Creating an Image Captioning System Enhanced with a Text-to-Speech Component to Enhance Accessibility for Individuals with Visual Impairments.

Abstract:

Individuals who are visually impaired often face challenges when navigating unfamiliar environments. To help address this issue, a new system has been proposed that utilizes computer vision, natural language processing, and deep learning techniques. By leveraging transfer learning and an attention mechanism, the technology is capable of generating accurate image captions that enable visually impaired individuals to better comprehend their surroundings and move with greater ease. The findings of this research indicate that this novel approach to assistive technology has the potential to substantially improve the quality of life for visually impaired individuals. The combination of transfer learning, attention mechanisms, and deep learning employed by the technology has produced high-quality image captions that have proven to be very effective. The study's results suggest that this system represents a significant advancement in the field of assistive technology and has the potential to make a meaningful impact in the lives of those with visual impairments.

Introduction:

The goal of the rapidly expanding study field of image caption generation, which falls under the broad categories of computer vision and natural language processing, is to automatically produce descriptions of visual content in natural language. A more natural approach for people to interact with visual data is made possible by this technology, which enables computers to recognise and describe the content of an image using human-like language. The potential applications of image caption generation are vast, including image retrieval, content-based image indexing, and assistive technologies for individuals with visual impairments.

In recent years, advances in deep learning, transfer learning, and attention mechanisms have substantially enhanced the image's accuracy and quality generation. By combining transfer learning and attention mechanisms, researchers have developed more sophisticated models that can produce high-quality image captions that accurately describe the content of an image. This has opened up numerous applications, including assistive technologies for individuals with visual impairments.

To this end, I have implemented image caption generation and text-to-speech technology in my project to assist individuals with visual impairments. Using deep learning and computer vision techniques, I created a methodology for creating image captions that completely captures the meaning of the image. Through the use of text-to-speech technology, the created captions are then transformed into spoken words, allowing people who are blind to more easily understand their environment. A big image dataset that was used for training the model, allowing it to

recognize and describe a wide range of visual content accurately. The text-to-speech system has been optimized to provide clear and natural-sounding speech, making it easy for visually impaired individuals to comprehend the captions.

Overall, the advancements in image caption generation and text-to-speech technology have significant implications for individuals with visual impairments. The technology can help them better understand their environment and navigate with greater ease, leading to greater independence and a higher quality of life. Additionally, this technology has the potential to revolutionize the way we interact with visual data, opening up new opportunities for research and applications in various fields.

