CREATE CAPTION BY EXTRACTING FEATURES FROM IMAGE AND VIDEO USING DEEP LEARNING MODEL

GOUTAM DUTTA

PhD. Research Scholar,

Amity University, Kolkata – 700135.

Abstract : The images, videos captured using various devices and images available in various sources like internet, news articles, social media does not have proper description through which a human can understand those without observing closely. It is also very difficult to write syntactically correct sentence describing the images for a large set of images and videos manually. Sometimes the description may vary based on each individual person's perception, mood and interpretation at the time of interpreting the image in the form of sentence. This may lead to forming a caption with inaccurate, imperfect, error prone description for the images or videos which is not acceptable in some scenarios where the accuracy is the primary criteria for the other systems to react or act upon.

The aim here is to generate a sentence for an image by identifying the features by using deep learning techniques and also for the video frames to generate sentence using the same model used for the generation of image captioning. The feature extraction from the images is done by using pre-trained deep learning model available like VGG16, Densenet121, InceptionV3 etc.

The vocabulary was built by processing the description of the images that is available as a part of the dataset and removing the stop words. The deep learning model is built and trained to generate syntactically correct sentence by looking into the extracted features and predicting the next set of sequence of words and interpreting those sequence of word into sentence. The experiments done on several datasets on the model and measured the accuracy of the model by measuring the fluency of the language it learns solely from image descriptions. The accuracy of model and smoothness or command of language model learns from image training on various sources of data from Flickr8k, Flickr30k, Video frames etc. BLEU score is generated to compare the model performance. To extract the feature from the images, deep learning neural network technique convolution neuron network (CNN) is used and recurrent neural network (RNN), such as a Long Short-Term Memory network (LSTM) is used to generate the sequence of word.

The generation of cation for the video, key-frames are extracted from the video by passing through the key frame extraction framework built inside the application. The sequence of key-frames extracted from video is fed into the same image captioning model that was trained for generating the caption for the images. The captions are generated from the frames extracted from the video by using the predefined pre-trained model by feeding the images.

Keywords: Video Image Caption, Image Caption using Deep Learning Techniques, Image Segmentation, Key Frame Extraction.

I. INTRODUCTION

Video captioning or video image captioning or identifying a scene from the video there are various methodology, technique developed using various technology from the past few years and those were not very effective and suffer one or the other issue with respect to accuracy, time and correctness of the sentences. The pain of the video captioning is that one needs to observe the video multiple times to come up with a caption manually. Also if someone wants to refer a specific scene from a video, to pin point the exact video frame manually one has to view the complete lengthy video multiple times and have to find the specific frame of the video. It's difficult and also a time consuming frustrating process of searching the same by repeatedly going through the video. It's easy if there is a tool or some kind of software which can generate the sequence of frame and correctly provide a caption of the frame automatically. The objective can be achieve without wasting long time and reducing the manual effort if machine learning technique helps to generate the sentences from the video automatically. Anyone who is looking for a specific frame from the video can easily go through the caption and correctly point at the video frame by simply using the software application developed using the deep leaning techniques. The objective can be achieve by using the modern day technology, infrastructure and the concept of using the Deep Learning, to process the video by consuming minimal amount of time by the system and correctly pointing out the video frame by generating the sentences out of it.

Deep learning methods have demonstrated the capability to generate the state of the art results on caption generation problems. The generated output from the deep learning methods is expected to describe in a single sentence what is shown in the image i.e. the features of the image, the objects present in the image, their properties, the actions being performed in the image and the interaction between the objects, etc. But to replicate this behaviour to automate the same using the an artificial system is a huge task, as compared to any other image processing problem and hence the use of complex and advanced techniques such as Deep learning techniques have been used to solve the complicated task like this for the purpose of this study.

1.1 Theoretical framework

In the thesis image caption model is created to extract the features from the standard set of images based on Flickr8k and Flick30k dataset. The captions are loaded, cleaned, processed and stored into a file for further processing. The same file is used to create vocabulary of the words to be used to generate the captions and kept it in a dictionary. The model which is used for generating caption from the static images is used to generate captions for the video frames. The video frames are generated by using key-frame extraction technique. The audio is not considered as a source of input to the model for the formation of sentences using the natural language.

One of the challenges was building huge vocabulary to generate accurate meaningful sentences and the other one was converting video images in meaningful sequence of frames. Considering these two challenges, the observation are scanned on the most of the recent work and found that most of those are either not scalable or have a huge dataset dependencies. To solve the issue, sequence of key frames of the video is extracted and passed to the encoder to extract the feature, and then it is sent to the decoder to predict the next word and by combining all the generated word a sentence is formed. Due to the huge video space, text datasets and processing time, the model performs poor in case of generalization. So there is a need for powerful image captioning model, infrastructure to be in place to work well in generating sentences.

The entire thesis work can be summarized as per the below diagram. From the below figure one can easily understand that a convolutional neural network (CNN) is used to encode the images and a recurrent neural network (RNN), such as a Long Short-Term Memory network (LSTM), is used to encode the text sequence generated to generate the next word using the sequence generator. In case of video files, the input video is fed into a key frame extraction algorithm, which gives output of sequence of key frames. Those key frames are fed into an image captioning model used earlier and trained as part of generating caption from the static image files. The efficiency of the output has been measured and compared by using Bi Lingual Evaluation Understudy (BLEU) for both images to caption generation as well as video to caption generation.

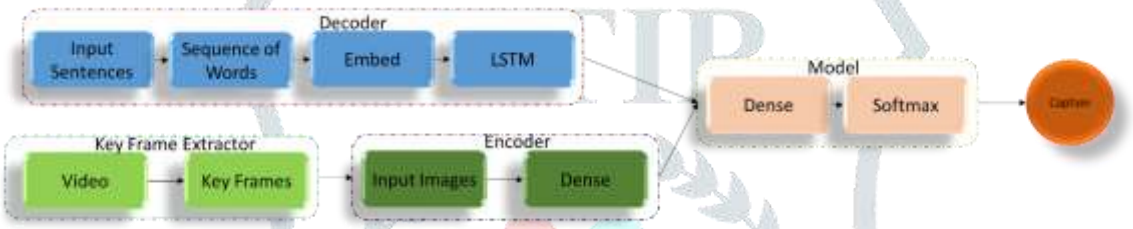

Figure 4.1.1 Flow diagram of the implementation approach

II. LITERATURE REVIEW

A Comprehensive Survey [1] of Deep Learning for Image Processing was made on deep leaning methods for image captioning, where various deep learning methods were studied and evaluation criteria's were observed and compared. Image captioning requires to acknowledge the important objects, their attributes and their relationships in a picture. It also needs to generate syntactically and semantically correct sentences. Deep learning-based techniques are built, tested and made to enhance the capability of handling the complexities and challenges of image captioning. In this paper it is observed that it only talks about the theory and no practical study and modelling done and also this paper does not implement the video captioning by using the pre-trained model.

