IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 14, 2018, accepted February 28, 2018, date of publication March 19, 2018, date of current version April 18, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2814075

Natural Language Description of Video Streams Using Task-Specific Feature Encoding

ANIQA DILAWARI[1], MUHAMMAD USMAN GHANI KHAN[1,2], AMMARAH FAROOQ[2], ZAHOOR-UR-REHMAN[3], SEUNGMIN RHO[4], AND IRFAN MEHMOOD[5]

[1]Department of Computer Science and Engineering, University of Engineering & Technology at Lahore, Lahore 54890, Pakistan
[2]Al-Khawarizmi Institute of Computer Science, University of Engineering & Technology at Lahore, Lahore 54890, Pakistan
[3]COMSATS Institute of Information Technology Attock, Attock 43600, Pakistan
[4]Department of Media Software, Sungkyul University, Anyang 430-742, South Korea
[5]Department of Software, Sejong University, Seoul 143-747, South Korea

Corresponding authors: Seungmin Rho (smrho@sungkyul.edu) and Irfan Mehmood (irfan@sejong.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2016R1D1A1A09919551.

ABSTRACT In recent years, deep learning approaches have gained great attention due to their superior performance and the availability of high speed computing resources. These approaches are also extended towards the real time processing of multimedia content exploiting its spatial and temporal structure. In this paper, we propose a deep learning-based video description framework which first extracts visual features from video frames using deep convolutional neural networks (CNN) and then pass the derived representations into a long-short term memory-based language model. In order to capture accurate information for human presence, a fine-tuned multi-task CNN is presented. The proposed pipeline is end-to-end, trainable, and capable of learning dense visual features along with an accurate framework for the generation of natural language descriptions of video streams. The evaluation is done by calculating Metric for Evaluation of Translation with Explicit ORdering and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores between system generated and human annotated video descriptions for a carefully designed data set. The video descriptions generated by the traditional feature learning and proposed deep learning frameworks are also compared through the ROUGE scores.

INDEX TERMS Convolutional neural network, encoder-decoder, LSTM, natural language generation, TRECVid 2007/2008, video description, video to text.

I. INTRODUCTION

With the exponential growth of digital multimedia, huge volumes of video data are being generated and are becoming a central part of the big data domain. The reason for this rise, is the increase in storage space and internet bandwidth. It has stimulated development for a broad range of video understanding applications. Among these applications, automatically describing a video in a natural language sentence requires understanding the video content (humans, key objects and their actions) and its spatio-temporal relationships. The automatic generation of video descriptions is a challenging task that requires computer vision and machine learning techniques in addition to natural language generation capabilities. There are a number of prospective applications that bring vision and language together, such as video retrieval (enhanced video search), automatic video subtitling, and navigation systems for the blind and visually impaired.

Even though the nature of this problem is challenging, there are many instances where image descriptions [1]–[5] have been proposed in constrained environments. Most of the previous work is based on taking static visual content as input and generating a single/multiple English sentence(s) as output. These image description approaches have mostly been used for short video segments consisting of a single human action or homogenous scene settings. The lessons learned from these approaches and the temporal characteristic of the video resulted in new approaches to solve the problem of automatic video description.

Deep learning has been applied to a variety of solutions in the last decade. In this paper, we propose to translate videos into natural language descriptions using a deep neural network where we have rectified the shortcomings of the previous approaches by providing a framework which can handle a variety of video streams. Deep networks are capable of learning powerful features directly from the raw data. They give us relief from traditional hand-engineered features. A proposed framework is based on encoder-decoder formation where encoder extracts representations from a given video and decoder translates those features into natural language sentences.

The main contributions of this research are as follows:

- We have introduced a deep learning video description framework that uses task-specific feature learning.
- The proposed framework generates descriptions for a higher number of high level features (HLFs) extracted from video streams compared to previous methods where these HLFs are quite limited.
- A video description task is presented as both machine translation (METEOR scores) and summarization (ROUGE scores) task.

