Experimenting Encoder-Decoder Architecture for Visual Image Captioning

Hasan Asif(✉)

State University of New York Buffalo, Buffalo, USA
hasanasi@buffalo.edu

Abstract. Image captioning is an essential task in artificial intelligence that predicts the description of a given input image. In recent years, both computer vision and natural language processing witnessed huge advancement making it capable to extract high-level semantic information and process it to structure new sentences based on that. A myriad of research has been done discussing the use of deep learning in image caption tasks. This paper explores the various combination of CNN and RNN modules in Encoder-decoder architecture to find the best image caption generator. The different models were compared using BLEU and CIDEr metrics. The proposed model with fine-tuned parameters showed a BLEU as high as 67.2, 59.8, 53, 44.7, and CIDEr score was 46 with Restnet50v2 model as an encoder and GRU as a decoder. After a little model hyperparameter tuning on Batch size and learning rate, an improvement of 15% and 12.5% was achieved in CIDEr and BLEU-4score.

Keywords: Image captioning · Deep learning · Sequence modelling

1 Introduction

The recent advancement in the area of deep learning especially CNN has motivated researchers to attend to the issue of image captioning. An image description generation is a demanding artificial intelligence problem where a word corpus must be brought out for a given image. It requires the incorporation of both methods from computer vision intelligence to comprehend the content of the image and a language model from the field of natural language processing to construct the understanding of the image into words in the right order. Recently, deep learning methods have achieved state-of-the-art results on examples of this problem.

Describing images automatically has a wide range of applications and covers a variety of tasks. Autonomously generating semantically rich descriptions from the image pixel has a handful of benefits. As a huge amount of unlabeled images upload every day to make the most use of this data it must be annotated earlier it was done manually by humans describing images but it was very time-consuming and boring for large databases. Automatic description generation is carried out by e-commerce companies: leveraging its capability to generate an accurate description of the listed product images. This technology contributes to image annotation for visually impaired people. Fortunately,

D. S. Guru et al. (Eds.): ICCR 2021, CCIS 1697, pp. 200–212, 2022.
https://doi.org/10.1007/978-3-031-22405-8_16

this technology can facilitate people having an issue in eyesight by providing descriptions of what's around them. In order to achieve this level of image caption generator, 3 modularity changes happen within the system i.e. first image is converted to text and, finally, these texts are read out by the system. In addition to this, various social media companies use image captioning technology to generate meaningful full text of the uploaded image by the user. This helps their models to refine and learn to detect novel patterns and objects.

The capability of autonomous image description generation can be of great importance in many fields/companies which extract information from videos for editing, adding subtitles, or summarization. Doing manually such a task can be very tedious; generation sentences from the ongoing scene in the video can be a boon for many video streaming sites and news websites to get headlines or descriptions of any viral video.

A human can easily comprehend the context behind the image and transmute the information in the form of sentences, while computers amalgamate two major fields of artificial intelligence computer vision and natural language processing to successfully complete this task. The issue associated with computer-based language generation is to make an algorithm that can discern the knowledge of scene, color, and object from the image. These computer-generated captions should be error-free as well as coherent to the main subject, discussed the state of an object, and uses a range of vocabulary to reduce redundancy.

To tackle the above-mentioned challenge many models have been proposed and quite promising results have been achieved in past few years. Work in the field image captioning field can be categorized into 3 approaches.

A: Retrieval based approach.
B: Template based approach.
C: Deep Learning based approach.

Primarily, image captioning was explored using retrieval based approaches. A plethora of primitive image captioning models are leveraging such an approach. In this approach, the input image is compared against a huge dataset using feature space by equating a Markov Random Field [1], and the caption is copied from the image which has the least semantic distance. Ordonez et al. 2011 [2] harness the caption for query image by finding the nearest visual match to the query image and pass-on the caption to the given input image. A year later, Kuznetsova et al. 2012 [3] presented a more refined model using a slightly different method to rank the images in the dataset against the query image. Kuznetsova introduced visual context and scene extraction and perform a manifold retrieval process for every entity/scene detected in the input image. However, Mason and Charnaik et al. 2014 [4] employed probability density estimation for language modeling which accurately describes what is in the image. This method outperforms many state-of-the-art methods by tackling the issue of grainy feature maps and makes it efficient where even entity detection systems perform poorly. The main issue associated with retrieval based approach is the need for a rich and great variety of images to have a close match to the query image. This constraint leads to more vague and incomplete sentences if used with small datasets [5]. Due to the explicit and inevitable