A Neural Image Caption Generation with Weighted Training and Reference [2] method based on the encoder-decoder framework, named Reference based Long Short Term Memory (R-LSTM), aiming to lead the model to generate more descriptive sentence form the given image by introducing reference information into it. The framework additionally maximize the score between the generated captions by the captioning model and the reference information extracted from the neighbouring images of the next image, this may reduce the misrecognition problem. A novel evaluation function by combining the likelihood with the consensus score is used to fix misrecognition and make the generated sentences more natural sounding. The Caption Generation using Weighted Training weights were assigned to the words according to the correlation between words and images and scores were given to generate the sentence. In this paper it is observed that it only used limited dataset and does not implemented by using different model. This paper does not implement the video captioning by using the pre-trained model.

Image Caption Generation Using Deep Learning Technique [3] using Encoder Decoder model to generate sentences for generating novel descriptions from images. In this model Convolutional Neural Network (CNN as encoder) is used for feature extraction from image as well as Recurrent Neural Network (RNN as decoder) is used for sentence generation. The Image caption generation is grouped into three categories. The first category consists of template based methods and the priority is given to detect objects, actions, scenes and attributes. The second category consists of transfer based caption generation methods. In this approach fetches visually similar images and then the captions of these images are used for query image. The third category neural network-based methods come from the recent advantages in machine translation. In this the goal is to convert an image into sentence which explains it rather than translating a sentence from a source language into a required format. Finally the descriptions or captions obtained from the model are categorized. In this paper only limited dataset for the experimentation and it has not been implemented by using different machine learning pre-trained model and video captioning by extracting frames or any unsupervised learning technique.

A Hierarchical Recurrent Neural Network [4] is used decomposing paragraphs into their corresponding sentences. In the thesis the model was developed that decomposes both images and paragraphs into their constituent parts and by detecting semantic regions in the images and using a hierarchical recurrent neural network is used to provide a reason about language. Dense captioning methodology was used as captioning approach and introduced the task of describing images with long, descriptive

paragraphs. It was presented a hierarchical approach for generation that leverages the compositional structure of both images and language. Region-level knowledge used to transfer the feature into paragraph captioning and the paragraph was generated by combining multiple captions for the same image. This paper does not implement the caption generation by using different pre-trained model and video captioning.

A dense captions was used for transforming the problem of complex image retrieval into a dense captioning and scene graph matching issue by using structured language descriptions for retrieval.

The Image retrieval by dense caption reasoning [5] address the issues of separating the complex image retrieval. The use of reasoning image dense captions by transforming the problem of complex image retrieval into a dense captioning and scene graph matching issue by using structured language descriptions for retrieval has been observed in the study. The experimental results on a novel proposed large-scale content-based image retrieval dataset demonstrate the rationality and effectiveness of the method. A dense caption reasoning strategy is used for image retrieval, which involves two stages of it. One is the dense caption generation, and second is the scene graph construction and reasoning. A novel method for scene graph matching is presented and a novel CBIR dataset is proposed to use for large-scale content-based image retrieval. It also covers the searching images by caption reasoning.

Situational Awareness from Social Media Photographs Using Automated Image Captioning [6] paper that was studied, argues that one can take advantage of the information that is spread across the various social networks like Twitter by analysing the messages posted. In order to provide a realistic solution for the post that was made on those sites can be useful for decision makers during particular events by using situational awareness methodology. Through the generation of English captions assigned to geo-referenced photographs extracted from tweets, and by using these captions to infer important features of those photographs, social monitoring approaches based on the images can produce prompt contextual information that helps in disaster monitoring and response. This paper reported on experiments that show the potential of social monitoring approaches based on image captioning. A pre-existing algorithm for image captioning to identify word features from Twitter images were used. Considered two image captioning algorithms used, namely the Microsoft Computer Vision API (MAPI) and Neural Talk by Stanford Vision Lab. The image captioning algorithm named Show and Tell is used to classify the images as relevant or non-relevant to the disaster.

In the Learning CNN-LSTM Architectures for Image Caption Generation [7], there are two main approaches to Image Captioning, one is bottom-up and another is top-down. Main challenges were in the field of Image Captioning is overfitting the training data. In CNN-LSTM architecture, modelled after the NIC architecture, top-down approaches were considered by using deep convolutional neural network to generate a vectorised representation of an image and then using Long Short-Term Memory (LSTM) caption were generated. Implemented a generative CNN-LSTM model and overfitting is alleviated and using hyper parameter tuning using dropout and number of LSTM layers. Deep convolutional neural network is used to generate a vectorised representation of an image and finally Long-Short-Term Memory (LSTM) network used to generates captions.

A parallel-fusion RNN-LSTM architecture for image caption generation [8], by combining the advantages of simple RNN and LSTM the thesis was conducted. In the thesis it was presented a novel parallel-fusion RNN-LSTM architecture, which obtains better results than a dominated one and improves the efficiency as well. The proposed approach that is presented here, divides the hidden units of RNN into several same-size parts, and lets them work in parallel. Then, it was merge their outputs with corresponding ratios to generate final results. The models were based on deep convolutional networks and recurrent neural networks and a novel parallel-fusion RNN-LSTM architecture is used. Training the model was done using the NeuralTalk1 platform by using Flickr8k dataset.

A Neural Image Caption Generator [9], is a generative model, which is based on a deep learning recurrent architecture. This combines the recent advancement in computer vision technology and machine translation and is used to generate natural sentences describing an image by predicting words. The model is trained to maximize the likelihood of the target description sentence for the training image given as input to the model. The various experiments on several datasets shows the accuracy of the model and the fluency of the languages that it learns solely from image descriptions. A generative model based on the deep learning recurrent architecture that utilizes the recent advances in computer vision and machine learning translation technology. BLEU-1 score is used to measure the improvements qualitatively and quantitatively on accuracy. A neural and probabilistic framework is used to generate descriptions from images.

