Analysis of Image Captioning Approaches from a Deep Learning Perspective

Monika Chawla* and Rashmi Agrawal

School of Computer Application, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India
E-mail: monika.chawla12@gmail.com; rashmi.sca@mriu.edu.in
**Corresponding Author*

Received 08 January 2025; Accepted 01 May 2025

Abstract

Every day, millions of images are seen through all these mediums-social media, news headlines, advertisements. People have an implicit sense of what those images represent, but for machines, it is possible to generate meaningful insights only through complex algorithms. Captioning images forms one of the most basic applications of AI and pertains to textual descriptions that can help in enabling functionalities like automatic indexing, CBIR, and accessibility. Deep learning models demonstrated potential capabilities to automatically learn features to generate semantically rich captions that are coherent; however, template-based, and retrieval-based approaches find it challenging to implement flexibly to produce ultra-high-detail, context-specific captions. The techniques here, such as CNNs, extract visual features while RNNs and LSTMs generate descriptive text. The higher-level architectures included are the encoder-decoder frameworks and compositional models that provide further enhancement by aligning visual data and textual data. The paper briefly discusses deep learning techniques categorized into structure and application-based categories and tests the performance of

Journal of Mobile Multimedia, Vol. 21_3&4, 363–378.
doi: 10.13052/jmm1550-4646.21341

benchmark datasets such as Flickr8k, Flickr30k, and MSCOCO. However, much remains to be done in terms of building models robust to complex and diverse visual content; thus, it is observed that there are challenges that carry forward to the work on multimodal integration and attention-based mechanisms to be improved in terms of better quality and accuracy by the captions.

Keywords: Image-captioning, AI, deep learning, CNN, RNN and LSTM.

1 Introduction

On a regular basis, people see and use thousands of images from various sources, including online services like Google Images and Facebook, news organizations, marketers, and more. Although these images are immediately intelligible to humans without any additional context, machines require mechanisms to interpret and describe them automatically. Captioning photographs is essential for computer-aided image indexing and aids Content-Based Image Retrieval (CBIR), and it has commercial, defence, educational, digital library, web search, and social media applications that produce contextual descriptions like location, clothing, and activity. As an element of artificial intelligence, image captioning has become a focal point in generating textual descriptions of images, which include object identification, scene recognition, and the generation of syntactically and semantically coherent sentences. Features are learned to understand images using traditional machine learning or deep learning methods. With traditional methods, handcrafted features such as Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and classifiers such as Support Vector Machines (SVMs) are used. These are less adaptive for various datasets. Deep learning learns the features automatically from the data and hence is extremely powerful when working with large sets of images.

Convolutional Neural Networks (CNNs) employ deep features, Softmax classifies, and Recurrent Neural Networks (RNNs) are used for caption generation. Despite the tremendous growth in deep learning-based image captioning, there is a significant deficit in comprehensive reviews compared to other methods. This paper fills the gap by focusing on deep learning-based image captioning, which has been propelled forward by vast and diverse datasets.

2 Image Captioning Methods

These describe three broad image captioning paradigms: template-based, retrieval based, and novel caption generation. The template-based paradigm fills pre-set templates with labelled objects, actions, and properties to create the captions. While flexible, this method is not adaptable and has increasingly been replaced by more parsing-based language models. Retrieval-based image captioning selects captions from a pre-existing database primarily by finding images with similar features and then retrieving their corresponding captions. While this method can obtain captions, these captions often lack significant semantic information since it primarily utilizes paraphrased text rather than original descriptions. Advanced deep learning techniques employ image caption generation models to consider visual content and create novel captions. In this approach, uniqueness, semantic accuracy, and specificity are more crucial than in retrieval methods. Deep learning paradigms, such as supervised, unsupervised, and reinforcement learning, encapsulate various learning activities in captioning. Some methodologies prefer an encoder-decoder structure, while others adopt a more complex compositional method that integrates attention mechanisms and semantic concepts with diverse artistic elements. Though most deep learning algorithms used to caption images rely on long short-term memory (LSTM) networks, a few models utilize CNNs or RNNs. This method of deep learning makes it possible to create captions for entire scenes or object information in any given image, making it flexible and giving it many possibilities for descriptive content.

