Image and Video Caption Generation

Harshitha Loganathan[1], Manju Bhashini. D[2],Grace Elizabeth[3] , Sangeetha Krishna[4], Salini. R [5]

Abstract

This consideration distinguishes different strategies for building encoder-decoder systems based on profound learning that gives full normal dialect caption of video arrangements. Recurrent Neural Network (RNNs) examines the data provided by the details in the image to produce a logical, sequential representation of the image. This study also provides evaluation criteria for determining the efficiency of video captioning models with different datasets used for image and video captions. Deep-learning based approaches are currently used for video processing, but this paper looks into the various methods for generating natural language.

Keywords: CNN, deep learning, LSTM, RNN, image captioning, video captioning, attention

1. Introduction IMAGE :

This inquire about proposes a framework that creates relevant depictions of objects in pictures to make a important articulation. Profound learning can be utilized to consequently supplant manual comments, decreasing human mistake and exertion. Real-world employments for captions produced from pictures incorporate helping the daze, permitting programmed, less costly labeling of millions of pictures transferred to the Web each day, providing proposals in altering programs, helping virtual colleagues, ordering pictures, helping the outwardly disabled, helping clients on social media, and numerous other normal dialect preparing applications. To prepare an picture caption era demonstrate, a considerable dataset with adequate get to to important information is required.

VIDEO:

Vedio subtitling is the technique for delivering an ordinary Lingo sentence for a given video.It requires various establishment ideas and recognizing their occasions inside the video,as well as unraveling the removed visual information into a conceivable and correct familiar lingo portrayal.The most difficulties in appreciating recordings and making normal english sentences are realizing the fine movement focal points of vedio substance conjointly the natural of various articles,Learning predominant portrayals of video between the video space and lingo space,and situating the development distinguished iniside the vedio. To resolve these issues,numerous ways have been suggested,counting siginificant learning-based and layout based methodogies.

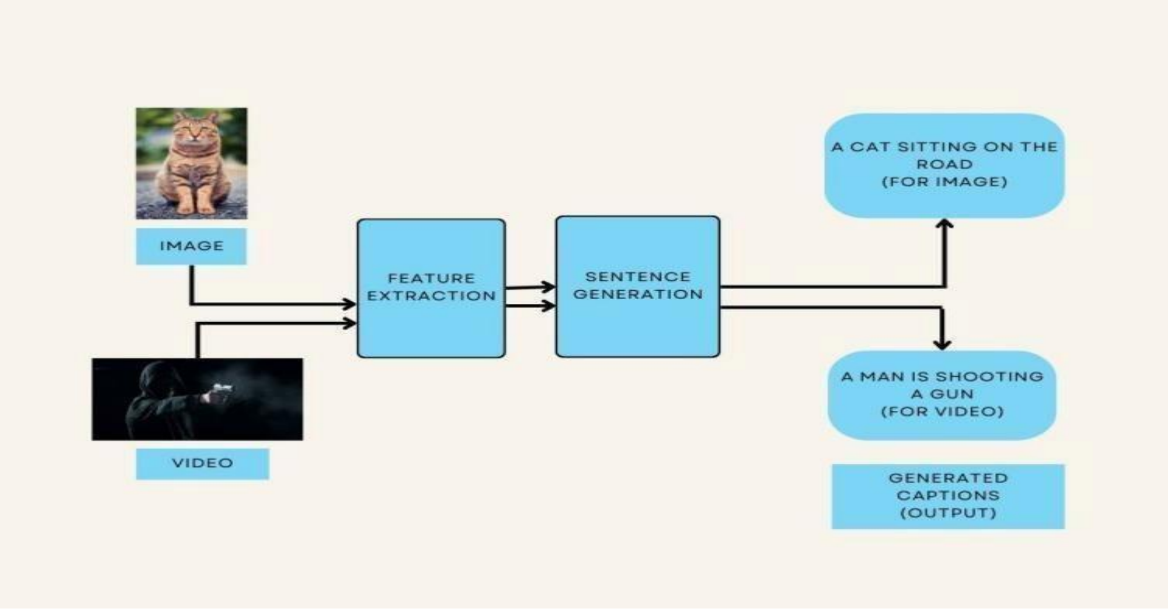


IMAGE
VIDEO
FEATURE EXTRACTION
SENTENCE GENERATION
A CAT SITTING ON THE ROAD (FOR IMAGE)
A MAN IS SHOOTING A GUN (FOR VIDEO)
GENERATED CAPTIONS (OUTPUT)

2. Literature Survey

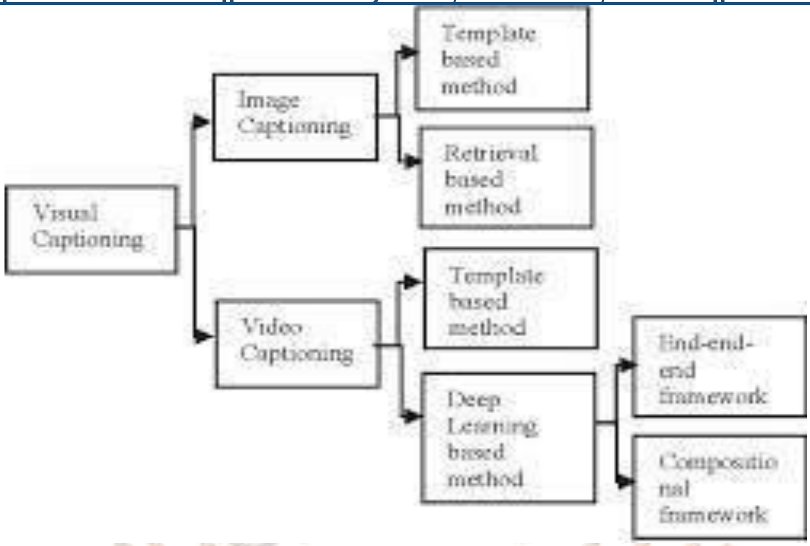
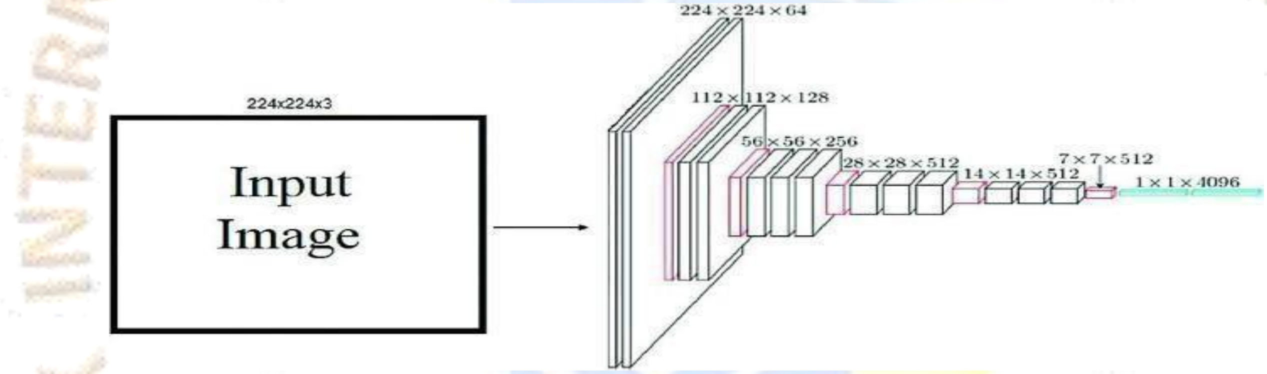


IMAGE CAPTIONING:

Recent interest in image captioning has been split into two groups: retrieval and template matching. In retrieval-based techniques, images are chosen from a library and fitted with the description of the test image. In template matching methods, items and activities are identified and compared to the template. O. Vinayals[54] given an end-to-end system to create the printed depictions for pictures by moving from the RNN encoder to the CNN encoder.

