

NNEval: Neural Network based Evaluation Metric for Image Captioning

Naeha Sharif¹, Lyndon White¹, Mohammed Bennamoun¹, and
Syed Afaq Ali Shah^{1,2}

¹ The University of Western Australia, 35 Stirling Highway, Crawley, Western
Australia

{naeha.sharif, lyndon.white}@research.uwa.edu.au,
mohammed.bennamoun@uwa.edu.au

² School of Engineering and Technology, Central Queensland University, Australia
s.shah@cqu.edu.au

Abstract. The automatic evaluation of image descriptions is an intricate task, and it is highly important in the development and fine-grained analysis of captioning systems. Existing metrics to automatically evaluate image captioning systems fail to achieve a satisfactory level of correlation with human judgements at the sentence level. Moreover, these metrics, unlike humans, tend to focus on specific aspects of quality, such as the n-gram overlap or the semantic meaning. In this paper, we present the first learning-based metric to evaluate image captions. Our proposed framework enables us to incorporate both lexical and semantic information into a single learned metric. This results in an evaluator that takes into account various linguistic features to assess the caption quality. The experiments we performed to assess the proposed metric, show improvements upon the state of the art in terms of correlation with human judgements and demonstrate its superior robustness to distractions.

Keywords: image captioning; automatic evaluation metric; neural networks; correlation; accuracy; robustness

1 Introduction

With the rapid advancement in image captioning research [25], [20], [12], [38], [39], [40], [29], [41], the need for reliable and efficient evaluation methods has become increasingly pressing. Describing images in natural language is an ability that comes naturally to humans. For humans, a brief glance is sufficient to understand the semantic meaning of a scene in order to describe the incredible amount of details and subtleties about its visual content [41]. While a reasonable amount of progress is made in the direction of replicating this human trait, it is still far from being solved [31], [17]. Effective evaluation methodologies are necessary to facilitate the fine-grained analysis for system development, comparative analysis, and identification of areas for further improvement.

Evaluating image description is more complex than it is commonly perceived, mainly due to the diversity of acceptable solutions [18]. Human evaluations can

serve as the most reliable assessments for caption quality. However, they are resource-intensive, subjective and hard to replicate. Automatic evaluation metrics on the other hand, are more efficient and cost effective. However, the automatic metrics currently in use for caption evaluation, fail to reach the desired level of agreement with human judgements at the sentence level [11], [21]. According to the scores of some metrics, the best machine models outperform humans in the image captioning task³ (Microsoft COCO challenge [8]), portraying an illusion that image captioning is close to being solved. *This reflects the need to develop more reliable automatic metrics which capture the set of criteria that humans use in judging the caption quality.*

Some of the automatic metrics which are commonly used to evaluate image descriptions, such as BLEU [33], METEOR [5] and ROUGE [27], were originally developed to assess *Machine Translation/Summarization* systems. Whereas, in the recent years CIDEr [37] and SPICE [4] were developed specifically for the image caption evaluation task and have shown more success as compared to the existing ones. All of these metrics output a certain score representing the similarity between the candidate and the reference captions. While there are a number of possible aspects to measure the quality of a candidate caption, all of the aforementioned metrics, rely on either lexical or semantic information to measure the similarity between the candidate and the reference sentences.

Our motivation to form a composite metric is driven by the fact that human judgement process involves assessment across various linguistic dimensions. We draw inspiration from the *Machine Translation* (MT) literature, where learning paradigms have been proposed to create successful composite metrics [6], [7]. Learning-based approach is useful because it offers a systematic way to combine various informative features. However, it is also accompanied by the need of large training data. To avoid creating an expensive resource of human quality judgements, we make use of a training criteria as inspired by [9], [24], which involves the classification of a caption as “human generated or machine generated”. This enables us to utilize the available human generated and machine generated data for training (Section 4.1).

While it is hard to find a globally accepted definition of a “good caption”, we hypothesize that the captions that are closer to human-generated descriptions can be categorized as acceptable/desirable. The better a captioning system, the more its output will resemble human generated descriptions. Moreover, the question of a caption being human or machine generated has the added advantage of being answered by the existing datasets containing human reference captions for corresponding images. Datasets such as MS COCO [8], Flickr30K [34], and Visual Genome [23] have multiple human generated captions that are associated with each image. These captions along with those generated by machine models can be used to train a network to distinguish between the two (human or machine), thus overcoming the need of the labour intensive task of obtaining human judgements for a corpus.

