Image Captioning - A Naive Approach

Deepak Kumar, Deepak Kumar Rai, Jnaneshwar, Manthan Singh
HMR Institute of Technology and Management, New Delhi

December 2020

Abstract: Image captioning is a complicated research area of Arti cial Intelligence (AI) which requires a functional and robust model that generates a caption for any image. Image captioning is a fundamental task which requires not only semantic understanding of images but also the interactions between the objects present in the image.Another task is to understand the visual language dynamics and to translate these relations into sensible captions . In this paper , we have proposed an architecture employing the use of multilayer Convolutional Neural Network (CNN) which is used for image processing and extracting features from the image. It generates an embedding which repre-sents the image which is passed to Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords . We showcase the accuracy of our model using Flickr30K dataset which contains 31,783 images with 5 captions for each image. We show that our model gives better results using the bleu metric. Keywords Image Captioning CNN RNN LSTM Flickr30k BLEU Inceptionv3

1 Introduction

Humans can see the world and can detailed descriptions of the scene before their eyes . Computer vision aims at incorporating this ability of humans to di erentiate multiple scenes and images by giving a detailed machine generated description . Thus, image captioning is basically describing various objects present in a scenario and also describing the relationship between these objects with respect to their surroundings. While some previous models [21,22,19] have been proposed to address the problem of image captioning, they rely on either use sentence templates, or treat it as retrieval task through ranking the best matching sentence in database as caption. Those approaches usually su er di culty in generating variable length and novel sentences. Recent work [1] after the advent of neural network and deep learning has given a boost to this area and has delivered promising results. Our model uses pretrained InceptionV3 (Convolutional Neural Network) to generate embeddings from the input image. We will then use those embeddings to generate captions using RNN. The system is trained by showing it hundreds of thousands of images that were captioned

manually by humans, and it often re-uses human captions when presented with scenes similar to what its seen before. Training requires lot of computational power (around two days on GPUs), so thats why we have used Google Colab to train our models on GPUs. Training will be performed using the Flickr30k dataset. This is an extension of Flickr8k It describes 31,783 images of people involved in everyday activities and events and each image captions 5 captions .Its obtained from the Flickr website by University of Illinois at Urbana, Champaign .[24]

2 Related Work

Various approaches have been described in the past to solve image captioning tasks. We have a done a detailed analysis of various methodologies applied in the past to generate captions and also compared their e ciency. 1) Three spaces: One of the most signi cant works involves de ning three spaces namely image, meaning and sentence space. Mapping is done from image and sentence space to meaning space to check whether the caption generated makes sense. We nd the degree of similarity between images and generated sentences and the outputs are stored as triplets (image, action, object) and score is evaluated to nd out how accurate the caption is. If the sentence generated and image are highly similar the output score generated will also be higher. This model wasnt as accurate and had a lot of drawbacks.[2] 2) CNN-RNN: The advent of neural network and deep learning has given a boost to this area. CNN (Convolutional neural network) is used for image processing and it nds out all the details about the image at hand such as brightness, height, width, edges, etc. Various features of the image are extracted using this. RNN (Recurrent neural networks) is used to generate the actual caption using LSTM (Long short term memory). The output generated by CNN is passed on as input to RNN which in turn generates captions. [4] 3) Visual attention: In this approach maximum attention is given to the main object of the image and the caption is generated around it. In real world there are lots of other objects and scenarios in the picture also theres noise and clutter present but all these features dont make it to the RNN only the most important ones do.[5] 4) Novel object caption network method: This method has clubbed the image captioning datasets with other sources such as object recognition datasets to improve accuracy. The captioning model consists of two di erent LSTM layers, the rst one is the topdown LSTM layer and the second LSTM layer is the language model. The model achieved a state-of-the-art performance by achieving a BLEU-4/Spice/Cider scores of 36.9,21.5,117.9 respectively Most of the work done in image captioning is based on combining CNN with other type of models.[6]

A. Other Related work The extension of the works of image captioning can be seen in some of the areas given below : 1) Captioning Evaluation: Image captioning has been arduous due to its vague nature. Human evalua-tion is a better way to obtain captions but it is costly. Therefore, there are certain metrics like METEOR,ROGUE,BLEU etc used nowadays instead of human judgement. One

of the metrics that is becoming popular nowadays is SPICE which gives higher correlation and the results are comparable to human judgement. An important observation here was all the aforementioned metrics compare reference captions and the ones weve given without considering the image. The model weve developed rather takes image as an input and we get scores for each candidate captions and the best one is selected. 2) Adversarial Training and Evaluation: Generative adversarial networks (GANs) is another technique which is used to generate image captions. GANs are especially useful in telling apart human and machine generated captions. The major di erence comes in the function of di erentiator which is used here for generation whereas we have used it in our model for evaluation. A good caption generator should make it di cult to nd out whether it is machine generated or written by humans.

