Challenges in Region-Specific Image Captioning: A Deep Learning Approach

Anonymous ACL submission

Abstract

Region-specific image captioning is the task of generating a caption from an image such that the caption is about the specific region in that image. This paper describes the challenges involved in region-specific image captioning and provides several methods to utilize the region-specific features to enhance the quality of the captions in addition to utilizing the features from the whole image. Our experiments on real-world data sets demonstrate that generating region-specific captions is challenging even after utilizing the information specific to the region. We analyze the variables impacting the quality of the captions which include the bounding box size and the region-specific feature extractor.

English Text: the snow is white. Hindi Text: बर्फ सफेद है

Malayalam Text: മഞ്ഞ് വെളുത്തതാണ് Gloss: Snow is white

Figure 1: Sample image with specific region and its description for caption generation. Image taken from Hindi Visual Genome (HVG) and Malayalam Visual Genome (MVG) (Parida et al., 2019)

1 Introduction

Image Captioning is the process of automatically describing an image with one or more natural language sentences (Hrga and Ivašić-Kos, 2019; Bernardi et al., 2016). It is a fundamental task related to a generic or specific image; and with the advent of deep neural networks, this area has made significant progress (Karpathy and Fei-Fei, 2015). The image captioning technology helps to create an applications such as: *i)* automation and acceleration of the close captioning process for digital content production[1]; *ii)* analysis of images and automatically generation of rich and detailed attributes for online catalogs [2]; and *iii)* supportive applications for the visually impaired people (Makav and Kılıç, 2019; Wang et al., 2020).

The neural encoder-decoder model has been found to be effective in image caption generation—where the encoder encodes the input image into a compact representation and the decoder transforms this representation into natural language describing the image (Pedersoli et al., 2017). Quite often, a specific region (containing one or more objects) is central to the image, and thus, is required to be described through the captioning process rather than focusing the entire image for the same. Although the existing models perform well at describing the whole image; their performance towards describing a part of the image or generating captions for a given specific region is relatively poor. If a region is cropped from the image and then subjected to the process of caption generation, it may result in poor captions as a single object (small region) may not be able to generate significant features towards the caption generation process. Also, if the selected region happens to be very small, it may require explicit resizing or padding prior to encoding its features. This process may as well result in poor quality features. The major challenge lies in the fact the generating meaningful captions for the part of

[1] https://artificialintelligence.oodles.io/blogs/ai-powered-image-caption-generator/

[2] https://evergreen.team/articles/automatic-image-captioning.html

an image (such as red bounding box in Fig 1) requires capturing the context of the whole image. For example, in Fig 1, it is difficult to predict “snow” for the region without considering the context from the whole image. The major motivation of our work is to address the issue of generating captions for a given region selecting relevant features from the whole image.

In this work, we propose a novel approach for generating region-specific captions through the fusion of features of the given region and features computed over the entire image. We demonstrate that a concatenation of weighted combination of these two sets of image features is a simple and effective mechanism to generate region-specific captions for an image. The attention-based LSTM is used to obtain the captions from fused features. The proposed approach was tested in multilingual (English, Hindi, and Malayalam) scenarios. Additionally, our analysis shows that the proposed approach is robust to the size or dimensions of the specific region of interest.

The major contribution of our work includes:

- Highlight the issues involved in region-specific image captioning and the existing evaluation metrics.
- Propose a novel approach to build encoder-decoder model for captioning using the image-level and region-specific features.
- Discuss the benefits and possible NLP applications using the proposed approach.

2 Related Work

In (Li et al., 2017), Li *et al.* proposed global-local attention (GLA) framework by integrating local representation at object-level with global representation at image-level through an attention mechanism. The proposed approach performed state-of-the-art performance on the MS-COCO dataset based on automatic evaluation metrics. A geometric attention-based model for image captioning which incorporates information about the spatial relationships between input detected objects through geometric attention has been proposed in (Herdade et al., 2019). A dense captioning model proposed by Johnson et al. (2016) to describe the regions of an image that is close to our work but our dataset has a single region with its caption is available. Nakayama *et al.* (Nakayama et al., 2020) proposed an English-to-Japanese multimodal corpus *F30kEnt-JP* with many-to-many phrase-to-region linking aiming to promote multilingual image captioning and multimodal machine translation.

