Volume 64, Issue 2, 2020

Journal of Scientific Research

Institute of Science,

Banaras Hindu University, Varanasi, India.

English to Hindi Multi Modal Image Caption Translation

Jagroop Kaur*[1], Gurpreet Singh Josan[2]

*[1]Department of Computer Science and Engineering, Punjabi University Patiala, jagroop_80@rediffmail.com

[2]Department of Computer Science, Punjabi University Patiala, josangurpreet@pbi.ac.in

***Abstract*—When it comes to describe image by words, the possibilities are numerous. Generating caption automatically from image is a well researched area. The concept is now advanced to utilise multiple models for translating captions in another language. One of the major hurdle in machine translation is word sense disambiguation. Visual cues from image can be helpful in disambiguate source words. System are being developed to generate image descriptions in a target language from an input image and one or more descriptions in source language. This paper describes the multi modal Neural Machine Translation systems for translating image captions from English to Hindi. Various multi-modal architectures were explored using local visual feature, global visual features, attention mechanisms, and pre-trained embedding. The features obtained form various models are integrated in various ways. We also tried re-ranking method. The systems are evaluated on BLEU score, RIBES score and AM/FM score. Re-ranking method proves to be best over all our other methods.**

***Index Terms*—Automatic Image Caption, Multi Modal Machine Translation, Deep Neural Network, Image Feature Extraction**

I. Introduction

Image is worth of thousand words. When it comes to describe image by words, the possibilities are numerous. An image can be described in number of ways. Each description can highlight different aspect of an image. This process of describing image using text is called image captioning. Captioning comes under broader term of image annotation which is a process of assigning metadata to an image. Annotation may include keywords, indexes or captions.

A good caption should enable the reader to understand the image and its relevance to the topic. It enables the people to understand story behind image. Image captions are not just a piece of text. They helps to grab attention of reader and compel them to explore the related article. Captions are text which can be scanned. Thus if an image has a caption, it is more likely that a search engine can find it. With recent advancements in computers, researchers aim to develop automatic image captioning system. Such system find applications in various areas. Automatic captioning can be useful to generate text from images. Many Natural Language Processing (NLP) applications which extract summary from a given text data can be augmented by automatic captioning system to obtain better summaries. Another use can be describing a video frame by frame. A text to speech system can be augmented with automatic image captioning system to explain the image to Visually impaired people. Tremendous amount of images has been clicked and uploaded by users daily which can be indexed and classified automatically using such system.

Captioning an image is not an easy task for human being themselves. The area fascinates lot of researchers to work on this problem. Detection of features from image is the primary step in all the approaches till date. Hossain et al. (2019) describes nicely the various techniques for generation of image captions. Researchers applied different techniques ranging from manual to automatic for identification of features from images. Captions can be generated from predefined templates where words to fill the templates are selected based on features detected in template. In another approach, captions are extracted from existing pool of captions. A pool of images along with their captions is collected. The caption of the image which has similar features with image in question is selected as caption. In another approach, multiple models are used to generate caption of an image. The image features are used with language model to generate captions. Each approach has its pros and cons. Template based approach is characterised by grammatically correct caption but fails to generate variable length captions. Captions generated form existing pool are general one and does not generate image specific captions. Captions generated with multiple models are syntactically and semantically more accurate than all previous approaches.

With deep neural network, systems for caption generation can be trained and it is now quite possible to

∗*Corresponding Author*

DOI: 10.37398/JSR.2020.640238

generate an understandable caption from a given image. For example Vinyals et al. (2014) Xu et al. (2015), You et al. (2016) achieved good results for caption generation. On the other hand, machine translation is one of the oldest area of research. Deep neural network also cause the dramatic advancements in this area with great success. Recently researchers are attracted toward multi-modal translation task where the image caption in source language is translated in target language using the cues from both image and text. For example Bernardi et al. (2016), Feng and Lapata (2010),Yang et al. (2015) etc. successfully applied multi-modal architecture for various NLP tasks. A great push on research in this area was given by introducing first shared task on multi modal Machine Translation by involving multilingual component in 2016Lucia et al. (2016). The shared task was to generate image descriptions in a target language from an input image and one or more descriptions in source language.

