IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DESKTOP BASED APPLICATION FOR CAPTIONING IMAGE AND VIDEO USING DEEP LEARNING

[1]Fazilhaq Akhondzada, [2]Mr. Hemant Sharma,

[1]Research scholar, [2].Assistance Professor,
[1] Department of Computer science & Engineering,
[1]Vivekananda Global University, Jaipur, India

Abstract: Methodologies that depend on them Deep Learning has a lot of potential for applications that produce captions or a short summary for images and video frames automatically. In imaging technology, image and video captioning are mentally difficult challenges. Automated caption development for photos and videos for human being with different levels of vision impairment; Among the technology fields are automatic metadata production for videos and images that can be used by online services; robot vision systems; and many others.in The focus of this research is on the algorithmic overlap between images and videos discerption.
Index Terms **– Caption, Image Caption, Video Caption, CNN, RNN, LSTM**

1. INTRODUCTION

The role of image processing has been played and shall continue to be played. There are many uses, including facial recognition [4] and scene perception, to name a few. However, most researchers have relied on imaging techniques that performed best on static structures in controlled environments using advanced hardware [32][33]. Deep learning-based CNNs have had a huge effect on the field of image captioning in recent years, making for a lot more versatility. In the sense of deep learning, we will aim to illustrate recent developments in the area of image and video processing. Several researchers have shown interest in improving deep learning model architecture, implementations, and analysis since 2012. Deep learning science and technique have been around for many years, but recently, a growing amount of digital data and the use of strong graphics processing units have intensified growth of deep learning technology [2]. For humans, creating a descriptive scene in a picture or video clip is a crucial activity [1]. Researchers have been experimenting with ways to combine the science of interpreting human language with the science of automated retrieval and processing of data in order to create devices with this capability. Owing to the additional task of identifying the objects and behaviors in the image and constructing concise sentences based on the contents identified, captioning of images and videos require more work than recognition of image. The development of this method opens up many possibilities in a variety of application areas of our daily lives, such as assisting persons with vision impairments, self-driving cars, sign language processing, human-robot communication and interaction, and so on. So, when it comes to automatically creating sentences that describe a picture and a clip of video, there are usually 2 components: an encoder and a decoder. We have used CNN, RNN, and LSTM in this case. A convolutional Neural Network is used by the Encoder to retrieve artefacts and characteristics from an image and clip of video. A neural network is required for the decoder to produce a natural sentence based on the available data.

2. Methodology

Encoder and Decoder are two modules that are used to automatically generate natural language sentences that describe a picture and clip of video. Each part's architecture is described in detail here. The Encoder makes use of a convolutional Neural Network to extract artefacts and features from the image. A picture in a video or photograph A neural network is required for the decoder to produce a natural sentence.

Convolutional Neural Network: To figure out about thousands of artefacts from a finite number of photos [1], a model with a large learning ability is needed. Deep learning [2], [3] describes computer models that are made up of several computing layers that are used to gain image-based data representations. Convolutional Neural Networks based on deep learning are used in a variety of applications, including image identification, Image identification is used to operate a wide range of visual functions, including comprehending image information. There are a number of well-known CNN models [2] focused on object recognition [4], [5], [6], and segmentation [7] which are widely applied to caption image and video architecture to retrieve visual details.

Sequence models such as the recurrent neural network [8] are commonly used in speech identification, processing of natural language, and other fields. Machine translation [9], name entity recognition, analyzing of DNA sequence, recognizing of activity in video, and classification of sentiment are examples of supervised learning problems that can be addressed with sequence models. Long Short-Term Memory is a particular structure of RNN that has been shown in numerous experiments to be robust and efficient for modelling long-range dependencies. LSTM may be used as a component of complex systems. A memory cell is a dynamic unit of Long Short-Term Memory. binary central unit having a set of self-connection is integrated into each memory cell [10]. as it contains 3 gates (forget, input, and output), LSTM has been shown to be more efficient and reliable than a standard RNN in the

past. Recurrent neural networks with Long Short-Term Memory could be applied to produce dynamic sequences with long-range structure [11], [12].

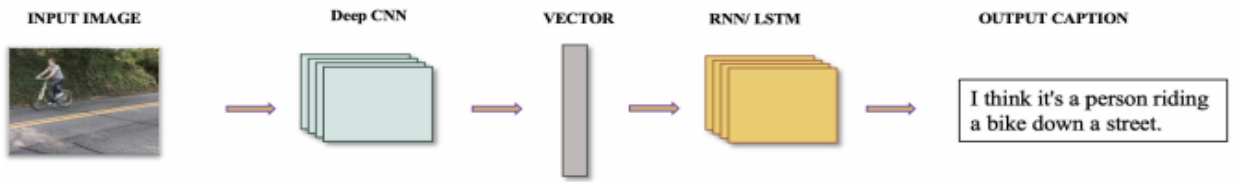

Figure 1: The encoder-decoder architecture

Figure 1 the encoder-decoder architecture was used in the primary efforts at image captioning as a research subject. The picture is encoded into a function vector by a deep learning algorithm. The model of language uses the vector of input to produce sentences that gives description about image

3. Image captioning

A picture in the eyes of a person consists of various colors that are used to compose various scenes. However, most pictures are painted with pixels in three channels from a computer's perspective. Different types of data are trending throughout the neural network to construct a vector and perform different operations on the mention features. This is demonstrated that by embedding an input image into a fixed-length vector, CNNs could generate a great depiction of the picture that could be used for a variety of visual tasks such as object identification, detecting object, and segmenting objects [13]. As a result, CNNs are often used as image encoders in image captioning techniques based on frameworks of encoding and decoding. Recurrent neural network gains past information by nonstop flow of the secret layer, which contains greater training capabilities and can outperform mining profounder linguistic skills such as implicit semantics and syntax info in the World Series [14]. In the hidden layer state, a RNN easily represents a dependent relation between various position words in historic knowledge. The encoder component of an encoder-decoder system for image captioning is a CNN model for retrieving features of image in the part of decoding, the machine insets the word vector expression into to the recurrent neural network model. Representation of every expression is done by single hot vector, which is then transformed into the same dimension as the image function using the word embedding model. The image captioning problem can be expressed as a binary (I, S) problem, Here (I) denotes single graph and S contains target words sequence, S=S1,S2, and Si is a word retrieved from data. The aim to train is to increase the probability approximation of target definition p(S|I) for the generated statement's goal and the target statement's matching more closely.

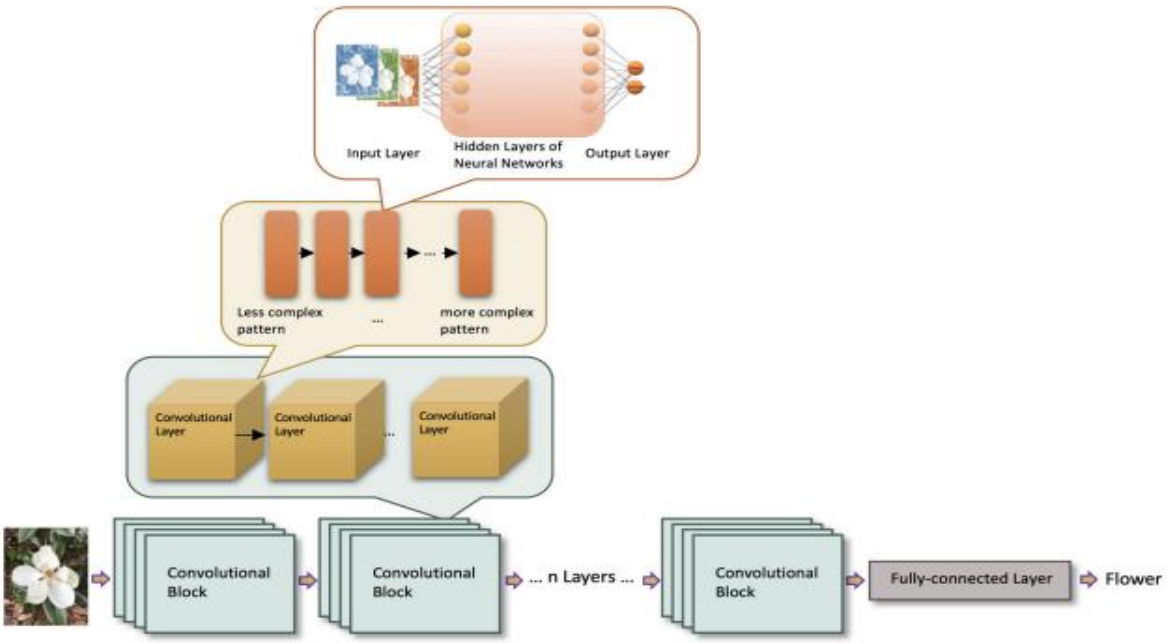

Figure 2: The overall design of a CNN indicates that each Convolutional Block is made up of n Convolutional layers

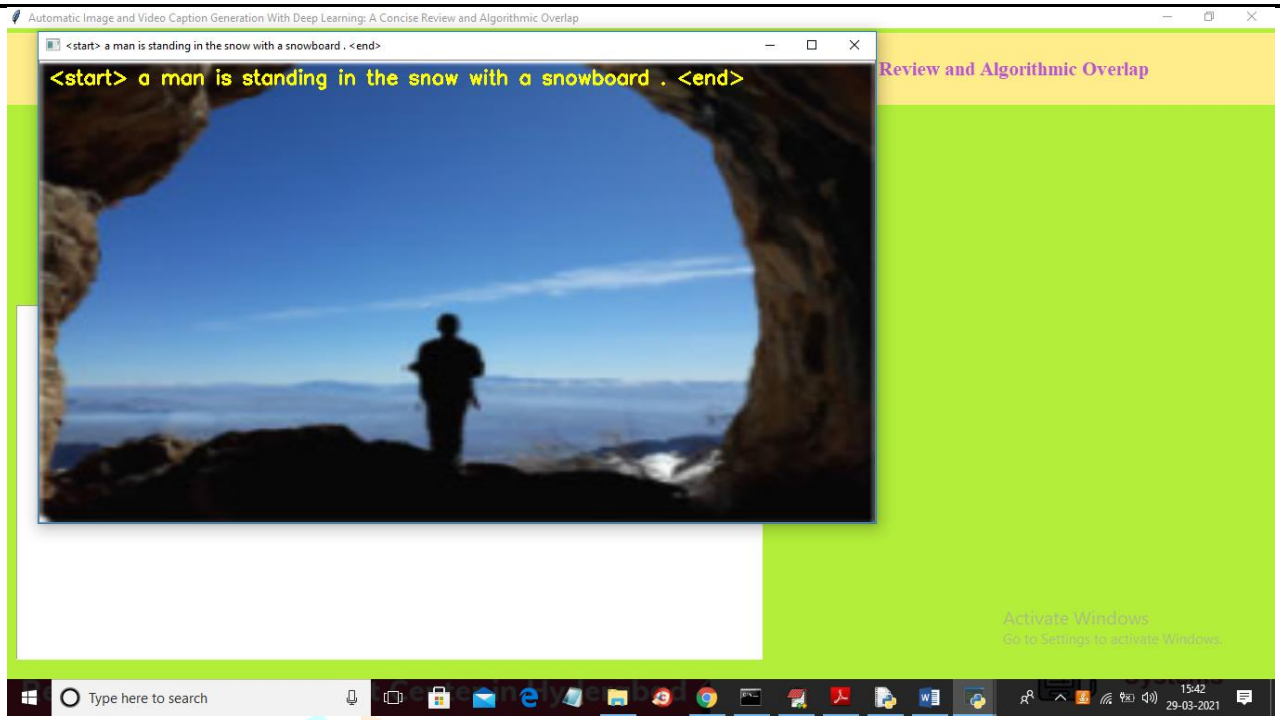

Figure 3: Our system output for captioning

4. Image Captioning Datasets

1) Flickr

More than 8000, 9000, and 30000 photographs are included in the Flickr8K, 9K, and 30K datasets, respectively. Amazon Mechanical Turk is used to annotate each picture with five separate sentences. The Flickr8K database primarily has images of humans and animals, where the Flickr30k database has images of people engaged in daily tasks. Five sentences are written for each picture [15], [16].

2) COCO

Lin et al. [17] introduced a fresh dataset for identifying and segmentation of common items in their natural environments. The MSCOCO dataset comprises 25000000 labelled illustrations in 328000 images, 91 types of object with 82 of those containing further 5K labelled instances, and 5 captions for every photo [15], [18].

Video captioning

For most people, unfolding a video in natural language is very simple, but for computers, it is difficult. It's problematic to determine the importance of the visual features and the adopted language model to the final definition from a methodological standpoint, so it is difficult to categorize the algorithms or models.

Image captioning (as explained in Section II-A) can be applied to keyframes of the video and a minor sample of the frames within the keyframes to achieve video captioning (Figure 4). The encoder-decoder system used for generating photo caption generator could be used as video caption generator as well (Figure 1 and 5).

Figure 4: key frames illustration

The first step is to comprehend the artefacts. With deep learning models, this focuses on visual detection and retrieves the performer, behavior, and action's object (for example person and activity recognition) from a clip of video. The clip of video is inserted in as frames' series, which are then interpreted as photos. As a result, each clip has a set of frames which are images of input. The retrieved data from the video is then placed in a common function vector. The second stage receives this vector. The caption generation stage describes what has been retrieved in a understandable language, therefore the objects mapping found in the initial phase. A mixture of CNN and RNN models is amongst the most important deep learning architectures used for video captioning. Long-term Recurrent Convolutional Networks (LRCNs) were suggested by Donahue et al. as a model for identification of visual and definition that incorporates convolutional layers and long-range temporal recursion while being end-to-end trainable. Datasets, using BLEU as a metric of description likeness they tested the video definition method on the TACoS multivariate dataset, ranking the results with the bilingual evaluation understudy -4 metric. Using LSTM helps with the modelling of the video as a length of variable input source, which is a big plus. Though the Long Short-Term Memory outdone statistical model-based methods, this could not be trained from start to end [19]. Venugopalan et al. [20] made use of the S2VT system (a sequence-to-sequence method for converting video to text), which combines convolutional neural network and Long Short-Term Memory models. The S2VT

architecture encodes and decodes a series of frames into a statement. a set of frames that is decoded into a sentence, their model tested on dataset of YouTube, the MPII-MD dataset, and the M-VAD dataset. They compared machine-generated descriptions to human-generated descriptions using METEOR and BLEU. The findings indicate substantial changes in human grammar assessments. Language alone makes a significant contribution; thus, it is essential to concentrate on both language and visual aspects in order to produce better descriptions. Later approaches, such as attention processes [21], have used a similar structure. In comparison to previous models, deep learning has produced significantly better results, with most approaches aiming to produce single sentence out of a clip of video having single bold case. Dense captioning, on the other hand, was proposed by Krishna et al. [22], and It aims to detect different events in a video by combining temporal proposals of interest and describing each one in natural language. This model implemented a novel approach of caption generator which makes use of historical and future contextual details to explain all events. They used the Activity Net Captions dataset to test the model. In images, Activity Net captions turn sentence statements from object-centric to action-centric in clips. It is not intended to deal with the problem of one sentence generator.

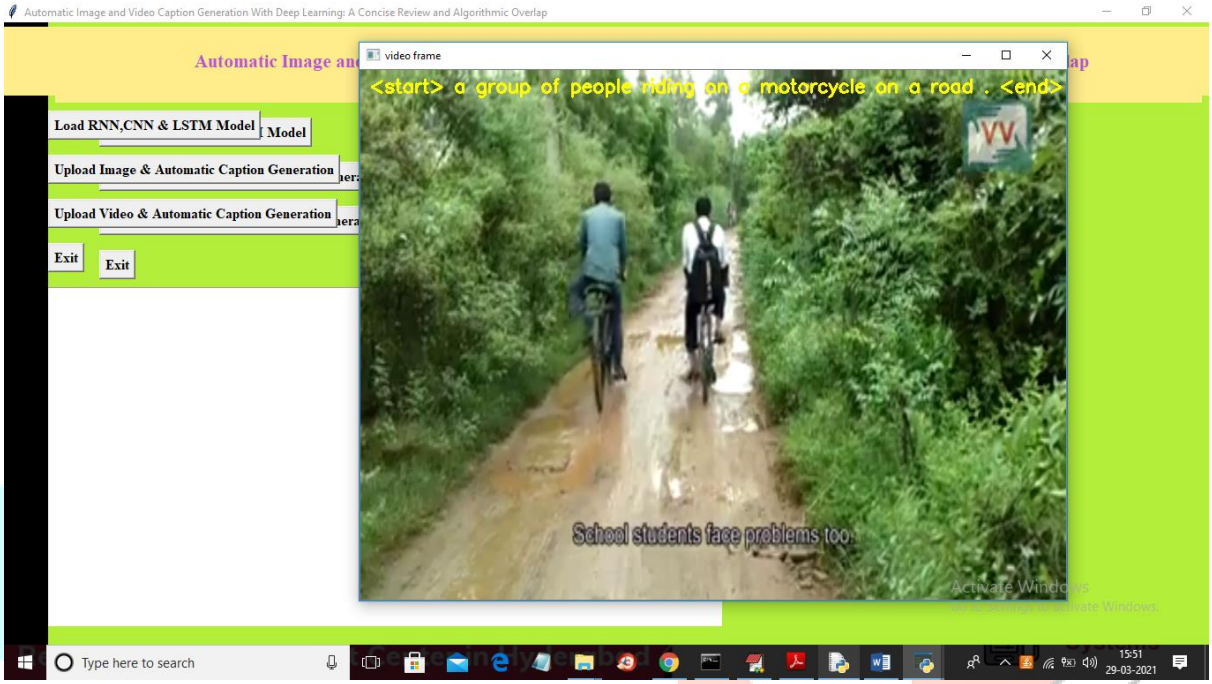

Figure 5: output for video captioning

Zhou et al. [23] model is the most identical to Krishna et al. in terms of dense video caption generator. However, this model, which consists of an encoder and two decoders, suggested an end-to-end modifier model for dense video caption generator. Over the encoding feature, which transforms the event proposal into a computable shield to guarantee that the proposal and captioning are consistent during training, the captioning decoder uses a masking network to focus its attention on the proposal event. This model also includes a self-attention mechanism. Deep reinforcement networks, a new research field for video captioning, is another line of work.

The Hierarchical Reinforcement Learning system, proposed by Wang et al. [24], aims to produce one or multiple sentences for a series of nonstop actions. An attention module is used for encoding as well for decoding in this model. On every metric, the novel HRL method outperformed all other algorithms. As a result, the HRL agent needs more focus space exploration and the use of features from several modalities. Ding et al. [25] introduced innovative strategies for applying long video segmentation in 2019, which can significantly reduce retrieval time. To increase the efficacy of video segmentation, repetitive frame of video recognition based on spatio-temporal interest points (STIPs) and a novel super-frame segmentation are merged. The filteblue long clip is then segmented superframe by superframe to locate the fascinating clip of the lengthy video. The saliency identification and LSTM variant network was used to translate keyframes from the most impactful segments to video captioning. Finally, in addition to the conventional LSTM, the focus mechanism is used to select more critical material. In this segment, we looked at a few techniques that were sorted chronologically based on the most recent convolutional neural network, Long Short-Term Memory, and attention-based methods were used by them. Since they use various methods, strategies, and datasets, we won't be able to evaluate them. Nonetheless, thanks to the methodology, detailed datasets, and captions, consistency and accuracy improve year after year.

5. Image and Video Captioning Evaluation Metrics

1) BLEU

Evaluation in two languages Understudy is a precision-based metric for automated machine translation assessment which relates well with human's assessment and contains a little marginal cost per run [15], [26]. For nominee sentences involving comparison sentences, BLEU has various n-grams-based variants.

2) METEOR

An automated metric for evaluating translation hypotheses is Metric for Evaluation of Translation with Explicit ORdering. It is based on a simplified principle of unigram matching [15], [18], [27], [28] between machine-produced translations and human-produced reference translations.

3) CIDEr
Consensus-based Image Description Evaluation [29] allows for an impartial comparison of machine generation methods based on their human-similarity, with not making subjective decisions about how content, syntax, saliency, and other factors are weighted in relation to one another. CIDEr was originally designed to assess image caption generation functions, but it is now often applied in video caption generation processes.

4) ROUGE
The content of a overview is determined by comparation it to other sum-ups formed by humans, according to Recall-Oriented Understudy for Gisting Evaluation [30]. ROUGE, like BLEU, has a variety of n-grams-based variants.

5) SPICE
Anderson et al. [31] proposed the Semantic Propositional Image Caption Generation Assessment, a new semantic assessment criterion that assesses how well image captions retrieve objects, features, and their relationships. In comparison to previously stated metrics, it coincides more with individual judgments of semantic consistency.

Conclusion and Future Work

Many templates for creating captions for photographs and short videos have been suggested and published in recent years. While these models aid in the advancement of technology, they are subject to inaccuracies due to basic limitations, limiting their use in real-world circumstances. Many of the earlier models used various algorithms and methodologies to handle image captioning and video captioning. In this paper, we used techniques for video caption generation that used image caption generation approaches as building blocks. As a result, the video caption generation procedure is regarded as a list of image caption summarization. For the reasons mentioned above, we have only concentrated on the algorithmic overlap in video and photo caption generator in this paper. It's difficult to compare various DP models for video and image caption generator in general. This is attributable to the element that computer scientist uses a variety of image datasets, criteria, classification processes, preprocessing, and structure variations, among other things. Despite their large variations, we formulated on the basis overlap between these approaches in this report. Many apps will benefit from a secure, precise, and instantaneous image and video caption generation process. Researchers are attempting to grant the machines eyes. Machines must first learn to see. Then they assist us with seeing more clearly. We can not only use the robots because of their intellect, but we will also communicate with them in areas we can only dream of. Captioning devices for images and videos may be an integral aspect of Assistive Technology for individuals with hearing or vision impairments. The captions will be used as meta-data for search engines, extending the functionality of the search engine to new heights. In certain implementations, captions can be used as part of suggestion processes. As previously mentioned, the latest technology for video and image caption generation frequently provides inaccurate captions. There is a lot of space for growth and change. Picture, video, and audio fusion and editing will result in more precise captions. There are audio-to-word converters on the market, and they're very good. Another issue with video captioning is that it is a computationally expensive task. Just really short videos (less than a few seconds in length) can be captioned using modern technologies. Using the next generation of GPUs and explicitly parallelizing algorithms (targeted at GPU computer architectures), For longer videos, We're getting closer to having hard data. Designing and developing a technique that allows viewers to order video captions at various levels of depth is a fantastic prospect in the field of video caption generation. Though, on can agree that the most basic and difficult study issue with video caption generation is that separate captions can be created for the same video based on different perceptions, just like two people can come up with two different views/descriptions after viewing the same video. We assume that by learning related topics and making the mechanism more engaging, we can solve this basic issue.