Region-specific Image Captioning Although much work has been done in image caption generation few researchers tried to generate a caption for a given specific region as input. For region-specific Hindi caption generation using the Hindi Visual Genome (HVG) dataset (Parida et al., 2019), (Meetei et al., 2019) used VGG16 for feature extraction and fed to RNN model with beam search. Their model obtained a very low BLEU score result of *2.59* on the evaluation test and *0* on the challenge test. The HVG dataset has a single reference caption available for a specific region for evaluation and the challenge test harder (consisting of many ambiguous words in English selected using a semi-automatic approach) compared to the evaluation test set due to which many researchers obtained very low BLEU (Papineni et al., 2002) score on this dataset on automatic evaluation metric though comparatively well based on human evaluation (Laskar et al., 2019; Meetei et al., 2019; Parida et al., 2020; Laskar et al., 2020; Nakazawa et al., 2019, 2020; Kaur and Josan, 2020). The HVG dataset serves in the Workshop on Asian Translation (WAT)[3] for the Multimodal image captioning task (Hindi) since 2019. The task includes the generation of “Hindi" captions for the given image, a region, and its captions in Hindi as shown in the Figure 1. The evaluation scores by the WAT participants on the HVG dataset for the “Hindi caption" task are summarized in the table Table 1.

3 Proposed Method

The task of captioning a complete image has recently been studied by several researchers (Yang and Okazaki, 2020; Yang et al., 2017; Lindh et al., 2018; Staniūtė and Šešok, 2019;

[3] https://lotus.kuee.kyoto-u.ac.jp/WAT/

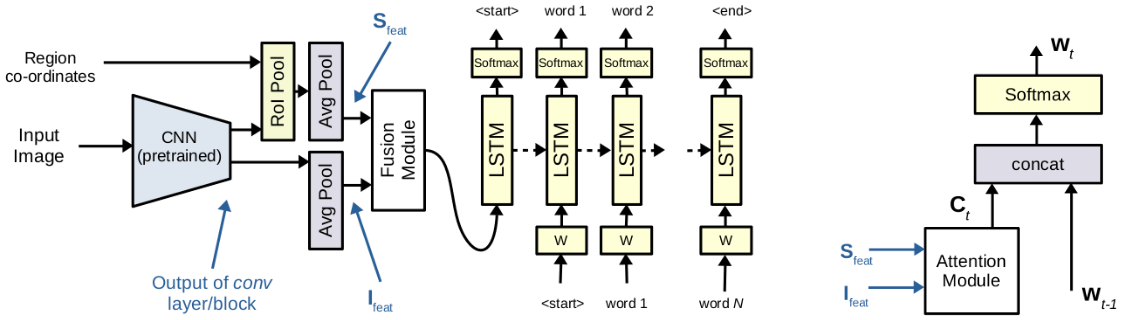
Author	Dataset	Method	BLEU	Human Evaluation
(Meetei et al., 2019)	HVG(EVTest)	VGG-RNN	2.59	51.78
	HVG(CHTest)	VGG-RNN	0.00	44.46
(Laskar et al., 2019)	HVG(CHTest)	merge-model	0.00	26.54
(Parida et al., 2020)	HVG (EVTest)	InceptionRes-NetV2	0.78	47.16
	HVG(CHTest)	InceptionRes-NetV2	0.00	52.10

Table 1: Image caption (Hindi) results on HVG dataset. The Human evaluation results are based on the WAT shared task (Nakazawa et al., 2019, 2020). *EVTest* and *CHTest* represent HVG Evaluation Test set and Challenge Test set, respectively.

