Multimedia Systems (2023) 29:2951–2959
https://doi.org/10.1007/s00530-022-00902-0

SPECIAL ISSUE PAPER

Image and audio caps: automated captioning of background sounds and images using deep learning

M. Poongodi[1] · **Mounir Hamdi**[1] · **Huihui Wang**[2]

Received: 11 November 2021 / Accepted: 23 January 2022 / Published online: 26 February 2022

Abstract
Image recognition based on computers is something human beings have been working on for many years. It is one of the most difficult tasks in the field of computer science, and improvements to this system are made when we speak. In this paper, we propose a methodology to automatically propose an appropriate title and add a specific sound to the image. Two models have been extensively trained and combined to achieve this effect. Sounds are recommended based on the image scene and the headings are generated using a combination of natural language processing and state-of-the-art computer vision models. A Top 5 accuracy of 67% and a Top 1 accuracy of 53% have been achieved. It is also worth mentioning that this is also the first model of its kind to make this forecast.

Keywords Computer vision · Image to caption · Scene recognition · Image analysis · Social networks

1 Introduction

The ability to naturally depict the picture with legitimately framed English sentences is an atypical test error, but it could have an incredible effect, for example, by assisting visibly disabled individuals in better comprehending the pictures on the Internet. This mistake is mostly connected to the overall picture assembly or, on the other hand, the concession assignments, which were a critical component of the PC vision network. In fact, an impression should not only capture the items in a picture, but also express how these items interact with each other and their characteristics and exercises In addition, the semantic learning mentioned above must be linked to a characteristic dialect such as English, which means that, despite visual understanding, a display of a dialect is necessary. Most of the past efforts have proposed fastening the existing arrangements of the above sub-issues to transform them from a picture to a portrait. Interestingly, we could desire to show a unique joint model in this work that receives an image I as input and is prepared to increase the likelihood of constructing an objective disposition of words, each stemming from a guaranteed word reference, that adequately describes the picture. The fundamental impetus for our research arises from recent advances in machine interpretation, in which the goal is to convert a sentence S written in a source dialect to its target dialect interpretation T by enhancing the source dialect $p(T\|S)$.

Machine interpretation has also been accomplished for many years by arranging isolated errors (interpreting words separately, adjusting words, reordering, and so on). However, ongoing work has shown that interpretation should be possible using recurrent neural networks (RNNs) in a much less complex way and still achieve the best in class execution. The source sentence is examined by an RNN ”encoder”, which turns it into a rich longitudinal vector picture, which is employed as the underlying shrouded condition of an RNN ”decoder”, which generates the objective sentence. We propose to pursue this rich formula by replacing the RNN encoder with a deep convolutional neural system (CNN). It has been proved over the past several years that CNNs can produce a rich picture of an input image by

✉ M. Poongodi
dr.m.poongodi@gmail.com

Mounir Hamdi
mhamdi@hbku.edu.qa

Huihui Wang
hwang@sbu.edu

[1] Department of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[2] CyberSecurity Program, St. Bonaventure University, St. Bonaventure, NY 14778, USA

putting it into a fixed length vector, which can then be utilized for a number of purposes.

The fields of question recognition, speech recognition, and machine interpretation were reformed through the development of monstrous marked data sets and adapted profound portraits. However, the equivalent progress in regular sound comprehension assignments has not yet been achieved. We credit this incompletely for the absence of expansive marked sound data sets, which are often both expensive and questionable. We trust that substantial sound information can advance characteristic sound understanding, as well. In this paper, we use more than a year of in-the-wild sounds to learn semi-rich sound portrayals. We propose to scale up by taking advantage of the normal vision and sound synchronization to take an acoustic portrait of unlabeled videos. Unlabeled video has the favorable position that it can be gained financially on a monstrous scale, but still contains useful sound flags. Later progress in PC vision enabled machines to accurately perceive scenes and protests in images and recordings. We demonstrate to exchange visual learning in sound utilizing unlabeled video as a scaffold.

We present a profound convolutionary organization that adapts directly to raw waveforms of sound or, in other words, prepared by the exchange of information from vision to sound.

- In our studies, we show that the representation learned by our system best acquires class accuracy in three standard acoustic scene characterization datasets.
- It is possible, because we can use a lot of unlabeled sound information to prepare further systems without critical overfitting, and our analyses propose further models perform better.
- Perceptions of the illustration suggest that the system also adapts abnormal state locators, for example, by perceiving feathered tweets of creatures or cheering groups, even though it is prepared specifically from sound without ground truth marks.
- The research contained in this paper has a wide range of effects from simple projects to potential government websites to detect unwanted image content.

.

2 Related work

From this survey, we concluded that although many existing systems perform the task of either identifying the scene in the image or generating a caption for the image individually, the models have a low level of precision. It can also be noted that the models do not perform the task of generating the caption for the image and simultaneously recognizing the scene in the image. A model is needed that simultaneously performs the tasks with greater precision and this is the focus of this paper. The problem of creating regular dialect representations from visual information in PC vision has been examined for quite some time now, mainly for video. This has led to complex [1–5].