The rest of the paper is organized in the following sections. Section 2 gives the overview on the related work for video description problems. Sections 3 and 4 provide detailed explanations of the proposed method, dataset and experimental setup. Results are presented in Section 5 followed by a discussion and conclusion in the final sections.

II. RELATED WORK

Early work on generating textual descriptions of videos used a two stage process that first identifies a subject, verb and object (SVO) triplet from the video frame and then generates a sentence based on a template [1], [2]. This requires training of individual classifiers to recognize humans and their attributes, objects, the scene, and the interaction between them. This information is combined with a language model to estimate the most relevant content to generate a sentence. While this problem is simplified, it still requires the selection of appropriate objects and actions. Also a template based approach does not effectively combine attributes for a good description. Moreover, missing, erroneous and misidentified information extracted from the video frames leads to disjointed descriptions.

Natural language description of images has received considerable attention and now the focus is shifting towards video descriptions using deep learning approaches. The simple idea for video descriptions is to use convolutional neural networks (CNN) to encode video content and recurrent neural networks (RNN) to decode it into a textual sentence. RNNs are networks with loops that allow information to persist. These networks have been used for a variety of problems such as language modeling, speech recognition, image or video captioning, description, translation and more.

The most cited state-of-the-art deep learning technique for video captioning is hierarchical RNN (hRNN) [6]. Yu *et al.* utilized this method to generate multiple sentences from a long video containing more than one event. These types of videos cannot be described semantically in one sentence, but rather give a mundane description. For example, instead of saying 'Person A is throwing a ball towards Person B. Person B catches the ball and high-fives his teammates', a method that can only generate a single sentence will sum up this event as 'people are playing with a ball'. This not only misses out most of the details but it is also uninformative and yields unexciting results. The notion of this hierarchical framework is to make use of the temporal dependency between the sentences in a paragraph. The semantic context of the previous sentence is taken into consideration while generating the next sentence. There are two generators in this framework. A sentence generator, focuses on spatial and temporal information present in a precise time interval of a video while generating a single sentence. A paragraph generator models dependency between the sentences.

RNNs conserve information in memory over a period of time. Nonetheless, they do not solve problems that require long-term temporal dependencies. The problem of vanishing and exploding gradients is also evident during the training of these networks. LSTM networks are special types of RNNs which work for numerous tasks and perform better than the standard version. An LSTM unit contains a memory cell that can sustain information for a longer period of time.

In literature, we have found many instances where different versions of LSTMs are used to solve the problem of video description. Long *et al.* [7] have suggested an LSTM with a multi-faceted attention model which takes temporal, motion and semantic properties. Visual features include temporal properties obtained from a pre-trained ResNet-152 model and motion properties extracted using a pre-trained C3D model. Three models have been used to mine semantic attributes: nearest neighbor (NN) search, Support Vector Machine (SVM) training of 100 common attributes in training data set and hierarchical recurrent neural encoders (HRNE) for a subject and verb prediction based on temporal features. Experiments on MSVD and MSR-VTT show results that are competitive with human results. Similarly, Zanfir *et al.* [8] also utilized semantic information to generate video captions. The semantic expressive sentences are produced by learning how to find subject, verb and object (S, V, O) tuples using pre-trained object detector models. The Attention-LSTM takes image classifiers and learned sematic attributes in (SVO) form to visually establish each of the words from the sentence it creates. Experiments have been performed on YouTube datasets, and competitive results were achieved from the baseline methods.

Song *et al.* [9] suggested a hierarchical LSTM approach with adjusted temporal attention to solve video captioning. This framework uses temporal attention to choose frames that will predict the visual words, and the adjusted temporal attention to choose whether to rely on the language context or visual information. Experiments were performed on MSVD and MSR-VTT datasets to test the effectiveness of this approach.

