COSMic: A Coherence-Aware Generation Metric for Image Descriptions

Mert İnan
University of Pittsburgh
mert.inan@pitt.edu

Piyush Sharma
Google Research
piyushsharma@google.com

Baber Khalid
Rutgers University
baber.khalid@rutgers.edu

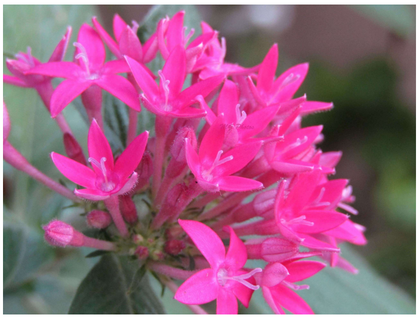
Radu Soricut
Google Research
rsoricut@google.com

Matthew Stone
Rutgers University
mdstone@cs.rutgers.edu

Malihe Alikhani
University of Pittsburgh
malihe@pitt.edu

Abstract

Developers of text generation models rely on automated evaluation metrics as a stand-in for slow and expensive manual evaluations. However, image captioning metrics have struggled to give accurate learned estimates of the semantic and pragmatic success of output text. We address this weakness by introducing the first discourse-aware learned generation metric for evaluating image descriptions. Our approach is inspired by computational theories of discourse for capturing information goals using coherence. We present a dataset of image–description pairs annotated with coherence relations. We then train a coherence-aware metric on a subset of the Conceptual Captions dataset and measure its effectiveness—its ability to predict human ratings of output captions—on a test set composed of out-of-domain images. We demonstrate a higher Kendall Correlation Coefficient for our proposed metric with the human judgments for the results of a number of state-of-the-art coherence-aware caption generation models when compared to several other metrics including recently proposed learned metrics such as BLEURT and BERTScore.



Caption		Coh.	CIDEr	COSMic
Model	first flower of the year	Story	0.000	0.653
Human	close-up of pink flowers	Visible		

Figure 1: A comparison of the scores for a generated (Model) caption that has a different coherence relation than the reference (Human) caption. "Coh." represents the coherence labels for generated and reference captions. Our coherence-aware metric COSMic is aware of the different information goals for these captions, and assigns a more adequate score when comparing the Model caption against the Human caption. In this case where a caption that does not just describe the image but elaborates on it, our metric recognizes that the model output is potentially successful (Photo credit: Moorthy Gounder)

1 Introduction

An investigation of the descriptions used with images on the web shows that image descriptions can have different functions and goals (Kruk et al., 2019a; Alikhani et al., 2020). For instance, captions may describe visible entities, activities and relationships, provide background information that goes beyond what's visible, or report the writer's own subjective reactions to what's displayed. By drawing on such diverse examples, image captioning models can learn the different inferential links between text and images and use that information at generation time to produce descriptions that can fulfill different discourse goals and inject the desired context into their output (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2014; Anderson et al., 2016a).

So far, however, efforts to develop such expressive captioning models have been hindered by the lack of automatic metrics that can evaluate their output with respect to their information goals in context. Previous approaches to automatic caption evaluation have mostly focused on n-gram measures of similarity to reference output (Vedantam et al., 2014); such surface-level models fail to deal with the lexical and syntactic diversity of image descriptions. More recent approaches more closely approximate semantic similarity using word embedding-based techniques. These models show

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3419–3430
November 7–11, 2021. ©2021 Association for Computational Linguistics

robust performance and achieve a higher correlation with human judgments than that of previous metrics. Nevertheless, they too fail to generalize to the different kinds of content that successful descriptions may exhibit across different goals and contexts. That is, they cannot distinguish reasonable descriptions that happen to differ from reference output in their goals and perspective, from problematic descriptions that hallucinate inappropriate content or context.

To bridge this gap, we present a coherence-aware embedding-based generation metric that learns to respect diverse discourse goals without penalizing captions that are purposefully generated to fulfill different purposes or communicate background information. Figure 1 demonstrates this capability by presenting an example image and captions with different coherence labels together with their scores.

Our approach to modeling discourse goals is based on the framework of discourse coherence theory (Hobbs, 1985), which characterizes the inferences that give discourse units a coherent joint interpretation using a constrained inventory of coherence relations. In particular, we use the taxonomy for image–text coherence developed by Alikhani et al. (2020), which for example includes *Visible*, *Story* and *Subjective* relations between the text and the image. A description and an image stand in a *Visible* relation if the text includes information that is recognizably depicted in the image. *Subjective* captions react to the content of the image and *Story* captions provide a free-standing description of the circumstances depicted in the image similar to the *Narration* relation in text. Our metric is learned in part from a new dataset of 4000 images with descriptions labeled with different coherence labels in this taxonomy.