Reference

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘Imagenet classification with deep convolutional neural networks,’’ in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105.
[2] S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabnia, ‘‘Dissection of deep learning with applications in image recognition,’’ in Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI), Dec. 2018, pp. 1132–1138.
[3] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, p. 436, 2015.
[4] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767.
[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.
[6] R. Girshick, ‘‘Fast R-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.
[7] B. Romera-Paredes and P. H. S. Torr, ‘‘Recurrent instance segmentation,’’ in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 312–329.
[8] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ‘‘Empirical evaluation of gated recurrent neural networks on sequence modeling,’’ 20
[9] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ‘‘Learning phrase representations using RNN encoder-decoder for statistical machine translation,’’ 2014, arXiv:1406.1078. [Online].
[10] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[11] R. Kiros, R. Salakhutdinov, and R. S. Zemel, ‘‘Unifying visualsemantic embeddings with multimodal neural language models,’’ 2014, arXiv:1411.2539.
[12] Y. Wu et al., ‘‘Google’s neural machine translation system: Bridging the gap between human and machine translation,’’ 2016, arXiv:1609.08144.
[13]. Sermanet, Pierre, et al. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks." Eprint Arxiv (2013)
[14]. Sundermeyer, M., et al. "Comparison of feedforward and recurrent neural network language models." IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 8430-8434. (2013)

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.
[16] A. Karpathy, A. Joulin, and L. Fei-Fei, ''Deep fragment embeddings for bidirectional image sentence mapping,'' in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 1889–1897.
[17] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ''Microsoft COCO: Common objects in context,'' in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 740–755.
[18] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, ''Microsoft COCO captions: Data collection and evaluation server,'' 2015, arXiv:1504.00325.
[19] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, ''Long-term recurrent convolutional networks for visual recognition and description,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2625–2634
[20] S. Venugopalan, L. Anne Hendricks, R. Mooney, and K. Saenko, ''Improving LSTM-based video description with linguistic knowledge mined from text,'' 2016, arXiv:1604.01729.
[21] A. Torabi, C. Pal, H. Larochelle, and A. Courville, ''Using descriptive video services to create a large data source for video annotation research,'' 2015, arXiv:1503.01070
[22] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, ''Densecaptioning events in videos,'' in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 706–715
[23] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, ''End-to-end dense video captioning with masked transformer,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8739–8748.
[24] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, ''Video captioning via hierarchical reinforcement learning,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4213–4222.
[25] S. Ding, S. Qu, Y. Xi, and S. Wan, ''A long video caption generation algorithm for big video data retrieval,'' Future Gener. Comput. Syst., vol. 93, pp. 583–595, Apr. 2019.
[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ''BLEU: A method for automatic evaluation of machine translation,'' in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2001, pp. 311–318.
[27] M. Denkowski and A. Lavie, ''Meteor universal: Language specific translation evaluation for any target language,'' in Proc. 9th Workshop Stat. Mach. Transl., 2014, pp. 376–380
[28] S. Banerjee and A. Lavie, ''Meteor: An automatic metric for MT evaluation with improved correlation with human judgments,'' in Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization, 2005, pp. 65–72
[29] R. Vedantam, C. L. Zitnick, and D. Parikh, ''CIDER: Consensus-based image description evaluation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4566–4575.
[30] C. Lin, ''Rouge: A package for automatic evaluation of summaries,'' in Proc. Text Summarization Branches Out, 2004, pp. 74–81.
[31] P. Anderson, B. Fernando, M. Johnson, and S. Gould, ''Spice: Semantic propositional image caption evaluation,'' in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 382–398
[32] H. R. Arabnia and M. A. Oliver, ''Fast operations on raster images with SIMD machine architectures,'' in Computer Graphics Forum, vol. 5, Hoboken, NJ, USA: Wiley, 1986, pp. 179–188, doi: 10.1111/j.1467- 8659.1986.tb00296.x
[33] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, ''OCR as a service: An experimental evaluation of google docs OCR, tesseract, ABBYY finereader, and transym,'' in Proc. Int. Symp. Vis. Comput., in Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10072. Springer, 2016, pp. 735–746