WAT2019: English-Hindi Translation on Hindi Visual Genome Dataset

L. Sanayai Meetei
NIT Silchar
loisanayai@gmail.com

Thoudam Doren Singh
NIT Silchar
thoudam.doren@gmail.com

Sivaji Bandyopadhyay
NIT Silchar
sivaji.cse.ju@gmail.com

Abstract

A multimodal translation is a task of translating a source language to a target language with the help of a parallel text corpus paired with images that represent the contextual details of the text. In this paper, we carried out an extensive comparison to evaluate the benefits of using a multimodal approach on translating text in English to a low resource language, Hindi as a part of WAT2019 (Nakazawa et al., 2019) shared task. We carried out the translation of English to Hindi in three separate tasks with both the evaluation and challenge dataset. First, by using only the parallel text corpora, then through an image caption generation approach and, finally with the multimodal approach. Our experiment shows a significant improvement in the translation with the multimodal approach than the other approach.

1 Introduction

Hindi is the lingua franca in the Hindi belt of India, written in the Devanagari script, an abugida. It consists of 11 vowels and 33 consonants. Both Hindi and English belong to the same language family, Indo-European, but follows different word order. Hindi follows the Subject Object Verb (SOV) order while English follows the Subject Verb Object (SVO) order.

In addition to communication, learning a language covers a lot more things. It spreads culture, traditions, and conventions. A machine translation(MT) is the process of automatically generating a target human language from a source human language. With big companies such as Google offering decent translation to most of the high resource languages, interlingual communication becomes easy. The application of machine translation, can also be applied in our daily healthcare services (Wołk and Marasek, 2015; Yellowlees et al., 2015), government services, disaster management, etc. The methodology of machine translation system where the traditional statistical machine translation (SMT) (Koehn et al., 2007) is replaced by the neural machine translation (NMT) system, a MT system based on artificial neural network proposed by (Kalchbrenner and Blunsom, 2013), results to a better translation. Using deep learning and representation learning, NMT translate a source text to a target text. In the encoder-decoder model of NMT (Cho et al., 2014), the encoder encodes the input text into a fixed length of input vector and the decoder generates a sequence of words as the output text from the input vector. The system is reported to learn the linguistic regularities of both at the phrase level and word level. With the advancement in Computer Vision, the work on generating caption of an image is becoming popular. In an image caption generation model, a deep neural network based model is used to extract the features from the image, the features are then translated to a natural text using a language model.

Recently, research work on incorporating the features extracted from the image along with the parallel text corpora in a multimodal machine translation(MMT) is carried out in many shared translation task. The impact of combining the visual context in the MMT system has shown an increase in the robustness of machine translation (Caglayan et al., 2019). As a part of the shared task WAT2019, the main objective of our task is carry out the translation of English to Hindi. The remaining of this paper is structured as follows: Sec-

tion 2 describe the related works, Section 3 illustrate the system architecture used in our model. Section 4 and Section 5 discuss the experimental setup and the result analysis respectively. Finally, concluding with our findings and the future scope of the work in Section 6.

2 Literature Review

With the introduction of neural machine translation, many approaches of the NMT model is carried out to improve the performance. Initially, because of the use of a fixed-length input vector, the encoder-decoder model of NMT suffers during the translation of long text. By introducing an attention mechanism (Bahdanau et al., 2014), the source text is no longer encoded into a fixed-length vector. Rather, the decoder attends to different parts of the source text at each step of the output generation. In their experiment (Bahdanau et al., 2014) of English to French translation task, the attention mechanism is observed to improve the translation performance of long input sentences.

The NMT translation of English to Hindi is carried out by (Mahata et al., 2019; Singh et al., 2017). Mahata et al. (2019) evaluate the performance of NMT model over the SMT system as a part of MTIL2017[1] shared task. The author reported that NMT performs better in short sentences while SMT outperforms NMT in translating longer sentences.