³ <http://cocodataset.org/#captions-leaderboard>

In our proposed framework, we cast the problem of image description evaluation as a classification task. A multilayer neural network is trained with an objective to distinguish between human and machine generated captions, while using the scores of various metrics based on lexical/semantic information as features. In order to generate a score on a continuous scale of $[0, 1]$, we use the confidence measure obtained through class probabilities, representing the believability of a caption being human-produced or otherwise. The proposed framework offers the flexibility to incorporate a variety of meaningful features that are helpful for evaluation. Moreover, with the evolution of image captioning systems, sensitive and more powerful features can be added in time. *To the best of our knowledge, this is the first “learning-based” metric designed specifically to evaluate image captioning systems.* Our main contributions are:

1. A novel learning-based metric, “NNEval”, to evaluate image captioning systems.
2. A learning framework to unify various criteria to judge caption quality into a composite metric.
3. A detailed experimental analysis reflecting various aspects of NNEval, its ability to correlate better with human judgements at the sentence level and its robustness to various distractions.

2 Related Work

2.1 Automatic Evaluation Metrics

The significance of reliable automatic evaluation metrics is undeniable for the advancement of image captioning systems. While image captioning has drawn inspiration from MT domain for encoder-decoder based captioning networks [38], [39], [40], [42], [29], it has also benefited from the automatic metrics that were initially proposed to evaluate machine translations/text summaries, such as BLEU [33], METEOR [10] and ROUGE [27]. In order to evaluate the quality of candidate captions, these metrics measure the similarity between candidate and reference captions, which is reported as a score (higher score reflects better caption quality).

In recent years, two metrics CIDEr [37] and SPICE [4] have been developed specifically to evaluate image captioning systems. CIDEr measures the consensus between the candidate and reference captions, primarily using lexical information. SPICE on the other hand, uses semantic information in the form of a scene-graph to measure the similarity between candidate and reference sentences. Both SPICE and CIDEr have improved upon the commonly used metrics such as BLEU, ROUGE and METEOR in terms of mimicking human judgements. However, there is still a lot of room for improving the sentence-level correlation with human scores [21]. Authors in [28] showed that optimizing the captioning models for a *linear combination* of SPICE and CIDEr scores can lead to better captions. This *linear combination* of metrics was termed as SPIDEr (SPICE + CIDEr). However, SPIDEr was not assessed for its correlation with human

judgements. Recently, [21] suggested the use of a distance measure known as “Word Movers Distance” (WMD) [26] for image caption evaluation, highlighting various strengths of this metric against the existing ones. WMD, which was originally developed to measure distance between documents, uses the word2vec [30] embedding space to determine the dissimilarity between the two texts.

2.2 Deterministic vs Learned Metrics

The currently used automatic metrics for image captioning, judge the caption quality by making deterministic measurements of similarity between candidate and references captions. These metrics tend to focus on specific aspects of correspondence, such as common sequences of words or the semantic likeness (using scene graphs). Moreover, these deterministic metrics fail to achieve adequate levels of correlation with human judgements at the sentence level, which reflects the fact that they do not fully capture the set of criteria that humans use in evaluating caption quality. One way to capture more features for evaluation is to combine various indicators, each of which focuses on a specific aspect, to form a fused metric [28].

Machine learning offers a systematic way to combine stand-alone deterministic metrics (or features related to them) into a unified one. In the literature related to MT evaluation, various learning paradigms have been proposed and the existing learned metrics can broadly be categorized as, *Binary functions*, - “which classify the candidate translation as good or bad” [24], [15] and *Continuous functions*, - “which score the quality of translation on an absolute scale” [3]. It is also shown that machine learning can be used to successfully combine stand-alone metrics and/or linguistic features to create composite evaluation metrics, showing a higher correlation with human judgments compared to the individual metrics [15], [3], [7].

2.3 Features

The features used by the learning-based metrics can be scores of the stand-alone metrics (such as BLEU, NIST, METEOR, and TER) and/or other numerical measurements reflecting, the lexical, the syntactic or the semantic similarity between candidate and reference captions. Various combinations of features have been proposed for the above mentioned paradigms in MT [3], [13], [16]. Moreover, combining meaningful linguistic features has shown promising results in metric evaluation campaigns, such as WMT (Workshop on Machine Translation) [6], [7]. Therefore, we hypothesize that a learning-based framework can be helpful in creating customized, reliable and efficient evaluators for captioning as well. We propose a neural network-based metric which combines the judgement of various existing metrics through a learning framework. Our work is more conceptually similar to the work in [24], which induces a *human-likeness* criteria. However, it differs in terms of the learning algorithm as well as the features used. In [24], a SVM classifier was trained with Gaussian Kernels to discriminate human and machine-like translations, using lexical features together with scores of individual metrics WER (Word Error Rate) and PER (Position-independent word Error

Rate) [36]. In contrast, *we propose the first neural network-based framework to induce a metric for caption evaluation*. Our feature set consists of individual metric scores, some of which are from captioning specific metrics and others are from metrics used in MT. We also include newer state-of-the-art MT metric ‘WMD’ as a part of our feature set. We believe that the novel combination of metrics used as features will allow our learned composite metric to correlate well with the human judgements.