Z. Gan[55] built a semantic compositional organize utilizing semantic concepts, and Q.You[56] proposed a novel approach that combines top-down and bottom-up methodologies employing a semantic consideration show. The visual featuresare extricated employing a convolutional neural organize.



VIDEO CAPTIONING :

Natural language image descriptions have drawn a lot of interest, and this project is concentrating on video descriptions. the basic concept for describing the video content will make use of deep learning techniques. Convolutional neural networks (CNN) are used in the first phase to encode the vector representation of the video, and recurrent neural networks (RNN) are used in the second step to decode the vector into a textual description. This process produces natural language descriptions for video content. In several applications, including video indexing, language modeling, machine translation, and others, these deep learning networks considerably produce positive results.

For video captioning, N. Xu [57] suggested a dual stream- RNN model that is utilized to investigate and combine the hidden states of the semantic and visual streams. By utilizing the hidden states of vector representation and semantic concepts separately by using two modalities specific RNN called Attentive MultiGrained Encoder (AMGE), the proposed model improves the local feature learning process along with the global semantic feature, and it makes the video representation effective for producing video captions. To create textual descriptions, the DualStream RNN decoder fuses the two streams from AMGE. Utilizing both the past and back dispersions, J. Tune [59] proposed the Multi-modal Stochastic Repetitive Neural Arrange (MS-RNN) to produce various sentences for the same occasion. This approach was too utilized to resolve the issue.

The foremost present day procedure to capture transient structure is 3D CNN[58], which employments max and cruel pooling over each video outline to construct vector representation. To encode video information into a video representation and infuse it into a dialect demonstrate, a joint visual modeling methodology joining forward and in reverse LSTM and CNN is used.13 S. Venugopalan [60] recommended a brand-new structure for end-to-end sequencing that would give rundowns for brief movies.

To create sentences for photographs and recordings, L. Gao[61] proposed a progressive LSTM with versatile consideration (hLSTMat) engineering, which employments a spatial consideration or temporal consideration instrument to choose a particular range of the outline to seek for words that are related.

This study of the composing gives a complete idea of the various methodologies used to frame textural depictons for accounts.CNN and RNN frameworks are used to remember the course of action and give the formed depiction.Joining picture and video sources has gained ground the making of normal portrayals fot the video.Y.Xu[63] spread the world about Progressive VLAD Layer a SeqVLAD,which delivers a predominant portrayal of video.

By using CNN and RNN framework ,Y.Holder[62] suggested the LSTM unit with traded semantic characteristics (LSTM- TSA) framework to remove the semantic features from the films.

Shared GRU-RCN , an upgraded type of the Gated Redundant Unit of the Tedious Convolutional Sort out(RCN),was wanted to remember spatial and common undertaking.The proposed approaches in this concentrate astonishingly increase the efficiency of the video subtitling handle,however there are as yet various unsure issues that grant experts to focus on video subtitling and convey human- like protrayals.

3. Proposed System

There are typically two parts needed to automatically produce natural language phrases that describe an image or a video clip:an encoder and a decoder .Here , we go into detial about each parts architecture. A convolutional neural network is used by an encoder to extract objects and features from an image or video frame. A neural arrangement is required for the decoder to produce a normal sentence based on the available data.

A. Encoder

Pre-prepared models VGG 16,RESNET and Inception areused inside the encoder of the system to analyze the comes about achieved from every one of them.The encoder is used to remove the picture's idea vector,which depicts the pictures's substance. A definitive completely related layer's yield consolidates a gauge of(none,4096).The idea vector is ready through a thick framework work to diminish it from 4096 to 512 since the embedding estimation of the decoder is 512.The GRU units in each layer are given this thought vector as their starting state .The tanh capability is trailed by the thick guide capability.Tanh capability is a strategic sigmoid capability that has been scaled to give a result scope of -1 to 1.

The logistic sigmoid function is given as: Tanh function is given as:

$\tanh(x) = 2g(2x) - 1$

Pre-trained VGG model: Information about the full image is collected rather than classifying the image by stripping the softmax layer of the VGG model. The VGG model's hidden layers are made up of two convolutional layers, followed by alayer called max pooling that cuts the size in half. To deliver a one- dimensional yield that's provided to completely associated layers, this design is repeated three times and after that smoothed. After being contracted by the thick outline layer, the yield of the moment completely associated layer is utilized as the decoder's starting state for GRU layers.

B. Decoder

The decoder comprises of the tokenize, implanting layer, GRU layers, and thick layer. The tokenize layer changes over numbers tokens into numbers tokens, whereas the inserting layer turns numbers tokens into vectors of 128 floating-point numbers. The grouping vectors are cushioned to guarantee they are of the same length or greatest length of the grouping.

GRU's 3 gates are :

Forget gate, input gate, output gate.

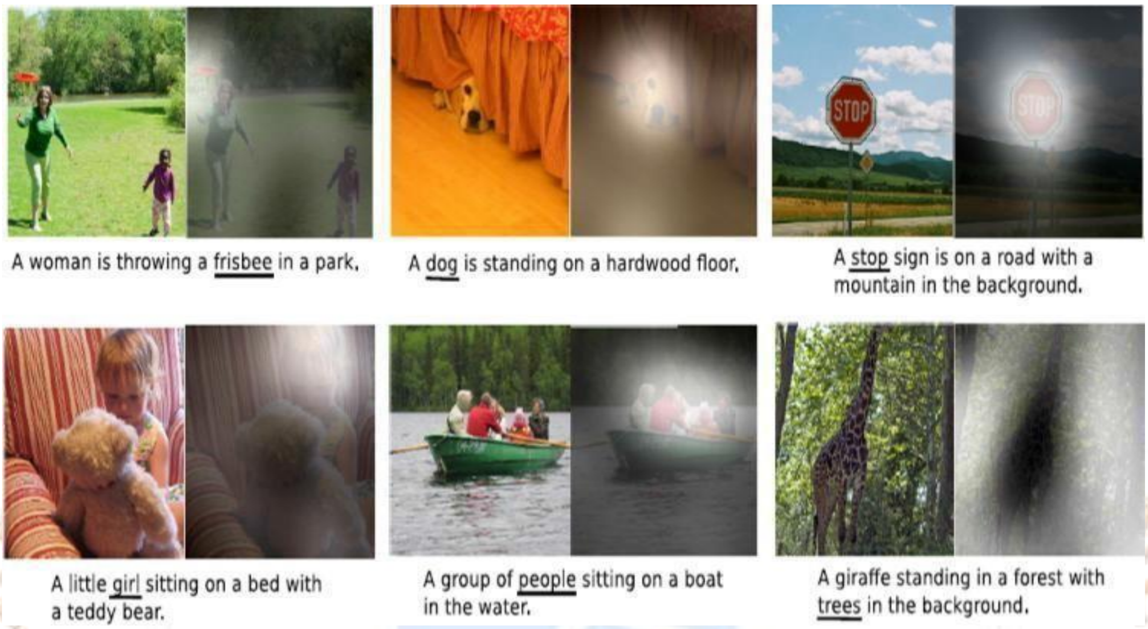
After passing through the token to the word instance of the tokenizer object, the dense layer receives the output of the GRU network and converts the sequences into integer tokens.

Steps per epoch =

total number of captions in the training dataset batch size

ATTENTION

The addition of attention, a method that helps the algorithm concentrate on important aspects of the image while avoiding redundant content, has significantly improved image captioning. Attention is the weighted sum of encoder outputs and is used to assign a weight to each image pixel using feature maps and a hidden state. Decoder focuses on most important portion of image. The attention weights are visualized to show what areas of the image the model is focusing on while producing a specific word.



