TRANSFORMING IMAGE UNDERSTANDING WITH AUTOMATED CAPTION GENERATION USING DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

Saasupalli Sukanya[1], Bollaram Charan[2], K Maheshwar Reddy[3], Mr. G Sathish[4]
[1,2,3] UG Scholar, Dept. of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4] Assistant Professor, Dept. of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
sukanyasaasupalligmail.com

Abstract

Early methods of describing images relied on human input, such as manual tagging and writing descriptions. Basic image classification techniques provided limited labels. As computer vision evolved, traditional feature extraction methods (like SIFT and HOG) were used for recognizing objects but lacked the ability to generate comprehensive descriptions. The objective of this project is to develop a system that automatically generates descriptive captions for images by leveraging deep learning models and natural language processing techniques, enhancing image understanding and accessibility. Before AI, traditional systems used basic image annotation tools where humans manually tagged images with metadata. Text-based image search engines relied on keywords provided by users. Descriptions were written manually, limiting scalability and accuracy. The manual generation of image captions is time-consuming, inconsistent, and inefficient for large datasets. Traditional systems lacked the ability to provide contextual or meaningful descriptions for complex images. The rapid increase in visual data and the need for efficient, scalable image description solutions motivate the exploration of AI models for automating caption generation, making visual content more accessible and understandable. The proposed system aims to enhance image understanding through automated caption generation by integrating deep learning and natural language processing (NLP) techniques. It employs a convolution neural network (CNN) as the encoder to extract salient features from images, effectively capturing essential visual elements. These features are then input into a recurrent neural network (RNN) or long short term memory (LSTM) model, which serves as the decoder to generate coherent and contextually relevant captions. The NLP component is crucial for interpreting the extracted image features and generating human-readable text, allowing the system to understand linguistic structures and semantics. The system is trained on a comprehensive dataset of images with paired descriptive captions, enabling it to learn the intricate relationship between visual content and language.

Key words: *AI, SIFT, HOG, CNN, NLP, RNN, LSTM etc.*

1. INTRODUCTION

The transformation of image understanding through automated caption generation represents a significant evolution in computer vision and artificial intelligence. Historically, methods for describing images relied heavily on human intervention, including manual tagging and crafting of descriptions, which were often simplistic and prone to errors. Early techniques in image classification provided limited labeling capabilities, falling short in offering meaningful insights. As computer vision progressed, traditional feature extraction methods, such as Scale-Invariant. Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), were utilized for object recognition but failed to generate comprehensive and contextual descriptions of images. In India, the rapid growth of visual content—driven by social media, e-commerce, and digital marketing example files the pressing need for efficient image description solutions. Statistics show that the country has one of the highest rates of internet penetration, leading to an overwhelming amount of visual data being generated daily.

Thus, the objective of this project is to develop a system that automatically generates descriptive captions for images by leveraging deep learning models and natural language processing techniques, significantly enhancing image understanding and accessibility. Applications include improving search engine capabilities, assisting visually impaired individuals, and enhancing content creation for digital platforms.

The objective behind this research stems from the growing demand for efficient and effective image description systems in today's digital landscape. With the exponential increase in visual content generated across platforms, there is an urgent need for solutions that can automate the captioning process. Traditional methods are no longer adequate, as they are limited in their ability to provide contextual and detailed descriptions.

By integrating deep learning and natural language processing techniques, this research aims to create a robust automated caption generation system that enhances accessibility and understanding of visual information. Moreover, as the field of artificial intelligence continues to advance, leveraging these technologies will not only improve image captioning but also open up new avenues for applications in various sectors, such as healthcare, e-commerce, and education.consistent diagnosis, particularly valuable in areas with limited radiology expertise.

The Early techniques in image classification provided limited labeling capabilities, falling short in offering meaningful insights. As computer vision progressed, traditional feature extraction methods, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), were utilized for object recognition but failed to generate comprehensive and contextual descriptions of images.

In India, the rapid growth of visual content—driven by social media, e-commerce, and digital marketing—exemplifies the pressing need for efficient image description solutions. Statistics show that the country has one of the highest rates of internet penetration, leading to an overwhelming amount of visual data being generated daily. Thus, the objective of this project is to develop a system that automatically generates descriptive captions for images by leveraging deep learning models and natural language processing techniques, significantly enhancing image understanding and accessibility, for object recognition but failed to generate comprehensive and contextual descriptions of images.