Automated Image Captioning with ConvNets and Recurrent Nets [10], in this thesis, the input dataset of images and 5 sets of sentence descriptions were collected. During the training stage, the images are fed into the model as input to RNN and RNN is used to predict the sequence of words in the form of sentence. During the prediction stage, a pre-held set of images is passed to RNN and RNN generates the sentence by extracting features by one by one word. Convolutional Neural Networks is used to extract the feature from the image and ranking model is used to detect the class in the image. Based on the ranking and the extracted feature caption were generated for each image.

III. RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes methodology used, sample of the study, data and sources of data, artificial intelligence and machine learning models used. The details are as follows;

3.1 Population and Sample

In the thesis image caption model is created to extract the features from the standard set of images based on Flickr8k and Flick30k dataset. The captions are loaded, cleaned, processed and stored into a file for further processing. The same file is used to create vocabulary of the words to be used to generate the captions and kept it in a dictionary. The model which is used for generating caption from the static images is used to generate captions for the video frames. The video frames are generated by using key-frame extraction technique.

3.2 Data and Sources of Data

Flickr8k contains 8,091 images collected from Flickr. Most of these images depict humans performing various activities. Among 8,091 images 6,000 is used for training, 1,000 is used for validation and testing.

Flickr30k contains 31,783 images collected from Flickr. Most of these images depict humans performing various activities. Each image is paired with 5 cross-sourced captions.

MSCOCO It is the largest image captioning dataset, containing 82,783 training images 40,504 validation images and 40,775 testing images. This dataset contains images that have multiple objects in the context of complex scenes. Each image has 5 human annotated captions and evaluated the model on entire validation set.

UCF-101 It is used for extracting the frames from the video using key-frames extraction and used to generate the caption for the image frames. UCF-101 contains 13,421 videos which converted into 90,065 images. Also other random video like Tom and Jerry was also used to test the model working. The two videos got extracted to 208 and 185 key frame images.

3.3 Preparation of text data

3.3.1 Data understanding

Each image is labelled with at least five predefined captions for Flickr8k, Flickr30k, MS COCO dataset. There is no standardized split on the datasets, so 80-20 split has been followed. Every line contains the <image name>#i <caption>, where 0≤i≤4. The image name is the name of the image, caption number (0 to 4) is the caption for each image and the actual caption. The created dictionary contains the image name as key and list of 5 corresponding captions as values. The video data does not contain any caption list. So there is no processing of the text for the video images.

```
1000268201_693b08cb0e.jpg#0     A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1     A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2     A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3     A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4     A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0     A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1     A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2     A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3     Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4     Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0     A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl
.
1002674143_1b742ab4b8.jpg#1     A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2     A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow
on it .
1002674143_1b742ab4b8.jpg#3     There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4     Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0     A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1     A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2     a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3     A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4     man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0     A man in an orange hat starring at something .
1007129816_e794419615.jpg#1     A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2     A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3     A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4     The man with pierced ears is wearing glasses and an orange hat .
```

Figure 3.1 Data in a file

3.3.2 Data cleaning

The descriptions for each images has been extracted and performed some basic cleaning activity, like lower casing all the words, removing specials tokens, removing punctuations, removing words that are one character and words with numbers. After cleaning the descriptions the vocabulary has been created and then transformed into dictionary where image name is the key and descriptions are the values. The below image shows after the sentences got cleaned and stored in a file.

```
1000268201_693b08cb0e child in pink dress is climbing up set of stairs in an entry way
1000268201_693b08cb0e girl going into wooden building
1000268201_693b08cb0e little girl climbing into wooden playhouse
1000268201_693b08cb0e little girl climbing the stairs to her playhouse
1000268201_693b08cb0e little girl in pink dress going into wooden cabin
1001773457_577c3a7d70 black dog and spotted dog are fighting
1001773457_577c3a7d70 black dog and tricolored dog playing with each other on the road
1001773457_577c3a7d70 black dog and white dog with brown spots are staring at each other in the street
1001773457_577c3a7d70 two dogs of different breeds looking at each other on the road
1001773457_577c3a7d70 two dogs on pavement moving toward each other
1002674143_1b742ab4b8 little girl covered in paint sits in front of painted rainbow with her hands in bowl
1002674143_1b742ab4b8 little girl is sitting in front of large painted rainbow
1002674143_1b742ab4b8 small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it
1002674143_1b742ab4b8 there is girl with pigtails sitting in front of rainbow painting
1002674143_1b742ab4b8 young girl with pigtails painting outside in the grass
1003163366_44323f5815 man lays on bench while his dog sits by him
1003163366_44323f5815 man lays on the bench to which white dog is also tied
1003163366_44323f5815 man sleeping on bench outside with white and black dog sitting next to him
1003163366_44323f5815 shirtless man lies on park bench with his dog
1003163366_44323f5815 man laying on bench holding leash of dog sitting on ground
1007129816_e794419615 man in an orange hat starring at something
1007129816_e794419615 man wears an orange hat and glasses
1007129816_e794419615 man with gauges and glasses is wearing blitz hat
1007129816_e794419615 man with glasses is wearing beer can crocheted hat
1007129816_e794419615 the man with pierced ears is wearing glasses and an orange hat
```

Figure 3.2 Cleaned data

3.4 Preparation of image data

Images are the input to the model and the input images are converted in the form of vector and then that is fed as input to the model. The image is converted into a fixed sized vector which then fed as input to the neural network. The extraction of image features is done by using various model like VGG16, DenseNet121, InceptionV3 etc. After the extraction of the features is done then it is saved on the disk using pickle file. The below images is displaying the extracted features list by using VGG16 on Flickr8k dataset.

```
>109671650_f7bbc297fa.jpg
>1095980313_3c94799968.jpg
>1095476286_87d4f8664e.jpg
>1095580424_76f0aa8a3e.jpg
>1095590286_c654f7e5a9.jpg
>1096097967_ac305887b4.jpg
>1096395242_fc69f0ae5a.jpg
>1096165011_cc5eb16aa6.jpg
>109738763_90541ef30d.jpg
>109823397_e35154645f.jpg
>1105959054_9c3a738096.jpg
>110595925_f3395c8bd6.jpg
>109823395_6fb423a90f.jpg
>109823394_83fcb735e1.jpg
>109738916_236dc456ac.jpg
>1107471216_4336c9b328.jpg
>1104133405_c04a00707f.jpg
>1107246521_d16a476380.jpg
>1100214449_d10861e633.jpg
>1112212364_0c48235fc2.jpg
Extracted Features: 8101
Image Features: 8101
```

Figure 3.3 Extracted feature

3.5 Processing the caption

In this section how the captions are predicted has been described. The caption generations are the main objective of the thesis that is finally predicted in the process of caption generation. So during the training period, captions will be target variables that the model is learning to predict. The prediction of caption happens word by word. By encoding each word into a fixed sized vector is achieved, in this case the maximum caption length. The output word represents the probability distribution of the expected word across the whole dictionary of possible words.

