# Nepali Image Captioning

Aashish Adhikari†
*School of EECS*
*Oregon State University*
Corvallis, Oregon
adhikara@oregonstate.edu

Sushil Ghimire†
*Dept. of Computer Engineering, Pashchimanchal Campus*
*Tribhuvan University, Institute Of Engineering*
Pokhara, Nepal
72bct647.sushil@wrc.edu.np

***Abstract*—Neural networks have seen a surge in applications ranging image recognition, time series prediction, and image captioning among others. The task of image captioning combines two fields of machine learning - computer vision and natural language processing. Several works have been done for image captioning in dominant languages like English and Mandarin with impressive results. The challenge, however, increases for Nepali, a language with a complex grammatical structure. With inherently complex grammar, inputting an image and generating the description of the image in Nepali with correct grammar is hard to achieve. Also, a standard data set does not exist for such tasks in Nepali. This work on image captioning in Nepali is the first of its kind that establishes a baseline for tasks of this sort and aims to encourage other researchers to pursue this line of research. For this, it makes public the data set that was generated during this process so as to provide a standard data set for future works. This work builds on top of the model proposed by [24] and generates image descriptions in Nepali. Describing the contents of an image in Nepali has several applications. Additionally, this work serves as a gateway to more challenging problems such as a video descriptor system and image search for search engines in Nepali. This work utilizes two encoder-decoder architectures, one with visual attention and another without visual attention. We empirically show the loss and perplexity of model performances using different optimizers. The captions generated are agreeable and coherent with the images in general and leave room for improvements in the future.**

## I. Introduction

Due to the advances in training a neural network [26] and the availability of large-scale data sets for different tasks [25], [27], there has been a surge in research for several deep learning applications including image captioning in recent years. Generating a caption that describes an image in a sensible manner is key to scene understanding in artificial intelligence. As such, image captioning is the process of extracting coherent features of an image using an image-based model and describing that image in a natural language using a language-based model that makes use of the features extracted by the image-based model.

Nepali is spoken by 16 million native speakers and nearly 25 million people worldwide [7]. Describing the content of an image in Nepali has several potential applications in Nepal including assisting visually-impaired to describe the contents of an image, image-based search for search engines, and real-time anomaly detection among others. Moreover, the world is rapidly moving towards making cities smart with the likes of several ambitious plans including the "Smart Cities Mission"" [28] in India and the smart city initiative taken by Pokhara metropolitan city [29], the largest metropolitan in Nepal. To this end, visually describing contents of an image sees several key applications in a smart city - providing better customer service to differently abled people, automatically translating the contents of an image to foreigners at key tourist spots, and many more. However, with inherently complex grammatical structure, inputting an image and generating the description of an object in Nepali with correct grammar is harder to achieve. Image captioning for a complex language is tricky in comparison to languages like English and several challenges exist. This work serves as a baseline upon which other researchers can build in the future and compare for improvements.

Hindi and Bangla are some of the closest languages to Nepali in terms of grammatical structure among other languages upon which image captioning has been tried [30], [31]. Image captioning in Nepali has not been tried before in the midst of the complexity in grammar as well as a lack of data set required unlike English with comparatively easier grammatical structure and several rich data sets including MS-COCO [25] and Flickr30k [32]. To overcome this, the authors have manually generated Nepali captions from the MS-COCO data set. Looking at the richness of potentials that arise from image captioning in Nepali language, the authors make use of two models to produce content descriptions in Nepali and analyze the results.

This paper stands as a baseline for image description generator in Nepali upon which new researches can be conducted. Further, it opens doors to similar but more challenging problems including video description generation in Nepali. Another contribution of this work is that the data set compiled and curated during this work has been released for public research purpose. Researchers can use the data set and better the research in this direction.

What follows hereafter is a description of the relevant work done in the literature. Then we explain the models deployed in this work which is followed by the training regime and the qualitative and the quantitative results obtained. Finally, we conclude by explaining the limitations, applications, and possible extensions that can be made in this domain.

†Both authors have equal contributions.