In inaugurating the study of coherence-aware generation metrics, we make the following specific contributions. In Section 3 we present two different, annotated datasets for training and testing a coherence-aware metric. We present a model to score a generated caption given the image, reference caption, and the discourse goals of both these captions (Section 4). We compare this metric to previous ones using a common methodology, ranking the performance of several different caption generation systems on out-of-domain images—relying on a new benchmark out-of-domain test set, which we publish, providing reference captions for a subset of OpenImages (Kuznetsova et al., 2020b). Our experiments demonstrate that among all these metrics, our proposed metric has the highest correlation with human judgments.

2 Related work

There are diverse ways of characterizing the contributions of text and imagery. Gao et al. (2015) investigate the genre of image captions and Huang and Kovashka (2016) study the persuasive implicit relationships between text and images. Kruk et al. (2019b) study the emotional links between text and images. Otto et al. (2019) present an annotated dataset of text and imagery that compares the information load in text and images. However, we build on works that study information-level inferences between discourse units in different modalities such as comic book panels (McCloud, 1993), movie plots (Cumming et al., 2017), and diagrammatic elements (Hiippala et al., 2021). In particular, we use Alikhani et al. (2020)'s relations that characterize inferences between text and images.

Coherence-aware models have benefited several NLP tasks such as gesture interpretation (Lascarides and Stone, 2009; Pustejovsky and Krishnaswamy, 2020), text summarization (Xu et al., 2019), machine comprehension (Gao et al., 2020). The majority of these works use Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and Penn Discourse TreeBank (PDTB) (Prasad et al., 2008b) datasets to learn and predict these relations between two adjacent text spans. In this line of work, we are the first to present a coherence-aware generation metric.

The most widely used automatic evaluation metrics are ngram-based, which compute the exact number of ngram matches between reference and generated text (Cui et al., 2018). Examples of such metrics that are commonly used for evaluating the output of captioning, translation and summarization models are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), . The major problem of the n-gram similarity metrics is that they give no credit to synonym matches of reference n-grams, even if those words are common and used appropriately in the generated text. Embedding-based metrics such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2020) designed to address this limitation are closer to human ratings. BLEURT is a data-intensive training scheme that is based on BERT (Devlin et al., 2019) fine-tuned on human

Facade of a glass building.

A pink flower bush in a garden.

The underside of the Arc de Triomphe.

Close-up of a fly sitting on a daisy.

Man sitting by his artwork looking at a large statue of a man on a horse in a royal courtyard.

Woman with an umbrella reading a book sitting in the grass in front of a city skyline.

Cowboy on a horse and cowboy on the ground working together to lasso a calf in a pen.

Black and white artwork painted on a blue wall.

Figure 2: Examples of the ground truth captions that we collected for the COIN dataset. (Photo credits from left to right, top to bottom: Sharron Mollerus, Northfielder, George M. Groutas, davebloggs007, Tim Adams, Brisbane City Council, Colin Brown, Guilhem Vellut)

ratings of generated text. BERTScore, however, computes the similarity score as the average of cosine similarities between predicted tokens and their top matching reference tokens. These metrics however, do not respect the information goal and the purpose for which the model has generated the text. We address this problem by introducing the first coherence-aware generation metric. Similar to SPICE (Anderson et al., 2016b) and VIFIDEL (Madhyastha et al., 2019) we use the information encoded in images. We further propose the addition of coherence relations that facilitate learning with fewer samples by a multimodal metric using pre-trained BERT and ViLBERT.

3 Data Collection

We collect two datasets: human judgments for image captions that are generated by coherence-aware captioning systems using Conceptual Captions dataset; and ground-truth labels for the Open Images dataset. With Conceptual Captions corpora we fine-tune ViLBERT with ratings and show that addition of coherence relations can make automated scoring closer to human scoring. We use OpenImages corpora to reinforce that multimodality and coherence relations have significant contributions to scoring out-of-domain datasets, as well.

Protocol We hired two expert linguists for data annotation and designed an annotation website to facilitate the annotation procedure. They are native English speakers who identify themselves as of White and Latino ethnicity. The code [1] of the annotation website, and the details of the protocol is publicly available. The study has been approved by our institution's human subject board.