Sennrich et al. (2015) introduced an effective approach of preprocessing for NMT task where the text is segmented into subword units. The NMT model supports open-vocabulary translation where sequences of subword units encoded from the rare and unknown words are used. The proposed approach is reported to perform better than the back-off to a dictionary look-up (Luong et al., 2014) in resolving the out of vocabulary translation problem.

An automatic image caption generation system is a system that generates a piece of text that describes an input image. Kiros et al. (2014) introduced a multimodal neural network based image caption generation model. The model makes use of word representations and image features learned from deep neural networks. In the work by Vinyals et al. (2015), the authors proposed a neural and probabilistic framework for image caption generation system consisting of a vision Convolution Neural Network (CNN) followed by a language generating Recurrent Neural Network(RNN) trained to increase the likelihood of the generated caption text.

Calixto et al. (2017) reported a research work on various multimodal neural machine translation (MNMT) models by incorporating global features extracted from the image into attention based NMT. The author also evaluated the impact of adding synthetic multi-modal, multilingual data generated using phrase-based statistical machine translation(PBSMT) trained on the dataset from Multi30k (Elliott et al., 2016). The model where the image is used to initialize the encoder hidden state is observed to perform better than the other models in their experiment. The research work of MNMT for Hindi is very recent. Koel et al. (2018) report a MNMT work on English to Hindi translation by building a synthetic dataset generated using a phrase based machine translation system on a Flickr30k (Plummer et al., 2017) dataset.

3 System Architecture

In our model, the dataset from the Hindi Visual Genome[2] are used for three separate tasks: 1) Translation of English-Hindi using only the text dataset, 2) Generate the captions from the image, 3) Multimodal translation of English-Hindi using the image and the parallel text corpus. Figure 1 shows a brief representation of our working model. Following of this section illustrates the details of the dataset, the various methods used in our implementation for the three tasks.

3.1 Dataset

Hindi Visual Genome, HVG: The dataset used in our work is from the HVG (Parida et al., 2019) as a part of WAT2019 Multi-Modal Translation Task[3] . The dataset consists of a total of 31525 randomly selected images from Visual Genome (Krishna et al.,

---

[1] https://nlp.amrita.edu/mtil_cen/

[2] https://ufal.mff.cuni.cz/ hindi-visual-genome/

[3] https://ufal.mff.cuni.cz/ hindi-visual-genome/wat-2019-multimodal-task
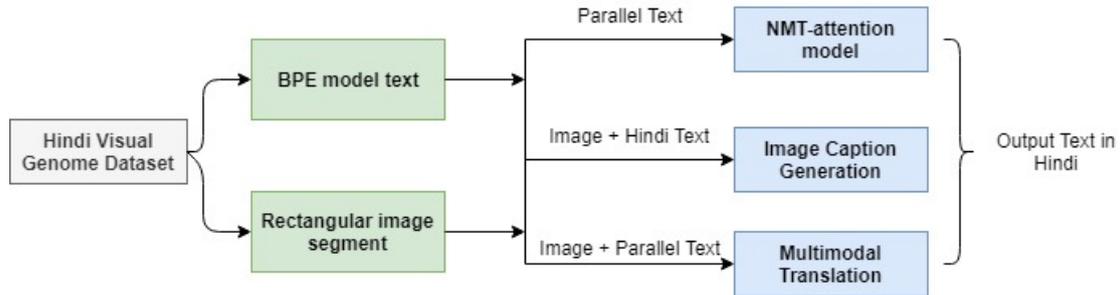
Figure 1: System Architecture

Dataset distribution	Items
Training set	28932
Development set	998
Evaluation set	1595
Challenge set	1400

Table 1: Hindi Visual Genome dataset details.

2017) and a parallel image caption corpus in English-Hindi for selected image segments. The details of the HVG corpus is shown in Table 1. Each item in Table 1 comprises of a source text in English, its translation in Hindi, the image and a rectangular region in the image. The text dataset represent the caption of the rectangular image segment.