Miyazaki and Shimizu, 2016; Wu et al., 2017). However, generating a caption for a specific region in the image is non-trivial. Most of the proposed architectures for the generation of caption for complete images consist of two modules: an image-based feature extractor, and a language-based model that transforms image features (or *embeddings*) into a sequence of natural words. A naive method of obtaining captions for a specific region in an image is to train an existing image caption network by providing only the cropped region as input. This method, however, does not consider the context or semantic relationship of a specific region with its surroundings (or the entire image). The caption is likely to be more meaningful when the context is well-captured and also incorporated in the generation of captions. Therefore, the features of the entire image essentially play an important role in shaping up the captions of its specific region. With these objectives, we propose the caption generation architecture that consists of a feature fusion module in addition to the image-based encoder and two variants of LSTM-based decoder. The overall architecture is provided in Fig. 2a. The details of each block are described below.

Image Encoder: Several recent methods of image caption generation have advocated the use of deep CNNs as feature extractors for images (Xu et al., 2015). A typical deep CNN consists of several layers or blocks of convolution (conv) and pooling layers, followed by one or more fully connected layers. The outputs of final (or pre-final) conv layers represent complex features from images learned hierarchically. These features are learned over local small regions (also known as receptive fields), and the overall receptive field expands as the features move to higher conv layers. The expansion is primarily due to pooling and conv (through kernel and stride values) operations– both of which preserve the relative spatial relations. Therefore, we can nominally correlate the subset of features at the final conv layer with the spatial region in the input image. Here, the term subset is used with reference to spatial dimensions of the output of the corresponding conv layer; whereas the channel (or depth) dimension remains unaltered. We use this simple idea to obtain the features of a region through the ROI Pooling mechanism. Let $\mathbf{F} \in \mathbb{R}^{MNC}$ be the features of the final conv layer of a pre-trained image CNN where C represents the number of channels or maps, and M, N are the spatial dimensions of each feature map. For a given input image dimensions, we compute the scaling factors in x and y dimensions from the knowledge of $(M,\ N)$, and thus, also compute the corresponding coordinates of the region bounding box in $\mathbf{F}$, say $(m,\ n)$. This procedure helps extract a subset of image features, $\mathbf{F}_s \in \mathbb{R}^{mnC}$ that predominantly consists of features from the region of interest. The subset $\mathbf{F}_s$ is obtained through the region of interest (RoI) pooling (Girshick, 2015). It should be noted that the dimensions of the input image, as well as the bounding box, are not constant; and therefore, the spatial dimensions of $\mathbf{F}_s$ also vary for every image. To bring consistency in region features, we apply spatial pooling with specific stride (≥ 1) values. However, we do not modify the channel dimensions of $\mathbf{F}_s$. The final features, thus obtained, are linearized to form a single column vector. We denote the region-subset features as S_{feat}. The features of the complete image are nothing but $\mathbf{F}$. We apply spatial pooling on this feature set to reduce their dimensionality, and obtain the linearized vector of full-image features denoted as I_{feat} in the subsequent discussions.

Fusion Module: To generate efficient and meaningful captions for a region of the image,



(a) (b)

Figure 2: Architecture of the proposed region-specific image caption generator. (a) shows the baseline consisting of a pre-trained image CNN as feature extractor, followed by a fusion module, and then an LSTM-based decoder (without attention mechanism) to obtain captions. (b) is the attention mechanism for LSTM-based decoder. The output of this module, $\mathbf{W}_t$, is fed to the LSTM blocks.