2. LITERATURE SURVEY

Oriol vinyals this seminal paper introduced an end-to-end deep learning approach to automatically generating image captions. The authors developed a neural image captioning model based on an encoder-decoder framework using a convolution neural network (CNN) as the image encoder and a recurrent neural network (RNN) with long short-term memory (LSTM) units as the decoder. The model achieved state-of-the-art results on the MSCOCO dataset, demonstrating the ability to generate descriptive and semantically meaningful captions.

Kelvin xu this paper introduced an attention-based model for image captioning, allowing the network to focus on different parts of the image while generating each word in a caption. The use of soft and hard attention mechanisms significantly improved caption quality, making the model more interpretable and enhancing performance on benchmark datasets.

Steven J. Renneie theauthors proposed a reinforcement learning-based optimization technique called self- critical sequence training (SCST), which improves caption generation by directly evaluation metrics such as CIDEr. By refining the training process, SCST enhances the quality.

Peter Anderson this work introduced a novel attention mechanism combining bottom-up and top-down attention for image captioning and visual question answering (VQA). By allowing the model to focus on salient image regions, the approach improved caption quality and interpretability, setting new benchmarks on the MSCOCO dataset.

Steven Hoi the BLIP framework leveraged large-scale retraining to improve vision-language tasks, including image captioning. By utilizing a bootstrapped approach to align visual and textual representations, BLIP achieved superior performance across multiple captioning benchmarks.

Haotian Liu this study explored multimodal learning by integrating large language models (LLMs) with visual understandingfor tasks like image captioning and question answering. By leveraging contrastive and generative learning strategies, LLaVA demonstrated significant improvements in caption fluency and accuracy.

Kelvin Xu this paper introduced an attention-based model for image captioning, where the network learns to focus on relevant image regions while generating each word in a caption. It demonstrated improvements in caption quality and interpretability.

3. PROPOSED METHODOLOGY

To overcome the limitations of traditional image captioning methods, a deep learning-based automated caption generation system is proposed. This system leverages Convolution Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory LSTM) network an meaningful.

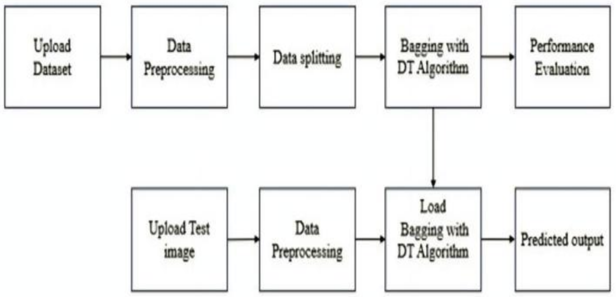


Fig 1: System Architecture

The proposed algorithm aims to automate the image captioning using AI technologies, specifically Convolution Neural Networks (CNNs) and the RNN (Recurrent Neural Network) algorithm. This system is designed to provide continuous, accurate monitoring and classification of images using real-time image feeds, thus addressing the limitations of traditional manual enforcement methods. The following steps outline the research procedure and implementation details. The model achieved state-of-the-art results by effectively capturing contextual relationships within an image. This system leverages Convolution Neural Networks (CNNs) for image feature extraction.

Dataset

The foundation of the proposed algorithm begins with the collection of a comprehensive dataset. This dataset includes a variety of the dataset consists of 8,091 images paired with 40,456 captions, where each image has multiple human-written descriptions. The dataset is structured in a way that each image filename is linked to multiple textual descriptions, capturing different perspectives. The images cover various scenes, including people, animals, objects, and activities, making it a diverse dataset for training deep learning models in automated image captioning. This involves several key steps. First, any null or missing values in the dataset are identified and removed to ensure the integrity of the data.

Dataset Pre-processing

Once the dataset is collected, preprocessing is conducted to prepare the data for training. This involves several key steps. First, any null or missing values in the dataset are identified and removed to ensure the integrity of the data. Null values can skew the training process and lead to inaccurate predictions, so their removal is essential. Preprocessing is done for both images and text to enhance model performance. For images, they are resized and normalized before being fed into a convolution neural network (CNN). For text (NLP preprocessing), captions are tokenized, converted to lowercase, and filtered for special characters. Stop words are removed, and padding created by mapping words to unique integer values for efficient processing.

