"Image Annotation Using Neural Networks"

MINI PROJECT REPORT

FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

BY

Anupam Jaiswal(IIT2014038)
Yash Jain (IIT2014043)
Shivam Awasthi (IIT2014155)

UNDER THE SUPERVISION OF

Dr. Sonali Agarwal
Assistant Professor
IIIT-ALLAHABAD

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD
(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE
NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000
OF THE GOVT. OF INDIA)
A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED
BY GOVT. OF INDIA

Candidate's Declaration

We hereby declare that the work presented in this project report entitled “Image Annotation Using Neural Networks”, submitted as 7th semester B-Tech IT mini project is an authenticated record of our original work carried out from August 2017 to December 2017 under the guidance of Dr. Sonali Agarwal. Due acknowledgements have been made in the text to all the resources and frameworks used.

Signature:

Date: 22 November, 2017

Anupam Jaiswal (IIT2014038)

Yash Jain (IIT2014043)

Shivam Awasthi (IIT2014155)

Supervisor's Certificate

This is to certify that the project work "Image Annotation Using Neural Networks" is a bonafide work of Anupam Jaiswal (IIT2014038), Yash Jain (IIT2014043), Shivam Awasthi (IIT2014155) who carried out the project work under my supervision.

Signature

Dr. Sonali Agarwal Date: 22 November 2017

Acknowledgment

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Dr. Sonali Agarwal for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would like to express our special gratitude and thanks to the panel under Prof. Anupam Agarwal, Dr. Vrijendra Singh, and Dr. Jagpreet Singh for their never ending support.

Abstract

We plan to present a model that detects objects in an image and provide a natural language description of its region using object detection. Our approach uses datasets of images and their sentence descriptions to learn about the correspondences between the image regions and their natural language descriptions. Our model will be based on a combination of Convolutional Neural Networks over image regions and Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. This Multimodal Recurrent Neural Network architecture uses the inferred alignments to learn to generate novel descriptions of image regions. We will then demonstrate that our alignment model produces state of the art results in image description generation experiments on some datasets.

Table of Contents

1. Introduction

Humans just by having a quick look at an image can acquire immense amount of details from the image regarding the objects present in it, the background and the context of image. However , the same task is a lot difficult for computers and visual models to gain insights from an image. Most of the work that has been done in the field of image and visual modelling focussed majorly on classifying the image into a fixed and limited set of labels and great results have been already achieved in this field. These models however are quite restrictive due to the limited and closed vocabulary of visual concepts because of which they cannot generate description of a image as a human would do.

There have been approaches developed to generate image descriptions but often they depend on the hard-coded sentence templates and labels which puts a restriction on proper sentence generation for any arbitrary image.

So, in this paper we aim to work towards generating sentence description for any image independent of any hard coding of template and visual concepts. We plan to devise a model such that it should be free of assumptions about specific hard-coded templates, rules or categories and instead rely on learning from the training data. The challenges we face includes that the dataset available online which contains lakhs of images provides descriptions in natural language sentences about an image these descriptions multiplex mentions of several entities whose locations in the images are unknown.
We plans to present a model that aligns sentence snippets to the visual regions that they describe through a multimodal embedding. We then treat these correspondences as training data for a second, multimodal Recurrent Neural Network model that learns to generate the snippets.

2. Motivation

Image annotation is a topic that is often researched and various works has been done to improve the accuracy in classification. There have been approaches developed to generate image descriptions but often they depend on the hard-coded sentence templates and labels which puts a restriction on proper sentence generation for any arbitrary image. Then comes the method using neural networks, recently devised that performs better than previously devised techniques, we use a similar architecture to study about the neural networks employed for image annotation and hence construct our own set of neural networks to annotate a new image.

3. Problem definition and Objectives

The main aim of the project is to generate annotation for images . The major challenge is designing a model that generate these annotations that appropriately describe the contents of the image and its representation in the domain of natural language. The model should learn from the training images and their annotations and should not rely on hard coded rules and categories.