3.5.1 Data Generator

The most important part of this application is generating the captions by using some kind of generator function which is responsible for implementing the same. The responsibility of the generator is to generate the result when it is called. The generator loops runs over and keeps yielding batches of input-output pairs and step parameter that allows to tune how many images of input-output pairs to generate for each batch. The data generator works as follows:

First, the image vector and the first word as input and try to predict the second word.

Second, image vector and the first two words as input and try to predict the third word. The actual English text of the caption is not passed, rather sequence of indices where each index represents a unique word is passed and zero padding is added at the end of each sequence. The target word sequence is predicted based on the input sequence.

The data generator will process the image and it's corresponding caption as follows:

For Example1: "startseq a little girl climbing into a wooden playhouse endseq"

For Example2 "startseq a girl going into a wooden building endseq"

The vocabulary that will be built based on the above sentences could be as follows:

Vocabulary = {little, girl, climbing, going, into, wooden, playhouse, building}

After doing the indexing the above vocabulary may look like as below:

Index = {1: 'a, 2: 'little', 3: 'girl', 4: 'climbing', 5: 'into', 6: 'wooden', 7: 'playhouse', 8: 'going', 9: 'building', 10: 'startseq', 11: 'endseq' }

The data matrix will then look as follows of one of the images and captions:

Table 3.1 Feature vector extraction logic

	Image feature vector	Caption	Target word
1	Input_Image_1	startseq	a
2	Input_Image_1	startseq a	girl
3	Input_Image_1	startseq a girl	going
4	Input_Image_1	startseq a girl going	into
5	Input_Image_1	startseq a girl going into	wooden
6	Input_Image_1	startseq a girl going into wooden	building
7	Input_Image_1	startseq a girl going into wooden building	endseq

The data matrix will then look as follows after replacing the indices:

Table 3.2 Feature vector indexing logic

	Image feature vector	Caption sequence	Target word sequence
1	Input_Image_1	[10]	1
2	Input_Image_1	[10,1]	3
3	Input_Image_1	[10,1,3]	9
4	Input_Image_1	[10,1,3,9]	6
5	Input_Image_1	[10,1,3,9,6]	1
6	Input_Image_1	[10,1,3,9,6,1]	9
7	Input_Image_1	[10,1,3,9,6,1,9]	11

The data matrix will then look as follows after padding:

Table 3.3 Feature vector sequencing logic

	Image feature vector	Caption sequence	Target word sequence
1	Input_Image_1	[10,0,0 …. , 0]	1
2	Input_Image_1	[10,1, 0,0 ….,0]	3
3	Input_Image_1	[10,1,3,0,0 …., 0]	9
4	Input_Image_1	[10,1,3,9,0,0 …..,0]	6
5	Input_Image_1	[10,1,3,9,6,0,0 …….,0]	1
6	Input_Image_1	[10,1,3,9,6,1,0,0 …….,0]	9
7	Input_Image_1	[10,1,3,9,6,1,9,0,0 …..,0]	11

3.5.2 Sequence Generator

This is one of the most important part of this application in caption generation. The responsibility of the sequence generator is to generate the sequence of indices from the description text. Each index represents a unique word and it is passed by adding padding at the end of each sequence. The target word sequence is predicted based on the input sequence and it is returned to the data generator for predicting the next word.

3.6 Deep learning model

This section describes the various image captioning model used in this thesis. The model consists of two parts, one is image extractor in the form of vector and other is the language processor in the form of sequence. The representation of the image vector comes through convolution neural network (CNN). In this case the pre-processed models are used to extract the features from the images and the captions are predicted out by this model by feeding the extracted features into dense layer. The language processor with a recurrent neural network (RNN), takes summary of previous words, to generate next word. This layer handles the text input followed by LSTM layer. The LSTM output is interpreted by dense layer one output at a time. Combining this two creates a merged model for generating text from the image. The output from the feature extractor and the output from the sequence processor comes out to be as output of a fixed length vector that is the length of a maximum sequence. These are concatenated together and processed by an LSTM and dense layer before a final prediction is made. The model predicts a probability distribution across the vocabulary. A softmax activation function is used and a categorical cross entropy loss function is minimized while fitting the network.

GlobalMaxPooling2D layer is used in this model to summarize the presence of each and every feature in an image. The image and feature map produce the same output when down sampled with maximum pooling.

Dense layer is a fully connected layer, meaning a linear operation in which every neurons in a layer are connected to every input to the next layer by a weight.

The RepeatVector layer is used like an adapter to fit the encoder and decoder parts of the network together. It is configured to repeat the fixed length vector one time for each time step in the output sequence.

Embedding layer is used on text data to capture the relationship in language that are very difficult to capture. It requires the input data to be integer encoded, so that each word is represented by integer.

LSTM layer is used to process the data passing on information as it propagates forward by remembering it longer time. The sequence of hidden state outputs are accessed when predicting a sequence of outputs with a dense output layer.

TimeDistributed layer is used to apply the same layer to several inputs and it produce one output per input to get the result in time. So basically it apply the same transformation for a list of input data.

Adam is used to as optimizer which is an extension to stochastic gradient descent and categorical cross entropy is used to calculate the loss function.

Below diagram represents the model summary:

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 7, 7, 512)	0	
input_2 (InputLayer)	(None, 34)	0	
global_max_pooling2d_1 (GlobalM	(None, 512)	0	input_1[0][0]
embedding_1 (Embedding)	(None, 34, 50)	1488450	input_2[0][0]
dense_1 (Dense)	(None, 128)	65664	global_max_pooling2d_1[0][0]
lstm_1 (LSTM)	(None, 34, 256)	314368	embedding_1[0][0]
repeat_vector_1 (RepeatVector)	(None, 34, 128)	0	dense_1[0][0]
time_distributed_1 (TimeDistrib	(None, 34, 128)	32896	lstm_1[0][0]
concatenate_1 (Concatenate)	(None, 34, 256)	0	repeat_vector_1[0][0] time_distributed_1[0][0]
lstm_2 (LSTM)	(None, 500)	1514000	concatenate_1[0][0]
dense_3 (Dense)	(None, 500)	250500	lstm_2[0][0]
dense_4 (Dense)	(None, 29769)	14914269	dense_3[0][0]