3 Deep Learning Captioning Techniques for Images

Deep learning-based image captioning is possible in a visual space, where text captions and image features are processed separately, or in a multimodal space, where images and text share a common representation to decode language. In this, visual information obtained from images by CNNs is fed to a language model such as a RNN, LSTM network, gated recurrent unit (GRU), or Transformer to produce captions. This kind of methodology is based exclusively on vision data to create understandable captions, the models being usually trained on large corpora consisting of images and captions. It is applied mostly in image search, annotation of content, and accessibility. Figure 1 categorizes deep learning image captioning methods into several criteria that involve visual space, multimodal space, dense captioning, supervised learning, encoder-decoder structure, and compositional structure. It has

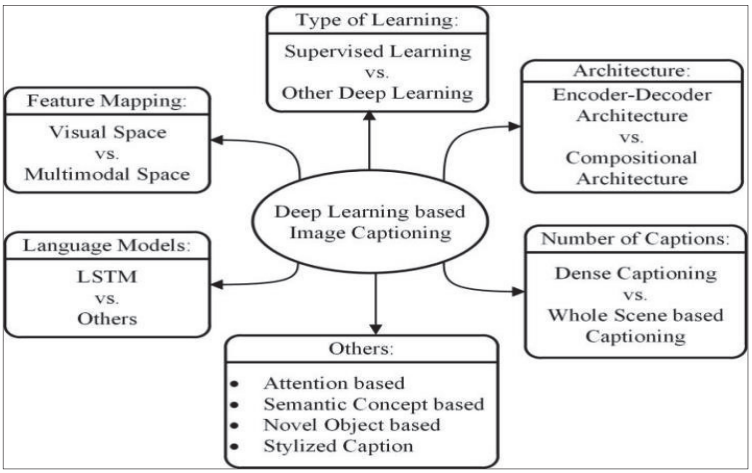

Figure 1 Classification of deep learning-captioning techniques for images.

also an "Others" category that includes attention, semantic concept, stylized, and novel object captioning techniques.

3.1 Multimodal Space Image Captioning

In contrast, multimodal space-based image captioning combines both visual and textual information to create captions. It commonly consists of a setup with a language encoder, a visual feature extractor (typically as a deep convolutional neural network), a multimodal space into which image and text information feed, and a language decoder. The language encoder produces dense word embeddings that retain semantic context, while the multimodal space aligns and integrates visual and text-based information so that captions appear coherent and contextually appropriate. The image captioning technique that operates within the multimodal space can be understood using the structure outlined in Figure 2.

This multimodal space-based representation is then passed on to the language decoder to produce the captions, which in turn decode the multimodal representation. Such a method usually goes through two steps: learning by the deep neural networks for visual and textual features in a shared multimodal space, thereby enabling joint image and text representation. Second, the captions are generated by the language decoder from the shared representation derived from the multimodal space. Kiros et al. [1] presented one of the first methods in this area, using convolutional neural networks (CNNs) to

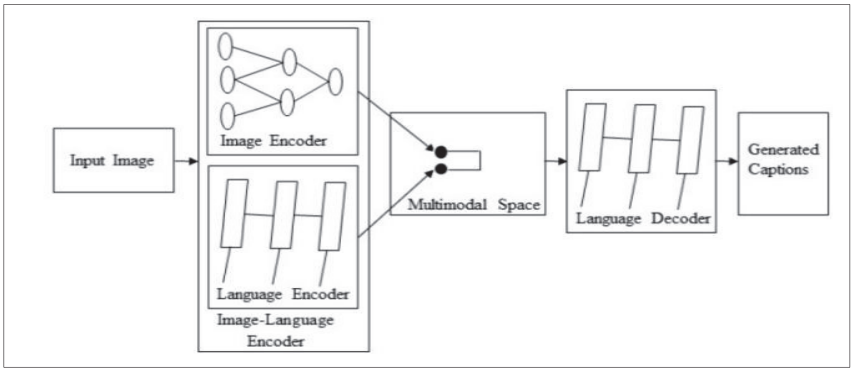

Figure 2 Architecture of multimodal space-based image captioning.