disadvantage of retrieval based approach other researcher pave their research in Template based approach [6–8]. In the template base method, the image is visually inspected and then the caption is generated by selecting the word as per the template defined. The template is defined as the sentence having a number of fixed empty blocks to produce a description of the image. Yang et al. [6] uses this approach and finds the most likely word to match the specific template(scene, noun, verb, and preposition) by using the hidden Markov model. In another paper, Kulkarni et al. [7] employed Conditional Random Field (CRF) to generate an image feature that highlights attributes, objects, and spatial relationships of different objects. And, at the last step, the caption template was filled as per the content. Similarly, Li et al. [8] use context information from the image to define triplet template format(adjective, preposition, adjective) This approach of making description using specified template lacks variable size description thus, resulting in loss of information in a number of occasions. In some aspect retrieval based approach is better than the template based approach but when it comes to more grammatically correct sentences template based method performed fairly good. Motivated by the advancement in the deep learning field, the deep learning model replaces the early shallow model. A lot of novel deep learning frameworks have been presented by researchers to do image caption in numerous fields is discussed in detail in the next section.

The aim of this paper is to present a model which can accurately and concisely define what is in the image. To do so we designed an encode-decoder framework containing CNN and LSTM units. In this paper, we extensively test the gamut of the combination between high-level feature extractor (VGGnet, ResNet50 V2, InceptionV3, and Xception) and language models(LSTM and GRU) and finds the best module for encoder and decoder by comparing different seq2seq combinations against BLEU and CIDEr score. Along with this, the paper provides a summary of approaches for image captioning from retrieval based to recent ones.

2 Related Work

In the past few years, visual caption generator has garnered huge attention from researchers due to huge image data generating every day. Automatic image captioning suits the best for the task of labeling unseen data on a large scale. In this section, we will discuss automatic image captioning technology using the most recent methods (Deep learning-based approach) in detail.

In a related paper author [9] attempts to solve the issue of image captioning without human interaction using encoder-decoder architecture. as high as 53.5 BLEU-4 score was achieved by using CNN feature extractor that consists of a pre-trained InceptionResnetv2 model as an encoder and GRU as the decoder. The proposed model was compared with different architecture on the basis of BLEU-4 and meteor score.H. Chen. et al. [10] founds the application of encoder and decoder in crack detection and improved the accuracy by using a switch between encoder and decoder module. WANG. et al. [11] finds the novel approach of the seq2seq model in long-term traffic prediction. The experiment showed improved results using a hard attention layer. The model learns long-time step patterns more accurately.use of LSTM significantly reduced the effect of vanishing gradient and gradient explosion during the learning phase.

Wang et al. [12] presented a model using bidirectional LSTM for image description. The model employed CNN and two different LSTM networks, leveraging it to make more contextually rich sentences. Rennie et al. [13] tried to optimize the image captioning task using novel self-critical sequence training(SCST) and the result are quite commendable tested on the MS COCO dataset. The baseline structure of the model contains ResNet50 V2 and LSTM as an encoder-decoder module. Aneja et al. [14] proposed a rational combination of Convolution Neural Net(CNN) based language approach for text generation to address the issue of vanishing gradient in the LSTM unit. The model performed equally well as compared to the CNN-LSTM combination. Additionally, significant improvement in training time was recorded. Kiros et al. [15] used multimodal learning where the model taught using image and text together with AlexNet for image understanding. The proposed method removed the need for templates, structures, or constraints instead uses state of art feature extractor(AlexNet). Yao et al. [16] described a novel architecture of GoogleNet incorporated with high-level attributes (LSTM-A). In his paper author tried to predict the attributes and inserted them into the LSTM to find the optimum node for insertion different LSTM-A models were tested. This architecture of Yao et al. extracts better semantic relationships, hence generate a more contextually accurate description.

Author [17] discussed the autonomous description generator for the input image by using a manifold framework to detect the human-object pair. The proposed method used a hybrid deep learning approach to get the insight from the image and outline the human object pair combined with probabilistic language model outperformed on many benchmark datasets. Captioning of an image is not an easy task to address this issue scholars recommended an Adversarial Networks and Reinforcement Learning [18] based framework to rectify the issue of bias in caption generation. The proposed model was trained on the COCO dataset showed significant improvement in evaluation.

Author [19] find the application of image description generator in medial domain i.e. deployed model generates a description of chest x-rays images using adversarial reinforcement learning and outstrip many previously employed methods. The proposed architecture comprises of 3 main components: encoder, decoder, and reward module, to extract feature VGG16 model was used in an encoder, whereas LSTM was used with attention mechanism in the decoder.

