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

Anonymous ACL submission

Abstract

abstract

1 Introduction

Lately, there has been a lot of improvement in the task of image captioning (Bernardi et al., 2016) with the 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 generation of multi-sentence image descriptions, which are highly informative, thus, include descriptions of a large variety of image objects, attributes, etc., 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 automatically 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 diverse set of *independent, one-sentence captions*, with each describing image as a whole. These captions are very likely to mention identical objects due to the nature of the task ("describe image with a single sentence"), and diversity is measured in terms of how different object descriptions are from one caption to another (e.g. 'man' can be described as a 'person', 'human'). 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 requirements: a generative model must (a) produce the set of sentences with reasonable mentions of a variety of image objects (sentence-level diversity), (ii) demonstrate the ability to use many different words to describe objects without unnecessary repetitions (word-level diversity).

Our primary research question is as follows: does supplying 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 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 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 automatic evaluation, showing that automatic metrics, which are targeted towards measuring *accuracy* of paragraphs rather than *diversity*, do not necessarily pick the most diverse paragraph as the most accurate.

2 Related Work

Human visual attention Humans are quite efficient in detecting objects and separating them from the rest of the visual scene (Ullman, 1987). We are also fluent in using visual attention, which allows us to single out particular objects in the visual scene, significantly reducing our perceptual load when needed (Lavie et al., 2004), therefore,

preventing us from being overwhelmed by typically complex real-world visual scenes. Our ability to attend to particular parts of the environment is based on both bottom-up information (low-level visual stimuli) and top-down information (high-level goal-related stimuli, discourse knowledge) (Zarcone et al., 2016). Stimuli that attracts our attention is said to be *salient* (relevant). Salience of objects affects our *surprisal* towards particular visual input: discourse-salient entities cause less surprisal (e.g. 'bed' in bedroom) unlike the visually salient objects (e.g. 'large pink elephant' in bedroom).

Attention has also been employed in formal theories of interaction. One of the approaches has been proposed by Dobnik and Kelleher (2016), who link attention with judgements as defined in Type Theory of Records (Cooper, 2008). They introduce a Bayesian-based framework in which attention controls the extent to which context induced judgements ($\sim$ task-based top-down information) are utilized by an agent. This allows for topic modelling at each timestamp in interaction. Thus, it follows along the lines of our proposal about attention using both contextual and low-level visual information in selecting relevant information for each individual sentence in the image paragraph.

Neural image paragraph captioning The task of generating more complex descriptions of images such as paragraphs has been introduced in Krause et al. (2017) along with the dataset of image-paragraph pairs. The paper adopts a hierarchical structure for the model of paragraph generation: sentence RNN is conditioned on visual features and unrolled for each sentence in the paragraph, giving a sentence topic as its output. Then, each of these topics is used by another RNN to generate actual sentences. We start by implementing this hierarchical image paragraph model, since it inherits the modular nature of human image paragraph production (given an image, plan structure of your paragraph and identify its sentence topics, then incrementally generate sentences). Liang et al. (2017) use similar hierarchical network in addition 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 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 optimize a target evaluation metric for image paragraph generation. Lastly, Wang et al. (2019) use convolutional auto-encoder for topic modelling based on region-level image features. They demonstrate that extracted topics are more representative and contain information relevant for sentence generation. In this paper we similarly model better topic representations. However, we use additional language representations as part of the input to our topic generator, which is an LSTM.

Language attention in language and vision models [Nikolai: wordy section, needs to be shorter, keep all information here for now]

Only a limited number of models for image captioning has been supplied with both visual and background information for caption generation. You et al. (2016) detect visual concepts found in the scene (objects, attributes, etc.) 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 generator about how relevant a particular concept is at each timestamp. Different to their model, we do not use any attribute detectors to identify objects in the scene, instead relying on the output of the model pre-trained for the task of dense captioning. Lu et al. (2016a) emphasize that image is not always useful in generating some function words ('of', 'the', etc.). 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 is focused on a similar task: we are interested in deciding which type information is more important 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 language attention on region captions and use it to assist recurrent word generation in producing sentences in a paragraph. They embed region descriptions into the same embedding space that their word RNN is operating on. While we also believe that feeding information about semantic concepts found in an image is beneficial for the model, we propose to employ transfer learning. We use hidden states of

