When image tells a story: the role of visual and semantic information for generating paragraph descriptions

Anonymous ACL submission

Abstract

Generating multi-sentence image descriptions is a challenging task, which requires a good model to produce coherent and accurate paragraphs, describing salient objects in the image. In this paper, we argue that multiple sources of information are beneficial when depicting visual scenes with long sequences: (i) perceptual information, (ii) semantic (language) information about how to describe what is in the image. We also compare the effects of using two different pooling mechanisms on either a single modality or their combination. As we demonstrate here, the model which utilises both visual and language inputs can be used to generate *accurate* and *diverse* paragraphs when combined with particular pooling mechanism. The results of our automatic and human evaluation show that learning to embed semantic information along with visual stimuli into the paragraph generation model to achieve better results is not that trivial, raising a variety of proposals for future experiments.

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; Lu et al., 2016) 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.

When humans describe images, especially over longer discourses, they take take into account (at least) two sources of information that interact with each other: (i) perceptual information as expressed by visual features and (ii) cognitive reasoning that determines the communicative intent of the text and the use of language (Kelleher and Dobnik, 2019). Perceptual information mainly determines *what* to refer to while the reasoning mechanisms tell us *how* and *when* to refer to it. Both mechanisms interact: that a particular object is described at a particular point of discourse and with particular words depends not only on its perceptual salience but also whether that object should be referred to at that point of the story that the text is narrating which is its discourse salience. Compare for example: "two cows are standing in the field", "there are trees in the field" and "a few of them are close to the trees". The selection and the order of the relevant features are described by a cognitive mechanism of attention and memory (Lavie et al., 2004; Dobnik and Kelleher, 2016).

In this paper, we investigate the interplay between visual and textual information (reflecting background knowledge about the world and communicative intent) and their ability to generate natural linguistic discourses spanning over several sentences. Our primary research question is as follows: does using both visual and linguistic information improve *accuracy* and *diversity* of generated paragraphs? We experiment with several types of inputs to the paragraph generator: visual, language or both. We also investigate the effects of different kinds of information fusion between visual and textual information using either attention or max-pooling. We demonstrate that multimodal input paired with attention on these modalities benefits model's ability to generate more diverse and ac-

Figure 1: Example image with generated paragraphs from our models (incl. ground truth description):
HUMAN: *There are several cars parked along a street. There are many trees in a field in front of the street. There are small blue parking meters on the sidewalk next to the street.*
IMG+MAX : *There are several cars parked on the road. There are cars parked on the street. There are trees behind the street.*
LNG+MAX : *There are several cars on the street. There are trees on the street. There are trees on the street.*
IMG+LNG+MAX : *There are several cars on the street. There are two cars on the street. There are cars parked on the sidewalk.*
IMG+ATT : *There are several cars parked on the street. There are two cars parked on the road. There are two cars parked on the road.*
LNG+ATT : *There are several signs on the street. There are signs on the street. The pole is white.*
IMG+LNG+ATT : *There is a parking meter on a sidewalk. There are cars next to the street. There is a parking lot next to the street.*

curate paragraphs. Fig. 1 shows an image with generated texts from both human and our models.

We evaluate the accuracy and diversity of our paragraphs with both automatic metrics and human judgements. We also argue that, as some previous work shows (van der Lee et al., 2019), *n*-gram-based metrics might be unreliable for quality evaluation of generated texts. The generated paragraph can be accurate as of the image, but because it does not match the ground truth, this would score low based on the automatic evaluation. To provide a different view on paragraph evaluation, we asked humans to judge the subset of generated paragraphs across several criteria, more specifically described in Section 3.4 and Appendix A.

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). In this work, the focus is on generating a diverse set of *independent, one-sentence captions*, with each describing image as a whole. Each of these captions might refer to 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, as argued above, a good image paragraph model must also introduce diversity at the sentence level, describing *different scene objects* throughout the paragraph. Here, we define *paragraph diversity* with two essential conditions. First, a generative model must demonstrate the ability to use relevant words to describe objects without unnecessary repetitions (*word-level diversity*). Secondly, it must produce a set of sentences with relevant mentions of a variety of image objects in an appropriate order (*sentence-level diversity*).

Producing structured and ordered sets of sentences (e.g. *coherent paragraphs*) 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; Deemter, 2016). The coherence of generated text depends on several NLG sub-tasks: *content determination (selection)*, the task of deciding which parts of the source information should be included in the output description, and *text structuring (micro-planning)*, 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, first as paragraph visual-textual topics and then as individual sentences within paragraphs.

