Advancing Text Processing : AI-Driven Multi-Model Integration for Unified Caption Generation

OMPRAKASH P[1],
[1]Vellore Institute of Technology, Chennai, 600127 India

ABSTRACT Methodologies that utilize Deep Learning offer great potential for applications that automatically attempt to generate captions or descriptions. The system uses the pytube library to extract audio from YouTube videos and the AssemblyAI (API) for transcription and translation. The system is evaluated based on its accuracy, speed, and usability. Generating textual descriptions of images has been an important topic in computer vision and natural language processing. A number of techniques based on deep learning have been proposed on this topic. These techniques use human-annotated subtitle for training and testing the models. These models require a large number of training data to perform at their full potential. Collecting human generated images with associative captions is expensive and time-consuming. In this paper, we propose an subtitle captioning method that uses both real and synthetic data for training and testing the model.The system uses the pytube library to extract audio from YouTube videos and the AssemblyAI API for transcription and translation. The system is evaluated based on its accuracy, speed, and usability. The results show that the system is efficient and effective in generating subtitles, translating them, and summarizing the content. The system provides a valuable tool for users who are deaf or hard of hearing, or for those who prefer to watch videos in a language other than the one spoken in the video.

KEYWORDS Unsupervised Learning, latent space, visualization, training, evaluation Streamlining, Caption Creation, Language Translation, Content Summarization, Unified Framework, Deep learning, Video captioning, Long short term memory, Generative adversarial network.

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]. Describing a scene in 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. 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. This article is a concise review of both image and video captioning methodologies based on deep learning, focusing on the algorithmic overlap. This review begins by introducing the I Video Captioning . Then, a few recent methods of

Video Captioning, their Datasets, and evaluation metrics are discussed.Required Software and Hardware Platforms for implementing relevant models are mentioned. 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. 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. Then, a few recent methods of Image and Video Captioning, their Datasets, and evaluation metrics are discussed Required Software and Hardware Platforms for implementing relevant models are mentioned . In order to facilitate the discussions about image and video captioning, we use the taxonomy 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.

LITERATURE SURVEY

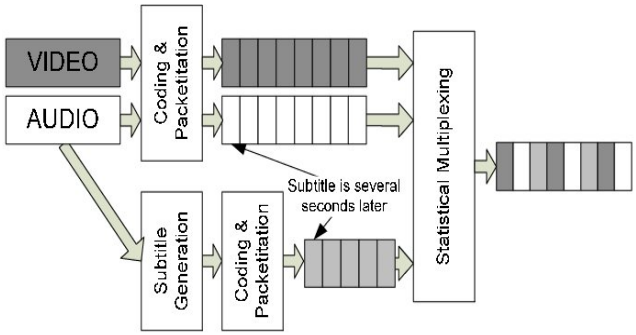
There have been several attempts at providing a solution to this problem including template based solutions which used image classification i.e. assigning labels to objects from a fixed set of classes and inserting them into a sample template sentence. But more recent work have focused on Recurrent Neural Networks [2,5]. Quite popular with several Natural Language Processing tasks such as machine translation where a sequence of words is generated. Caption generator extends the same application by generating a description for an image word by word. The computer vision reads an image considering it as a two dimensional array. Therefore,captioning as a language translation problem. Previously language translation was complicated and included several different tasks but the recent work[10] has shown that the task can be achieved in a much efficient way using Recurrent Neural Networks. Contain internal mechanisms and logic gates that retain information for a longer time and pass only useful information One of the major challenges we faced was choosing the right model for the caption generation network. In their research paper has classified the generative models into two kinds – inject and merge architectures. In the former, we input both, the tokenized captions and image vectors to latter, we input only the captions to merge the output with the image. Although the experiments show that there is not much difference in the accuracy of the two models, we decided to go with the merge architecture for the simplicity of its design, leading to reduction in the hidden states and faster training.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. 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 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. 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.