Conceptual Captions Score Annotation We have collected ratings on the quality of different image descriptions with coherence labels for a subset of 1000 images from the Conceptual Captions (CC) training dataset (Ng et al., 2020). With this paper, we are publishing this dataset as a benchmark for evaluation metrics that are coherence-aware. The set-up of the data collection is as follows: CC images are input into a caption-generation model created by Alikhani et al. (2020). This model generates coherence-aware descriptions for input images in 4 different coherence classes of `Meta`, `Visible`, `Subjective`, `Story`. These 4,000 image/caption pairs are then presented to human annotators who are asked to select the correct coherence label for each pair:

- *Meta:* the caption talks about when, where, and how the picture is taken. *Meta-talk* in Schiffrin (1980)
- *Visible:* the caption is true just by looking at the picture. *Restatement* relation in Prasad et al. (2008a).
- *Subjective:* the captions is the matter of opinion. *Evaluation* relation in Hobbs (1985).
- *Story:* text and image work like story and illustration. *Occasion* relation in Hobbs (1985).

[1] `https://github.com/Merterm/COSMic`

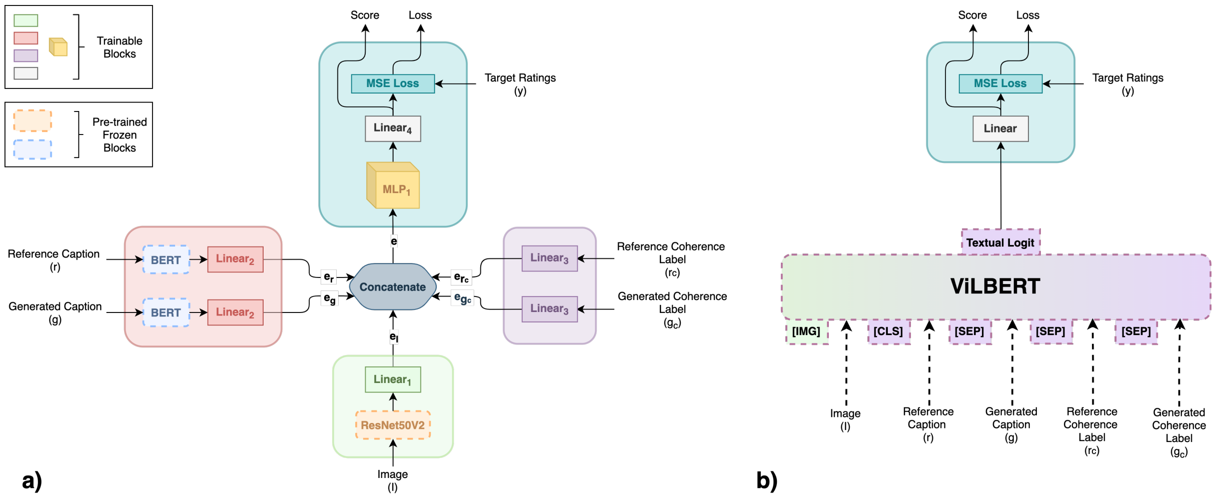


Figure 3: An illustration of different flavors of COSMic that outputs a score for the generated caption given the image, reference caption, and the coherence-labels for both the captions. (a) COSMic Vanilla uses only global textual and visual features, while (b) COSMic ViLBERT uses combined visio-linguistic features with both local and global focus. This model takes into account the information goals (determined by coherence-labels) for both the captions when comparing the generated caption to the reference for evaluation.

After the annotator selects a specific coherence label from the above, we ask them to rate the quality of the captions, given the label, on a scale of 1 to 5. We use these annotations as training data for our coherence-aware captioning metric, COSMic. We call this data we annotated RaCCoon (Ratings for Conceptual Caption).

To calculate the Cohen's κ agreement measure, we selected 150 images randomly and assigned them to two annotators. The Kappa coefficient is $\kappa = 0.89$ which indicates a substantial agreement (Viera and Garrett, 2005)

OpenImages Ground Truth Captions To create an out of domain test set we asked our annotators to write *Visible* captions for 1,000 images[2] from the OpenImages dataset (Kuznetsova et al., 2020a). We call this dataset COIN (**C**orpus of **O**pen**I**mages with **N**atural descriptions). A sample of these ground truth captions written by our expert linguists are presented in Figure 2. We use this dataset to test COSMic and other learned metrics in Section 5 and present our benchmark results in Table 1.