Frameworks made of visual crude identifiers and an organized formal dialect, e.g., oi possibly graphs or rational frameworks that are also changed to a regular dialect using standard-based frames. Such frameworks are strongly structured by hand, moderately weak and have only been illustrated in restricted areas, for example, scenes for movement or sports [6–14]. The problem of still image depiction with normal content has intrigued more late. The use of late advances in the recognition of articles, their qualities, and fields allows us to conduct regular dialect age frameworks, although they are limited in their expressiveness and others. Use identifiers to construe a triple of components of the scene or, in other words, messages using formats. Start with recognition and sort a last portrait using phrases containing identified articles, links.

A more unpredictable chart of past triplet identifications. However, with age of content based on format. All the more intense dialect models were also used depending on dialect parsing. The above methodologies have the ability to depict images ”in the wild”, but are strongly manually planned and inflexible with regard to the age of content. An expansive group of works has tended to place images for a given image. Such methods depend on the possibility of also inserting images into a similar vector space. For a picture question, pictures that lie near the picture in the inserting space are recovered. Most firmly, neural systems are used to co-implant pictures and phrases or even picture crops and subphrases, but they do not attempt to create new pictures When all is said in fact, the above methodologies can not depict already hidden items, even though the individual items can be seen in the preparation information. They also refrain from focusing on the question of how large a produced portrait is We join profound networks for picture order in this work with repetitive arrangement systems demonstrating the creation of a solitary system that produces pictures.

The RNN is prepared for this single ” end-to-end ” organization [15–24]. The model is driven by late succession age achievements in machine interpretation, with the distinction that we give a picture managed by a convolutionary net instead of starting with a sentence [25–38], who uses a neural net to predict the next word, given the picture and past words, but a feedforward one. A continuous work by [39] uses the repetitive NN for the equivalent assignment of expectations. This is essentially the same as the present proposition, since there are various vital contrasts: we use an even more intense RNN demonstration and specifically make the visual contribution to the RNN display, which

makes it workable for the RNN to monitor the content clarified items. Because of these seemingly unimportant contrasts, our framework generously achieves better results on the established benchmarks. Ultimately, [40] propose the creation of a common multimodal installation space using a ground-breaking PC vision display and an encoding LSTM. Instead of our methodology, they use two independent pathways (one for images, one for content) to characterize a joint implant and, although they can generate messages, their approach is deeply tuned for positioning [38, 41–53].

3 Proposed work

In this paper, we propose a neural and probabilistic structure for images. Later developments in machine interpretation have shown that given an intense succession. It is conceivable to achieve cutting-edge results by simply amplifying the likelihood of correct interpretation in a "conclusion to-end" form—both for preparation and deduction. These models make use of an intermittent neural system that codes the variable length contribution to a settled dimensional vector and use this representation to "interpret" the coveted yield phrase. In this respect, it is normal to use a similar methodology in which, given a picture (rather than an information sentence in the source dialect), a similar standard of "making an interpretation of" is applied in its representation. Subsequently, we propose to expand the probability of the right picture given the picture using the accompanying details

$$\theta^* = \text{argmax}\theta \sum (I, S)\log_p(S/I;\theta) \quad . \tag{1}$$

In Eq. (1), theta is our model's parameters, In a picture, S is its right translation. Since S speaks to a sentence, the length of it is unlimited. Accordingly, the affix guideline is usually applied to demonstrate the joint probability over S0, SN, and SN . A preparation precedent is combined at the time of preparation, and we improve the total log probabilities as shown in the entire preparation set using a stochastic angle plunge.

It is common to use a recurrent neural network (RNN) to show how a constant length hidden state or memory ht communicates the variable amount of words we need up to $t1$. After viewing additional information, this memory is updated by doing non-direct tasks. Two crucial structural choices must be made to enhance the above RNN: what is the right kind of f and how pictures and words are promoted as data sources xt. We employ a LongShort Term Memory (LSTM) network for f, which performed well in sequence class execution (see Fig. 1). This model is shown in the next section. To represent images, we employ a Convolution Neural Network (CNN). They were widely utilized and studied for film studios, and they are today

