Multimodal Image Paragraph Generation: Utilising Linguistic Information in Generating Diverse Image Descriptions

Anonymous ACL submission

Abstract

abstract

1 Introduction

The quality of automatically generated image captions (Bernardi et al., 2016) has been continuously improving as evaluated by a variety of metrics. These improvements include use of neural networks (Kiros et al., 2014; Vinyals et al., 2014), attention mechanisms (Xu et al., 2015) and more fine-grained image features (Anderson et al., 2017). More recently, a novel open-ended task of image paragraph generation has been proposed by Krause et al. (2017). This task requires the generation of multi-sentence image descriptions, which are highly informative, thus, include descriptions of a large variety of image objects, and attributes, which makes them different from standard single sentence captions. In particular, a good paragraph generation model has to produce descriptive, detailed and coherent text passages, depicting salient parts in an image.

In this paper, we focus on learning to generate more *diverse* image paragraphs. In language and vision literature, "diversity" of image descriptions has been mostly defined in terms of lexical diversity, word choice and *n*-gram based metrics (Devlin et al., 2015; Vijayakumar et al., 2016; Lindh et al., 2018; van Miltenburg et al., 2018). These criteria are focused on generating a diverse set of *independent, one-sentence captions*, with each describing image as a whole. Each of these captions might mention identical objects due to the nature of the task ("describe an image with a single sentence"). Then, diversity is measured in terms of how different object descriptions are from one caption to another (e.g. a man can be described as a 'person' or 'human' in two different captions). However, a good image paragraph model must also introduce diversity on the sentence level, since describing *different scene objects* throughout the paragraph is what makes it more informative than single sentence captions. Here, we define *paragraph diversity* with two essential conditions. First, a generative model must produce a set of sentences with reasonable mentions of a variety of image objects (**sentence-level diversity**). Second, it must demonstrate the ability to use many different words to describe objects without unnecessary repetitions (**word-level diversity**).

To improve on the first point, for each generated sentence we learn to attend to salient objects in the scene., Our primary research question is as follows: does supply image paragraph models with both visual and background (linguistic) information improve **diversity** of generated paragraphs? We expect these types of input to be complementary in generating varied paragraphs. We experiment with several types of inputs to the paragraph generator: visual, language or both. We also investigate the effects of using either attention or max-pooling on image regions as a way of representing an image as a whole. We demonstrate that multimodal input paired with attention on these modalities benefits model's ability to generate more diverse paragraphs. We evaluate the diversity of our paragraphs with both automatic metrics and human judgements.

Additionally, we note that paragraphs must be *accurate* in describing an image. For completeness, we also report results of the automatic evaluation, showing that automatic metrics, which aim to measure *accuracy* of paragraphs rather than *diversity*, do not necessarily pick the most diverse paragraph as the most accurate and vice versa.

2 Related Work

Discourse Structure Producing structured and ordered sets of sentences (e.g. *coherent para-*

graphs) has been a topic of research in NLG community for a long time with both formal theories of coherence (Grosz et al., 1995; Barzilay and Lapata, 2008) and traditional rule-based model implementations (Reiter and Dale, 2000). The coherence of generated text depends on several NLG subtasks: *content determination (selection)*, the task of deciding which parts of the source information should be included in the output description, and *text structuring*, the task of ordering selected information (Gatt and Krahmer, 2017). We believe that the hierarchical structure of our models reflects the nature of these tasks. First, the model attends to the image objects and defines both their salience and order of mention and then it starts to realise them linguistically. [Nikolai: more references about discourse structure are needed] More recently, Ilinykh et al. (2019) collected the dataset of image description sequences (mini-discourses). These mini-discourses are especially useful for learning models which can plan and realise such ordered and structured data.

Neural image paragraph captioning The task of generating image paragraphs has been introduced in Krause et al. (2017) along with the dataset of image-paragraph pairs. The authors hierarchically construct their model: sentence RNN is conditioned on visual features to output sentence topics. Then, each of these topics is used by another RNN to generate actual sentences. Our models are based on this hierarchical model. However, we substantially change its structure by removing the end of paragraph prediction.