## II. Related Work

In the beginning, there existed two dominant methods for caption generation [3] - either generating caption templates which were filled in based on the results of object detection and attribute discovery [33], [34] or retrieving similar-captioned images from a large database, then modifying these retrieved captions to fit the query [35]. However, these methods have gone out of standard practice and different methods have been proposed to generate image captions since the first neural network model was proposed by [36] for this purpose, many of which are based on the use of recurrent neural networks (RNNs) [37]. Specifically, the use of a special kind of recurrent neural network known as Long Short Term Memory (LSTM) network [38] has allowed promising results in this domain. An LSTM is a kind of a recurrent neural network with an addition of a gating mechanism that sets it apart from traditional recurrent neural networks. It has the ability to preserve long-term memory [8] and thus addresses the short-term memory problem of a typical RNN [38].This is especially important in tasks like natural language processing (NLP) [4]–[6], object detection [8], [9], and machine translation among others. The dominant sequence translation models are based on complex convolutional and recurrent neural networks in an encoder-decoder configuration [40], [41].

One class of works in image captioning is inspired by machine translation [42]–[44]. While several works used attention mechanism [11], [45] was the first to use visual attention for image captioning. Attention mechanism in encoder-decoder architecture shows promising results in image captioning [11] for English language data sets.

Yesko barerma ramro sanga janna parxa, especially attention mechnism vneko k ho vanne ma.

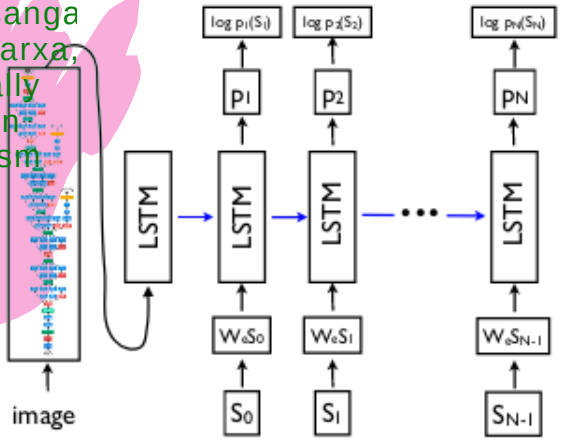


Figure 1:Encoder-decoder architecture for image captioning [52]

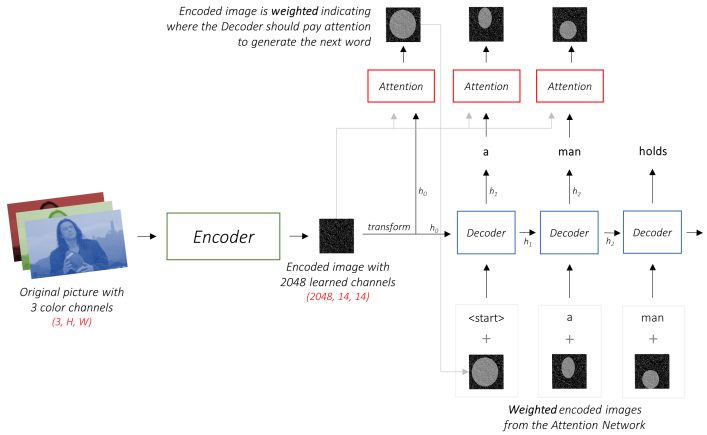


Figure 2:Encoder-decoder architecture with attention for image captioning [53]

## III. Model Architectures

In this paper, we propose two models for Nepali image descriptor, model A and model B. Model A uses a plain encoder-decoder architecture where the encoder is a pre-trained ResNet model [46] and the decoder is a plain LSTM network. Model B is a Show, Attend, and Tell model with Visual attention [24] architecture.

Bujina ni dost!

### A. Model A : Plain Encoder-Decoder

As shown in figure 1, when an image is fed into an image-based model, it produces a vector representing different visual features of that image which along with individual words in the caption are fed into an LSTM for training. For this model, the encoder is a ResNet-50 model with pre-trained weights. Every input image is transformed into a tensor of 224✕222 with RGB channels. Each image is further horizontally flipped with the probability of 0.5 and normalized. The decoder is a plain LSTM sequence-to-sequence model. The choice of LSTM instead of plain RNN or GRU [47] comes from an empirical evaluation through experiments performed. The LSTM has the embedding size of 512 with every intermediate encoder hidden state size of 512. Table 1 summarizes the configuration for Model A's training.

| | |
|---|---|
| Batch size | 128 |
| Number of epochs | 5 |
| Embedding size | 512 |
| Hidden size | 512 |
| Minimum words in caption | 5 |
| Loss function | Cross-Entropy |
| Framework | Pytorch |
| Optimizer | Adam, Rectifier Adam, ASGD |