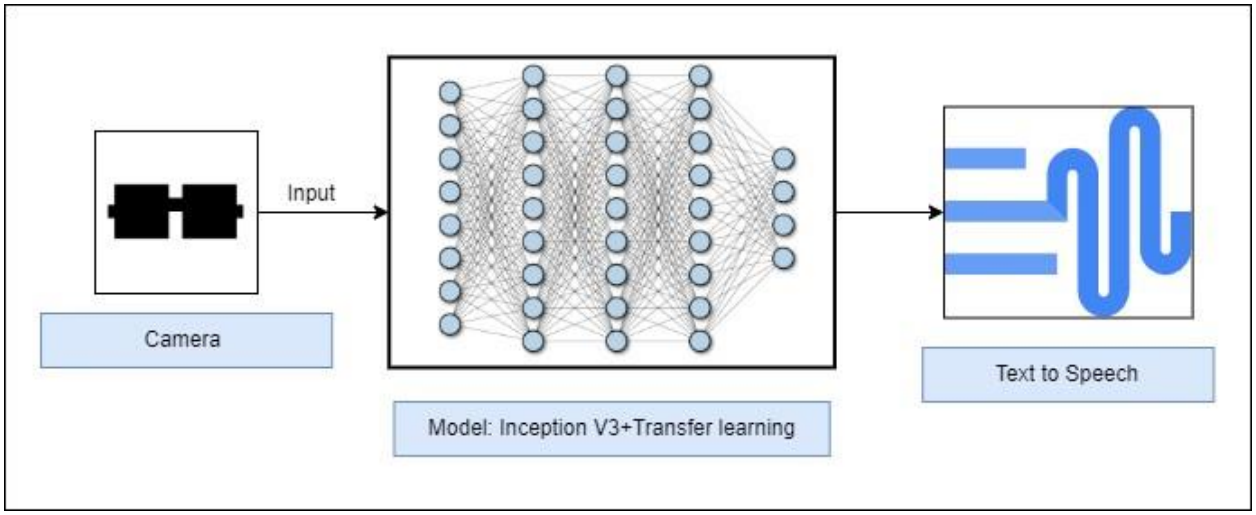


Fig. 1 System architecture for image captioning

Literature Review:

It is necessary to create a navigation system for the visually impaired, that is fully functioning.
S.R Alwi and M.N. Ahmad's study [2] looks at the daily requirements of the blind people who live in urban areas. The results, which include a walking stick and dark glasses, demonstrate that people who are blind or visually handicapped need a portable navigation system that may help them cross roadways and prevent dangerous accidents.

Every time a human is shown an image he or she involuntarily labels the image, this process is done by our brain to describe the image and the process of describing the image through words is called image captioning. Machines are now being trained to perform similar tasks. There are various methodologies and algorithms that have been used and are being used for image captioning, these methodologies are also efficient enough to caption videos and not just images.

Youtube video auto-captioning is a live example of the same. The solutions for precise image captioning can help break the barrier of languages for a lot of digital and image-based information while it will also benefit the blind by describing the image or videos to them.

The past papers have tackled the problem of image captioning by either using recurrent neural network technology, generally long short-term memory (LSTM)[3], or by using a CNN-RNN-based framework [4].Convolutional neural networks are used for image encoding, and recurrent neural networks are used for image decoding, in the CNN-RNN pair method. Image features or vectors that are derived from the images via CNN, these vectors are then feeded to RNN layers and networks and serve as the input for the same. The NLTK libraries are then used to generate the captions. Generally these algorithms pick the closest and most relevant Verbs, Nouns and Prepositions to describe the image provided as the input. Thus the captions, although well equipped with the vocabulary to describe the image, lack the grammar and caption forming technique to create a sound and understandable complete caption. Few of the other papers suggest using V3 as the encoder module and Gated Recurrent neural networks[GRU] as their decoder while also using attention mechanisms [5].

These methodologies however focus on the main object in the image and fail to generate captions for multiple center focus objects that require more than a sentence of caption. These modules also require high volumes of training data set for them to perform efficiently since they are based on deep learning techniques, generating such high volumes of data can be an ordeal as they would have to be updated frequently based on the exposure the algorithm faces.

The algorithms generally use Xception it is responsible for image feature extraction and helps the machine understands and figures out the main object to be described. This problem is one of the key sub-problems of the problem statement to be tackled efficiently as without adequate and accurate feature extraction the caption can fail to provide reliable knowledge. The feature extraction solves the major two problems with the captioning of an image. First, it prevents the machine from describing irrelevant features in the image that would make the caption long and more than a few sentences, second it helps in keeping focus on the main object in the image. For example let's take an example of an image of a girl walking a dog, with the help of feature extraction (Xception, CNN or etc) the algorithm would extract the center focus(the girl and the dog) it would drop the other irrelevant features like the road they are walking on or the trees or other objects surrounding them. Then the caption would be generated by relating the two nouns (girl and dog) with a verb (walking).

In the paper published by subrata das, Lalit jain [6]. This paper basically uses CNN-RNN based models for tackling the problem. The author proposes using the Inception model for encoding the images and LSTM algorithm for reducing the gradient descent of the algorithm. Although there are quite a few solutions out there the core basics of almost all solutions rely on using CNN or RNN and these algorithms have their drawbacks. The CNN-CNN algorithm has high loss and hence can provide inaccurate captioning on an image. The CNN-RNN algorithm or procedure solves the problem of high loss however it fails to be fast when it comes to the training time. The algorithm of such kind also needs more than usual amounts of data which would eventually reduce the complete efficiency of the algorithm. Both these methods face the problem of vanishing gradient descent. Gradient is equal to the ratio of change of weights to the

change to error of an algorithm in a deep neural network. The vanishing gradient descent disables the RNN module to store the words for a longer run and disables the algorithm to provide longer captions, usually more than a few sentences thus, reducing the solutions' effectiveness and precision.