3 NNEval

In this section, we describe the proposed metric in detail. The overall architecture of NNEval is shown in Fig. 1.

3.1 Proposed Approach

To create a machine learning-based metric that is well aligned with human evaluations, we frame our learning problem as a classification task. We adopt a training criteria based on a simple question: “is the candidate caption human or machine generated?” The human generated captions are still easily distinguishable from the machine generated ones [31], [17], as the former are of superior quality. Image captioning would be a solved problem, if outputs of image captioning systems were of such a high quality that they could not be distinguished from human generated captions. Exploiting this gap in quality, our trained classifier can set a boundary between human and machine produced captions. Furthermore, in order to obtain a continuous output score, instead of a class label, we use the class-probabilities. These probabilities represent the degree of confidence about a candidate belonging to one of the two classes. Thus the resulting evaluator’s output can be considered as some “measure of believability” that the input caption was human produced.

Another possible way to create a learned metric could be to directly approximate human judgement scores as a function of some feature set which is generated over the input captions. However, this approach would require the availability of a large training corpora containing human-evaluated candidate captions and their reference counterparts. The development of such a resource can be very difficult, time consuming and even prohibitive [24]. Framing our learning problem as a classification task allows the creation of a training set from existing datasets containing human reference captions for given images [9]. The human generated captions paired with various machine generated ones, for the given images, can serve as the training examples for the metric. Thus, mitigating the need of obtaining expensive manual annotations. Moreover, such a dataset can be updated easily by including the outputs of more evolved models without incurring any additional cost.

We use a fully connected multilayer feed-forward neural network as our learning algorithm to build the proposed metric. We describe the details of NNEval’s architecture and the learning task in Section 3.3, whereas the features used for NNEval in the following Section:

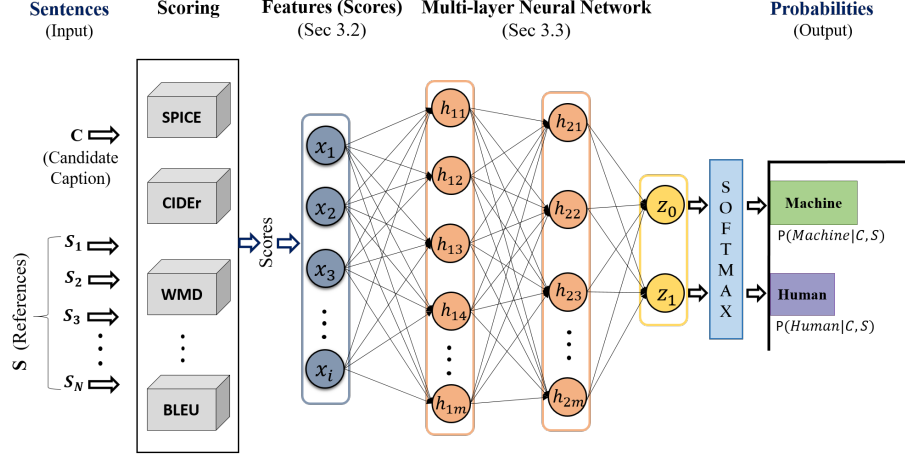


Fig. 1. Overall architecture of NNEval

3.2 NNEval Features

In our proposed framework, the candidate “C” and reference sentences “S” are not fed directly as an input to the neural network, instead, a set of numeric features are extracted from them as shown in Figure 1. Only the feature vector is given as an input to the neural network, not allowing the network to directly analyse the candidate and reference sentences. Each entity in the feature vector corresponds to a quality score generated by an individual metric for the given candidate. Metrics that we use to generate our feature vector are found to be statistically different from each other [21] and complement each other in assessing the quality of the candidate captions. Our basic feature set consists of the scores of the following metrics:

SPICE[4] estimates the caption quality by first converting both candidate and reference texts to a semantic representation known as a “scene graph”; which encodes the objects, attributes and relationships found in the captions. Next, a set of logical tuples are formed by using possible combinations of the elements of the graphs. Finally, an F-score is computed based on the conjunction of candidate and reference caption tuples.

CIDEr[37] measures the consensus between candidate and reference captions using n-gram matching. N-grams that are common in all the captions are down-weighted by computing the *Term Frequency Inverse Document frequency* weighting. The mean cosine similarity between the n-grams of the reference and candidate captions is referred as $CIDEr_n$ score. The final CIDEr score is computed as the mean of $CIDEr_n$ scores, with $n = 1, 2, 3, 4$, which we use as a feature.

BLEU[33] evaluates the candidate captions by measuring the n-gram overlap between the candidate and reference texts. BLEU score is computed via geometric averaging of modified n-gram precisions scores multiplied by a brevity

penalty factor to penalize short sentences. We use four variants of *BLEU* i.e., *BLEU*₁, *BLEU*₂, *BLEU*₃ and *BLEU*₄ scores as our features.