4 Method

The goal of a coherence-aware image captioning metric is to predict a score for the generated caption given the image, reference caption, and coherence relations of one generated caption and one reference caption. This metric function M can be formalized as predicting a score s as follows:

$$s = M(I, g, r, g_c, r_c; \theta) \tag{1}$$

where the metric is defined by parameters θ, and where the model inputs are defined as I being the image being captioned, g and r the generated and reference captions, respectively. g_c and r_c are the coherence relations for g, r respectively.

We now describe the architecture of our coherence-aware image captioning metric, COSMic (**CO**herence-**S**ensitive **M**etric of **i**mage **c**aptions). It has two flavors — a ViLBERT-based model pre-trained on large multimodal data, and a baseline Vanilla version, as illustrated in Figure 3. Both are trained on RaCCoon training data (Section 3) with normalized human annotated rating to obtain the model's target score.

4.1 COSMic ViLBERT

ViLBERT (Lu et al., 2019) is a multimodal feature learning model pre-trained on 3.3 million Conceptual Captions image and captions data. It is trained for masked multi-modal learning and multi-modal alignment prediction and demonstrates strong performance on several downstream multimodal tasks such as VQA, VCR, grounding, and image retrieval. For this reason we use a pre-trained ViLBERT to embed our multimodal inputs shown in Equation 1 with changes to incorporate both the captions and coherence relations.

[2] The same subset, named T2, was used for the CVPR-2019 Workshop on Conceptual Captions, www.conceptualcaptions.com.

For input image (I), we use the same process as ViLBERT. We use a Faster R-CNN (Ren et al., 2016) model pre-trained on Visual Genome (Krishna et al., 2016) to detect objects regions and extract features. The sequence of these image features is denoted as I' with 100 bounding box features where each element is R^{2048}. Similar to ViLBERT, we use the special token *[IMG]* to denote the beginning of the bounding box features list.

For input captions (g, r) and coherence labels (g_c, g_r), the sequence begins with the special token *[CLS]* followed by input text embeddings. Each of our text inputs are tokenized and embedded using ViLBERT's input text pre-processing and denoted as g', r', g'_c, g'_r for g, r, g_c and g_r respectively. Note that the coherence labels are processed as text inputs such as *"Visible"* and *"Story"* which allows the model to use its pre-trained representations of these concepts. Each of these input sequences are separated by the special token *[SEP]* to form our input sequence.

Hence, our input to ViLBERT is of form:

$$v = (\textit{[IMG]}, I', \textit{[CLS]}, r', \textit{[SEP]}, g', \textit{[SEP]}, r'_c, \textit{[SEP]}, g'_c)$$

We use a linear layer with sigmoid activation on ViLBERT's output text logits to compute COSMic's output metric score (s).

$$s = \text{Linear}(\text{ViLBERT}(v)) \tag{2}$$

During training, we fine-tune ViLBERT and the output linear layer in an end-to-end fashion by minimizing the Mean-Squared error between the output score, s and the corresponding reference score, y, on the RaCCoon dataset.

4.2 COSMic Vanilla

The COSMic ViLBERT approach above takes advantage of multimodal pre-training on the Conceptual Captions dataset to embed the image and text inputs. As a simpler baseline, we now present COSMic Vanilla which independently embeds the input image and text to be later combined for score computation with no end-to-end training.

To extract image features, we use a ResNet50v2 (He et al., 2015) model pre-trained on ImageNet (Deng et al., 2009) and linearly transform the global image representation to 512-dimensional space.

$$e_I = \text{Linear}_1(\text{AveragePool}(\text{ResNet}(I))) \tag{3}$$

In our textual feature extraction module, we embed g and r independently with a pre-trained BERT-Large-512 model. We use the *[CLS]* token embedding as 1024 dimensional caption-level representation in each case and transform them to 512-dimensional space.

$$\begin{aligned} e_g &= \text{Linear}_2(\text{BERT}_{\text{CLS}}(g)) \\ e_r &= \text{Linear}_2(\text{BERT}_{\text{CLS}}(r)) \end{aligned} \tag{4}$$

In our coherence label embedding module, g_c and r_c are each represented as one-hot vectors such that the dimensions correspond to labels *Meta*, *Visible*, *Subjective* and *Story*. Each is embedded into a 512-dimensional space.

$$\begin{aligned} e_{g_c} &= \text{Linear}_3(g_c) \\ e_{r_c} &= \text{Linear}_3(r_c) \end{aligned} \tag{5}$$

We thus obtain the 5 vectors (each R^{512}), representing one of the inputs of Equation 1. We concatenate and use a feed-forward network with progressively smaller hidden layers of sizes $[512, 256, 128, 64, 32, 16, 8]$, each with ReLU (Agarap, 2018) activation. The output score, s, is computed by a final linear layer on top of the above network.

$$\begin{aligned} e &= \text{concat}([e_I, e_g, e_r, e_{g_c}, e_{r_c}])) \\ s &= \text{Linear}_4(\text{MLP}_1(e)) \end{aligned} \tag{6}$$

where $e \in R^{2560}$ and $s \in R$.