3.2 Byte Pair Encoding (BPE)

BPE, a data compression technique proposed by Gage (1994) iteratively replaces the common pairs of bytes in a sequence with a single, unused byte. To handle an open vocabulary problem, we followed the word segmentation algorithm described at (Sennrich et al., 2015) where characters or character sequences are merged instead of common pairs of bytes. For example, the word "booked" is split into "book" and "ed", while "booking" is split into "book" and "ing". The resulting tokens or character sequences allows the model to generalize to new words. The method also reduces the overall vocabulary.

3.3 Neural Machine Translation

The neural machine translation uses RNN encoders and decoders where an encoder maps the input text to an input vector then a decoder decodes the vector into the output text. Following the attention mechanism of (Bahdanau et al., 2014), a bidirectional RNN in the encoder and, an alignment model paired with a LSTM in the decoder model is used.
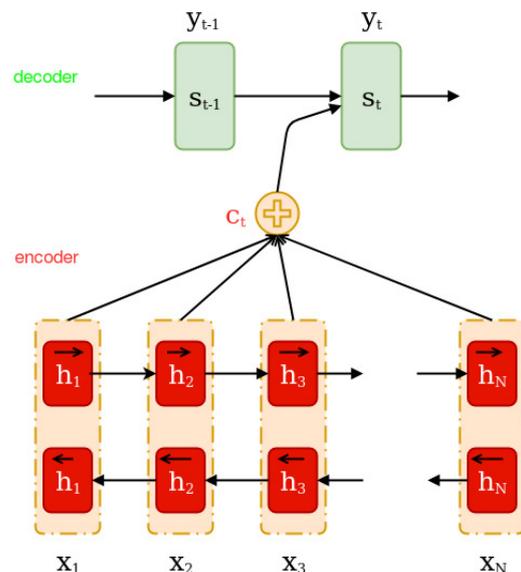
Figure 2: Neural Machine Translation model with attention mechanism

Figure 2 illustrate the attention model trying to generate the *t-th* target word y_t from a source sentence $(x_1, x_2, .., x_N)$ where the forward RNN encoder generates a forward annotation vectors sequence $(\vec{h}_1, \vec{h}_2,...,\vec{h}_N)$ and the backward RNN encoder generates a backward annotation vectors sequence $(\overleftarrow{h}_1, \overleftarrow{h}_2,...,\overleftarrow{h}_N)$. The concatenation of the two vectors gives the annotation vector at the time step i, as $h_i = [\vec{h}_i; \overleftarrow{h}_i]$. The attention mechanism learns where to place attention on the input sequence as each word of the output sequence is decoded.

3.4 Image Caption Generation

With the hypothesis of CNN drawn from human visual handling framework, CNN provides a set of hierarchical filtering on image.

CNN in the end is able to extract latent features that represents a semantic meaning to the image. The combination of CNN with RNN makes use of the spatial and temporal features. A neural network based caption generator for an image using CNN model followed by RNN with BEAM Search(BS) for generating the language (Vinyals et al., 2015) is used in our system.
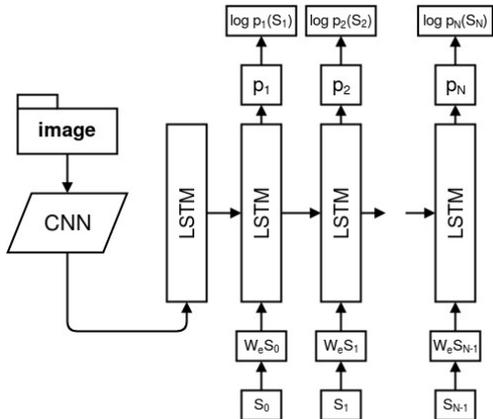
Figure 3: Image caption generation model