Total params: 18,580,147
Trainable params: 18,580,147
Non-trainable params: 0

Figure 3.4 Model summary for VGG16

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 7, 7, 1024)	0	
input_2 (InputLayer)	(None, 34)	0	
global_max_pooling2d_1 (GlobalM	(None, 1024)	0	input_1[0][0]
embedding_1 (Embedding)	(None, 34, 50)	1488450	input_2[0][0]
dense_1 (Dense)	(None, 128)	131200	global_max_pooling2d_1[0][0]
lstm_1 (LSTM)	(None, 34, 256)	314368	embedding_1[0][0]
repeat_vector_1 (RepeatVector)	(None, 34, 128)	0	dense_1[0][0]
time_distributed_1 (TimeDistrib	(None, 34, 128)	32896	lstm_1[0][0]
concatenate_1 (Concatenate)	(None, 34, 256)	0	repeat_vector_1[0][0] time_distributed_1[0][0]
lstm_2 (LSTM)	(None, 500)	1514000	concatenate_1[0][0]
dense_3 (Dense)	(None, 500)	250500	lstm_2[0][0]
dense_4 (Dense)	(None, 29769)	14914269	dense_3[0][0]

Total params: 18,645,683
Trainable params: 18,645,683
Non-trainable params: 0

Figure 3.5 Model summary for DenseNet121

Layer (type)	Output Shape	Param #	Connected to
input_6 (InputLayer)	(None, 5, 5, 2048)	0	
input_7 (InputLayer)	(None, 28)	0	
global_max_pooling2d_3 (GlobalM	(None, 2048)	0	input_6[0][0]
embedding_3 (Embedding)	(None, 28, 50)	25050	input_7[0][0]
dense_9 (Dense)	(None, 128)	262272	global_max_pooling2d_3[0][0]
lstm_5 (LSTM)	(None, 28, 256)	314368	embedding_3[0][0]
repeat_vector_3 (RepeatVector)	(None, 28, 128)	0	dense_9[0][0]
time_distributed_3 (TimeDistrib	(None, 28, 128)	32896	lstm_5[0][0]
concatenate_5 (Concatenate)	(None, 28, 256)	0	repeat_vector_3[0][0] time_distributed_3[0][0]
lstm_6 (LSTM)	(None, 500)	1514000	concatenate_5[0][0]
dense_11 (Dense)	(None, 500)	250500	lstm_6[0][0]
dense_12 (Dense)	(None, 501)	251001	dense_11[0][0]

Total params: 2,650,087
Trainable params: 2,650,087
Non-trainable params: 0

Figure 3.6 Model summary for InceptionV3

Below diagram represents the visualised structure of the network layers:

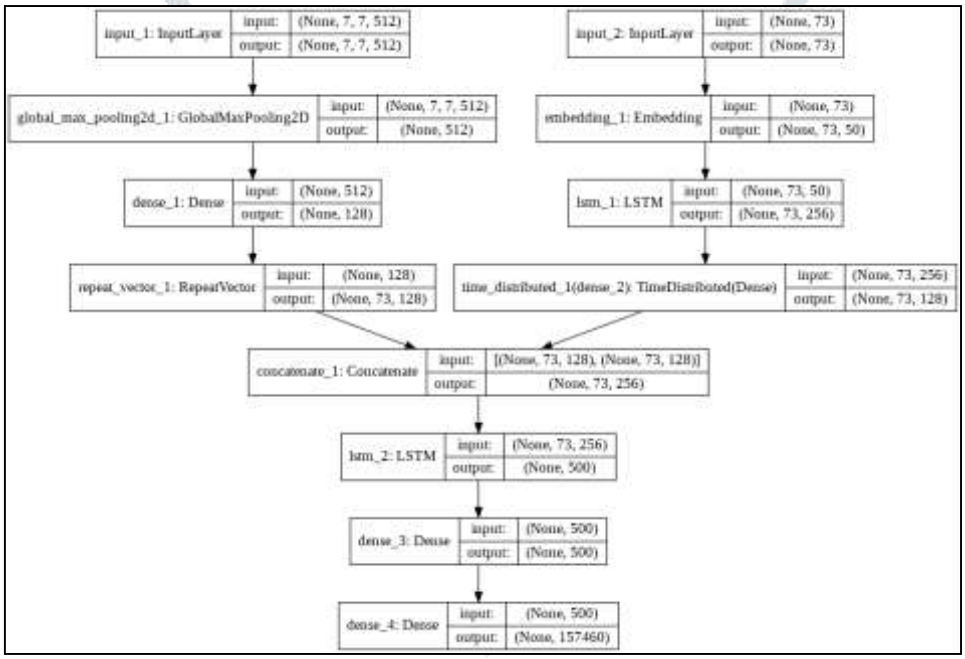

Figure 3.7 Model network layer structure

There are 8 layers that has been designed in the model to extract the feature as well as predicting the sequence of words and forming sentence out of it to fit it as caption to the input image. The below section explains the above captioning model used.

The image feature extractor expects input image features to be a vector of 7, 7, 512 elements. These are fed into GlobalMaxPooling2D and then into Dense layer to produce 128 element representation of the image. Then the input is fed to RepeatVector which produces output of 73, 128.

The sequence processor expects input sequences with a pre-defined length (73 words) which are fed into a 50 dimensionality of Embedding layer that uses a mask to ignore padded values. This is followed by an LSTM layer with 256 memory units.

The final model merges the vectors from both input models using an addition operation by concatenating output from the feature extractor and sequence processor. This is then fed to a Dense 500 neuron layer LSTM and 500 unit Dense followed by a 500 neuron Dense at the end of the network. The final output of dense layer that makes a Softmax prediction over the entire output vocabulary for the next word in the sequence.

3.7 Key-frame extraction sequence

A video is represented as multiple frames represented by f1, f2, f3, ……, fN. The calculated difference in low-level visual information between the two consecutive frames are denoted by d1, d2, d3, ……, dN, where di = abs (fi - fi-1). If the difference between two consecutive frame di is greater than some threshold value, the frame will consider as key-frame. In this thesis the considered standard deviation of differences is σd = SD (d1, d2, ……, dN) and that set the threshold as n * σd, where n depends on

video. If value of n is small the output will generate more number of key-frames, and for larger values of n the output will generate small number of key-frames. The implementation for extracting the key frames from the video openCV library is used to extract feature from video files.

IV. RESULTS AND DISCUSSION

4.1 Captioning results

This section covers the image captioning output from the various sources of images as well as videos. The results are shown after taking sample results from the development environment. This section also shows the extracted key-frames from the video sources based on the sigma value.