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

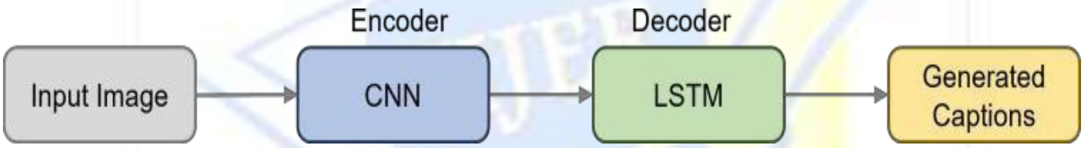
A stop sign is on a road with a mountain in the background.

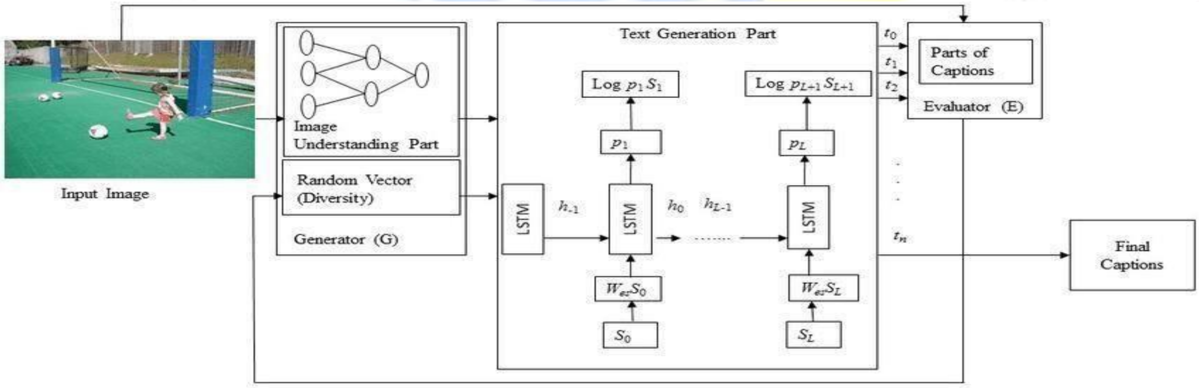
A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

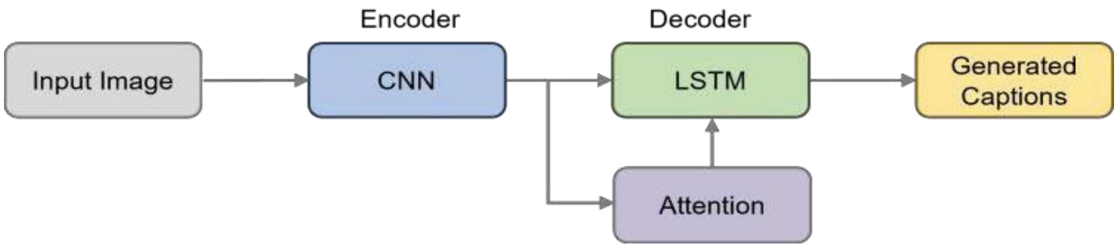
A giraffe standing in a forest with trees in the background.

CAPTIONING BY WITHOUT USING ATTENTION





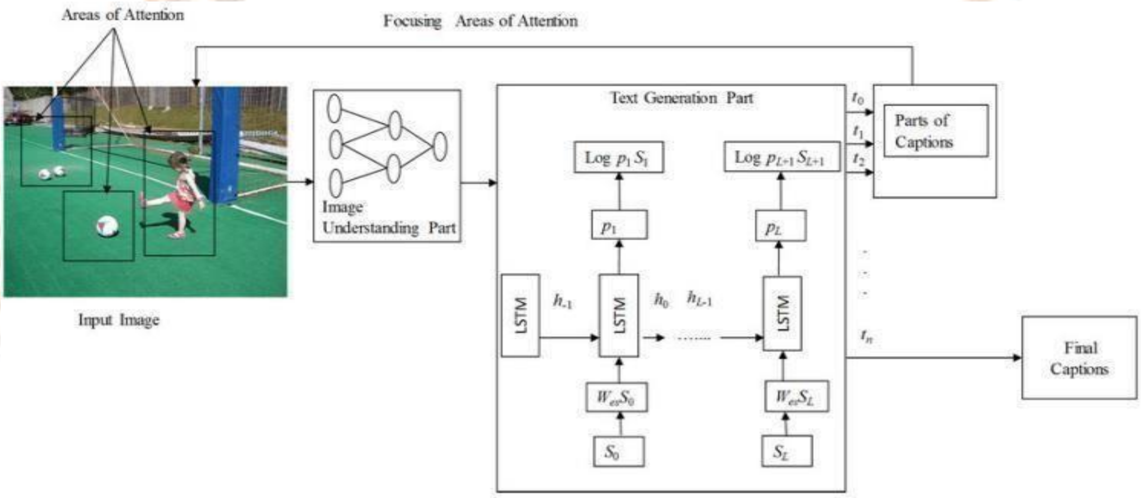
CAPTIONING BY USING ATTENTION



Captioning Methodologies

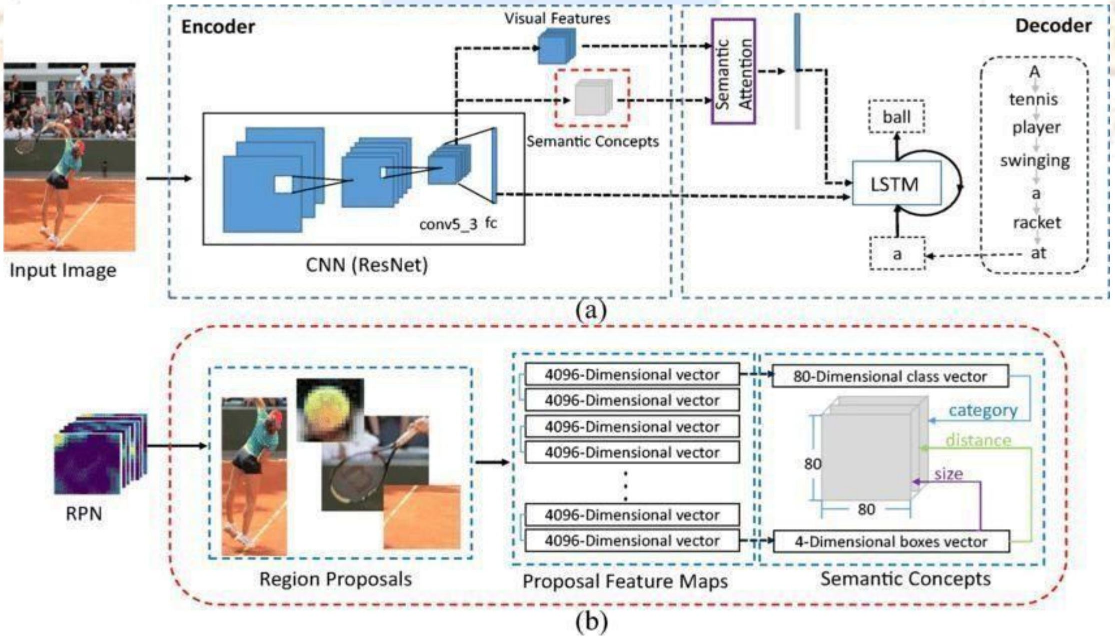
Convolutional neural networks: Memorizing thousands of things from a huge number of images requires demonstrators with tremendous learning power[3].Deep learning describes acomputer model with several processing layers arranged to store representationsof information in images[2],[18].One of the many applications of deep learning-based convolutional neural systems is image verification. Image confirmation is used formany visual tasks such as understanding the data in your images.Some well-known CNN models[2] based on dissenting opinions[1],[19],[20] and the revelation of nuclear fission [21] have been proposed in image and videocaptioning schemes to remove visual information used whenever possible.