2 Approach

Overview For our experiments we implement and adapt the hierarchical image paragraph model by Krause et al. (2017).[1] To prepare input features,

[1]The authors have not publicly released the code of their model and hence the model implementation is based on our

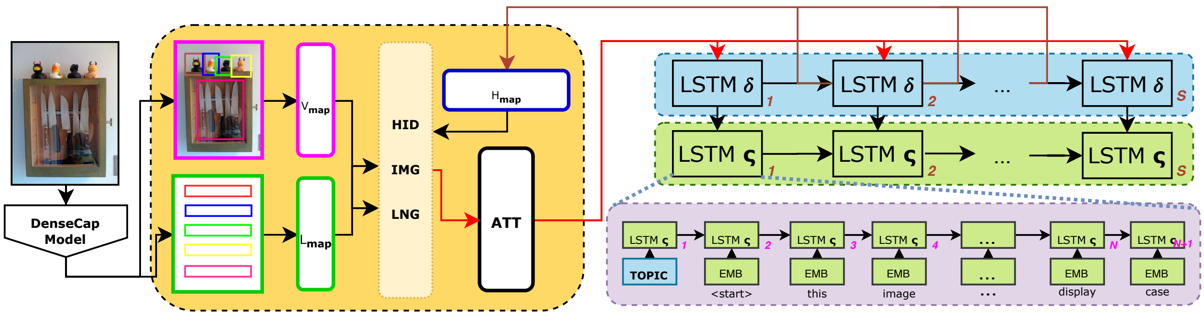


Figure 2: Multimodal paragraph generator architecture. The orange area on the left is the learned space where two modalities are attended to (vision in purple, language in green). The mapped features are concatenated together and passed to the attention mechanism, that outputs a vector which is used as an input to the discourse LSTM (in blue, marked with δ). The attention module also uses the last hidden state of the discourse LSTM at each timestamp. The sentence LSTM (in green, marked with ς) is given the sentence topic and word embeddings. Due to limited space, we omit the linear layer and the softmax layer which are used to predict the next word from the output of the sentence LSTM.

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 multi-modal space*, in which we learn mappings from both text and vision features. Lastly, we concatenate both modalities and attend to them to form a multi-modal 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 the topic of each sentence from the multi-modal 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. 2. Our model is different from the model by Krause et al. (2017) in the following ways: (i) we use either max-pooling or attention in our models, (ii) we do not learn to predict the end of the paragraph, but generate the same number of sentences as we find in ground-truth paragraph per each image, (iii) we use semantic information about objects in the visual scene. The focus of our work is not to improve on the results of Krause et al. (2017) but to investigate the effects of different multi-modal fusion on the accuracy of the diversity of paragraph descriptions.

2.1 Input Features

Visual Features We use DenseCap region detector (Johnson et al., 2016)[2] to identify salient image regions and extract their convolutional features. 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 $\{v_1, \ldots, v_M\}$, 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 at 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 background information about detected objects. Specifically, we condition a pre-trained LSTM on region features to output a set $Y = \{y_m, \ldots, y_M\}$ with $y_m \in \mathbb{R}^{1\times T\times H}$, where $T = 15$ and $H = 512$. We condense each vector over the second dimension T, which determines the maximum number of words in each description. We achieve this by

interpretation of their paper.

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

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 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, we concatenate these mappings to form the multimodal vector f, which is then concatenated with the mapping from the hidden state. We have experimented with fusing two attended modalities into a single vector via an additional linear layer but observed no improvement. We also tried to use modality-dependent attention (*early attention*) as such setting has shown to produce good joint representation for the task of multimodal machine translation (Caglayan et al., 2016, 2019), which is very similar to image captioning in its nature. However, this set-up provided us with worse scores of automatic metrics. Therefore, here we use *late attention*: attending to the visual and textual features when they are already concatenated.

As shown in Eq. 1, at each timestamp t we concatenate mapped features from both modalities to output the multimodal vector $mult_t$, where $t \in \{1,...,S\}$ and S is the maximum number of sentences to generate. We use δ to refer to the discourse LSTM and ς when referring to the sentence LSTM. Concatenation, the logistic sigmoid function and element-wise multiplication are indicated with $\oplus$, σ and $\odot$ respectively. We set S depending on the number of sentences in the ground-truth paragraph with the maximum $S = 6$. Then, as Eq. 2 indicates, we generate attention weights for our multimodal vector $mult_t$. We use additive (concat) attention mechanism and concatenate multimodal representation with the previous hidden state of the discourse LSTM. Finally, as in Eq. 3, we obtain a weighted multimodal vector $f \in \mathbb{R}^{1 \times H}$, which encapsulates and merges salient information from attended visual and textual modalities.

