applied sciences

Article

Natural Language Description of Videos for Smart Surveillance

Aniqa Dilawari [1], Muhammad Usman Ghani Khan [1], Yasser D. Al-Otaibi [2], Zahoor-ur Rehman [3], Atta-ur Rahman [4] and Yunyoung Nam [5,*]

[1] Department of Computer Science, University of Engineering & Technology, Lahore 54890, Pakistan; aniqa.dilawari@gmail.com (A.D.); usman.ghani@kics.edu.pk (M.U.G.K.)
[2] Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah 21911, Saudi Arabia; yalotaibi@kau.edu.sa
[3] Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan; xahoor@gmail.com
[4] Department of Computer Science, College of Computer and Information Technology, Imam Abdulrahman bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia; aaurrahman@iau.edu.sa
[5] Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea
* Correspondence: ynam@sch.ac.kr

Citation: Dilawari, A.; Khan, M.U.G.; Al-Otaibi, Y.D.; Rehman, Z.-u.; Rahman, A.-u.; Nam, Y. Natural Language Description of Videos for Smart Surveillance. *Appl. Sci.* **2021**, *11*, 3730. https://doi.org/10.3390/app11093730

Academic Editors: Jerry Chun-Wei Lin, Stefania Tomasiello and Gautam Srivastava

Received: 9 March 2021
Accepted: 9 April 2021
Published: 21 April 2021

Abstract: After the September 11 attacks, security and surveillance measures have changed across the globe. Now, surveillance cameras are installed almost everywhere to monitor video footage. Though quite handy, these cameras produce videos in a massive size and volume. The major challenge faced by security agencies is the effort of analyzing the surveillance video data collected and generated daily. Problems related to these videos are twofold: (1) understanding the contents of video streams, and (2) conversion of the video contents to condensed formats, such as textual interpretations and summaries, to save storage space. In this paper, we have proposed a video description framework on a surveillance dataset. This framework is based on the multitask learning of high-level features (HLFs) using a convolutional neural network (CNN) and natural language generation (NLG) through bidirectional recurrent networks. For each specific task, a parallel pipeline is derived from the base visual geometry group (VGG)-16 model. Tasks include scene recognition, action recognition, object recognition and human face specific feature recognition. Experimental results on the TRECViD, UET Video Surveillance (UETVS) and AGRIINTRUSION datasets depict that the model outperforms state-of-the-art methods by a METEOR (Metric for Evaluation of Translation with Explicit ORdering) score of 33.9%, 34.3%, and 31.2%, respectively. Our results show that our framework has distinct advantages over traditional rule-based models for the recognition and generation of natural language descriptions.

Keywords: CNN; multitask feature learning; bidirectional long short-term memory (LSTM); TRECVid 2007/2008; video captioning; smart surveillance; agriculture; intrusions

1. Introduction

There is an exponential increase in digital multimedia, resulting in the generation of enormous amounts of video data. This can be used to understand videos that have inspired development for a broad range of applications. The growing rate of multimedia content uploaded on the Internet involves automatic interpretation and description of the videos for the retrieval of important information. This can also be useful in surveillance, security, human–computer interaction, robotic intelligence and even helps visually impaired people. Among these applications, an automatic description of videos in a natural language is gaining interest, where we give a video to the deep learning framework that converts it into one or multiple sentences.

The problem of automatically describing videos has been explored for several years. The first rule-based method [1] described human actions and activities with natural language in a restricted setting. The semantic features obtained from the video were linked with syntactic constituents, such as the subject, verb or objects, and then interpreted into natural language sentences. A sentence is formed by filling the predetermined templates with part-of-speech tags. The major problem faced at that time was bridging the semantic gap when converting videos into text. There are several papers [2–4] that applied relatively similar rule-based systems on different datasets that contained large instances of objects in diverse situations. Later, more complex rules were applied in [5] that contained a relatively large vocabulary to generate sentences. These approaches require monotonous work when the data is huge. To eliminate this problem, statistical models [6,7] have been used, which can train even larger datasets that contain many lexical entries and many hours of videos. However, the results of these approaches were lacking with large datasets like Microsoft Common Objects in Context (MS COCO) [8].

Lately, a lot of research was carried out for applying deep learning methods to video captioning. Encouraging results have been attained by using a convolutional neural network (CNN) or a recurrent neural network (RNN) for labeling [9], image captioning [10] and video descriptions. These approaches have been applied to short video segments that contain action, anomaly detection or scene surroundings.