we need to consider the features of the region S_{feat} along with the features of the entire image I_{feat}. The combining of feature vectors is crucial in generating descriptions for the region. The region-level features capture details of the region (objects) to be described; whereas image-level features provide an overall context. Therefore, a concatenated form of both feature vectors is an obvious yet effective choice for the input ($\mathbf{f}$) to the decoder. A simple concatenation assigns equal weightage to both features, S_{feat} and I_{feat}; which may not necessarily result in captions focused on the intended region. In this work, we investigate an idea of the concatenation of weighted features from region and image for region-specific caption generation. The fused feature, $\mathbf{f}$, can be represented as:

$$\mathbf{f} = [\alpha\, S_{\text{feat}}; (1-\alpha)\, I_{\text{feat}}] \qquad (1)$$

where α represents the weightage parameter in $[0.50, 1]$. Although α can theoretically be as low as zero (which indicates discarding the region-level features), we consider only the range where S_{feat} receives equal or higher weightage than I_{feat} resulting in higher importance to the region over its surroundings for the present task. When α is set to 0.50, both feature vectors receive equal weightage, which is akin to representation, before feeding them to the decoder. For $\alpha = 0.66$, the region-level features are weighted twice as high as the entire image-level features. At another extreme with α equal to one, the captions are generated using only region-level features, and the image-level features are discarded.

The weighing of a feature vector simply scales the magnitude of the corresponding vector without altering its orientation. Unlike the fusion mechanisms based on weighted addition, we do not modify the complex information captured by the features (except for scale); however, its relative importance with respect to the other set of features is adjusted for better caption generation. The fused feature $\mathbf{f}$ with the dimensionality of the sum of both feature vectors are then fed to the LSTM-based decoder.

LSTM Decoder: In the proposed approach, the encoder module is not trainable, it only extracts the image features however the LSTM decoder is trainable. We used LSTM decoder using the image features for caption generation using greedy search approach (Soh). We used the cross-entropy loss during decoding (Yu et al., 2019).

Attention-LSTM Decoder For the task of region-based caption generation, the region to focus on is known *a priori*. Therefore, the requirement for deciding which parts to focus on most (which is what an attention module generally does in a decoder) is apparently not critical. However, it would be useful to learn which components (subregion or whole image) to focus on and based on this generate the next token conditioned on the previous token and the set of subregion and the image features. Thus, we employ an attention-based decoder that generates each token by first deciding which component to focus on using an attention module (Bahdanau et al., 2014). The attention module takes as input the two image feature vectors, S_{feat} and I_{feat}, and the last hidden state $\mathbf{h}_t$ of the LSTM and computes the context vector $\mathbf{c}_t$ which is then concatenated to the input of the LSTM in the previous time step $\mathbf{w}_{t-1}$ to generate $\mathbf{w}_t$. This leads the modified decoder architecture as shown in Fig. 2b whereas the encoder and fusion module from original architecture (Fig 2a) remain unchanged. The process of attention-based decoding is outlined below. Let W_1, W_2, V and W be trainable weight matrices. First, the hidden state $\mathbf{h}_t$ is used to obtain $\mathbf{s}_t$ and $\mathbf{i}_t$ for the subregion and the image respectively.

$$\mathbf{s}_t = \tanh\left(W_1 S_{\text{feat}} + W_2 \mathbf{h}_t\right) \tag{2}$$

$$\mathbf{i}_t = \tanh\left(W_1 I_{\text{feat}} + W_2 \mathbf{h}_t\right) \tag{3}$$

The attention scores a_s and a_i are then obtained by applying softmax,

$$[a_s, a_i]_t = \text{softmax}([V\mathbf{s}_t, V\mathbf{i}_t]), \tag{4}$$

where V maps $\mathbf{s}_t$ and $\mathbf{i}_t$ to a single dimension. The context vector is then obtained by summation after scaling with attention scores: $\mathbf{c}_t = a_s S_{\text{feat}} + a_i I_{\text{feat}}$. Finally, the logits are obtained by concatenating $\mathbf{c}_t$ and $\mathbf{w}_{t-1}$ and projecting onto a space with dimension equal to the vocabulary size: $\mathbf{w_t} = W[\mathbf{c}_t; \mathbf{w}_{t-1}]$.

4 Experiment and Results

This section discusses our experimental settings and obtained results.

4.1 Datasets

Our experiments are based on the HVG and MVG multimodal datasets.