$$mult_t = [W_m^V V_t \oplus W_m^L L_t] \tag{1}$$

$$\alpha_t^{mult} = softmax(W_a^L tanh(mult_t \oplus W_h h_{t-1}^{\delta}) \tag{2}$$

$$f_t = [\alpha_t^{mult} \odot mult_t] \tag{3}$$

2.2 Discourse LSTM

Our discourse-level LSTM is responsible for modelling multi-modal topics of each of the individual sentences in the paragraph. At each timestamp, it is conditioned on the weighted multimodal 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, the discourse LSTM has to simultaneously complete at least two tasks: produce a topic with a relevant combination of visual and linguistic information for each sentence, while preserving some type of *ordering* between the topics. Such topic ordering is essential for keeping a natural transition between sentences (discourse items) in the paragraph (discourse). We expect attention on the combination of two modalities to assist the discourse LSTM in its multiple objectives since attention weights specific parts of the input as more relevant for a particular sentence. We expect that this allows discourse LSTM to learn better sentence representations and sentence order.

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

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

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

$$f_t^{\delta} = \beta \odot f_t \tag{5}$$

2.3 Sentence LSTM

Our sentence-level LSTM is a single-layer LSTM that generates individual sentences in the paragraph. We run the sentence LSTM S times. Each time we use a concatenation of the corresponding hidden state of the discourse LSTM with the learned embeddings of the words in the target sentence y_s as its input:

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

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, where N is the number of words to generate. At each step, its hidden state is used to predict a probability distribution over the words in the vocabulary. We set $N = 50$. The final set of sentences is concatenated together to form a paragraph.

2.4 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:

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

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 the first sentence, the hidden states of both LSTMs are initialised with zeros. For every subsequent sentence, both LSTMs use the last hidden states generated for the previous sentence for each respective layer. 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 an optimiser and choose the best model based on the validation loss (early stopping). For decoding we use beam search (Freitag and Al-Onaizan, 2017) with beam width $B = 2$ (we tested several values for the beam width $B \in \{2, 4, 6, 8, 10\}$).

3 Experiments and Evaluation

3.1 Models

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) (Krause et al., 2017). The **IMG** model is conditioned only on the mapped visual features, while the **LNG** model only uses the mapped semantic information to generate paragraphs. The **IMG+NLG** is conditioned on both mapped visual and semantic information. All models with **+ATT** use late attention on either uni-modal or multi-modal features. We also test another configuration of the models with max-pooling of input features across M regions, represented by mapping from either language features $x = W_m^L L_t$ or visual features $x = W_m^V V_t$:

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

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

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

3.2 Metrics

Typically, a variety of n-gram based automatic metrics is used to measure the correctness/accuracy of image captions. We evaluate our models with the following 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 lexical diversity, we report self-BLEU (Zhu et al., 2018) which is sometimes referred to as mBleu (Shetty et al., 2017). This metric evaluates how much one sentence resembles another by calculating the 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 as references. A lower score indicates more diversity, e.g. fewer n-gram matches between compared sentences. We also calculate the diversity metric introduced in (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 combined with CIDEr-based kernel metric measures semantic differences between sentences as well. More identified topics in paragraph sentences indicate a higher level of diversity. However, this intrinsic metric does not evaluate if the paragraph demonstrates discourse coherence in terms of how these topics are introduced and the quality of the generated sentences and their sequences (Section 1).

3.3 Results

As the results in Table 1 demonstrate, models which utilise both semantic and visual information (any **IMG+LNG** configuration) outperform their single modality variants in both attention and max-pooling settings. When using max-pooling, **IMG+LNG** model improves on CIDEr by 0.72 and METEOR by 0.10. Also, two-modal architecture is slightly lexically more diverse from the ground

Model Input	Type	WMD	CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
IMG	+MAX	7.48	25.66	11.20	24.51	13.67	7.96	4.51
LNG	+MAX	7.19	22.27	10.81	23.20	12.69	7.34	4.19
IMG+LNG	+MAX	**7.61**	**26.38**	**11.30**	**25.10**	**13.88**	**8.11**	**4.61**
IMG	+ATT	7.47	26.01	11.26	24.88	**13.99**	**8.13**	**4.67**
LNG	+ATT	7.20	22.11	10.82	23.20	12.55	7.16	3.97
IMG+LNG	+ATT	**7.54**	**26.04**	**11.28**	**24.96**	13.82	8.04	4.60

Table 1: Automatic evaluation results. Models are separated based on the input features (one modality / multi-modal) and type of the mechanism used to compactly describe content of the image (max-pooling / attention). Best scores for both **+MAX** and **+ATT** modes are shown in bold. The colour intensity indicates how good the score is compared to the other models' scores.