Paper [20] discussed the implementation of VGG16 based encoder-decoder network(Segnet) for semantic segmentation task with the Conditional Random Field(CRF) layer after the basic Seg-Net model. The model was trained on the CamVid dataset.

Encoder Decoder Architecture: Encoder-Decoder find its way back in 2012, by google for machine translation tasks [21]. Prior difficulty to the task where the previous context is necessary to predict the target, was done in a very nascent way. A critical advantage associated with the encoder-decoder model is that size of the input and output sequence can be different. This led the researcher to continue work in tasks like machine language translation, text-to-image conversion, and image caption where the length of input and output are different.

Image captioning is one of the tasks in which transferring of one sequence(image) into another sequence(caption) takes place. Therefore, the encoder-decoder framework

fits perfectly to tackle the problem. Kiros et al. [15] were the first to employed Encoder-Decoder architecture for image captioning task. Similar to his work, Vinyal et al. [22] presented his work of generating image descriptions using CNN as an encoder and LSTM as a decoder. In sequence2sequence there are mainly two-component one is an encoder and the other one is a decoder.

The Encoder-decoder model not only takes the current state into consideration while predicting the target output but also its neighbor state. The encoder consumes the image as inputs and produces a feature vector of the image called a context vector. Onwards, the decoder takes the context vector and predicts the words for each time step.

In general, the working of an encoder and decoder model for image captioning starts with a feature extraction of an input image. This feature map is achieved by employing Convolution Neural Network (CNN) as an encoder module. CNN layer extracts the feature of an input image from a fully connected layer(Global feature), these global features are then fed into the decoder. The decoder is responsible for the text generation from an input feature vector. This decoder could be Recurrent Neural Network (RNN), Long-short Term Memory (LSTM), and Gated Recurrent Unit (GRU) as the basic unit. The majority of the encoder-decoder image captioning models use Maximum Likelihood Estimation (MLE) as their learning method whereas, few used reinforcement learning [13] to reduce the exposure bias issue, Generative Adversarial Network [23] and contrastive learning [24] methods were also proposed (Fig. 1).

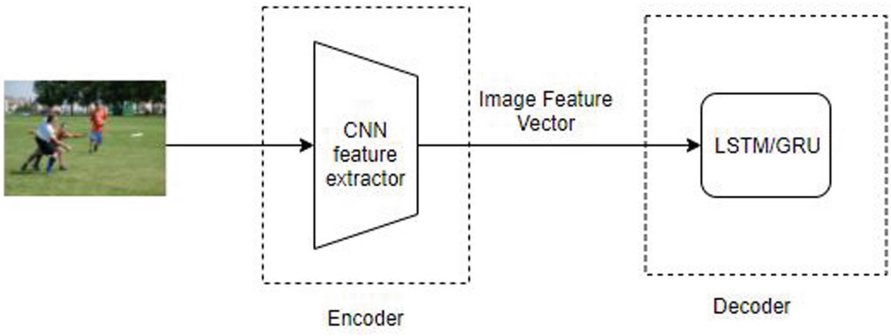


Fig. 1. Basic encoder decoder architecture

3 Proposed Model

In this section we will study about the proposed method. Here, we define the method of study. For each encoder decoder combination.

Feature Extractor: To represent the feature from the input image CNN was employed, ought to their success in feature extraction. The output from the Last fully connected layer gives the global representation of an image. This output is the vector input to the decoder

later. In this paper we have chosen many different CNN as an encoder such as VGGNet [25], ResNet50V2 [26], Inception [27]. The best results comes out with ResNet50V2 which was discussed later in the result. To achieve the feature representation vector from each encoder module we have to follow certain steps shown in Fig. 2.

This paper proposed ResNet50V2 model as an encoder having 50 layers deep neural network, such deep network trained quite well even without facing vanishing gradient problem. ResNet50V2 escape the problem of vanishing gradient in deep network by cleverly implementing switching connection between the layers. The idea of skip connection was introduced by the ResNet50V2 in 2016. Skipping connection between the layers helps ResNet50V2 model by permitting the flow of gradient to pass through another route. Since then ResNet-50 architecture are widely explored in computer vision tasks. Talking about the architecture, ResNet50 has over 23 million trainable parameters. The v2 variant of ResNet50V2 is similar to V1 in all ways except aa batch normalization and Relu activation applied before convolution operation. ResNet50V2 is the more modified and efficient version of RestNet50.