Methodology:

The goal of this project is to create a system that can generate a natural language description of an image, and then convert that description to speech to help blind people understand the content of the image. The system will use the InceptionV3 image recognition model and a transformer model for natural language generation.
Here are the steps to create the system:

A. Data Collection:
B. Data preprocessing:
C. Image Captioning Model:
D. Model Training:
E. Model Deployment:

Data Pre-processing

We used the popular Microsoft COCO 2017 (COCO) benchmark dataset to train the Transformer model for image captioning tasks. The dataset consists of 82k images, with each image having five human-annotated captions, resulting in a total of over 410k image-text pairs. Sample of 70000 images is used and out of that 56013 image caption pairs are used for training the transformer model and 13987 image caption pairs are used for testing the model.

Before model training Data pre-processing is implemented on images and text captions. Images cannot be directly given as inputs to any model for training. Images must be converted to Feature vectors first before feeding to model for training. Hence we have used the inception V3 model for image to vector conversion. We have used tensor flow keras inceptionV3 to implement the vectorization process. Any image provided as input is initially resized to 299X299 dimensions and then normalized by dividing the vectors by 255, these generated values images that are passed inside the inception V3 model that generates the final image vectors.

Text Captions are also pre-processed. Pre-processing involves lowering the text, removing the punctuations and adding start and end tokens to mark the beginning and end of every caption. Next step is vectorising the captions. This will generate the unique number for each and every word as neural networks can only deal with numbers.

Image Captioning Model

The Transformer is a powerful deep learning model architecture that was introduced by Google researchers in a 2017 paper titled "Attention is All You Need". It is designed for natural language processing (NLP) tasks such as language translation, sentiment analysis, and text summarization, among others.

The Transformer model is based on the concept of self-attention, where the model pays attention to different parts of the input sequence to determine the importance of each part for generating the output sequence. This attention mechanism allows the model to capture long-term dependencies between different parts of the sequence, which is particularly important for NLP tasks where the meaning of a word or phrase can depend on its context.

The Transformer architecture consists of a series of encoder and decoder layers, each of which includes a self-attention mechanism. The encoder processes the input sequence and generates a sequence of hidden states, while the decoder generates the output sequence based on the encoder's hidden states and a context vector that summarizes the input sequence.

The Transformer model has proven to be highly effective for a wide range of NLP tasks, and it has been widely adopted in the research and industry communities. It has also been the basis for many state-of-the-art language models, including GPT-3, one of the largest and most powerful language models to date.

Model training

Training a Transformer for image captioning typically involves the following pipeline:

1. Data Preparation: The first step is to prepare the data for training. This involves collecting a large dataset of images and their corresponding captions. The dataset should be diverse, and the images and captions should be aligned so that each caption corresponds to the correct image.

2. Pre-processing: The next step is to pre-process the data. This involves converting the images to a format that the model can work with, such as resizing them to a uniform size and normalizing the pixel values. The captions are typically tokenized and converted to a numerical format that the model can process.

3. Model Architecture: The Transformer architecture is adapted for image captioning by modifying the input to include the image features in addition to the text input. This can be done using a pre-trained convolutional neural network (CNN) to extract the image features, which are then fed into the Transformer encoder along with the tokenized captions.

4. Loss Function: The loss function used to train the model is typically the cross-entropy loss, which measures the difference between the predicted caption and the ground-truth caption. The goal of the training is to minimize this loss.

5. Training: The model is trained using stochastic gradient descent (SGD) or a similar optimization algorithm. During training, the model is presented with an image and a corresponding caption, and it tries to generate a caption that is as close as possible to the ground-truth caption. The model is updated after each batch of images and captions is processed.