For each high-level feature (HLF), the state-of-the-art methods used different methods of feature extraction, such as for human action recognition (star skeleton and a hidden Markov model), object detection (Haar features) and age (facial features). This might result in the feature extraction of narrow domains, rigid template-based sentences and missing information, thus making it difficult to produce correct results for large data.

Our proposed deep learning model is a single carefully designed and trained network that can extract various HLFs simultaneously with language model incorporation. The model focuses on better visual information extraction from the frames, which is reflected in the results. We have captured the power of semantic feature extraction with a relatively simple deep learning architecture. This has been achieved by fine-tuning a multitask CNN, which can learn dense features for a scene, human, dress, object and action. In terms of computational complexity, the proposed model is quite inexpensive for the prediction of unseen data. This paper suggests a deep neural network with multitask feature learning to interpret videos into natural language descriptions. The key contributions are as follows:

- A multitask deep learning video description framework has been presented that uses learning to extract robust information, which describes the visual scene, persons, objects and their interaction;
- The proposed framework extracts multiple high-level features (HLFs) from videos compared with traditional approaches;
- We have provided a detailed evaluation of this framework and compared video descriptions with its preceding bottom-up approach, presented by [11];
- We have also gathered the UET (University of Engineering & Technology) Video Surveillance (UETVS) dataset [12] and annotated it with a text description;
- Evaluation of this framework is conducted on three datasets: the 2007/2008 TREC video benchmark [13], UET Video Surveillance (UETVS) and agriculture surveillance datasets.

The rest of the paper is arranged into the subsequent sections. Section 2 details the literature survey for the video captioning and description problem. Section 3 provides a detailed explanation of the proposed multitask natural language description framework. Section 4 discusses the TRECVid dataset, self-generated agriculture and the video surveillance datasets used for the experiment and its setup. Sections 5 and 6 present the results and discuss the evaluation scores, respectively.

2. Literature Survey

The prior video description approaches were established on finding subject–verb–object (SVO) triplets from the video frames, and captions were produced through a language model which was built on predefined sentence templates. This requires training multiple classifiers that could identify anomalies, activities, human beings and their properties, such as emotion or gender, objects, actions and scene settings. These extracted features are combined with a language model to generate sentences from templates. In [14], the authors extracted SVO triplets and learned a semantic structure on subjects, objects and verbs by using a multichannel SVM (Support Vector Machine) to predict these values. Once the triplet values are found, a sentence is formed using templates. A factor graph model was proposed in [15] that integrated visual detection with language statistics to learn the subject–verb–object–place (SVOP) tuple for a video. The sentences are generated using a template-based approach. Although the problem in both cases is simplified, this still entails choosing suitable objects and actions. Additionally, the sentences generated from this approach are not flexible enough to generalize unseen, erroneous and misidentified data and cannot satisfy the richness of natural language.

Based on the reasons above, research on image or video description generation has shifted to deep learning approaches. The simple plan for description generation is to use a CNN for encoding visual information and an RNN for decoding natural language sentences. CNNs are like a simple neural network, but the architecture makes a clear supposition that the inputs are images, which allows certain properties to be encoded. These networks are mostly used in image and video recognition systems. RNNs are networks that have looping capability and grant information to prevail. These networks are extremely popular for multiple problems such as speech recognition, modeling languages, image or video captioning, translation and description.

The hierarchical recurrent neural network (hRNN) [16] is the most credited state-of-the-art deep learning technique for video captioning and description. This network can generate sentences (multiple sentences) or a paragraph for a lengthy video which contains multiple scenarios or scenes. These types of videos cannot be depicted in a single sentence, but rather give a dull description. The idea of this framework is to utilize time-based dependency among sentences in a paragraph. While generating the next sentence, the semantic context of the previous sentence is put into use.

RNNs keep information in memory over a specific time period. However, it is hard to resolve problems that need long temporal dependencies, such as understanding the present frame of a video that requires information from the previous frame or frames. Long short-term memory (LSTM) networks are a specific kind of RNN that works for several tasks and performs better. An LSTM unit comprises a memory cell that can withstand data for a lengthier time period.

Rapid development has been made in previous years to learn image features, and numerous pretrained models are offered. However, these deep features cannot be used for videos due to the absence of temporal information. The CNN and LSTM were integrated in [17] to learn spatial (space) and temporal (time) information from videos. Features are mined from a 2D CNN and input into an LSTM network to translate the time-based information of the videos. There are other variations of LSTM, such as BiLSTM (Bidirectional Long short-term memory) [18]. The authors proposed a bidirectional LSTM with a soft attention mechanism. This conserve the global, temporal and visual information of an input video, and the attention process helps in identifying the most important words in the content.