Recurrent Neural Networks: The repetitive neural organize (RNN) and other arrangement models are regularly utilized in discourse acknowledgment, characteristic dialect handling,and other spaces [22]. Grouping models can be utilized to handle a assortment of directed learning issues, such as opinion examination, named substance recognizable proof, DNA grouping investigation, and machine interpretation [23].The gated repetitive unit (GRU), a gating strategy for RNNs, was created by Cho et al. in 2014 [23]. The vanishing slope issue influences the fundamental RNN approach (a trouble in preparing counterfeit neural systems). The vanishing angle issue is effectively illuminated with gated repetitive units. As a result, neural systems can presently record conditions over aimpressively more extensive run [22].



Long Short-Term Memory (LSTM): As a special RNNstructure, LSTM has been appeared in various tests to be solid and compelling for recreating long-range conditions. It is conceivable to utilize LSTM as a building square for perplexing plans.

A memory cell could be a complex unit found in long short- term memory. Each memory cell encompasses a settled self-connection shaped around a center direct unit [24]. Since LSTM incorporates three entryways (disregard, upgrade, and yield), it has generally been illustrated to be more powerful and successful than a ordinary RNN. Repetitive neural systems with long-range structures can be utilized to deliver complex arrangements [25], [26].



A. IMAGE CAPTIONING METHODOLOGIES

There are numerous approaches to captioning images. Deep neural networks (DNNs) replaced earlier techniques like retrieved-based [9] or template-based [17] models. Deep neural networks are the foundation of modern techniques. There are two steps involved in creating an automatic caption for an image.

The data must first be taken out of the image and entered into a feature vector. At this phase, deep learning models are used to recognize images. The second stage receives the feature vector after that. In the subsequent stage, caption creation, the retrieved information is described in grammatically sound natural language sentences. As a result, we divided DNN- based methodologies into the subcategories that each one usesbased on the basic framework. Here, a survey of current automatic picture captioning systems based on deep learningis covered.

Encoder-decoder models have driven to a breakthrough in captioning pictures and recordings, changing over a fixed- size highlight vector to a arrangement of words. This is the encoder-decoder pipeline paradigm that Kiros et al. [25]introduced. When employing the 19-layer Oxford convolution network, they achieved new records for results. The attention mechanism was developed to improve picture captioning by aligning portions of captions to visuals. Possibly the best descriptions yet have been generated [25]. One of the profound learning models, the considerationdemonstrate was propelled by one of the foremost interestinghighlights of the human visual framework. The attention-based approach picks up the capacity to concentrate ondifferent ranges of the picture. When an image contains a lotof clutter, this is essential.

Sadly,this can bring about a deficiency of information utilized for more extravagant and more nitty-gritty captioning.Utilizing the BLEU and METEOR measrements,Xu et al.provide cutting-edge execution .They showed that learned arrangements closely match human instinct and that learned consideration can be utilized to expand the interpretability of the model-age process.The utilization of visual consideration is an invigorating worldview for additional examination.It’s been quite a while since I’ve anticipated a book to such an extent.

The method combines rich semantic features from the image with an RNN that can selectively focus on them to extract more detailed information from an image. They experimented with their techniques on various datasets and then created a captioning system utilizing the LSTM network. The final 1024- dimensional convolution layer of the Google Net [2] CNN model is where the picture feature vector is created. In the input and output levels of their technique, they additionally use the RNN module.

They made an effort to improve the caption by integrating global and local information and utilizing a wealth of fine- grain visual semantic characteristics. The findings demonstrate that the algorithm regularly beats state-of-the-art methods using a variety of evaluation parameters. In the following study, we notice that Fu et al. [27] suggested an image caption system that takes advantage of the structural similarities between phrases and images. This system's abilityto coordinate the creation of captions and the shifting ofattention across different visual regions is one of its contributions. The addition of scene-specific contexts to LSTM, which modify language models for word production to suit particular scenario types, is another benefit. A first analysis and representation of an image with numerous visualregions from which visual information might be derived are done in that system. A scene vector, a global visual context taken from the entire image, also controls the neural network model. It makes a scene-specific language model choice based on intuition while producing text. They tested captions using several well-known datasets, including the MSCOCO, Flickr8K, and Flickr30K, to determine how they performed in the metrics. Performance is improved by either scene- specific contexts or region-based attention alone, but combining the two yields much greater gains. There are three primary parts to this approach. Input/output word embeddings make up the first and last components respectively. Nevertheless, their CNN-based technique uses masked convolutions while the middle component in the RNN example uses LSTM or GRU units. This element has a feed-forward design and no recurrent operations. Their CNN with attention (Attn) produced similar results. The activations of the conv-layer were also used to test an attention mechanism with attention parameters.

The CNN+Attn technique produced better results than the LSTM baseline, and ResNet features improved performance on the MSCOCO.

The image encoder in the retrieval and FC captioning models is Resnet-101. After being trained using the object and attribute annotations from Visual Genome, a Faster R-CNN with ResNet-101's output is used to extract the spatial features [47].

GRU-RNN is utilized to encode substance into captions, but wealthier and more moved sources of planning signals are required to advance caption generators' planning.

Anderson et al. [39] have proposed a combined bottom-up and top-down thought instrument that would enable thought to be calculated at the level of objects and other outstanding visual districts. Bottom-up thought is executed utilizing speedier R-CNN with ResNet-101 [2], a common appearance of a bottom-up thought component. The top-down componentgauges an thought scattering over the visual regions utilizinga task-specific setting. The weighted typical of picture highlights over all regions is at that point utilized to calculate the gone to incorporate vector. The bunch finished CIDEr, BLEU-4 scores of 117.9 and 36.9 on the MSCOCO dataset, setting a unused standard for the challenge. They additionally won the 2017 VQA Challenge.

Yao et alproposed .'s Chart Convolutional Frameworks too Long Short-Term Memory (GCN- LSTM) illustrate may be a dynamic arrange that prominently updates CIDEr-D execution on the COCO testing set while checking both semantic and spatial dissent affiliations into picture encoder.

Moreover, Cornia et al[50] .'s proposal for a Transformer- based architecture is available. The Meshed- Memory Transformer is a multi-layer architecture that uses permanent memory vectors to learn and encode prior information. It uses a learned gating system to weigh contributions from multiple levels at each stage, establishing a mesh communication structure between the encoder and decoder layers. This method sets a new standard for COCO and tops the online scoreboard.

The same year, Pan et al. [51] introduced the unified attention block, also known as the X-Linear attention block, which uses bilinear pooling to selectively take use of visual input or carry out multimodal reasoning.

X-Linear Attention Networks (X-LANs) are deep learningmodels that integrate X-Linear attention block(s) in a novel way to take advantage of higher-order intra- and intermodal interactions. The COCO benchmark trials demonstrate thattheir X-LAN achieves the best published CIDEr performanceto date of 132.0% on the COCO Karpathy test split. The Generative Adversarial Network (GAN) is a new frameworkfor estimating generative models using an adversarial process. Image generation has been done successfully using GAN. They can create natural images that are essentially indistinguishable from actual photographs [41], [42], [42],[48]. Utilizing Conditional Generative Ill-disposed Systems(CGAN), Dai et al. [43] presented a unused system that at the same time trains a generator to create portrayals that are subordinate on pictures and an evaluator to decide how well a depiction matches the visual substance. They utilized G-MLE, a generator prepared on MLE, and G-GAN, a comparative generator that's based on conditional GAN definitions. The chosen image encoder is VGG16. The ATTEND-GAN model was proposed in 2019 and uses both the developed attention- based caption generator and the adversarial training method on the SentiCap dataset to generate stylish captions that are similar to those of humans in two-stage architecture. The architecture makes advantage of spatial-visual data produced by the ResNet- 152 network and draws inspiration from the Wasserstein GAN (WGAN). Anderson et al. have used attention-based techniques to compare models on the MSCOCO dataset. Table 2 shows the outcomes of these techniques on the dataset, which suggests that Anderson et al. did well.