Liang et al. (2017) also use the hierarchical network, but also with adversarial discriminator, that forces model to generate realistic paragraphs with smooth transitions between sentences. Chatterjee and Schwing (2018) also address cross-sentence topic consistency by modelling the global coherence vector, conditioned on all sentence topics. Different from these approaches, Melas-Kyriazi et al. (2019) employ self-critical training technique (Rennie et al., 2016) to directly optimise a target evaluation metric for image paragraph generation. Lastly, Wang et al. (2019) use convolutional autoencoder for topic modelling based on region-level image features. They demonstrate that extracted topics are more representative and contain information relevant to sentence generation. In this paper, we similarly model better topic representations. However, we use additional semantic representations of image objects as part of the input to our topic generator. Lin et al. (2015) has proposed a non-neural approach to generate texts describing images. However, this approach is infeasible due to its dependence on multiple components: visual scene parsing, generative grammar for learning from training descriptions, and an algorithm, which analyses scene graphs and extracts semantic trees to learn about dependencies across sentences.

Language representation for image captioning Several proposed models for image captioning are conditioned on both visual and background information. You et al. (2016) detect visual concepts found in the scene (objects, attributes) and extract top-down visual features. Both of these modalities are then fed to the RNN-based caption generator. Attention is applied on detected concepts to inform the generator about how relevant a particular concept is at each timestamp. Different from their model, we do not use any attribute detectors to identify objects in the scene. Instead, our model uses the output of another model pre-trained for the task of dense captioning. Lu et al. (2016) emphasise that image is not always useful in generating some function words ('of', 'the'). They introduce adaptive attention, which determines when to look at the image and when it is more important to use the language model to generate the next word. In their work, the attention vector is the mixture of visual features and visual sentinel, a vector obtained through the additional gate function on decoder memory state. Our model focuses on a similar task: we are interested in deciding which type information is more relevant at a particular timestamp, but we also look at how *merging* two modalities into a single representation performs and how it affects attention of the model. Closest to our work is the work by Liang et al. (2017), who apply attention to region description representation and use it to assist recurrent word generation in producing sentences in a paragraph. Similar to our approach, they also supply their model with embeddings of local phrases used to describe image objects. However, they use textual phrases directly, while we are using hidden representations from the model trained to generate such phrases (Johnson et al., 2016). Besides, here we explore a different route of *where* to use language information: we use phrase representations to define sentence topics to choose from (topic selection) rather than directly guide the generation of words (micro-planning).

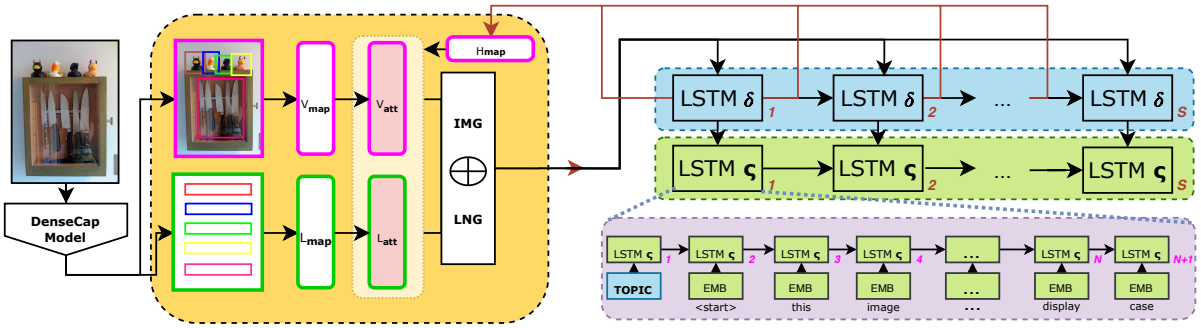

Figure 1: Multimodal paragraph generator architecture. Orange area on the left side is the learned space where two modalities are attended to (vision framed with purple, language framed with green). The attended features are concatenated and used as input to the discourse LSTM (coloured in blue, indicated with δ) to produce sentence topics. Also, the last hidden state of discourse LSTM is used by attention module at each timestamp. Sentence LSTM (coloured in green, indicated with ς) is given the sentence topic and word embeddings. Due to limited space, we omit linear layer and softmax layer which are used to predict the next word from the output of the sentence LSTM.