METEOR[5] judgement is based on the unigram overlap between the candidate and reference captions. It matches unigrams based on their meanings, exact forms and stemmed forms. Whereas, the metric score is defined as the harmonic mean of unigram precision and n-gram recall.

WMD[26] measures the dissimilarity between two sentences as the minimum amount of distance that the embedded words of one sentence need to cover to reach the embedded words of the other sentence. More formally, each sentence is represented as a weighted point cloud of word embeddings $d \in R_N$, whereas the distance between two words i and j is set as the Euclidean distance between their corresponding word2vec embeddings [30]. To use it as feature, we convert this distance score to similarity by using a negative exponential.

We use the MS COCO evaluation code [8] to implement all of the above metrics except for WMD. To implement WMD, we use the Gensim library script [35]. We also map all the feature values (scores) in the range of $[-1, 1]$, using the min-max normalization.

3.3 Network Architecture and Learning task

Given a candidate caption C and a list of references $S = \{S_1, S_2, S_3 \dots S_N\}$, the goal is to classify the candidate caption as human or machine generated. We model this task using a feed-forward neural network, whose input is a fixed length feature vector $\mathbf{x} = \{x_1, x_2, x_3, \dots x_i\}$, which we extract using the candidate caption and corresponding references (Section 3.2), and its output is the class probability, given as:

$$\mathbf{y}_k = \frac{e^{z_k}}{e^{z_0} + e^{z_1}}, k \in \{0, 1\} \quad (1)$$

Where z_k represents un-normalized class scores (z_0 and z_1 correspond to the machine and human class respectively). Our architecture has two hidden layers and the overall transformations in our network can be written as:

$$\mathbf{h}_1 = \varphi(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \quad (2)$$

$$\mathbf{h}_2 = \varphi(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2) \quad (3)$$

$$\mathbf{z}_k = \mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3 \quad (4)$$

\mathbf{W}_l and \mathbf{b}_l , are the weights and the bias terms between the input, hidden and output layers respectively. Where, $\mathbf{W}_l \in R^{N_l \times M_l}$, $\mathbf{b}_l \in R^{M_l}$ given $l \in \{1, 2, 3\}$. Moreover, $\varphi(\cdot) : R \rightarrow R$ is the non-linear activation function, given as:

$$\varphi(x) = \max(x, 0) \quad (5)$$

We use $P(k=1|\mathbf{x})$ as our metric score, which is the probability of an input candidate caption being human generated. It can be formulated as:

$$P(k=1|\mathbf{x}) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}} \quad (6)$$

The cross entropy loss for the training data with parameters, $\theta = (\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ can be written as:

$$J_{\theta} = -\frac{1}{p} \sum_{s=1}^p \log\left(\frac{e^{z_{\tilde{y}}^s}}{e^{z_0^s} + e^{z_1^s}}\right) + \beta L(\theta) \quad (7)$$

In the above equation $z_{\tilde{y}}^s$ is the activation of the output layer node corresponding to the true class \tilde{y} , given the input \mathbf{x}^s . Where, $\beta L(\theta)$ is a regularization term, that is commonly used to reduce model over-fitting. For our network we use L_2 regularization [32].

3.4 Gameability

A common concern in the design of automatic evaluation metrics is that the system under evaluation might try to optimize for the metric score, leading to undesirable outcomes [37], [4]. The resulting captions in such case might not be of good quality as per human judgement. However, by “gaming the metric”, a captioning system can achieve a higher than deserving performance, which may lead to false assessments. For instance, a metric that only takes into account the lexical similarity between the candidate and reference captions, might be gamed to assign a higher than deserving score to a caption that just happens to have many n-gram matches against the reference captions. Since, NNEval itself is a composition of various metrics, it has a built-in resistance against systems which have gamed only one or few of the subset metrics. Having said that, the potential of a system gaming against all, or a subset of features is still plausible.

4 Experimental Settings

To train our metric, a dataset which contains both human and machine generated captions of each image is required. We create a training set by sourcing data from Flickr30k dataset [43]. Flickr30k dataset consists of 31,783 photos acquired from Flickr⁴, each paired with 5 captions obtained through the Amazon Mechanical Turk (AMT). For each image in Flickr30k dataset we choose three amongst the five captions to use as human generated candidate captions. Whereas, we obtain machine generated captions of the same images, using three image captioning models, which achieved state-of-the-art performance when they were published [38], [39], [29]. In Section 4.1, we describe the training set-up for these image captioning models. In Section 4.2 and Section 4.3, we provide the details of the training and validation sets used for NNEval. The early stopping criteria and the network parameters for NNEval are discussed in Section 4.4 and Section 4.5 respectively.

