IMAGE INSCRIPTION AND INTONATION - A NEURAL NETWORK APPROACH

Arunachalam. T.K.S
Computer Science and Engineering
College of Engineering – Guindy, Chennai
tksarunachalam2508@gmail.com
Srihari. S
Computer Science and Engineering
College of Engineering – Guindy, Chennai
srihari961112@gmail.com

Publication History
Manuscript Reference No: **IRJCS/RS/Vol.08/Issue08/AUCS10082**
Received: 25, July 2021
Accepted: 03, August 2021
Published: 09, August 2021
Citation: Arunachalam & Srihari (2021). IMAGE INSCRIPTION AND INTONATION - A NEURAL NETWORK APPROACH. International Research Journal of Computer Science, VIII, 172-177.
doi: https://doi.org/10.26562/irjcs.2021.v0808.003
Peer-review: Double-blind Peer-reviewed
Editor: Dr.A.Arul Lawrence Selvakumar, Chief Editor, IRJCS, AM Publications, India

Abstract: Digital communication technologies have greatly influenced and expanded the way humans interact. The progress of information technology has opened wider opportunities for communication. Social networks have become the modern-day social communities connecting people from different parts of the globe, sharing images and videos on these platforms. By creating virtual communities, digital communication has expanded the scope of communication eliminating barriers. We aim to make further progress in this arena by describing an image in the form of audio to visually impaired people. A certain section of differently abled people is unfortunately isolated from this world. In-order to combat this issue we have come up with a system that describes an image shown in the form of plain text using an encoder-decoder architecture and is integrated with an end-to-end lexical articulator which produces a vocal description of the given image.

Keywords: neural network image inscription intonation

I. INTRODUCTION

Technology has become an integrated part of our daily lives over the past decades. The efficient processing and association of different multimodal information is a very important research field with a great variety of applications, such as human computer interaction, knowledge discovery, document understanding, etc. Computerized elucidation of an image has been one of the primary goals of computer vision. Not only must description generator models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language. Hence, to tackle this conundrum we present the development of a novel methodology to extract meaningful information from images, in the form of short descriptions. The results can be further run through a lexical articulator engine to offer full sustainability. This way, a fully independent experience could be delivered to visually impaired people. The dataset which we are planning to use for training and testing the model is MS-COCO 2014 (Microsoft Common Objects in Context) which contains more than 80000 labelled images. For converting still images into natural language text sentences we must first start from the understanding of the context of an image and secondly how this context is expressed into natural language. Thus, for understanding the image a feature extracting convolutional neural network can be used. With the aid of recurrent neural networks, the extracted features can be transformed into a suitable textual description. These results are finally fed to a lexical articulator module which produces the output of the system in the form of speech. Portraying the contents of a picture using precisely framed sentences is fundamentally harder than the deeply scrutinized arenas like object recognition and classification of images. The greatest test is having the option to make a portrayal that should catch the articles contained in a picture, yet in addition express how these objects identify with one another. The goal of image inscribing is to automatically depict an image in natural language sentences. This is a task that facilitates the amalgamation of computer vision and natural language processing, so its principal challenges emerge from the need of translating between two discrete, yet typically combined modalities.

In the first place, it is important to identify objects on the scene and decide the relations among them and then express the picture content accurately with meaningful sentences. The manner in which the descriptions are formulated is still very different from how humans depict pictures since individuals depend on good judgment and experience, and call attention to significant subtleties. Automatically produced depictions can likewise be utilized for content-based recovery or in social media communications.