3 Approach

Overview For our experiments we implement and adapt the hierarchical image paragraph model by (Krause et al., 2017).[1] To prepare input features, we utilise the pre-trained model for dense captioning (Johnson et al., 2016) in two ways. First, we use it to extract convolutional features of identified image regions. We also use its hidden states from the RNN layer as language features. In the original model, these states are used to generate region descriptions; therefore, these vectors represent semantic information about objects. We construct *a multimodal space*, in which we learn mappings from both text and vision features and attend to produced vectors. Lastly, two attended modalities are concatenated to form a multimodal vector, which is used as an input to the paragraph generator. Our paragraph generator consists of two components: discourse-level and sentence-level LSTMs (Hochreiter and Schmidhuber, 1997). First, the discourse-level LSTM learns each sentence topic from the multimodal representation, capturing information flow between sentences. Second, each of the topics is used by sentence-level LSTM to generate an actual sentence. Finally, all generated sentences per image are concatenated to form a final paragraph. An overview of our model and a more detailed description is shown in Fig. 1. Note that different from Krause et al. (2017), we do not learn to predict the end of the paragraph. Instead, we generate the same number of sentences for each image, as we find in its ground-truth paragraph. We leave the task of predicting the number of sentences to generate in a paragraph for future work.

3.1 Input Features

Visual Features We use DenseCap region detector (Johnson et al., 2016)[2] to identify salient image regions and extract their convolutional features. We provide only a brief overview of this procedure: first, a resized image is passed through the VGG-16 network (Simonyan and Zisserman, 2014) to output a feature map of the image. A region proposal network is conditioned on the feature map to identify the set of salient image regions, which are then mapped back onto the feature map to produce corresponding map regions. Each of these map regions is then fed to the two-layer perceptron, which outputs a set of the final region features $\{v_1, ..., v_M\}$, where $v_m \in \mathbb{R}^{1\times D}$ with $M = 50$ and $D = 4096$. This matrix $V \in \mathbb{R}^{M\times D}$ provides us with fine-grained image representation on the object level. We use this representation as features of visual modality.

Language Features In the dense captioning task, a single layer LSTM is conditioned on region features to produce descriptions of these regions in natural language. We propose to utilise its outputs as language features, using them as additional semantic information about detected objects. Specifically, we condition pre-trained LSTM on region features to output a set $Y = \{y_m, ..., y_M\}$ with $y_m \in \mathbb{R}^{1\times T\times H}$, where $T = 15$ and $H = 512$.

[1] The authors have not publicly released the original code of the model.

[2] Available at: https://github.com/jcjohnson/densecap

We normalise each vector over the second dimension T, which determines the maximum number of words in each description. We achieve this by summing all elements across this dimension and dividing the result by the actual length of the corresponding region description, which we generate from Y. The final matrix $L \in \mathbb{R}^{M \times H}$, contains language representations of M detected regions.

Multimodal Features To jointly utilise different modalities (textual and visual), we use methods from multimodal machine translation, which is similar to the task of image captioning in its nature. In particular, we build on work by Caglayan et al. (2016, 2019), who demonstrate that using modality-dependent linear layers in multimodal attention mechanism helps achieve better results as evaluated by automatic metrics for the task of multimodal machine translation. First, we learn two different mappings, using V_{map} for vision and L_{map} for language. These linear projections learn to embed modality-specific information into the attention space. Then, two attention mechanisms are trained on each modality to learn relevant features from each modality. Finally, weighted attended features are concatenated into the multimodal vector f. We have experimented with fusing two attended modalities into a single vector via additional linear layer, but observed no improvement.

Specifically, as formulas in Eq. 1 and Eq. 2 demonstrate, we first attend to mapped modality features at each timestamp t, where $t \in \{1, ..., S\}$ and S is the maximum number of sentences to generate. We set $S = 6$. Last hidden state from discourse LSTM is used at each timestamp to inform the network about previous discourse context. Concatenation, logistic sigmoid function and element-wise multiplication are indicated with $\oplus$, σ and $\odot$ respectively. We also use δ to refer to the discourse LSTM and ς for denoting sentence LSTM.

$$\alpha_t^L = softmax(W_a^L tanh(W_h h_{t-1}^{\delta} + W_m^L L_t)) \quad (1)$$

$$\alpha_t^V = softmax(W_a^V tanh(W_h h_{t-1}^{\delta} + W_m^V V_t)) \quad (2)$$

Finally, as Eq. 3 indicates, we obtain a single multimodal vector $f \in \mathbb{R}^{1 \times H}$, which encapsulates and merges salient information from attended visual and textual modalities:

$$f_t = [\alpha_t^L \oplus \alpha_t^V] \quad (3)$$

3.2 Discourse LSTM

Our discourse-level LSTM is responsible for modelling topics of each of the individual sentences in the paragraph. At each timestamp, it is conditioned on the multimodal feature vector f_t, and its output is a set of hidden states $\{h_1, ..., h_S\}$, where each state is used as an input to the sentence-level LSTM. In its nature, sentence LSTM has to simultaneously complete at least two tasks: produce a topic with relevant information for each sentence, while preserving some type of *ordering* between topics. Such topic ordering is essential for keeping a smooth transition between sentences (discourse items) in the paragraph (mini-discourse). We expect attention on two modalities to assist discourse LSTM in its multiple objectives since attention alleviates the task of weighing specific parts of the input as more important for a particular sentence. This allows discourse LSTM to learn more precise sentence representations and sentence order.