Fig 1 : Image with caption - *"Two dogs are playing in snow"*

4. Literature Survey

In research paper **[1]**, descriptions of images is prepared, focusing on what is important in it. Here, a model is described that computes a score mapping an image to a sentence. This score is used to map a descriptive sentence to a particular image, or vice versa. The score is calculated by comparing the meaning inferenced from the image to meaning from the sentence. The meaning is obtained from a unique procedure that is learned using data. A dataset - "PASCAL Sentence data set" - consisting of human-annotated images is used here. The estimate of meaning here is weak, but it is sufficient to produce very good quantitative results.

In **[2],** the aim is to understand images that are accompanied with text in the form of sentential descriptions. A model is proposed based on conditional random field for semantic parsing which identifies which objects are present in the image, their spatial extent and semantic segmentation, and generates text and image information as input. Automatic parsing of the sentences and extraction of objects and their relationships is done, and incorporated into the model. This approach is used on the UIUC dataset. **[2]** shows segmentation improvements of 12.5% over the visual only model.

In paper **[3]**, the problem of automatically generating a description of an image from its annotation is discussed. According to this paper, the results reported by the previous approaches are not very promising since they either use computer vision techniques to determine the labels or use the available descriptions of the training images to transfer/compose a new description for the test image. With this insight, an approach to generate image descriptions from image annotation is shown. **[3]** shows that with accurate object and attribute detection, human-like descriptions can be generated. NLP, Knowledge-based Information Systems, Information retrieval and Natural Language Generation are the key technologies used here. An extensive task-based evaluation is performed to analyze the results.

"Bidirectional retrieval" of images and sentences through a "multi-modal embedding" of visual and natural language data is used in **[4]**. Unlike other models that directly map images or sentences into a common space, the model used in this paper works on a finer level and embeds objects and fragments of sentences (typed dependency tree relations) into a common space. Experimental results shows that reasoning on both the global level of images and sentences and the finer level of their respective fragments significantly improves performance significantly.

The main motive of **[5]** is that while visual sense is present in images, it is the semantic features that are important and not image pixel data. This paper explores the use of human-generated abstract scenes made from clipart for learning human sense. The dataset is created from various cliparts. "MS COCO" training set is used. The approach used is defined as "joint text and vision model", which is followed by a training procedure. Vision and text alignment functions, text alignment score and vision alignment score are defined within the model to achieve the aim. Many strong baselines that use text alone are experimented upon - WikiEmbedding, COCOEmbedding, ValText, LargeVisualText, BigGenericText (Bing). Finally, two evaluation metrics are used to evaluate the results - AP and rank correlation.

[6] aims at using Recurrent neural network for language modelling. For our purpose, we need to additionally condition these models on images to generate meaningful sentences for each image. In **[6]** , the main goal is to demonstrate the use of RNN that are trained using Hessian Free (HF) Optimizer for predicting the next character in a stream of text. Instead of using standard RNN , Multiplicative RNN (MRNN) architecture was used. MRNNs were trained on over a hundred of megabytes of text using 8 Graphics Processing Units in parallel to perform significantly better than one of the best word agnostic character-level language models. **[6]** used three datasets - sequence of characters from the English Wikipedia, collection of articles from the NewYork Times , a corpus of machine learning papers consisting of every NIPS and JMLR paper till then. It was found that MRNNs learn surprisingly good language models using only 1500 hidden units, and unlike other approaches such as the sequence memoizer, they are easy to extend along various dimensions.

In **[7]** , image descriptions are generated using multimodal language model and sets a baseline when no additional structures are used. Multimodal language models are the models of natural language that can be conditioned on other modalities.An image-text multimodal neural language model can be used to retrieve images given a sentence query or retrieve image descriptions given an image.Unlike the past approaches that have worked in the field to generating image descriptions, **[7]** makes no use of structured models, syntactic trees or templates. Instead, it relies on word representations learned from millions of words and then the model is conditioned on high-level image features learned from deep neural networks. **[7]** then show how to learn word representations and image features together by jointly training our language models with a CNN. **[7]** has used three datasets with image text descriptions to perform the tasks : IAPR TC-12, Attributes Discovery, and the SBU datasets. Capabilities of the models were demonstrated through quantitative retrieval evaluation and through Bleu scores that are used for automated evaluation of statistical machine translation and has been used in **[7]** to measure

similarity of descriptions. It obtained improved Bleu scores to existing approaches for sentence generation.