Proposed Algorithm -Convolution Neural Network (CNN)

CNNs process image frames by passing them through a series of convolution layers that extract hierarchical features, such as edges, textures, and shapes, from the images. These features are crucial for identifying different elements in images. NLP enhances this process by dividing each image frame into a grid and applying bounding boxes to identify and localize objects within each grid cell. The algorithm then classifies these images into predefined categories. CNN’s real-time processing capability allows for immediate detection and classification of multiple images in a single pass, making it highly efficient for caption generation. It demonstrated improvements in caption quality and interpretability.

Performance Comparison

To evaluate the effectiveness of the proposed system, a performance comparison is conducted between the traditional manual method and the AI-based approach using CNN and RNN. Key metrics such as detection accuracy, processing speed, and false positive rates are measured. The AI-driven system is expected to outperform the manual approach by providing more consistent and accurate captioning, and significantly reducing the time required to identify data. The comparison is crucial in demonstrating the superiority of the proposed system, particularly in its ability to operate continuously without fatigue and to process vast amounts of data in real time. CNN's real-time processing capability allows for immediate detection and classification of multiple images in a single pass, making it highly efficient for caption generation.

4. EXPERIMENTAL ANALYSIS

Figure 2 shows the GUI (using stream lit)

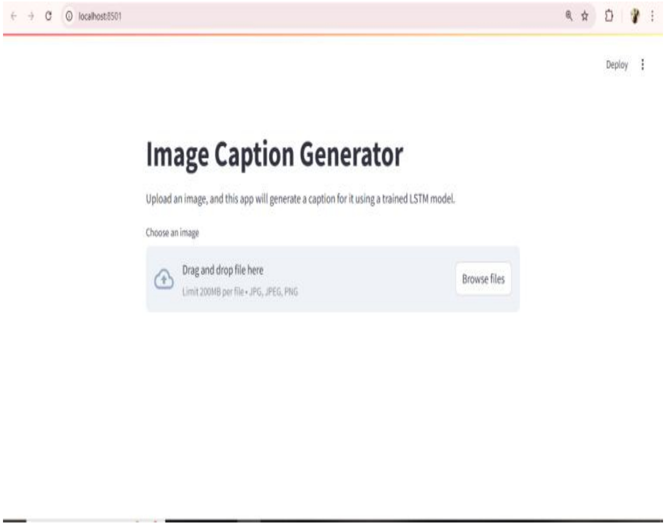


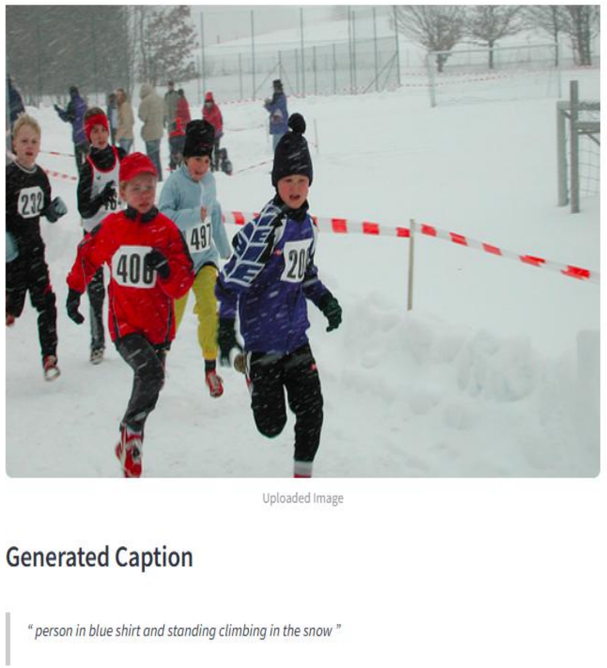
Image Caption Generator

Upload an image, and this app will generate a caption for it using a trained LSTM model.

Choose an image

Figure 2: GUI (using stream lit)

The predict_caption function takes an image's feature vector and generates a caption word by word. The process starts with the start seq token and continues until the end seq token is predicted or the maximum caption length is reached. The predicted caption is compared with the actual captions to evaluate the model's performance.