Pan *et al.* [10] presented LSTM with transferred semantic attributes (LSTM–TSA) architecture where the semantic features were extracted from images and videos using CNN and RNN frameworks. In this approach, the semantic features helped to learn the sequence for video captioning. The semantic features in images reflect the static objects and scenes whereas the video exudes temporal dynamics. These two sources are merged together to improve video captioning. Experiments have been carried out on MSVD, M-VAD and MPII-MD datasets. Results showed top performance on

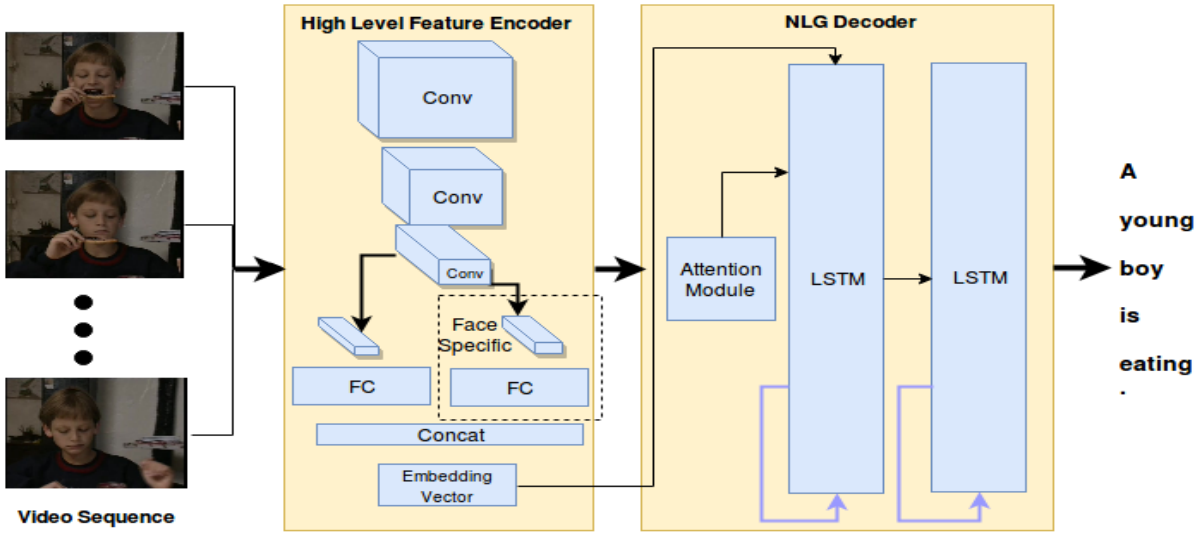


FIGURE 1. Overview of proposed deep learning framework for video description generation; face specific feature encoding helps in accurately distinguishing human related features.

MSVD dataset and higher results on M-VAD and MPII-MD datasets.

Gated Recurrent Units (GRU) are similar to LSTMs but they have a simplified architecture. They use a set of gates to control information flow, but do not use separate memory cells, hence, there a few number of gates. GRUs have been used frequently in the context of machine translation. A GRU is comparatively new, trains faster and is computationally efficient because of its non-complex structure. Both LSTMs and GRUs prevent the vanishing gradient problem.

Ballas *et al.* [11] suggested an approach to acquire spatial and temporal properties in videos from intermediate visual representations known as percepts using GRU networks. The visual percepts are extracted using VGG-16 CNN which are pre-trained on ImageNet dataset and adjusted on UCF101. This approach was validated on YouTube2Text and UCF101 datasets and showed equivalent results in comparison to the new approaches.

In this work, we propose a CNN and LSTM based video description generation system focusing more on the accurate prediction of human facial features. This framework is useful for many applications like surveillance, tracking and identification, in the case of limited memory resources.