To understand the role of each component of this implementation, we further deconstruct each module in ablation experiments described in Table 2.

4.3 Coherence-aware Captioning Systems

In order to experiment with COSMic, we generate our own captions. In this section we describe the coherence-aware captioning systems used to generate these image captions for the training and testing of COSMic.

For our base captioning system, we use the state-of-the-art coherence-aware captioning system introduced by (Alikhani et al., 2020). It uses a Transformer-based (Vaswani et al., 2017) encoder-decoder architecture where the encoder inputs are (1) global image features, (2) image labels, and (3) coherence label. The coherence-label also serves as the first input token for the decoder which generates the output captions. We set the coherence label to the groundtruth relation at training time, and the desired relation at inference time. We use the Conceptual Captions dataset (Sharma et al., 2018) with machine-generated coherence labels for

System		Avg. Hum. Rating	Metrics											
Model	Coh. Label		B_1	B_2	M	R_L	C	S	BR	BS-F	COSMic Vanilla	COSMic ViL-BERT	COSMic Vanilla+	COSMic ViL-BERT+
BUTD	Visible	2.191	.163	.077	.049	.160	.092	.030	-.877	.863	.706	.796	.522	.641
Base	Visible	3.532	.050	.025	.019	.066	.020	.002	-1.114	.862	.696	.777	.516	.614
	Meta	3.213	.041	.000	.012	.063	.012	.000	-1.059	.863	.548	.727	.505	.602
	Subj.	2.830	.033	.012	.011	.057	.017	.000	-1.197	.849	.323	.421	.358	.403
	Story	2.915	.029	.000	.017	.058	.013	.000	-1.304	.842	.533	.629	.482	.527
Lite	Visible	3.298	.028	.011	.013	.053	.011	.000	-1.101	.863	.684	.784	.515	.604
	Meta	2.830	.026	.010	.008	.055	.015	.000	-1.084	.859	.548	.748	.511	.565
	Subj.	2.298	.039	.012	.019	.066	.024	.003	-1.217	.849	.364	.451	.379	.419
	Story	2.426	.036	.000	.018	.062	.021	.000	-1.362	.842	.568	.666	.499	.519
Kendall's Correlation (τ)		1.000	.071	.154	.036	-.036	-.571	-.052	.286	.445	.571	.546	.667	**.764**

Table 1: System-level scores for 9 different image captioning systems as evaluated by human annotators and various captioning metrics. Bottom-Up Top-Down (BUTD) is trained on COCO, while others are trained on the Conceptual Captions (CC) dataset. The evaluation however is conducted on COIN dataset, which is out-of-domain for both COCO and CC. This domain shift causes the n-gram based metrics (e.g. BLEU, ROUGE, CIDEr) to assign very low scores to otherwise correct captions (See Table 4). Whereas embedding based metrics (e.g. BLEURT, BERTScore and COSMic) do not suffer from this limitation. Since all metrics have different scales, instead of absolute scores, we use Kendall Rank Correlation to measure agreement with human scores. Model names are abbreviated as follows: B_1: $Bleu_1$, B_2: $Bleu_2$, M: METEOR, R_L: $ROUGE_L$, C: CIDEr, S: SPICE, BR: BLEURT, BS-F: BERTScore F1. COSMic models with '+' denote application of data augmentation to remove training data bias. More metrics and detailed results can be found on the code repository.

training this captioning system. To obtain the coherence labels above, we closely follow (Alikhani et al., 2020) to train a coherence classifier on the Clue dataset (Alikhani et al., 2020) that provides around 4K human annotated (image, caption, relation) triplets. We present two caption-generation systems in this section.

Base-systems family A family of 4 captioning systems is created by setting the coherence-label to *Meta*, *Visible*, *Subjective* or *Story* in the base captioning model described above. These are considered different captioning systems because the information content and discourse goals, as controlled by the coherence label, are different.

Lite-systems family We remove the global image features from the base model's input to obtain a smaller, light-weight (lite) model. Similar to the base model, we obtain a family of 4 captioning systems by changing the coherence-label.