HVG: For every sample image, the HVG dataset provides a region (as a rectangular bounding box) and its bi-lingual (English and Hindi) segments (captions). The training set contains nearly 29K segments. Further 1K and 1.6K segments are provided in development and test sets, respectively. Additionally, a challenge test set of 1400 segments given which was created by searching for (particularly) ambiguous English words based on the embedding similarity and manually selecting those where the image helps to resolve the ambiguity (Parida et al., 2019).

MVG: The MVG[4] dataset is an extension of the HVG dataset for supporting Malayalam, which belongs to the Dravidian language family (Kumar et al., 2017). The dataset size, images are the same as HVG. While HVG contains bilingual (English and Hindi) segments, MVG contains bilingual (English and Malayalam) segments.

4.2 Pretrained Models

To extract the image features we have considered two commonly used CNN architectures that have proved to be highly successful at general image classification:

ResNet: One of the extremely successful CNN architectures proposed to solve the issue of diminishing gradient where the idea is to skip the connection and pass the residual to the next layer so that the model can continue to train. ResNet exists in several variants based on the number of deep layers (He et al., 2016). We have considered a 50-layer model, also known as ResNet-50, as a feature extractor. Additionally, we have extracted features as the output of the third (L3) and fourth (L4) blocks of the ResNet-50 for our experiments.

VGG: Being one of the top-ranked architectures for image classification. The architecture derives its name from the Visual Geometry Group (at Oxford University) who proposed this idea. Similar to ResNet, the VGG architecture is also available in multiple variants; we have chosen the 19-layer model referred to as VGG-19 (Simonyan and Zisserman, 2014).

4.3 Training

The image and subregion features are concatenated after scaling based on the weightage parameter α. The dimensionality of subregion and the image level features vectors is 2048 for ResNet L4 and VGG models, while for the ResNet L3 features, the subregion feature vector size is 1024. We set α to 0.0 for the baseline experiment, where only the image level features have non-zero values. When α is 1.0, only the subregion features have non-zero values. We set α to 0.5 and 0.66 to fuse subregion features and image features when both have

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533

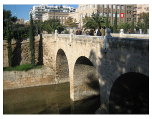
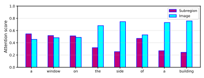
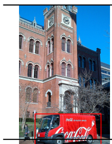
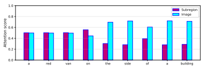
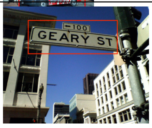
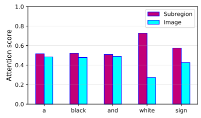
		gold : window on a building LSTM alpha=0.0 : a large boat on the water LSTM alpha=0.5 : a large white building LSTM alpha=1.0 : the man is wearing a hat Att LSTM : a window on the side of a building
		gold : a van on the side of street LSTM alpha=0.0 : a clock on the building LSTM alpha=0.5 : a red and white bus LSTM alpha=1.0 : the bus is yellow Att LSTM : a red van on the side of a building
		gold : black and white street signs LSTM alpha=0.0 : a sign on a pole LSTM alpha=0.5 : a sign on the street LSTM alpha=1.0 : a man with a smile on his face Att LSTM : a black and white sign

Table 2: Captions generated using the proposed model from the ResNet L3 image features. These are some positive results using attention based model.

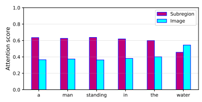
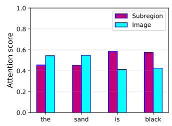
		gold : two elephants standing in the water LSTM alpha=0.0 : a elephant in the water LSTM alpha=0.5 : the elephant is eating LSTM alpha=1.0 : man riding a skateboard Att LSTM : a man standing in the water
		gold : The sand is brown LSTM alpha=0.0 : a large blue sky LSTM alpha=0.5 : a red and white fire hydrant LSTM alpha=1.0 : the ground is made of asphalt Att LSTM : the sand is black

Table 3: Captions generated using the proposed model from the ResNet L3 image features. These are some negative results using attention based model.