III. PROPOSED METHODOLOGY

The overview of the proposed framework is presented in Figure 1. It is a typical encoder-decoder based formation used for natural language processing and machine translation. The encoder stage is based on the convolutional neural network which is responsible for extracting visual content information from the video frames. The visual content vector is then passed to a long short-term memory (LSTM) based language model to generate a textual description of the video. To generate semantically more appropriate descriptions, an attention module is also introduced in the framework. Implementation details of each stage are given in the following subsections.

A. HIGH LEVEL FEATURE ENCODING USING CNN

For a given video sequence, a high level feature encoding stage is required to properly understand and extract the visual characteristics of the objects and persons present in the specific video frame. For this purpose, a convolutional neural network (CNN) is used as its hierarchical structure, which enables it to learn features on multiple scales that are invariant to spatial changes. Its structure exploits the local spatial regions for feature learning and max-pooling takes care of incorporating global features.

CNNs provide promising visual understanding [19], [20]. Most of the recent applications are built on using pre-trained CNN models. The choice of model depends on the context, requirements and resources of the application being developed. Most of the models are trained on the largest ImageNet dataset containing 1.2 million images for the object recognition task. Besides the object recognition, these models are powerful enough to extract the visual information required for any vision problems.

In the proposed work, we have used the famous VGG-16 network due to its superior performance in ILSVRC [18]. It has a homogeneous network structure with 3×3 dimensional convolving filters. For the passing feature vector to decoder stage, we have chopped off the top fully connected layer of the network and replaced it with a linear layer, which converts the FC7 features into an embedding vector size of 256. The model is trained on object centric data. We have also fine-tuned the CNN model while training the combined pipeline. In order to take out maximum information related to humans i.e. age, gender and emotion, we have also trained task specific top CNN layers for this purpose. The combined features are robust and the final output performs

surprisingly well on the TRECVid dataset. The weights for the face specific task are updated along with the combined training in a supervised manner. For this purpose, each video is also loosely annotated for the three face related features mentioned above. The video inputs are passed to the CNN frame by frame and features of all frames are fused and concatenated with the face vector to form one single vector by the max pooling operation before the formation of the embedding vector.

B. LONG SHORT-TERM MEMORY DECODER

Recurrent neural networks (RNN) and LSTMs [16] are the building blocks of deep learning based sequence modeling applications. This is due to their ability to take the input sequential data and predict the upcoming member of the sequence. Caption and description generation tasks are paying even more attention to these models and making multiple architectural variations for better performance. However, in our design we are using a basic version of LSTM with a soft attention model [1].

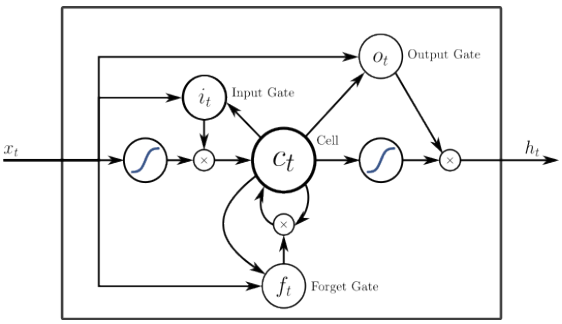


FIGURE 2. Structure OF LSTM. (Reprinted from [17]).

The structure of LSTM presented in Figure 2, is composed of an internal memory unit called cell (ct), a hidden layer output (ht) and several gates associated with these units. The data present in the cell remains hidden from the outside processing. The output of LSTM corresponds with the state of hidden unit. The overall structure of the LSTM network is similar to the multi-layer neural network, which has multiple hidden layers with numerous nodes. However, each node possesses a recurrence relationship. The function of the gates is to preserve and update the state of the cell and output accordingly. The main advantage of LSTM lies in flowing gradients in the backward pass, which prevents the vanishing gradients issue prevailing in the basic RNNs. For the task of video captioning, we have used two layers of LSTM with 512 hidden nodes each. The input to the first layer is the embedding vector from CNN and the attention vector from the soft attention module. The attention module emphasises the distribution over the spatial location in order to focus the specific part of the image at a given time instance. A set of vocabulary was generated using handwritten annotations. Each video was annotated by 40 different annotators with varying levels of words complexity. A word occurring less than 5 times is assigned to a special unknown <UNK> token. A total of 5600 sentences were processed. Target captions have been fed to the first layer during training. However, at the time of the test, the output generated a word which was fed back to the first layer to predict the next word in the caption.