Language Model: Given the input feature vector, a decoder is responsible for generating word. To do so, we tested LSTM [28] and GRU [29] as a decoder or language model. Recurrent Neural Network perform exceptionally good in sequence generation. Due to the major issue of vanishing gradient associated with RNN, we tested out both LSTM and GRU as the basic unit as well. And, the result were significantly impressive.

GRU: This version of RNN is the newer than LSTM, first introduced in 2014. Basic Recurrent Neural Network suffers from short-term memory problem which means for longer sentences or sequence it forgets the previous data and remembers only last few sequence. This issue was first resolved storing long-term memory in cell state in Long-short Term Memory (LSTM) cell whereas, the hidden sate store the short term memory. GRU is the lighter modified version of LSTM, which contains both long and short term memory in hidden sate as shown in fig. Typical GRU cell contains two gates: update gate and reset gate. Update gate is responsible for how much of past memory to store meanwhile, Reset gate perform the forget operation on the stored memory. Reset gate takes the hidden state h_{t-1} and the current word x_t and then sigmoid function(σ_g) on the top of it to give the reset gate value r_t as in Eqs. 1–3 Similarly, update gate performs, the same weight operation however the weights are not same. z_t is the value we get from update gate. Hence, GRU will use these value r_t, z_t, h_{t-1} and x_t to get new hidden state as well as the output vector (h_t).

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \hat{h}_t \tag{3}$$

W, U, and b are the parameter matrices and vector. Whereas represents Hadamard product in the Eq. 3.

4 Model Training and Dataset

Different encoder decoder sets were created and trained to find out the most effective one. Details are listed in the table below. Each model was roughly tested with both LSTM/GRU in the decoder and the best one was selected based on overall BLEU and CIDEr score Then, the best architecture was selected for further analysis. All models training was done in the Google Colab environment. For every model training process, we have used Adam optimizer. Experiment results illuminate that ResNet50V2 and GRU architecture outperform other models. The experiment was repeated for batch sizes between 8 to 32, with learning rate ranges from .01 to .0001. Word2vec embedding was utilized for word embedding. It generates a 200 dimension vector for each token. During the experiment, LSTM and GRU state size was initialized to 512. We had used Flickr8k [30] dataset, which includes 8000+ image samples which are further divided into training, testing, and validation set. Every image in the dataset is paired with the 5 captions. Overall, it contains somewhere around 40000+ captions. Mean length of the caption comes out 11 and the longest sequence contains 37 words. All the image samples were collected from the photo storage website Flickr.

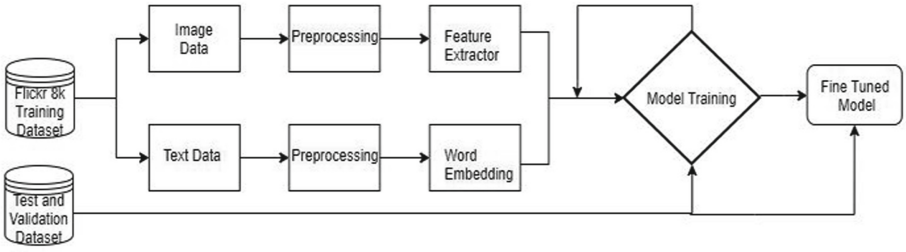


Fig. 2. Model training process

Table 1. Number of image sample in dataset

Set	Size
Training set	6000
Testing set	1000
Validation set	1000

Table 2. Various encoder decoder combination we tested in our experiment

Model no	Encoder	Decoder
1	VGG-16	LSTM/GRU
2	VGG-19	LSTM/GRU
3	ResNetV2-50	LSTM/GRU
4	InceptionV3	LSTM/GRU
5	Xception	LSTM/GRU

5 Evaluation Metrics

BLEU Score [31]: BLEU stands for Bilingual Evaluation Understudy score proposed by Kishore papineni 2002. It ranges from 0 to 1. Higher the score better the generated text. BLEU score implies the similarity between the generated text and actual text. Effectiveness of the generated description was evaluated by counting the match between the n-gram of generated description and the n-gram of the original caption. Similar grams are irrespective of the index in the sentence. BLEU score can evaluate the performance of many deep learning applications for example machine translation, text summarizer, image captioning, and many more. The main drawback of using the BLEU score is its inability to grasp the context or grammar of the generated text. Thus, sometimes it even shows a high correlation between the original text and grammatically incorrect or opposite meaning sentences.

$$\text{BLEU} = \text{BP} * \exp\left(\sum_{n=1}^{N} W_n \log P_n\right)$$

BLEU or BLEU Overall is a geometric mean of n-gram scores from 1 to 4.