Their approach fared better than earlier works. The rationale is that it makes use of an attention mechanism thatconcentrates just on the image's pertinent objects. Also, we discovered that a technique's performance can change depending on the metrics, settings, and datasets used. Here, we attempted to assess them in light of the many techniques they have employed. To improve the accuracy of captioning the information in photos, however, there is still much workto be done in this area of ongoing study.

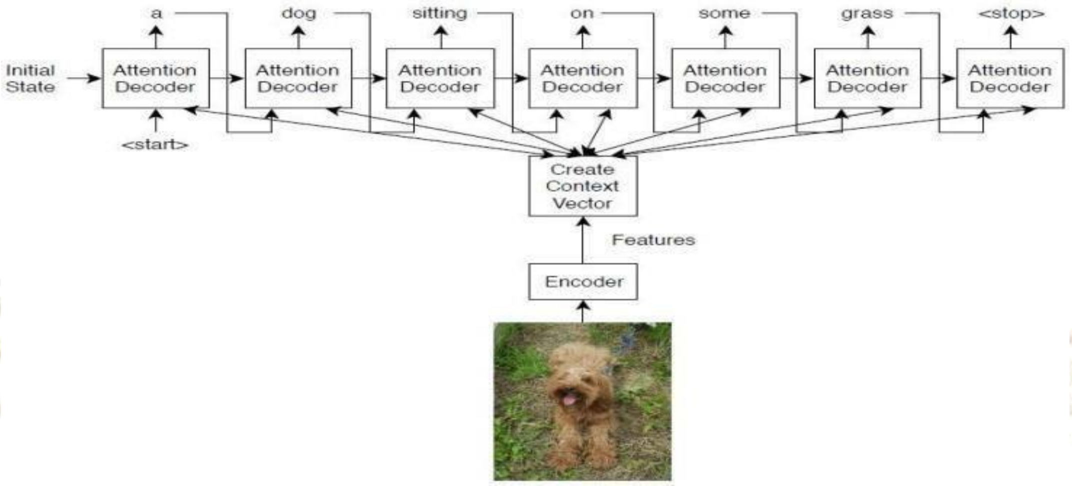


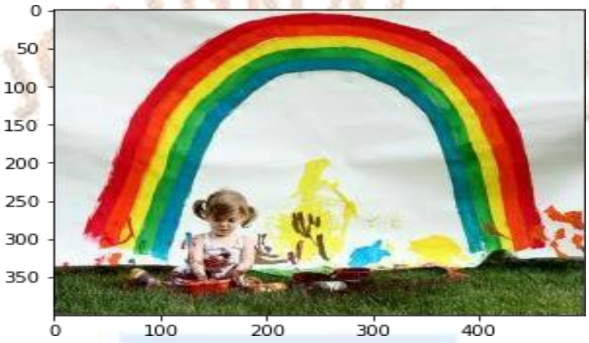
Image Captioning Datasets

A few datasets, including Flickr8K [11], Flickr9K [6], Flickr30k [6], [11], and Microsoft COCO [6], [7], are frequently used to assess and compare image captioning approaches

1) Flickr

More than 8000, 9000, and 30000 photographs, respectively, are included in the Flickr8K, 9K, and 30K databases. Amazon Mechanical Turk is used to annotate each image with five distinct words. The Flickr30k dataset primarily includes photographs of people participating in normal activities and events, while the Flickr8K dataset primarily includes images of people and animals. Five sentences are given for every picture [6], [11].

An example of an image and its five captions from Flicker8k dataset

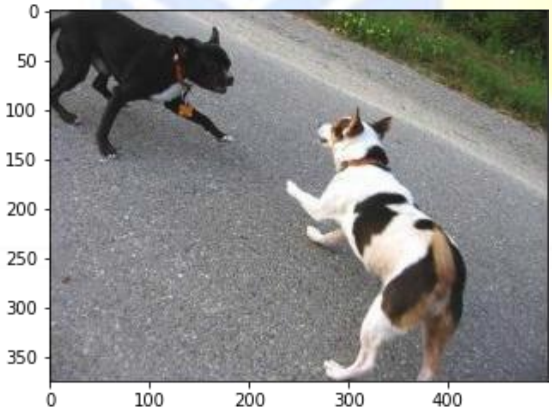


An illustration of a picture and its five subtitles fromFlicker8k dataset

-----------------------------ACTUAL--------------------------

a. Girl covered in paint sits in front of paintedrainbow with her hands in bowl
b. There is girl with pigtails sitting in front ofrainbow painting
c. Little girls is sitting in front of large paintedrainbow
d. Young girl with pigtails painting outside in thegrass
e. Small girl in the grass plays with fingerpaints infront of white canvas with rainbow on it
----------------------------PREDICTED----------------------

Little girl in pink dress is lying on the side of thegrass



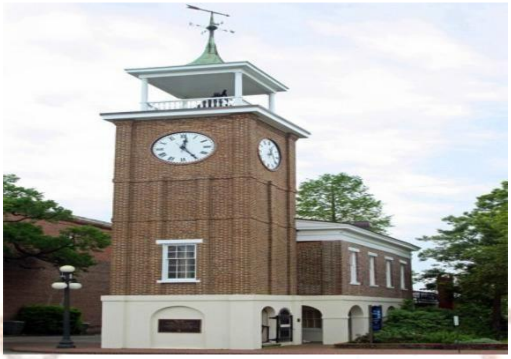
-----------------------------ACTUAL--------------------------

A. Black dog and spotted dog are fighting
B. Two dogs on pavement moving toward each other
C. Black dog and white dog with brown spots arestaring at each other in the street
D. Black dog and tri-colored dog playing with eachother on the road
E. Two dogs of different breeds looking at each otheron the road
------------------------PREDICTED--------------------------

Two dogs play with each other in the grass

2) COCO

Lin et al. [29] given a unused dataset for classifying and fragmenting common family things in their characteristic situations. The Microsoft Common Objects in Setting (MSCOCO) dataset incorporates 91 thing categories with more than 5,000 recognized illustrations, five captions for each picture, 2.5 million labeled events over 328k photos,and 91 question sorts [6], [7].



An illustration of a picture and its five subtitles from COCOdataset

a. A building with a balcony and a clock

b. A clock tower has a balcony around it

c. A very tall clock tower with a minty roof

d. The top of a tower with a balcony and clock

B. Video captioning methodologies

Portraying a film in plain dialect can be troublesome formachines, because it is troublesome to set up the commitments of the acknowledged dialect demonstrate and visual elements to the ultimate depiction, making it troublesome to classify models or calculations.

By utilizing picture captions to the video keyframes and a brief test of the outlines in between the keyframes, video captioning can be finished. The encoder-decoder structure forpicture captioning that was already displayed can too be utilized for video captioning. The foremost critical points of interest in this content are the two forms included in naturallycreating normal dialect sentences that portray video data. Knowing things is the primary step, taken after by extricatingthe entertainer, activity, and question of the activity from the video clip with a center on visual acknowledgment utilizing profound learning models. The video clip is given as a collection of outlines, alluded to as pictures, or outlines. As a result, each clip contains a few input image-based outlines. At that point, a common include vector is filled with the information from the clip that was extricated. The second step receives this vector as input. The moment arrange, caption era, maps the objects recognized within the past organize by portraying what is extricated in linguistically exact characteristic dialect sentences. Deep learning architectures are used for encoding and decoding.