6. Evaluation: Once the model has been trained, it is evaluated on a separate validation dataset to determine its performance. This involves generating captions for a set of images and comparing them to the ground-truth captions. The performance is typically measured using metrics such as BLEU, METEOR, and ROUGE.

7. Fine-tuning: The model can be further fine-tuned on a smaller dataset of images and captions to improve its performance on specific tasks or domains.

Overall, training a Transformer for image captioning requires a large and diverse dataset, careful pre-processing, and modifications to the model architecture to incorporate image features. With these steps, a Transformer can generate high-quality captions that accurately describe the content of an image.

Model Deployment/Generating Captions for new images

We have written an Inference function, unlike training data we do not have training captions while testing the model for new images, hence [Start] token is given to the model and it continues to predict tokens till [end] special token is predicted.

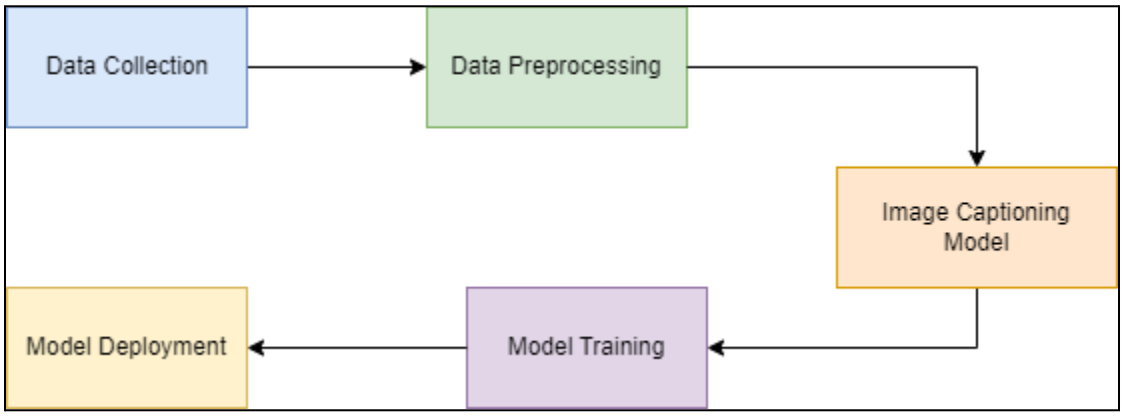


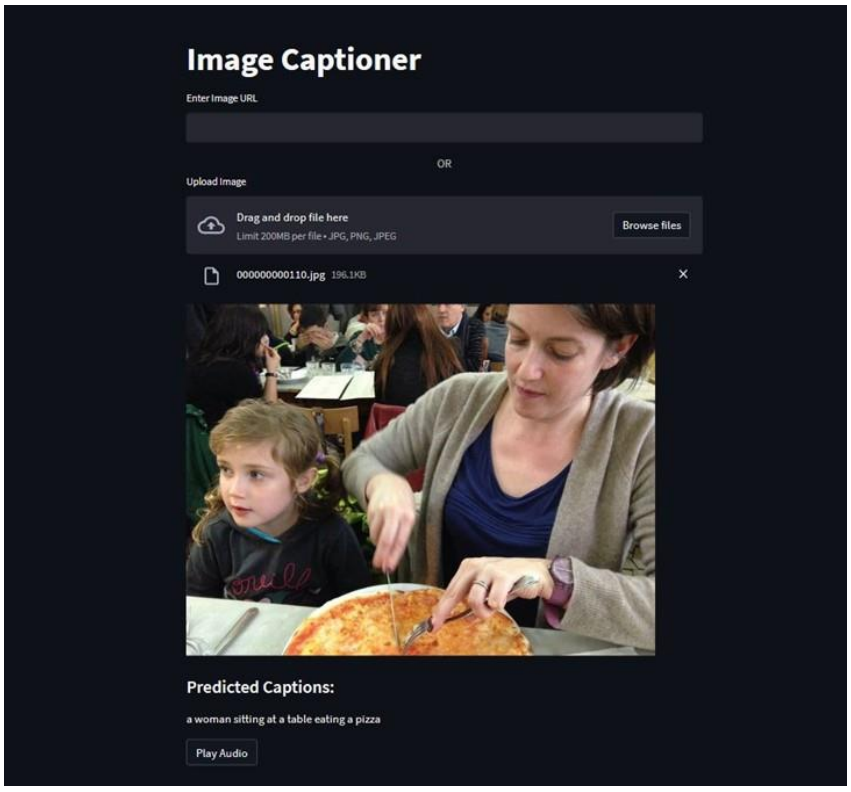
Fig.2. The flowchart of the model

Text to Speech Module.

Speech to text is one of the most crucial and distinguishing features of this project. Text captions generated are converted to speech using pyttsx3 module of python, This module is compatible with both python 2 and python 3. This can be applied as a descriptive tool for people with vision impairment. Project is capable of generating image captions based on input image and converting the generated captions to speech which could help people with vision impairment to get description of scenes.

Result:

Throughout this research, we made use of a deep learning model based on Inception V3 architecture for generating captions from images. The aim of this project was to enhance accessibility for individuals who might face difficulties in visualizing images by providing spoken descriptions. The model was trained on a vast dataset of images and corresponding captions, utilizing transfer learning for better accuracy. Following caption generation, we used Google Text-to-Speech API for converting the textual description into speech. The resulting audio file contained a natural-sounding spoken description of the image that could be played back for visually-impaired individuals or anyone who prefers an auditory description. This project highlights the potential of deep learning and TTS technologies for promoting inclusivity by improving accessibility. Future advancements in image captioning and TTS can lead to even more precise and natural-sounding image descriptions, creating a more inclusive digital environment for everyone.