CIDEr Score [32]: Consensus-based image description evaluation BLEU score which is traditionally used for machine translation task, whereas CIDEr metric was wholly generated for image caption task. CIDEr metric shows higher similarity with human judgment score. CIDEr score was calculated by measuring how frequently n-gram in the generated description are matching the reference sentence, the same thing is calculated for non-matching words. And then, the most frequent n-gram across all captions are penalized which calculating correlation using TF-IDF weighting for every n-gram.

$$\text{CIDEr}_n(C_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(C_i) \times g^n(S_{ij})}{||g^n(C_i)||||g^n(S_{ij})||}$$

$g^n(a)$ is a vector formed by TF-IDF scores of all n-grams in "a". C_i and S_i refers to the generated caption and the original caption respectively (Tables 1 and 2).

Results: Table 3 Shows the overall performance of different model in their default settings, the optimum result was shown by the restnet50v2 and GRU model, which suggests that ResNet50v2 is quite well in capturing image feature as compared to the other feature

extractor. Furthermore, between GRU and LSTM, GRU provides better description and pays more attention to the scene, and object as compared to LSTM generated model. Only VGG variants performs well with LSTM otherwise GRU outperform in every metrics.

Table 3. Comparison of various architecture on Flickr 8k dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
VGG-16+LSTM	55.1	44	38	31.6	32.4
VGG-19+LSTM	63	54.9	45.6	33.3	35.7
ResNetV2-50+GRU	63.2	55.8	48.1	39.7	40.4
InceptionV3+GRU	61	54.1	43	35	39.1
Xception+GRU	58	49.9	30.6	25.7	17.4

Table 4. Comparison of different batch size on ResNetV2-50+GRU architecture

Batch size/score	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
4	66	55.5	49.4	41	38
8	65.9	56.7	48	40.1	43.4
16	59.3	48.6	35	19	34.1

Table 5. Comparison of different learning rate on ResNetV2-50+GRU architecture

Learning rate/score	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
0.01	60	50.4	40.1	31	31.3
0.001	57	51	40.2	36.8	40.2
0.0001	59.6	40.3	34	30	33.1

Table 6. Different metrics score of the proposed fine tuned mode

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr
ResNetV2-50+GRU	67.2	59.8	53	44.7	46

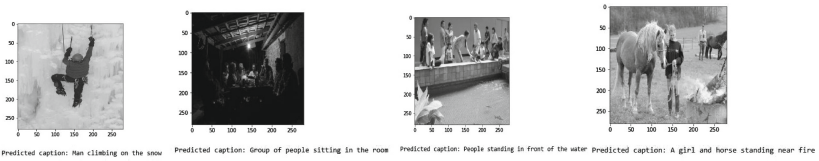
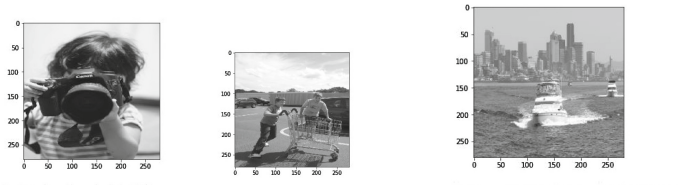
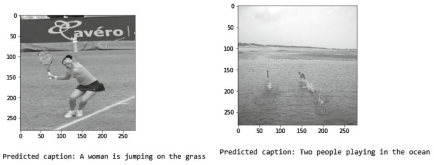
Table 4. Shows the effect of different batch sizes on the Evaluation metrics. It was evident that increasing the batch had a detrimental effect on the both BLEU-n and CIDEr score. This was due to the overfitting thus model's ability to generate effective caption degrades. The best BLEU score was achieved for batch size 4 whereas CIDEr score is the highest for batch size 8. A significant drop in BLEU-4 was recorded from 40.1 at batch size 4 to 19 at batch size 16. In Table 5. Different learning rates were tested on ResNet50ve and GRU model. Overall, model performance decreases as the learning rate decreases. Various fluctuations were recorded but the best learning rate comes out to be .001. Presented results hints that a batch size of 4 and a learning rate of .001 could generally be expected to give out the most optimum description of the image.

After all the parameter tuning was done, the test was reconstructed by setting the most optimum parameters for the selected ResNet50V2 and GRU architecture, and the results achieved were quite impressive as shown in Table 6. Approximately as high as 15% increase in the CIDEr score was achieved similarly 12.5% increment was recorded in BLEU-4 score.