[8] studies the problem of holistic scene understanding in which the scene type, objects and their spatial support in the image is inferred.It propose a unified framework to classify an image by recognizing, annotating and segmenting the objects within the image. It develops a hierarchical generative model that classifies the overall scene, recognizes and segments each object component, as well as annotates the image with a list of tags. Thus, A hierarchical model is developed to unify the patch-level, object-level, and scene-level information. This is the first model that performs all three tasks in single framework. The model presented in **[8]** jointly explains images through a visual and a textual model. Visually relevant objects are represented by regions and patches, while visually irrelevant textual annotations are influenced directly by the overall scene class. **[8]** also defines a fully automatic learning framework that is able to learn from noisy data on the web such as images and tags of the image from Flickr . The effectiveness of framework is then demonstrated by comparing its performance with other algorithms like Alipr and Corr-LDA. **[8]** significantly outperforms other state of-the-art algorithms in all three tasks.

In **[9]** , Frome et al. associate words and images through a semantic embedding. It describe a new deep visual-semantic embedding model trained to identify visual information obtained from unannotated text-objects using both labeled image data as well as semantic. The goals of **[9]** is to develop a vision model that makes semantically relevant predictions that generalizes to classes outside of its labeled training set which is called as zero-shot learning. It explicitly maps images into a rich semantic embedding space. It has three training phases - training a simple neural language model for learning semantically-meaningful, dense vector representations of words, training a deep NN for visual object recognition and then constructing a deep visual-semantic model by taking the lower layers of the pre-trained visual object recognition network and re-training them to predict the vector representation of the image label text as learned by the neural language model. The datasets that are used in **[9]** are ILSVRC 2012 1K dataset and ImageNet 2011 21K dataset. The model presented in **[9]** performs comparably to popular image classifiers when they were evaluated on flat 1-of-N metrics. **[9]** demonstrates that semantic knowledge improves zero-shot predictions achieving hit rates of up to 18% across thousands of labels never seen by the model.

The major aim of **[10]** is object detection within images at a large scale.In this work, we propose a neural network model for detection, which predicts a set of class-agnostic bounding boxes .It trains a detector which is called DeepMultiBox with the aim of generating small number of bounding boxes that are object candidates. Each box has a

score associated with it that tells the likelihood of containing any object of interest within that box. The model is robust is a sense that it can handle a variable number of instances for each class . It also allows for cross class generalization at the highest levels of the network. The datasets that are used in **[10]** are PASCAL VOC 2007 edition and ILSVRC 2012 edition. The object recognition based on this model works fairly well on the mentioned datasets even though it uses only the top few predicted locations in each image.

[11] measured object detection performance on PASCAL VOC dataset by implementing the same. By measuring the performance on datasets and comparing it with previously available performances on same database while applying computer vision techniques such as SIFT and HOG it concludes that its approach using CNN is better than previously used computer vision techniques. Its object detection system consists of three modules. The first generates category-independent region proposals. These proposals define the set of candidate detections available to their detector. The second module is a large convolutional neural network that extracts a fixed-length feature vector from each region. The third module is a set of class specific linear SVMs. Since approach combines region proposals with CNN's, they called the proposed method as R-CNN: Regions with CNN features.

In **[12]** a model is introduced for bidirectional retrieval of images and sentences through a multi-modal embedding of visual and natural language data. Described model works by embedding fragments of images and fragments of sentences into a common space. In addition to a ranking objective seen in previous work, this allows us to add a new fragment alignment objective that learns to directly associate these fragments across modalities. By extensive evaluation they showed that reasoning on both the global level of images and sentences and the finer level of their respective fragments significantly improved performance on image-sentence retrieval tasks. In **[12]** main emphasis is on training a model on a set of images and their associated natural language descriptions such that they can later retrieve annotations from image and vice versa. They formulated a structured, max-margin objective for a deep neural network that learns to embed both visual and language data into a common, multimodal space. As told above they fragmented both natural language annotation and image into fragments, the neural network they designed thus have to explicitly reason about their latent, inter-modal correspondences. Improvements over previously reported methods were seen when tested on image-sentence retrieval tasks on Pascal1K, Flickr8K and Flickr30K datasets. Also as our project is on unidirectional retrieval of annotations from images this paper could also help in extending our current project.