extract visual features. The authors used multimodal neural language models, including the Modality-Biased Log-Bilinear Model and the Factored three-way Log-Bilinear Model. These methods leverage high-level image features and word embeddings, thereby not using additional templates or restrictions. Though, neural language models are limited by very large-scale data and long-term memory handling. Subsequently, Kiros et al. [1] utilized long short-term memory (LSTM) networks for phrase encoding to develop the idea of a structure-content neural language model (SC-NLM). Captioning is adaptive, and this technique outperformed earlier techniques and demonstrated great improvement over earlier techniques. In [2], Karpathy et al. presented an alternative perspective with another deep multimodal model that embeds image fragments and sentences for the bidirectional retrieval task. The essence behind this decomposition of images into objects and sentences into dependency trees allows for finer alignment between images and text than previous methods. However, it has its limitations; the dependency trees are quite complex and are often unsuitable for mapping visual objects, especially regarding complex phrases and ambiguous relationships.

Mao et al. [3] presented the m-RNN (multimodal Recurrent Neural Network) for captioning images, where deep convolutional and recurrent networks are merged in the multimodal layer to generate captions through the computation of the probability distribution of the subsequent word. Five layers are used in the architecture: two layers of word embeddings, a recurrent layer, a multimodal layer, and a SoftMax layer. Mao et al.'s m-RNN is better than Kiros et al.'s Log-Bilinear model in using a recurrent architecture with variable-length context and learning word representations from training data to generate more accurate captions related to the images and enhancing

the recurrent layer's performance. Chen et al. [4] presented a multimodal image captioning approach where, in addition to caption generation from images, the model can also reverse-project visual features from captions. This approach is significantly distinct from others since it employs a recurrent visual hidden layer for backward projection [5].

3.2 Dense Image Captioning

Dense image captioning differs from traditional methods because it produces a variety of captions to depict different parts of the image, whereas traditional methods provide a single caption that sums up the entire scene. Traditional systems combine various elements of the image to produce a caption; however, they sometimes fail to provide an explanation for each area. Johnson et al. [6] introduced DenseCap as a dense image captioning method that detects prominent areas of an image and produces meaningful descriptions for each detected area. The major steps in dense captioning are: (1) Creating region proposals from an image (2) Passing the features of these regions through a convolutional neural network (3) Passing the features extracted into a language model to generate captions for each region.

DenseCap uses fully convolutional network with dense localization layer and LSTM model. Unlike certain approaches, e.g., Fast R-CNN, DenseCap does not need an external Region Proposal Network (RPN). It makes a single effective forward pass to estimate areas of interest and utilizes differential spatial soft attention and bilinear interpolation rather than the ROI pooling mechanism for smooth area selection and effective backpropagation. While dense captioning explains regions much more objectively and in detail, it is plagued by overlapping regions and ambiguity in detecting each visual concept that falls within dense regions. Yang et al. [7] introduced one more dense captioning approach that solves same problem. It uses an inference mechanism that aggregates visual features and predicted captions to identify accurate bounding boxes. Moreover, it incorporates a context fusion mechanism, wherein context features are aggregated with region-specific visual features to generate high-level semantic descriptions. Figure 3 depicts the architecture of the dense image captioning approach.

3.3 Encoder-decoder method for image captioning

Encoder-decoder approach to image captioning methods is like neural machine translation, utilizing convolutional neural networks (CNNs) to extract image features, which are passed to a long short-term memory

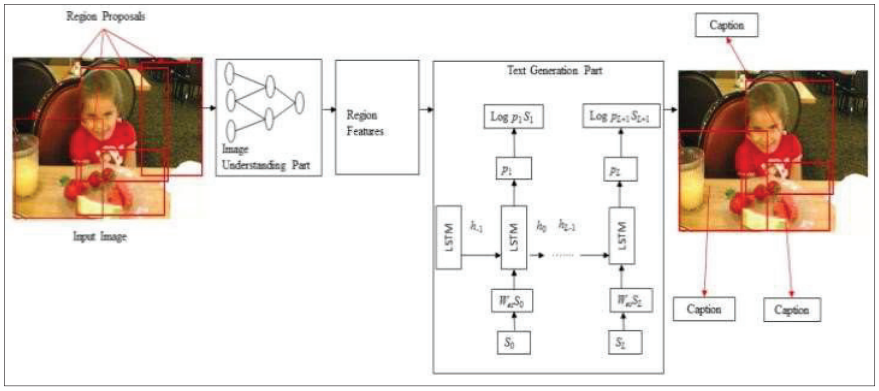

Figure 3 Architecture of dense image captioning method [8].