In 2019, the challenge posed by organizer of the task is extended to Indic language also where the input language is English and target language is Hindi. The task aims to check how visual cues helps in disambiguate source words. Along with the image, coordinates of object about which the caption describes are also given. This paper presents our approach for this multi modal machine translation task. Different multi modal architectures were experimented for translating captions from English to Hindi. The next section discuss related literature followed by description of data set and our methodology. Last section discuss the results followed by conclusion.

II. Related Work

Kulkarni et al. (2013), Karpathy and Li (2014), Vinyals et al. (2014) showed that image captioning is a solvable problem. The first attempt to use multi modal translation using visual context was found in Elliott et al. (2015) and Hitschler et al. (2016). Hitschler et al. (2016) proposed a caption generation approach based on monolingual data set. They suggest to use the caption of most similar image in target space. The source image feature were compared with target images in database and captions of the most similar images are extracted for re-ranking. Liu et al. (2016) propose that instead of using whole image feature in caption generation, detect objects from image and serve the sequence of detected objects as the source sequence of the RNN model. Multi modal machine translation attracts the focus of researcher after WAT 2016 when first multi-modal shared task has been conducted by organizers of WAT 2106 Lucia et al. (2016). In the competition, Hitschler et al. (2016) uses statistical machine translation system, Libovický et al. (2016) and Shah et al. (2016) uses Moses, while rest uses neural machine translation Huang et al. (2016); Hokamp and Calixto (2016); Calixto et al. (2016); Caglayan et al. (2016); Rodríguez Guasch and Costa-jussà (2016). Huang et al. (2016) experimented with local as well as global features of image while rest of the teams uses global features. VGG19 was the choice of most participants for extracting image feature except Caglayan et al. (2016) uses Resnet50. Huang et al. (2016) tested three different hypothesis by appending image features with head/tail of text and parallel dissipating with LSTM network. Hokamp and Calixto (2016) uses image features to initialize the target side decoder. Calixto et al. (2016) tried integrating separate attention mechanisms over the source language and visual features. The system using Moses for translation outperforms the NMT systems producing 53.2, 34.2, and 48.7 respectively for Meteor, BLEU and TER. It was also observed that visual information provides marginal improvement.

Similar task were also organized in 2017 Elliott et al. (2017) and 2018 Barrault et al. (2018) with added language pairs. Çağlayan et al. (2017) system's produced best results in shared task of 2017. Their system uses separate attention over source text and image features. They also proposed neural machine translation system where global visual features were multiplicatively interacted with word embedding. For image feature extraction Resnet50 had been employed. Grönroos et al. (2018) were able to produce better results in shared task of 2018 but it attributed to underlying good quality of training samples.

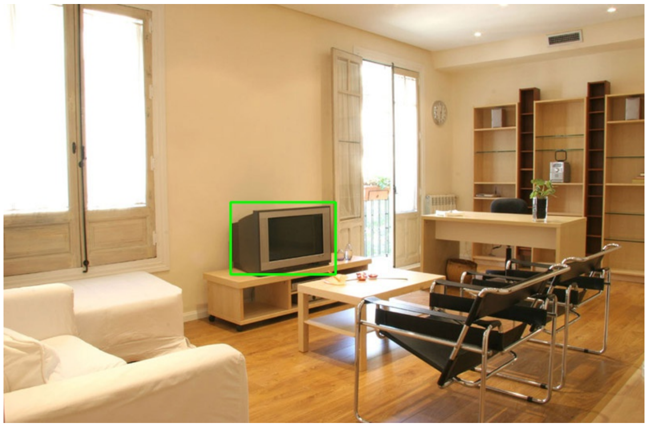
In 6th workshop during 2019, the organizers of WAT include a multi modal task where English caption of image has to be translated to Hindi using textual and visual information Nakazawa et al. (2019). Sanayai Meetei et al. (2019) used multi modal machine translation on English-Hindi language pair and conclude that using image features improves translation. Similarly, Laskar et al. (2019) uses transformer model augmented with image features to translate captions. The image features were extracted using VGG19 model. The translation system apply doubly-attentive decoder to predict sentences, which shows better performance than text-only translations. The overall impression remains same that visual feature does not help much in translation task.