II. RELATED WORK

[1] A framework of encoder/decoder system with attention mechanism has been considered to overcome the limitation of considering the image's scene as a whole. Attention mechanism considers the spatial aspects of the image and allows the decoding process to focus on the emphasis and details of the input image at each time step while the output sequences are being produced. Inception V3 model has been used as the encoder and a GRU with state size of 512 is being using in the decoder. The model has been compared with Microsoft's Caption Bot, an online caption generator. This does not reflect the true nature of the system since a standard metric has to be found to calculate the performance of the system.
[2] A neural framework has been proposed for generating captions from images derived from probability theory. By using a powerful mathematical model, the probability of the correct translation for both inference and training has been maximized and better results are obtained. The accuracy of model and command of language model learnt from image descriptions is tested on different datasets. Sometimes the generated sentence seems to lose track or differ completely from that of the original image content.
[3] Analysed different deep neural network-based image caption generation approaches and pretrained models to conclude on the efficient model with fine-tuning. A number of pre-trained models like VGG 16, RESNET and Inception were used to compare the results. Performed a comparative study by which the approach encompassing attention is found to perform better. The tanh activation in attention gives the smoother parts of the sub region over the inner dot product approach. Used BLEU-1 as the sole metric to evaluate the performance of the system which fails to provide the appropriate view.
[4] Based on a recurrent neural network with modified LSTM cells with an additional gate responsible for image features. This modification results in generation of more accurate captions. The Mixed-7c layer from Google Inception had been chosen and appended the average pooling layer having a 2048-dimensional output for image description. The probability of the correct caption for the given image is maximized by optimizing the sum of the log probabilities for the whole training set using stochastic gradient descent.
[5] The attention-based approach improved caption generation accuracy, but it did not always focus on appropriate image parts. The attention accuracy is uncorrelated to the caption accuracy directly. The ordinary attention mechanism does not help us to understand how a caption generation system generates a word.

III. METHODOLOGY

An iterative approach had been chosen for the implementation of the problem statement. The first goal was to understand the nature of the dataset and pre-process it. The input MS- COCO 2014 dataset is of size 25 GB. In-order to deal with this huge dataset and the constrained computing resources we make use of the dynamic programming paradigm by caching the values, the first time the dataset is downloaded, in-order to make access faster the subsequent times.

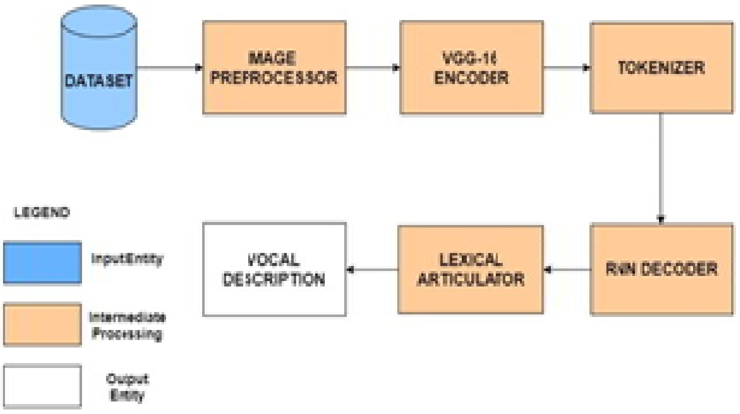


Fig. 1 Abstract System Architecture

A. Image Pre-processor

On loading the dataset, the images and the corresponding captions are then segregated and stored separately. The images then undergo normalization followed by scaling to finish the pre-processing. On the other hand, the captions are encoded in a dictionary and are thus pre-processed so that it could be used by the tokenizer. Caching is used to persist the data so it can be reloaded very quickly and easily.

The COCO data-set contains a large number of images and various data for each image stored in a JSON-file. On loading the image, it is resized as specified. Further the images are scaled so that their pixels fall between 0.0 and 1.0. The pycocotools have been put into use. It is a Python API that assists in loading, parsing and visualizing the annotations in COCO.