In Section 5, we study the order in which several image captioning metrics rank these 8 systems. The goal is to identify the metric that agrees the most with the groundtruth rankings based on human assessments.

4.4 COCO-trained Captioning System

COSMic's training data, RaCCoon, is based on Conceptual Captions and it is coherence-aware. To test the model's generalization capability, we use a captioning system trained on MS COCO (Chen et al., 2015). Since COSMic expects an input coherence label, and COCO captions are *Visible* style by design, we set the label to *Visible*. Specifically, we use the Bottom-Up Top-Down (BUTD) Attention model (Anderson et al., 2018). This helps study how well COSMic generalizes to other captioning datasets and coherence-agnostic captioning systems.

5 Experiments

Here, we describe the experimental setup to compare COSMic with other metrics. As outlined in Section 3 and 4, we use the RaCCoon data to train our models, and COIN to test COSMic and other metrics. We have several baseline metrics that we compare to, which can be found on Table 1.

5.1 Model Training Setup

We implement COSMic—as described in Section 4—with PyTorch (Paszke et al., 2019) and train on a GTX1080 GPU. We pre-compute BERT[3] and ResNet[4] features using their TensorFlow (Abadi et al., 2015) implementations. We use the pub-

[3] https://github.com/google-research/bert

[4] https://www.tensorflow.org/api_docs/python/tf/keras/applications/ResNet50V2

lic ViLBERT[5] implementation. We use a batch size of 4, and a learning rate of 2×10^{-6} for fine-tuning ViLBERT and use RAdam optimizer and stop the training when the validation score does not change for 3 epochs. For COSMic Vanilla, we train with a batch-size of 10, Adam optimizer (Kingma and Ba, 2017) with a base learning rate of 10^{-3} that decays by a factor of 10^{-2} every 10 epochs. We observe that the Vanilla converges in approximately 100 epochs and ViLBERT converges in 9 epochs. ViLBERT has 250 million parameters. COSMic Vanilla includes 3,062,913 trainable parameters. Pre-trained BERT-Large and ResNet50V2 have an additional 350 million parameters. The setup for coherence-aware captioning models to obtain machine-generated captions for our study is the same as (Alikhani et al., 2020).

5.2 Baseline Captioning Metrics

To benchmark COSMic, we compare it with other learned metrics. In this section we describe these various metrics traditionally used for measuring image captioning systems. None of these metrics were designed to support the coherence relations of the reference or generated captions. These serve as baselines for COSMic.

N-gram based The most popular image captioning metrics are based on precision and recall of n-grams from generated and reference captions. We compare with Bleu$_1$, Bleu$_2$, Bleu$_3$, Bleu$_4$ (Guo and Hu, 2019), ROUGE$_L$ (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016b). We compute these using their popular open-source implementation[6].

BLEURT We use a pre-trained BLEURT model[7] as a baseline for our work. Unlike N-gram based approaches, BLEURT uses BERT-based word embeddings which are robust to variations in surface word realizations between the reference and generated captions. We do not do any fine-tuning for this baseline.

BERTScore BERTScore[8] uses a pre-trained BERT model to embed the reference and generated captions. Text-level similarity scores are then computed by matching the tokens' output embeddings.

Please note that for both BERT-based baselines above (BLEURT, BERTScore), we use the BERT-Large-512 size model.

5.3 COIN-based Evaluation Setup

We use each baseline metric and COSMic to score the 8 different image captioning systems described in Section 4 on the same set of test images with reference captions. Note that the range and scale of each metric is different, however they are all monotonously increasing functions of model quality. So in our study, we do not analyze the absolute score assigned by these metrics, but only their ranks. We also ask human annotators to rank these 8 captioning systems on the same set of test images. The ranks assigned by a higher performing metric will align better with the ranks from human annotators.

Since the captioning systems above are trained on Conceptual Captions or COCO, we use image/caption pairs from COIN for an out-of-domain evaluation. A subset of 50 random images is used to rank the captioning systems as described above, resulting in 400 machine-generated captions total for the 8 captioning systems. These were then evaluated by human annotators using the process described in Section 3. The human-scored system level performance for each captioning system on this test set is reported in Table 1 in "Average Human Rating".

We measure the alignment between metric-assigned and human-assigned scores using the Kendall (Kendall, 1938) correlation coefficient. In order to calculate the score, we first aggregate all the sample scores and average them. Then we calculate the Kendall tau score using the SciPy 1.7.1 implementation. The score is calculated between two vectors, first of which is the average human ratings for 8 models and the second being the investigated metric scores for 8 models in the following order:[Base$_{Visible}$, Base$_{Meta}$, Base$_{Subjective}$, Base$_{Story}$, Lite$_{Visible}$, Lite$_{Meta}$, Lite$_{Subjective}$, Lite$_{Story}$]. Due to the small sample size, Kendall correlation is the most suitable correlation measure.