[13] describes about The ImageNet Large Scale Visual Recognition Challenge that is a benchmark in object category classification and detection in images. The challenge has been run annually from 2010 to present. This paper highlight keys breakthroughs in categorical object recognition, provide a detailed analysis of the current state of the field of large-scale image classification and object detection, and compare the state-of-the-art computer vision accuracy with CNN's and human accuracy.
[14] proposes a model which is semi supervised which segments and annotates images using very few labeled images and to relate image region to text labels it uses a large unaligned text corpora. Having a newspaper article about an image and few labeled images are necessary to provide a pixel level labeling of objects. The given model is inspired by the observation in a text corpora words share certain context and and must have some similarity with visual objects. This paper describes images using visual words, a new region-based representation. This model is based on kernelized canonical correlation analysis which finds a mapping between visual and textual words by projecting them into a latent meaning space. Kernels are derived from context and adjective features inside the respective visual and textual domains. Articles from New York times were used in this paper for textual features and to apply this method a dataset such as that of flicker was used. The model outsmarted other existing approaches to annotate an image and in segmentation it compares with method that more training data than the paper's approach.

Bidirectional mapping between image and its annotation is implemented and examined in **[15]**. The described approach uses Recurrent neural network. Image can also be retrieved from given image descriptions using the same model described in this paper. A novel recurrent visual memory was used that automatically learns to remember long-term visual concepts to help in both visual feature reconstruction and sentence generation. They evaluated their approach on several tasks. These include sentence generation, sentence retrieval and image retrieval. State-of the-art results are shown for the task of generating novel image descriptions. They demonstrated their method on numerous datasets including the PASCAL sentence dataset, Flickr 8K, Flickr 30K, and the Microsoft COCO dataset. They demonstrates results as measured by BLEU and METEOR on PASCAL 1K.

4.1 Summarization -

S.No	Title	Frameworks/ Algorithms	Tools/ dataset	Summary of research outcome
1	Every Picture Tells a Story: Generating Sentences from Images	Mapping of Images to meanings	PASCAL Sentence data set	Estimate of meaning here is weak, but produces good quantitative results.
2	A sentence is worth a thousand pixels	Conditional random field model	UIUC dataset	Improvements of 12.5% over the visual only model.
3	From Image Annotation to Image Description	NLP, Knowledge based Information Systems, Information retrieval, Natural Language Generation	PASCAL Sentence data set	Human like descriptions are produced. Task based evaluation is performed.
4	Deep Fragment Embeddings for Bidirectional Image Sentence Mapping	Bidirectional retrieval of images and sentences through a multi-modal embedding of visual and natural language data	Pascal1K, Flickr8K, Flickr30K datasets	Improved performance than approaches that directly map image and sentences into a common space.
5	Learning Common Sense Through Visual Abstraction	Joint text and vision model	Clipart and MS COCO	Two evaluation metrics are used- AP and rank correlation.
6	Generating Text with Recurrent Neural Networks	Recurrent Neural Networks, Hessian Free Optimization	Dataset of stream of characters from wiki , articles from New York	Learnt surprisingly good language models that could be easily extended to various dimensions .

			Times and corpus of Research papers.	
7	Unifying visual-semantic embeddings with multimodal neural language models	Multimodal neural language model , Joint image text feature model	IAPR TC-12, Attributes Discovery, SBU datasets	Obtained improved Bleu scores to existing approaches for sentence generation for images
8	Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework	Hierarchical generative model , Automatic training framework	Flickr dataset	Outperforms other state of-the-art algorithms in the field of image annotation.
9	Devise: A deep visual-semantic embedding model	Deep neural network , neural language model, visual semantic model , Image Classification	ILSVRC 2012 1K dataset, ImageNet 2011 21K dataset	Semantic knowledge improves zero-shot predictions achieving high hit rates across thousands of labels never seen by the model.
10	Scalable Object Detection using Deep Neural Networks	Deep Neural Network , Bounding Boxes , Object detection	VOC 2007 , ILSVRC 2012 datasets	It showed competitive results with other state of art object detectors but the computation cost is large.