Conclusion:

The paper describes an assistive system in order to help the blind and visually impaired. The development of an image captioning technology that provides real-time scene descriptions for visually impaired and blind individuals, especially for low-resource languages without dedicated image caption databases, is presented in this paper. The feasibility of the approach was confirmed through experiments with a visually unimpaired individual. Future work includes deploying the system on smartphones to reach a wider audience and incorporating contextual information into the image captioning model to generate relevant descriptions for the visually impaired and blind.

Reference:

1. Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Global Health 2020. doi: 10.1016/S2214-109X(20)30425-3.
2. S. R. A. W. Alwi and M. N. Ahmad, “Survey on outdoor navigation system needs for blind people,” 2013 IEEE Student Conference on Research and Development, Putrajaya, Malaysia, 2013, pp. 144-148, doi: 10.1109/SCOReD.2013.7002560.
3. ShuangLiu, Liang Bai, Yanli Huand Haoran Wang, “Image Captioning based on Deep Neural Networks” https://doi.org/10.1051/matecconf/201823201052
4. H. Yanagimoto and M. Shozu, "Multiple Perspective Caption Generation with Attention Mechanism," *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, Kitakyushu, Japan, 2020, pp. 110-115, doi: 10.1109/IIAI-AAI50415.2020.00031.
5. V. Agrawal, S. Dhekane, N. Tuniya and V. Vyas, "Image Caption Generator Using Attention Mechanism," *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.
6. Ansar Hani Najiba Tagougui and Monji Kherallah "Image caption generation using a deep architecture" International Arab Conference on Information Technology ACIT 2019.
7. Himanshu Sharma Manmohan Agrahari Sujeet Kumar Singh Mohd Firoj and Ravi Kumar Mishra "Image captioning: a aomprehensive survey" International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control PARC 2020.
8. Adela Puscasiu Alexandra Fanca Dan-Ioan Gota and Honoriu Valean "Automated image captioning" IEEE International Conference on Automation Quality and Testing Robotics AQTR 2020.
9. Phyu Phyu Khaing and May The’ Yu "Attention-based deep learning model for image captioning: a comparative study" International Journal of Image Graphics and Signal Processing June 2019.
10. O. Vinyals A. Toshev S. Bengio et al. "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge" IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 39 no. 4 2017.
11. Asifullah Khan Anabia Sohail Umme Zahoora and Aqsa Saeed Qureshi "A survey of the recent architectures of deep convolutional neural networks" Artificial Intelligence Review 2020.