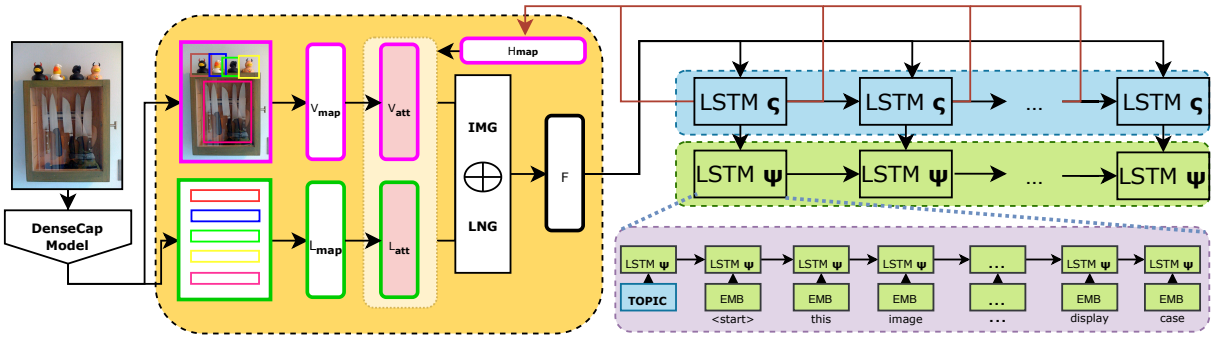

Figure 1: Multimodal paragraph generator architecture. Orange block on the left side is the learned space where two modalities are attended (vision framed with purple, language framed with green). The attended features are concatenated and passed to the linear layer for fusion. The output of the fusion layer F is used by sentence-level LSTM (coloured in blue, indicated with ς) to produce sentence topics. Also, last hidden state of sentence LSTM is used by attention module at each timestamp. Word-level LSTM (coloured in green, indicated with ω) 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.

the RNN trained for the task of dense captioning (Johnson et al., 2016) as our background information representation. Outside of image paragraph captioning, Lu et al. (2016b) have proposed a joint image and question attention model for the task of visual question answering. [Nikolai: Any work on language attention in visual dialog? I think one sentence with some citations would be nice to have.]

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 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 fused 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 more detailed description are shown in Fig. 1. Note that different from Krause et al. (2017), we do not learn to predict 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.

[1] The original code of the model has not been publicly released by the authors.

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

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 $H = 512$. 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 fuse 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 important features from each modality. Lastly, weighted attended features are concatenated and passed to another linear layer, which learns to integrate and fuse information into the multimodal vector.

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 additionally inform 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.

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

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, thus allowing 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. Thus, the input to discourse LSTM is computed as follows:

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

where W_b is a learnable model parameter. 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 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:

$$x_s^\varsigma = [h_s^\delta \oplus Ey_s] \quad (5)$$

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 M times, 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. During decoding, we use beam search (Freitag and Al-Onaizan, 2017) with beam width $B = 10$ to generate each sentence word by word. We empirically found this beam width to produce better automatic evaluation scores. 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:

$$loss^{s}(x, y) = -\sum_{i=1}^{S}\sum_{j=1}^{M_i} log(p_{j,s}) \quad (6)$$

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 visual features, while **LNG** model uses 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 $x = W_m^V V_t$:

$$x_s^{s} = max_{i=1}^{M}(x) \quad (7)$$

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

$$x_s^{s} = [max_{i=1}^{M}(W_m^L L_t) \oplus max_{i=1}^{M}(W_m^V V_t)] \quad (8)$$

All models with **+ATT** use attention on either unimodal or multimodal features. Also, we experiment with controlling for minimal length in generated sentence during beam search. Large beam size is known to favour shorter sentences to the longer ones (Yang et al., 2018). To avoid this problem, similar to Klein et al. (2017), we ensure that each generated sentences consists of at least C words. We empirically find that setting $C = 10$ provides us with better automatic metric scores. Note that this number is close to the average sentence length in ground-truth paragraphs (11.91).

5 Evaluation

5.1 Metrics

Accuracy 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 six different metrics: CIDEr (Vedantam et al., 2014), METEOR (Denkowski and Lavie, 2014), and BLEU-{1, 2, 3, 4} (Papineni et al., 2002). [Nikolai: add a note about WMD and calculate WMD]

Diversity Majority of the diversity metrics for image descriptions are based on calculating *n*-gram statistics as well as comparing vocabulary size, POS usage etc. Here, we report vocabulary size of our models and also the number of unique nouns which are generated as an approximation of how many unique objects are mentioned in the paragraphs. We use spaCy (Honnibal and Johnson, 2015) to identify nouns in generated paragraphs. We also calculate Self-BLEU (Zhu et al., 2018) or sometimes referred to as mBleu (Shetty et al., 2017). This metric has been specifically proposed to assess the similarity between two sentences, and it can be used to measure how much one sentence resembles another. This is in particular important for the paragraphs, in which sentences should not

Model	CIDEr		METEOR		BLEU-1		BLEU-2		BLEU-3		BLEU-4	
	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-
Krause et al. (2017)	-	13.52	-	15.95	-	41.90	-	24.11	-	14.23	-	8.69
IMG	20.36	13.83	13.75	11.28	37.87	25.08	21.10	13.92	12.23	8.14	7.09	4.63
IMG+ATT	**21.07**	14.52	13.74	11.63	37.53	26.72	20.82	15.01	12.13	8.87	7.12	5.18
LNG	19.86	12.13	13.65	10.79	37.39	23.55	20.74	12.80	11.89	7.41	6.92	4.24
LNG+ATT	20.30	11.37	13.76	11.00	37.68	24.48	20.76	13.32	11.85	7.65	6.82	4.31
IMG+LNG	20.75	13.08	**13.85**	11.15	37.97	24.54	21.09	13.70	12.22	8.05	**7.18**	4.69
IMG+LNG+ATT	20.99	13.83	13.81	11.21	**38.18**	25.01	**21.31**	13.86	**12.25**	8.08	7.12	4.67

Table 1: Scores for automatic evaluations metrics computed for the test set. C^+ indicates control for the minimum number of words in generated sentences, C^- similarly indicates the opposite. Scores from the original hierarchical model are reported for completeness (beam search with C^-).