Table 1

### B. Model B : Encoder-Decoder with Attention

By using the attention mechanism in encoder-decoder architecture, we not just use the thought vector produced at the

end of the encoder training but use all the intermediate vectors the encoder produces after appropriately weighting them allowing us to give more attention to the most relevant vector corresponding to the object of concern in the image. As shown in figure 2, every intermediate vector of the image model the encoder produced will be used by the decoder language model while generating the caption. For this architecture, the encoder is an InceptionV3 model with pre-trained weights [48]. A GRU network with attention [49] is used as a decoder. The choice of GRU instead of an RNN or an LSTM comes from experimental evaluations performed. Table 2 summarizes the configuration for Model B's training.

| | |
|---|---|
| Batch size | 64 |
| Number of epochs | 20 |
| Embedding size | 256 |
| Hidden size | 512 |
| Minimum words in caption | 5 |
| Loss function | Cross-Entropy |
| Framework | Keras |
| Optimizer | Adam, Rectifier Adam, |

Table 2

## IV. Methodology

- Data Set Collection: The data set was generated on top of the existing Microsoft COCO (Common Objects in Context) [25]. The original data set consists of 100,000+ image-caption pairs. There are 82,783 image-caption pairs in the training set, 40,504 image-caption pairs in the validation set, and 40,775 in the test set.
- Caption Translation to Nepali: For each given English caption, we used Google Translate service to convert it to a Nepali caption. Thus, the training data itself presents some discrepancy in the compositionality of the captions used as targets.
- Pre-processing: Several translated captions turned out to be nonsensical and many image-caption pairs were filtered subjectively. For this purpose, the authors first used an existing API for spelling check [50] and then verified each target's compositionality. The data set can be accessed here.
- Training and Testing: Models A and B were trained with several configurations and only the best-configuration results are reported. It is to note that the work used validation set for testing instead of the test set and the hyperparameters were chosen empirically.

## V. Results

The results for different optimizers used are illustrated in tables 3 and 4. We also indicate perplexity, the degree of uncertainity a model has while assigning probabilities.

Intriguingly, model A without visual attention performed better than model B with visual attention hinting that visual attention is not always the optimal method for image captioning. Model A achieves a loss of 0.02 and a perplexity of 0.88 with RAdam which is significantly better than the results with Adam or ASGD for the same model. The experiment shows that model B achieves a minimum loss of 0.06 with a perplexity of 1.006 with Adam. Further, perplexity of model A is lower than of model B indicating that model A generalizes better to unseen samples. While visual attention is an applicable tool for English image captioning, its capability appears to be limited for languages like Nepali with intricate grammatical structures.

Interesting!

Table 3 illustrates the results in terms of the loss function used and perplexity for model A.

| Optimizer | Loss | Perplexity |
|---|---|---|
| ADAM | 0.09300 | 1.09412429 |
| RADAM | 0.0218 | 0.88853 |
| ASGD | 0.049287 | 1.381937 |

Table 3

Table 4 illustrates the results in terms of the loss function used and perplexity for model B.

| Optimizer | Loss | Perplexity |
|---|---|---|
| ADAM | 0.06777 | 1.0069894 |
| RADAM | 0.07793 | 1.0078235 |

Table 4

Language modelling is a task where efficiency increases with the increasing availability of the data set and a sound training regime. The results on the test samples depend on the priors seen during the training regime. More importantly, the target captions used for this study were results of the Google Translate service and thus not perfectly accurate themselves. Thus, this study is limited by the data set used, the specific training regime followed, and the correctness of caption translation from English to Nepali. As a result, the quality of the captions generated for test samples is not impeccable. The performance could be enhanced using a larger data set, possibly a combination of several publicly-available data sets like Flickr30k and MS-COCO. Unlike machine-translated Nepali captions, using captions generated in Nepali manually for all the targets of the training samples should increase the performance of the models presented by a great margin. Further, the models presented here were trained to reduce the cross entropy loss and the behavior with a different loss function was not analyzed. The authors believe that the results can further be improved for both the models with further tuning of the architectures and training. Hence, the findings of this study are limited to a finite data set, machine-translated captions, and a fixed training regime and should not be generalized unconditionally to any other data set or language. Figures 5,6, and 7 are some qualitative results with agreeable captions generated by model A.