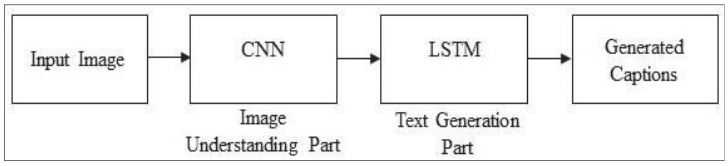

Figure 4 Architecture of encoder-decoder based captioning of image.

(LSTM) network to form words in sequence. Vinyals et al. gave us the Neural Image Captioning model, while the issue of the vanishing gradient was treated by Jia et al., presenting the gLSTM with better support for long sentences and adaptable control over the caption's length. More structural details may be observed from Figure 4.

Mao et al. [9] suggested an approach that produces accurate descriptions, i.e., referring expressions for objects or image regions, from MS COCO dataset references. Due to the superficial and one-way design of CNN-RNN models lacking depth in context, some researchers developed this model employing a more complicated method known as Bi-LSTM-based architecture. This adjustment facilitates better visual and linguistic interaction through the addition of context both before and after a CNN and an LSTM network on either side.

3.4 Compositional Image Captioning Architecture

Compositional image-based methods have many functional components. In such methods, the CNN architecture first learns semantic concepts from

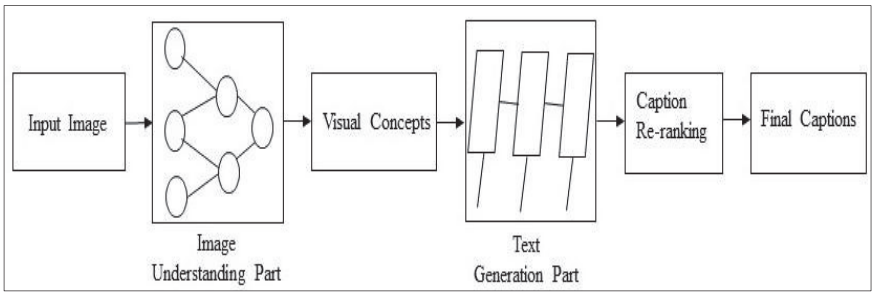

Figure 5 Compositional network-based captioning architecture.

the images. Second, with the help of a language model, candidate captions are generated. The candidate captions are re-ranked using deep multimodal similarity models to select the best final caption. Such a process usually has the following workflow:

- Utilizing CNN for Image Feature Extraction.
- Identifying Visual Concepts from the extracted features, including attributes and object recognition.
- Generating a Caption Pool by integrating extracted data with a language model.
- Optimizing Captions through a multimodal deep similarity model to select the most accurate and contextually relevant captions.

Fang et al. [10] presented the compositional network-based image captioning paradigm, as demonstrated in Figure 5. This consists of multimodal similarity and visual detectors trained from an image captioning set with a common vocabulary of top 1,000 most used terms in the training set represented as nouns, adjectives, and verbs to describe the image captions. Attention here is to sub-regions of an image rather than the entire image.

Related CNNs, e.g., AlexNet [5] and VGG16, learn sub-region features corresponding to potential vocabulary words. A maximum entropy (ME) language model produces captions and is trained with multiple instance learning (MIL) to learn discriminative visual cues. The captions are sorted using linear weighting and minimum error rate training (MERT) optimization. A deep multimodal similarity model (DMSM) calculates similarity between fragments and sentences and facilitates shared vector representation and comparison for the selection of improved captions.