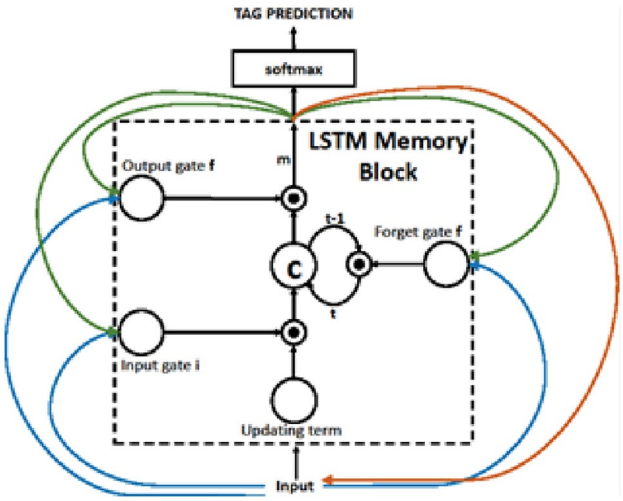

Fig. 1 Long short-term memory (LSTM architecture)

the finest in class for detecting and detecting demonstrations. CNN's judgment applies a novel technique to cluster standardization and produces the best results in the group rivalry [54]. Furthermore, they seemed to summarize several tasks, such as scene characterization using exchange techniques [55]. The sentences are uttered with an implantation model in mind.

3.1 Flowchart of the algorithm

The information layer is not made from complete neurons, but essentially comprises the qualities in the information record, which contribute to the next layer of neurons. The next layer is known as a hidden layer; there may be some hidden layers. The last layer is the yield layer where each class has a hub. A lonely breadth through the system results in an incentive to each yield hub and the record is relegated to which class hub has the highest appreciation.

3.2 Preparing an artificial network

In the preparation stage, the right class is known for each record (this is called the directed preparation), and the yield hubs can then be relegated to the " amend " values " 1 " for the right class hub and " 0 " for the others. By and by it, estimates of 0.9 and 0.1 were better discovered individually. It is therefore conceivable to think about the system's figured qualities for yield hubs to these " right " values and to determine a blunder for each hub (the " Delta " rule. These error terms are then used to modify the weights in the covered layers, so that ideally, whenever the yield is estimated, the " right " values are closer.

3.3 The iterative learning process

A key element of neural systems is an iterative learning process in which information cases (columns) are displayed to each system and the weights associated with information appreciation are balanced every time. In this stage of learning, the system learns by changing its weights, so that it can anticipate the right class mark of information tests. Neural system learning is also referred to as 'connectionist learning' of associations between the units. Neural systems 'favorable conditions incorporate their high resistance to boisterous information and their ability to arrange designs on which they have not been prepared. The most famous calculation of the neural system is the back-generating calculation proposed in the 1980s. When a system for a specific application is organized, this system is ready to be prepared. The underlying weights (represented in the following area) are arbitrarily selected to start this procedure. The preparation or learning begins at that point. The system, using the weights and capacities in the hidden layers, then thinks about the subsequent yields against the coveted yields, forms records in the preparation information. Errors are then generated through the framework, so that the framework changes the weights to be applied to the following record. This procedure repeatedly occurs when the weights are constantly changed. A similar arrangement of information is commonly handled during the preparation of a system, as the weights of the association are persistently refined. Note that you never learn a few systems. This could be based on the fact that the information does not contain the specific data from which the coveted output is determined. In addition, systems do not unite if there is not enough information to finish learning. Instead, sufficient information should be available to ensure that piece of information can be kept as an approval set.

3.4 Feedforward, back propagation

A few autonomous sources (Werbor; Parker; Rumelhart, Hinton, and Williams) produced the feedforward, backspread design in the mid-1970s. This free cooperation was the result of the expansion of articles and talks at various meetings that animated the entire business. For complex, multilayered systems, this synergistically created backproliferation engineering is currently the most prominent, powerful and simple tool demonstration. Its highest quality is in non-direct answers to difficult problems. The operation of the regeneration system has an information layer, a yield layer, and something like a hidden layer. There is no hypothetical containment point on the quantity of covered layers, but normally only a few. Some work has been done that shows that a maximum of five layers (one info layer, three covered layers, and a yield layer) are required to deal with problems of any complexity. Each layer is entirely related to the successor layer. As noted above, the preparation process typically uses some variation of the Delta rule, starting with the calculated contrast between the actual yields and the coveted yields. Using this blunder, association weights are increased to the extent that they are a scaling factor for worldwide accuracy in the times of error. This means that the data sources, the yield, and the coveted yield must all be available for a single hub in a similar handling component. The intricate part of this learning component is for the framework to determine which input has most contributed to off-base yield and how this component is changed to correct the error. An inert hub would not add to the error and would not have any reason to change its weights. To address this problem, input preparation is connected to the system information layer and the desired yields are analyzed at the yield layer. In the course of the learning process, a forward compass is produced by the system and the yield of each component is processed layer by level. The difference between the yield of the last layer and the coveted yield is backward to the previous layer(s), generally altered by the exchange subsidiary, and the weights of the association are regularly balanced using the Delta rule.

The measurement of accessible information sets the upper head for the amount of handling components in the covered layer(s). To determine this limit, use the quantity of cases in the information index and gap by the whole quantity of hubs in the information and yield layers in the system. At this point, the result is again isolated by a scaling factor in the range of five and ten. More important scaling factors are generally used for less loud information. The chance that you use such a large number of fake neurons remembers the preparation set. In the event that this happens, information speculation will not occur, making the system futile with new information indexes.