The encoder-decoder framework [19] for generating video captions has been proposed. The encoder extract features and encodes the input into vectors. Two models have been used to extract features from videos. In the first approach, a video is viewed as a series of 2D images, from which frame-level features are drawn out using the GoogleNet (CNN) model that is trained on ImageNet. In the second approach, a video is considered as 3D data which are a series of video segments, and it treats variation with respect to time for feature extraction. Here, segment-level features are mined using dense trajectories and a 3D CNN (C3D) model pretrained on the Sports 1M dataset. The pooling method used for these two approaches merges various frame or segment features into one video feature vector. The decoder generates captions using these feature vectors. The LSTM network based on the language model is trained with some tweaks. This architecture can generate multiple captions which pick the best candidate from the group. The authors concluded that this model is appropriate for the MSR-VTT (MSR-Video to Text) dataset, which consists of various types of videos. The results demonstrate that the captions generated from this model closely relate to human evaluation.

Real-world images cannot be explained by just one label due to rich information such as the attributes, scenes, objects and parts of objects that are contained in it. The proposed combined CNN-RNN [20] approach for multi-label image classification learns not only semantic but also co-occurrence dependency. This approach was tested on three benchmark multi-label datasets—MS COCO, NUS-WIDE and VOC PASCAL 2007—with promising results. Many classifiers are unable to perform when the training set is small. Building large, labeled data requires a lot of time and effort. Generative models can produce realistic data samples to bridge this gap. A new model, the category sentence generative adversarial network (CS-GAN) was proposed in [21]. This model combines reinforcement learning, a GAN and an RNN to learn and reproduce the sentence structure.

CNN models tend to lose spatial information and might misclassify objects in different orientations and proportions. To solve this problem, capsule networks are introduced which not only work for images, but also for natural language processing (NLP) problems. A capsule network to solve the multi-label text classification problem was proposed in [22]. The experiments showed promising results using low-capacity resources.

In this work, we have proposed a CNN- and LSTM-based multi-line video caption generation system. However, we have focused on learning task-specific features for the scene, human (age, gender and emotion), object, dress and action descriptions. We have tested this on TRECVid, an agriculture and university surveillance dataset. Such a framework is helpful for many applications such as surveillance, intrusion detection and alert systems, person re-identification and video-to-text summarization.

3. Proposed Methodology

The block diagram of the multitask deep learning framework is shown in Figure 1. It comprised two primary parts: a convolutional neural network (CNN)-based feature learning stage and a long short-term memory (LSTM) network stage which was jointly tuned and trained. The implementation information about each stage is provided in the following subsections.

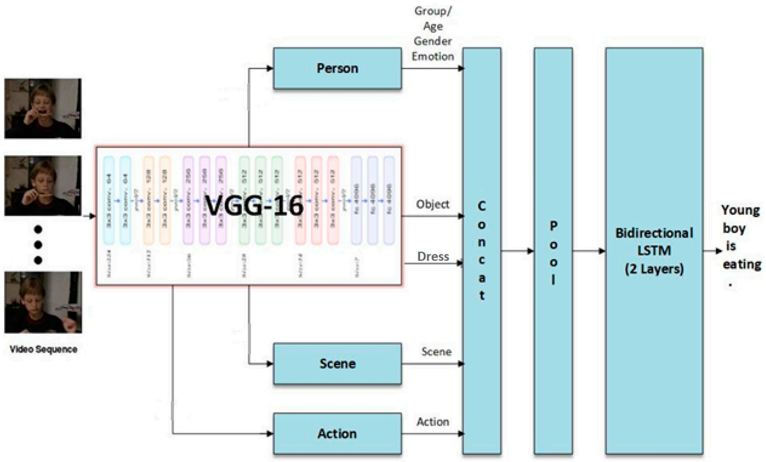


Figure 1. Block diagram of the multitask learning framework.

3.1. Multitask Learning for HLF Extraction

The available literature provides evidence of the power of CNNs for automatic feature extraction in object classification and detection tasks. Compared with traditional feature learning, CNNs learn the filters and representations useful for distinction in a hierarchical manner while training. This characteristic is inspired by the visual cortex of human beings. These networks extract features by moving from simple to complex structures. By sliding and convolving, it incorporates local features, and pooling cares for learning global features. This in turn gives the benefit of spatial invariance property and parameter sharing, which would otherwise cost huge memory. The trained CNN models are reproducible as well as reusable. This is especially helpful in cases with small amounts of data. Among the pretrained models, the Imagenet model [23] formed the basis of most of the recent frameworks.