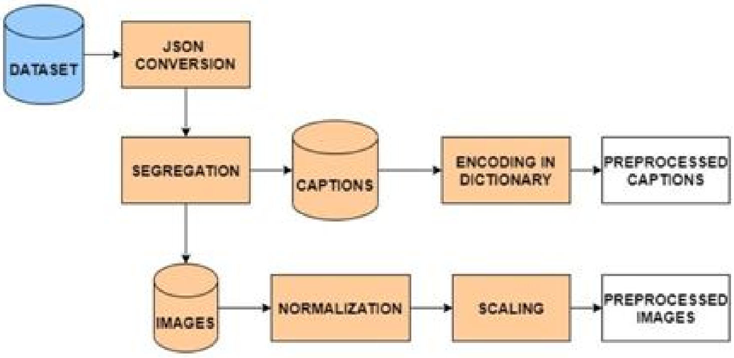


Fig. 2 Image Pre-processor Architecture

B. VGG-16 Encoder

In caption generation, we have to extract features from raw images. A pretrained convolutional neural network is used with the help of transfer learning thereby reducing the training time and utilizing the large amounts of training data. We employ VGG16 as the pretrained convolutional neural network. VGG16 has been trained with ImageNet dataset already and can classify images into 1000 object classes. It extracts enough features from an input image and the features are useful to generate the description of an image. The architecture of VGG16 consists of two modules: a feature extraction module which consists of 13 convolutional layers and a classification module which consists of 3 fully-connected layers. We capture the 15th layer in VGG16 as a feature vector of an input image representing the final output from this feature extraction module. The size of an image feature vector is 4096 x 1 x 1.

C. Tokenizer

As a consequence of the inability of neural networks to straightforwardly work on textual-information, a two-step process has been adopted to translate the text into numerical form. Before handling the content, the start and the end of each text series is set apart to monitor the captions. Then the process is commenced by changing the text into integer. The token-sequences are padded with zeros such that each has the same length before being fed to the decoder. This is followed by converting the integer-tokens into vectors of floating-point numbers using an embedding-layer. Integer tokens take upon values among 0 and the size of the vocabulary, but the recurrent neural network can't deal with values in such a wide range. Word embeddings is an alternate to one-hot encoding along with dimensionality reduction. This sequence of real valued vectors obtained from the embedding layer helps us to understand the words, their correlation in a better way.

D. RNN Decoder

Recurrent neural networks are used for decoding the extracted features from the image. Gated Recurrent Unit (GRU) is a sophisticated recurrent unit used to capture dependencies of various time scales, process memories of sequential data by storing previous inputs in the internal state of networks and plan from the history of previous inputs to target vectors in principle. GRU consists of Update and reset gates, these gates are responsible for regulating the information to be kept or discarded at each time step.

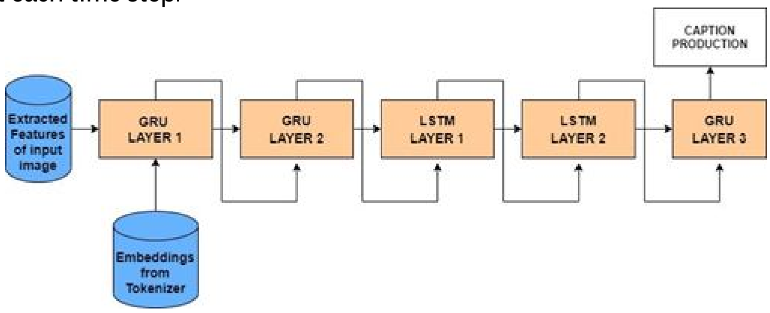


Fig. 3 RNN Architecture

E. Lexical Articulator

The final module of this system brings into play a state-of-the-art Lexical Articulator. Being fed with the textual description of images from the decoder, the language to be translated into is stipulated. Further we specify the pace of the vocal description we wish to obtain. Next, we create an mp3 instance of the description using the GTTS package. The terminal step in this module is the creation of an audio object which produces the required vocal output and enables us to play the articulated description.