3.5 Preprocessing of images

As a first stage, the area of interest (in this example, the hand) is removed from the picture. The background noise is removed as a result of this procedure. Defining a binary mask for a particular area of interest is the best way to define a region of interest. After that, the mask is utilized to extract an item from the image. There are several data annotation tools that let us specify the coordinates of important spots to manually generate a mask. Manually annotating each picture would be a lengthy task due to the magnitude of the data collection for this study. The test data are also subjected to the same procedure. As a result, a more automated technique is sought. The photos in the data set are not equally contrasted, as demonstrated in Figs. 2 and 3. As a result, before the pictures are given into the semantic image segmentation model, they must be contrast equalized. To increase the picture quality and

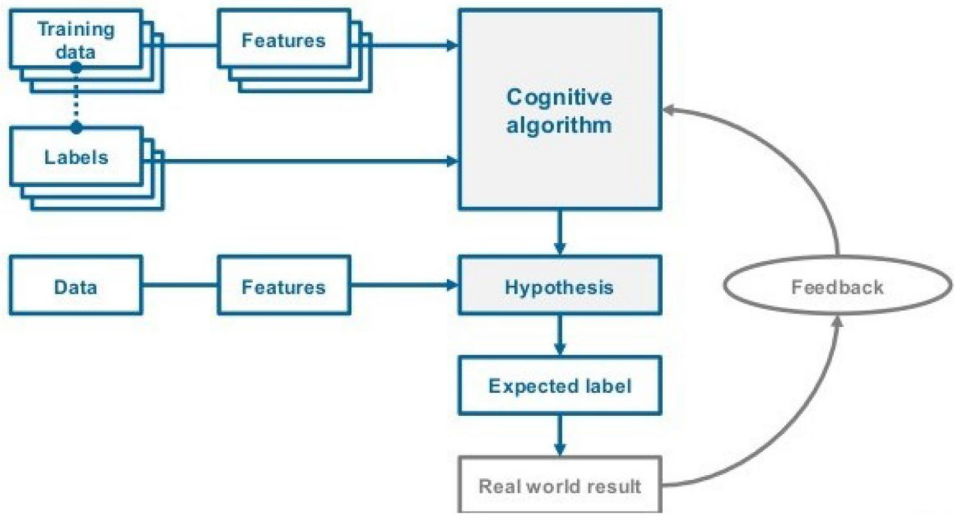

Fig. 2 System architecture of network

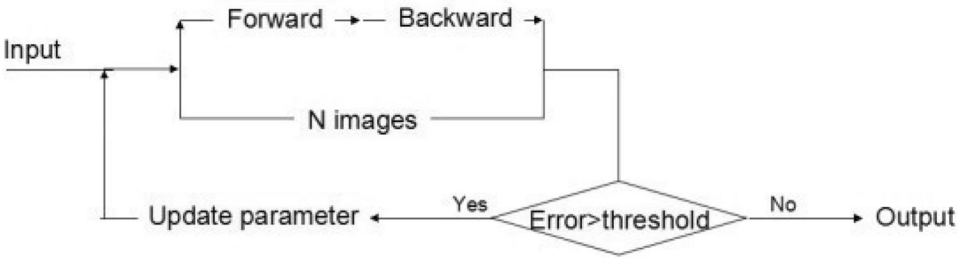

Fig. 3 Intial iteration of the deep learning network

Table 1 Dice coefficient results on a fourfold cross-validation split

Model	Split-0	Split-1	Split-2
Caption	47.23	49.27	52.43
Scene	53.29	54.21	53.89

characteristics, contrast amplification is performed. This is accomplished using a method called contrast-limited adaptive histogram equalization, which was developed by [56]. CLAHE (contrast-limited adaptive equalization of histograms) is a version of the [57] Adaptive Histogram approach for improving contrast. The first step is to convert the picture color space from RGB to LAB before using this method. In this color space, the L component stands for lightness, a for green–red, and b for yellow–blue. The goal of this color space is to bring human eyesight together. The initial step in using this technique is to convert the image's color space from RGB to the LAB color space. In this color space, the L component stands for lightness, a for green–red, and b for yellow–blue. The goal of this color space is to bring human eyesight together. It conforms to the human experience of lightness, although it ignores a few factors (Figs. 4, 5, and 6) (Table 1).

The equation of the loss function is given as

$$L = H - \log(\mathrm{J}). \tag{2}$$

3.6 Network structure

The network was trained for a duration on 3 days on AWS 2.8 × GPUs for 5 million iterations. The dataset was also cleaned for any inconsistencies. As it can be seen, the two images are pretty similar to each other. The more similar they are, the higher the accuracy of our developed model. This is a sign that a bit more training will lead to amazing results

4 Experimental results

4.1 Dataset description

The dataset was collected from various sources. First, the flikr image dataset with over 12000 images along with their captions. The dataset was divided as shown in Table 2.