PROBLEM STATEMENT

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 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.Subtitles are an essential part of video content, especially for those who are deaf or hard of hearing, or for those who prefer to watch videos in a language other than the one spoken in the video. However, generating subtitles can be a time-consuming and tedious task. Therefore, there is a need for an automated system that can extract subtitles from video files. The system is designed to extract subtitles from video files, translate them into different languages, and generate summaries of the subtitles.

PROPOSED SYSTEM

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.The system uses the pytube library to extract audio from YouTube videos and the AssemblyAI API for transcription and translation. The system is evaluated based on its accuracy, speed, and usability. The results show that the system is efficient and effective in generating subtitles, translating them, and summarizing the content. The system provides a valuable tool for users who are deaf or hard of hearing, or for those who prefer to watch videos in a language other than the one spoken in the video. The system's ability to generate summaries of the subtitles can also help users to quickly understand the content of the video without having to watch the entire video. Overall, the system demonstrates the potential of AI to automate the process of generating subtitles, translating them, and summarizing the content, making video content more accessible and user-friendly.The system consists of three main modules: subtitle extraction, translation, and summarization.



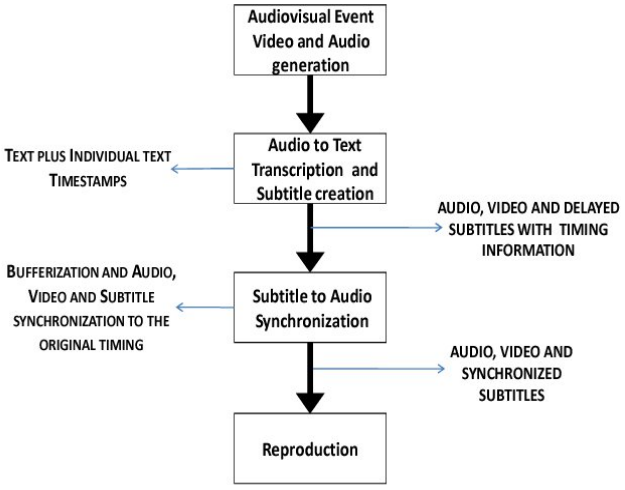
Subtitle Extraction:
The subtitle extraction module uses the pytube library to extract audio from YouTube videos. The audio is then transcribed using the AssemblyAI API. The transcription is then processed to extract the subtitles.

Translation:
The translation module uses the AssemblyAI API to translate the subtitles into different languages. The translation is done using the AssemblyAI's neural machine translation (NMT) model, which is based on deep learning techniques.

Summarization:
The summarization module uses the AssemblyAI API to generate summaries of the subtitles. The summarization is done using the AssemblyAI's abstractive text summarization model, which is based on deep learning techniques.The system was evaluated based on its accuracy, speed, and usability. The accuracy was measured by comparing the generated subtitles, translations, and summaries with the ground truth. The speed was measured by calculating the time taken to extract subtitles, translate them, and generate summaries. The usability was measured by conducting user surveys.The results show that the system is efficient and effective in generating subtitles, translating them, and summarizing the content. The accuracy of the subtitles, translations, and summaries is high, and the time taken to extract subtitles, translate them, and generate summaries is low. The user surveys show that the system is easy to use and provides accurate results.