11	Rich feature hierarchies for accurate object detection and semantic segmentation	Approach combines region proposals with CNN's, called as R-CNN: Regions with CNN features.	PASCAL VOC Dataset	Concludes that its approach using CNN is better than previously used computer vision techniques.
12	Deep Fragment embeddings for Bidirectional Image Sentence Mapping	Multi-modal embedding of visual and natural language data. Model works by embedding fragments of images and fragments of sentences into a common space	Pascal 1K, Flickr 8K, Flickr 30K datasets	Concludes that reasoning on both the global level of images and sentences and level of their respective fragments significantly improved performance on image-sentence retrieval tasks.
13	ImageNet Large Scale Visual Recognition Challenge	Provides a detailed analysis of the current state of the field of large-scale image classification and object detection	ILSVRC Dataset	Compares the state-of-the-art computer vision accuracy with CNN's.
14	Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora	Model is based on kernelized canonical correlation analysis which finds a mapping between visual and textual words by projecting them into a latent meaning space	Articles from New York times and Flickr database	Model outsmarted other existing approaches to annotate an image and in segmentation it compares with method that more training data than the paper's approach.

15	Learning a Recurrent Visual Representation for Image Caption Generation	Uses Recurrent neural network. A novel recurrent visual memory was used that automatically learns to remember long-term visual concepts to help in both visual feature reconstruction and sentence generation.	PASCAL sentence dataset, Flickr 8K, Flickr 30K, and the Microsoft COCO dataset.	They evaluated their approach on several tasks. These include sentence generation, sentence retrieval and image retrieval.

4.2 Conclusion of Literature Survey -

Various works in the field of Image captioning have been done, researches are going on for finding techniques that could outperform state of the art techniques. The best techniques emerged in recent times in ILSVRC 2014, when a group of participants used neural networks namely convolutional and recurrent neural network and using such a model resulted in higher accuracy in annotation than the existing computer vision methods, so we decide to employ similar technique to achieve our objectives.

5. Proposed Methodology

We propose to make use of training dataset of images and their corresponding description to learn about inter-modal correspondences between language and visual data.

1. We will use a deep convolutional neural network model with the aim to convert an image into corresponding vector that would be fed into RNN.
2. We will then develop a Recurrent Neural Network architecture that makes use of the data vector for an input image and generates its natural language sentence description.
3. Finally, we will train the model on this model and evaluate its performance on a new dataset of images using appropriate metrics like BLEU score.

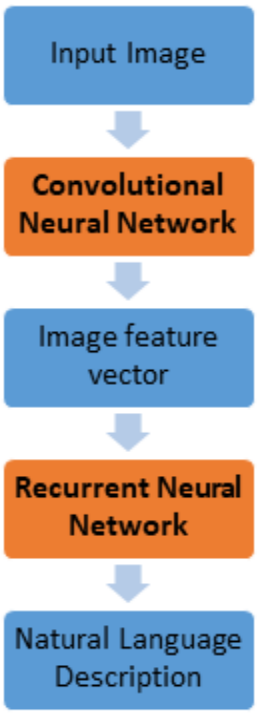


Fig 2: Flowchart of Proposed Approach

6. Hardware and Software requirements

6.1 Hardware Requirements -

- Machine with GPU sufficient enough for training of over 1,00,000 images .

6.2 Tools , Libraries and Programming Language -

- Linux Based OS.
- Python Programming language.
- Numpy
- Scipy
- argparse
- Other relevant python libraries.