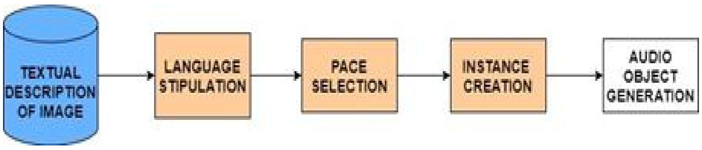
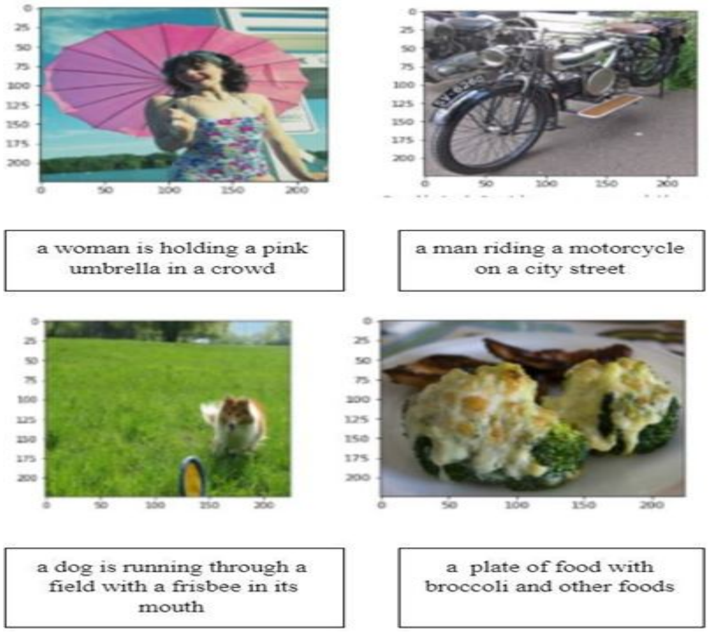


Fig. 4 Abstract System Architecture

IV. RESULTS AND ANALYSIS

A. Results obtained on the two different algorithms

Using VGG16 model as an encoder with 16 hidden layers and GRU network (using 3 GRU layers) as decoder with number of epochs set to 300 for optimum performance taking 82783 images from MS-COCO dataset as training dataset and vocabulary size of 10000 unique words, Fig.5 shows prediction for an image in validation set.



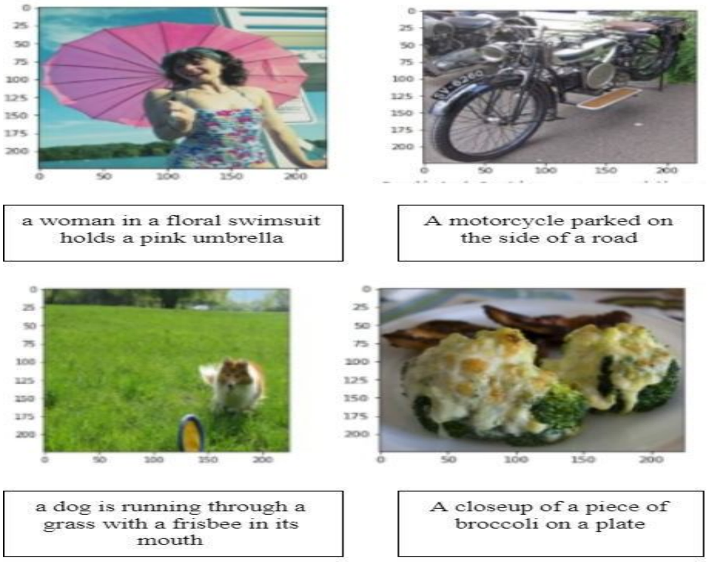
a woman is holding a pink umbrella in a crowd