III. Dataset

We use Hindi Visual Genome data set provided by Parida et al. (2019). Total number of samples for training were 29000. Further 1000 samples were provided for validation and 1600 samples for evaluation task. A separate challenge set of 1400 samples was also provided. This challenge set contains ambiguous English words based on the embedding similarity along with the image where it helps to resolve the ambiguity. Each source language sentence is accompanied by an image with a bounding box. The source text describes about the portion of image bound by box. Only one target sentence was provided against each source sentence. See fig 1 for example.

IV. Models

For a given input of an image, a rectangular region in that image and a short English caption of the rectangular region, system needs to translate the caption to Hindi. The problem falls in sequence to sequence translation domain and encoder decoder architecture has been successfully



Source Caption: A big tv on a stand
Reference Given: "एक स्टैंड पर एक बड़ा टीवी"

Fig. 1. Sample image and caption from dataset

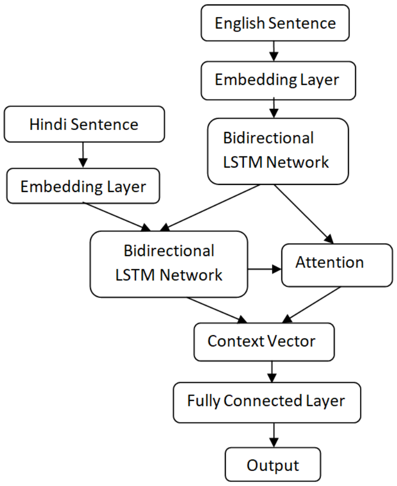


Fig. 2. Architecture for caption generation from text only

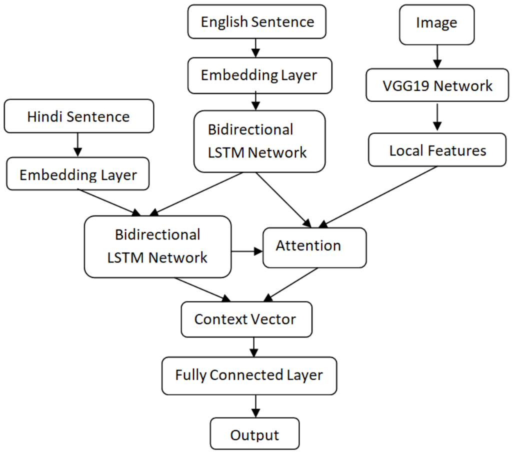


Fig. 3. Architecture for caption generation from text using image feature for attention calculations

applied to such problems Sutskever et al. (2014). The aim here is to translate source caption to target caption looking cues from the image. Thus we need to incorporate image features in translation process. In literature, number of methods are mentioned to incorporate image features with text features. Some combine image feature with word feature, some use image feature as attention while other uses image feature as a part of input sequence. Tanti et al. (2017b) and Tanti et al. (2017a) presents a comprehensive overview on this. As both image and coordinate of object in image are provided, multiple ways exist for using image features. For the task in hand we tried various models as follow:

A. Model 1(Baseline)

The baseline model is a simple seq2seq model using bidirectional LSTM on encoder side and unidirectional LSTM on decoder side (see figure 2). In baseline, image features are not utilized. Luong's dot attention Luong et al. (2015) has been used to get the weightage of source words on target selection. Embedding size for both English and Hindi has been kept 256. Hidden units of encoder LSTM layer are kept at 1024 and decoder LSTM 2048. The output of LSTM is combined with context which was calculated using luong's dot product method. The combined context is passed through dense layer producing the target word. Training was done using teacher enforcement technique. During prediction, one token is passed through network and next token was predicted by the model. Beam search method with beam size 3 has been used to select the final output.