IV. EXPERIMENTAL SETUP

A. DATASET

For the experiments, we used the dataset prepared by Khan *et al.* [12]. The dataset comprised of videos taken from 2007/2008 TREC video benchmarks. The total number of video segments was 140, while the length of individual video segments ranges from is from 10 to 30 seconds with single camera shot. Annotations were completed by 40 humans. They ranged from one to seven sentences and were referred to as hand annotations. These videos are categorized in seven groups:

1) Action: Humans performing an activity e.g. walking or eating

2) Close up: Frontal view of a human face that shows emotions or expressions e.g. happy or angry

3) News: Scene settings that show a reporter, anchor or weather boards

4) Meeting: Gathering of humans that are interacting with one another; object settings such as the presence of objects commonly found in meeting scenes like chairs, tables and curtains etc.

5) Grouping: The presence of multiple humans but not in any particular scene setting

6) Traffic: The presence of vehicles, traffic signals

7) Scene Category: Indoor, outdoor, room, building, roads

B. TRAINING

All experiments were performed using an Intel core i7 CPU system with 16 GB memory. The framework was developed using PyTorch [21] deep learning framework. Both CNN and LSTM models were jointly trained and optimized by Adam optimizer with a batch size of 16 (due to memory limitation) using cross-entropy loss. The loss function converged in approximately 78 hrs (∼3 days) of training with a learning rate of 0.001. Each video frame was converted to a dimension of 256 × 256 before passing to CNN. The video was fed to the network with a rate of 4 frames per second. The data split ratio was selected as 75/25% for training and testing purposes respectively.

V. RESULTS

A. Machine Generated Annotation Samples

The results of three videos randomly selected from the test set of the dataset are presented in Figure 3 to 5. Each figure shows three video descriptions: deep learning (DL) based framework, machine annotation taken from Khan and Gotoh [15] and hand annotation 1 (randomly selected from the set of 40 hand annotations). Our framework clearly showed the improvements in the video description compared to the machine annotations from [15].

FIGURE 3. Video montage taken from Closeup category – Video no. 'MS206410 ' from the 2007 rushes video summarization track.

FIGURE 4. A road traffic video montage taken from '20041101 110000 CCTV4 NEWS3 CHN' from TRECVid 2006.

FIGURE 5. A video montage taken from the 'Action' category – '20041101 160000 CCTV4 DAILY NEWS CHN' from TRECVid 2007.

1) FACE CLOSEUP

The frontal face view is the main focus of this group. Here the DL framework correctly identified the emotion and gender of the person, along with other objects. The hand annotations described more, for instance, the identity of the person who was a policeman, clothing information and scene settings.

DL Framework: A serious man is standing in front of a building. He is talking. There are many humans around him.

Machine annotation [12]: This is an outdoor scene. A serious man is speaking. There are many humans in the background. Sample Hand Annotation: This is a scene out of a court. A man with brown hair, wearing a formal suit is explaining his points to someone else. He looks quite serious and sad. He has swollen face. A police man and a lady with a hat is present in the background.

2) TRAFFIC SCENE

The description contains a correct vehicle identification and scene setting. It has produced better results compared to the machine annotations. The hand annotations have more detail such as the vehicle color and other objects (e.g. a bridge, a flyover, road signs and neighboring buildings).

DL Framework: Many cars and a bus are travelling on the road. This is a traffic scene.

Machine annotation: There are many cars. Cars are moving. There is a bus.