a man riding a motorcycle on a city street

a dog is running through a field with a frisbee in its mouth

a plate of food with broccoli and other foods

Fig. 5 Predicted caption and generated audio using Algorithm – 1



a woman in a floral swimsuit holds a pink umbrella

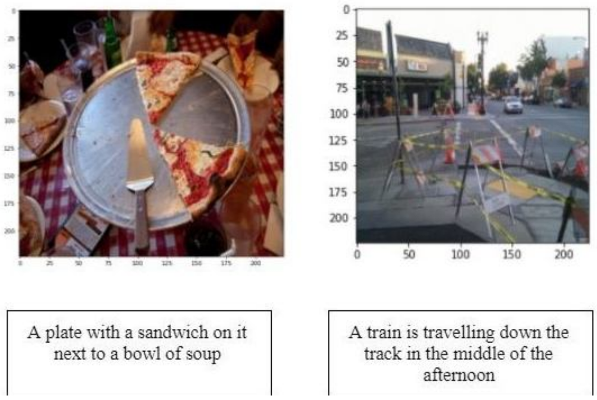
A motorcycle parked on the side of a road

a dog is running through a grass with a frisbee in its mouth

A closeup of a piece of broccoli on a plate

Fig. 6 Predicted caption and generated audio using Algorithm - 2

Using VGG16 model as an encoder with 16 hidden layers and the decoder network comprising of a fusion of 2 GRU layers, 2 LSTM layers and 1 GRU layer, with number of epochs set to 300 taking 82873 images from MS-COCO dataset as training dataset and vocabulary size of 10000 unique words, Fig.6 depicts outputs are observed. There are certain images where the performance of the model is just above par like those in Fig.7. While for some images it needs more training.



A plate with a sandwich on it next to a bowl of soup

A train is travelling down the track in the middle of the afternoon

Fig.7 Outputs that can be improved

B. Analysing the effect of dropout

The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. Inputs not set to 0 are scaled up by 1/(1- rate) such that the sum over all inputs is unchanged. The dropout parameter in the Long Short Term Memory cells has This led to an increase in the training loss due to the penalty imposed by regularization. Adding dropout makes the model generalize well for unseen data.

C. Analysing the effect of the number of layers

It is observed that the time taken to train the model for an epoch increases as the number of layers in the neural network increases. In the Algorithm-1 where the decoder consisted of 3 Gated Recurrent Units it took around 480 seconds to train for an epoch. Upon the addition of 2 Long Short-Term Memory cells, there was a 25% increase in the training time, where it took around 600 seconds to train for an epoch. The larger the number of epochs, the more time it takes for the algorithm to train, but it also raises the chances of resulting in a better map function. Sparse categorical cross entropy is the best loss function to use as it produces the index of a category index of the most likely matching category and not the one hot encoder (10,000 sized vector). This saves a lot of memory used for computation and representation. Generally, Adam optimizer is believed to work well for Computer Vision tasks, but we found out RMSprop to con verge efficiently for our tasks. We have used RMSprop with the learning rate of 0.001.

D. Analysis of Performance Metrics

From Table. I and Table. II it can be inferred that the algorithm 2 where the decoder is composed of a combination of 2 Gated Recurrent Units, 2 Long Short-Term Memory and 1 Gated Recurrent Unit performs better than the algorithm 1 where the decoder is composed of 3 Gated Recurrent Units. The two algorithms are compared among 5 metrics - BLEU, METEOR, ROUGE-L, CIDER and SPICE and from Table.II we can observe that the Algorithm 2 outperforms the other in all the metrics.