Model Input	Type	mBLEU	self-CIDEr
IMG	+MAX	**50.63**	76.43
LNG	+MAX	52.24	75.59
IMG+LNG	+MAX	52.09	**76.46**
IMG	+ATT	51.82	75.51
LNG	+ATT	50.93	76.41
IMG+LNG	+ATT	**47.42**	**78.39**
GT	-	18.84	96.51

Table 2: Automatic paragraph diversity evaluation. mBLEU stands for the average score between all self-BLEU scores for *n*-grams (1, 2, 3, 4). Self-CIDEr stands for the average score of the LSA-based diversity metric. We also include ground-truth scores calculated from the test set (GT, coloured in blue). Best models are shown in bold. All scores are multiplied by 100 for better interpretability.

truth paragraphs, according to the WMD scores. This result comes at no decrease in other metrics, concerned with lexical accuracy.

When replacing max-pooling with late attention, we observe that the **IMG** model reaches the highest scores in BLEU-{2, 3, 4}, while finishing second in all other metrics. However, **IMG+LNG** model does not seem to benefit from the attention that much, reaching lower scores in comparison to its version with max-pooling. Interestingly, semantic information is beneficial to WMD, CIDEr and METEOR, which also take into account the syntactic structure of the sentences.

Table 2 contains the scores of the lexical diversity metrics. The best (i.e. the lowest) mBLEU scores are achieved by models which use either a visual modality with max-pooling (**IMG+MAX**) or both modalities with attention (**IMG+LNG+ATT**). The best self-CIDEr scores are achieved by both bi-modal architectures. In addition, **IMG+LNG+ATT** strongly outperforms all other models in both lexical diversity metrics: mBLEU is reduced by 3.21% indicating a smaller *n*-gram overlap between paragraph sentences, while self-CIDEr increases by 1.93% demonstrating that attention in the model which uses multimodal features helps to generate a more diverse set of sentences in terms of topicality.

Note that automatic metrics do not necessarily capture the quality of generated texts. This is because the intrinsic metrics might not be the best indicator for identifying clear differences in diversity and accuracy of the generated texts. In addition, such diversity metrics as mBLEU under-represent the diversity, being unable to take into account semantic differences between sentences. Therefore, we conduct a human evaluation experiment to achieve a better understanding of which input features and which pooling mechanism assists in the generation of both accurate and diverse paragraphs.

3.4 Human Evaluation

In the human evaluation task we are interested in the following properties of generated paragraphs covering both accuracy and diversity aspects: word choice, object salience, sentence structure and paragraph coherence. 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). 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 (those with the high reputation and task acceptance rate) and controlled for the number of tasks a single worker is able to submit (we set it to 30). We paid 0.15$ per task to a single worker. Finally, we obtained judgements from 154 unique Master workers for 1,750 image paragraphs overall. For each judgement criteria, we took the average score across all models; the results are shown in Table 3.

As shown by human evaluation, looking at the overall mean, the multi-modal information does

Input	Type	WC	OS	SS	PC	Mean
IMG	+MAX	31.58	38.24	**59.57**	**37.87**	41.81
LNG	+MAX	29.64	36.43	56.43	36.95	39.86
IMG+LNG	+MAX	**34.20**	**38.72**	57.85	37.06	41.95
Mean	+MAX	31.80	37.79	57.95	37.29	-
IMG	+ATT	36.91	45.10	69.34	32.27	45.90
LNG	+ATT	**37.06**	**46.78**	**72.95**	**40.88**	49.41
IMG+LNG	+ATT	33.81	37.67	45.37	34.71	37.89
Mean	+ATT	35.92	43.18	62.55	35.95	-
GT	-	89.83	87.36	83.07	84.78	-

Table 3: Human evaluation results. WC, OS, SS, PC stand for word choice, object salience, sentence structure and paragraph coherence. Each value in the table is the average of all scores for the corresponding criterion. The mean values per each model and type of pooling mechanism are coloured in light cyan.