Predictions

1. Type: inside

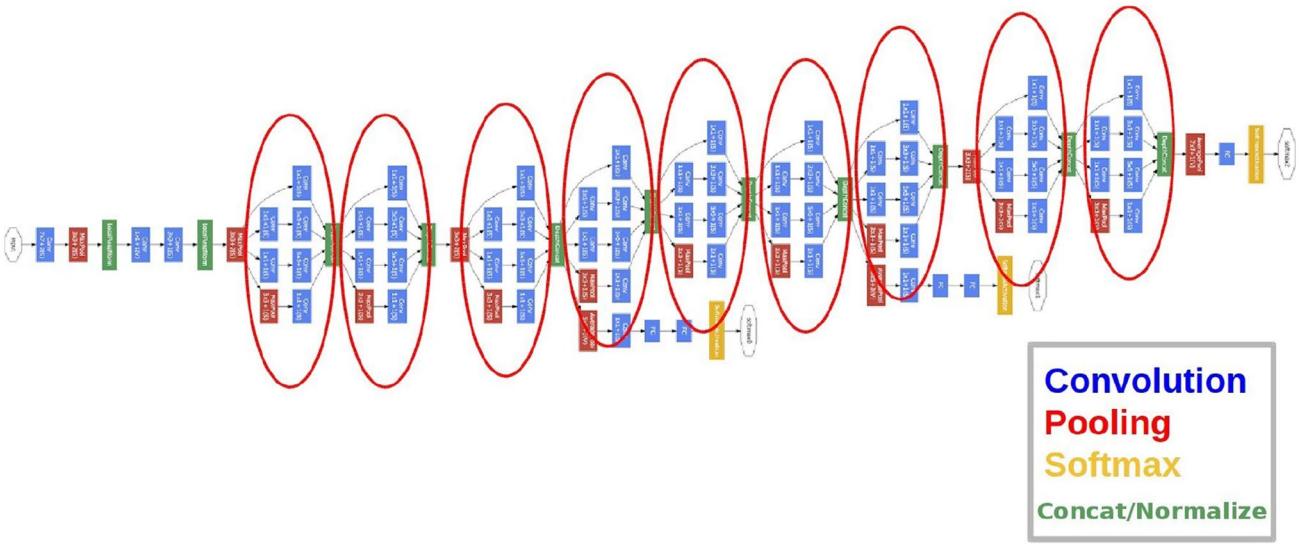

Fig. 4 Caption generator architecture

Fig. 5 Sample image taken as input

Fig. 6 Heatmap for prediction

Table 2 Description of the flikr 2K dataset

Train	Validation	Test
12364 images	1253 images	536 images
12019 images	958 captions	287 captions

Table 3 Model performance on flikr 2K dataset

Validation set for scene recognition			
Top 1% Acc (%)	Top 1% Acc (%)	Top 1% Acc (%)	Top 1% Acc (%)
53.42	83.21	48.23	56.47
67.2	72.83	62.32	63.13
48.23	54.43	46.23	56.93

2. Scene categories:eatery (0.690), people (0.163)
3. Scene attributes: no sun, closed area, artificial, speaking, inside illumination, cloth, congregation, speaking, working (Table 3).

5 Conclusion

From the Tiny Image dataset, to ImageNet [58] and Spots [59] , moreover, the rise of multi-million-thing datasets [60–62] has enabled hungry machine learning information calculations to achieve close human semantic characterization of visual examples as articles and scenes. With its class integration and high models diversity, setting to control advance on scene understanding issues. Such issues could incorporate deciding the activities occurring in a given

situation, spotting conflicting articles or then again human practices for a specific place, and foreseeing future occasions or the reason for occasions given a scene.

6 Future work

Since we have been so successful with this model, even though we have so few resources, this model certainly has a lot of potential. The authors see that this model is used in a wide range of applications from social networking sites to public websites. Since we have been so successful with this model, even though we have so few resources, this model certainly has a lot of potential. The authors see that this model is used in a wide range of applications from social networking sites to public websites. At present, we lack intelligent technology capable of detecting and understanding image content. The authors believe that this is the need for the hour and is highly imperative at a time when elections are even biased because of the text in the online images.

Funding Open Access funding provided by the Qatar National Library.

References

1. Baccour, E., Foufou, S., Hamila, R., Hamdi, M.: A survey of wireless data center networks. In: 2015 49th Annual conference on information sciences and systems (CISS), pp. 1–6. (2015). https://doi.org/10.1109/CISS.2015.7086853
2. Poongodi, M., Bose, S.: Design of intrusion detection and prevention system (IDPS) using DGSOTFC in collaborative protection networks. In: Fifth international conference on advanced computing (ICoAC), vol. 2013, pp. 172–178. (2013). https://doi.org/10.1109/ICoAC.2013.6921946
3. Mhamdi, L., Hamdi, M.: Scheduling multicast traffic in internally buffered crossbar switches. In: 2004 IEEE international conference on communications (IEEE Cat. No.04CH37577), vol. 2, pp. 1103–1107. (2004). https://doi.org/10.1109/ICC.2004.1312671
4. Poongodi, M., Vijayakumar, V., Rawal, B., Bhardwaj, V., Agarwal, T., Jain, A., Ramanathan, L., Sriram, V.P.: Recommendation model based on trust relations and user credibility. J. Intell. Fuzzy Syst. **36**(5), 4057–4064 (2019)
5. Poongodi, M., Hamdi, M., Vijayakumar, V., Rawal, B.S., Maode, M.: An effective electronic waste management solution based on blockchain smart contract in 5G communities. In: 2020 IEEE 3rd 5G World Forum (5GWF), pp. 1–6 (2020). https://doi.org/10.1109/5GWF49715.2020.9221346
6. Pun, K., Hamdi, M.: Distro: a distributed static round-robin scheduling algorithm for bufferless Clos-Network switches. In: Global telecommunications conference, GLOBECOM '02, vol. 3, IEEE, pp. 2298–2302. (2002). https://doi.org/10.1109/GLOCOM.2002.1189041
7. Poongodi, M., Vijayakumar, V., Chilamkurti, N.: Bitcoin price prediction using ARIMA model. Int. J. Internet Technol. Secur. Trans. **10**(4), 396–406 (2020)
8. Xia, Q., Hamdi, M., Letaief, K.B.: Open-loop link adaptation for next-generation IEEE 802.11n wireless networks. IEEE Trans. Veh. Technol. **58**(7), 3713–3725 (2009). https://doi.org/10.1109/TVT.2009.2013234
9. Poongodi, M., Bose, S.: Detection and Prevention system towards the truth of convergence on decision using Aumann agreement theorem. Proc. Comput. Sci. **50**, 244–251 (2015)
10. Wang, L., Wu, K., Xiao, J., Hamdi, M.: Harnessing frequency domain for cooperative sensing and multi-channel contention in CRAHNs. IEEE Trans. Wirel. Commun. **13**(1), 440–449 (2014). https://doi.org/10.1109/TWC.2013.120413.130767
11. Poongodi, M., Bose, S.: A firegroup mechanism to provide intrusion detection and prevention system against DDoS attack in collaborative clustered networks. Int. J. Inf. Secur. Priv. **8**(2), 1–18 (2014)
12. Xia, Q., Jin, X., Hamdi, M.: Cross layer design for the IEEE 802.11 WLANs: joint rate control and packet scheduling. IEEE Trans. Wirel. Commun. **6**(7), 2732–2740 (2007). https://doi.org/10.1109/TWC.2007.06019
13. Poongodi, M., Bose, S.: The COLLID based intrusion detection system for detection against DDOS attacks using trust evaluation. Adv. Nat. Appl. Sci **9**(6), 574–580 (2015)
14. Poongodi, M., Bose, S., Ganeshkumar, N.: The effective intrusion detection system using optimal feature selection algorithm. Int. J. Enterp. Netw. Manag. **6**(4), 263–274 (2015)
15. Lin, D., Liu, Y., Hamdi, M., Muppala, J.: FlatNet: towards a flatter data center network. In: 2012 IEEE global communications conference (GLOBECOM), pp. 2499–2504. (2012). https://doi.org/10.1109/GLOCOM.2012.6503492
16. Poongodi, M., Sharma, A., Hamdi, M., Maode, M., Chilamkurti, N.: Smart healthcare in smart cities: wireless patient monitoring system using IoT. J. Supercomput. 1–26 (2021)
17. Xia, Q., Hamdi, M.: Contention window adjustment for IEEE 802.11 WLANs: a control-theoretic approach. In: 2006 IEEE international conference on communications, pp. 3923–3928. (2006). https://doi.org/10.1109/ICC.2006.255694
18. Poongodi, M., Bose, S.: Stochastic model: reCAPTCHA controller based co-variance matrix analysis on frequency distribution using trust evaluation and re-eval by Aumann agreement theorem against DDoS attack in MANET. Cluster Comput. **18**(4), 1549–1559 (2015)
19. Ma, M., Hamdi, M.: Providing deterministic quality-of-service guarantees on WDM optical networks. IEEE J. Sel. Areas Commun. **18**(10), 2072–2083 (2000). https://doi.org/10.1109/49.887926
20. Poongodi, M., Hamdi, M., Malviya, M., Sharma, A., Dhiman, G., Vimal, S.: Diagnosis and combating COVID-19 using wearable Oura smart ring with deep learning methods. Pers. Ubiquitous Comput. 1–11 (2021)
21. Hamdi, M., Lee, C.K.: Dynamic load-balancing of image processing applications on clusters of workstations. Parallel Comput. **22**(11), 1477–1492 (1997). https://doi.org/10.1016/S0167-8191(96)00054-3. (**ISSN 0167-8191**)