Hyperparameter	**Value**
embedding size	128
hidden size	256
number of layers	{1, 2, 4}
dropout	0.3
learning rate	{0.1, 0.01, 0.001, 0.0001}

Table 4: Values of model hyperparameters.

non-zero values. The concatenated features are projected to space with the same dimensionality as the token embedding dimension, which we set to 128, using a linear layer. This serves as input to both the LSTM decoder and the Attention-LSTM decoder which generate tokens autoregressively.

We have tokenized training captions on the word level and built the vocabulary, which is fixed for all the experiments for a given language. We experimented with NLTK tokenizer (English), Moses tokenizer[5] and the corresponding detokenizer (English), and the Indic NLP tokenizer (Kunchukuttan, 2020) and the corresponding detokenizer (Hindi and Malayalam) and found that any choice of the tokenizer and detokenizer leads to very similar results.

We have implemented the experiments using PyTorch, with the choice of hyperparameters as provided in Table 4. The decoder was trained using Adam optimizer (Kingma and Ba, 2014). For each training experiment, we computed the adopted early stopping criterion based on the BLEU scores on the development set. The model with the best BLEU score (dev set) was chosen to report the scores. That captions are generated token-by-token using greedy decoding till the end-of-sentence token is generated or the maximum sequence length is reached, which we set to 20.

The results for the test and challenge sets are shown in Tables 5, 6 and 7 for generating English, Hindi, and Malayalam captions respectively. As per the automatic evaluation metric (BLEU), the attention-based LSTM decoder overall performs better as compared to the LSTM-based decoder. Based on the evaluation score, image feature extraction using ResNet performs better in comparison to VGGNet, and evaluation test set performance is better as compared to challenge test set for

[5] https://github.com/alvations/sacremoses

Method	ResNet L3 image features EVTest	CHTest	ResNet L4 image features EVTest	CHTest	VGGNet image features EVTest	CHTest
LSTM						
I ($\alpha = 0.0$)	2.5	0.4	3.0	0.3	2.7	0.6
S ($\alpha = 1.0$)	1.0	0.5	2.5	0.6	2.3	0.4
I + S ($\alpha = 0.5$)	**3.3**	0.6	2.4	0.7	2.7	0.6
I + S ($\alpha = 0.66$)	2.2	0.3	2.9	0.6	3.2	0.7
Attention-LSTM	2.4	1.0	2.9	**1.1**	2.0	0.6

Table 5: English caption generation results in terms of BLEU scores.

Method	ResNet L3 image features EVTest	CHTest	ResNet L4 image features EVTest	CHTest	VGGNet image features EVTest	CHTest
LSTM						
I ($\alpha = 0.0$)	1.9	0.6	1.8	0.4	1.7	0.4
S ($\alpha = 1.0$)	1.5	0.3	1.3	0.3	1.2	0.3
I + S ($\alpha = 0.5$)	2.0	0.5	2.0	0.5	1.7	0.3
I + S ($\alpha = 0.66$)	1.9	0.7	1.8	0.7	1.8	0.8
Attention-LSTM	1.5	**0.9**	**2.5**	0.7	1.4	0.7

Table 6: Hindi caption generation results in terms of BLEU scores.

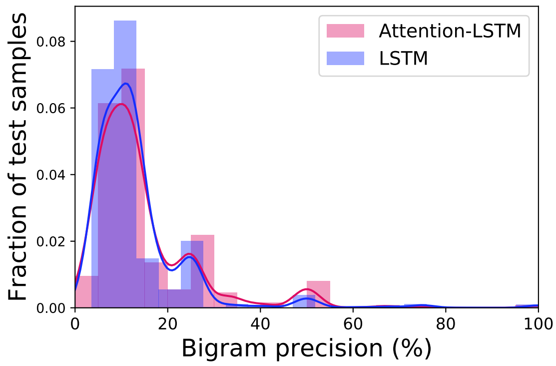


Figure 3: Distribution of caption level bigram precisions for caption generation.

all three (English, Hindi, and Malayalam) languages.