Similar to Xu et al. (2015), we also learn a gating scalar β and apply it to f_t:

$$\beta = \sigma(W_b h_{t-1}^{\delta}), \quad (4)$$

where W_b is a learnable model parameter. Thus, the input to discourse LSTM is computed as follows:

$$f_t^{\delta} = \beta \odot f_t \quad (5)$$

We do not implement doubly stochastic regularisation, since this would force the model to pay equal attention to modality features, eliminating the purpose of learning to attend to the most salient parts of its input.

3.3 Sentence LSTM

Our sentence-level LSTM is a single-layer LSTM tasked to generate all sentences in the paragraph. We run sentence LSTM S times, and use the concatenation of the corresponding hidden state of discourse LSTM with the learned embeddings of the words in the target sentence y_s as its input:

$$x_s^{\varsigma} = [h_s^{\delta} \oplus E y_s] \quad (6)$$

Our word embedding matrix $E \in \mathbb{R}^{K \times H}$ is learned from scratch, K is the vocabulary size. This is different from Krause et al. (2017), who use word embeddings and LSTM weights from the pre-trained DenseCap model. We have also experimented with transferring DenseCap weights

and embeddings into our model but observed no significant improvement.

At each timestamp t, our sentence LSTM is unrolled $N+1$ times (we set $N=50$, which is the number of words to generate), and at each step, its hidden state is used to predict a probability distribution over the words in the vocabulary. The final set of sentences is concatenated together to form a paragraph.

3.4 Learning Objective

We train our model end-to-end with image-paragraph pairs (x, y) from the training data. Our training loss is a simple cross-entropy loss on the sentence level:

$$loss^{S}(x,y) = -\sum_{i=1}^{S}\sum_{j=1}^{M_i} log(p_{j,s}) \quad (7)$$

where $p_{j,s}$ is the softmax probability of the j^{th} word in the i^{th} sentence given all previously generated words for the current sentence $y_{1:j-1,i}$. For our first sentence, hidden states for both LSTMs are initialised with zeros. For every next sentence, both LSTMs use last hidden states generated for the previous sentence from the corresponding layers. During training, we use teacher forcing and feed ground-truth words as target words at each timestamp. We use Adam (Kingma and Ba, 2014) as our optimiser and choose the best model based on the validation loss (early stopping).

4 Experiments

Here we describe six configurations of our model, which we train, validate and test on the released paragraph dataset splits (14,575, 2,487, 2,489 for training, validation and testing respectively). **IMG** model is conditioned only on mapped visual features, while **LNG** model uses mapped semantic information to generate paragraphs. These models do not learn to attend to its input features. Instead, we max-pool input features across M regions, represented by mapping from either language features $x = W_m^L L_t$ or visual features $x = W_m^V V_t$:

$$x_s^S = max_{i=1}^{M}(x) \quad (8)$$

Similarly, in **IMG+LNG** model we apply max-pooling on both modalities and concatenate them into a single vector:

$$x_s^S = [max_{i=1}^{M}(W_m^L L_t) \oplus max_{i=1}^{M}(W_m^V V_t)] \quad (9)$$

All models with **+ATT** use attention on either unimodal or multimodal features.

During decoding, we use beam search (Freitag and Al-Onaizan, 2017) and experiment with forcing the model to generate a minimum number of words C in each sentence. Along with the n-gram penalty, which we leave for future work, we believe that setting C would provide us with more varied and diverse sentences. Similar to Klein et al. (2017), we ensure that each generated sentences consists of at least C words by setting `p(<end>) = -1e20` if it has been chosen by the beam before the sentence minimal length C is achieved. We tested a range of values for beam width $B \in \{2, 4, 6, 8, 10\}$ and several values for $C \in \{7, 8, 9, 10, 11\}$. Based on the CIDEr score, we chose to set $C = 2$ and $C = 9$. Note that C is close to the average sentence length in ground-truth paragraphs (11.91).