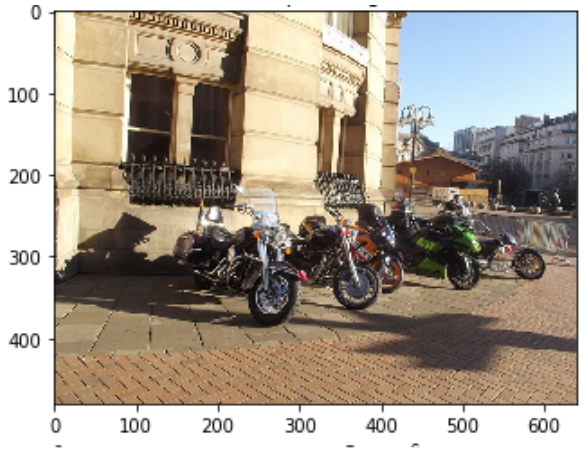


Figure 5


Figure 6

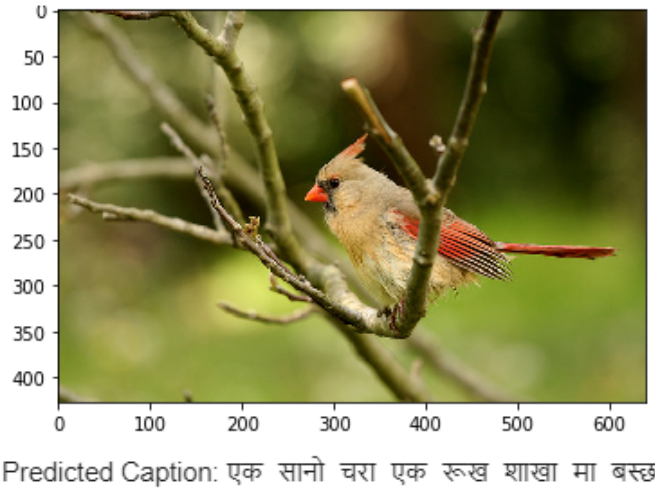


Figure 7

Figures 8, 9, and 10 show the captions generated by model B for the same test samples. As illustrated, model B generates captions that are compositionally not as sound as the captions given by model A.

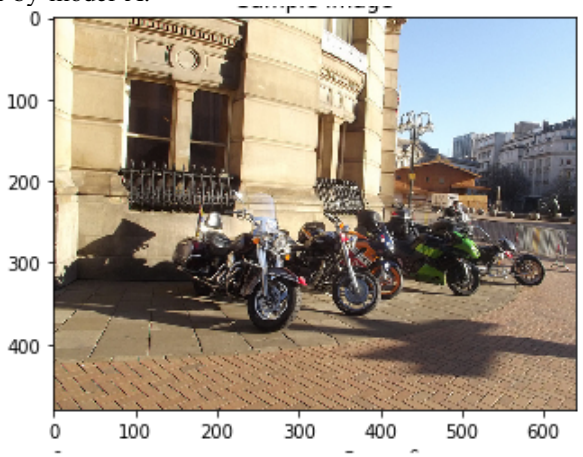


Figure 8


Figure 9


Figure 10

Nonetheless, model A also failed at different test samples and returned nonsensical captions as shown in figures 11 and 12. The quality of the captions generated depends on the data set that the model learned from and it is reasonable that the model generates meaningless captions for samples that appear off the distribution observed during training.