12. Jojo John Moolayil Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python New York:Apress 2019.
13. Shailja Gupta Manpreet Kaur Sachin Lakra and Mayank Khattar "Application-based attention mechanisms in deep learning" International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) ICRITO September 2020
14. Murk Chohan Adil Khan Muhammad Saleem Mahar Saif Hassan Abdul Ghafoor and Mehmood Khan "Image captioning using deep learning: a systematic literature review" International Journal of Advanced Computer Science and Applications vol. 11 no. 5 2020.
15. Yong Yu Xiaosheng Si Changhua Hu and Jianxun Zhang "A review of recurrent neural networks: LSTM cells and network architectures" Neural Computation vol. 31 no. 7 July 2019.
16. Kanchan M. Tarwani and Swathi Edem "Survey on recurrent neural network in natural language processing" International Journal of Engineering Trends and Technologies IJETT vol. 48 no. 6 June 2017.
17. J. Vaishnavi and V. Narmatha, "Video Captioning based on Image Captioning as Subsidiary Content," *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Bhilai, India, 2022, pp. 1-6, doi: 10.1109/ICAECT54875.2022.9807935.
18. J. Park M. Rohrbach T. Darrell and A. Rohrbach "Adversarial inference for multi- -sentence video description" CVPR pp. 6591-6601 2019.
19. S.H. Abdulhussain S.A.R. Al-Haddad M.I. Saripan B.M. Mahmmod and A. Hussien "Fast temporal video segmentation based on krawtchouk-tchebichef moments" IEEE Access vol. 8 pp. 72347-72359 2020.
20. S.H. Abdulhussain A. Rahman Ramli B.M. Mahmmod M. Iqbal Saripan S.A.R. Al-Haddad T. Baker et al. "A fast feature extraction algorithm for image and video processing" International Joint Conference on Neural Networks (IJCNN) pp. 1-8 2019.
21. J. Lei L. Wang Y. Shen D. Yu T.L. Berg and M. Bansal "Mart: memory-augmented recurrent transformer for coherent video paragraph captioning" ACL 2020.
22. A. Kojima T. Tamura and K. Fukunaga "Natural language description of human activities from video images based on concept hierarchy of actions" International Journal of Computer Vision vol. 50 no. 2 pp. 171-84 2002.
23. L. Yao A. Torabi K. Cho N. Ballas C. Pal H. Larochelle et al. "Describing videos by exploiting temporal structure" Proceedings of the IEEE international conference on computer vision pp. 4507-4515 2015.
24. V. Iashin and E. Rahtu "A better use of audio-visual cues: dense video captioning with bi-modal transformer" British Machine Vision Conference (BMVC) 2020.
25. V. Iashin and E. Rahtu "Multi-modal dense video captioning" CVPR Workshops pp. 958-959 2020.
26. D. L. Chen and W. B. Dolan "Collecting highly parallel data for paraphrase evaluation" ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies vol. 1 pp. 190-200 2011.
27. G. A. Sigurdsson X. Wang A. Farhadi I. Laptev and A. Gupta "Hollywood in Homes : Crowdsourcing Data" arXiv pp. 1-17 2016.
28. R. Krishna K. Hata F. Ren L. Fei-Fei and J. C. Niebles "DenseCaptioning Events in Videos" Proceedings of the IEEE International Conference on Computer Vision vol. 2017-Octob pp. 706-715 2017.
29. N. Xu A. Liu Y. wong Y Zhang W. Nie Y. Su et al. "Dual-Stream Recurrent Neural Network for Video Captioning" IEEE Transactions on Circuit and systems for video technology 2018.