Hand annotation 1: This is a video of a high way in the day time. There are many cars and a bus. There is a fly-over in the far end.

3) ACTION SCENE

The main characteristics of this category includes the identification of humans and their activities. We can see in the description generated by the DL framework the correct identification of humans and their actions which leads to a well phrased description. The machine annotation, on the other hand, has repetitive information which can be compiled as one sentence.

DL Framework: A man and a woman are talking in an outdoor scene. The woman is sitting on a chair and the man is standing in front of her. Behind them there is a car and a bus.

Machine annotation: A man is standing while a woman is sitting; this is an outdoor scene. There is a bus and a car in the background.

Hand annotation 1: A man and woman are talking to each other in a parking place; Both are wearing formal clothes. A bus is passing by in the background. Both man and woman look serious.

B. Evaluation Metrics

The video to textual description problem can be modeled as either a machine translation or machine summarization problem. To quantitatively evaluate the performance of our framework, we have used METEOR [13] and ROUGE [14] because of their robust performance. METEOR (Metric for Evaluation of Translation with Explicit ORdering) a standard evaluation measure for machine translation makes use of harmonic mean of unigram recall and precision. It has several features such as stemming and synonym matching. This score uses both recall and precision. The value of this measure lies between 0 and 1. ROUGE compares word to word matching between the candidate (computer generated) and a set of reference (human generated) translations. If several reference translations are present, the METEOR results are measured for each individual translation and the best performing score is selected. Table I, shows the METEOR score for the dataset taken from [15]. The results show that the close-up category has the highest score because the framework learns face specific features.

TABLE 1. METEOR score.

action	closeup	in/out	grouping	meeting	news	traffic
31	39.0	29.2	32.5	33	30.8	33.7

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used for evaluating machine summarization tasks. Several variants of ROUGE include ROUGE-N (n-unigram), ROUGE-L (longest subsequence) and ROUGE-S (skip bigram). Table 2, shows the ROUGE results where a higher score indicates that the DL framework generated description and the hand annotations are closely matched. We have compared the ROUGE score from the DL framework and the machine annotations generated by using a bottom-up approach in [15]. We can see that the close-up group has

TABLE 2. ROUGE score between deep learning framework descriptions and hand annotations.

	Action	Close-up	In/Out	Group	Meeting	News	Traffic
ROUGE 1	0.5521	0.8015	0.5329	0.6661	0.6951	0.6512	0.6887
ROUGE 2	0.5166	0.7692	0.4455	0.6020	0.5700	0.5803	0.5512
ROUGE 3	0.4703	0.6967	0.3958	0.5422	0.5619	0.5083	0.5305
ROUGE L	0.5498	0.6100	0.5201	0.6614	0.6915	0.5566	0.6661
ROUGE W	0.5010	0.5978	0.4990	0.6212	0.5233	0.5492	0.6456
ROUGE S	0.4992	0.5526	0.4289	0.5145	0.5425	0.5104	0.6125
ROUGE SU	0.5052	0.5822	0.5010	0.6330	0.4929	0.5641	0.6531

the highest score because the framework has learned the face specific features separately. Similarly, the dataset categories–news, meeting, grouping and traffic–have good results.

VI. CONCLUSION

The deep learning based framework is presented for the task of natural language description of video sequences. The proposed framework generated descriptions, which were compact, scalable and versatile. The comparison using the standard measures of machine translation and summarization proves the superiority of the proposed model in comparison to previous approaches.

In the future, we are planning to extend the architecture of LSTM and use different variations of RNN, such as GRU. We then plan on comparing the results with these extensions to see which performs best with our data. We will also perform more rigorous experiments and evaluation on large data and include more HLFs for multi-task feature learning.