TABLE I - COMPARISON OF BLEU SCORES

Metric	Algorithm-1	Algorithm-2
BLEU-1	0.497	0.580
BLEU-2	0.297	0.406
BLEU-3	0.179	0.280
BLEU-4	0.103	0.193

TABLE II - COMPARISON OF PERFORMANCE METRICS

Metric	Algorithm-1	Algorithm-2
BLEU-4	0.103	0.193
METEOR	0.165	0.195
ROUGE-L	0.318	0.396
CIDER	0.471	0.600
SPICE	0.128	0.133

V. CONCLUSIONS

We have fabricated an end-to-end neural network system which mechanically views an image and provides a suitable description in simple language. It is based on a convolution neural network which, by encoding the input image into a compact representation, forwards it to a recurrent neural network that generates a corresponding description. The generated description is then being fed to a state-of-the-art lexical articulator which produces the corresponding vocalized version. Thus, a visual image is mapped to a textual narrative. It is observed that the additional LSTM cells added to work increases the model efficiency. Two models, one comprising only the GRU and the other the fusion of LSTM and GRU have been trained on the same MSCOCO image train dataset. The final loss value recorded is 0.67 for the algorithm involving only GRU and 0.60 for the other algorithm involving the amalgamation of GRUs and LSTMs.

The measurement of average value of different metrics on the same dataset with these different models was done. These metrics have shown that the new RNN model can generate the image captions more accurately. This new framework provides users with controllability in generating intended captions for images, which may inspire exciting applications. Albeit the depictions produce energizing outcomes, we believe this is only the start. The future heading will consider a framework that can all the more explicitly portray the recognizable depictions of traffic signs and clinical pictures. Applying an unsupervised approach for both images as well as text to enhance the image inscribing process seems quite promising. More explorations can be done to the natural language processing when dealing with the formation of sentences in-order to ensure a cogent description.

REFERENCES

1. A. Hani, N. Tagougui, and M. Kherallah, "Image caption generation using a deep architecture," in 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246–251
2. C. Amritkar and V. Jabade, "Image caption generation using deep learning technique," in 2018 Fourth International Confer- ence on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–4.
3. V. Kesavan, V. Muley, and M. Kolhekar, "Deep learning based automatic image caption generation," in 2019 Global Confer- ence for Advancement in Technology (GCAT), 2019, pp. 1–6.
4. A. Poghosyan and H. Sarukhanyan, "Short-term memory with read-only unit in neural image caption generator," in 2017 Computer Science and Information Technologies (CSIT), 2017, pp. 162–167.
5. M. Shozu and H. Yanagimoto, "Attention analysis in caption generation," in 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), 2019, pp. 95–98.
6. A. Puscasiu, A. Fanca, D.-I. Gota, and H. Valean, "Automated image captioning," in 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), 2020, pp. 1–6.
7. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2016.
8. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651–4659.
9. V. Atliha and D. Sˇesˇok, "Comparison of vgg and resnet used as encoders for image captioning," in 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 2020, pp. 1–4.
10. M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion rnnlstm architecture for image caption generation," in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 4448–4452.
11. C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in 2019 IEEE International Conference on Data Mining (ICDM), 2019, pp. 728–737.
12. B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 256–270, 2020.
13. G. Hoxha, F. Melgani, and J. Slaghenau ffi, "A new cnnrnn framework for remote sensing image captioning," in 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), 2020, pp. 1–4.
14. M. Yang, J. Liu, Y. Shen, Z. Zhao, X. Chen, Q. Wu, and C. Li, "An ensemble of generation- and retrieval-based image captioning with dual generator generative adversarial network," IEEE Transactions on Image Processing, vol. 29, pp. 9627– 9640, 2020.
15. N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic- oriented image captioning based on order-embedding," IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743– 2754, 2019.
16. M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 32–44, 2019.
17. S. Takada, R. Togo, T. Ogawa, and M. Haseyama, "Generation of viewed image captions from human brain activity via unsupervised text latent space," in 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 2521–2525.
18. S. M. Xi and Y. I. Cho, "Image caption automatic generation method based on weighted feature," in 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), 2013, pp. 548–551.
19. Y. Feng and M. Lapata, "Automatic caption generation for news images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797–812, 2013.
20. Y. Zhenyu and Z. Jiao, "Image caption method combining multi-angle with multi-modality," in 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), 2019, pp. 24–30.