Combining CNN and RNN models is one of the preeminent utilized significant learning plans for video captioning. Long-term Tedious Convolutional Frameworks (LRCNs), which combine convolutional layers and long-range common recursion and are end-to-end trainable, were proposed by Donahue et al. as a appear for visual affirmation and depiction. Three vision issues were taken into thought: activity affirmation, picture and video depiction, and picture delineation. The consider utilized LRCN, BLEU, and TACoS multilevel datasets to assess the picture building on the COCO and Flickr30k datasets. They utilized the BLEU-4metric to assess the video portrayal method, and utilized LSTM to appear the video as a variable-length input stream. In spite of the reality that the LSTM performed better than strategies based on real models, it was still not trainable in a persistent way [12].
Able to observe that Venugopalan et al. [13] utilized the CNN and LSTM models combined S2VT methodology (a sequence-to-sequence approach for a video- to-text). A course of action of traces are encoded by the S2VTdesigning, which at that point deciphers them into a sentence. The YouTube dataset, MPII-MD, and M-VAD were utilized to compare their appear. They overviewed the execution by separating the machine-generated delineations with human ones utilizing METEOR and BLEU. The revelations illustrate a perceivable alter in how people judge linguistic use. As lingo alone makes a basic influence, it is imperative to concentrate on both verbal and visual components to form overwhelming depictions. Following techniques have joined thought components and a comparable framework [14].

Profound learning has significantly beaten more seasoned models and the lion's share of procedures. The lion's share of strategies endeavored to extricate one line from a video clip that contained as it were one striking event. Conflictingly, dense captioning, which was created by Krishna et al. [15], points to perceiving a few occasions that happen in a video bycollaboratively restricting worldly thoughts of intrigued and after that depicting them in characteristic dialect. This strategy included a brand-new captioning module that employments chronicled setting from later and past occasions to depict everything at once. The show was created utilizing the ActivityNet Captions dataset. When depicting expressionsin recordings, the ActivityNet captions move from being object-centric in pictures to being action-centric. But its expecting arrangement isn't the single-sentence producing issue.

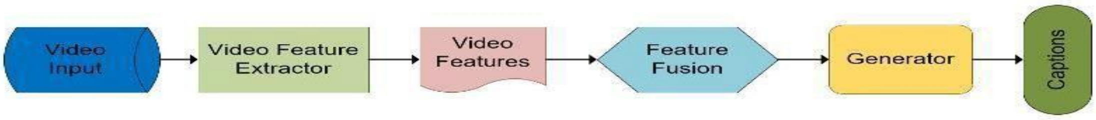
Zhou et al. [52] model is the one that uses dense video captioning the most similar to Krishna et alThis model provides an end-to-end transformer paradigm for dense video captioning, with an encoder and two decoders. The event proposal is converted into a differentiable mask by the captioning decoder's use of a masking network. This model also uses a self-attention mechanism, and deep reinforcement networks are an area of study for video description.

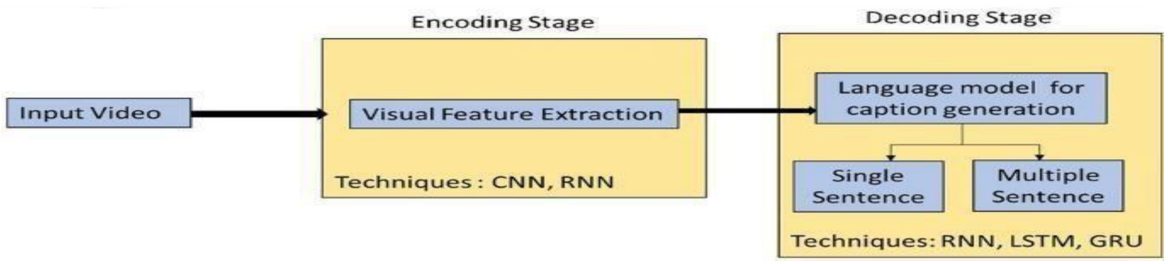
The Hierarchical Reinforcement Learning approach was introduced by Wang et al. [16] and tries to produce one or more sentences given a series of one or more continuous actions. On every metric, the innovative HRL technique performed better than all the previous methods. As a result, the HRL agent needs to explore more of the attention space and make use of features from other modalities.

In arrange to successfully abbreviate recovery time, Ding et al.

[40] displayed novel approaches for the utilization ofamplified video division in 2019. A brand-new super- frame division and excess video outline distinguishing proof based on spatiotemporal intrigued focuses (STIPs) are combined to move forward the viability of video division. The foremost captivating area of the filterable, long motion picture is along these lines extricated utilizing superframe segmentation. Keyframes from the foremost significant sections are changed over into video captions with the help of saliency discovery and an LSTM variation organization. The consideration component is at that point utilized to supplement the conventional LSTM with extra pivotal information. This approach is assessed utilizing the BLEU, Meteor, and Rouge for the picture captioning component and is benchmarked utilizing the VideoSet dataset. However, when it comes to assignments like question location at moo resolutions or for small objects, the dialect show still performs gravely when compared to individuals. Mun et al[44] .'s approach to thick video captioning is comparable to the work of Krishna et al[15] .'s in that it models the transient reliance between occasions in a video and employments the visual and etymological setting from prior events to tell a compelling story. In arranging to make a caption utilizing RNN frameworks, the exceedingly connected occasions that make up a scene were sent into a successive captioning organize. For a single check, they utilized the Single-Stream Transient Activity show to accumulate a few suggestions. The proposed strategy accomplishes exceptional METEOR comes about on the ActivityNet Captions dataset. In Sung Park et al[45] .'s system, antagonistic systems were used to make a discriminator that assessed the movie's visual centrality, dialect differences, familiarity, and coherence overexpressions by presenting GAN into DL. Creating more precise, captivating, and influential multi-sentence video depictions is made less difficult by GAN. The discriminator should evaluate the depictions that the generator (G) came up with for a certain video. They propose building D from three distinctive discriminators, each centering on one of the previously mentioned objectives. They classify this plan as a blended discriminator.

We evaluated a few techniques in this part, arranging them chronologically depending on the most recent techniques CNN, LSTM, and attention-based have employed. Since they employ various methodologies, procedures, and datasets, we have no intention of comparing them. Yet, the algorithms, large datasets, assigned captions, and hardware improvements all contribute to the performance and accuracy improvement every year.





Video Captioning Datasets

Methods for video captioning are evaluated using a variety ofdatasets. Just a few of them are mentioned here, and they are grouped into five categories based on the topic of the videos: People, Open Subjects, Social Media, Cookery, and Cinema.

People: The Charades dataset [30] is composed of 15 scenarios with a total of 40 objects and 30 actions. The app Charades was created by Sigurdsson et al. and contains 9,848 motion pictures (7,985 for preparing and 1,863 for testing), with each video enduring an normal of 30 seconds. 157 actions are described by 66,500 annotations in the dataset.

Additionally, it offers 27,847 descriptions for every video.

Open Subject: The **1,970** Youtube accounts inside the Microsoft Video Depiction dataset (Chen and Dolan,2011)have sentences that have been incorporated by people (600 films for testing,100 accounts for underwriting and 1200 accounts for arranging).The normal runtime of every video inside the MSVD assortment is somewhere between 10 and 25 seconds.Krishna et al [15] given the ActivityNet inscriptions dataset, a basic norm for thick subtitling conditions.It has 20k development images that have lasted for 849 hours and close to 100K Portrayals.By Xu et al,the MSR-VIT dataset was given (representing MSR-Video to Content).A paid video showing Motors 257 most notable requests every one which joins 118 account,were consolidated to make this.The 41.2 long periods of 10k web video cuts that MSR-VIT gives with an incorporation of up to 200k clasp sentence sets , cover twenty explicit classifications.

Social media: A dataset called VideoStory [31] is used to convey the narratives behind social media videos. It has 123k phrases in 20k videos totaling 396 hours of video.

Cooking: There are 65 fine-grained culinary operations within the Cooking dataset [32] from the Max Board Established for Informatics (MPII). Each of the 44 movies within the collection has an normal runtime of 600 seconds. Literarily Commented on Cooking Scenes (TACoS), which give coherent printed depictions for high-quality recordings and cover 26 fine-grained cooking movements in 127 clips, were to begin with presented by Regnery et al. [4]. Zhou et al.