Similarly, our work also adopts visual geometry group (VGG)-16 [24] and branches out task-specific features. This network is sequential but follows a homogeneous filter structure throughout its depth. The 16-layer architecture includes 13 convolutional layers with 3 fully connected (FC) layers on top. A total of five tasks are learned jointly, including action, person-related, objects, dress and scene information. For this task, each video is annotated in a multi-labeled fashion as well. For object recognition, the pure VGG-16 model is applied, and the last FC layer is fine-tuned for the purpose. All the FC layers are fine-tuned for scene recognition tasks. For person or face close-up features, a task-specific branch is taken out after the second pool layer with a map size of 56 × 56. This branch is further multi-labeled for the gender, age and emotion of the person present in the video clip.

Multitask learning within the branch is highly recommended for large-scale data. The first step of the face task is based on a faster RCNN (Region Based Convolutional Neural Network) [25]. If multiple regions are detected for the face, then the video gets labeled as a group of people or, to some extent, a count of people (e.g., three persons or four persons). If a close-up scene is detected, then the above-mentioned features are extracted through training the network. Action recognition is also performed by extracting out parallel branches after the second-last pooling layer. All the obtained features are concatenated and pooled to an embedding vector size of 512 before passing into a language model.

3.2. Description Generation Using Bidirectional LSTM

A recurrent neural network (RNN) and LSTM [26] are the basic sequence modeling blocks in deep learning. However, many variants derived from these basic versions have been used for various tasks, and it is still an open area of research to work out newer and better architectures. The main task of these models is to predict the next member of the incoming sequence data. Due to improved gradient flow properties, LSTM models are preferred over vanilla RNNs, which have exploding and vanishing gradient problems. The main difference in the structure is the presence of an internal cell unit (ct) and a gated structure.

The role of the gates is to maintain and modify the state of the cell and hidden layer output (h_t) according to the previous hidden state and present input. The values of the cell state are invisible to the other network and remain intact unless the forget gate signal applies.

For video description generation, we have used the bidirectional LSTM architecture. Bidirectional architecture processes the information forward and backward in time and hence can capture better contextual information from the data. We have used two-layered LSTM, each layer having 512 hidden nodes. The embedding vector from the CNN is passed to the first layer together with the attention vector [27], which gives the grouping over the spatial location to concentrate the segment of an image at a certain time. A set of vocabulary has been created by using human handwritten annotations. Every video has been interpreted by 40 human annotators with a varied level of expertise, hence producing a variety of words. A word which occurred less than five times was appointed an unknown <UNK> token. A total of 22,500 sentences were processed. The joint training ensured filling of the gap between visual clues and language.

4. Experimental Setup

The proposed architecture was assessed with a dataset which was manually devised from the 2007/2008 TREC video benchmark for video description. It consisted of 140 video segments used in the experiments. Each video segment duration was between 10 and 30 s with a sole camera shot. Videos were annotated by 40 human beings, and each video was described in two to seven sentences, which were stated as human annotations. They were classified into seven categories:

- Activity: Humans carrying out some tasks, such as walking or sitting;
- Close up: Human face with the front view that displays expressions or sentiments (e.g., happy or sad);
- News: Reporter, anchor or weatherboards shown in a scene or setting;
- Meeting: Assembly of humans showing an interaction with one another, and the existence of objects normally seen in meeting scenarios (e.g., chairs, tables, mics and projectors);
- Groups: Several humans in any scene or setting;
- Traffic: Automobiles and traffic sign;
- Scene Category: Scene locations more prominent than human actions (e.g., indoor, outdoor, parks and kitchens).

The second dataset was on UET Video Surveillance (UETVS), which consisted of 1200 surveillance videos and corresponding multi-line textual descriptions. The videos were gathered from four locations (Boys Student Service Center (BSSC), Girls Student Service Center (GSSC), Al-Khawarizmi Institute of Computer Science and the UET Bus Stand). CCTV (Closed-Circuit Television) security cameras were installed in these locations. Each location consisted of 300 videos. The span of each video lied between 7 and 10 s. The frames per second (fps) rate of each video clip was 25. These videos were described by professional English writers, with descriptions ranging between 3 and 6 sentences.

The third dataset was self-collected (AGRIINTRUSION), which contained around 100 videos. These videos were taken from YouTube. The videos contained the type of intrusions that could happen on agricultural land, such as animals, birds, vehicles or humans. These images were annotated and described with multiple sentences.