Table 1 summarizes deep learning-based techniques for image captioning in brief, using the names of methods and categories of adopted deep neural

Table 1 Below the list summary of deep learning imaging captioning approaches (ms=multimodal space, dc=dense captioning, eda=encoder-decoder architecture, ca=compositional architecture)

Image Encoder	Language Model	Category	Reference
AlexNet	LBL	MS, EDA	Kiros et al. 2014 [8]
AlexNet, VGGNet	LSTM SC-NLM	MS, EDA	Kiros et al. 2014 [11]
VGGNet	RNN	EDA	Chen et al. 2015 [12]
AlexNet, VGGNet	MELM	CA	Fang et al. 2015 [13]
VGGNet	LSTM	EDA	Jia et al. 2015 [14]
VGGNet	RNN	MS, EDA	Karpathy et al. 2015 [15]
GoogLeNet	LSTM	EDA	Vinyals et al. 2015 [16]
VGGNet	LSTM	DC, EDA	Johnson et al. 2016 [17]
VGGNet	LSTM	EDA	Mao et al. 2016 [18]
VGGNet	LSTM	CA	Wang et al. 2016 [19]
VGGNet	LSTM	DC, EDA	Yang et al. 2016 [20]
VGGNet	LSTM	CA	Anne et al. 2016 [6]
GoogLeNet	LSTM	EDA	Yao et al. 2017 [21]
ResNet	LSTM	EDA	Lu et al. 2017 [22]
ResNet	LSTM	CA	Gan et al. 2017 [23]
VGGNet	RNN	EDA	Pedersoli et al. 2017 [24]
VGGNet	Language CNN & LSTM	EDA	Gu et al. 2017 [25]
VGGNet	LSTM	CA	Yao et al. 2017 [26]
ResNet	LSTM	EDA	Rennie et al. 2017 [27]
VGGNet	LSTM	CA	Vsub et al. 2017 [28]
Inception-V3	LSTM	EDA	Zhang et al. 2017 [29]
VGGNet	Language CNN	EDA	Aneja et al. 2018 [5]
VGGNet	Language CNN	EDA	Wang et al. 2018 [30]

networks in encoding image data as well as the respective language models used for description.

4 Result Analysis

Although this paper does not include any sort of formal experimental analysis, we still give a very brief analysis of the experiment results and the effectiveness of various techniques as presented. Table 2 presents a comparison of the performance of some image captioning techniques on some various benchmark datasets. Each row represents a specific dataset, indicating which image was used, along with the captioning technique and method applied. The columns present BLEU-1 and BLEU-2 scores, as these

Table 2 Comparative performance analysis of image captioning techniques across multiple benchmark datasets

Datasets	Captioning Techniques on Images	Methods	BLEU1	BLEU2
FLICKR8K	MULTIMODAL SPACE-BASED	MAO ET AL. 2015 [31]	0.565	0.386
FLICKR8K	ENCODER-DECODER ARCHITECTURE	JIA ET AL. 2015 [14]	0.647	0.459
FLICKR30K	MULTIMODAL SPACE-BASED	MAO ET AL. 2015 [31]	0.600	0.410
FLICKR30K	ENCODER-DECODER ARCHITECTURE	JIA ET AL. 2015 [14]	0.646	0.466
MSCOCO	ENCODER-DECODER ARCHITECTURE	WU ET AL. 2018 [32]	0.740	0.560
MSCOCO	ENCODER-DECODER ARCHITECTURE	REN ET AL. 2017RL [33]	0.713	0.539
MSCOCO	ENCODER-DECODER ARCHITECTURE	LU ET AL. 2017 [22]	0.742	0.580

are standard measures for evaluating the quality of captions in comparison to the reference captions. For example, in the "Flickr8k" dataset, both "MULTIMODAL space-based" and "Encoder-Decoder Architecture" techniques were employed. The former utilized the approach of Mao et al. from 2015, while the latter used the approach of Jia et al. from the same year. For each technique, the corresponding BLEU-1 and BLEU-2 scores have also been provided. Similarly, performance metrics for the "Flickr30k" and "MSCOCO" datasets are included, indicating that various techniques and methods have been evaluated. In essence, the table provides a comparison of different techniques in image captioning across various datasets, which in turn helps assess their effectiveness in terms of BLEU scores.