5 Analysis and Discussion

Though image caption generation considers the full image and its associated captions for building a model and generating captions for any input images, this work focused on building a model to generate captions for a given specific region. We used the automatic evaluation metric BLEU for a comparison purpose with other researchers' reported scores for the "Hindi image captioning task" using the HVG dataset as shown in Table 1. For the challenge test set, our proposed model obtained a better result (see Table 6).

Impact of attention decoder The overall evaluation results using BLEU scores for the LSTM decoder vs the Attention-LSTM decoder are very similar as shown in Tables 5, 6 and 7. To analyze this further we evaluate generated captions by calculating the bigram precision score at caption level. The fraction of captions for different bigram precision scores is shown in Fig 3 where we compare the English test set captions generated using the LSTM decoder with $\alpha = 0.5$ and the Attention-LSTM decoder, both using ResNet L3 image features. The comparison shows that captions generated using the Attention-LSTM model are more precise as its curve lies below the LSTM curve for lower precision values. For higher precision values, the trend reverses where the Attention-LSTM curve lies above, suggesting that more captions have high precision scores in comparison to LSTM. In addition to generating better captions, the attention-based decoding also enables us to examine the component (subregion or whole image) that the model focuses on while generating tokens. The sample captions and the attention scores are shown in Table 8. The captions generated without attention in different settings ($\alpha = 0.0, 0.5, 1.0$) are also shown for comparison.

Impact of bounding box size The size of the bounding box is an important factor that varies across images.[6] To measure the impact of the bounding box size on caption generation, we computed sentence level BLEU scores on the English test set for different sizes shown

[6] For more than half of the test samples, the bounding box size is less than 10% of the image size.

Method	ResNet L3 image features		ResNet L4 image features		VGGNet image features	
	EVTest	CHTest	EVTest	CHTest	EVTest	CHTest
LSTM						
I ($\alpha = 0.0$)	0.4	0.1	0.3	0.6	0.9	0.3
S ($\alpha = 1.0$)	0.3	0.1	0.6	0.6	1.4	0.8
I + S ($\alpha = 0.5$)	1.0	0.2	0.5	0.2	0.4	0.2
I + S ($\alpha = 0.66$)	1.1	0.6	1.1	0.2	0.7	0.2
Attention-LSTM	**2.5**	**1.1**	2.3	0.6	2.4	**1.1**

Table 7: Malayalam caption generation results in terms of BLEU scores.

	Ref: the snow is white. Caption(I): a man on a snowboard. Caption(S): a on a. Caption(I+S): a group of people skiing.	Ref: बर्फ सफेद है Caption(I): एक आदमी एक सर्फ़बोर्ड पर है Gloss: a man on the surfboard Caption(S): एक आदमी एक पहाड़ी के नीचे स्कीइंग कर रहा है Gloss: a man skiing below the mountain Caption(I+S): बर्फ में स्कीयर Gloss: skier in the snow

Table 8: Sample captions (English and Hindi) generated for the challenge test set using our model in different settings (I: Image only ($\alpha = 0$), S: Subregion only ($\alpha = 1.0$), and S+I: Image + Subregion ($\alpha = 0.66$). The outputs are taken from the models based on the best performance (BLEU score).

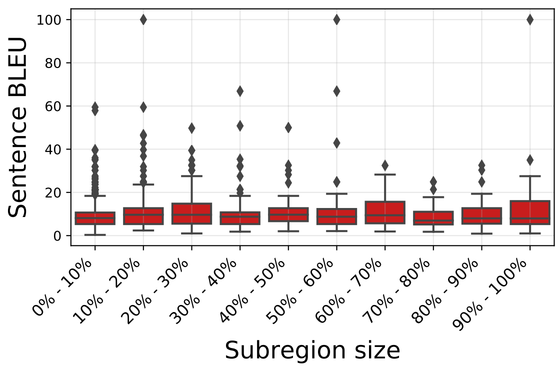


Figure 4: Box plot of sentence level BLEU scores across different subregion sizes.

as a fraction of the image size in Fig 4. The Attention-LSTM with ResNet L3 features was used to generate the captions. The box plot shows that on average the sentence level BLEU scores are very close for different bounding box sizes, indicating that the attention-based caption generation model is robust to bounding box size.