help the generation of better paragraphs when using max-pooling. The **IMG+LNG** model with max-pooling might be a beneficial choice (scores first in two criteria out of four) in terms of word choice and identification of salient objects. The performance of the **IMG+LNG** model with max-pooling is close to the performance of the **IMG** model while the performance of the **LNG** model is slightly lower. Overall, attention is judged as more advantageous in general than max pooling, having higher mean scores across all criteria compared to the mean scores of max-pooling models. However, here the **IMG+LNG** model is outperformed by both uni-modal models. The **LNG** model which utilises semantic information and uses attention is judged as the best configuration by humans, which is in line with some previous work that reports strong bias on the semantic information (Agrawal et al., 2017). Note that while its performance is close to the **IMG** model in terms of word choice and object salience, the improvement of the **LNG** model is much more expressed in terms of sentence structure and paragraph coherence, categories where would expect that semantic information matters most. Interestingly, max-pooling does not seem to have the same effect on utilisation of semantic information: the **LNG+MAX** model achieves the lowest scores. A possible explanation for this is that when using max-pooling, the same semantic information is chosen for every sentence topic. At the same time, attention learns to select different semantic information for a sequence of topics. This appears to affect semantic features more than visual features.

Overall, the results indicate that both visual and semantic information are beneficial for the generated paragraphs as they affect different evaluation categories differently. The main challenge lies in information fusion of visual and semantic information in the model with attention. We believe that these results suggest the following future experiments: (i) detailed investigation of early vs. late attention (when to fuse two modalities and how), (ii) as van Miltenburg et al. (2017) argue, more control over human evaluation can provide us with better, more precise human judgements, (iii) training with other decoding strategies such as top-k sampling or nucleus sampling (Holtzman et al., 2019).

4 Related Work

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 and also remove the end of paragraph prediction.

Liang et al. (2017) also use the hierarchical network, but with an 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 auto-encoder for topic modelling based on region-level image features. They demonstrate that extracted topics are more representative and contain information relevant to sentence generation. We also model topic representations, but 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 depends 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 existing 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. Our approach does not use any attribute detectors to identify objects in the scene. Instead, we use the output of another pre-trained model 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 a mixture of visual features and visual sentinel, a vector obtained through the additional gate function on decoder memory state. Our model is guided by their approach: we are interested in deciding which type of 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). Also, our approach explores a different application of semantic information encoded in language: we use phrase representations to define sentence topics to choose from (topic selection) rather than directly guide the generation of words (micro-planning).

5 Conclusion

In this paper, we addressed the problem of generating both accurate and diverse image paragraphs. We demonstrated that utilising both visual and linguistic information might benefit the quality of generated texts depending on the pooling mechanism that is used. We showed that intrinsic evaluation metrics are insufficient for evaluation of paragraphs as they focus on lexical choice and do not capture human level of judgement (**LNG+ATT** is judged as the best model in human evaluation, while it is not among the leaders according to the automatic evaluation). We believe that our work is a good starting point for further investigation of the ways multiple sources of information about the world can be merged for learning generation of high-quality multi-sentence stories, describing real-world visual scenes.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2017. Don't just assume; look and answer: Overcoming priors for visual question answering. *arXiv*, arXiv:1712.00377 [cs.CV]:1–15.

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.

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*.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. The MIT Press, Cambridge, Massachusetts and London, England.

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.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in Type Theory with Records. In *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA.

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.

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*.

John D. Kelleher and Simon Dobnik. 2019. Referring to the recently seen: reference and perceptual memory in situated dialogue. In *CLASP Papers in Computational Linguistics: Dialogue and Perception – Extended papers from DaP2018 Gothenburg*, pages 41–50.

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.

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*.

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.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

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. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

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:

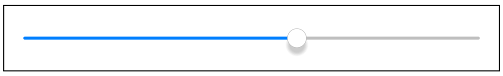
- choice of words: does the text correctly describe objects and events in the scene and with the right detail?
- relevance: does the text describe relevant objects and events in the scene?
- sentence structure: do the sentences have a good and grammatical structure?
- coherence: does the test progresses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

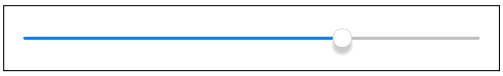
DESCRIPTION: there are two cows standing in the field. there are trees behind them.

How well do you agree with the following statements?

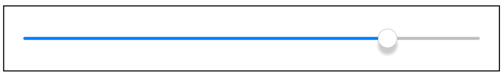
1. The description contains words that correctly refer to the objects and events in the image



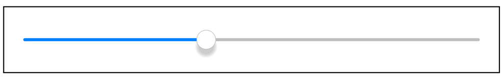
2. The description is referring to the relevant/important parts of the image.



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).