7. Implementation

7.1 Convolutional Neural Network -

Convolutional neural networks (CNNs) are the current state-of-the-art model architecture or image classification tasks. CNNs apply a series of filters to the raw pixel data of an image to extract and learn higher-level features, which the model can then use for classification. CNNs contains three components:

- Convolutional layers, which apply a specified number of convolution filters to the image. For each subregion, the layer performs a set of mathematical operations to produce a single value in the output feature map. Convolutional layers then typically apply a ReLU activation function to the output to introduce nonlinearities into the model.
- Pooling layers, which downsample the image data extracted by the convolutional layers to reduce the dimensionality of the feature map in order to decrease processing time. A commonly used pooling algorithm is max pooling, which extracts subregions of the feature map (e.g., 2x2-pixel tiles), keeps their maximum value, and discards all other values.
- Dense (fully connected) layers, which perform classification on the features extracted by the convolutional layers and downsampled by the pooling layers. In a dense layer, every node in the layer is connected to every node in the preceding layer.

Typically, a CNN is composed of a stack of convolutional modules that perform feature extraction. Each module consists of a convolutional layer followed by a pooling layer. The last convolutional module is followed by one or more dense layers that perform classification. The final dense layer in a CNN contains a single node for each target class in the model (all the possible classes the model may predict), with a softmax activation function to generate a value between 0–1 for each node (the sum of all these softmax values is equal to 1). We can interpret the softmax values for a given image as relative measurements of how likely it is that the image falls into each target class.

We used VGG-16 pre trained cnn for extracting image features of the given images.

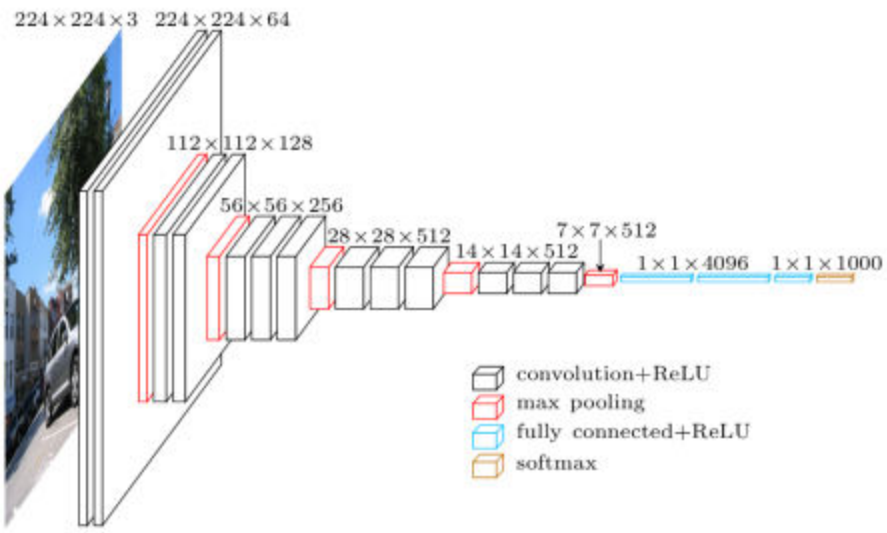


Fig 3 : Architecture of VGG-16 CNN

This network is characterized by its simplicity, using only *3×3* convolutional layers stacked on top of each other in increasing depth. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier. The "16" stand for the number of weight layers in the network.

7.2 Recurrent Neural Network -

The main idea of RNN is to make use of sequential information in a natural language. This means that to predict the next word we require the knowledge of previous words that have come before. They are called recurrent because they do the same task for every word of the sequence, and the output depends on the previous computations. RNN have their own memory which stores the information about what has been calculated so far.

Language Modelling -

Given m words, probability of observing the sentence is -

$$\mathbf{P}(w_1, w_2, \ldots, w_m) = \Pi \mathbf{P}(w_i \mid w_1, \ldots, w_{i-1})$$

i.e, Probability of a sentence is the product of probabilities of each word given the words that came before it, eg, the probability of the sentence "He went to buy some chocolate" would be the probability of "chocolate" given "He went to buy some", multiplied by the probability of "some" given "He went to buy", and so on.