The video description is the explanation of nonverbal elements in a video that include facial expressions, scenery, actors' clothing or body language. It provides a description of visual elements that are important to perceive the plot, which will specifically be of service to visually impaired people. The annotation process was done on three levels: students, teachers and professional experts. Every video was annotated by three different people based on the expertise level defined above. These HLF and video descriptions were cross-checked, and discrepancies were corrected by a data annotation specialist.

The video description experiments were performed on a GeForce GTX 1080 Ti which had 11 GB of GPU memory. The multitask framework was formed using PyTorch [28]. The CNN and LSTM models were trained and optimized together using a stochastic gradient descent optimizer with a batch size of 32. We used cross-entropy loss, which converged with the annealing learning rate scheme. The TRECViD dataset took around 62 h (~2.5 days) of training time, and UETVS took around 120 h (~5 days) of training. The AGRIINTRUSION dataset took around 45 h of training time. All video frames were regenerated to 256 × 256 dimensions before being input into the CNN. Each video in the dataset was fed to the network at 4 frames per second. The data distribution consisted of 75% (training) and 25% (testing).

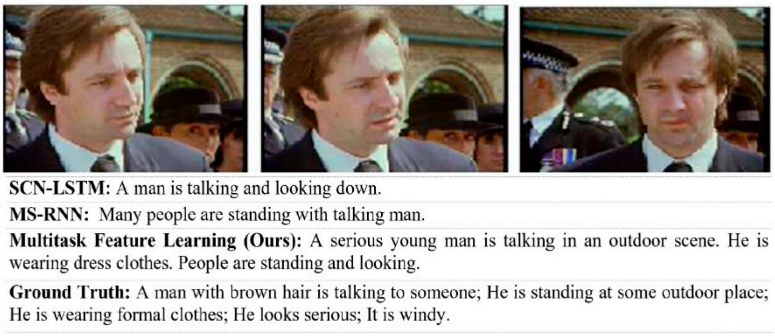
5. Experiments and Results

This section details the results of the TRECViD, UETVS and AGRIINTRUSION datasets. We compare the results of this multitask learning-based framework with two baseline models—an LSTM semantic compositional network (SCN-LSTM) and multimofal stochastic RNN (MS-RNN)—and video descriptions using deep neural networks.

A long short-term memory semantic compositional network (SCN-LSTM) [29] was originally proposed for image description and later extended for videos. The spatial and temporal features of the video clips were characterized using 2D and 3D dimensional CNN visual features for every video frame in an input clip. Mean pooling was used for the features and then put together. The semantic concept was fed into the network, which was extracted using a semantic detector. This produced a detailed description of a video.

The multimodal stochastic (MS) RNN [30] model generated different video descriptions using multimodal LSTM. This mined visual and text features using backward stochastic LSTM.

Figure 2 shows the results of two randomly selected videos from the TRECViD dataset. It shows the video descriptions generated by the multitask deep learning framework, SCN-LSTM, MS-RNN and human annotations.



SCN-LSTM: A man is talking and looking down.

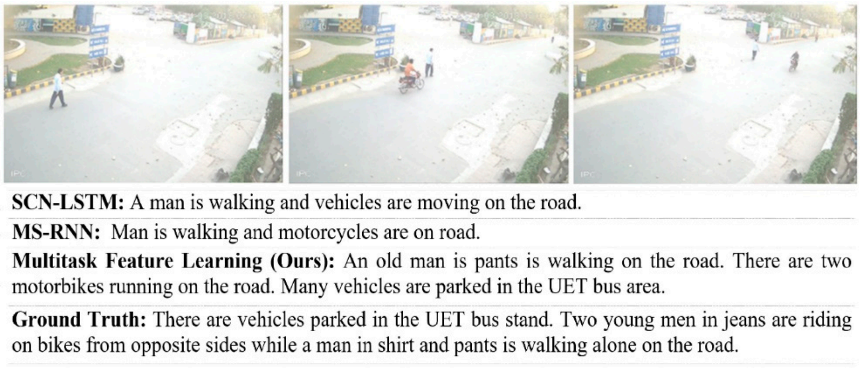
MS-RNN: Many people are standing with talking man.

Multitask Feature Learning (Ours): A serious young man is talking in an outdoor scene. He is wearing dress clothes. People are standing and looking.

Ground Truth: A man with brown hair is talking to someone; He is standing at some outdoor place; He is wearing formal clothes; He looks serious; It is windy.

Figure 2. Face close-up video.

Similarly, the results from the UETVS dataset are shown in Figure 3. The video descriptions generated from our multitask feature learning (MFL) framework showed a more accurate and improved version in comparison with the other approaches.