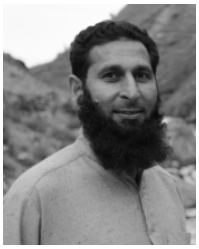
REFERENCES

[1] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[2] H. Fang *et al.*, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1473–1482.
[3] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2407–2415.
[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.
[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
[6] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2016, pp. 4584–4593.
[7] X. Long, C. Gan, and G. de Melo. (2016). "Video captioning with multi-faceted attention." [Online]. Available: https://arxiv.org/abs/1612.00234
[8] M. Zanfir, E. Marinoiu, and C. Sminchisescu, "Spatio-temporal attention models for grounded video captioning," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 104–119.
[9] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. T. Shen. (2017). "Hierarchical LSTM with adjusted temporal attention for video captioning." [Online]. Available: https://arxiv.org/abs/1706.01231
[10] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6504–6512.
[11] N. Ballas, L. Yao, C. Pal, and A. Courville. (2015). "Delving deeper into convolutional networks for learning video representations." [Online]. Available: https://arxiv.org/abs/1511.06432
[12] M. U. G. Khan, R. M. A. Nawab, and Y. Gotoh, "Natural language descriptions of visual scenes: Corpus generation and analysis," in *Proc. Joint Workshop Exploiting Synergies Between Inf. Retr. Mach. Transl. (ESIRMT) Hybrid Approaches Mach. Transl. (HyTra)*, 2012, pp. 38–47.
[13] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Statist. Mach. Transl.*, 2014, pp. 376–380.
[14] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, 2004, pp. 1–8.
[15] M. U. G. Khan and Y. Gotoh, "Describing video contents in natural language," in *Proc. Workshop Innov. Hybrid Approaches Process. Textual Data*, 2012, pp. 27–35.
[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[17] Wikipedia. *Long Short-Term Memory*. Accessed: Jan. 5, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Long_short-term_memory
[18] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[19] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[20] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[21] Pytorch. *Pytorch: Tensors and Dynamic Neural Networks in Python With Strong GPU Acceleration*. Accessed: Dec. 2, 2017. [Online]. Available: http://pytorch.org/

ANIQA DILAWARI is currently pursuing the Ph.D. degree from the Department of Computer Science and Engineering, University of Engineering & Technology at Lahore, Lahore, Pakistan. Her current research interests include image processing, natural language processing, pattern recognition, and deep learning in image/video analysis, under the supervision of Dr. M. U. G. Khan.

MUHAMMAD USMAN GHANI KHAN received the Ph.D. degree from Sheffield University, U.K., concerned with statistical modeling for machine vision signals, specifically language descriptions of video streams. He is currently an Associate Professor with the Department of Computer Science and Engineering, University of Engineering & Technology at Lahore, Lahore, Pakistan, where he is the Head of the National Center for Artificial Intelligence, Al-Khwarizmi Institute of Computer Science.

AMMARAH FAROOQ received the master's degree in computer engineering from the National University of Sciences and Technology in 2017. She is currently a Research Officer and a Team Lead at the Computer Vision and Machine Learning Lab, Al-Khwarizmi Institute of Computer Science. Her research interests include deep learning, pattern recognition, medical image analysis, and artificial intelligence.



ZAHOOR-UR-REHMAN has completed his educational and academic training at the University of Peshawar, Foundation University Islamabad, and UET Lahore, Pakistan. He joined COMSATS Institute of Information Technology as an Assistant Professor in 2015. He has experience both in academia and research. Along with teaching responsibilities, he is an active researcher and Reviewers of various conferences and reputed journals.

SEUNGMIN RHO is currently a Faculty Member of the Department of Media Software, Sungkyul University, South Korea. His current research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management, and computational intelligence.

IRFAN MEHMOOD has been involved in IT industry and academia in Pakistan and South Korea for over a decade. He is currently serving as an Assistant Professor with the Department of Software, Sejong University. His sustained contribution at various research and industry-collaborative projects gives him an extra edge to meet the current challenges faced in the field of multimedia analytics. Specifically, he has made significant contribution in the areas of visual surveillance, information mining, and data encryption.