B. Model 2 (Image Feature Combined with multiplicative way)

This model extended the baseline system by utilizing the features of image in attention vector. The caption in the source language is describing the portion of image whose coordinates are given. Thus local features from only that part of image described by bounding box are extracted using VGG19 network. These features are termed as local image feature. The last layer of VGG19 network which produces feature vector of size 4096 has been utilized for this task. These features are passed through Dense network to downgrade them to 2048. These features are utilized to calculate context vector as per equation

$$Score(h_t, \overline{h_s}, f_i) = h_t^T.\overline{h_s}.f_i \qquad (1)$$

$$\alpha_{ts} = \frac{exp\ (Score(h_t, \overline{h_s}, f_i)}{\sum_{s'=1}^{S} exp\ (Score(h_t, \overline{h_s}, f_i)} \qquad (2)$$

$$c_t = \sum_s \alpha_{ts}.\overline{h_s} \qquad (3)$$

Where f_i is image features, α_{ts} is attention vector and c_t is context vector. This context vector is concatenated with the output of decoder LSTM to predict the next token in sequence (see figure 3). All other setup remain same.

C. Model 3 (Image Features Combined with Additive Way)

In this model, context vector is created by additive ways instead of multiplicative way using following equations:

$$Score1(h_t, \overline{h_s}) = h_t^T.\overline{h_s} \tag{4}$$

$$\alpha_{ts} = \frac{exp\ (Score(h_t, \overline{h_s})}{\sum_{s'=1}^{S} exp\ (Score(h_t, \overline{h_s})} \tag{5}$$

$$c_{t'} = \sum_s \alpha_{ts}.\overline{h_s} \tag{6}$$

$$Score2(h_t, f_i) = h_t^T.f_i \tag{7}$$

$$\alpha_{is} = \frac{exp\ (Score(h_t, f_i)}{\sum_{s'=1}^{S} exp\ (Score(h_t, f_i)} \tag{8}$$

$$c_i = \sum_s \alpha_{is}.\overline{h_s} \tag{9}$$

$$c_t = c_{t'} + c_i \tag{10}$$

Here score for image and text is calculated separately and context vector for text and image are obtained using equations 4-9. Finally these individual context vectors are concatenated (Equ. 10) to produce final context vector. These vectors are used to predict next target word.

D. Model 4 (Using Both Local and Global Image Features)

We feel that context of object in image can also help to resolve the ambiguity and may play a vital role in selecting appropriate token at target side. So we decided to use the features of whole image and termed it as global features. The global image features are used to initialize the hidden state of encoder. Local image features are used to calculate context vector using dot operation (see figure 4).

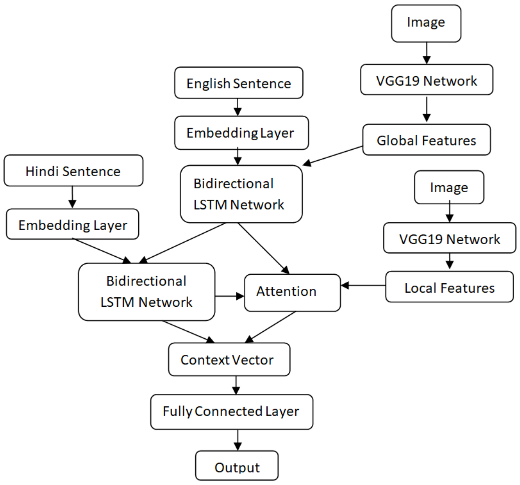


Fig. 4. Architecture for caption generation from text using image feature for attention calculations

E. Model 5 (Using Pre-Trained Embedding)

This model utilizes pre-trained glove 300 vector embedding [1] on encoder side. On decoder side, Fastext 300 embedding [2] for Hindi is used to initialize the embedding. Both side embeddings are further trained on training data. Rest of the settings are same as previous model 4.

F. Model 6 (Bridging Source and Target)

Following Kuang et al. (2018), we can shorten the distance between source and target words and thus strengthen the association, by bridging source and target word embeddings. This model tries to bridge the target side with source side by determining the most likely source word aligned to it and use the word embedding of this source word to support the prediction of the target hidden state of the next target word to be generated. Rest of the settings are same as previous model.