SCN-LSTM: A man is walking and vehicles are moving on the road.

MS-RNN: Man is walking and motorcycles are on road.

Multitask Feature Learning (Ours): An old man is pants is walking on the road. There are two motorbikes running on the road. Many vehicles are parked in the UET bus area.

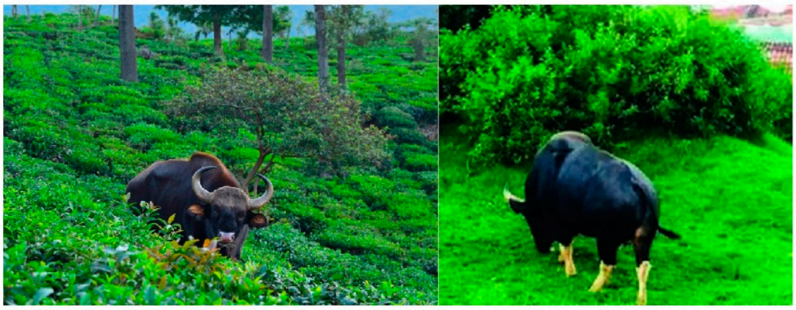
Ground Truth: There are vehicles parked in the UET bus stand. Two young men in jeans are riding on bikes from opposite sides while a man in shirt and pants is walking alone on the road.

Figure 3. UET parking area.

The results from the AGRIINTRUSION dataset are shown in Figure 4. Again, the video descriptions generated from our framework were more detailed and better in comparison with the other approaches.

5.1. Face Close-Up Scene

Our MDL (Multitask Deep Learning) framework accurately recognized the human face, emotion, dress, gender and surrounding objects. Hand annotations were more descriptive, such as the identity of the person (e.g., policeman), precise details of the clothing, such as a woman's hat and formal suit, and more detailed settings were mentioned.



SCN-LSTM: A cow is walking on the plants.

MS-RNN: One cow is moving on the ground.

MultiTask Feature Learning (Ours): A bison is eating in the fields.

Ground Truth: A bison is walking through the lush green fields. It is eating the crops.

Figure 4. Intrusion video.

5.2. UETVS

The video description produced for the parking scene video contained the correct identification of humans, objects, dress and scenic location. The human annotation or the ground truth had more detail (e.g., human profile, such as a man wearing shalwar kameez or jeans and a shirt). This surveillance video was taken from a CCTV camera that was mounted at a high place, where the human profiles and actions could be identified through the eager human eye. Still, our MDL framework produced a coherent and consistent video description based on the high-level features.

5.3. AGRIINTRUSION

The video description produced for the agriculture intrusion contained the correct identification of objects and the scene's location. The human annotation, or the ground truth, contained the object that was present in the video frames (e.g., a bison is in the green fields). Our MDL framework still produced a coherent and consistent video description based on the high-level features.

6. Evaluation

The performance of our deep learning framework can be investigated using two standard evaluation metrics: METEOR [31] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [32]. METEOR is a measure used for evaluating many features, such as synonym matching and stemming. This measure uses the harmonic mean for recall and precision. The range of this measure is between 0 and 1. If there was more than one reference translation, then the METEOR results were calculated for every single translation, and the top score was chosen.

Table 1 shows the METEOR scores of the multitask feature learning framework for video descriptions with TRECViD, UETVS and AGRIINTRUSION. It can be seen in the table that our model outperformed the state-of-the-art methods. This is because this framework has specialized training of data for interpreting scenes and object settings.

Table 1. METEOR score (%).

Methods	TRECViD	UETVS	AGRIINTRUSION
SCN-LSTM	26.4	32.5	27.5
MS-RNN	31.5	33.1	30.6
MFL (Ours)	33.9	34.3	31.2

The ROUGE metric compared words between the machine-generated (deep neural network) and human written annotations. It is commonly used for evaluating machine summarization tasks. There are many variations of ROUGE, such as N-gram (ROUGE-N), skip bigram co-occurrence with unigrams (ROUGE-S) and longest subsequence (ROUGE-L). Table 2 shows the ROUGE scores between the multitask feature learning-generated video descriptions and reference hand annotations done by humans. ROUGE 1–3 shows an N-gram comparison between the reference and framework descriptions.

Table 2. ROUGE scores for the multitask deep learning framework for TRECViD.