A key measure of the success of an automatic evaluation metric is whether it makes the same decision about which system is better in a head-to-head evaluation as we would get from a human-subjects

[5] https://github.com/facebookresearch/vilbert-multi-task

[6] https://github.com/tylin/coco-caption

[7] https://github.com/google-research/bleurt

[8] https://github.com/Tiiiger/bert_score

evaluation. If each system is evaluated based on its average score, then success comes when the average computed metric correlates closely with the average human-ranking. In particular, we measure the alignment between metric assigned and human assigned scores using the Kendall score, following the work of (Sellam et al., 2020).

6 Results

Table 1 presents the results of the COIN-based study. The last row reports the Kendall correlation coefficient between the scores assigned by the metric and humans.

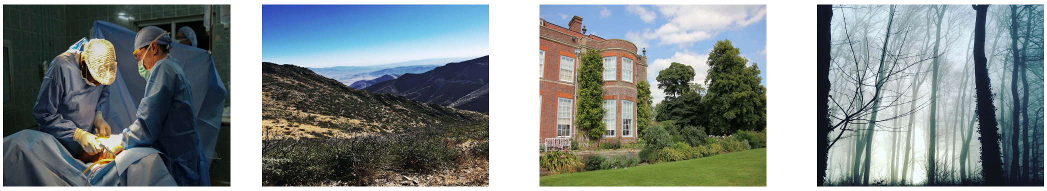
All N-gram based metrics, such as BLEU and CIDEr, fail to adapt to the out-of-domain ground-truth captions from COIN. This results in a relatively flat distribution of system-level scores concentrated close to 0, and hence low correlation coefficients. CIDEr has a highly negative Kendall's τ, which denotes a strong negative association with human judgements. This is partly due to low ($\sim$0.01) and hence noisy CIDEr scores. (Figure 4 provides example cases that illustrate this argument.)

Embedding-based methods, BLEURT and BERTScore, do not suffer from this limitation resulting in more meaningful scoring of systems and hence higher correlation with human scores. However, by design, both these metrics are agnostic to coherence-labels and the input image. COSMic, which is coherence-aware, obtains the highest correlation with human scores. COSMic ViLBERT has the highest Kendall's correlation among all of our models. COSMic Vanilla performs the second best among our models and it performs better than the rest of the models in terms of Kendall's correlation.

Data Augmentation The raw RaCCoon training data has a coherence-level bias as demonstrated by the average COSMic score for each class — *Visible* (0.622), *Meta* (0.459), *Subjective* (0.236) and *Story* (0.397). This reflects the human annotators' bias towards liking *Visible* captions the most, and *Subjective* captions the least, which is expected. However, training COSMic on this data injects the same coherence-bias into the model which is undesirable. As presented in Table 1, both flavors of COSMic (without the '+') assign high scores to *Visible* captioning systems.

To mitigate this issue, we algorithmically augment the training data to bring the average scores for each coherence class to comparable values. We achieve this by pairing images with random captions from the coherence class and assigning them a score of 0. This is a valid training sample because the randomly sampled caption does not describe the said image and serves as a negative sample. With these operations, the class bias is significantly reduced — *Visible* (0.459), *Meta* (0.439), *Subjective* (0.328) and *Story* (0.425). The COSMic columns in Table 1 with '+' denote that this data augmentation approach improves ranking of captioning systems leading to better alignment with human judgements.

Ablation Study Table 2 reports the performance of COSMic Vanilla without coherence-labels and/or the image as model inputs. We find that removal of image features affects COSMic's performance, showing the important contribution of images. The performance deteriorates significantly when the coherence-labels are removed from the model ("No r_c, g_c" column in Table 2). This demonstrates that COSMic successfully integrates coherence-relations in the caption scoring process.



Reference	two men in scrubs performing surgery.	mountains in front of a clear blue sky.	large brick building next to a green lawn and big trees.	a foggy forest.
Generated	surgeons operating on a patient.	mountain range as seen from the trail.	the front of the house.	light shining through the trees.