G. Model 7 (Rerank)

Shen et al. (2004), Neubig et al. (2015) and Imamura and Sumita (2017) suggests that re-ranking improves the translation quality. So we also decided to use re-ranking of output of all the above models. It was observed that for different sequences different model produces correct sentences. We use Eng-Hindi Dictionary provided by IIT Bombay[3]. We extracted only English and corresponding Hindi meaning from this lexicon. Besides we use IIT Bombay monolingual Hindi corpus[4] to generate ngram model. The Hindi sentences of training dataset are also used for generating ngrams. We filter the candidate sentences based on bag of words of target language. For every source sentence, we generate a bag of words of target language. Only those sentences from candidates were selected whose words has maximum match in bag of words. Further tie was broken by language model. The remaining sentences were scored by language model using equation:

$$r = \lambda_1 * TProb + \lambda_2 * BProb + \lambda_3 * UProb \tag{11}$$

Where TProb, BProb and UProb are Tri, Bi and Unigram probabilities of sentence and λ_1, λ_2 and λ_3 are weights. We use 0.6, 0.3, 0.1 for λ_1, λ_2 and λ_3 respectively. The sentence getting highest score was selected as final output.

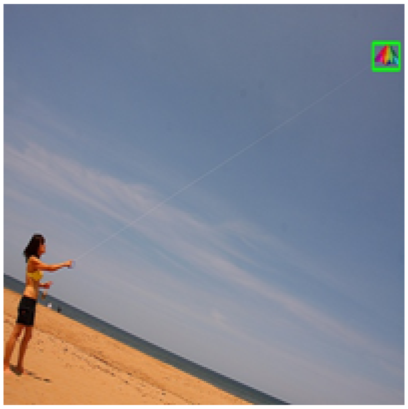
V. Experimental Setup and Results

Two sets of test data termed as Evaluation Set and Challenge Set were provided by Nakazawa et al. (2019). Challenge set consist of selective ambiguous words to make it harder. BLEU score was calculated using nltk toolkit[5]. The results of BLEU-4 are as shown in table I. Results on challenge set are little bit lower than evaluation set attributed to the complex and ambiguous nature of input

[1] https://nlp.stanford.edu/projects/glove/
[2] https://fasttext.cc/docs/en/crawl-vectors.html
[3] http://www.cfilt.iitb.ac.in/wordnet/webhwn/
[4] http://www.cfilt.iitb.ac.in/iitb_parallel/
[5] http://www.nltk.org/api/nltk.translate.html

TABLE I
BLEU Score of Evaluation Set and Challenge Set

Models	Evaluation Set	Challenge Set
Model 1	46.23%	40.93%
Model 2	47.90%	41.28%
Model 3	43.24%	38.86%
Model 4	47.88%	40.62%
Model 5	48.47%	42.04%
Model 6	47.42%	41.59%
Model 7	48.37%	42.52%



Source Caption: Colorful Kite
Reference Given: "कॉलोफ्र्यूल किताब"
Correct output: "रंगीन पतंग"

Fig. 5. Wrong Reference Translation and Correct Output

TABLE II
RIBES Score of Evaluation Set and Challenge Set

Models	Evaluation Set	Challenge Set
Model 5	0.67644	0.487897
Model 7	0.69288	0.507192

TABLE III
AM/FM Score of Evaluation Set and Challenge Set

Models	Evaluation Set	Challenge Set
Model 5	0.707520	0.632060
Model 7	0.722110	0.659840

sentences. In both cases, improvement from baseline was noticed but it is not significant. From results it seems that contribution of image features in selecting correct target word is not significant. Upon observing output manually, we noticed that out put is indeed correct. Reason for low score is just the choice of word which is not in reference text. Training data set is not standardized. There are number of spellings exist in training data for a same word e.g. for word ”zebra” there exist three different variant i.e. "ज़ेब्रा", "ज़ेबरा" and "झेब्रा". If our system outputs any of them but in reference other variant has been used, it cause low BLEU score. Treating all of them as correct, output can be treated as correct. Similarly, lot of words in training data are transliterated instead of translation for example words विंडो, खिडकी for input ”window”. Both translated and transliterated versions are present in training data making hard for system to learn. In some cases, the reference is totally wrong. For example in figure 5, for the caption ”Colorful kite”, the reference translation given was "कॉलोफ्र्यूल किताब " (Colorful book). Here ”Colorful” is transliterated and ”Kite” is wrongly translated as "किताब" which means ”book”. This cause wrong learning of model. These kind of mismatch attributed to low BLEU score. Although the provided reference is wrong, but proposed model is able to produce correct output "रंगीन पतंग" (Colorful Kite). This gives us reasons to believe that image features are helpful in translating source sentence to target. In one of output, for source sentence ”A Glass of Wine”, baseline model produces "शराब का कांच" whereas proposed model produces "शराब का एक गिलास". Clearly choice of words is affected by image features. Here the word ”Glass” is ambiguous. Baseline system which is not using image features select wrong output "कांच" where as model using image feature is able to choose correct word "गिलास". We selected model 5 and model 7 for final evaluation. Besides BLEU, two more metrics i.e. Rank-based Intuitive Bilingual Evaluation Score (RIBES) and Adequacy-Fluency Metric (AM/FM) has been used to check the system performance. RIBES is another automatic translation scoring method which considers word reordering also. AM/FM considers both semantic correctness and grammatical fluency of output. Both scores are given in tables II and III.