	Activity	Close-Up	Indoor/Outdoor	Groups	Meeting	News	Traffic
ROUGE 1	0.6721	0.8115	0.8249	0.7661	0.7922	0.7515	0.7787
ROUGE 2	0.6366	0.7792	0.7585	0.7020	0.7433	0.7073	0.7302
ROUGE 3	0.5603	0.7067	0.6898	0.6422	0.6654	0.6653	0.6595
ROUGE-L	0.5698	0.6200	0.8141	0.7614	0.8015	0.7606	0.7361
ROUGE-W	0.5510	0.6278	0.7909	0.7212	0.7673	0.7232	0.7056
ROUGE-S	0.5192	0.5826	0.7679	0.6145	0.7505	0.6804	0.6695
ROUGE-SU	0.5552	0.6022	0.8221	0.7330	0.7111	0.7431	0.7131

Table 3 shows the ROUGE scores between the multitask feature learning-generated descriptions and reference hand annotations done by the humans for the UETVS dataset. The rouge scores were calculated for the four locations where the CCTV cameras were installed. This dataset is one of a kind which, to the best of our knowledge, did not exist previously. As there is no gold standard dataset that deals with the textual description of surveillance videos, we were not able to provide any comparison.

Table 3. ROUGE scores for the multitask deep learning framework for the UETVS dataset.

	GSSC	BSSC	KICS	UET Parking
ROUGE 1	0.7521	0.7715	0.7855	0.7916
ROUGE 2	0.6933	0.7392	0.7321	0.7620
ROUGE 3	0.6203	0.7027	0.6898	0.7102

Table 4 shows the ROUGE score between the multitask feature learning-generated video description and human annotations for the AGRIINTRUSION dataset. The scores show that there was significant similarity between the two.

Table 4. ROUGE scores for the multitask DL framework for the AGRIINTRUSION dataset.

	AGRIINTRUSION
ROUGE 1	0.6506
ROUGE 2	0.5660
ROUGE 3	0.5022

In the TRECViD dataset, the scene-based categories of indoor or outdoor, meeting, groups and traffic had the highest scores due to the superior learning capability of the VGG network for scene and object settings. The activity category also saw a gain in performance compared with our previous experiment [33] of 12%. However, it still required incorporating carefully handled action recognition techniques to outperform the state of the art for action. Scores for the close-up features were comparable to our former experiment [33] but high compared with traditional hand-engineered techniques. However, multitask learning with basic fine-tuning had a positive impact on the overall results for video description generation.

It was noted that the age and emotion information in our MDL framework was missing due to the mounted surveillance camera in the UETVS dataset, which captured scenes from afar. The proximity and video quality were not high, which led to minute details being missed in our video description feature extraction.

For the AGRIINTRUSION dataset, our MDL framework was not able to provide results for objects that were not part of the training set. The agriculture dataset needed objects that were domain-specific, such as tractors, crop types and lotuses.

7. Conclusions

In this paper, we have developed a multitask learning-based deep neural network framework. The CNN was trained on benchmark datasets to extract visual features (human attributes such as age, gender, dress, emotion, objects and their interactions) and produce labeled information. This is passed on to the sentence generation model (LSTM) that learns the structure of the sentence with their labels.

Unlike machines, humans tend to write more rich descriptions, catching every nitty-gritty detail in the video. Our resulting model produced condensed, scalable and useful descriptions with reasonable results for the three datasets (TRECViD, UETVS and AGRIINTRUSION). Using the standard metrics of machine translation and summarization demonstrated that the video descriptions generated by our deep learning framework were better than the state-of-the-art approaches, which are consistent with the results reported above.

8. Future Work

The video description problem is not completely solved yet. Although our proposed model has shown promising results in comparison with the state-of-the-art approaches, it still is far from how humans caption data. The proposed base model (CNN-LSTM) can be improved by implementing the latest techniques, such as reinforcement learning, generative adversarial networks (GANs) or capsule networks. The classifiers require a large set of training data to produce accurate results. For this, generative models such as GANs can also be implemented to bridge this gap. In addition, CNN models tend to classify items incorrectly if the input image is in a different orientation or percentage. To solve this problem, capsule networks can used to provide accurate classification of high-level features. Dense captioning is one direction which shows a natural mapping of temporal events, the same type of structuring as the human brain. The feature learning process can be improved by adding audio data, which can help increase activity detection.

Author Contributions: Conceptualization, A.D., and M.U.G.K.; methodology, Y.D.A.-O., Z.-u.R. and A.D.; project administration, Y.N., M.U.G.K. and A.-u.R.; validation, Z.-u.R. and Y.N.; visualization, A.D.; writing—original draft, A.D.; writing—review and editing, M.U.G.K., Y.N. and Z.-u.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Soonchunhyang University Research Fund.