4.2 Key-frame extraction sample results

Key-frame extraction technique is applied on a video of 2 minutes 59 seconds. The total number of frames in the video is 186. By choosing n = 6 got 6 key-frames out of those 186 frames which is shown in Fig. From the figure it is observed that those 6 frames are sufficient to describe the video.

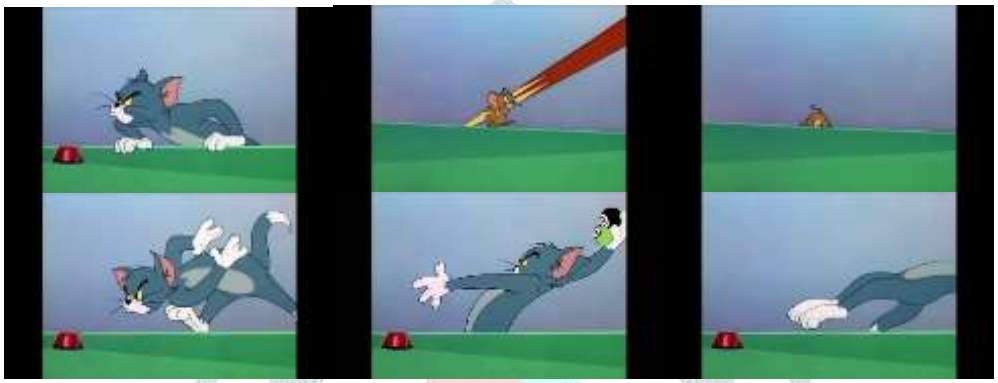

Figure 4.1 Visualization of key frames for n = 6

These 6 key-frames will be fed to the image captioning model, to generate the sentences by using the model generated using VGG16, DenseNet121, and InceptionV3 etc.

4.3 Image and video frame caption

In the thesis the relevant dataset like Flickr8k, Flickr30k, MSCOCO is applied on the various predefined model like VGG16, DenseNet161, InceptionV3. Also To investigate how well the model works in the wild, and collected random images and visualized generated caption manually to use common sense to find how well the sentences are formed in terms of the context of the images and features inside the images. Below are the some of the samples:

4.4 Flickr8k sample caption generated by the model

Figure 4.2 Generated caption for Flickr8k dataset

4.5 Flickr30k sample caption generated by the model

Figure 4.3 Generated caption for Flickr30k dataset

4.6 Tom and Jerry video key-frame sample caption generated by the model

Figure 4.4 Generated caption for Tom & Jerry video extraction

4.7 Evaluation metrics

The various model performance of the model used is shown below in the form of diagram and individual model performance is captured in the subsequent below sub-sections. It has been observed that low dataset corresponds to poor quality of caption generations. When trained on the entire dataset, the correctness of the test image caption improves significantly. The below results is shown based on the less than 1000 image data.

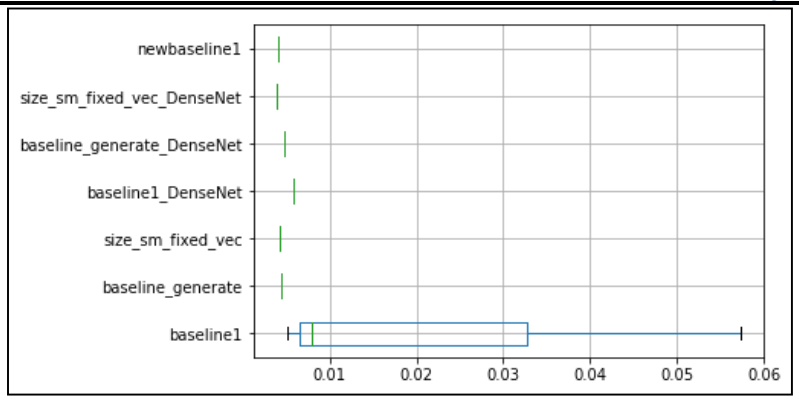

Figure 4.5 Visualization of model performance

4.8 Flickr8k BLEU score

Table 4.1 BLEU score for VGG16 on Flickr8k

Model Name	BLEU	
VGG16	Train	Test
baseline1	0.172	0.9923
findImage	0.172	0.9923
newbaseline1	0.889	0.4006
baseline_generate	0.1048	0.4351
size_sm_fixed_vec	0.098	0.4217

Table 4.2 BLEU Score for DenseNet161 on Flickr8k

Model Name	BLEU	
DenseNet161	Train	Test
baseline1	0.172	0.9923
findImage	0.172	0.9923
newbaseline1	0.0889	0.4006
baseline_generate	0.1048	0.4351
size_sm_fixed_vec	0.098	0.4217

Table 4.3 BLEU Score for InceptionV3 on Flickr8k

Model Name	BLEU	
InceptionV3	Train	Test
baseline1	0.172	0.9923
findImage	0.172	0.9923
newbaseline1	0.0889	0.4006
baseline_generate	0.1048	0.4351
size_sm_fixed_vec	0.098	0.4217

4.9 Flickr30k BLEU score

Table 4.4 BLEU Score for VGG16 on Flickr30k

Model	BLEU	
VGG16	Train	Test
baseline130k	0.400	0.650
baseline_generate30k	.75	0.5
size_sm_fixed_vec30k	0.3	.35

4.10 Overall loss and accuracy

This section captures the overall loss and accuracy that has been observed from as an outcome of the model. By observing the graph conclusion can be drawn that the loss can be minimized and the accuracy can be improved by increasing the training data. The below loss graph is created out of the model. From the below graph it is displayed that the model achieved nearly 80% of accuracy, however the model loss is also increasing with the increase in no of epochs.

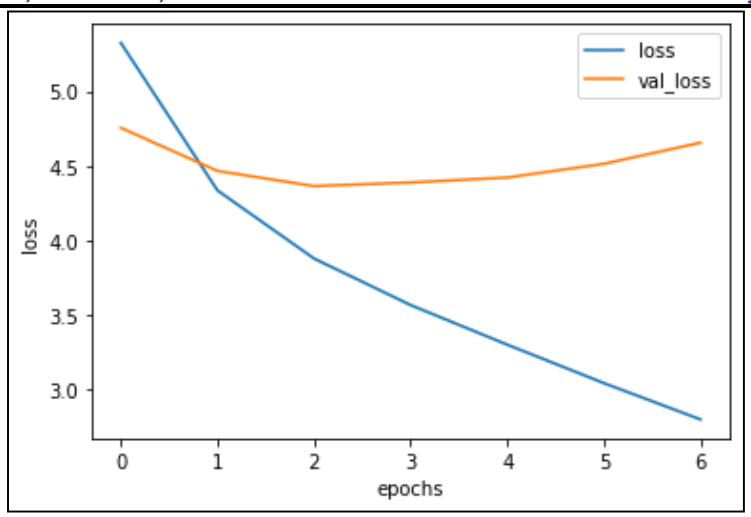

Figure 4.6 Loss graph

The below accuracy graph is created out of the model. The below graph it can be represented that overfitting issue is due to the fact that use of small size of sample data. This leads to the poor performance on classifying unseen data.