Results show that despite low BLEU score, the output of our proposed model is quite adequate and fluent. Human evaluation also confirms this fact. Direct Assessment(DA) method as suggested by GRAHAM et al. (2017) has been used for human evaluation. Three annotators were asked to assign a score from 0 to 100 to each candidate. The main benefit of bilingual evaluation is that the reference is not needed for the evaluation. The evaluators are shown both the image and the source English text. The evaluators are asked to indicate to what extent the meaning is preserved. The collected DA scores are averaged for each model. The results are shown in table IV.

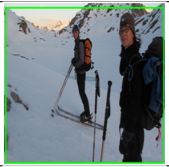
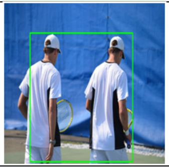
TABLE IV
Human Evaluation Result of Evaluation Set and Challenge Set

Models	Evaluation Set	Challenge Set
Model 5	60.22	47.06
Model 7	62.42	48.06

VI. Discussion

Despite low BLEU score, it has been observed that output of proposed model is quite good. Some sample outputs are shown in table V. The reference translation is either wrong or not of good quality. For example पानी का एक बर्तन is better choice than कांच in first case. Similarly reference of second case is not correct but proposed model is able to use the cues from image to produce a fluent and correct sentence. Third reference literally means ”Two people are walking on mountain”. As it can be seen in image clearly that they are not walking but posing for a

TABLE V
Some Sample Translations

	Source Caption: a water glass on a table Reference Given: "एक मेज पर एक कांच" Correct output: "एक मेज पर पानी का एक बर्तन"
	Source Caption: a girl playing tennis Reference Given: "लड़की टेनिस खेलने " Correct output: "एक लड़की टेनिस खेल रही है"
	Source Caption: two people hiking mountain Reference Given: "दो व्यक्ति पहाड पर पैदल चल रहे है" Correct output: "दो व्यक्ति स्कीइंग कर रहे हैं"
	Source Caption: there are two players in the court Reference Given:"क़ोर्ट में दो खिलाड़ी हैं" Correct output:"कोर्ट में टेनिस खिलाड़ी हैं"

picture standing still. Skees can be seen clearly in given picture. Taking cues from the image our model is able to predict that two people are skiing and accordingly producing the output "दो व्यक्ति स्कीइंग कर रहे हैं" which is not wrong. Failing case are those which have no or few instances in training data like "date and time of photo" or "date stamp of photograph". These kinds of sentences are in challenge set but not in training set. So our model fails to translate them properly.

VII. Conclusion

Sequence to Sequence models have been successfully applied for machine translation task. This modal is further extended to make it multi modal system where both features of image and text are combined together to translate a source language to target language. This paper discuss our model for shared task on multi modal machine translation from English to Hindi. Features of images are extracted using VGG19 Network. The given captions represent some part of image which is described by bounding box. Features of whole image are extracted and termed as global feature. Feature of bounding box image were extracted separately and termed as local features. Various models have been explored for the task. The best model produce BLEU score of 48.37 whereas RIBES and AM/FM score are 0.69288 and 0.722110 for evaluation set and 0.507192 and 0.659840 for challenge set. Although BLEU score is not significant but RIBES and AM/FM score shows that the output of our model is quite correct and fluent. The low score is due to number of reasons like wrong reference translation or multiple versions of spellings etc. In future, we would like to check the performance with other networks for image feature extraction like Resnet50. Improving on training data set is one area to be focused in future. Besides, spelling normalization may also improve the results. Character level features can also be explored in future.