5 Evaluation

5.1 Metrics

Typically, a variety of n-gram based automatic metrics is used to measure the correctness/accuracy of image captions. Here, we evaluate our models across several metrics: CIDEr (Vedantam et al., 2014), METEOR (Denkowski and Lavie, 2014), BLEU-{1, 2, 3, 4} (Papineni et al., 2002), and Word Mover's Distance (Kusner et al., 2015; Kilickaya et al., 2017). To measure diversity, we report self-BLEU (Zhu et al., 2018), which sometimes referred to as mBleu (Shetty et al., 2017). This metric evaluates how much one sentence resembles another by calculating BLEU score between sentences. We calculate self-BLEU as follows: we split each generated paragraph into sentences and use one sentence as a hypothesis, and the other sentences are regarded as references. A lower score indicates more diversity, e.g. less *n*-gram matches between compared sentences. We also calculate the diversity metric introduced by Wang and Chan (2019). This metric applies Latent Semantic Analysis (Deerwester et al., 1990) to the weighted n-gram feature representations (CIDEr values between unique pairs of sentences) and identifies the number of topics among sentences. Compared to self-BLEU, which measures n-gram overlap, LSA and CIDEr-based diversity metric measures semantic differences between captions as well. More identified topics in paragraph sentences indicate higher level of diversity.

Model	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG	7.48	25.66	11.20	24.51	13.67	7.96	4.51
LNG	7.19	22.27	10.81	23.20	12.69	7.34	4.19
IMG+LNG	7.61	26.38	11.30	25.10	13.88	8.11	4.61

(a) Scores of automatic evaluation metrics for **models without attention / with max-pooling**. Best scores are coloured in green.

Model	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG+ATT	7.56	26.72	11.40	25.56	14.28	8.36	4.85
LNG+ATT	7.34	25.11	10.97	24.28	13.36	7.67	4.35
IMG+LNG+ATT	7.36	24.96	11.02	24.30	13.44	7.80	4.46

(b) Scores of automatic evaluation metrics for **models with attention**. Best scores are coloured in red.

Table 1: Automatic evaluation of generated paragraphs.

Model	mBLEU	self-CIDEr
IMG	0.50630	0.76438
LNG	0.52245	0.75591
IMG+LNG	0.52090	0.76465
IMG+ATT	0.51501	0.76423
LNG+ATT	0.50387	0.76890
IMG+LNG+ATT	0.52247	0.75131
GT	0.18847	0.96514

Table 2: Scores for diversity measures for different paragraph models. mBLEU stands for the average score between all self-BLEU scores for n-grams (1, 2, 3, 4). Self-CIDEr stands for average score of the LSA-based diversity metric. GT (ground truth) stands for corresponding diversity metric scores for the test set. Higher intensity of blue colour indicates better scores.

5.2 Results

As results in Table 1a demonstrate, we gain on all n-gram-based metrics when the model utilises semantic information along with the visual representations (**IMG+LNG**) compare to models, which use single modality (**IMG** or **LNG**). Also, the distance between ground-truth and generated paragraphs in word embedding space increases when we use both modalities, as indicated by WMD score. In the context of generating *diverse* paragraphs, we see it as an improvement: **IMG+LNG** model does not only generate more accurate paragraphs, but also learns to slightly deviate its output's content from the content in ground-truth descriptions without hurting its accuracy.

However, as Table 1b indicates, when replacing max-pooling with attention for modality processing, we observe that **IMG+ATT** performs the best, leaving multimodal architecture (**IMG+LNG+ATT** slightly behind. This indicates that using attention might actually hurt the accuracy when using both visual and language modalities. On the contrary, **IMG+ATT** and **LNG+ATT** seem to benefit from attention, improving on the scores compared to their versions that use max-pooling. We hypothesise that since attention looks at *muilti-ple* region representations, it is prone to choose regions, which might be redundant between two modalities (learning to attend to the same object representation in both spaces, for example). Max-pooling, however, is less likely to do so, since it picks *a single* most important region representation, therefore, reducing a chance that the max-pooled visual and language representations overlap with each other. As for **IMG**-based and **LNG**-based models, they seem to benefit from attention since each of them uses a single modality, making it impossible to output redundant information from two different modalities. In unimodal scenario, attention produces more informative representations, compared to max-pooling, and in the multimodal scenario, attention misinforms the sentence LSTM, likely choosing non-complementary representations from two modalities. This might occur due to the nature of the input features as well: both visual and language inputs represent the same objects.