4.1 Dataset for Image Captioning Models

The models that we use to obtain machine generated captions for our training set are: **1)** Show and tell [38], **2)** Show, attend and tell (soft-attention) [39],

⁴ <https://www.flickr.com/>

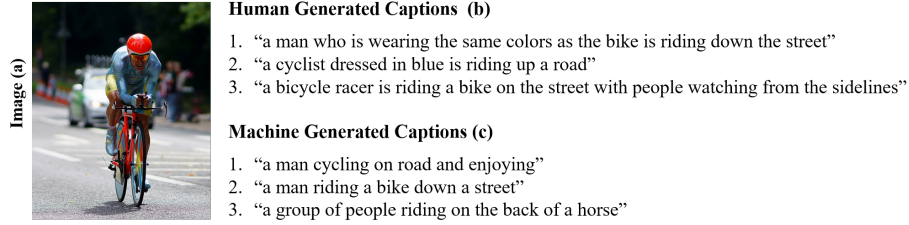


Fig. 2. shows, (a) an image from Flickr30k dataset, (b) human generated captions for the corresponding image, and (c) captions generated by machine models [38], [39], [29] for the given image.

and **3)** Adaptive attention [29]. We use publicly available official codes of these captioning models⁵ and train them on MS COCO dataset [8], which is one of the largest image captioning datasets. The models that are trained on a large dataset tend to give a better performance when tested on an unseen dataset. MS COCO dataset consists of training, validation and testing set containing 82,783, 40,504 and 40,775 images respectively. Each image in these sets is associated with five or more captions (collected through AMT), except for the testing set. We combine the MS COCO training and validation sets and use this combined set for the training of captioning models, while reserving 10,000 image-caption pairs for validation and testing purposes. We train the image captioning models using the original experimental protocols to achieve close to their reported performances.

4.2 Training Set for NNEval

We use the trained image captioning models discussed above to generate captions for the images in Flickr30k dataset. For each image, we obtain three machine generated captions, one from each model. Moreover, we randomly choose three captions amongst the five human produced captions, which were originally paired with their respective image in Flickr30k, to use as human generated captions. This provides us with an equal number of human and machine generated candidate captions per image. Figure 2 shows an example of human and machine produced candidate captions for a given image. In order to obtain reference captions for each candidate caption, we again utilize the human written descriptions of Flickr30k. For each machine-generated candidate caption, we randomly choose four out of five human written captions which were originally associated with each image. Whereas, for each human-generated candidate caption, we select the remaining four out of five original AMT captions.

We make sure that there is no overlap between each human candidate caption and its corresponding reference captions. In Figure 3, a possible pairing scenario is shown to demonstrate the distribution of the candidate and reference captions. If we select S1 as the candidate human caption, we choose S2, S3, S4, S5 as its

⁵ We thank the authors of these captioning approaches for making their codes publicly available.

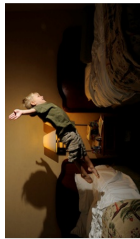
Image		Human Generated Captions		
		S_1 : “a boy is leaping from one bed to another in a room”		
		S_2 : “a small blond child jumping from one bed to another”		
		S_3 : “a little blond boy jumps from one bed to another”		
		S_4 : “a young blond boy is jumping from bed to bed”		
		S_5 : “a small boy leaps from one bed to another”		
		Machine Generated Captions		
		M_1 : “a man is jumping in the air on a bed”		
		M_2 : “a man in a white shirt and a tie”		
		M_3 : “a person jumping a bed in the air”		
(a)		(b)		(c)

Fig. 3. shows an image (a), its corresponding human and machine generated captions (b), and candidate (human and machine generated captions) and reference pairing for the given image in the training set (c).

references. Whereas, when we select M_1 as a candidate machine caption, we randomly choose any of the four amongst S_1, S_2, S_3, S_4, S_5 as references. While different sorts of pairing strategies can be explored, we leave that to future work. Moreover, the reason why we select four references for each caption is to exploit the optimal performance of each metric. Most of these metrics have been tested and reported to give a better performance with a larger number of reference captions [33], [10], [37], [4].

4.3 Validation Set for NNEval

For our validation set we draw data from Flickr8k [43], which consists of 8,092 images, each annotated with five human generated captions. The images in this dataset mainly focus on people and animals performing some action. This dataset also contains human judgements for a subset of 5,822 captions corresponding to 1000 images in total. Each caption was evaluated by three expert judges on a scale of 1 (the caption is unrelated to the image) to 4 (the caption describes the image without any errors).

From our training set we remove the captions of images which overlap with the captions in the validation and test sets (discussed in Section 5), leaving us with a total of 132,984 non-overlapping captions for training the NNEval model.