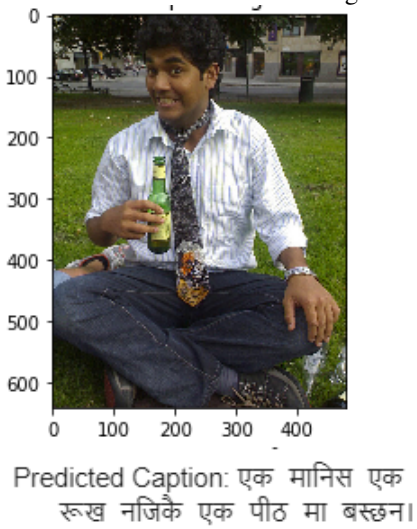


Figure 11

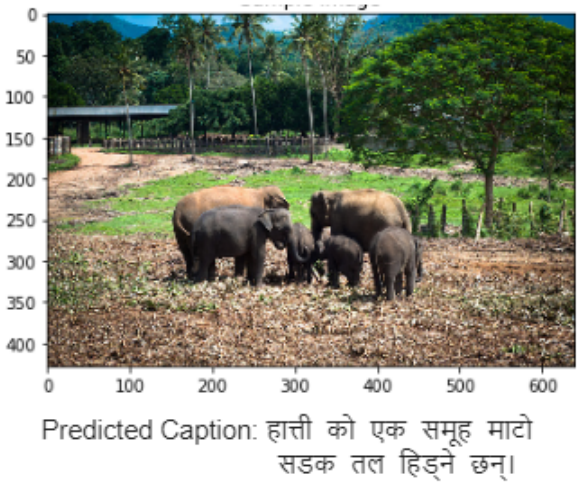


Figure 12

## VI. Conclusion and Future Work

We here present a baseline for end-to-end Nepali image-captioning, which when fed with an image can describe the contents of the image reasonably in Nepali. Another contribution of this paper is that the data set compiled and curated for this study has been publicly released to better the research in this direction. We showed that visual-attention model does not always outperform a plain encoder-decoder without visual attention for a complex language like Nepali. Nonetheless, this study is limited to the data set used, correctness of the machine-translated Nepali captions, and the specific training regime followed and the quality of the captions generated is not impeccable. As shown, even a reasonably good model fails to generate sensible captions for the test samples that are off the training distribution that the model learned from. With an extensive training data set, manually generated training captions in Nepali, and proper fine-tuning, researchers should be able to come up with more compositionally-valid captions for the test inputs to compare against the baseline presented in this paper. Further, the models presented here were trained using a single type of loss function and the variation in the efficacy of the models with the variation in the loss function used can be an interesting extension to this work. The presented model should be useful for various key applications - both commercial and personal, such as using a mobile application to help a visually-disabled individual to perceive the surrounding by playing the sentence generated using the model. This work can further be extended to several complex tasks including a video descriptor to summarize the contents of a video in Nepali and an image-based search engine where the input image is converted to a set of keywords for internet search using the model presented here among others.

## References


[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recog., pages 28742883, 2016.

[2] Guyon, I., Albrecht, P., Le Cun, Y., Denker, J. S., and ubbard W., H. (1991). design of a neural network character recognizer for a touch termin al. Pattern Recognition, 24(2):105.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. ICML, 2015

[4] C. Qian, Z. Xiao-Dan, L. Zhen-Hua, W. Si, J. Hui, and I. Diana, Enhanced lstm for natural language inference, In ACL, 2017.

[5] J. P. Chiu and E. Nichols, Named entity recognition with bidirectional lstm-cnns, arXiv preprint arXiv:1511.08308, 2015.

[6] Y. Ma, H. Peng, and E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm, in AAAI, 2018.

[7] Wikipedia contributors. "Nepali language." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 24 Aug. 2019. Web. 3 Sep. 2019.

[8] Sherstinsky, A., 2018. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv preprint arXiv:1808.03314.

[9] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat., 26(5):16511686, 1998.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. ICML, 2015.

[12] Ross Girshick. Fast r-cnn. In Proc. IEEE Int. Conf. on Computer Vision, pages 1440 1448, 2015.

[13] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In Proc. European Conf. on Computer Vision, pages 734750, 2018.

[14] Tsung-Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[15] IU Zeyu,MA Longlong,WU Jian, et al. Chinese Image Captioning Method Based on Multimodal Neural Network[J]. , 2017, 31(6): 162-171.

[16] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. "An Empirical Study of Language CNN for Image Captioning." ICCV, 2017.

[17] Jiuxiang Gu, Jianfei cai, Gang Wang, and Tsuhan Chen. "Stack-Captioning: Coarse-to-Fine Learning for Image Captioning." arXiv preprint arXiv:1709.03376 (2017).

[18] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In European conference on computer vision. Springer, 1529.

[19] Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv:1412.6632, December 2014.

[20] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.

[21] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.

[22] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. CoRR, abs/1410.1090, 2014.

[23] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.

[24] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740755. Springer, 2014

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[27] Tensorflow, mnist and your own handwritten digits. https://medium.com/@o.kroeger/ tensorflow-mnist-and-your-own-handwritten-digits-4d1cd32bbab4, 2016. [Online; accessed 16-July-2019].

[28] Roy, Sandip, and Debabrata Sarddar. The Role of Cloud of Things in Smart Cities. ArXiv:1704.07905 [Cs], Apr. 2017. arXiv.org, http://arxiv.org/abs/1704.07905.