Acknowledgment

We acknowledge the organizers of WAT 2019 for providing data and supporting evaluation process to obtain results for this paper.

References

Çağlayan, O., Aransa, W., Bardet, A., Garcia Martinez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., Weijer, J., 2017. Lium-cvc submissions for wmt17 multimodal translation task, in: Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pp. 432–439.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., Frank, S., 2018. Findings of the Third Shared Task on Multimodal Machine Translation, in: THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18), Brussels, Belgium. pp. 308 – 327. URL: https://hal.archives-ouvertes.fr/hal-02008843, doi:10.18653/v1/W18-6402.

Bernardi, R., Çakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B., 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. CoRR abs/1601.03896. URL: http://arxiv.org/abs/1601.03896, arXiv:1601.03896.

Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., van de Weijer, J., 2016. Does multimodality help human and machine for translation and image captioning?, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 627–633. URL: http://www.aclweb.org/anthology/W/W16/W16-2358.

Calixto, I., Elliott, D., Frank, S., 2016. Dcu-uva multimodal mt system report, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 634–638. URL: http://www.aclweb.org/anthology/W/W16/W16-2359.

Elliott, D., Frank, S., Barrault, L., Bougares, F., Specia, L., 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. CoRR abs/1710.07177. URL: http://arxiv.org/abs/1710.07177, arXiv:1710.07177.

Elliott, D., Frank, S., Hasler, E., 2015. Multi-language image description with neural sequence models. CoRR abs/1510.04709. URL: http://arxiv.org/abs/1510.04709, arXiv:1510.04709.

Feng, Y., Lapata, M., 2010. Topic models for image annotation and text illustration, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational

Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 831–839. URL: http://dl.acm.org/citation.cfm?id=1857999.1858124.

GRAHAM, Y., BALDWIN, T., MOFFAT, A., ZOBEL, J., 2017. Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering 23, 3–30. doi:10.1017/S1351324915000339.

Grönroos, S., Huet, B., Kurimo, M., Laaksonen, J., Mérialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., Vázquez, R., 2018. The memad submission to the WMT18 multimodal translation task. CoRR abs/1808.10802. URL: http://arxiv.org/abs/1808.10802, arXiv:1808.10802.

Hitschler, J., Schamoni, S., Riezler, S., 2016. Multimodal pivots for image caption translation, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany. pp. 2399–2409. URL: https://www.aclweb.org/anthology/P16-1227, doi:10.18653/v1/P16-1227.

Hokamp, C., Calixto, I., 2016. Multi-modal neural machine translation using minimumrisk training. URL: https://www.github.com/chrishokamp/multimodal_nmt.

Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., 2019. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys 51, 1–36. doi:10.1145/3295748.

Huang, P.Y., Liu, F., Shiang, S.R., Oh, J., Dyer, C., 2016. Attention-based multimodal neural machine translation, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 639–645. URL: http://www.aclweb.org/anthology/W/W16/W16-2360.

Imamura, K., Sumita, E., 2017. Ensemble and reranking: Using multiple models in the NICT-2 neural machine translation system at WAT2017, in: Proceedings of the 4th Workshop on Asian Translation (WAT2017), Asian Federation of Natural Language Processing, Taipei, Taiwan. pp. 127–134. URL: https://www.aclweb.org/anthology/W17-5711.

Karpathy, A., Li, F., 2014. Deep visual-semantic alignments for generating image descriptions. CoRR abs/1412.2306. URL: http://arxiv.org/abs/1412.2306, arXiv:1412.2306.

Kuang, S., Li, J., Branco, A., Luo, W., Xiong, D., 2018. Attention focusing for neural machine translation by bridging source and target embeddings, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia. pp. 1767–1776. URL: https://www.aclweb.org/anthology/P18-1164, doi:10.18653/v1/P18-1164.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L., 2013. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2891–2903. doi:10.1109/TPAMI.2012.162.