4.4 Early Stopping

NNEval is optimized over the training set for a maximum of 500 epochs, and tested for classification accuracy on the validation set after each epoch. While the loss function is used during the training period to maximize the classification accuracy, we are primarily interested in maximizing the correlation with human judgements. As accuracy is not a perfect proxy for correlation [24], we use early stopping based on Kendalls τ (rank correlation), which is evaluated on the validation set after each epoch. We thus terminate (early-stop) the training when the correlation is maximized. Since each caption in the validation set is paired with three judgements, we use the mode value of these three judgements to evaluate the correlation coefficient.

Table 1. Caption-level correlation of evaluation metrics with human quality judgements. All p-values (not shown) are less than 0.001

Metric	Pearson	Spearman	Kendall
BLEU-1	0.373	0.366	0.269
BLEU-4	0.223	0.360	0.267
ROUGE-L	0.381	0.376	0.279
METEOR	0.448	0.451	0.337
CIDEr	0.440	0.479	0.359
SPICE	0.475	0.482	0.376
SPIDEr	0.467	0.495	0.381
NNEval	0.532	0.524	0.404

4.5 Network Parameters

We use Adam optimizer [22] to train our network, with an initial learning rate of 0.25 and a mini-batch size of 75. We initialize the weights for our network by sampling values from a random uniform distribution [14]. Furthermore, we set the size of each of the hidden layers \mathbf{h}_1 and \mathbf{h}_2 (Section 3.3) to 72 nodes. The NNEval architecture is implemented using TensorFlow library [1].

5 Results and Discussion

To analyse the performance of our proposed metric, compared to the existing captioning metrics, we devise three sets of experiments, each judging a different aspect. *First* and foremost, we judge the metrics ability to correlate with human judgements (Section 5.1). *Second*, we observe how accurate it is in terms of distinguishing between two candidate captions given the human consensus over the pair (Section 5.2). *Third*, we observe the metric’s ability to deal with various distractions introduced in the candidate sentences (Section 5.3). In the latter two experiments, we report the accuracy instead of the correlation.

5.1 Correlation with Human Judgements

The purpose of designing automatic metrics is to replace human judgements. Therefore, the most desirable characteristic of an automatic evaluation metric is its strong correlation with human scores [44]. A stronger correlation with human judgements indicates that a metric captures the features that humans look for, while assessing a candidate caption. In order to measure the sentence-level correlation of our proposed metric with human judgements we use the COMPOSITE dataset [2] which contains human judgements for 11,985 candidate captions and their image counterparts. The images in this dataset are obtained from MS COCO, Flickr8k and Flickr30k datasets, whereas, the associated captions consist of human generated captions (sourced from the aforementioned datasets) and machine generated captions (using two captioning models [2], [19]). The candidate captions of images are scored for correctness on the scale of 1 (low relevance) to 5 (high relevance) by AMT workers. To ensure that the performance

Table 2. Comparative accuracy results on four kinds of caption pairs tested on PASCAL-50S

Metric	HC	HI	HM	MM	AVG
BLEU-1	53.5	95.6	91.1	57.3	74.4
BLEU-4	53.7	93.2	85.6	61.0	73.4
ROUGE-L	56.5	95.3	93.4	58.5	75.9
METEOR	61.1	97.6	94.6	62.0	78.8
CIDEr	57.8	98.0	88.8	68.2	78.2
SPICE	58.0	96.7	88.4	71.6	78.7
SPIDEr	56.7	98.5	91.0	69.1	78.8
NNEval	60.4	99.0	92.1	70.4	80.5

of NNEval is evaluated on unseen data, we remove the 771 image-caption pairs from this test set (which were overlapping with our validation set), leaving a total of 11,214 pairs for evaluation. Following the approach in [21], we report Pearsons r , Kendalls τ and Spearman's ρ correlation coefficients for commonly used caption evaluation metrics along with a newer metric SPIDEr (linear combination of SPICE and CIDEr) [28] .

The results in Table 1 show that NNEval outperforms the existing automatic metrics for captioning by a decent margin in terms of linear (Pearson) and rank based (Spearman and Kendall) correlation coefficients. *This is an improvement in the current state of the art.*

5.2 Accuracy

We follow the framework introduced in [37] to analyse the ability of a metric to discriminate between a pair of captions with reference to the ground truth caption. A metric is considered accurate if it assigns a higher score to the caption preferred by humans. For this experiment, we use PASCAL-50S [37], which contains human judgements for 4000 triplets of descriptions (one reference caption with two candidate captions). Based on the pairing, the triplets are grouped into four categories (comprising of 1000 triplets each) i.e., Human-Human Correct (HC), Human-Human Incorrect (HI), Human-Machine (HM), Machine-Machine (MM). The human judgements in PASCAL-50S were collected through AMT, where the workers were asked to identify the candidate sentence which is more similar to the given reference in the triplet. Unlike the previous study of [?], the AMT workers were not asked to score the candidate captions but to choose the best candidate. We follow the same original approach of [37] and use 5 reference captions per candidate to assess the accuracy of the metrics and report them in Table 2. The slight variation from the previously reported results [37], [4] might be due to the randomness in the choice of references.

The results in Table 2 show that on average, NNEval is ahead of the existing metrics. In terms of individual categories, it achieves the best accuracy in differentiating between Human-Human Incorrect captions. We believe that the reason that has contributed to this improvement, is that our validation set had all human generated captions, which were scored by human judges for their relevance

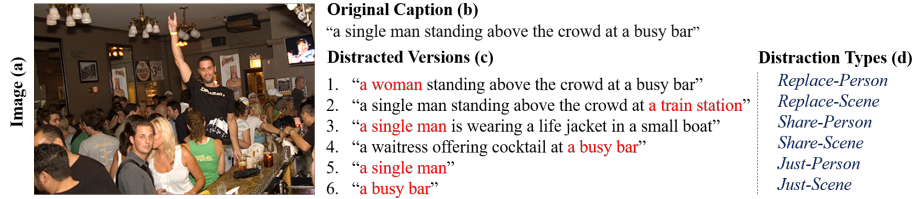


Fig. 4. shows an image (a), corresponding correct caption (b), distracted versions of the correct caption (c), and type of distraction in each caption (d).

to the image. Moreover, by using early stopping (Section 4.2), we selected the model which achieved the best correlation with human judgements. Hence, our model was optimized for this specific case of Human-Human Incorrect scenario. As evident from the results in Table 2, *NNEval is the most consistently performing model, with the highest average accuracy*. Note that HC is the hardest category as all the metrics produced the lowest accuracy in this category. In HC category, NNEval comes in only marginally behind the best performing metric METEOR. NNEval outperforms the three captioning specific metrics (CIDEr, SPICE and SPIDER) in three out of four categories, and is second only to SPICE in MM category with minor difference in the achieved accuracy.

5.3 Robustness

The authors in [17] introduced recently a dataset to perform a focused evaluation of image captioning systems with a series of binary forced-choice tasks, each designed to judge a particular aspect of image captions. Each task contains an image paired with two candidate captions, one correct and the other incorrect (distracted version of correct caption). For our evaluation, a robust image captioning metric should mostly choose the correct over the distracted one, to show that it can capture semantically significant changes in words and can identify when a complete sentence description is better than a single Noun Phrase. In [21], the authors used this dataset to perform their robustness analysis of various image captioning metrics. Following their approach, we also use the same dataset. However, we report the performance on six different tasks instead of the four reported in [21], namely **1)** Replace Person, **2)** Replace Scene, **3)** Share Person, **4)** Share Scene, **5)** Just Person and **6)** Just Scene. An example of each of the six tasks is shown in Figure 4. For the replace-scene and replace-person task, given a correct caption for an image, the incorrect sentences (distractors) were constructed by replacing the scene/person (first person) in the correct caption with different scene/people. For the share-person and share-scene tasks, the distractors share the same scene/task with the correct caption. However, the remaining part of the sentence is different. The just-scene and just-person distractors only consist of the scene/person of the correct caption.

We evaluate the metric scores for each correct and distracted version against the remaining correct captions that are available for an image in the dataset. The average accuracy scores for the caption evaluation metrics are reported in

Table 3. Comparative accuracy results on various distraction tasks

Metric	Replace Person	Replace Scene	Share Person	Share Scene	Just Person	Just Scene	AVG	Worst-case accuracy
BLEU-1	84.9	78.1	87.5	88.2	87.5	98.4	87.4	78.1
BLEU-4	85.9	75.2	83.5	82.4	54.9	67.7	72.1	54.9
ROUGE-L	83.3	71.1	86.8	86.8	83.4	94.1	84.1	71.1
METEOR	83.7	75.1	92.4	91.4	91.9	98.4	89.3	75.1
CIDEr	89.9	95.0	94.1	93.1	73.3	81.5	85.7	73.3
SPICE	84.0	76.0	88.5	88.8	78.1	92.0	83.6	76.0
SPIDEr	89.7	95.0	94.7	93.6	76.6	86.1	89.3	76.6
NNEval	90.2	91.8	95.1	94.0	85.8	94.7	91.9	85.8
#Instances	5816	2513	4594	2619	5811	2624	Total: 23977	

Table 3. The last row of Table 3 shows the numbers of instances tested for each category. It can be seen that NNEval outperforms the other metrics in three categories i.e., replace-person, share-person and share-scene task. Note that *NNEval is again the most consistent performer among all metrics. It has the best performance on average, and it also has the highest worst-case accuracy amongst all metrics.* Thus, we conclude that NNEval is overall the most robust metric.