[29] Joshi, Manish. SMART CITY IN NEPAL: CONCEPT AND INDICATORS. www.academia.edu, https://www.academia.edu/33684114/SMART CITY IN NEPAL CONCEPT AND INDICATORS. Accessed 5 Sept. 2019.

[30] Rahman, Motiur, et al. Chittron: An Automatic Bangla Image Captioning System. ArXiv:1809.00339 [Cs], Sept. 2018. arXiv.org, http://arxiv.org/abs/1809.00339.

[31] Parida, Shantipriya, et al. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. ArXiv:1907.08948 [Cs], July 2019. arXiv.org, http://arxiv.org/abs/1907.08948.

[32] Visual Geometry Group Home Page. http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/flickr100k.html. Accessed 5 Sept. 2019.

[33] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In IEEE Conference on Computer Vision and Pattern Recognition.

[34] Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In The SIGNLL Conference on Computational Natural Language Learning.

[35] Kuznetsova, Polina, Ordonez, Vicente, Berg, Alexander C, Berg, Tamara L, and Choi, Yejin. Collective generation of natural image descriptions. In Association for Computational Linguistics. ACL, 2012.

[36] Kiros, Ryan, et al. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. ArXiv:1411.2539 [Cs], Nov. 2014. arXiv.org, http://arxiv.org/abs/1411.2539.

[37] Sherstinsky, Alex. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. ArXiv:1808.03314 [Cs, Stat], Aug. 2018. arXiv.org, http://arxiv.org/abs/1808.03314.

[38] Hasim Sak, Andrew W. Senior, and Franoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[39] In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors, Interspeech, pages 338342. ISCA, 2014.

[40] Song, Kaitao, et al. MASS: Masked Sequence to Sequence Pre-Training for Language Generation. ArXiv:1905.02450 [Cs], May 2019. arXiv.org, http://arxiv.org/abs/1905.02450.

[41] Ren, Shuo, et al. Explicit Cross-Lingual Pre-Training for Unsupervised Machine Translation. ArXiv:1909.00180 [Cs], Aug. 2019. arXiv.org, http://arxiv.org/abs/1909.00180.

[42] Wang, Mingxuan, et al. Towards Linear Time Neural Machine Translation with Capsule Networks. ArXiv:1811.00287 [Cs], Nov. 2018. arXiv.org, http://arxiv.org/abs/1811.00287.

[43] Crego, Josep, et al. SYSTRANs Pure Neural Machine Translation Systems. ArXiv:1610.05540 [Cs], Oct. 2016. arXiv.org, http://arxiv.org/abs/1610.05540.

[44] Bahdanau, Dzmitry, et al. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv:1409.0473 [Cs, Stat], Sept. 2014. arXiv.org, http://arxiv.org/abs/1409.0473.

[45] Vaswani, Ashish, et al. Attention Is All You Need. ArXiv:1706.03762 [Cs], June 2017. arXiv.org, http://arxiv.org/abs/1706.03762.

[46] He, Kaiming, et al. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [Cs], Dec. 2015. arXiv.org, http://arxiv.org/abs/1512.03385.

[47] Chung, Junyoung, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv:1412.3555 [Cs], Dec. 2014. arXiv.org, http://arxiv.org/abs/1412.3555.

[48] Szegedy, Christian, et al. Rethinking the Inception Architecture for Computer Vision. ArXiv:1512.00567 [Cs], Dec. 2015. arXiv.org, http://arxiv.org/abs/1512.00567.

[49] Szegedy, Christian, et al. Rethinking the Inception Architecture for Computer Vision. ArXiv:1512.00567 [Cs], Dec. 2015. arXiv.org, http://arxiv.org/abs/1512.00567.

[50] Nepali Spell Checker. https://www.nepalilanguage.org/spellcheck/. Accessed 5 Sept. 2019.

[51] Perplexity. Wikipedia, 30 Aug. 2019. Wikipedia, https://en.wikipedia.org/w/index.php?title=Perplexityoldid=913229334.

[52] Image Captioning Using Encoder-Decoder. https://kharshit.github.io/blog/2019/01/11/image-captioning-using-encoder-decoder. Accessed 5 Sept. 2019.

[53] Vinodababu, Sagar. Show, Attend, and Tell — a PyTorch Tutorial to Image Captioning: Sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning. 2018. 2019. GitHub, https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning.