IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 26, 2020, accepted November 24, 2020, date of publication December 4, 2020,
date of current version December 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042484

Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap

SOHEYLA AMIRIAN[1], KHALED RASHEED[1,2], THIAB R. TAHA[1], AND HAMID R. ARABNIA[1]
[1]Department of Computer Science, University of Georgia, Athens, GA 30602, USA
[2]Institute of Artificial Intelligence, University of Georgia, Athens, GA 30602, USA

Corresponding author: Soheyla Amirian (amirian@uga.edu; soh.amirian@gmail.com)

ABSTRACT Methodologies that utilize Deep Learning offer great potential for applications that automatically attempt to generate captions or descriptions about images and video frames. Image and video captioning are considered to be intellectually challenging problems in imaging science. The application domains include automatic caption (or description) generation for images and videos for people who suffer from various degrees of visual impairment; the automatic creation of metadata for images and videos (indexing) for use by search engines; general-purpose robot vision systems; and many others. Each of these application domains can positively and significantly impact many other task-specific applications. This article is not meant to be a comprehensive review of image captioning; rather, it is a concise review of both image captioning and video captioning methodologies based on deep learning. This study treats both image and video captioning by emphasizing the algorithmic overlap between the two.

INDEX TERMS Deep learning, image captioning, video captioning, long short term memory, generative adversarial network.

I. INTRODUCTION

Image processing has played and will continue to play an important role in science and industry. Its applications spread to many areas, including visual recognition [1] and scene understanding [2], to name a few. Before the advent of Deep Learning, most researchers used imaging methods that worked well on rigid objects in controlled environments with specialized hardware [3]–[12]. In recent years, deep learning-based convolutional neural networks have positively and significantly impacted the field of image captioning allowing a lot more flexibility. In this article, we attempt to highlight recent advances in the field of image and video captioning in the context of deep learning. Since 2012, many researchers have participated in advancing the deep learning model design [13], applications, and interpretation [14]. The science and methodology behind deep learning have been in existence for decades, but an increasing abundance of digital data and the involvement of powerful GPUs have accelerated the development of deep learning research in recent years. Convenient software development libraries such as TensorFlow and PyTorch, the open-source community, large labeled datasets like MSCOCO, Flicker, TACoS, LSMDC [15], [16], and splendid demonstrations simulate and model the explosive growth of the deep learning field.

Describing a scene in an image or a video clip is a highly demanding task for humans. To create machines with this capability, computer scientists have been exploring methods to connect the science of understanding human language with the science of automatic extraction and analysis of visual information. Image captioning and video captioning need more effort than image recognition, because of the additional challenge of recognizing the objects and actions in the image and creating a succinct meaningful sentence based on the contents found. The advancement of this process opens up enormous opportunities in many application domains in real life, such as aid to people who suffer from various degrees of visual impairment, self-driving vehicles, sign language translation, human-robot interaction, automatic video subtitling, video surveillance, and more. This article surveys the state of the art approaches with a focus on deep learning models for image and video captioning. The models and the generated

The associate editor coordinating the review of this manuscript and approving it for publication was Hazrat Ali.

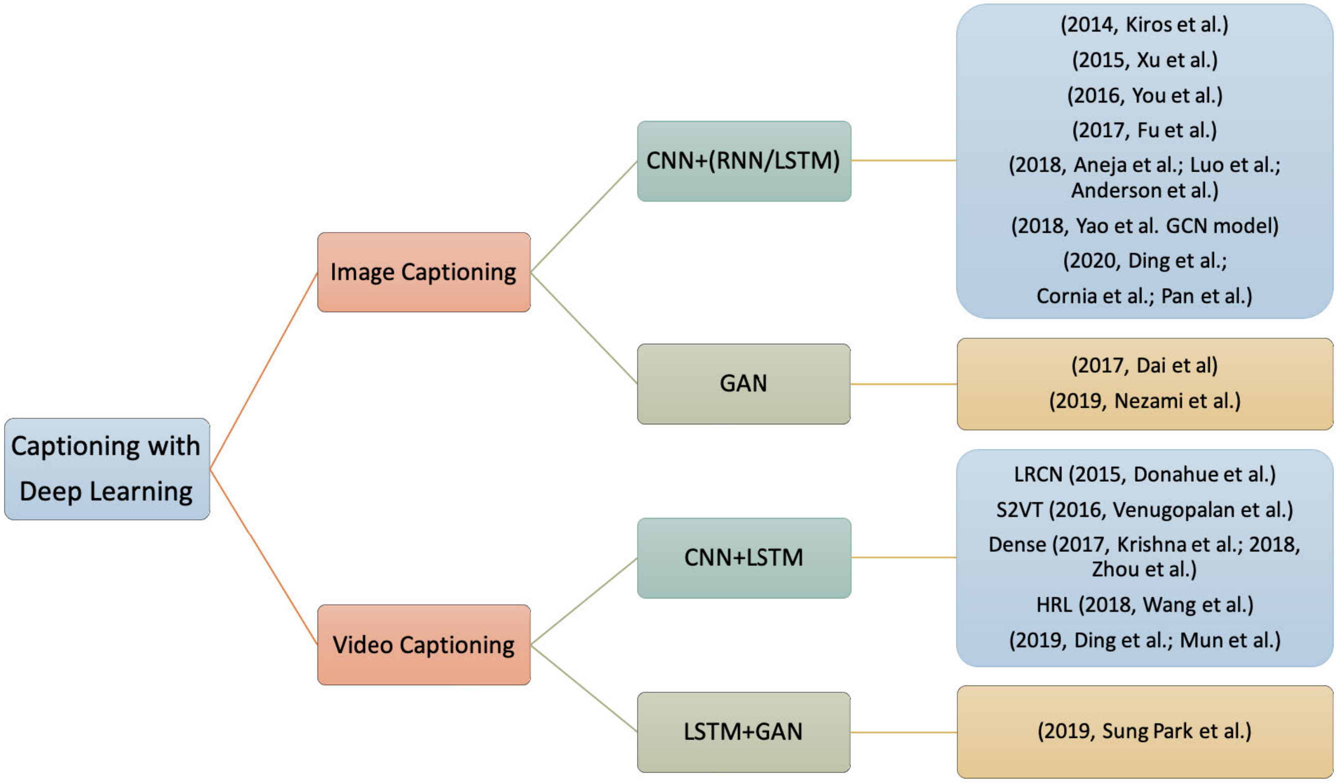

Captioning with Deep Learning
Image Captioning
Video Captioning
CNN+(RNN/LSTM)
GAN
CNN+LSTM
LSTM+GAN
(2014, Kiros et al.)
(2015, Xu et al.)
(2016, You et al.)
(2017, Fu et al.)
(2018, Aneja et al.; Luo et al.; Anderson et al.)
(2018, Yao et al. GCN model)
(2020, Ding et al.;
Cornia et al.; Pan et al.)
(2017, Dai et al)
(2019, Nezami et al.)
LRCN (2015, Donahue et al.)
S2VT (2016, Venugopalan et al.)
Dense (2017, Krishna et al.; 2018, Zhou et al.)
HRL (2018, Wang et al.)
(2019, Ding et al.; Mun et al.)
(2019, Sung Park et al.)

FIGURE 1. **The Taxonomy of the reviewed papers in this research.**

captions are evaluated by using BLEU, METEOR, CIDEr [17]–[19], and other evaluation metrics.

This article is a concise review of both image and video captioning methodologies based on deep learning, focusing on the algorithmic overlap between the two. This review begins by introducing the Image and Video Captioning in Section II. Then, a few recent methods of Image and Video Captioning, their Datasets, and evaluation metrics are discussed in Section III. Required Software and Hardware Platforms for implementing relevant models are mentioned in Section IV. Finally, a Case Study is presented in Section V. In order to facilitate the discussions about image and video captioning, we use the taxonomy shown in Figure 1. Figure 1 shows the conventional methods currently utilized in image and video captioning as well as the corresponding and relevant publications. In summary, the main contributions of this article include, a concise review of both image captioning and video captioning approaches based on deep learning. More specifically, the contributions include:

- A concise review of different architectures used for image and video captioning;
- The utilization of image captioning methods as building blocks to construct a video captioning system - i.e., Treating image captioning as a repetitive subset of video captioning;
- Review of hardware requirements and software frameworks for implementing an image/video captioning architecture;
- A novel application (case study) of video captioning, namely, the automatic generation of "titles" for video clips.

II. IMAGE AND VIDEO CAPTIONING

Many impressive studies have been done about image captioning [20]–[23]. Image captioning is often regarded to be the process of generating a concise description of objects and/or information about the scenes in an image. Some examples (images and their corresponding captions) are shown in Figure 2. Often, captions of images are generated manually. Automating this process would be a significant contribution. A system that automatically generates image captions can be utilized in many applications. Examples include: enhancing the accuracy of search engines; recognition and vision applications; enriching and creating new image datasets; enhancing the functionality of systems similar to Google Photos; and enhancing the optical system analysis of self-driving vehicles. In image captioning, the main challenges include the process of extracting visual information from the picture and the process of transforming this visual information into a proper and meaningful language. Captioning research started with the classical retrieval [20] and template-based [29] approaches in which Subject, Verb, and Object are detected separately and then joined using a sentence template. However, the advent of Deep Learning and the tremendous advancements in Natural Language Processing have equally and positively affected the field of captioning. Hence, the latest approaches follow

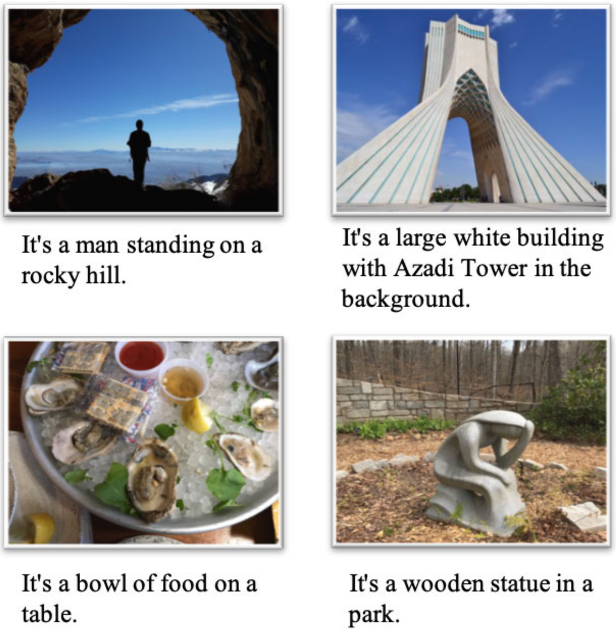

It's a man standing on a rocky hill.
It's a large white building with Azadi Tower in the background.
It's a bowl of food on a table.
It's a wooden statue in a park.

FIGURE 2. Some examples of image captioning. Each caption describes the image above it. These captions are generated with the model presented in [71] and the images are taken by the authors.

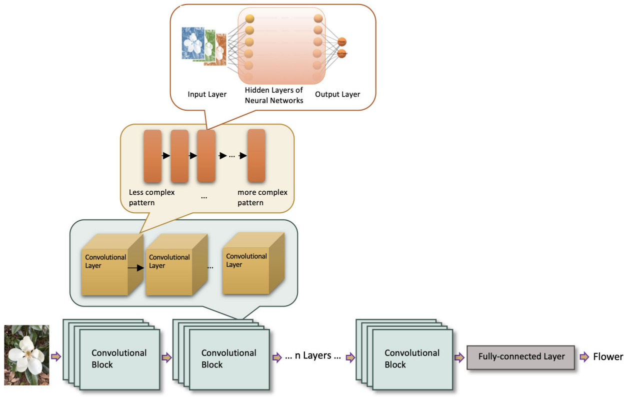


FIGURE 3. Overall architecture of Convolutional Neural Network that shows each Convolutional Block consists of n Convolutional layers and each of these Convolutional layers is built up of convolutions with filters.

deep learning-based architectures that encode the visual features with Convolutional Neural Networks and decode with a language-based model, which translates the features and objects given with an image-based model to a meaningful sentence. We dissect the image captioning process and models in Section III.

Video description is the automatic generation of meaningful sentences that describes the events in a video. Many researchers present different models on video captioning [24]–[28], mostly with limited success and many constraints. Video captioning can also be achieved by applying image captioning methods to the video frames as images. The advancement of video description opens up opportunities in a wide range of applications like human-robot interaction, automatic video subtitling, and video surveillance. Section III provides a detailed discussion of the video captioning process and recent models.

III. CAPTIONING METHODOLOGIES

Automatically generating natural language sentences describing an image or a video clip generally has two components: Encoder and Decoder. Here we specifically explain the architecture of each part. The Encoder utilizes a convolutional Neural Network, which extracts the objects and features from an image or video frame. For the decoder, a neural network is needed to generate a natural sentence based on the available information.

Convolutional Neural Network: A model with a large learning capacity to learn about thousands of objects from a large number of images [14] is needed. Deep learning presents computational models that are composed of multiple processing layers to learn representations of data in images [13], [30]. Deep learning-based Convolutional Neural Networks plays a key role in many applications, one of which is image recognition (See Figure 3). Image recognition is used to perform a large number of visual tasks, such as understanding the content of images. Several well-known models [13] in the field of CNNs based on object detection [1], [31], [32] and segmentation [33] exist that are heavily used in image captioning and video captioning architecture to extract the visual information.

Recurrent Neural Networks: Sequence models like recurrent neural network (RNN) [34] have widely been utilized in speech recognition, natural language processing, and other areas. Sequence models can address supervised learning problems like machine translation [35], name entity recognition, DNA sequence analysis, video activity recognition, and sentiment classification. Gated recurrent unit (GRU) is a gating mechanism in RNN, introduced by Cho *et al.* [35] in 2014. The basic RNN algorithm runs into a vanishing gradient problem (a difficulty in training artificial neural networks). The gated recurrent units are an effective solution for addressing the vanishing gradient problem. They allow neural networks to capture a much longer range dependencies [34]. The advantage of the GRU is that it is a simple model, therefore it is easy to build a big network with GRU. Also, it only has two gates, as a result, it computes quickly.

Long Short Term Memory: LSTM, as a special RNN structure, has proven to be stable and powerful for modeling long-range dependencies in various studies. LSTM can be adopted as a building block for complex structures. The complex unit in Long Short Term Memory is called a memory cell. Each memory cell is built around a central linear unit with a fixed self-connection [36]. LSTM is historically proven more powerful and more effective than a regular RNN since it has three gates (forget, update, and output). Long Short Term Memory recurrent neural networks can be used to generate complex sequences with long-range structure [37], [38].

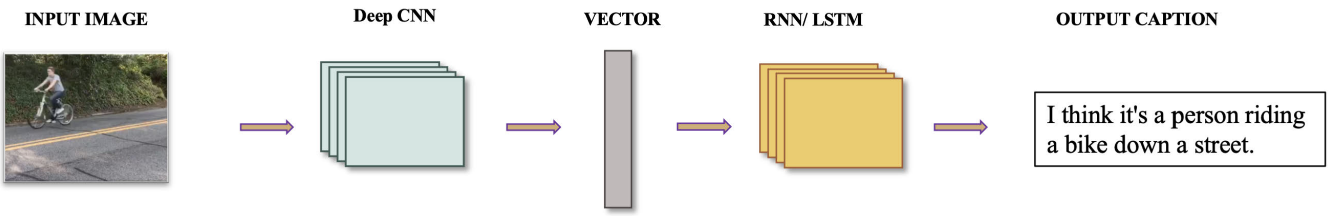


FIGURE 4. **The early attempts of image captioning as an active research area exploit the encoder-decoder architecture. A deep learning model encodes the image into a feature vector. The language model takes the input vector to generate a sentence that describes the image, leading to promising results for this task.**

A. IMAGE CAPTIONING METHODOLOGIES

Many methods for image captioning are there. Earlier methods, prior to deep neural networks (DNNs), were retrieved-based [20] or template-based [29] models. Recent methods are based on deep neural networks. Generating an automatic caption for describing an image has two stages. First, the information needs to be extracted from the image and put it in a feature vector. This stage focuses on visual recognition through deep learning models. Then the feature vector is fed into the second stage. The second stage is caption generation which is describing what is extracted in a grammatically correct natural language sentence (See Figure 4). So, we classified DNN-based methods based on the main framework into subcategories that they respectively use. Here, a review of recent deep learning-based works for automatic image captioning is discussed. All are summarized with more details about the evaluation results in Table 2.

A breakthrough in image and video captioning occurred in 2014 through the application of encoder-decoder models. Kiros *et al.* [37] introduced an encoder-decoder pipeline model in which an encoder network takes the image or video as an input and extracts a fixed-size feature vector that a decoder network maps to a sequence of words. They set new best results when using the 19-layer Oxford convolutional network. Then, a series of innovations such as attention mechanism have been introduced to boost image captioning by encouraging more interactions between the two different modalities. They were developing an attention-based model that jointly learns to align parts of captions to images. The generated descriptions are arguably the nicest ones to date [37]. The attention model is one of the models used in deep learning that got from one of the most curious facets of the human visual system. The attention-based model learns to focus on different parts of the image. This is crucial when much clutter is in an image. However, this may cause losing information which could be useful for richer and more descriptive captions. Xu *et al.* [17] proposed the attention-based approach that gives the state of the art performance on three benchmark datasets using the BLEU and METEOR metric (See section III-C). They showed how the learned attention can be exploited to give more interpretability to the model generation process and demonstrate that the learned alignments correspond well to human intuition. Their model encourages future work in using visual attention. Next, You *et al.* [21] proposed a model of semantic attention that learns to selectively focus on the semantic attributes in the image. The algorithm combines top-down and bottom-up strategies to extract richer information from the image and fuses them with an RNN that can selectively attend on rich semantic attributes detected from the image. They performed their method on different datasets, and the captioning system was implemented based on the LSTM network. The image feature vector is extracted from the last 1024 dimensional convolutional layer of the GoogleNet [13] CNN model. Furthermore, their framework employs attention at both input and output layers to the RNN module. Their effort was exploiting abundant fine-grain visual semantic aspects and fusing global and local information for generating a better caption. The results show that the algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics. We see in the next research, Fu *et al.* [39] proposed the image caption system that exploits the parallel structures between images and sentences. One contribution of this system is that it aligns the process of generating captions and the attention shifting among the visual regions. Another is that it introduces the scene-specific contexts to LSTM that adapt language models for word generation to specific scene types. In that system, an image is first analyzed and represented with multiple visual regions from which visual features are extracted. The visual feature vectors are then fed into an LSTM network which predicts both the sequence of focusing on different regions and the sequence of generating words based on the transition of visual attention. The neural network model is also governed by a scene vector, a global visual context extracted from the whole image. Intuitively, it selects a scene-specific language model for generating text. They evaluated captions in BLEU-n, METEOR, ROUGE-L and CIDEr-D metrics by testing on several popular datasets, including the MSCOCO, Flickr8K, and Flickr30K (See Table 2). Either region-based attention or scene-specific contexts alone improve performance but combining the two provides a further improvement.

Researching more with CNN and LSTM models, in 2018, Aneja *et al.* [40] developed a convolutional image captioning technique with existing LSTM techniques and analyzing the differences between RNN based learning and their method.

TABLE 1. The summary of a few recent works for Image Captioning.

Model	Architecture	Evaluation (on MSCOCO)	Comment
(2014, Kiros et al.)	CNN+LSTM encoder–decoder Attention-based	Image Annotation result R@1 R@5 R@10 Med r 23.0 50.7 62.9 5 (OxfordNet) on Flickr30K. R@K is Recall@K (high is good). Med r is the median rank (low is good).	The generated descriptions are arguably the nicest ones to date.
(2015, Xu et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 71.8 50.4 35.7 25.0 23.04	They encourage future work in using visual attention.
(2016, You et al.)	CNN+RNN Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR 0.709 0.537 0.402 0.304 0.243 Using the ground-truth visual attributes	They need to experiment with phrase-based visual attributes with their distributed representations.
(2017, Fu et al.)	VGG/Alex/ResNet + LSTM Attention-based	ENSEMBLE result BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr-D 72.4 55.5 41.8 31.3 24.8 53.2 95.5	Either region-based attention or scene-specific contexts improve performance. Combining these two modeling ingredients provides a further improvement.
(2017, Dai et al)	GAN VGG/G-MLE/G-GAN	BLEU-3 BLEU-4 METEOR ROUGE L CIDEr SPICE E-NGAN E-GAN G-MLE: 0.393 0.299 0.248 0.527 0.1020 0.199 0.464 0.427 G-GAN: 0.305 0.207 0.224 0.475 0.795 0.182 0.528 0.602	E-NGAN regard G-GAN as the best generator. This framework also provides an evaluator that is more consistent with human's evaluation.
(2018, Aneja et al.)	CNN+LSTM (ResNet152)	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr 0.725 0.555 0.41 0.299 0.251 0.532 0.972	ResNet is capable of encoding better feature vectors for images. The Meteor and CIDEr results are comparable with previous works.
(2018, Luo et al.)	ResNet101 + LSTM ATTN+CIDER+DISC	BLEU-4. ROUGE. METEOR CIDEr SPICE 0.3274 0.2574 **0.5457** 1.0231 0.1939	Incorporating a discriminability loss, in training image caption generators improves the quality of resulting captions. Their model needs more sophisticated visual semantic embedding model.
(2018, Anderson et al.)	CNN+LSTM Faster R-CNN; ResNet101 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L CIDEr SPICE c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 80.2 95.2 64.1 88.8 49.1 79.4 36.9 68.5 27.6 36.7 57.1 72.4 117.9 120.5 21.5 71.5	They obtained first place in the 2017 VQA Challenge.
(2018, Yao et al.)	GCN+LSTM ResNet101 Attention-based	BLEU-1 BLEU-4 METEOR ROUGE-L CIDEr-D SPICE 80.9 38.3 28.6 58.5 128.7 **22.1**	They build graphs over the detected objects in an image based on their spatial and semantic connections.
(2019, Nezami et al.)	ATTEND-GAN	BLEU-1 BLEU-2 BLEU-3 BLEU-4 ROUGE-L METEOR CIDEr SPICE 56.55 33.85 20.80 13.05 44.45 18.35 62.85 16.05	It also adds sentiment and naturalness to the sentences.
(2020, Ding et al.)	CNN+LSTM (VGG-19)	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr 0.748 0.525 0.365 0.235 0.235 0.505 1.041	They introduced the theory of attention in psychology to image caption generation.
(2020, Cornia et al.)	CNN+LSTM Faster R-CNN; ResNet101 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 81.6 **96.0** 66.4 **90.8** 51.8 **82.7** 39.7 **72.8** 29.4 39.0 59.2 **74.8** 129.3 132.1	Novelty in using a stack of memory-augmented encoding layers and a stack of decoder layers.
(2020, Pan et al.)	CNN+LSTM SENet-154 Attention-based	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE CIDEr c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 c5 c40 **81.9** 95.7 **66.9** 90.5 **52.4** 82.5 **40.3** 72.4 29.6 39.2 **59.5 75** **131.1 133.5**	They used a novel unified X-Linear attention block for image captioning. Using SENet-154 makes it proceed other models.

This technique contains three main components. The first and the last components are input/output word embeddings respectively. However, while the middle component contains LSTM or GRU units in the RNN case, masked convolutions are employed in their CNN-based approach. This component is feed-forward without any recurrent function. Their CNN with attention (Attn) achieved comparable performance. They also experimented with an attention mechanism with attention parameters using the conv-layer activations. The results on CNN+Attn method were improved relative to the LSTM baseline. For better performance on the MSCOCO they used ResNet features and the results show that ResNet boosts the performance. The results on the MSCOCO with Resnet101 and Resnet152 were comparable to previous works. Table 2 shows that the METEOR and CIDEr results are outstanding, therefore better captions. Then, we see Ding *et al.* [72] introduced the same architecture of CNN by VGG-19 and LSTM on the MSCOCO dataset, but with the theory of attention [17] in psychology to image caption generation with two types of attention mechanisms: The

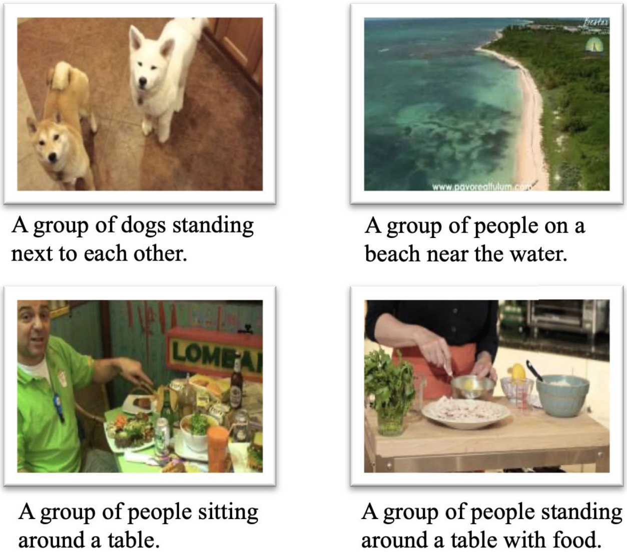
TABLE 2. The summary of a few recent works for Video Captioning.

Model	Architecture	Evaluation	Comment
LRCN (2015, Donahue et al.)	Long-term recurrent convolutional networks.	BLEU 28.8 on TACoS	It was still not trainable in an end-to-end fashion.
S2VT (2016, Venugopalan et al.)	A sequence to sequence approach. CNN+LSTM.	BLEU-4 METEOR **42.1** **31.4** on Youtube, MPII-MD and M-VAD	The contribution of language alone is considerable.
Dense (2017, Krishna et al.)	Attention mechanism. Dense-captioning, multiple events.	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 26.45 13.48 7.12 3.98 9.46 24.56 on ActivityNet	No single-sentence generation scenario.
(2018, Zhou et al.)	CNN+LSTM Attention based, Dense-captioning.	BLEU-3 BLEU-4 METEOR 4.76 2.23 10.12 End-to-end Masked Transformer, on ActivityNet	This model is able to produce proposal and description simultaneously.
HRL (2018, Wang et al.)	Hierarchical Reinforcement Learning. attention module.	BLEU-4 METEOR ROUGE-L CIDEr 41.3 28.7 **61.7** **48.0** on MSR-VTT	Outperformed all the other algorithms. Still needs a boost.
(2019, Ding et al.)	CNN+LSTM	BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR ROUGE-L **76.4** **56.3** **43.6** 31.7 26.5 53.5 on MSCOCO	The result that has been provided is for the image caption part.
(2019, Mun et al.)	A streamlined approach. C3D+GRU+RNN	BLEU-1 BLEU-2 BLEU-3 BLEU-4 CIDEr METEOR 17.92 7.99 2.94 0.93 30.68 8.82 on ActivityNet	Algorithm generates captions for events sequentially conditioned on the prior ones by detecting highly correlated events in a video.
(2019, Sung Park et al.)	LSTM+GAN	BLEU-4 METEOR CIDEr-D 10.02 16.69 21.07 GAN results on ActivityNet	A hybrid discriminator consists of three individual experts: language, one for relating the sentence to the video, and one pairwise, across sentences.

stimulus-driven for monitoring salient information by Color stimulus- driven, Dimension stimulus-driven and location perception stimulus-driven for attention detection. And the concept-driven that is a classical question-guided attention mechanism. This approach enhances the encoder framework to suit complex scenes.

At the same year, Luo *et al.* [66] presented a method that has been trained with the COCO dataset for the Encoder part. For the image encoder in retrieval and FC captioning model, Resnet-101 is used. The spatial features are extracted from the output of a Faster R-CNN with ResNet-101, trained with an object and attribute annotations from Visual Genome [67]. The retrieval model uses GRU-RNN to encode text. The captions generated with this model describe valuable information about the images. However, richer and more diverse sources of training signal may further improve the training of caption generators. For experimenting the output, we implemented their method and some results are shown in Figure 5.

Also, Anderson *et al.* [58] proposed a combined bottom-up and top-down attention mechanism that enables attention to be calculated at the level of objects and other salient image regions. The bottom-up attention uses Faster R-CNN with ResNet-101 [13], which represents a natural expression of a bottom-up attention mechanism. The top-down mechanism uses a task-specific context to predict an attention distribution over the image regions. The attended feature vector is then computed as a weighted average of image features over all


A group of dogs standing next to each other.

A group of people on a beach near the water.

A group of people sitting around a table.

A group of people standing around a table with food.

FIGURE 5. These captions are generated with the model presented in [66] and the images are scenes from the ActivityNet dataset.

regions. Their results on the MSCOCO dataset present a new state-of-the-art for the task, achieving CIDEr, BLEU-4 scores of 117.9, and 36.9, respectively. Demonstrating the broad applicability of the method, they applied the same approach to Visual Question Answering, and obtained first place in the 2017 VQA Challenge. In a novel architecture, Yao *et al.* [76] proposed Graph Convolutional Networks plus Long Short-Term Memory (GCN-LSTM) model that integrates both semantic and spatial object relationships into image encoder which more remarkably increases CIDEr-D performance on COCO testing set.

We also have Cornia *et al.* [73] which proposed a Transformer-based architecture. The architecture composed of a stack of memory-augmented encoding layers and a stack of decoder layers. Image regions and their relationships are encoded in a multi-level fashion, in which low-level and high-level relations are taken into account. The model can learn and encode a priori knowledge by using persistent memory vectors. The generation of the sentence, done with a multi-layer architecture, exploits both low-level and high-level visual relationships instead of having just a single input from the visual modality. This is achieved through a learned gating mechanism, which weights multi-level contributions at each stage. They name this model Meshed-Memory Transformer as this creates a mesh connectivity schema between encoder and decoder layers. Based on their results, this approach achieves a new state of the art on COCO, ranking first in the on-line leaderboard.

The same year, Pan *et al.* [74] presented a unified attention block or X-Linear attention block, that employs bilinear pooling to selectively capitalize on visual information or perform multimodal reasoning. In addition, they present X-Linear Attention Networks that novelly integrates X-Linear attention block(s) to leverage higher order intra- and inter-modal interactions. The experiments on COCO benchmark shows that their X-LAN obtains the best published CIDEr performance of 132.0% on COCO Karpathy test split so far. By endowing Transformer with X-Linear attention blocks, CIDEr is boosted up to 132.8%.

The models mentioned above are all heavily utilized in image caption generation. Many other deep learning models have the potential to be used for applications such as image caption generation; one such model is Generative Adversarial Network. In 2014, for the first time, Goodfellow *et al.* proposed a new framework for estimating generative models via an adversarial process, in which they simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. GAN has been successfully used in image generation. They can produce natural images almost indistinguishable from real photos [60], [61], [61], [68]. Dai *et al.* [62] presented a new framework based on Conditional Generative Adversarial Networks (CGAN), which jointly learns a generator to produce descriptions conditioned on images and an evaluator to assess how well a description fits the visual content. This work proposed a different task for the GAN method. They have a strategy stemming from Reinforcement Learning, which allows the generator to receive early feedback along the way. In their method, they implemented G-MLE: a generator trained based on MLE that is used to produce the descriptions and G-GAN, the same generator, which is based on the conditional GAN formulations. For both G-MLE and G-GAN, VGG16 is used as the image encoder. They considered multiple evaluation metrics, including six conventional metrics BLEU-n, METEOR, ROUGE L, CIDEr, SPICE, and two additional metrics relevant to their formulation: E-NGAN and E-GAN, particularly using their framework. This method was the first to apply GAN. We believe GAN has significant potential in image captioning. In 2019, Nezami *et al.* [63] proposed the ATTEND-GAN model. Their contribution is to generate human-like stylistic captions in a two-stage architecture, with ATTEND-GAN using both the designed attention-based caption generator and the adversarial training mechanism on the SentiCap dataset. The architecture of the ATTEND-GAN model uses spatial-visual features that are generated with ResNet-152 network and the caption discriminator is inspired by the Wasserstein GAN (WGAN).

So far, we briefly reviewed a few methods, according to the common approaches that they have used. For a fair comparison of the models, Table 2 shows the results of attention-based methods on the MSCOCO dataset, the common dataset that they have utilized. With this comparison, we could state that Anderson *et al.* performed well on the MSCOCO dataset. Their method outperformed previous works. The reason is that it uses the attention mechanism which focuses only on relevant objects of the image. Also, We found that the performance of a technique can vary across different metrics, parameters, and datasets. Here, we tried to analyze them based on the different methods they have used. However, image captioning still remains an active research and it has a long way to go in improving the accuracy of captioning the information in images (See Figure 6).

A.1 IMAGE CAPTIONING DATASETS

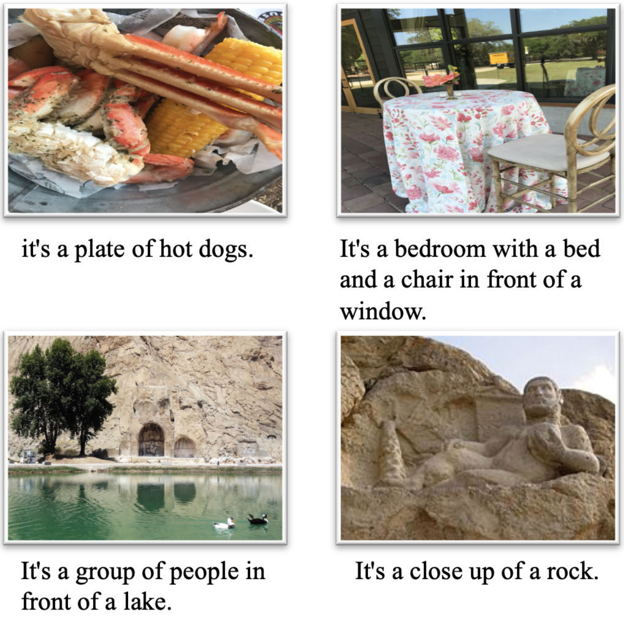
A few datasets are widely used to evaluate and compare image captioning methods: Flickr8K [23], Flickr9K [17], Flickr30k [17], [23] and Microsoft COCO [17], [18].

1) Flickr

The Flickr8K, 9k, and 30k datasets contain more than 8000, 9000, and 30000 images, respectively. Each image is annotated using Amazon Mechanical Turk with 5 independent sentences. The Flickr8K dataset mainly contains human and animal images, while the Flickr30k dataset contains humans involved in everyday activities and events. For each image, five sentences are provided [17], [23].

2) COCO

Lin *et al.* [41] presented a new dataset for detecting and segmenting objects found in everyday life in



it's a plate of hot dogs.

It's a bedroom with a bed and a chair in front of a window.

It's a group of people in front of a lake.

It's a close up of a rock.

FIGURE 6. Examples of poor image captioning generated by state-of-the-art systems. These captions are generated with the model presented in [71] and the images are taken by the authors.

their natural environments. Microsoft Common Objects in COntext (MSCOCO) dataset contains a total of 2.5 million labeled instances in 328k images, 91 object categories with 82 of them having more than 5,000 labeled instances, and five assigned captions to each image [17], [18].

B. VIDEO CAPTIONING METHODOLOGIES

Describing a video in natural language is a trivial task for most people, but a challenging one for machines. From the methodological perspective, categorizing the models or algorithms is challenging because it is difficult to assert the contributions of the visual features and the adopted language model to the final description.

Video captioning can be achieved by applying image captioning (as discussed in Section III-A) to the video keyframes and a small sample of the frames in-between the keyframes (See Figure 7). The encoder-decoder framework discussed for image captioning can also be extended to video captioning (compare Figure 4 and 8). Overall, generating natural language sentences describing the video content automatically has two stages. The first stage is understanding the objects. This focuses on visual recognition with deep learning models and extracts the performer, action, and the object of the action (e.g. human and activity detection) from the video clip. The video clip is fed as a series of frames that are considered as images. So, we have a series of frames in each clip that are input images. Then the extracted information from the clip is put in a common feature vector. This vector is fed into the second stage. The second stage is caption generation which is describing what is extracted in a grammatically correct natural language sentence, thus mapping the objects identified in the first stage. Here, we bring a combination of deep learning architectures for the encoding and decoding stages (See Table 2).

One of the most common architectures in deep learning that is used for video captioning is a combination of CNN and RNN models. Donahue *et al.* proposed Long-term Recurrent Convolutional Networks (LRCNs), a model for visual recognition and description which combines convolutional layers and long-range temporal recursion and is end-to-end trainable. They considered three vision problems: activity recognition, image description, and video description. LRCN processes the variable-length visual input with a CNN, whose outputs are fed into a stack of recurrent sequence models, which finally produce a variable-length prediction. They evaluated the image architecture on the COCO and Flickr30k datasets, using BLEU as a measure of similarity of the descriptions. They evaluated the video description approach on the TACoS multilevel dataset, using the BLEU-4 metric for scoring the results. The advantage of using LSTM here is that it allows them to model the video as a variable-length input stream. Although the LSTM outperformed the statistical model-based approaches, it was still not trainable in the end-to-end fashion [24]. We see that Venugopalan *et al.* [25] used the S2VT method (a sequence to sequence approach for a video to text), which is a combination of CNN and LSTM models. The S2VT architecture encodes a sequence of frames and decodes them to a sentence. They compared their model on the YouTube dataset, MPII-MD, and M-VAD (See Table 3). They evaluated the performance using METEOR and BLEU to compare the machine-generated descriptions to human ones. The results show significant improvements in human evaluations of grammar. The contribution of language alone is considerable; hence, it is important to focus on both language and visual aspects to generate

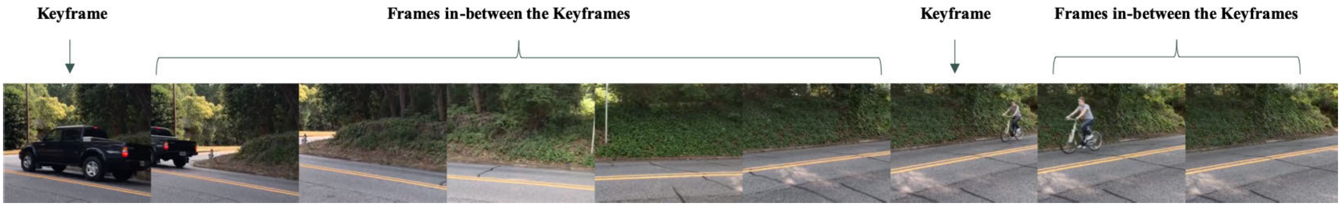

Keyframe
Frames in-between the Keyframes
Keyframe
Frames in-between the Keyframes

FIGURE 7. Video: Keyframes and Frames in-between the Keyframes (Keyframe is a frame used to indicate the beginning or end of a change made to a parameter).

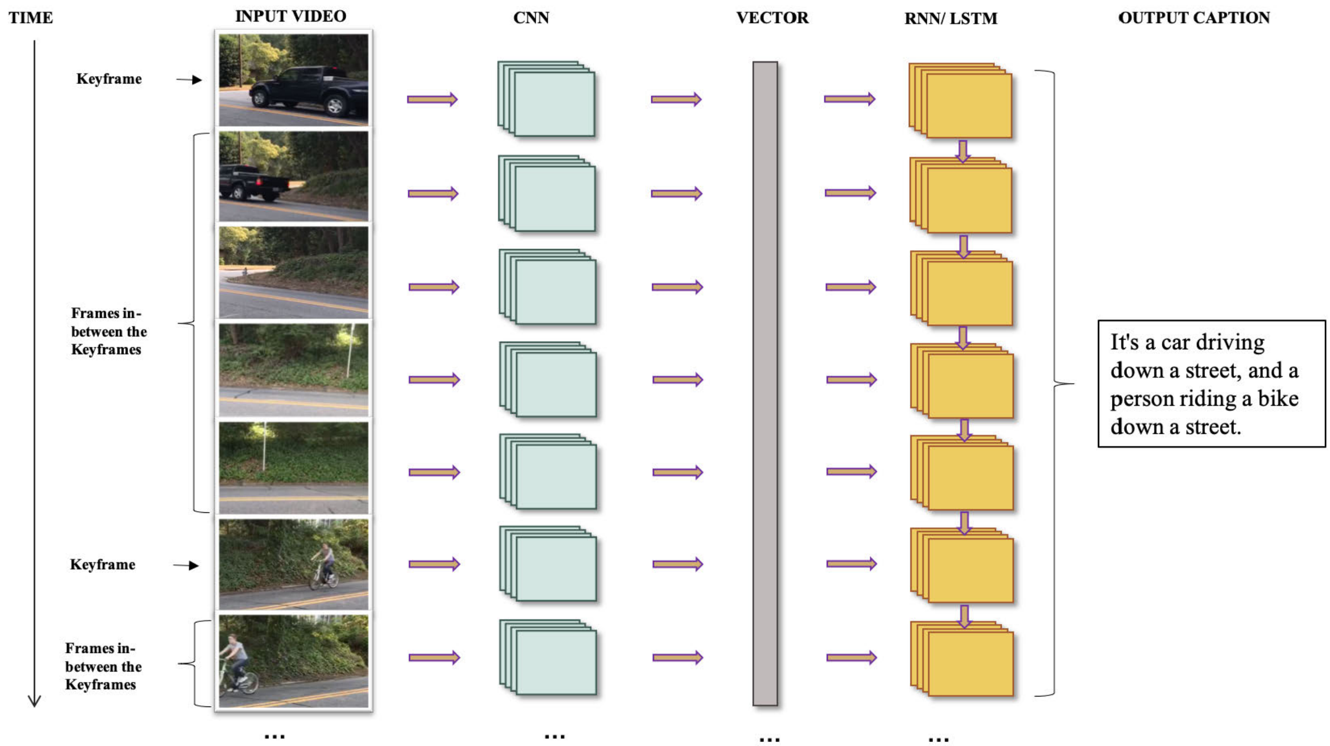

TIME
INPUT VIDEO
CNN
VECTOR
RNN/ LSTM
OUTPUT CAPTION
Keyframe
Frames in-between the Keyframes
Keyframe
Frames in-between the Keyframes
It's a car driving down a street, and a person riding a bike down a street.
...
...
...
...

FIGURE 8. This is a basic structure for video captioning models. Each DCNN takes a frame of the video as an image, then encodes the frame into a common feature vector between all the other video frames. The language model takes the vector to generate a sentence or a paragraph that describes the video.

better descriptions. Later methods have adopted a similar framework, including attention mechanisms [26].

Deep learning has achieved much better results compared to previous models, and most methods aimed at producing one sentence from a video clip containing only one bold event. Krishna *et al.* [27], however, presented Dense-captioning, which focuses on detecting multiple events that occur in a video by jointly localizing temporal proposals of interest and then describing each with natural language. This model introduced a new captioning module that uses contextual information from past and future events to jointly describe all events. They implemented the model on the ActivityNet Captions dataset (See Table 3 and Section III-B). The captions that came out of ActivityNet shift sentence descriptions from being object-centric in images to action-centric in videos. It does not aim to solve the single-sentence generation scenario, though.

The most similar work to Krishna *et al.* in using dense video captioning model is Zhou *et al.* [75] model. However, this model proposed an end-to-end transformer model for dense video captioning, and is composed of an encoder and two decoders. The captioning decoder employs a masking network to restrict its attention to the proposal event over the encoding feature which converts the event proposal to a differentiable mask to ensure the consistency between the proposal and captioning during training. Furthermore, this model employs a self-attention mechanism.

Another line of work is deep reinforcement networks, a relatively new research area for video description. Wang *et al.* [28] presented the Hierarchical Reinforcement Learning method that aims to generate one or more sentences for a sequence of one or more continuous action. In this model, both the encoder and decoder are equipped with an attention module. The novel HRL method outperformed all the other algorithms on all metrics. Hence, the HRL agent needs more exploration in terms of attention space and utilizing features from multiple modalities.

In 2019, Ding *et al.* [59] proposed novel techniques for the application of long video segmentation, which can effectively shorten the retrieval time. Redundant video frame detection based on the spatio-temporal interest points (STIPs) and a novel super-frame segmentation are combined to improve the effectiveness of video segmentation. After that, the super-frame segmentation of the filteblue long video is performed to find the interesting clip of a long video. Keyframes from the most impactful segments are converted to video captioning by using the saliency detection and LSTM variant network. Finally, the attention mechanism is used to select more crucial information to the traditional LSTM. This method is benchmarked on the VideoSet dataset and evaluated with the BLEU, Meteor, and Rouge on the image captioning part. However, the language model still has a large performance gap from humans in cases such as small object recognition or object recognition at lower resolutions. Similar to Krishna *et al.* [27] work, Mun *et al.* [64] proposed

TABLE 3. Some of the Video Caption Datasets.

Domain	Dataset	Total duration
People	Charades [42]	82h
Open	MSVD [43], ActivityNet Captions [27], MSR-VTT [44]	5.3h, 849h, 41.2
Social Media	VideoStory [45]	396h
Cooking	MPII [46], TACoS [15], YouCook2 [47]	490m, 15.9h,176h
Movie	LSMDC [16], MPII-MD [48], M-VAD [26]	158h, 73.6h, 84.6h

a dense video captioning framework, that models temporal dependency across events in a video explicitly and leverages visual and linguistic context from prior events for coherent storytelling. They have used Single-Stream Temporal Action model to get some proposals at a single scan, then by implying PtrNet, the highly correlated events that makeup an episode fed into a sequentional captioning network to produce a caption by RNN systems. The proposed technique achieves outstanding performances on the ActivityNet Captions dataset in terms of METEOR. By injecting GAN to DL, Sung Park *et al.* [65] applied Adversarial Networks in their framework by designing a discriminator to evaluate visual relevance to the video, language diversity, fluency, and coherence across sentences. GAN helps to generate more accurate, diverse, and coherent multi-sentence video descriptions. The task of discriminator (D) is to score the descriptions generated with the generator (G) for a given video. They propose to compose D out of three separate discriminators, each focusing on one of the above tasks. They denote this design as a hybrid discriminator.

In this section, we reviewed a few methods ordered chronologically according to the recent methods of CNN, LSTM, and attention-based that they have used. Table 2 shows the performance of these methods. We do not intend to compare them because they are using different approaches, techniques, and datasets. Nevertheless, the performance and accuracy are getting better each year due to the methods, extensive datasets and captions that are assigned, and also the advancements in hardware.

B.1 VIDEO CAPTIONING DATASETS

Many datasets are used to evaluate video captioning methods. Here, we mention just a few of them and classify them into five domains based on the video contents: People, Open Subjects, Social Media, Cooking, and Movie (See Table 3).

People: The Charades dataset [42] is built up by combining 40 objects and 30 actions in 15 scenes. Sigurdsson *et al.* proposed Charades, which contains 9,848 videos (7,985 for training and 1,863 videos for test purposes) with an average length of 30 seconds of people's daily activities. The dataset comprises of 66,500 annotations describing 157 actions. It also provides 27,847 descriptions covering all the videos.

Open Subject: The Microsoft Video Description dataset (Chen and Dolan, 2011) contains 1,970 YouTube clips (1,200 videos for training, 100 videos for validation, and 670 videos for testing) with human-annotated sentences. The duration of each video in the MSVD dataset is typically between 10 to 25 seconds. On average, 41 descriptions for each video [43] are there. Krishna *et al.* [27] presented the ActivityNet Captions dataset, a large-scale benchmark for dense-captioning events, which contains 20k videos amounting to 849 hours with 100k total descriptions. Xu *et al.* [44] presented the MSR-VTT dataset (standing for MSR-Video to Text). This is created by collecting 257 popular queries from a commercial video search engine, with 118 videos for each query. MSR-VTT provides 41.2 hours of 10K web video clips with 200K clip-sentence pairs in total, covering a list of 20 categories.

Social media: VideoStory [45] is a dataset for telling the stories of social media videos. It contains 20k videos amounting to 396 hours of video with 123k sentences.

Cooking: Max Plank Institute for Informatics (MPII) Cooking dataset [46] presents 65 fine-grained cooking activities. The dataset is comprised of 44 videos with an average length of 600 seconds per clip. Regneri *et al.* [15] presented Textually Annotated Cooking Scenes (TACoS), which provides coherent textual descriptions for high-quality videos, and contains 26 fine-grained cooking activities in 127 videos. In 2018, Zhou *et al.* [47] collected a large-scale procedure segmentation dataset with procedure segments temporally localized and described; they used cooking videos and named the dataset YouCook2. It contains 176 hours of runtime comprised of 2000 videos that are nearly equally distributed over 89 recipes from Africa, America, Asia, and Europe.

Movie: The Large Scale Movie Description Challenge (LSMDC, Rohrbach *et al.*, 2017) dataset [16], which provides transcribed and aligned Audio Description and script data sentences, is based on 200 movies and has 128,118 sentences with aligned clips (around 150 hours of video in total). LSMDC is based on the MPII-MD dataset and the M-VAD dataset, which were initially collected independently but are presented jointly in this work. The MPII Movie Description (MPII-MD) [48] dataset contains a parallel corpus of over 68K sentences and video snippets from 94 HD movies. The Montreal Video Annotation dataset (M-VAD) [26] includes over 84.6 hours of paired video and sentences from 92 DVDs.

C. IMAGE AND VIDEO CAPTIONING EVALUATION METRICS

Captions are evaluated using the BLEU, METEOR, CIDEr, and other metrics [17]–[19]. These metrics are common for comparing the different image and video captioning models and have varying degrees of similarity with human judgment [42].

1) BLEU

BiLingual Evaluation Understudy is a method of automatic machine translation evaluation that is a precision-based metric, correlates highly with human evaluation, and has a little marginal cost per run [17], [49]. BLEU has different n-grams based versions for candidate sentences concerning the reference sentences.

2) METEOR

Metric for Evaluation of Translation with Explicit ORdering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations [17], [18], [50], [51].

3) CIDEr

Consensus-based Image Description Evaluation [19] enables an objective comparison of machine generation approaches based on their human-likeness, without having to make arbitrary calls on weighing content, grammar, saliency, etc. concerning each other. CIDEr was first developed specifically for evaluating image captioning tasks, but it is also used in video captioning methods.

4) ROUGE

Recall-Oriented Understudy for Gisting Evaluation [52] determines the quality of a summary by comparing it to other summaries created by humans. ROUGE, similar to BLEU, has different n-grams based versions.

5) SPICE

Anderson *et al.* [53] introduced Semantic Propositional Image Captioning Evaluation, a novel semantic evaluation metric that measures how effectively image captions recover objects, attributes, and the relations between them. It correlates more with the human judgment of semantic quality as compared to previously reported metrics.

6) WMD

Word Mover's Distance [54] measures the dissimilarity between two text documents. Therefore, the sensitivity of this metric when compared to BLUE, ROUGE, and CIDEr, is low about word order or synonym swapping, but, like CIDEr and METEOR, it provides a high correlation with human judgments.

IV. THE REQUIRED PLATFORM FOR IMPLEMENTATION

Deep Learning has dramatically improved the accuracy of image recognition. Image recognition is considered to be one of the most challenging problems in image science. In recent years, deep learning-based convolutional neural networks have positively and significantly impacted the field of image recognition allowing much flexibility. Deep Learning is responsible for many of the recent breakthroughs in image science, such as image and video captioning. Despite Deep Learning's popularity, it is difficult to accurately predict the time that it takes to train a deep learning network to solve a given problem. The training time can be seen as the product of the training time per epoch and the number of epochs that need to be performed to reach the desired level of accuracy. We define the features which could influence the prediction of execution time while performing the training. We categorize these features into layer, implementation, and hardware features. Each of these categories can contain almost an endless list of features. Layer (Algorithm or model) Features include Activation Function like ReLU, Softmax, and Tanh; Optimizer (e.g. Gradient Descent, Momentum, Adam); Batch Size (the number of training samples which are processed together as part of the same batch); Number of inputs to the layer, the neurons within the layer, Matrix, Kernel, Stride, and Padding size. Hardware Features include CPU, GPU, or TPU technology (regarding memory, clock, speed, and bandwidth) [55].

A. SOFTWARE REQUIREMENT

1) TENSORFLOW

TensorFlow is an end-to-end open-source platform for machine learning. TensorFlow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN and RNN with different settings. TensorFlow is designed for remarkable flexibility, portability, and high efficiency of equipped hardware [56].

2) PyTorch

PyTorch is a Python-based scientific computing package that serves two purposes: as a replacement for NumPy to use the power of GPUs and as a deep learning research platform that provides maximum flexibility and speed[1] [40].

3) KERAS

Keras is a high-level neural network API, written in Python, and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is the key to doing good research. Keras allows for easy and fast prototyping (through user-friendliness, modularity, and extensibility). Keras supports both convolutional networks and recurrent networks, as well as a combination of both. Keras runs seamlessly on CPU and GPU.[2]

B. HARDWARE REQUIREMENT

The science and methodology behind deep learning have been in existence for decades. In recent years, however, a significant acceleration in the utilization of deep learning has been due to an increasing abundance of digital data and the involvement of the powerful hardware.

[1] https://pytorch.org/tutorials/beginner/blitz/tensor-tutorial.html

[2] https://keras.io/

1) GPU

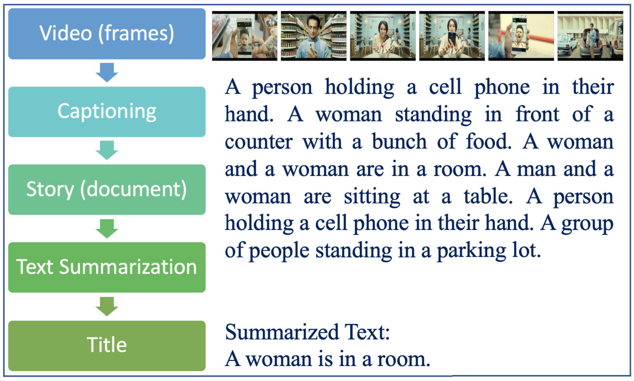
Compared to CPU, the performance of matrix multiplication on the Graphics Processing Unit is significantly better. With GPU computing resources, all the deep learning tools mentioned achieve much higher speedup when compared to their CPU-only versions [56]. GPUs have become the platform of choice for training large, complex Neural Network-based systems because of their ability to accelerate the systems. For example, it used to take a few days to train AlexNet (the work of Krizhevsky *et al.* [14] which outperformed all other image recognition approaches at the time [13]) on the ImagetNet dataset with an NVIDIA K40 machine. Now with DGX-2, the NVIDIA group can train AlexNet in a few minutes.[3] Shi *et al.* worked to evaluate the running time performance of a set of modern deep learning software tools and see how they perform on different types of neural networks and different hardware platforms. They showed that all tested tools can make good use of GPUs to achieve significant speedup over their CPU counterparts. No single software tool exists that can consistently outperform others. However, we have some opportunities to further optimize performance [56].

2) TPU

Tensor Processing Unit (Domain-Specific Architecture) is a custom chip that has been deployed in Google data centers since 2015. DNNs are dominated by tensors, so the architects created instructions that operate on tensors of data rather than one data element per instruction [57]. To reduce the time of deployment, TPU was designed to be a coprocessor on the PCI Express (PCIe) I/O bus rather than be tightly integrated with a CPU, allowing it to plug into existing servers just as a GPU does. The goal was to run whole inference models in the TPU to reduce I/O between the TPU and the host CPU. Minimalism is a virtue of domain-specific processors. Jouppi *et al.* show in their paper that the TPU leverages its advantages to run 15 times as fast as the K80 GPU, resulting in a performance/ Watt advantage of 29 times. While future CPUs and GPUs will surely run inference faster, a redesigned TPU using circa-2015 GPU memory would go three times faster and boost the performance/ Watt advantage to nearly 70 over the K80 and 200 over Haswell CPU [38], [57].

V. CASE STUDY: REAL-WORLD APPLICATION

In our earlier research, we proposed the use of deep learning models, image captioning, and NLP in an integrated manner for generating meaningful ‘‘titles’’ for videos. We now refer to this proposed system as DTVC (automatic generation of Descriptive Titles for Video Clips using deep learning). DTVC uses many different components - some of these components are presented in [13] and [68] and others are discussed in this article. For a more detailed discussion about DTVC, refer to [70]. To the best of our knowledge, our work reported in [70] is the first attempt in automatic generation of ‘‘titles’’ for videos.



A person holding a cell phone in their hand. A woman standing in front of a counter with a bunch of food. A woman and a woman are in a room. A man and a woman are sitting at a table. A person holding a cell phone in their hand. A group of people standing in a parking lot.

Summarized Text:
A woman is in a room.

FIGURE 9. An overall example of the proposed system, DTVC (automatic generation of Descriptive Titles for Video Clips using deep learning). illustrated architecture consists of two different, complementary processes: Video Captioning and Text Summarization.

In this research, we are proposing DTVC that could be applicable for the cinema industry, search engines, supervision cameras, etc. The framework with an example of the proposed system, DTVC is presented in Figure 9. The architecture consists of two different, complementary processes: Video Captioning and Text Summarization [69]. During the first process (Video Captioning), the system gets a video as its input. The keyframes of the video are selected. Each key-frame is captioned [66]. The collection of these captions results in a ‘‘story’’ describing the video. During the second process (Text Summarization), this generated ‘‘story’’ is fed into a process as a document. The document is summarized to one sentence using Text Summerization method. The manual generation of captions for video would involve a user/viewer to watch the whole video and take notes. Because of this, the manual generation of captions is considered to be a time-consuming task. The purpose of the proposed system is to automatically generate a title and also an abstract for a video clip without manual intervention. In this article, we have provided results based on our experimentation using video clips available from publicly.

VI. CONCLUSION AND FUTURE WORK

In recent years, many models have been proposed and presented to generate captions for images and short videos. Although, these models are helping to advance the technology, they suffer from inaccuracies due to fundamental constraints; resulting in limited use in practical situations. Many of the earlier models proposed, treat image captioning and video captioning differently using different algorithms and methodologies. In this article, we have focused on methodologies that perform video captioning by using image captioning methods as building blocks. Thus, the video captioning process is considered to be a compilation of the summarization of image captions. It is for the above reason that in this article, we only focused on the algorithmic overlap between image and video captioning. Therefore, this article is not meant to be a comprehensive review of image and video captioning; rather, it is a concise review of the algorithm over-

[3] https://devblogs.nvidia.com/tensor-core-ai-performance-milestones/

lap between the two. Furthermore, this article only considered those algorithms that used deep learning.

In general, comparing different deep learning models used for image and video captioning is difficult. This is due to the fact that researchers use different image datasets, different parameters, different classification methods, different preprocessing, different combinations of structures, and others. Despite the vast differences, in this study, we focused on the general overlap between these methods.

A reliable, accurate, and real-time video and image captioning method can be used in many applications. Researchers attempt to give sight to the machines. First, machines learn to see. Then, they help us to see better. We will not only use the machines because of their intelligence, but we will also collaborate with them in ways that we cannot even imagine. Image and video captioning systems can be used as an important part of Assistive Technologies that would help people with hearing or sight impairments. The captions can be used as meta-data for search engines which would take the search engine's functionality to a new dimension. Captions can be used as part of recommendation systems in many applications.

Future Research Direction and Broader Impact: As mentioned earlier, the current technologies used for image and video captioning often generate captions that are not very accurate. There is much room for improvement and enhancement. The fusion and processing of image, video, and audio would provide more accurate captions. Audio-to-Word converters are available, and they are quite reliable. Integrating an Audio-to-Word converter with a video and combining the captions/words generated via audio and video would generate more accurate and meaningful captions even though, an elaborate text/sentence summarization would have to be performed.

Another challenge with video captioning is the very compute intensive nature of the problem. With the current technology, only very short videos can be captioned (videos that are only a few seconds long). The use of the next generation of GPUs and with explicit algorithm parallelization (targeted at the GPU machine architectures), we can get closer to real-time performance for longer videos. A great opportunity in the area of video captioning is to design and develop a strategy that would permit users to request video captions at varying levels of detail. However, we believe that the most fundamental and challenging research problem with video captioning is the fact that different captions based on different interpretations can be generated for the same video - in the same way as two individuals can come-up with two different views/description by watching the same video. We believe that this fundamental problem can be addressed by studying relevant concepts and making the process more interactive.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[3] 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.

[4] S. M. Ehandarkar and H. R. Arabnia, "Parallel computer vision on a reconfigurable multiprocessor network," *IEEE Trans. Parallel Distrib. Syst.*, vol. 8, no. 3, pp. 292–309, Mar. 1997.

[5] H. Valafar, H. R. Arabnia, and G. Williams, "Distributed global optimization and its development on the multiring network," *Neural, Parallel Sci. Comput.*, vol. 12, no. 4, pp. 465–490, 2004.

[6] D. Luper, D. Cameron, J. Miller, and H. R. Arabnia, "Spatial and temporal target association through semantic analysis and GPS data mining.," in *Proc. IKE*, vol. 7, 2007, pp. 25–28.

[7] R. Jafri and H. R. Arabnia, "Fusion of face and gait for automatic human recognition," in *Proc. 5th Int. Conf. Inf. Technol., New Generat.*, vol. 1, Apr. 2008, pp. 167–173.

[8] H. R. Arabnia, W.-C. Fang, C. Lee, and Y. Zhang, "Context-aware middleware and intelligent agents for smart environments," *IEEE Intell. Syst.*, vol. 25, no. 2, pp. 10–11, Mar. 2010.

[9] R. Jafri, S. A. Ali, and H. R. Arabnia, "Computer vision-based object recognition for the visually impaired using visual tags," in *Proc. Int. Conf. Image Process., Comput. Vis., and Pattern Recognit. (IPCV). Steering Committee World Congr. Comput. Sci., Comput. Eng. Appl. Comput. (WorldComp)*, 2013, p. 1.

[10] L. Deligiannidis and H. R. Arabnia, "Parallel video processing techniques for surveillance applications," in *Proc. Int. Conf. Comput. Sci. Comput. Intell.*, Mar. 2014, pp. 183–189.

[11] E. Parcham, N. Mandami, A. N. Washington, and H. R. Arabnia, "Facial expression recognition based on fuzzy networks," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2016, pp. 829–835.

[12] 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.

[13] 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.

[14] 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.

[15] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

[16] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 94–120, 2017.

[17] 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.

[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*. [Online]. Available: http://arxiv.org/abs/1504.00325

[19] 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.

[20] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2010, pp. 15–29.

[21] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[22] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.

[23] 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.

[24] 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.

[25] 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*. [Online]. Available: http://arxiv.org/abs/1604.01729

[26] 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*. [Online]. Available: http://arxiv.org/abs/1503.01070

[27] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.

[28] 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.

[29] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using Web-scale n-grams," in *Proc. 15th Conf. Comput. Natural Lang. Learning. Assoc. Comput. Linguistics*, 2011, pp. 220–228.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[31] 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.

[32] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[33] B. Romera-Paredes and P. H. S. Torr, "Recurrent instance segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 312–329.

[34] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[35] 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]. Available: http://arxiv.org/abs/1406.1078

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: http://arxiv.org/abs/1411.2539

[38] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: https://arxiv.org/abs/1609.08144

[39] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2321–2334, Dec. 2017.

[40] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5561–5570.

[41] 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.

[42] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowd sourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 510–526.

[43] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2011, pp. 190–200.

[44] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[45] S. Gella, M. Lewis, and M. Rohrbach, "A dataset for telling the stories of social media videos," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 968–974.

[46] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1194–1201.

[47] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from Web instructional videos," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[48] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3202–3212.

[49] 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.

[50] 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.

[51] 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.

[52] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[53] 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.

[54] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.

[55] D. Justus, J. Brennan, S. Bonner, and A. S. McGough, "Predicting the computational cost of deep learning models," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 3873–3882.

[56] S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 99–104.

[57] N. Jouppi, C. Young, N. Patil, and D. Patterson, "Motivation for and evaluation of the first tensor processing unit," *IEEE Micro*, vol. 38, no. 3, pp. 10–19, May 2018.

[58] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[59] 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.

[60] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[61] T. Iqbal and H. Ali, "Generative adversarial network for medical images (MI-GAN)," *J. Med. Syst.*, vol. 42, no. 11, p. 231, Nov. 2018.

[62] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2970–2979.

[63] O. Nezami, M. Dras, S. Wan, C. Paris, and L. Hamey, "Towards generating stylized image captions via adversarial training," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 270–284.

[64] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6588–6597.

[65] J. S. Park, M. Rohrbach, T. Darrell, and A. Rohrbach, "Adversarial inference for multi-sentence video description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[66] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, "Discriminability objective for training descriptive captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6964–6974.

[67] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, ''Visual genome: Connecting language and vision using crowdsourced dense image annotations,'' *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[68] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, ''Image captioning with generative adversarial network,'' in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2019, pp. 272–275.

[69] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, ''Text summarization techniques: A brief survey,'' 2017, *arXiv:1707.02268*. [Online]. Available: http://arxiv.org/abs/1707.02268

[70] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, ''Automatic generation of descriptive titles for video clips using deep learning,'' in *Transactions on Computational Science & Computational Intelligence* (Advances in Artificial Intelligence & Applied Cognitive Computing). Springer, 2020. [Online]. Available: https://www.springer.com/series/11769

[71] Microsoft Research. (2019). Accessed: Apr. 4, 2019. [Online]. Available: https://caption.ai

[72] S. Ding, S. Qu, Y. Xi, and S. Wan, ''Stimulus-driven and concept-driven analysis for image caption generation,'' *Neurocomputing*, vol. 398, pp. 520–530, Jul. 2020.

[73] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, ''Meshed-memory transformer for image captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10578–10587.

[74] Y. Pan, T. Yao, Y. Li, and T. Mei, ''X-linear attention networks for image captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.

[75] 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.

[76] T. Yao, Y. Pan, Y. Li, and T. Mei, ''Exploring visual relationship for image captioning,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.

SOHEYLA AMIRIAN received the B.Sc. degree in computer software engineering in 2006 and the M.Sc. degree in information technology, engineering, computer networks from the Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2013. She is currently pursuing the Ph.D. degree (under the supervision of Prof. Arabnia; with Prof. Rasheed and Prof. Taha as Ph.D. committee members) with the University of Georgia, Athens, GA, USA. She was a Lecturer in computer science and a Course Coordinator of the IT Program with the Computer Science Department, Technical and Vocational University, a Lecturer with the University of Applied Science (Academic Center for Education, Culture and Research), and a Teacher at the Technical High School, Iran, from 2001 to 2016. She is currently an Instructor of record in computer science with the University of Georgia, where she has been investigating new ways in utilizing deep learning methodologies for imaging applications. She has received many scholarships from Anita Borg Grace Hopper, TAPIA Conference, and CRA-W Grad Cohort. Also, she was awarded the 2019 International Conference on Computational Science and Computational Intelligence CSCI 2019 Outstanding Achievement Award, and she was named as the Finalist of the 2020 National Center for Women and Information Technology (NCWIT) Collegiate Award.

KHALED RASHEED received the Ph.D. degree in computer science from Rutgers University, in January 1998. He is currently a Professor with the Department of Computer Science, University of Georgia. He is also the Director of the UGA Institute for Artificial Intelligence, the Director of the Evolutionary Computation and Machine Learning (ECML) Laboratory, and a member with the Georgia Informatics Institutes and the UGA Faculty of Robotics. He has authored more than 100 research articles. His research interests include artificial intelligence methods, including genetic algorithms, evolutionary computation, and machine learning; and artificial intelligence applications, including engineering design optimization, computational biology, and bioinformatics.

THIAB R. TAHA received the Ph.D. degree from Clarkson University, in 1982. He joined UGA, in 1982. He has been a Professor and the Head of the Computer Science Department, UGA, since July 2013. He is also the Director of the UGA CUDA Teaching and Research Centers and the Big Data Consulting Services and Training Center. He is also an Adjunct Faculty with the Institute of Bioinformatics, UGA. He has published more than 80 research articles and has given more than 100 invited talks or keynotes at international conferences. His research interests include scientific and distributed computing and software development for solving problems in nonlinear waves, optical fiber communication systems, biochemical reaction networks, and related areas. He received the M. G. Michael Award for Research in the Sciences at UGA, in 1985. He was the Fulbright Scholar from 1995 to 1996. He received several grants from NSF and DOE, and the ARO in support of his research. He is a Senior Editor of the *Mathematics and Computers in Simulation* journal, the Co-Editor-in-Chief of the APNUM journal, and the Chair and a Conference Coordinator of the IMACS International Conferences on Nonlinear Waves: Computation and Theory since 1999.

HAMID R. ARABNIA received the Ph.D. degree in computer science from the University of Kent, Canterbury, in 1987. He has been with the University of Georgia, Athens, GA, USA, since 1987, where he is currently a Professor Emeritus of computer science. He is a Fellow and an Advisor of the Center of Excellence in Terrorism, Resilience, Intelligence, and Organized Crime Research. His research interests include parallel and distributed processing techniques and algorithms, supercomputing, big data analytics (in the context of scalable HPC), imaging science (image processing, computer vision, and computer graphics), and other compute-intensive problems. His great research interest includes methodologies that promote cross-disciplinary education. He has authored extensively in journals and refereed conference proceedings. He has authored or coauthored about 200 peer-reviewed research publications and also 250 edited research books in his areas of expertise. Applications of his interests include health informatics, medical imaging, and security. His most recent activities include studying ways to promote legislation that would prevent cyberstalking, cyber harassment, and cyberbullying. He is also the Editor-in-Chief of *The Journal of Supercomputing* (Springer).