Laskar, S.R., Singh, R.P., Pakray, P., Bandyopadhyay, S., 2019. English to Hindi multi-modal neural machine translation and Hindi image captioning, in: Proceedings of the 6th Workshop on Asian Translation, Association for Computational Linguistics, Hong Kong, China. pp. 62–67. doi:10.18653/v1/D19-5205.

Libovický, J., Helcl, J., Tlustý, M., Bojar, O., Pecina, P., 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 646–654. URL: http://www.aclweb.org/anthology/W/W16/W16-2361.

Liu, C., Wang, C., Sun, F., Rui, Y., 2016. Image2text: A multimodal caption generator, in: ACM international conference on Multimedia (ACM MM). URL: https://www.microsoft.com/en-us/research/publication/image2text-a-multimodal-caption-generator/.

Lucia, S., S, F., K, S., D, E., 2016. A shared task on multimodal machine translation and crosslingual image description, in: in Proceedings of the First Conference on Machine Translation, WMT2016, colocated with ACL 2016, Association for Computational Linguistics(ACL), Berlin, Germany. pp. 543–553.

Luong, M., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. CoRR abs/1508.04025. URL: http://arxiv.org/abs/1508.04025, arXiv:1508.04025.

Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W.P., Kunchukuttan, A., Parida, S., Bojar, O., Kurohashi, S., 2019. Overview of the 6th workshop on Asian translation, in: Proceedings of the 6th Workshop on Asian Translation, Association for Computational Linguistics, Hong Kong, China. pp. 1–35. URL: https://www.aclweb.org/anthology/D19-5201, doi:10.18653/v1/D19-5201.

Neubig, G., Morishita, M., Nakamura, S., 2015. Neural reranking improves subjective quality of machine translation: NAIST at WAT2015, in: Proceedings of the 2nd Workshop on Asian Translation (WAT2015), Workshop on Asian Translation, Kyoto, Japan. pp. 35–41. URL: https://www.aclweb.org/anthology/W15-5003.

Parida, S., Bojar, O., Dash, S.R., 2019. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. Computación y Sistemas In print. Presented at CICLing 2019, La Rochelle, France.

Rodríguez Guasch, S., Costa-jussà, M.R., 2016. Wmt 2016 multimodal translation system description based on bidirectional recurrent neural networks with double-embeddings, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 655–659. URL: http://www.aclweb.org/anthology/W/W16/W16-2362.

Sanayai Meetei, L., Singh, T.D., Bandyopadhyay, S., 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset, in: Proceedings of the 6th Workshop on Asian Translation, Association for Computational

Linguistics, Hong Kong, China. pp. 181–188. doi:10.18653/v1/D19-5224.

Shah, K., Wang, J., Specia, L., 2016. Shef-multimodal: Grounding machine translation on images, in: Proceedings of the First Conference on Machine Translation, Association for Computational Linguistics, Berlin, Germany. pp. 660–665. URL: http://www.aclweb.org/anthology/W/W16/W16-2363.

Shen, L., Sarkar, A., Och, F.J., 2004. Discriminative reranking for machine translation, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA. pp. 177–184. URL: https://www.aclweb.org/anthology/N04-1023.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. CoRR abs/1409.3215. URL: http://arxiv.org/abs/1409.3215, arXiv:1409.3215.

Tanti, M., Gatt, A., Camilleri, K.P., 2017a. What is the role of recurrent neural networks (rnns) in an image caption generator? CoRR abs/1708.02043. URL: http://arxiv.org/abs/1708.02043, arXiv:1708.02043.

Tanti, M., Gatt, A., Camilleri, K.P., 2017b. Where to put the image in an image caption generator. CoRR abs/1703.09137. URL: http://arxiv.org/abs/1703.09137, arXiv:1703.09137.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2014. Show and tell: A neural image caption generator. CoRR abs/1411.4555. URL: http://arxiv.org/abs/1411.4555, arXiv:1411.4555.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. CoRR abs/1502.03044. URL: http://arxiv.org/abs/1502.03044, arXiv:1502.03044.

Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J., 2015. Stacked attention networks for image question answering. CoRR abs/1511.02274. URL: http://arxiv.org/abs/1511.02274, arXiv:1511.02274.

You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. CoRR abs/1603.03925. URL: http://arxiv.org/abs/1603.03925, arXiv:1603.03925.