30. J. Song Y. Guo L. Gao X. Li and H.T. Shen "From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning" IEEE Transactions on Neural Networks and Learning Systems 2018.
31. R. Shetty and J. Laaksonen "Frame-and segment-level features and candidate pool evaluation for video caption generation" ACM-MM pp. 1073-1076 2016.
32. P. Pan Z. Xu Y. Yang F. Wu and Y. Zhuang "Hierarchical recurrent neural encoder for video representation with application to captioning" CVPR pp. 1029-1038 2016.
33. Y. Yang J. Zhou J. Ai Y. Bin A. Hanjalic H.T. hen et al. "Video Captioning by Adversarial LSTM" IEEE Transaction on Image Processing 2018.
34. R. Shetty M. Rohrbach L. Anne Hendricks M. Fritz and B. Schiele "Speaking the same language: matching machine to human captions by adversarial training" ICCV pp. 4135-4144 2017.
35. T Lin M. Maire S. Belongie L. Bourdev and R. Girshick "Microsoft COCO: Common Objects in Context" 2015.
36. H. Yu J. Wang Z. Huang Y. Yang and W. Xu "Video paragraph captioning using hierarchical recurrent neural networks" IEEE CVPR pp. 4584-4593 2016.
37. L. Gao X. Li J. Song and H. T. Shen "Hierarchical LSTMs with Adaptive Attention for Visual Captioning" IEEE Journal of Latex Class Files vol. 14 no. 8 August 2015.
38. I. Goodfellow J. Pouget-Abadie M. Mirza B. Xu D. Warde-Farley S. Ozair et al. "Generative adversarial nets" NeurIPS pp. 2672-2680 2014.
39. B. Dai S. Fidler R. Urtasun and D. Lin "Towards diverse and natural image descriptions via a conditional GAN" ICCV pp. 2970-2979 2017.
40. X. Liang Z. Hu H. Zhang C. Gan and E.P. Xing "Recurrent topic-transition GAN for visual paragraph generation" ICCV pp. 3362-3371 2017.
41. J. Wang J. Fu J. Tang Z. Li and T. Mei "Show reward and tell: automatic generation of narrative paragraph from photo stream by adversarial training" AAAI 2018.
42. O. Vinyals A. Toshev S. Bengio and D. Erhan "Show and Tell: A Neural Image Caption Generator" IEEE CVPR pp. 3156-3164 2015.
43. Y. Quanzeng J. Hailin Z. Wang F. Chen and J. Luo "Image Captioning with Semantic Attention" IEEE CVPR 2016.
44. J. Donahue L. Anne Hendricks S. Guadarrama M. Rohrbach S. Venugopalan K. Saenko et al. "Long-term recurrent convolutional networks for visual recognition and description" Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2625-2634 2015.
45. Y. Zhou and H. nanjia "Accelerated masked transformer for dense video captioning" Elsevier Neurocomputing 2021.
46. G. H.Wang D. ji and H. B. Zhang "Multi-feature fusion refine network for video captioning" journal of experimental & theoretical artificial intelligence 2021.
47. D. L. Chen and W. B. Dolan "Collecting highly parallel data for paraphrase evaluation" ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies vol. 1 pp. 190-200 2011.
48. G. A. Sigurdsson X. Wang A. Farhadi I. Laptev and A. Gupta "Hollywood in Homes : Crowdsourcing Data" pp. 1-17 Jul 2016.
49. R. Krishna K. Hata F. Ren L. Fei-Fei and J. C. Niebles "DenseCaptioning Events in Videos" Proceedings of the IEEE International Conference on Computer Vision vol. 2017-Octob pp. 706-715 2017.
50. E. Boron A. Erdem N. I. Cinbis E. Erdem and P. Madhyastha "Leveraging Auxiliary Image Descriptions For Dense Video Captioning" Pattern Recognition Letters 2021.
51. R. Luo G. Shakhnarovich S. Cohen and B. Price "Discriminability objective for training descriptive captions" Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition pp. 6964-6974 2018.