In our implementation , each word $\mathbf{x_t}$ is represented as a vector of size of the vocabulary. Using this input information, hidden layers are built according to equations:

$$\mathbf{e_t = f\,(\,W_x \,.\, x_t + b_x\,)}$$
$$\mathbf{h_t^{forward} = f\,(\,e_t + W_h \,.\, h_{t-1}^{forward} + b_h\,)}$$
$$\mathbf{h_t^{backward} = f\,(\,e_t + W_h \,.\, h_{t+1}^{backward} + b_h\,)}$$
$$\mathbf{y_t = softmax\,(\,W_d \,.\,(\,h_t^{forward} + h_t^{backwatd}\,)\,+ b_d\,)}$$

RNN used here is bidirectional, i.e., forward as well as backward. Forward moves from left to right and backward moves from right to left. $\mathbf{h_t}$ is the hidden layer representation that depends on $\mathbf{h_{t-1}}$, previous information. $\mathbf{y_t}$ is the final predicted word. **W** and **b** are weights that are updated at every step. **f** is the activation function which can be set to either Rectified Linear Unit (**ReLU**) or **tanh**.

RNN training. The RNN takes a word ($\mathbf{x_t}$), the context from previous time steps ($\mathbf{h_{t-1}}$) to predict the next word ($\mathbf{y_t}$). Initially h_0 is set as zero vector, x_1 to a special START vector, and the desired label y_1 as the first word in the sequence. Then x_2 is set to the word vector of the first word and expect the network to predict the second word, etc. Finally, on the last step when x_T represents the last word, the target label is set to a special END token. The cost function is to maximize the log probability assigned to the target labels (i.e. Softmax classifier).

RNN at test time. To predict a sentence, we compute the image representation b_v (a vector), set h_0 as 0, x_1 to the START vector and compute the distribution over the first word y1. We sample a word from the distribution (or pick the argmax), set its embedding vector as x_2, and repeat this process until the END token is generated.

8. Dataset Description

The datasets that we will be working on are widely used in the field of Object and visual recognition and have been used in numerous research .

8.1 Flickr8k -

The dataset contains 8000 images.Each image has 5 independent captions that are human written. Multiple captions for each image are taken because there is a great amount of variance that is possible in the captions that can be written to describe a single image. This also helps satisfy the dynamic nature of images. There are multiple objects in the image but in a caption usually the main subject and either one or two of the secondary subjects are included in the caption. We use 1000 images for testing , 1000 for validation and rest for training.

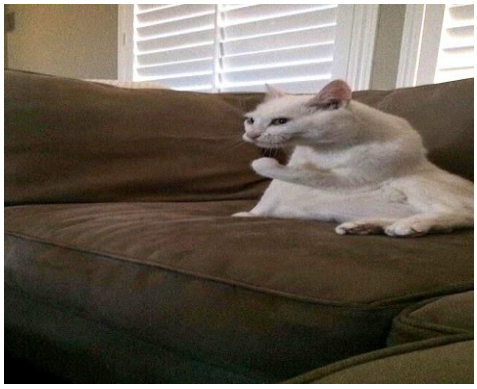
8.2 MS COCO -

Microsoft COCO is a large-scale object detection, segmentation, and captioning dataset obtained from the COCO Challenge held every year. It consists of 1,23,000 images. It has fewer categories but more instances per category, which enables better learning and makes this a richer dataset. Each image has 5 caption that are human written. We use 5000 images for testing , 5000 for validation and rest for training.

9. Results and Comparison

The model was trained separately on 2 datasets - flickr8k and MSCOCO . Now, we evaluate our model to caption images . First, we extract features using VGGNET cnn. Then, we run our RNN model to generate sentences for images with the aim of verifying that the model is rich enough to support the mapping from image data to sequences of words. We perform qualitative evaluation as well as we calculate the evaluation metric BLEU score. BLEU Score evaluates a candidate sentence by measuring how well it matches a set of five reference sentences written by humans which is provided in the dataset.

The results obtained on MSCOCO dataset is shown below-



A close up of a cat laying on a bed

A dog that is standing in the grass

A close up of a person holding a doughnut.