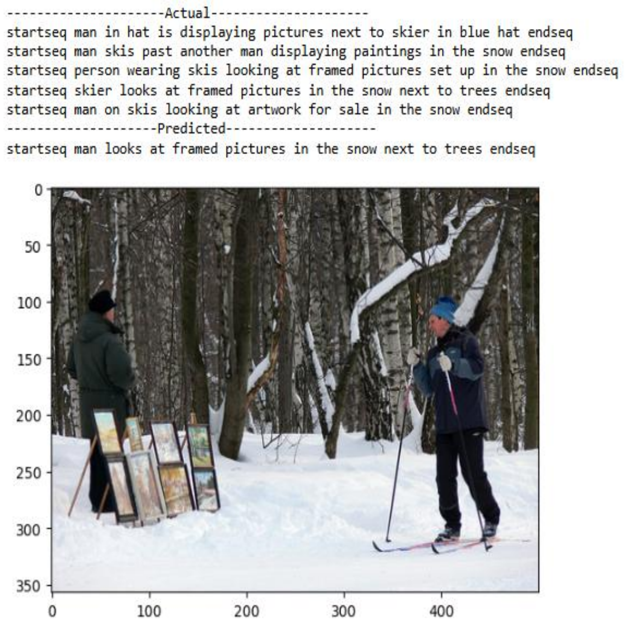


Generated Caption

> " person in blue shirt and standing climbing in the snow "

Figure 3: Prediction on Stream lit

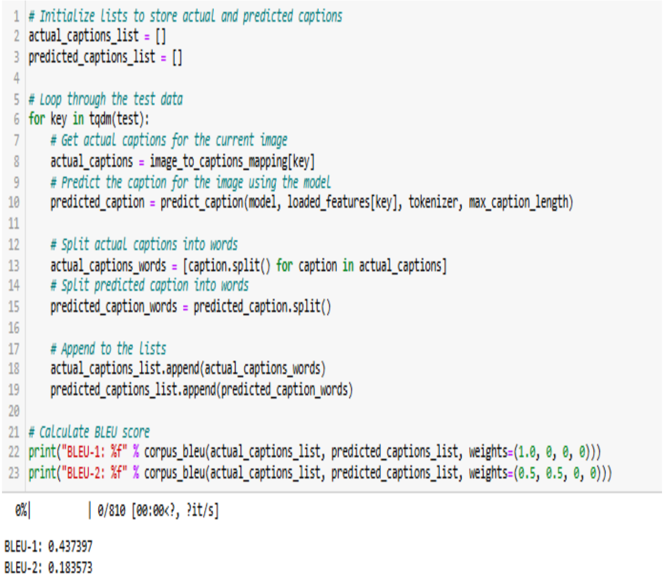
Figure 3The final step in the proposed algorithm involves using the trained CNN and RNN model to generate captions to the images. The model is fed real-time image feeds , and it applies the learned patterns to classify any data. The output predictions are then validated against the actual scenarios to assess the model's accuracy and reliability. The model's predictions are expected to align closely with real-world observations, demonstrating its ability to generalize from the training data to accurately detect violations in a variety of environments and conditions. This step is essential for validating the effectiveness of the system in practical, real-world applications, ensuring that it can be reliably deployed for generation of captions.



```
---------------------Actual---------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted--------------------
startseq man looks at framed pictures in the snow next to trees endseq
```

Figure 4: prediction on jupyter notebook

Figure 3To evaluate the effectiveness of the proposed system, a performance comparison is conducted between the traditional manual method and the AI-based approach using CNN and RNN. Key metrics such as detection accuracy, processing speed, and false positive rates are measured. The AI-driven system is expected to outperform the manual approach by providing more consistent and accurate captioning, and significantly reducing the time required to identify data. The comparison is crucial in demonstrating the superiority of the proposed system, particularly in its ability to operate continuously without fatigue and to process vast amounts of data in real time.



```
# Initialize lists to store actual and predicted captions
actual_captions_list = []
predicted_captions_list = []

# Loop through the test data
for key in tqdm(test):
    # Get actual captions for the current image
    actual_captions = image_to_captions_mapping[key]
    # Predict the caption for the image using the model
    predicted_caption = predict_caption(model, loaded_features[key], tokenizer, max_caption_length)

    # Split actual captions into words
    actual_captions_words = [caption.split() for caption in actual_captions]
    # Split predicted caption into words
    predicted_caption_words = predicted_caption.split()

    # Append to the lists
    actual_captions_list.append(actual_captions_words)
    predicted_captions_list.append(predicted_caption_words)

# Calculate BLEU score
print("BLEU-1: %f" % corpus_bleu(actual_captions_list, predicted_captions_list, weights=(1.0, 0, 0, 0)))
print("BLEU-2: %f" % corpus_bleu(actual_captions_list, predicted_captions_list, weights=(0.5, 0.5, 0, 0)))
```

```
  0%|          | 0/810 [00:00<?, ?it/s]

BLEU-1: 0.437397
BLEU-2: 0.183573
```