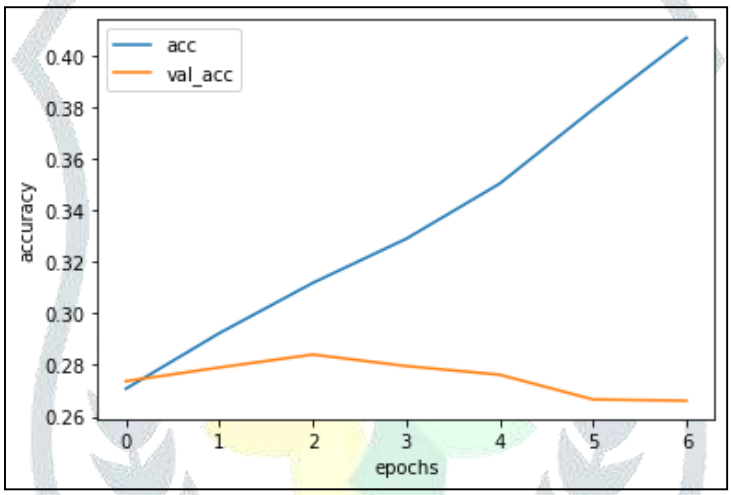

Figure 4.7 Accuracy graph

4.11 Evaluation of Individual Prediction

The below section captures the individual test image caption output along with the BLEU score against each of the predefined sentences. It is observed that the score improved drastically when while using the unsupervised learning (glove data) and the accuracy of the generation of sentences that have come out is really great.

Figure 4.8 Final prediction

The below sentence is generated from the frame extracted from the video images.

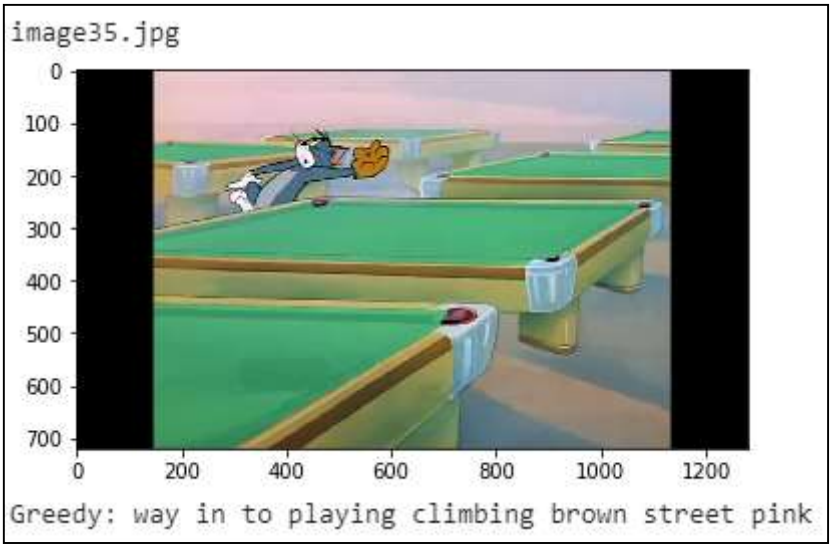

Figure 4.9 Video frame prediction

The below table represents the individual performance level on various predefined model captured in the experiment is given below.

Table 4.5 BLEU Score comparison for the dataset of Flickr8k

Model	Score
VGG16	75
InceptionV3	79
Densnet121	82

V. CONCLUSION

This thesis presented various models that have been used for image captioning. The thesis also show their performance on various datasets and analysed the challenges associated with these models. These models with key frame extraction algorithm generates sentences for the image frames extracted from the video. But key frame extraction is an ill-defined problem and due to the lack of better generalization in image captioning model, generated summary may not be so meaning all the times. The main observation from this thesis is that training with huge sentences and creating huge vocabulary or dictionary may solve the issue. The performance can be improved by executing the same model in higher configuration computing assets which can give far better result with higher accuracy with close to perfect sentences.

One of the challenge is choosing the value of n, smaller the value of n, the number of redundant frames increases, larger the value of n, smaller the number of key-frames which may not be able to capture the context properly. Selecting optimal value of n is a challenging effort and this may vary from video to video depending on the rate at which the frames in the video gets changed. Another problem is that insufficient vocabulary making the description meaningless. Some other issue is that images with complex scene, image captioning giving description unrelated to the content, which leads to generate meaningless summary.

The approach to the problem is unique in nature. Tried to generate the sentences from video by using the same model which is used to generate the sentences from images. The model which is neural network that can automatically generate appropriate sentences in natural language like English. This model can be further leveraged by training with languages other than English and performance can be measured. Also unsupervised learning can be leveraged for both the images as well as text for improving the image caption generation. The process of the automatic caption generation can be modified further by taking an approach of generating the sentences from a dictionary by extracting the each unique feature from the image and the same can be applied for generating sentences for the video frames. Secondly the same model can be tried by training on huge set of description, so that enough vocabulary should be available to correctly address the features extracted from the images. Although the approach is data independent, but due to the problem of unavailability of video frame sentences, model generates meaningless summary sometimes. To overcome the issue a common decoder should be used which can address the issue effectively. The same model can be tried by detecting the objects from the images by using various libraries and feeding the same in the model for generating the sentence.

VI. ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to entire UPGRAD team to give me the golden opportunity to do this wonderful project on the topic image and video captioning and research mentor, Nasib Ullah.

REFERENCES

[1] Liu, Xing & Liu, Weibin & Xing, Weiwei. (2019). Image Captioning with Emotional Information via Multiple Model. 174-177. 10.1109/ICEIEC.2019.8784585.

[2] MIYOSHI, Yuya & HAGIWARA, Masafumi. (2019). Automatic Affective Image Captioning System using Emotion Estimation. Transactions of Japan Society of Kansei Engineering. 10.5057/jjske.TJSKE-D-18-00071.