3 Architecture

We have used pretrained InceptionV3 (Convolutional Neural Network) to generate embeddings from the input image. We then used those embeddings to generate captions using RNN. The system is trained by showing it hundreds of thousands of images that were captioned manually by humans, and it often re-uses human captions when presented with scenes similar to what its seen before.

4 Inception V3

Inception-v3 is trained for the ImageNet Large Visual Recognition Challenge using the data from 2012. ImageNet, is a dataset of over 15 millions labeled high-resolution images with around 22,000 categories. This is a standard task in computer vision, where models try to classify entire images into 1000 classes, like "Zebra", "Dalmatian", and "Dishwasher". As we can see from the above gure, InceptionV3 is 42 layers deep and much more e cient than VGG-net. We extract the features from the lower convolutional layer of InceptionV3 giving us a vector of shape (8, 8, 2048). We squash that to a shape of (64, 2048). This vector is then passed through the CNN Encoder(which consists of a single Fully connected layer). We represent an image using the 4096 1 nal layer of InceptionV3, denoted as g(I) for an image I. We train a linear transformation of g(I) that maps it into the 256 1 input dimensions expected by our LSTM network. This entire pipeline of image representation generation is represented by: CN N(I) = W (I)g(I) + b(I) (1)

5 RNN

Output from InceptionV3 which is a 4096 image embedding vector is passed to a bidirectional LSTM which keeps on generating new words until the END token is generated. We initialize a recurrent neural network with initial state equal

to zero. We then feed the image representation CN N(I) in as the rst input of a dynamic length LSTM, i.e. x 1 = CN N(I). Each hidden state of the LSTM emits a prediction for the next word in the sentence, denoted by pt+1 = LST M(xt) for t = 0:::N 1. The model is fully described by the set of equations: x 1 = CNN(I) (2) xt = WeSt for t = 0:::N 1 (3) p t+1 = LST M(xt) for t = 0:::N 1 (4)

A. LSTM Caption Generator The LSTM function[25] above can be described by the following equations where LST M(xt) returns pt + 1 and the tuple (mt ; ct) is passed as the current hidden state to the next hidden state.

6 Implementation

The implementation phase comprises of data pre-processing and training explained below : A. Data Pre-processing Preprocessing involves loading the dataset and caching the output from the InceptionV3 model to the disk. 1) Dataset: The Flickr30k [24]dataset has become a standard benchmark for sentence-based image description. This is an extension of Flickr8k. It describes 31,783 images of people involved in everyday activities and events and each image captions 5 captions. It's obtained from the Flickr website by University of Illinois at Urbana, Champaign. Out of the all vision and language datasets , Flickr30k has the most syntactically complex sentences. It is also very good with the out of the domain data. Also, the Flickr30K corpus provides the most nouns compared to other datasets, which often correspond to object/stu categories in vision research, thus helpful in detecting objects. Here is an example of the dataset: 2) Pre-processing: Initially, all the images are passed through InceptionV3 which outputs 4096 x 1 image embedding vector. All these image embedding vectors are saved using np.save function to avoid passing data through InceptionV3 again and again which will increase the training speed. There is an annotation le which contains ImageId and Caption. We load this annotation le using the Pandas library

B. Training As we are using Pre-trained InceptionV3(CNN), we don't need to de ne any con guration for our CNN. We only have to de ne various con guration pa-rameters such as dimensions, for our RNN. As we already have image em-bedding which was saved using np.save function, we don't have to pass the image through InceptionV3 rst. We can directly use image embedding which was saved during preprocessing. This image embedding is an input to LSTM network. We pass this image embedding as input to LSTM network. Generate the caption. Calculate the loss using generated caption and original caption and propagate to modify weights of LSTM.