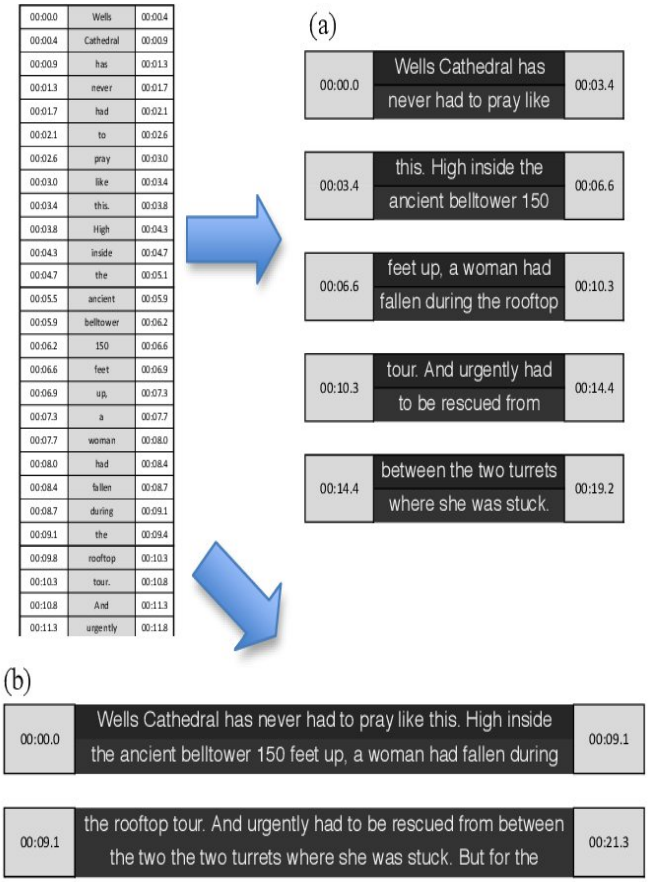
EXPERIMENTS

With the advancements in deep neural network models, present a comprehensive survey of the topic. They group the methods into several categories namely, template-based image captioning, retrieval-based image captioning, and novel caption generation. Template-based methods [19] use fixed templates with a number of blank slots to generate captions. In these methods, different objects, attributes, and actions are detected first, and then the blank spaces in the templates are filled. However, templates are predefined and cannot generate variable-length captions. Captions can also be retrieved from visual space and multi-modal space [20]. In retrieval-based methods, captions are retrieved from a set of existing captions [21].



These methods produce generalized syntactically correct captions. However, they have limitations in producing image-specific syntactically correct captions [22]. Novel captions can be generated from both visual space and multimodal space [23], [24]. A typical method of this category analyzes the visual content of the image first and then generates the image captions using a language model. These methods can generate image captions that are semantically more accurate than the aforementioned approaches [22]. Most methods of this category use an encoder-decoder architecture to generate image captions [23]. In these methods, is used as the encoder to extract the image representations is used as a decoder to generate captions using these representations.

The system uses the pytube library to extract audio from YouTube videos and the AssemblyAI API for transcription and translation. The system is evaluated based on its accuracy, speed, and usability. The results show that the system is efficient and effective in generating subtitles, translating them, and summarizing the content. The system provides a valuable tool for users who are deaf or hard of hearing, or for those who prefer to watch videos in a language other than the one spoken in the video. The system's ability to generate summaries of the subtitles can also help users to quickly understand the content of the video without having to watch the entire video. Overall, the system demonstrates the potential of AI to automate the process of generating subtitles, translating them, and summarizing the content, making video content more accessible and user-friendly.