5 Conclusion and Future Work

In this way, image captioning makes the visually impaired more accessible as it offers textual descriptions of pictures on websites. The access of images by content is facilitated by the text descriptions produced through image

captioning. In this paper, we conducted a systematic review and categorized the deep learning techniques that have spearheaded innovation in image captioning since its inception over a decade ago. Although there has been remarkable recent development in deep learning-based image captioning methods, the strong need for a fully robust approach to generate high-quality captions for almost all images is nonetheless an open problem. The constant appearance of novel deep architectures implies that research in automatic image captioning will be active for the foreseeable future.

References

[1] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14). 595–603.

[2] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. In Workshop on Neural Information Processing Systems (NIPS)).

[3] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR).

[4] Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.

[6] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3137–3146.

[7] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2016. Dense Captioning with Joint Inference and Visual Context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1978–1987.

[8] MD.Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for

Image Captioning. ACM Comput. Surv. 0, 0, Article 0 (October 2018), 36 pages.

[9] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 11–20.

[10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollą́r, Jianfeng Gao, Xiaodong He, Margaret Mitchell, and John C Platt. 2015. From captions to visual concepts and back. In proceedings of the IEEE conference on computer vision and pattern recognition. 1473–1482.

[11] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

[12] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[13] 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, 15–29.

[14] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.

[16] 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. 3156–3164.

[17] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4565–4574.

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition. 11–20.

[19] Minsi Wang, Li Song, Xiaokang Yang, and Chuanfei Luo. 2016. A parallel-fusion RNN-LSTM architecture for image caption generation. In Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 4448–4452.

[20] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2016. Dense Captioning with Joint Inference and Visual Context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 1978–1987.

[21] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In IEEE International Conference on Computer Vision (ICCV). 4904–4912.

[22] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3242–3250.

[23] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 1141–1150.

[24] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of Attention for Image Captioning. In Proceedings of the IEEE international conference on computer vision. 1251–1259.

[25] Jiuxiang Gu, GangWang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In Proceedings of the International Conference on Computer Vision (ICCV). 1231–1240.

[26] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 5263–5271.

[27] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) 1179–1195.

[28] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning

images with diverse objects. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1170–1178.
[29] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. 2017. Actor-critic sequence training for image captioning. arXiv preprint arXiv:1706.09601.
[30] Qingzhong Wang and Antoni B Chan. 2018. CNN+ CNN: Convolutional Decoders for Image Captioning. arXiv preprint arXiv:1805.09019.
[31] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR).
[32] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence 40, 6, 1367–1381.
[33] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep Reinforcement Learning-based Image Captioning with Embedding Reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 1151–1159.

Biographies

Monika Chawla completed her Master of Engineering (M.E.) in Computer Science from LIMAT, MD University, Rohtak, in 2007. She is currently pursuing doctoral research at Manav Rachna International Institute of Research and Studies in India, on a Ph.D. in Computer Science & Engineering.

She is currently working as an Assistant Professor in the School of Computer Applications, Manav Rachna International Institute of Research and Studies, India. Her research interests include machine learning, deep learning, image processing, artificial intelligence, and optimization methods. She has

published several research papers in national and international conferences and journals, and has also actively engaged in academic conferences, faculty development programs, and technical workshops.

She has contributed notably to software development, web technologies, and artificial intelligence applications, as well as to reviewing many of the top-ranked journals.

Rashmi Agrawal is PhD and UGC-NET qualified with 20 years of experience in teaching and research, working as Professor in Department of Computer Applications, Manav Rachna International Institute of Research and Studies, Faridabad, India. She is associated with various professional bodies in different capacities, life member of Computer Society of India and senior member of IEEE, She is book series editor of Innovations in Big Data and Machine Learning, CRC Press, Taylor and Francis group, USA and Advances in Cybersecurity in Wiley. She has authored/co-authored many research papers in peer reviewed national/international journals and conferences which are SCI/SCIE/ESCI/SCOPUS indexed. She has also edited/authored books with national/international publishers (Springer, Elsevier, IGI Global, Apple Academic Press, and CRC Press) and contributed chapters in books. Currently she is guiding PhD scholars in Sentiment Analysis, Educational Data Mining, Internet of Things, Brain Computer Interface and Natural language Processing. She is Associate Editor in Journal of Engineering and Applied Sciences and Array Journal, Elsevier.