The system is trained for 30 epochs with learning rate of 0.001 which decays exponentially. The loss value started at around 2.1 and converged to 0.36 after 30 epochs. Fig. 2 Loss vs Epoch C. Inference We created a simple API(a python function) which takes image path as input and returns the generated caption. We use the weights of the latest checkpoint le for generating the caption. This function is similar to our training function except we stop predicting when we

encounter the END token

7 Result

Here are some of the results obtained from our model: Fig. 3 Real Caption: a large dark skyscraper stands beside a large cathedral Prediction Caption: a large building with a clock on a building Fig. 4 Real Caption: a big crowd of people walking in the snow on their skis Prediction Caption: people are playing in the snow We have evaluated our model using the BLEU (Bilingual Evaluation Un-derstudy) which compares the refernce text with machine generated text. It's value ranges between 0 to 1. A perfect match between the 2 captions will lead to score 1 and a total mismatch will lead to score 0. Our model achieved a score of 0.61 which is quite commendable compared to the work done before in this eld.

8 Conclusion

The image captioning model described in this paper was implemented and we were able to generate moderately comparable captions with compared to human generated captions. CNN (Convolutional neural network) is used for image processing and it nds out all the details about the image at hand such as brightness, height, width, edges, etc. Various features of the image are extracted using this. RNN (Recurrent neural networks) is used to generate the actual caption using LSTM (Long short term memory). The output generated by CNN is passed on as input to RNN which in turn generates captions. The InceptionV3 model rst assigns probabilities to all the objects that are possibly present in the image. The model converts the image into word vector. This word vector is provided as input to LSTM cells which will then form sentence from this word vector. Although the results were satisfactory other alternate techniques such as attention models, GANs can be used to improve the performance of image captioning model.

9 Future Scope

Most of the work done in generating captions from images includes the use of Convolutional Neural Network as an important component of the framework. The drawbacks of Convolutional Neural Networks described as follows: CNN does not take into account the orientational and the spatial relation-ship of the features. It can be illustrated with the help of an example: In the given example the convolutional neural network will identify the two im-ages mentioned as a face, as compared to identifying and extracting both, the orientation and spatial relationship of the faces in the two images. Fig. 5 VGG-net architecture The approach used by convolutional neural networks to solve the above di culty is to use max pooling which leads to data loss from the image as it tends to take the

maximum value within the matrix. CNN is also slower in operation compared to maxpool. If the network is too deep with many hidden layers it'll take more time to train. Finding a faster method can improve the process further. Human beings are quite heterogeneous and the same image can lead to thousands of di erent captions by each person. One direction of future work could aim to capture the heterogeneous nature of human annotated captions and incorporate such information into captioning evaluation.

10 Refrences

[1] Andrej Karpathy and Li Fei-Fei,Deep Visual-Semantic Alignments for Generating Im-age Descriptions,664676,Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),jun 2015 [2] Ali Farhadi and Seyyed Mohammad Mohsen Hejrati and Mohammad Amin Sadeghi and Peter Young and Cyrus Rashtchian and Julia Hockenmaier and David A. Forsyth,Every Picture Tells a Story: Generating Sentences from Images,1529,Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, Septem-ber 5-11, 2010,Proceedings, Part IV,2010 [3] Peter Anderson 0001 and Xiaodong He and Chris Buehler and Damien Teney and Mark Johnson 0001 and Stephen Gould and Lei Zhang,Bottom-Up and TopDown Attention for Image Captioning and VQA,CoRR,abs/1707.07998,2017 [4] Oriol Vinyals and Alexander Toshev and Samy Bengio and Dumitru Erhan,Show and Tell: A Neural Image Caption Generator,CoRR,abs/1411.4555,2014 [5] Kelvin Xu and Jimmy Ba and Ryan Kiros and Kyunghyun Cho and Aaron C. Courville and Ruslan Salakhutdinov and Richard S. Zemel and Yoshua Bengio,Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,CoRR,abs/1502.03044,2015 [6] Subhashini Venugopalan and Lisa Anne Hendricks and Marcus Rohrbach and Raymond J. Mooney and Trevor Darrell and Kate Saenko,Captioning Images with Diverse Ob-jects,CoRR,abs/1606.07770,2016 [7] Yonghui Wu and Mike Schuster and Zhifeng Chen and Quoc V. Le and Mohammad Norouzi 0002 and Wolfgang Macherey and Maxim Krikun and Yuan Cao and Qin Gao and Klaus Macherey and Je Klingner and Apurva Shah and Melvin Johnson and Xi-aobing Liu and Lukasz Kaiser and Stephan Gouws and Yoshikiyo Kato and Taku Kudo and Hideto Kazawa and Keith Stevens and George Kurian and Nishant Patil and Wei Wang and Cli Young and Jason Smith0006 and Jason Riesa and Alex Rudnick and Oriol Vinyals and Greg Corrado and Macdu Hughes and Je rey Dean,Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,CoRR,abs/1609.08144,2016 [8] Saleema Amershi and Maya Cakmak and William Bradley Knox and Todd Kulesza,Power to the People: The Role of Humans in Interactive Machine Learning,The AI Magazine,35,105120,2014 [9] Geo rey Hinton and Li Deng and Dong Yu and George E.Dahl and Abdelrahman Mo-hamed and Navdeep Jaitly and Andrew Senior and Vincent Vanhoucke and Nguyen Patrick and Tara N. Sainath and Brian Kingsbury,Deep Neural Networks for Acous-tic Modeling in Speech Recognition,8297,IEEE Sig-