A beach with a lot of chairs and umbrellas

Fig 4 : Captions generated using our model trained on MSCOCO dataset

9.1 Qualitative Evaluation -

The model generates sensible descriptions of images (see Fig 4). Although , the captions generated not always give accurate results but most of the times , the captions fairly describe the image .In general, we find that a relatively large portion of generated sentences (60% with beam size 7) can be found in the training data. This fraction decreases with lower beam size; For instance, with beam size 1 this falls to 25%, but the performance also deteriorates in terms of the captions generated and their relevance with the image.

9.2 Comparison with other work -

We compare the results obtained by our model with other popular image captioning models . The comparison is done on the basis of BLEU Score which we obtained for different models from MSCOCO Captioning Challenge . We make a comparison on the basis of 4 Bleu Score - B-1, B-2, B-3, B-4 where each B-n is a Bleu Score that uses up to n-grams.

MSCOCO Dataset				
Model	**B-1**	**B-2**	**B-3**	**B-4**
Nearest Neighbor	48.0	28.1	16.6	10.0
LRCN	62.8	44.2	30.4	--
Google	71.3	54.2	40.7	30.9
Our Model	64.9	46.4	32.1	23

Table 1: Comparison of our model with other models on basis of BLEU Score. HIgher score is better.

Our model outperforms the nearest neighbour model where each test image is annotated with a sentence of the most similar training set image as determined by L2 norm over VGGNet features . Our model did not worked better than Google's model which uses GoogleNet to extract CNN features and instead of RNN, uses more complex LSTM .LRCN Model presented by Donahue et al. use a 2-layer factored LSTM and it appear to work worse than ours. From the comparisons and results , it was observed that our model works at par with other state of art models in the field of image captioning and yet it is more fast and simple.

10. Conclusions

We developed a model that generates captions for images using neural networks .We used a convolutional neural network VGGNet model to convert an image into corresponding vector to be fed into RNN. We then developed a Recurrent Neural Network architecture that makes use of the data vector for an input image and generates its natural language sentence description. Finally, we trained the model and evaluated its performance on MSCOCO and Flickr8k dataset using appropriate metrics like BLEU score. We observed that our model works at par with other state of art models in the field of image captioning and yet it is more fast and simple.

10.1 Future Scope -

Our approach consists of two separate models. Going directly from an image sentence dataset to image level annotations as part of a single model trained end-to-end remains an open problem.

Apart from this , instead of RNN , its more complex variant LSTM could be used which is more difficult to implement and takes longer time in training but is more efficient .

11. References

1. Every Picture Tells a Story: Generating Sentences from Images. *A. Farhadi, M. Hejrati, Md. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth. (2010)*
2. A sentence is worth a thousand pixels. *Sanja Fidler, Abhishek Sharma, Raquel Urtasun. (2013)*
3. From Image Annotation to Image Description. *Ankush Gupta, Prashanth Mannem. (2012)*
4. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Andrej Karpathy Armand Joulin Li Fei-Fei. (2014)*
5. Learning Common Sense Through Visual Abstraction. *Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, Devi Parikh. (2012)*
6. Generating Text with Recurrent Neural Networks. *I. Sutskever, J. Martens, and G. E. Hinton. (2011)*
7. Unifying visual-semantic embeddings with multimodal neural language models. *R. Kiros, R. Salakhutdinov, and R. S. Zemel. (2014)*
8. Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework. *L.-J. Li, R. Socher, and L. Fei-Fei. (2009)*
9. Devise: A deep visual-semantic embedding model. *A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean,T. Mikolov. (2013)*
10. Scalable Object Detection using Deep Neural Networks. *Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. (2014)*
11. Rich feature hierarchies for accurate object detection and semantic segmentation. *Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. (2014)*
12. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Andrej Karpathy, Armand Joulin, Li Fei-Fei. (2014)*
13. ImageNet Large Scale Visual Recognition Challenge. *Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla Michael Bernstein, Alexander C. Berg, Li Fei-Fei. (2015)*
14. Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora. *Richard Socher, Li Fei-Fei*
15. Learning a Recurrent Visual Representation for Image Caption Generation. *Xinlei Chen, C. Lawrence Zitnick. (2014)*