22. Poongodi, M., Hamdi, M., Varadarajan, V., Rawal, B.S., Maode, M.: Building an authentic and ethical keyword search by applying Decentralised (Blockchain) verification. In: IEEE INFOCOM 2020-IEEE conference on computer communications workshops (INFOCOM WKSHPS), IEEE, pp. 746–753 (2020)
23. Pan, Y., Hamdi, M.: Quicksort on a linear array with a reconfigurable pipelined bus system. In: Proceedings second international symposium on parallel architectures, algorithms, and networks (I-SPAN'96), pp. 313–319. (1996). https://doi.org/10.1109/ISPAN.1996.508999
24. Jeyachandran, A., Poongodi, M.: Securing cloud information with the use of Bastion Algorithm to enhance confidentiality and protection. Int. J. Pure Appl. Math. 118(24) (2018)
25. Wang, T., Su, Z., Xia, Y., Qin, B., Hamdi, M.: NovaCube: a low latency Torus-based network architecture for data centers. IEEE Global Commun. Conf. **2014**, 2252–2257 (2014). https://doi.org/10.1109/GLOCOM.2014.7037143
26. Poongodi, M., Al-Shaikhli, I.F., Vijayakumar, V.: The probabilistic approach of energy utility and reusability model with enhanced security from the compromised nodes through wireless energy transfer in WSN. Int. J. Pure Appl. Math. **116**(22), 233–250 (2017)
27. Poongodi, M., Vijayakumar, V., Ramanathan, L., Gao, X.-Z., Bhardwaj, V., Agarwal, T.: Chat-bot-based natural language interface for blogs and information networks. Int. J. Web Based Communities **15**(2), 178–195 (2019)
28. Poongodi, M., Malviya, M., Hamdi, M., Vijayakumar, V., Mohammed, M.A., Rauf, H.T., Al-Dhlan, K.A.: 5G based blockchain network for authentic and ethical keyword search engine. IET Commun. (2021)
29. Xia, Q., Hamdi, M.: Smart sender: a practical rate adaptation algorithm for multirate IEEE 802.11 WLANs. IEEE Trans. Wirel. Commun. **7**(5), 1764–1775 (2008). https://doi.org/10.1109/TWC.2008.061047
30. Poongodi, M., Malviya, M., Hamdi, M., Rauf, H.T., Kadry, S., Thinnukool, O.: The recent technologies to curb the second-wave of COVID-19 pandemic. IEEE Access **9**, 97906–97928 (2021)
31. Wang, T., Su, Z., Xia, Y., Hamdi, M.: Rethinking the data center networking: architecture, network protocols, and resource sharing. IEEE Access **2**, 1481–1496 (2014). https://doi.org/10.1109/ACCESS.2014.2383439
32. Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., Alyamani, H.: New York city taxi trip duration prediction using MLP and XGBoost. Int. J. Syst. Assur. Eng. Manag. 1–12 (2021)
33. Chan, M.-K., Hamdi, M.: An active queue management scheme based on a capture-recapture model. IEEE J. Sel. Areas Commun. **21**(4), 572–583 (2003). https://doi.org/10.1109/JSAC.2003.810499
34. Rawal, B.S., Manogaran, G., Singh, R., Poongodi, M., Hamdi, M.: Network augmentation by dynamically splitting the switching function in SDN. In: 2021 IEEE international conference on communications workshops (ICC Workshops), IEEE, pp. 1-6 (2021)
35. Pan, Y., Li, K., Hamdi, M.: An improved constant-time algorithm for computing the Radon and Hough transforms on a reconfigurable mesh. IEEE Trans. Syst. Man. Cybern. Part A Syst. Hum. **29**(4), 417–421 (1999). https://doi.org/10.1109/3468.769762
36. Poongodi, M., Nguyen, T.N., Hamdi, M., et al.: A measurement approach using smart-IoT based architecture for detecting the COVID-19. Neural Process. Lett. (2021). https://doi.org/10.1007/s11063-021-10602-x
37. Lin, D., Liu, Y., Hamdi, M., Muppala, J.: Hyper-BCube: a scalable data center network. IEEE Int. Conf. Commun. **2012**, 2918–2923 (2012). https://doi.org/10.1109/ICC.2012.6363759
38. Poongodi, M., Nguyen, T.N., Hamdi, M., Cengiz, K.: Global cryptocurrency trend prediction using social media. Inf. Process. Manag. **58**(6), 102708 (2021). https://doi.org/10.1016/j.ipm.2021.102708
39. He, H., Yang, H.: Deep visual semantic embedding with text data augmentation and word embedding initialization. Math. Probl. Eng. **2021**, 6654071 (2021). https://doi.org/10.1155/2021/6654071
40. Gong, X., Liu, X., Li, Y., Li, H.: A novel co-attention computation block for deep learning based image co-segmentation. Image Vis. Comput. **101**, 103973 (2020)