Figure 3 shows the LSTM model combined with a CNN image embedder and word embeddings. To predict each word of the sentence, the LSTM model is trained with the image and all preceding words as defined by $p(S_t|I, S_0, \ldots, S_{t-1})$. For an input image I and a caption description, $S = (S_0, \ldots, S_N)$ of I, the unrolling procedure of LSTM (Vinyals et al., 2015) is shown in the following equation:

$$x_{-1} = \text{CNN}(I) \qquad (1)$$
$$x_t = W_e S_t, \quad t \in \{0 \ldots N-1\} \qquad (2)$$
$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \ldots N-1\} \qquad (3)$$

A one-hot vector S_t of dimension equal to the size of the dictionary represent each word. A special start word, S_0 and a special stop word, S_N is used to mark the start and end of the sentence. The image with vision CNN and words by word embedding W_e are mapped to the same space as shown in Equation 1 and Equation 2 respectively. At instance $t = -1$, the image I is fed only once to deliver LSTM the content of the image. To generate the image caption, the BS iteratively examine the k best sentences up to time t as candidates for generating sentences of size $t+1$, keeping only the best k resulting from them.

3.5 Multimodal Machine Translation

In MMT, the image paired with the parallel text corpus is used to train the system. Using the multimodal neural machine translation (MNMT) model (Calixto et al., 2017), global features are extracted using a deep CNN based models.
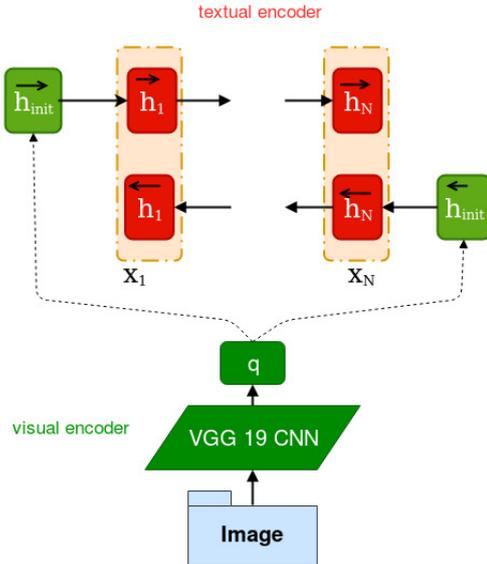
Figure 4: Multimodal translation model using image to initialize the hidden state of encoder

Using the global image feature vector ($q \in \mathbb{R}^{4096}$), a vector d is computed as follows:

$$d = W_I^2 \cdot (W_I^1 \cdot q + b_I^1) + b_I^2 \qquad (4)$$

where W and b are image transformation matrices and bias vector respectively.

With bidirectional RNN at the encoder, the features are used to initialize the hidden states of the encoder. As shown in Figure 4, two new single-layer feed-forward networks are used to initialize the states of forward and backward RNN rather than initializing encoder hidden states with $\vec{0}$ (Bahdanau et al., 2014) as:

$$\vec{h}_{init} = \tanh(W_f d + b_f) \qquad (5)$$

$$\bar{h}_{init} = \tanh(W_b d + b_b) \qquad (6)$$

with W_f and W_b as the multi-modal projection matrices that project the image features d into the encoder forward and backward hidden states dimensionality, respectively, and b_f and b_b as bias vectors.

4 Experimental Setup

The translation of English to Hindi on the HVG dataset is evaluated in three separate tasks:

- Using only the text dataset.

- Using only the image dataset.

- Using both the image and the text dataset.

To carry out the experiment, the dataset from the HVG is processed as described in the following Subsection 4.1.

4.1 Dataset Preparation

Text: The text dataset is processed into a BPE format as describe in Subsection 3.2. The encoding-decoding of the text dataset to and from subword units is carried out using the open-source tool[4].