[33] created the dataset YouCook2 in 2018 using cooking motion pictures to create a noteworthy sum of procedural division information that was transiently restricted and characterized. 2000 motion pictures with a combined length of 176 hours are equally conveyed among 89 dishes from Africa, America, Asia, and Europe.

Movie: The huge scle motion picture portrait challenge dataset(LSMDC,Rohrbach et a,2017)[5] contains a decoded and adjusted audio and script information set and 128,118 sets of adjusted clips from 200 movies .Included(plus about 150 hours of video).LSMDC is initially built on the freely authored MPII-MD and M-VAD datasets which are provided collectively in this work.The MPII Motion picture portrayal(MPII-MD)[34] dataset contains an identical corpus of over 68,000 words and video clips from 94 HD movies.The Montreal video comment record(M- VAD)[14]contains a set of 92 DVDs and over 84.6 hours of related films

C. IMAGE AND VIDEO CAPTIONING EVALUATION

The metrics BLEU, METEOR, CIDEr, and others are used to evaluate captions [6]–[8]. These criteria are frequently usedto compare the various image and video captioning technologies and vary in how closely they resemble human judgment [30].

1)BLEU

Bilingual Appraisal Understudy could be a technique for evaluating automatic machine interpretation that's precision- based, encompasses a moo negligible taken a toll per run, and a solid relationship with human evaluation[6], [35]. For candidate sentences of the reference sentences, BLEU has numerous n-grams- based varieties.

2)METEOR

The consensus-based picture delineation assessment[8](CIDer)is an objective comparison of machine- time methods based on their similarity to human language.It was originally developed for recording image captioning tasks but has gradually become widely used for video captioning strategies.

3)ROUGE

The quality of a diagram is overviewed utilizing Recall-Oriented Understudy for Gisting Assessment [36], which contrasts it with other rundowns composed by individuals. Comparative to BLEU, ROUGE offers distinctive n-grams- based assortments.

4)SPICE

Anderson et al.[37] presented a new semantic evaluation metric ,the semantic propositional image caption rating,that measures how effectively image captions recover objects,qualities and relatonships between them .This interacts more with human judgements of semantic quality than current coarse-grained estimates.

5)WMD

Word Mover's Distance [38] calculates how different two texts are from one another. Because of this, this metric is less sensitive to word order or synonym switching than BLUE, ROUGE, and CIDEr, but it still has a good correlation with human assessments, just like CIDEr and METEOR.

4. Result and analysis

A. Results Of Image Captioning Without Attention:

As shown in Figure 1, the image uses a VGG16 displayas an encoder with 16 secure levels and GRU orchestration(using 3 GRU levels).Shown in the image of the endorsement set provided. Images from the MS- COCO dataset as the planning dataset and dictionary measurements of 10,000 odd words.Figure 2 shows theimage of the test photo and the expected caption.

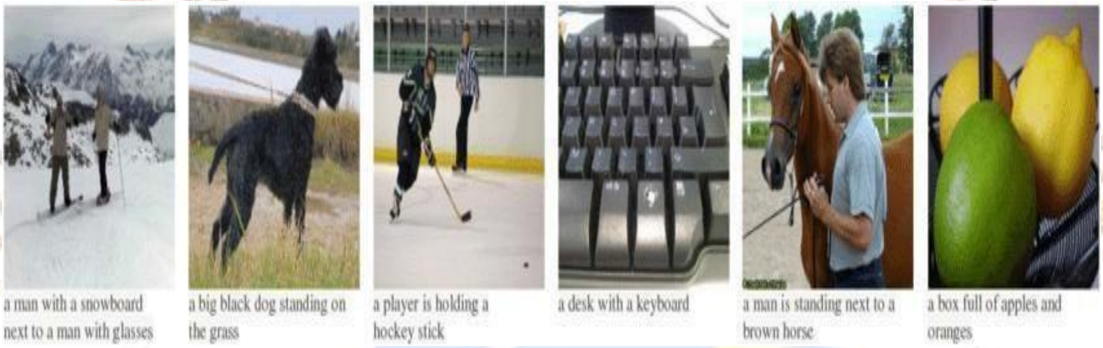

a man with a snowboard next to a man with glasses
a big black dog standing on the grass
a player is holding a hockey stick
a desk with a keyboard
a man is standing next to a brown horse
a box full of apples and oranges

Figure 1

PREDICTION FOR IMAGES IN VALIDATION SET

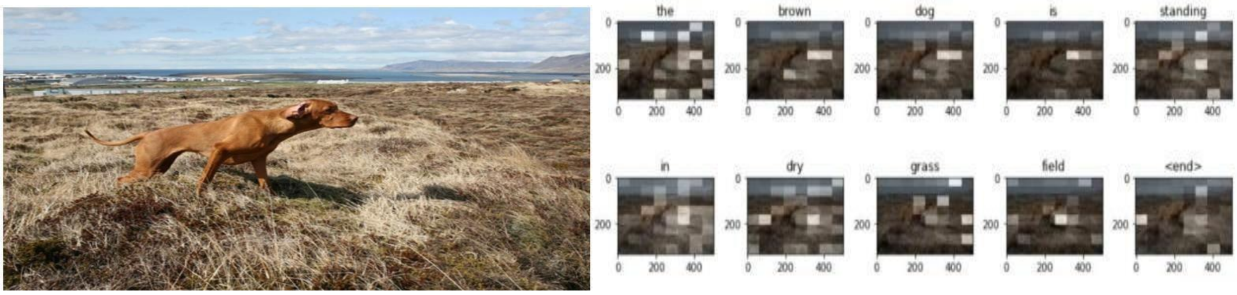
A small bird is sitting on a tree branch

–

Figure2

PREDICTION FOR RANDOM IMAGE

A. Results Of Image Captioning With Attention:



REAL CAPTION: BROWN DOG IN FIELD

PREDICTION CAPTION: THE BROWN DOG ISSTANDING IN DRY GRASSFIELD

5. Conclusion and future work

This paper centers on the algorithmic cover between picture and video captioning, and the methods utilized to attain videocaptioning utilizing picture captioning procedures as building pieces. A few models have been put forward to make captions for still photographs and brief movies, but they have limitations that make them wrong and restricted in real-world applications. The method of making video captions is seen as a combination of image caption outlines.

Since of this, the reason of this article isn't to supply a comprehensive examination of picture and video captioning, but or maybe to supply a brief depiction of how their calculations associated. Moreover, this article as it were considered profound learning-based calculations.

Deep-learning models utilized for captioning pictures andrecordings can be challenging to compare due to diverse picture datasets, parameters, classification methods, preprocessing settings, structure combinations, and other components. Our ponder centered on the zones where both procedures covered, and found that there are various employments for a framework of real-time, accurate, and dependable video and picture captioning.

Machines learn to locate and help in progressing our vision, permitting people to utilize them in other ways. Frameworks for captioning pictures and recordings can be an critical figure of assistive advances for those with hearing or vision disabilities.

Captions can be utilized as meta-data by search motors and proposal frameworks, growing the usefulness of the search engine.

The next conclusions were reached:

Different models require varying amounts of iterations for thesame dataset.

The number of epochs required to solve the "Image captioning" problem decreases as the network depth increases.There is a trade-off between the number of hidden layers and the execution time.

Several models and different numbers of epochs were used tomeasure the average value of various metrics on the same dataset.

The accuracy with which various measures assess system performance in comparison to captions created by humans was assessed.

Future Research Direction and Broader Impact:

The mistake in creating captions for pictures (or) pictures andrecordings covers up the imperative points of interest in this content. With advance improvement, more exact captions canbe produced by combination and preparing of pictures, soundand video. With the existing innovation, exceptionally brief recordings can be captioned because they require a parcel of captioning in them. The other era of GPUs and unequivocal strategy parallelization must be utilized to form captioning for longer recordings. Making an arrangement that would empower clients to inquire about video captions at different degrees of detail would be an incredible opportunity. Unmistakable subtitles inview of various clarifications can be made for a similar video and this chief issue can be settled by checking out German considerations and making the technique all the more natural.