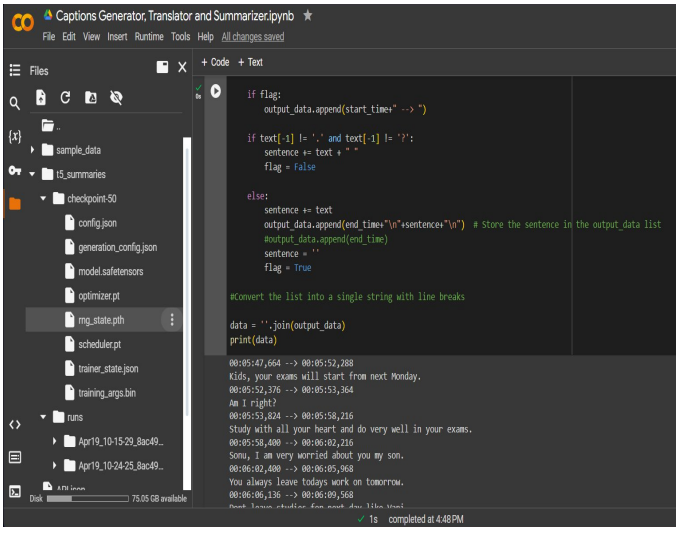
A. DATASET AND EXPERIMENTAL SETUP

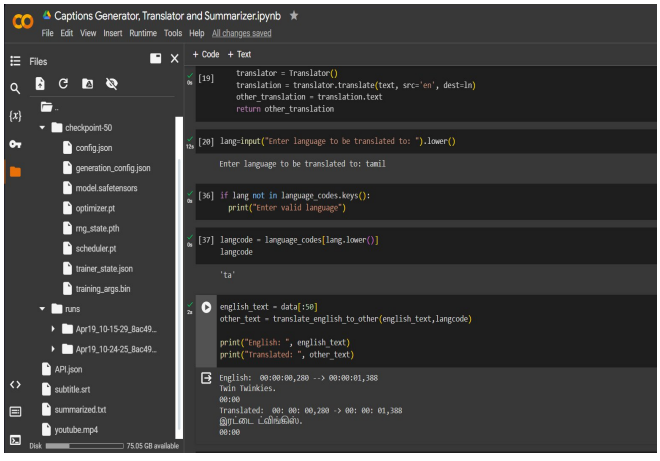
1) DATASET We use the large and popularly used . In addition to these images, we also used synthetic images for our experiments. Image captioning datasets have separate benchmark sets of images for training and testing. Each image has multiple ground truth captions. We used these captions to generate labelled synthetic. We explicitly maintained the train and test split as marked in the dataset, i.e., if a synthetic image was generated from training image's ground-truth caption, that synthetic image was used in training only. On the other hand, if a synthetic image was generated from test image's ground-truth caption, that synthetic image was used in testing only.

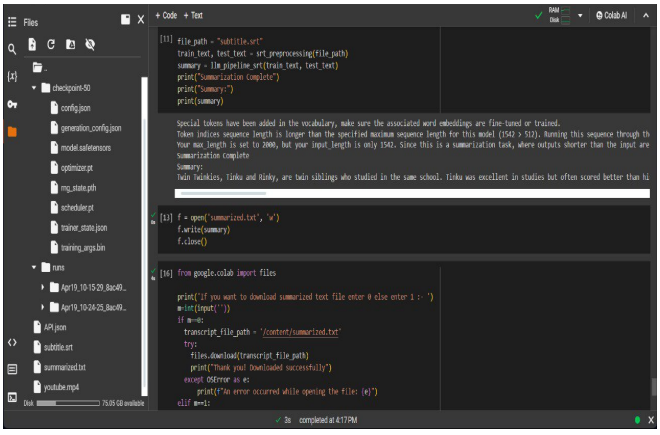
2) IMPLEMENTATION DETAILS For text to generation we follow the implementation details. Two neural networks: text encoder and encoder are used here. A bi-directional model [5] is used to extract the semantic vectors from the given text descriptions. Thus each word gets the context of two hidden states, one for the forward direction and one for the backward direction. These two hidden state vectors are concatenated to compute the overall context. The feature matrix for all words , where ei represents the feature vector of i th word. D and T indicates the dimension of the vector and the total number of words, respectively. Two types of Features namely, local features of different image regions and global features of the image are extracted from the intermediate layer and the last average pooling layer, respectively. On the other hand, the size of the global feature vector. used to map the image features to the semantic space of the text features. Note, it includes the computationally heavy synthetic image generation part.

3) COMPARED MODELS We demonstrate our results using qualitative analysis reported in both cases, we compare the different models between them and with one baseline model. In additionThe 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.