Example:

Raw text: outdoor blue mailbox receptacle

After processing: outdoor blue ma@@ il@@ box re@@ ce@@ p@@ ta@@ cle

Image: The image and description (English-Hindi pair) in HVG dataset are structured in such a format that, the caption describes only a selected rectangular portion of the image. With the image coordinates (X, Y. Width, Height) provided in the HVG dataset, the rectangular image segment from the original image is cropped as a part of processing. A sample is shown below in Figure 5.

(a)   (b)

Figure 5: (a) A sample image
(b) Image segment from (a) with *English caption:* woman with sunglasses holding a cellphone, *Hindi caption:* सेलफोन पकड़ने वाली स्त्री

---

[4]https://github.com/rsennrich/subword-nmt

With the model described in Section 3, the experimental setup for each of the three tasks are explained in the Subsections below.

4.2 NMT Text only Translation

Using the processed text data from Subsection 4.1, the translation of English-Hindi is carried out on a neural machine translation open-source tool based on OpenNMT (Klein et al., 2017). We used the attention mechanism of (Bahdanau et al., 2014). Along with other parameters such as learning rate at 0.002, Adam optimizer (Kingma and Ba, 2014), a dropout rate of 0.1, we train the system for 25 epoch.

4.3 Image Caption Generation

Our second task is to generate the caption of an image in Hindi. For this task, we trained our system (Subsection 3.4) with the processed images from Subsection 4.1 paired with its Hindi captions. For extracting the features from the image a 16-layer VGG (VGG16) model (Simonyan and Zisserman, 2014), pre-trained on the ImageNet dataset, is used. A 4096-dimensional vector generated by the VGG16 for each image is then fed to RNN Model with BEAM search. With BEAM search parameter set to three (number of words to consider at a time), the system is trained for 20 epoch.

4.4 Multimodal Translation

In our final task of multimodal translation of English to Hindi, the processed text and image dataset from Subsection 4.1 are fed into our model (Subsection 3.5). A pre-trained model, VGG19-CNN, is employed to extract the global features from the image. The system is trained for 30 epoch with a learning rate set to 0.002, dropout rate of 0.3 and using Adam optimizer.

5 Results and analysis

As a part of the Hindi Visual Genome (WAT2019 Multi-Modal Translation Task) shared task, we submitted in all the three task: 1) Text-only translation, 2) Hindi-only image captioning and 3) Multi-modal translation (uses both the image and the text), for the two types dataset (Parida et al., 2019): the Evaluation Test Set and the Challenge Test

Set . The experiment for the three tasks is carried out separately on both the test dataset.

Evaluation metrics: The evaluation of the translation system is carried out using three different techniques: AFMF (Banchs et al., 2015), BLEU (Papineni et al., 2002) score and RIBES (Isozaki et al., 2010).

Task	BLEU	RIBES	AMFM
TOT	20.13	0.57	0.61
HIC	2.59	0.15	0.41
MMT	28.45	0.63	0.68

Table 2: Results obtained in Evaluation Test Set.

Task	BLEU	RIBES	AMFM
TOT	5.56	0.37	0.46
HIC	0.00	0.08	0.38
MMT	12.58	0.48	0.55

Table 3: Results obtained in Challenge Test Set.

Table 2 and Table 3 shows the scores obtained by our system on the Evaluation Test Set and Challenge Test Set respectively. In Table 2 and Table 3, TOT, HIC, and MMT represents the text-only translation sub task system, automatic image caption generation system of Hindi-only image captioning sub task and the multi-modal translation (using both the image and the text) sub task system respectively. Three sample inputs with the different forms of an ambiguous word "stand" from the challenge test set and their outputs are shown in Table 4, Table 5 and Table 6.

From the above observations, we see that the results of multimodal translation outperforms the other methods. However, the evaluation of image caption generation is reported to achieve poor score. Reason being the evaluation metric used rely on the surface-form similarity or simply match n-gram overlap between the output text and the reference text, which fails to evaluate the semantic information describe by the generated text. Also, an image can be interpreted with different captions to express the main theme contained in the image. Hence, the poor performance report even though the generated caption text for the input image is observe to show reasonable quality of adequacy and fluency on ran-

dom human evaluation. We can conclude that, for the case of image caption generation, there is a need for a different type of evaluation metrics.