Model	Voc Size		# of NT		SB-1		SB-2		SB-3		SB-4	
	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-	C^+	C^-
IMG	378	430	284	313	80.33	70.62	71.75	58.99	64.84	50.85	58.71	44.11
IMG+ATT	296	435	295	315	79.72	71.07	71.05	59.69	64.05	51.53	57.92	44.63
LNG	399	431	294	310	77.58	68.69	68.26	57.24	60.74	49.41	54.45	42.79
LNG+ATT	269	430	295	308	77.85	69.78	68.44	58.35	61.23	50.42	55.31	43.85
IMG+LNG	283	420	295	299	79.22	71.10	70.56	59.85	63.65	51.79	57.72	45.05
IMG+LNG+ATT	413	**452**	300	**326**	77.81	**68.25**	68.19	**55.97**	61.06	**47.61**	55.23	**40.98**
GT	-	**5835**	-	**3865**	-	**48.66**	-	**27.95**	-	**15.82**	-	**8.70**

Table 2: Measures of diversity for different paragraph models. NT and GT stand for noun types and ground-truth paragraphs from the test set respectively. SB stands for self-BLEU and corresponding n-gram (1, 2, 3, 4).

be very similar with each other. Higher self-Bleu indicates less diversity, e.g., more n-gram matches between sentences. We calculate Self-BLEU as follows: we split each generated paragraph into sentences and use one sentence as hypothesis and the others are regarded as references. Then, BLEU score is calculated for every sentence in every paragraph, and the average BLEU score over the paragraphs is used as the Self-BLEU score of the whole set.

5.2 Results

Accuracy As Table 1 demonstrates, all our models which do not control for C show similar performance and slightly underperform the model by Krause et al. (2017) in all metrics except CIDEr. We believe that there are several reasons for such behaviour. First, our model slightly differs from the original hierarchical model: we do not use any pre-trained weights and do not learn to predict end of the paragraph. Also, we utilise attention in LSTM-based text generator, which, as has been shown by Liang et al. (2017) and Wang et al. (2019), significantly boosts performance in terms of CIDEr score.

The visible trend is that models which are forced to generate at least C words in a sentence perform much better, improving all accuracy-related scores. However, it is unclear whether this will have a positive affect on diversity measures as well: automatic evaluation metrics might be biased towards favouring longer sentences even if they are repetitive. [Nikolai: ref!]

We also conclude that adding modalities (IMG+LNG) somewhat improves scores by a small margin compared to using one of the two (either IMG or LNG). Also, attention seems to boost performance of either unimodal or multimodal systems. We note that image description systems might produce an accurate description, which does not necessarily correspond to the ground truth paragraph from the dataset, since there are multiple correct ways of describing an image. This would, in turn, affect the scores of automatic metrics, which compare generated descriptions with the ground truth ones, and, therefore, may not be good measures of accuracy.

Interestingly, LNG-based model performs slightly worse across almost all metrics. We believe that the quality of linguistic representations affects automatic scores, indicating that more complex linguistic representations are needed. Currently, language input is constructed from representations used to describe only specific regions in isolation. In the paragraph, however, these regions must be linked to each other to form a coherent whole. We leave experiments with such linguistics representations for the future work.

Diversity Note that our goal is to show that purely visually dependent models are less diverse than models which incorporate additional sources of information such as semantic information about objects or use attention. As Table 2 demonstrates, the model which uses both modalities and attention generates most diverse paragraphs compared to outputs of the other models. It generates more unique nouns, has bigger vocabulary size and better self-BLEU scores. However, it still performs much worse than the human-generated paragraphs. Based on the self-BLEU scores for models which are forced to generate at least C words, we conclude that these models tend to generate the same n-grams more often. Therefore, these models suffer from repetitiveness on word level, and do not conform with our criteria for paragraph diversity.

5.3 Diversity in decoding

diversity in terms of choosing best candidate from the pool of candidates. Experiment with choosing different candidates (not always the most probable) and report how diversity changes

5.4 Human Evaluation

Human evaluation on

6 Conclusion

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.

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.

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*.

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.

Robin Cooper. 2008. Type theory with records and unification-based grammar. In *Logics for Linguistic Structures, pages 9 – 34. Mouton de Gruyter*.

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.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records.

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.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words.

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

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. 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*.

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)*.

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.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of experimental psychology. General*, 133(3):339—354.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

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. 2016a. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016b. Hierarchical Question-Image Co-Attention for Visual Question Answering.

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.

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.

S. Ullman. 1987. Visual routines. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 298–328. Kaufmann, Los Altos, CA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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.

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.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the Beam Search Curse: A Study of (Re-)Scoring Methods and Stopping Criteria for Neural Machine Translation.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Alessandra Zarcone, Marten van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing.

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 Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

LaTeX-specific details: Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudocode or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.