Figure 5: Performance Evaluation of CNN

Figure 4The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of text generated by machine learning models, typically for machine translation or text generation tasks. It measures how closely the generated text matches one or more reference texts based on n-gram precision.BLEU-1 (0.437397): This score evaluates unigrams (single words) and shows how many words in the generated captions appear in the actual captions.BLEU-2 (0.183573): This score considers bigrams (two-word sequences), meaning it accounts for short phrases rather than just individual words.BLEU-1: 0.437 (~43.7%) → About 43.7% of the words in the generated captions match the actual captions.

5. CONCLUSION

The implementation of an image captioning model using NLP, Bi-directional LSTM, and CNN with VGG16 demonstrates significant advancements in artificial intelligence and deep learning. By extracting image features using VGG16 and generating captions with Bi- directional LSTM, the model effectively bridges the gap between computer vision and natural language processing. The BLEU score evaluation validates the model's performance, showing its ability to generate meaningful and contextually relevant captions. This approach enables machines to understand and describe images in a human-like manner, making it applicable in automated image annotation, accessibility for visually impaired individuals, and content-based image retrieval systems.

The combination of CNN for feature extraction and LSTM for sequential caption generation ensures that the model understands the spatial structure of images while maintaining the context of sentences. Despite its promising results, challenges such as handling complex image scenarios, improving grammatical accuracy, and reducing bias in caption generation remain. Enhancing the dataset, using transformer-based models, or integrating attention mechanisms can further improve caption quality. Overall, this work contributes to the ongoing research in deep learning and AI, paving the way for more sophisticated models in image captioning and multimodal AI applications.

This paper introduced an attention-based model for image captioning, where the network learns to focus on relevant image regions while generating each word in a caption. It demonstrated improvements in caption quality and interpretability. The manual generation of image captions is time-consuming, inconsistent, and inefficient for large datasets. Traditional systems lacked the ability to provide contextual or meaningful descriptions for complex images. The rapid increase in visual data and the need for efficient, scalable image description solutions motivate the exploration of AI models for automating caption generation, making visual content more accessible and understandable.

REFERENCES

[1]Vinyals, O., Toshev, A., Bengio, S., Erhan, D., "Show and Tell: A Neural Image Caption Generator", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, 2017, DOI: 10.1109/TPAMI.2016.2587640.

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", Proceedings of the 32nd International Conference on Machine Learning, 2015, arXiv:1502.03044.

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.,"Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, DOI: 10.1109/CVPR.2018.00117.

[4] Hossain, M. S., Sohel, F., Shiratuddin, M. F., Laga, H., "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, vol. 51, no. 6, pp. 1-36, 2019, DOI: 10.1145/3295748.

[5] Karpathy, A., Fei-Fei, L., "Deep Visual-Semantic Alignments for Generating Image Descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39,no.4,pp.664-676,10.1109/TPAMI.2016.2598339.

[6] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., "Self-Critical Sequence Training for ImageCaption,IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, DOI: 10.1109/CVPR.2017.131.

[7] Cornia, M., Baraldi, L., Cucchiara, R., "Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, DOI: 10.1109/CVPR.2019.00344.

[8] Huang, L., Wang, W., Chen, J., Qian, S., "Attention on Attention for Image Captioning", IEEE International Conference on Computer Vision (ICCV), 2019, DOI: 10.1109/ICCV.2019.00392.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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.,"Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, DOI: 10.1109/CVPR.2018.00117.

[17] Hossain, M. S., Sohel, F., Shiratuddin, M. F., Laga, H., "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, vol. 51, no. 6, pp. 1-36, 2019, DOI: 10.1145/3295748.

[18] Karpathy, A., Fei-Fei, L., "Deep Visual-Semantic Alignments for Generating Image Descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39,no.4,pp.664-676,10.1109/TPAMI.2016.2598339.

[19] Huang, L., Wang, W., Chen, J., Qian, S., "Attention on Attention for Image Captioning", IEEE International Conference on Computer Vision (ICCV), 2019, DOI: 10.1109/ICCV.2019.00392.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.

[20] 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.

[21] 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.

[22] 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.