6.References

[1] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767. [Online]. Available:http://arxiv.org/abs/1804.02767

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 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

[8] 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.

[9] 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.

[10] 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

[11] 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.

[12] 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
[13] 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
[14] 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
[15] 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.
[16] 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.
[17] 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.
[18] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521,no. 7553, p. 436, 2015
[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun.2014, pp. 580–587.
[20] R. Girshick, ‘‘Fast R-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.
[21] B. Romera-Paredes and P. H. S. Torr, ‘‘Recurrent instance segmentation,’’ in Proc. Eur. Conf. Comput. Vis.
Cham, Switzerland: Springer, 2016, pp. 312–329.
[22] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, ‘‘Empirical evaluation of gated recurrent neural networks on sequence modeling,’’ 2014, arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555
[23] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ‘‘Learning phrase representations using RNN encoder-decoder for statistical machine translation,’’ 2014, arXiv:1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078
[24] S. Hochreiter and J. Schmidhuber, ‘‘Long short-term memory,’’ Neural Comput., vol. 9, no. 8, pp. 1735– 1780, 1997.
[25] R. Kiros, R. Salakhutdinov, and R. S. Zemel, ‘‘Unifying visualsemantic embeddings with multimodal neural language models,’’ 2014, arXiv:1411.2539. [Online]. Available: http://arxiv.org/abs/1411.2539
[26] Y. Wu et al., ‘‘Google’s neural machine translation system: Bridging the gap between human and machine translation,’’ 2016, arXiv:1609.08144. [Online]. Available: https://arxiv.org/abs/1609.08144
[27] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, ‘‘Aligning where to see and what to tell: Image captioning with region-based attention and scenespecific contexts,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 39,no. 12, pp. 2321–2334, Dec. 2017
[28] J. Aneja, A. Deshpande, and A. G. Schwing, ‘‘Convolutional image captioning,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5561–5570.
[29] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, ‘‘Microsoft COCO: Common objects in context,’’ in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 740–755
[30] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, ‘‘Hollywood in homes: Crowd sourcing data collection for activity understanding,’’ in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 510–526.
[31] S. Gella, M. Lewis, and M. Rohrbach, ‘‘A dataset for telling the stories of social media videos,’’ in Proc.
Conf. Empirical Methods Natural Lang.Process., 2018, pp. 968–974.
[32] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, ‘‘A database for fine grained activity detection of cooking activities,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 1194–1201
[33] L. Zhou, C. Xu, and J. J. Corso, ‘‘Towards automatic learning of procedures from Web instructional videos,’’ in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 1–9.
[34] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, ‘‘A dataset for movie description,’’ in Proc.
IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2015, pp. 3202–3212.
[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ‘‘BLEU: A method for automatic evaluation of machine translation,’’ in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2001, pp. 311–318.
[36] C. Lin, ‘‘Rouge: A package for automatic evaluation of summaries,’’ in Proc. Text Summarization Branches Out, 2004, pp. 74–81
[37] P. Anderson, B. Fernando, M. Johnson, and S. Gould, ‘‘Spice: Semantic propositional image caption evaluation,’’ in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 382–398.
[38] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, ‘‘From word embeddings to document distances,’’ in Proc. Int. Conf. Mach. Learn., 2015, pp. 957–966.
[39] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, ‘‘Bottom-up and top- down attention for image captioning and visual question answering,’’ in Proc. IEEE/CVF Conf. Comput. Vis.
Pattern Recognit., Jun. 2018, pp. 6077–6086

[40] S. Ding, S. Qu, Y. Xi, and S. Wan, ‘‘A long video caption generation algorithm for big video data retrieval,’’ Future Gener. Comput. Syst., vol.93, pp. 583–595, Apr. 2019.
[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial nets,’’ in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.
[42] T. Iqbal and H. Ali, ‘‘Generative adversarial network for medical images (MI-GAN),’’ J. Med. Syst., vol. 42, no. 11, p. 231, Nov. 2018.
[43] B. Dai, S. Fidler, R. Urtasun, and D. Lin, ‘‘Towards diverse and natural image descriptions via a conditional GAN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2970–2979.
[44] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, ‘‘Streamlined dense video captioning,’’ in Proc.
IEEE/CVF Conf. Comput. Vis. Pattern Recognit.(CVPR), Jun. 2019, pp. 6588–6597.
[45] J. S. Park, M. Rohrbach, T. Darrell, and A. Rohrbach, ‘‘Adversarial inference for multi-sentence video description,’’ in Proc. IEEE/CVFConf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019.

[46] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, ‘‘Discriminability objective for training descriptive captions,’’ in Proc. IEEE/CVF Conf.

[47] Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6964–6974. R. Krishna,
Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, ‘‘Visual genome: connecting language and vision using crowdsourced dense image annotations,’’ Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017.

[48] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, ‘‘Image captioning with generative adversarial network,’’ in Proc. Int. Conf.Comput. Sci. Comput. Intell. (CSCI), Dec. 2019, pp. 272–275

[49] S. Ding, S. Qu, Y. Xi, and S. Wan, ‘‘Stimulus-driven and conceptdriven analysis for image caption generation,’’ Neurocomputing,vol. 398, pp. 520–530, Jul. 2020.

[50] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, ‘‘Meshed- memory transformer for image captioning,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10578– 10587.

[51] Y. Pan, T. Yao, Y. Li, and T. Mei, ‘‘X-linear attention networks for image captioning,’’ in Proc.
IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10971–10980

[52] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C.Xiong, ‘‘End-to-end dense video captioning with masked transformer,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8739–8748.

[53] T. Yao, Y. Pan, Y. Li, and T. Mei, ‘‘Exploring visual relationship for image captioning,’’ in Proc. Eur. Conf.
Comput. Vis. (ECCV), 2018, pp. 684–699.

[54] O. Vinyals, A. Toshev, S. Bengio, D. Erhan,“ Show and Tell: A Neural Image Caption Generator”, IEEE CVPR, pp. 3156-3164, 2015.

[55] Z. Gany, C. Gan, X. Hez, Y. Puy, K. Tranz, J. Gaoz, L. Cariny, L. Dengz, “Semantic Compositional Networks for Visual Captioning”, arXiv:1611.08002v2, 28 Mar 2017.

[56] Quanzeng You1, Hailin Jin2, Zhaowen Wang2, Chen Fang2, and Jiebo Luo “Image Captioningwith Semantic Attention”, IEEE CVPR, 2016.

[57] N. Xu, A. Liu, Y. wong, Y Zhang, W. Nie, Y. Su, M. Kankanhalli, “Dual-Stream Recurrent Neural Network forVideo Captioning”, accepted in IEEE Transactions on Circuit andsystems for video technology, Mar. 2018.

[58] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, X. Li, “Describing Video with Attention-Based Bidirectional LSTM”, accepted in IEEE Transactions on Cybernetics, 2019.

[59] J. Song, Y. Guo, L. Gao, X. Li, H.T. Shen, “From eterministic to Generative: Multimodal Stochastic RNNs for ideo Captioning” in IEEE Transactions on Neural Networks and Learning Systems, 2018.

[60] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, “Sequence to sequence- video to text”, In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 2015.

[61] L. Gao, X. Li, J. Song and H. T. Shen, Hierarchical LSTMs with Adaptive Attention for Visual Captioning”, accepted in IEEE Journal of Latex Class Files, Vol. 14, No. 8, August 2015.

[62] Y. Pan, T. Yao, H. Li and T. Mei, “Video captioning with Transferred Semantic Attributes”, IEEE CVPR, pp. 984-992,2017.

[63] Y. Xu, Y. Han, R. Hong, Q. Tian, “SequentialVideo VLAD: Training the Aggregation Locally and Temporally” in IEEE Transactions on Image Processing, Vol. 27, No. 10, October 2018.