Acknowledgments: The authors would like to thank the Intelligent Criminology Lab in the National Center of Artificial Intelligence at the Al-Khawarizimi Institute of Computer Science, UET, Lahore for providing the research platform and technical support.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Kojima, A.; Tamura, T.; Fukunaga, K. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vision* **2002**, *50*, 171–184. [CrossRef]
2. Lee, M.W.; Hakeem, A.; Hearing, N.; Zhu, S.C. Save: A framework for semantic annotation of visual events. In Proceedings of the Computer Vision and Pattern Recognition Workshops CVPRW'08, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
3. Khan, M.U.G.; Zhang, L.; Gotoh, Y. Towards coherent natural language description of video streams. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 664–671.
4. Hanckmann, P.; Schutte, K.; Burghouts, G.J. Automated textual descriptions for a wide range of video events with 48 human actions. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 372–380.

5. Barbu, A.; Bridge, A.; Burchill, Z.; Coroian, D.; Dickinson, S.; Fidler, S.; Michaux, A.; Mussman, S.; Narayanaswamy, S.; Salvi, D.; et al. Video in sentences out. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, 15–17 August 2012; pp. 102–112.
6. Rohrbach, A.; Rohrbach, M.; Qiu, W.; Friedrich, A.; Pinkal, M.; Schiele, B. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2014; pp. 184–195.
7. Chen, D.L.; Dolan, W.B. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 190–200.
8. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
9. Lin, J.C.W.; Shao, Y.; Djenouri, Y.; Yun, U. ASRNN: A recurrent neural network with an attention model for sequence labeling. *Knowl.-Based Syst.* **2021**, *212*, 106548. [CrossRef]
10. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
11. Khan, M.U.G.; Gotoh, Y. Describing video contents in natural language. In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, Avignon, France, 23–27 April 2012; pp. 27–35.
12. Dilawari, A.; Khan, M.U.G. UET Video Surveillance (UETVS) Dataset. Unpublished work. 2020.
13. Khan, M.U.G.; Nawab, R.M.A.; Gotoh, Y. Natural language descriptions of visual scenes: Corpus generation and analysis. In Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), Avignon, France, 23–27 April 2012; pp. 38–47.
14. Guadarrama, S.; Krishnamoorthy, N.; Malkarnenkar, G.; Venugopalan, S.; Mooney, R.; Darrell, T.; Saenko, K. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 2712–2719.
15. Thomason, J.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Mooney, R. Integrating language and vision to generate natural language descriptions of videos in the wild. In Proceedings of the COLING, 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014; pp. 1218–1227.
16. Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; Xu, W. Video paragraph captioning using hierarchical recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4584–4593.
17. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
18. Bin, Y.; Yang, Y.; Shen, F.; Xie, N.; Shen, H.T.; Li, X. Describing video with attention-based bidirectional LSTM. *IEEE Trans. Cybern.* **2018**, *49*, 2631–2641. [CrossRef] [PubMed]
19. Shetty, R.; Laaksonen, J. Frame-and segment-level features and candidate pool evaluation for video caption generation. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1073–1076.
20. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2285–2294.
21. Li, Y.; Pan, Q.; Wang, S.; Yang, T.; Cambria, E. A generative model for category text generation. *Inf. Sci.* **2018**, *450*, 301–315. [CrossRef]
22. Zhao, W.; Peng, H.; Eger, S.; Cambria, E.; Yang, M. Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1549–1559.
23. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* **2015**, *115*, 211–252. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Sun, X.; Wu, P.; Hoi, S.C. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing* **2018**, *299*, 42–50. [CrossRef]
26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
28. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G. Pytorch: Tensors and Dynamic Neural Networks in Python with Strong GPU Acceleration. 2017. Available online: https://github.com/pytorch/pytorch (accessed on 1 April 2021).
29. Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; Deng, L. Semantic compositional networks for visual captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5630–5639.
30. Song, J.; Guo, Y.; Gao, L.; Li, X.; Hanjalic, A.; Shen, H.T. From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *30*, 3047–3058. [CrossRef] [PubMed]

31. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 1 June 2014; pp. 376–380.
32. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the ACL-04 Workshop on Text Summarization Branches out, Barcelona, Spain, 25–26 July 2004; Volume 8.
33. Dilawari, A.; Khan, M.U.G.; Farooq, A.; Rehman, Z.U.; Rho, S.; Mehmood, I. Natural language description of video streams using task-specific feature encoding. *IEEE Access* **2018**, *6*, 16639–16645. [CrossRef]