6 Conclusion and Future Work

In this paper, we reported the evaluation of English-Hindi translation with different approaches as a part of WAT2019 shared task. It is observed that the multimodal approach of incorporating the visual features paired with text data gives significant improvement in translation than the other approaches. We also conclude that the same evaluation metrics used for the machine translation is not applicable to the automatic caption generation system, as the latter approach provides a good adequacy and fluency to the output text. In the future, we would like to investigate the impact of adding features in the BPE model. Furthermore, evaluating the system on a larger size of the dataset might give us more insight into the feasibility of the system in the real world applications.

Acknowledgments

References

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

Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521*.

Input Image and Text	Reference and Output by different Model Types
  Woman standing on tennis court	**Reference:** टेनिस कोर्ट पर खड़ी महिला Transliteration: tenis kort par khadee mahila **TOT:** टेनिस कोर्ट पर मनुष्य Transliteration: tenis kort par manushy *Translation: A man on a tennis court* **HIC:** एक व्यक्ति टेनिस खेल रहा है Transliteration: ek vyakti tenis khel raha hai *Translation: A person playing tennis* **MMT:** टेनिस कोर्ट पर खड़ी महिला Transliteration: tenis kort par khadee mahila *Translation: A woman standing on a tennis court*

Table 4: Sample 1 input and output

Input Image and Text	Reference and Output by different Model Types
  man stand on skateboard	**Reference:** आदमी स्केटबोर्ड पर खड़ा है Transliteration: aadmee sketabord par khada hai **TOT:** स्केटबोर्ड पर मनुष्य Transliteration: sketabord par manushy *Translation: Man on skateboard* **HIC:** व्यक्ति एक स्केटबोर्ड पर Transliteration: vyakti ek sketabord par *Translation: A person on a skateboard* **MMT:** व्यक्ति स्केटबोर्ड पर खड़ा है Transliteration: vyakti sketabord par khada hai *Translation: A person standing on a skateboard*

Table 5: Sample 2 input and output

Input Image and Text	Reference and Output by different Model Types
  A big tv on a stand	**Reference:** एक स्टैंड पर एक बड़ा टीवी Transliteration: ek staind par ek bada teevee **TOT:** एक सेलफोन पर एक बड़ा सा वैन Transliteration: ek selaphon par ek bada sa vain *Translation: A big van on a cellphone* **HIC:** इमारत के किनारे पर एक दीवार Transliteration: imaarat ke kinaare par ek deevaar *Translation: A wall on the side of the building* **MMT:** एक स्टैंड पर एक बड़ा टीवी Transliteration: ek staind par ek bada teevee *Translation: A big tv on a stand*

Table 6: Sample 3 input and output

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

N. Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, N. Federico, M.and Bertoldi, B. Cowan, W. Shen, R. Moran, C.and Zens, and C. Dyer. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. 2014. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2019. Mtil2017: Machine translation using recurrent neural network on statistical machine translation. *Journal of Intelligent Systems*, 28(3):447–453.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, 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*, Hong Kong. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *arXiv preprint arXiv:1907.08948*.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2017. Comparing recurrent and convolutional architectures for english-hindi neural machine translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 167–170.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Krzysztof Wołk and Krzysztof Marasek. 2015. Neural-based machine translation for medical text domain. based on european medicines agency leaflet texts. *Procedia Computer Science*, 64:2–9.

Peter Yellowlees, Steven Richard Chan, and Michelle Burke Parish. 2015. The hybrid doctor–patient relationship in the age of technology–telepsychiatry consultations and the use of virtual space. *International Review of Psychiatry*, 27(6):476–489.