6 Conclusion

Overall, in this paper, we have discussed the image captioning task through the different methods and proposed an encoder-decoder architecture based on ResNet50v2 with GRU for the same task. The promising result was shown by our model as compared to the previous method. Deep layer architecture of ResNet50v2 joint with GRU having capabilities to retain long-term memory powered this image captioning task. Perhaps, from the initial model selection stage, this combination clearly stands out from the others. With very minimal refinement, we achieved significant improvement(15% in case of the CIDEr score and 12.5% in case of the BLEU-4 score) in the quality of the description. Our presented encoder-decoder architecture, retrieve image information quite efficiently as depicted in captions that correspond to the content in the image. Table 7. Depicts some of the generated captions using the proposed model and classified based on their quality(good, average and bad captions).

Table 7. Table shows generated captions from our proposed model on flickr 8k dataset.

Caption Quality	Image and Caption
Good Captions	
Average Captions	
Bad Captions	

References

1. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 15–29. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_2
2. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems (NIPS), pp. 1143–1151 (2011)
3. Kuznetsova, P., Ordonez, V., Berg, A.: Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 1 July, pp. 359–368 (2012)
4. Mason, R., Charniak, E.: Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 592–598. Association for Computational Linguistics (2014)
5. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
6. Yang, Y., Teo, C., Daumé III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 444–454 (2011)
7. Kulkarni, G., et al.: Baby talk: understanding and generating simple image descriptions. In: Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), vol. 18, pp. 1601–1608 (2011)

8. Li, S., Kulkarni, G., Berg, T., Berg, A., Choi, Y.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL 2011), Portland, USA, pp. 220–228. Association for Computational Linguistics (2011)
9. Parikh, H., et al.: Encoder-decoder architecture for image caption generation. In: Proceedings of the 3rd International Conference on Communication System, Computing and IT Applications (CSCITA). IEEE (2020)
10. Chen, H., Lin, H., Yao, M.: Improving the efficiency of encoder-decoder architecture for pixel-level crack detection. IEEE Access (2019). https://doi.org/10.1109/ACCESS.2019.2961375
11. Wang, Z., Su, X., Ding, Z.: Long-term traffic prediction based on LSTM encoder-decoder architecture. IEEE Trans. Intell. Transp. Syst. (2020). https://doi.org/10.1109/TITS.2020.2995546
12. Wang, C., Yang, H., Bartz, C., Meinel, C.: Image captioning with deep bidirectional LSTMs. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 988–997. ACM (2016)
13. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1179–1195 (2017)
14. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5561–5570 (2018)
15. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: Proceedings of the 31st International Conference on Machine Learning (ICML), pp. 595–603 (2014)
16. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: IEEE International Conference on Computer Vision (ICCV), pp. 4904–4912 (2017)
17. Huo, L., Bai, L., Zhou, S.-M.: Automatically generating natural language descriptions of images by a deep hierarchical framework. IEEE Trans. Cybern. **52**, 1–12 (2021). https://doi.org/10.1109/TCYB.2020.3041595
18. Balakrishna, K.: WSN, APSim, and communication model-based irrigation optimization for horticulture crops in real time. In: Tomar, P., Kaur, G. (eds.) Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture, pp. 243–254. IGI Global (2021). https://doi.org/10.4018/978-1-7998-1722-2.ch015
19. Hou, D., Zhao, Z., Liu, Y., Chang, F., Hu, S.: Automatic report generation for chest X-ray images via adversarial reinforcement learning. IEEE Access **9**, 21236–21250 (2019). https://doi.org/10.1109/ACCESS.2021.3056175
20. de Oliveira Junior, L.A., Medeiros, H.R., Macedo, D., Zanchettin, C., Oliveira, A.L.I., Ludermir, T.: SegNetRes-CRF: a deep convolutional encoder-decoder architecture for semantic image segmentation. In: International Joint Conference on Neural Networks (IJCNN) (2018)
21. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html
22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164 (2015)
23. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: ICCV, pp. 2970–2979 (2017)
24. Dai, B., Lin, D.: Contrastive learning for image captioning. In: Advances in Neural Information Processing Systems, pp. 898–907 (2017)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

27. Szegedy, I.C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.308
28. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
29. Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Deep Learning and Representation Learning Workshop (2014)
30. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's mechanical turk. In: NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147 (2010)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, p. 311 (2002)
32. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015)