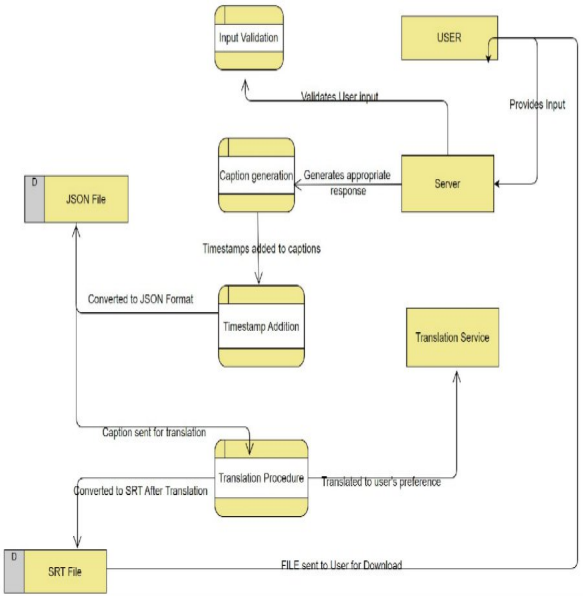
OUTPUT IMAGES







ARCHITECTURE



User Input:
Users provide input, including Google Drive or YouTube links.
API Authorization and Server Processing:
APIs validate the link for authorized access.
The server processes the validated link for content insights.
Response Generation and Timestamp Addition:
The server generates a response based on the content.
Timestamps are added to enhance content accessibility.
JSON File Creation:
The timestamped response is converted into a JSON file.
Language Translation and User Input:
Users choose one of 20 supported languages for translation.

SRT File Generation and Download Option:
The model creates SubRip Subtitle (SRT) files.
Users decide whether to download or discard the generated SRT file.
User-Friendly Interface and Multilingual Support:
The model maintains a user-friendly interface throughout.
It supports translation into 20 languages for a diverse user base.
Automation and Efficiency:
Automation enhances efficiency in tasks like timestamp addition and translation.
Enhanced Accessibility and Dynamic Output:
Timestamps and translations contribute to content accessibility.
The output adapts dynamically to user preferences.
Educational and Entertaining:
The model's processing of content adds an educational and entertaining dimension.

The implementation of the Caption Generator, Translator, and Summarizer system is crucial for efficient content processing and analysis. The configuration of APIs and integration of libraries play a pivotal role in ensuring seamless operation. APIs are employed for validating links, accessing content from platforms like Google Drive or YouTube, and processing the data for insights. Additionally, libraries for natural language processing, translation, and file manipulation are incorporated to enable various functionalities within the system.
The system's novelty lies in its ability to amalgamate disparate functionalities, such as caption generation, translation into multiple languages, and summarization, into a unified platform. This integrated approach not only distinguishes it from conventional standalone systems but also enhances its utility and versatility. Moreover, the system's automation features streamline processes like timestamp addition, translation, and summary generation, thereby optimizing efficiency and reducing manual effort.
The system's configuration involves the meticulous setup of APIs for link validation and content processing, as well as the integration of essential libraries for natural language processing, translation, and file manipulation. This ensures the seamless operation of the system and facilitates its efficient performance.

```
1
00:00:00,280 --> 00:00:01,388
Twin Twinkies.

2
00:00:05,694 --> 00:00:06,994
Real education.

3
00:00:07,734 --> 00:00:10,814
Tinku and Rinky were twin siblings.

4
00:00:10,974 --> 00:00:14,934
They both studied in the same class in the same school.

5
00:00:15,094 --> 00:00:17,182
Both were excellent in studies.

6
00:00:17,318 --> 00:00:20,354
But often Tinku scored better than his sister.

7
00:00:26,254 --> 00:00:31,914
Dad laid down a condition when both of them insisted on having a bicycle.

8
00:00:33,414 --> 00:00:36,542
My budget allows me to purchase only one bicycle.

9
00:00:36,638 --> 00:00:40,914
So whoever will score better between you two will get the bicycle.

10
00:00:41,534 --> 00:00:45,674
After hearing father's words, both the kids started studying.
```

DRAWBACKS

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.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.

CONCLUSION

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. 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 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.

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, ‘‘Densecaptioning 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.