6 Conclusion and Future Work

We propose NNEval, a Neural-Network based Evaluation metric which measures the quality of a caption across various linguistic aspects. Our empirical results demonstrate that NNEval correlates with human judgements better than the existing metrics for caption evaluation. Moreover, our experiments show that it is also robust to the various distractions in the candidate sentences. Our proposed framework, facilitated the incorporation of various useful features that contributed to the successful performance of our metric. In order to further improve NNEval to mimic human score, we intend to carry out a detailed analysis on the impact of various features on correlation and robustness. We plan to release our code in the coming months and hope that it will lead to further development of learning-based evaluation metrics and contribute towards fine-grained assessment of captioning models.

Acknowledgements

We are grateful to Nvidia for providing Titan-Xp GPU, which was used for the experiments. We would also like to thank Somak Aditya and Ramakrishna Vedantam for sharing their COMPOSITE and PASCAL-50S dataset respectively. This work is supported by Australian Research Council, ARC DP150100294.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., Fermüller, C.: Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding* (2017)
3. Albrecht, J.S., Hwa, R.: Regression for machine translation evaluation at the sentence level. *Machine Translation* **22**(1-2), 1 (2008)
4. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: *European Conference on Computer Vision*. pp. 382–398. Springer (2016)
5. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. pp. 65–72 (2005)
6. Bojar, O., Graham, Y., Kamran, A., Stanojević, M.: Results of the wmt16 metrics shared task. In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. vol. 2, pp. 199–231 (2016)
7. Bojar, O., Helcl, J., Kocmi, T., Libovický, J., Musil, T.: Results of the wmt17 neural mt training task. In: *Proceedings of the Second Conference on Machine Translation*. pp. 525–533 (2017)
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015)
9. Corston-Oliver, S., Gamon, M., Brockett, C.: A machine learning approach to the automatic evaluation of machine translation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. pp. 148–155. Association for Computational Linguistics (2001)
10. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: *Proceedings of the ninth workshop on statistical machine translation*. pp. 376–380 (2014)
11. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. vol. 2, pp. 452–457 (2014)
12. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al.: From captions to visual concepts and back (2015)
13. Giménez, J., Màrquez, L.: Linguistic features for automatic evaluation of heterogeneous mt systems. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. pp. 256–264. Association for Computational Linguistics (2007)
14. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. pp. 249–256 (2010)
15. Guzmán, F., Joty, S., Màrquez, L., Nakov, P.: Pairwise neural machine translation evaluation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. vol. 1, pp. 805–814 (2015)
16. Guzmán, F., Joty, S., Màrquez, L., Nakov, P.: Machine translation evaluation with neural networks. *Computer Speech & Language* **45**, 180–200 (2017)

17. Hodosh, M., Hockenmaier, J.: Focused evaluation for image description with binary forced-choice tasks. In: *Proceedings of the 5th Workshop on Vision and Language*. pp. 19–28 (2016)
18. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* **47**, 853–899 (2013)
19. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3128–3137 (2015)
20. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: *Advances in neural information processing systems*. pp. 1889–1897 (2014)
21. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E.: Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600* (2016)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* **123**(1), 32–73 (2017)
24. Kulesza, A., Shieber, S.M.: A learning approach to improving sentence-level mt evaluation. In: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. pp. 75–84 (2004)
25. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(12), 2891–2903 (2013)
26. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: *International Conference on Machine Learning*. pp. 957–966 (2015)
27. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004)
28. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. *arXiv preprint arXiv:1612.00370* (2016)
29. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. vol. 6 (2017)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. pp. 3111–3119 (2013)
31. van Miltenburg, E., Elliott, D.: Room for improvement in automatic image description: an error analysis. *arXiv preprint arXiv:1704.04198* (2017)
32. Ng, A.Y.: Feature selection, l1 vs. l2 regularization, and rotational invariance. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. pp. 78–. *ICML '04*, ACM, New York, NY, USA (2004). <https://doi.org/10.1145/1015330.1015435>, <http://doi.acm.org/10.1145/1015330.1015435>
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. pp. 311–318. Association for Computational Linguistics (2002)

34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Computer Vision (ICCV), 2015 IEEE International Conference on. pp. 2641–2649. IEEE (2015)
35. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), <http://is.muni.cz/publication/884893/en>
36. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., Sawaf, H.: Accelerated dp based search for statistical translation. In: Fifth European Conference on Speech Communication and Technology (1997)
37. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
38. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3156–3164. IEEE (2015)
39. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
40. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. *OpenReview* **2**(5), 8 (2016)
41. You, Q., Jin, H., Luo, J.: Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121* (2018)
42. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)
43. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* **2**, 67–78 (2014)
44. Zhang, Y., Vogel, S.: Significance tests of automatic machine translation evaluation metrics. *Machine Translation* **24**(1), 51–65 (2010)