nal Processing Magazine, v. 29, (6), November 2012 [10] Janardan Misra and Indranil Saha,Arti cial neural networks in hardware: A survey of two decades of progress,Neurocomputing,75,239255,2010 [11] Holger R. Maier and Graeme C. Dandy,Neural networks for the prediction and forecasting of water resource variables: a review of modelling issues and applica-tions,Environmental Modelling and Software,15,101124,2000 [12] Avinash N. Bhute and B. B. Meshram,Text Based Approach For Indexing And Retrieval Of Image And Video: A Review,CoRR,abs/1404.1514,2014 [13] Keiron OShea and Ryan Nash,An Introduction to Convolutional Neural Net-works,CoRR,abs 1511.08458,2015 [14] Zachary Chase Lipton and David C. Kale and Charles Elkan and Randall C. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks,4th International Confer-ence on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings , 2016 [15] Jurgen Schmidhuber,Deep learning in neural networks: An overview, Neural Networks,85117,2015 [16] Micah Hodosh and Peter Young and Julia Hockenmaier,Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,J. Artif. Intell. Res,47,853899,2013 [17] Tsung-Yi Lin and Michael Maire and Serge J. Belongie and Lubomir D. Bourdev and Ross B. Girshick and James Hays and Pietro Perona and Deva Ramanan and Piotr Dollar and C. Lawrence Zitnick,Microsoft COCO: Common Objects in Context,Computing Re-search Repository (CoRR),abs/1405.0312,2014 [18] Karen Simonyan and Andrew Zisserman,Very Deep Convolutional Networks for Large-Scale Image Recognition,Computer Science - Computer Vision and Pattern Recogni-tion,2014 [19] Polina Kuznetsova and Vicente Ordonez and Alexander C. Berg and Tamara L. Berg and Yejin Choi,Collective Generation of Natural Image Descriptions,359368,The Association for Computer Linguistics,2012 [20] Siming Li and Girish Kulkarni and Tamara L. Berg and Alexander C. Berg and Yejin Choi,Composing Simple Image Descriptions using Web-scale Ngrams,Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Port-land, Oregon, USA,220228,ACL,2011 [21] Girish Kulkarni and Visruth Premraj and Vicente Ordonez and Sagnik Dhar and Siming Li and Yejin Choi and Alexander C. Berg and Tamara L. Berg,BabyTalk: Understanding and Generating Simple Image Descriptions,IEEE Trans. Pattern Anal. Mach. Intell,2891 2903,2013 [22] Polina Kuznetsova and Vicente Ordonez and Tamara L. Berg and Yejin Choi,Treetalk: Composition and compression of trees for image description,TACL,2,351362,2014 [23] Ryan Kiros and Ruslan Salakhutdinov and Richard S. Zemel,Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models,CoRR,abs/1411.2539,2014 [24] Bryan A. Plummer and Liwei Wang 0009 and Chris M. Cervantes and Juan C. Caicedo and Julia Hockenmaier and Svetlana Lazebnik,Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,26412649,IEEE Computer Society,2015 [25] Xu Jia and Efstratios Gavves and Basura Fernando and Tinne Tuytelaars,Guiding the Long-Short Term Memory Model for Image Caption Generation,24072415,IEEE Com-puter Society,2015 [26] Christian Szegedy and Wei Liu 0015 and Yangqing Jia and Pierre Sermanet and Scott E. Reed and Dragomir Anguelov and Dumitru Erhan and Vincent Vanhoucke and Andrew

Rabinovich,Going Deeper with Convolutions,CoRR,abs/1409.4842,2014.