[3] Li, Sheng & Tao, Zhiqiang & Fu, Yun. (2019). Visual to Text: Survey of Image and Video Captioning. IEEE Transactions on Emerging Topics in Computational Intelligence. PP. 1-16. 10.1109/TETCI.2019.2892755.

[4] Hossain, Md & Sohel, Ferdous & Shiratuddin, Mohd Fairuz & Laga, Hamid. (2018). A Comprehensive Study of Deep Learning for Image Captioning.
[5] Ding, Guiguang & Chen, Minghai & Zhao, Sicheng & Chen, Hui & Han, Jungong & Liu, Qiang. (2018). Neural Image Caption Generation with Weighted Training and Reference. Cognitive Computation. 10.1007/s12559-018-9581-x.
[6] Amritkar, Chetan & Jabade, Vaishali. (2018). Image Caption Generation Using Deep Learning Technique. 1-4. 10.1109/ICCUBEA.2018.8697360.
[7] Krause, Jonathan & Johnson, Justin & Krishna, Ranjay & Li, Fei Fei. (2016). A Hierarchical Approach for Generating Descriptive Image Paragraphs.
[8] Wang, Wenshi. (2018). Video Indexing and Retrieval based on Key Frame Extraction. International Journal of Performability Engineering. 14. 10.23940/ijpe.18.08.p19.18241832.
[9] Pawaskar, Sailee & Laxminarayana, J.. (2018). Image Caption Generation A Comprehensive Survey. International Journal of Computer Sciences and Engineering. 6. 230-234. 10.26438/ijcse/v6i3.230234.
[10] Singam, Prof. (2018). Automated Image Captioning Using ConvNets and Recurrent Neural Network. International Journal for Research in Applied Science and Engineering Technology. 6. 1168-1172. 10.22214/ijraset.2018.3182.
[11] Herranz, Luis & Min, Weiqing & Jiang, Shuqiang. (2018). Food recognition and recipe analysis: integrating visual content, context and external knowledge.
[12] Wei, Xinru & Qi, Yonggang & Liu, Jun & Liu, Fang. (2017). Image retrieval by dense caption reasoning. 1-4. 10.1109/VCIP.2017.8305157.
[13] Monteiro, João & Kitamoto, Asanobu & Martins, Bruno. (2017). Situational Awareness from Social Media Photographs Using Automated Image Captioning. 203-211. 10.1109/DSAA.2017.59.
[14] Huang, Gao & Liu, Zhuang & van der Maaten, Laurens & Weinberger, Kilian. (2017). Densely Connected Convolutional Networks. 10.1109/CVPR.2017.243.
[15] Moses Soh. (2016). Learning CNN-LSTM Architectures for Image Caption Generation. Published In: Stanford University.
[16] Wang, Minsi & Song, Li & Yang, Xiaokang & Luo, Chuanfei. (2016). A parallel-fusion RNN-LSTM architecture for image caption generation. 4448-4452. 10.1109/ICIP.2016.7533201.
[17] Vinyals, Oriol & Toshev, Alexander & Bengio, Samy & Erhan, Dumitru. (2016). Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39. 1-1. 10.1109/TPAMI.2016.2587640.
[18] Baker, Bowen & Gupta, Otkrist & Naik, Nikhil & Raskar, Ramesh. (2016). Designing Neural Network Architectures using Reinforcement Learning.
[19] Mills, Susanna & White, Martin & Wrieden, Wendy & Brown, Heather & Adams, Jean. (2016). Home cooking practices, experiences, and perceptions: a qualitative study using photo-elicitation interviewing. The Lancet. 388. S78. 10.1016/S0140-6736(16)32314-5.
[20] Szegedy, Christian & Vanhoucke, Vincent & Ioffe, Sergey & Shlens, Jon & Wojna, ZB. (2016). Rethinking the Inception Architecture for Computer Vision. 10.1109/CVPR.2016.308.
[21] Vinyals, Oriol & Toshev, Alexander & Bengio, Samy & Erhan, Dumitru. (2015). Show and tell: A neural image caption generator. 3156-3164. 10.1109/CVPR.2015.7298935.
[22] Ren, Shaoqing & He, Kaiming & Girshick, Ross & Sun, Jian. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39. 10.1109/TPAMI.2016.2577031.
[23] Liu, Shuying & Deng, Weihong. (2015). Very deep convolutional neural network based image classification using small training sample size. 730-734. 10.1109/ACPR.2015.7486599.
[24] Souza, Celso & Pádua, Flávio & Nunes, Cristiano & Assis, Guilherme & Silva, Giani. (2014). A unified approach to content-based indexing and retrieval of digital videos from television archives. Artificial Intelligence Research. 3. 49-61. 10.5430/air.v3n3p49.
[25] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.
[26] Chan, Tsung-Han & Jia, Kui & Gao, Shenghua & Lu, Jiwen & Zeng, Zinan & Ma, Yi. (2014). PCANet: A Simple Deep Learning Baseline for Image Classification?. IEEE Transactions on Image Processing. 24. 10.1109/TIP.2015.2475625.
[27] Szegedy, Christian & Liu, Wei & Jia, Yangqing & Sermanet, Pierre & Reed, Scott & Anguelov, Dragomir & Erhan, Dumitru & Vanhoucke, Vincent & Rabinovich, Andrew. (2014). Going Deeper with Convolutions.
[28] Andrej Karpathy, Fei-Fei Li. (2013). Automated Image Captioning with ConvNets and Recurrent Nets. Published In: Stanford University.
[29] Agrawal, Ameeta & An, Aijun. (2012). Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012. 346-353. 10.1109/WI-IAT.2012.170.
[30] Cireşan, Dan & Meier, Ueli & Masci, Jonathan & Schmidhuber, Jürgen. (2012). Multi-Column Deep Neural Network for Traffic Sign Classification. Neural networks : the official journal of the International Neural Network Society. 32. 333-8. 10.1016/j.neunet.2012.02.023.
[31] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.
[32] Ejaz, Naveed & Tariq, Tayyab & Baik, Sung. (2012). Adaptive key frame extraction for video summarization using an aggregation mechanism. Journal of Visual Communication and Image Representation. 23. 1031–1040. 10.1016/j.jvcir.2012.06.013.
[33] Bhuyan, Manas & Narra, Chaitanya & Chandra, Darsha. (2011). Hand gesture animation by key frame extraction. 1-6. 10.1109/ICIIP.2011.6108947.