Although the proposed model generates meaningful and better captions in low BLEU scores for the HVG, and MVG datasets. As per our analysis, these are due to:

- Both HVG and MVG dataset contain a single gold caption for evaluation.
- The automatic evaluation metric *BLEU* is not suitable for image captioning task as in multilingual scenario the image can be described in various ways and deep learning-based model can able to produce much better caption than the reference (Cui et al., 2018). Some researchers proposing new metrics for automatic evaluation of image captions such as "SPICE", (Anderson et al., 2016) "TIGEr" (Jiang et al., 2020) and claims its closely matching with human evaluation.

6 Conclusions

In this work, we highlighted the challenges involved in generating meaningful caption for a defined region of an image. Our proposed approach can generate meaningful captions by fusing features of whole and region-specific images. The proposed approach can be useful for building NLP applications such as *i)* image labeling in different commercial/non-commercial applications (E-Commerce product labeling) including multilingual scenarios, *ii)* application for visually impaired persons.

The future work includes exploring the combination of different image and object feature sets, and subjective evaluation and linguistic analysis of captions generated in Hindi and Malayalam.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Euro-*

pean conference on computer vision*, pages 382–398. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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. *Journal of Artificial Intelligence Research*, 55:409–442.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5804–5812. IEEE.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *arXiv e-prints*, pages arXiv–1906.

I Hrga and M Ivašić-Kos. 2019. Deep image captioning: An overview. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 995–1000. IEEE.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2020. Tiger: Text-to-image grounding for image caption evaluation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2141–2152. 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*, pages 4565–4574.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jagroop Kaur and Gurpreet Singh Josan. 2020. English to hindi multi modal image caption translation. *Journal of Scientific Research*, 64(2).

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. Morphological analysis of the dravidian language family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 217–222.

Anoop Kunchukuttan. 2020. The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Multimodal neural machine translation for english to hindi. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113.

Sahinur Rahman Laskar, Rohit Pratap Singh, Partha Pakray, and Sivaji Bandyopadhyay. 2019. English to hindi multi-modal neural machine translation and hindi image captioning. In *Proceedings of the 6th Workshop on Asian Translation*, pages 62–67.

Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image caption with global-local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Annika Lindh, Robert J Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D Kelleher. 2018. Generating diverse and meaningful captions. In *International Conference on Artificial Neural Networks*, pages 176–187. Springer.

Burak Makav and Volkan Kılıç. 2019. A new image captioning approach for visually impaired people. In *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 945–949. IEEE.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4204–4210.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2020. Overview of the 7th workshop on asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

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.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4).

Shantipriya Parida, Petr Motlicek, Amulya Ratna Dash, Satya Ranjan Dash, Debasish Kumar Mallick, Satya Prakash Biswal, Priyanka Pattnaik, Biranchi Narayan Nayak, and Ondřej Bojar. 2020. Odianlps participation in wat2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 103–108.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE international conference on computer vision*, pages 1242–1250.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Moses Soh. Learning cnn-lstm architectures for image caption generation.

Raimonda Staniūtė and Dmitrij Šešok. 2019. A systematic literature review on image captioning. *Applied Sciences*, 9(10):2024.

Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Zhishen Yang and Naoaki Okazaki. 2020. Image caption generation for news articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1941–1951.

Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, and Yongfeng Huang. 2017. Image captioning with object detection and localization. In *International Conference on Image and Graphics*, pages 109–118. Springer.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.