Table 2 contains scores of our diversity metrics. Best mBLEU scores are achieved by models, which use a single modality. Both multimodal architectures (**IMG+LNG** and **IMG+LNG+ATT**) seem to be one of the least diverse in terms of n-gram overlap between sentences in generated paragraphs. However, mBLEU under-represents the diversity, being unable to take into account semantic differences between sentences. As results for LSA-based metric (self-CIDEr) indicate,

IMG+LNG model outperforms its unimodal versions, but performs worse than **LNG+ATT**. However, **LNG+ATT** model scores highest in diversity, but worse in accuracy among all attention-based models (Table 1b). Furthermore, both LNG-based models are the worst ones in terms of accuracy.

As previous work shows (Caccia et al., 2018; Holtzman et al., 2019), there is a trade-off between *quality* (accuracy) and *diversity* when generating natural language expressions. Here, we conclude that when using max-pooling, multimodal input benefits both accuracy and diversity of the generated outputs. On the contrary, attention seems to worsen both accuracy and diversity of these models, improving on accuracy and diversity metrics for unimodal counterparts. However, models which utilise single modality, benefit from attention in both accuracy and diversity.

5.3 Human Evaluation

To collect human judgements about paragraphs, we employed our experiment on Amazon Mechanical Turk. We randomly chose 10% of the images from our test set resulting in 250 images. For each of these images we gathered seven paragraphs (six from the models and one from the test set). Each of these paragraphs has been evaluated by the crowdsourcing workers across multiple criteria: word choice, object salience, sentence structure and paragraph coherence. We presented workers with the instructions shown in Appendix A. To ensure quality and variety of workers' judgements, we presented our tasks only to the Master workers (workers with the high reputation and task acceptance rate) and controlled for the number of tasks a single worker is able to submit. We paid 0.15$ per task to a single worker. Finally, we obtained judgements from X unique Master workers for 1750 image paragraphs overall. For each judgement criteria we took the average score across all models; the results are shown in Table 3.

Model	mBLEU	self-CIDEr
IMG	0.50630	0.76438
LNG	0.52245	0.75591
IMG+LNG	0.52090	0.76465
IMG+ATT	0.51501	0.76423
LNG+ATT	0.50387	0.76890
IMG+LNG+ATT	0.52247	0.75131
GT	0.18847	0.96514

Table 3: Average scores for 4 param

6 Conclusion

Future work: diversity in terms of choosing best candidate from the pool of candidates. Experiments with choosing different candidates (not always the most probable from the beam search) and report how diversity changes. Experiment with other language representations (more complex ones). End of paragraph prediction.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2018. Language GANs Falling Short.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal Attention for Neural Machine Translation.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation.

Moitreya Chatterjee and Alexander G. Schwing. 2018. Diverse and coherent paragraph generation from images. In *ECCV*.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

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, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language Models for Image Captioning: The Quirks and What Works.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*.

Albert Gatt and Emiel Krahmer. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P. Xing. 2017. Recurrent topic-transition gan for visual paragraph generation.

Dahua Lin, Chen Kong, Sanja Fidler, and Raquel Urtasun. 2015. Generating Multi-Sentence Lingual Descriptions of Indoor Scenes.

Annika Lindh, Robert J. Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D. Kelleher. 2018. Generating Diverse and Meaningful Captions.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.

Luke Melas-Kyriazi, Alexander Rush, and George Han. 2019. Training for Diversity in Image Paragraph Captioning.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the Diversity of Automatic Image Descriptions. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. Building natural language generation systems.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical Sequence Training for Image Captioning.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator.

Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation.

Qingzhong Wang and Antoni B. Chan. 2019. Describing like humans: on diversity in image captioning.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models.

A Human Evaluation: AMT Instructions

Short Summary: You are going to be shown an image and several sentences describing the image. Below you will see statements that relate to the image descriptions. Please rate each of these statements by moving the slider along the scale where 0% stands for 'I do not agree', 100% stands for 'I fully agree'.

Detailed Instructions: In general, you are required to judge image descriptions based on the following:

- choice of words: does the text correctly describe objects and events in the scene and with the right detail?
- relevance: does the text describe relevant objects and events in the scene?
- sentence structure: do the sentences have a good and grammatical structure?
- coherence: does the test progresses in a natural way forming a narrative?

You can enter any feedback you have for us, for example if some questions were not easy to answer, in the corresponding feedback field (right after the survey).

DESCRIPTION: there are two cows standing in the field. there are trees behind them.

How well do you agree with the following statements?

1. The description contains words that correctly refer to the objects and events in the image
2. The description is referring to the relevant/important parts of the image.
3. The sentences have a correct structure and are grammatical.
4. The sentences are well-connected and form a single story.

Write you feedback in the field below if you have any (not necessary).