41. Alharbi, A., Alyami, H., Poongodi, M., Rauf, H.T., Kadry, S.: Intelligent scaling for 6G IoE services for resource provisioning. PeerJ Comput. Sci. **7**, e755 (2021)
42. Song, H., Liu, Y., Wang, J.: UAS detection and negation. U.S. Patent US 2021/0197967 A1, Jul. 1 (2021)
43. Yue, X., Liu, Y., Wang, J., Song, H., Cao, H.: Software defined radio and wireless acoustic networking for amateur drone surveillance. IEEE Commun. Mag. **56**(4), 90–97 (2018). https://doi.org/10.1109/MCOM.2018.1700423
44. Yang, J., Wang, C., Jiang, B., Song, H., Meng, Q.: Visual perception enabled industry intelligence: state of the art, challenges and prospects. IEEE Trans. Ind. Inform. **17**(3), 2204–2219 (2021). https://doi.org/10.1109/TII.2020.2998818
45. Jiang, B., Yang, J., Lv, Z., Song, H.: Wearable vision assistance system based on binocular sensors for visually impaired users. IEEE Internet Things J. **6**(2), 1375–1383 (2019). https://doi.org/10.1109/JIOT.2018.2842229
46. Song, H., Srinivasan, R., Sookoor, T., Jeschke, S.: Smart Cities: Foundations, Principles and Applications, pp. 1–906. Wiley, Hoboken (2017).. (**ISBN: 978-1-119-22639-0**)
47. Sun, Y., Song, H., Jara, A.J., Bie, R.: Internet of Things and Big Data analytics for smart and connected communities. IEEE Access **4**, 766–773 (2016). https://doi.org/10.1109/ACCESS.2016.2529723
48. Song, H., Rawat, D., Jeschke, S., Brecher, C.: Cyber-Physical Systems: Foundations, Principles and Applications, pp. 1–514. Academic Press, Boston (2016).. (**ISBN: 978-0-12-803801-7**)
49. Liu, Y., Wang, J., Li, J., Niu, S., Song, H.: Class-incremental learning for wireless device identification in IoT. IEEE Internet Things J. (2021). https://doi.org/10.1109/JIOT.2021.3078407
50. Liu, Y., et al.: Zero-bias deep learning for accurate identification of Internet-of-Things (IoT) devices. IEEE Internet Things J. **8**(4), 2627–2634 (2021). https://doi.org/10.1109/JIOT.2020.3018677
51. Liu, Y., Wang, J., Niu, S., Song, H.: Deep learning enabled reliable identity verification and spoofing detection. In: Yu, D., Dressler, F., Yu, J. (eds.) Wireless Algorithms, Systems, and Applications: WASA 2020: Lecture Notes in Computer Science, vol. 12384. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59016-1_28
52. Niu, S., Liu, Y., Wang, J., Song, H.: A decade survey of transfer learning (2010–2020). IEEE Trans. Artif. Intell. **1**(2), 151–166 (2020). https://doi.org/10.1109/TAI.2021.3054609
53. Liu, Y., Wang, J., Li, J., Niu, S., Song, H.: Machine learning for the detection and identification of Internet of Things (IoT) devices: a survey. IEEE Internet Things J. (2021). https://doi.org/10.1109/JIOT.2021.3099028
54. Liu, M., Li, L., Hu, H., Guan, W., Tian, J.: Image caption generation with dual attention mechanism. Inf. Process. Manag. **57**(2), 102178 (2020)
55. Katiyar, S., Borgohain, S.K.: Comparative evaluation of CNN architectures for image caption generation. arXiv preprint. arXiv:2102.11506 (2021)
56. Kumar, A., Verma, S.: CapGen: a neural image caption generator with speech synthesis. In: Data Analytics and Management, pp. 605–616. Springer, Singapore (2021)
57. Xia, P., He, J., Yin, J.: Boosting image caption generation with feature fusion module. Multimedia Tools Appl. **79**(33), 24225–24239 (2020)
58. Zeng, X., Wen, L., Liu, B., Qi, X.: Deep learning for ultrasound image caption generation based on object detection. Neurocomputing **392**, 132–141 (2020)

59. Cheng, L., Wei, W., Mao, X., Liu, Y., Miao, C.: Stack-VS: stacked visual-semantic attention for image caption generation. IEEE Access **8**, 154953–154965 (2020)
60. Liu, X., Xu, Q.: Adaptive attention-based high-level semantic introduction for image caption. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **16**(4), 1–22 (2020)
61. Zhang, J., Li, K., Wang, Z., Zhao, X., Wang, Z.: Visual enhanced gLSTM for image captioning. Expert Syst. Appl. **184**, 115462 (2021)
62. Sur, C.: aiTPR: attribute interaction-tensor product representation for image caption. Neural Process. Lett. **53**(2), 1229–1251 (2021)