Figure 4: Illustration of COIN reference captions and corresponding outputs of the Base-Visible model. Though the generated captions are correct, an n-gram based metric such as CIDEr assigns them a very low score due to the variations in surface word realizations. See Table 1 for average scores over the test set. (Photo credits, from left to right: U.S. Army Africa, Gabriel, Fr James Bradley, Rosmarie Voegtli)

System		COSMic			
Model	Coh. Label	Full	No I	No c	No I & c
Base	Visible	.516	.447	.434	.442
	Meta	.505	.439	.442	.453
	Subj.	.356	.347	.438	.453
	Story	.505	.433	.436	.445
Lite	Visible	.515	.444	.434	.433
	Meta	.511	.434	.447	.464
	Subj.	.379	.367	.440	.459
	Story	.499	.440	.433	.442
Kendall's Corr. (τ)		**.667**	.546	-.222	-.415

Table 2: Ablation experiment results. "No I" represents "COSMic Vanilla without image features", "No r_c, g_c" represents "COSMic Vanilla without coherence label embeddings", finally "No I & No r_c, g_c" represents "COSMic Vanilla without coherence label embeddings and without image features".

7 Conclusion

Our work is the first step towards designing generation metrics that respect the information goal of the generated text. We observe that a small set of examples annotated with coherence relations can provide what is needed for learning a discourse-aware generation metric. Our findings have implications for designing context-aware multimodal metrics with criteria that are closer to human ratings for evaluating machine-generated multimodal content.

We have called attention to the challenge of learning robust generation metrics that can evaluate the output of the generation models considering the information goals. Our findings suggest that fine-tuning ViLBERT—originally trained with millions of images—with a smaller sample of coherence relations and expert-annotated scoring, automated metrics can score generated captions closer to a human rating. The presented dataset provides the opportunity for future research in the area of image description generation, designing discourse-aware metrics, and multimodal content evaluation. We hope that coherence-aware text generation metrics could be used for learning better generation models (such as abstractive summarization or story generation) and could be deployed directly in machine learning pipelines to help in optimizing hyper-parameters. Ultimately, it is intended to have a generalizable model that can use a labeling mechanism—not restricted to coherence labels— to improve applicability of generation metrics in different tasks.

8 Ethics

This paper describes a research prototype. We do not work with sensitive or personal data. Our protocol was approved by our ethics board. Human subjects participated voluntarily, undertook minimal risk, and were compensated fairly for their time. The dataset we produced is fully anonymized. Subjects consented to the distribution of their data as part of their participation in the research. Technologists should think carefully before deploying our ideas in production. Our work depends on pretrained models such as word and image embeddings. These models are known to reproduce and even magnify societal bias present in training data. Moreover, like many ML NLP methods, our methods are likely to perform better for content that is better represented in training, leading to further bias against marginalized groups. We can hope that general methods to mitigate harms from ML bias can address these issues.

A distinctive complication of our work is the fact that many image–text presentations involve writers expressing subjective opinions. By its nature, our evaluation metric assesses such subjective texts based on averages and trends across many users, which may be problematic. Although such judgments are ultimately matters of personal taste, they are nevertheless often grounds by which hierarchies of differences are culturally encoded and enforced. Thus, a deployed subjective-caption generation system could well be unfair to users, especially if those users are not confident in their own taste or critical towards the system's responses. Our evaluation metric is not sensitive to such harms.

Acknowledgements

The authors affiliated with Rutgers University were partly supported by NSF Award CCF-19349243. Thanks to Pitt Cyber for supporting this project and the authors from the University of Pittsburgh. We also acknowledge the Center for Research Computing at the University of Pittsburgh for providing the required computational resources for carrying out experiments at the University of Pittsburgh.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey

Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016a. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822.

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven C.H. Hoi, Caiming Xiong, Irwin King, and Michael Lyu. 2020. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2439–2449, Online. Association for Computational Linguistics.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. 2021. AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Lang. Resour. Evaluation*, 55(3):661–688.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse.

Xinyue Huang and Adriana Kovashka. 2016. Inferring visual persuasion via body language, setting, and deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 73–79.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019a. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019b. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and et al. 2020a. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020b. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Scott McCloud. 1993. *Understanding comics: The invisible art*. William Morrow.

Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. 2020. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*.

Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. 2019. Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 168–176. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008b. The Penn discourse treebank 2.0. In *LREC*. Citeseer.

J Pustejovsky and N Krishnaswamy. 2020. Situated meaning in multimodal dialogue: human-robot and human-computer interactions.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Deborah Schiffrin. 1980. Meta-talk: Organizational and evaluative brackets in discourse. *Sociological Inquiry*, 50(3-4):199–236.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

Anthony Viera and Joanne Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family medicine*, 37:360–3.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.