level correlation against existing unsupervised methods and, it also achieves state-of-the-art results on the pairwise classification task (HHI) on ABSTRACT-50S. A comprehensive correlation study against popular unsupervised metrics shows that WEmbSim is complementary to other commonly used measures.

Our work highlights the surprising effectiveness of word embeddings and WEmbSim can serve as a strong baseline for caption evaluation. This is in the same sense of a null model to any machine learning model. In order to justify the usefulness of a more complex model, it has to outperform the null model. Similarly, in order to justify a more complex metric, it has to perform better than WEmbSim.

Like many non-trivial methods for caption evaluation, WEmbSim does require an external resource for its definition. However, rather than the resource being a trained parser, or a lexical database, it is a set of pre-trained word embeddings. Such pre-trained embeddings are publicly available for over 150 languages [44]. In general, this makes our proposed method

more readily deployable than any other method requiring an external resource.

Acknowledgment

We are grateful to Nvidia for providing Titan-Xp GPU, which was used for our experiments. This work is supported by Australian Research Council, ARC DP150100294.

References

[1] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CSUR) 51.6 (2019): 1-36

[2] Kilickaya, Mert, et al. "Re-evaluating automatic metrics for image captioning." arXiv preprint arXiv:1612.07600 (2016)

[3] Sharif, Naeha, et al. "Learning-based composite metrics for improved caption evaluation." Proceedings of ACL 2018, student research workshop. 2018.

[4] Sober, Elliott. "The principle of parsimony." The British Journal for the Philosophy of Science 32.2 (1981): 145-156.

[5] Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi. "Unsupervised learning of sentence embeddings using compositional n-gram features." arXiv preprint arXiv:1703.02507 (2017).

[6] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[7] Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).

[8] Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka. "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't." Proceedings of the NAACL Student Research Workshop. 2016.

[9] Tshitoyan, Vahe, et al. "Unsupervised word embeddings capture latent knowledge from materials science literature." Nature 571.7763 (2019): 95-98.

[10] D'Haro, Luis Fernando, et al. "Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics." Computer Speech & Language 55 (2019): 200-215.

[11] Kågebäck, Mikael, et al. "Extractive summarization using continuous vector space models." Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC). 2014.

[12] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[13] Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[14] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[15] Sharif, Naeha, et al. "NNEval: Neural network based evaluation metric for image captioning." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[16] Sharif, Naeha, et al. "LCEval: Learned Composite Metric for Caption Evaluation." International Journal of Computer Vision 127.10 (2019): 1586-1610.

[17] Cui, Yin, et al. "Learning to evaluate image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[18] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[19] Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." European Conference on Computer Vision. Springer, Cham, 2016.

[20] Liu, Siqi, et al. "Improved image captioning via policy gradient optimization of spider." Proceedings of the IEEE international conference on computer vision. 2017.

[21] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[22] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005.

[23] Kusner, Matt, et al. "From word embeddings to document distances." International conference on machine learning. 2015.

[24] Jiang, Ming, et al. "TIGEr: Text-to-Image Grounding for Image Caption Evaluation." arXiv preprint arXiv:1909.02050 (2019).

[25] Madhyastha, Pranava, Josiah Wang, and Lucia Specia. "VIFIDEL: Evaluating the visual fidelity of image descriptions." arXiv preprint arXiv:1907.09340 (2019).

[26] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[27] Harris, Zelig S. "Distributional structure. WORD, 10: 146–162. Reprinted in J. Fodor and J. Katz, The structure of language: Readings in the philosophy of language." (1964): 33-49.

[28] White, Lyndon, et al. "How well sentence embeddings capture meaning." Proceedings of the 20th Australasian document computing symposium. 2015.

[29] Ritter, Samuel, et al. "Leveraging preposition ambiguity to assess compositional distributional models of semantics." Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. 2015.

[30] Dumais, Susan T., et al. "Using latent semantic analysis to improve access to textual information." Proceedings of the SIGCHI conference on Human factors in computing systems. 1988.

[31] Levy, Omer, and Yoav Goldberg. "Neural word embedding as implicit matrix factorization." Advances in neural information processing systems. 2014.

[32] White, Lyndon, et al. "Finding word sense embeddings of known meaning." 19th international conference on intelligent text processing and computational linguistics (CICLing). 2018.

[33] Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

[34] Zhang, Ying, and Stephan Vogel. "Significance tests of automatic machine translation evaluation metrics." Machine Translation 24.1 (2010): 51-65.

[35] Loper, Edward, and Steven Bird. "NLTK: the natural language toolkit." arXiv preprint cs/0205028 (2002).

[36] Kombrink, Stefan, et al. "Recurrent neural network based language modeling in meeting recognition." Twelfth annual conference of the international speech communication association. 2011.

[37] Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. "A simple but tough-to-beat baseline for sentence embeddings." (2016).

[38] Conneau, Alexis, et al. "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." arXiv preprint arXiv:1805.01070 (2018).

[39] Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. "Linguistic regularities in continuous space word representations." Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013.

[40] White, Lyndon, et al. "Modelling sentence generation from sum of word embedding vectors as a mixed integer programming problem." 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, 2016.

[41] Wang, Rui, Wei Liu, and Chris McDonald. "A matrix-vector recurrent unit model for capturing compositional semantics in phrase embeddings." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.

[42] Reisinger, Joseph, and Raymond Mooney. "Multi-prototype vector-space models of word meaning." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010.

[43] Hodosh, Micah, and Julia Hockenmaier. "Focused evaluation for image description with binary forced-choice tasks." Proceedings of the 5th Workshop on Vision and Language. 2016.

[44] Grave, Edouard, et al. "Learning word vectors for 157 languages." arXiv preprint arXiv:1802.06893 (2018).