Image Captioning using Machine Learning

Gaurav Hoskote (1225134352), Amey Bhilegaonkar(1225368924), Ninad Nale (1225710226), Apoorv Kakade (1225280290), Varad Deshmukh (1225369184)

December 8, 2022

Abstract

The process of providing a written description of a picture's content is known as Image Captioning. To generate captions, it combines computer vision and natural language processing. The project's objective is to study how the encoder-decoder Neural architecture works and how making changes to this neural network architecture - specifically the encoder, can affect the performance. Image Captioning is suitable for a wide range of applications, from social media to help those with visual impairments. Working on this project gave us clarity on different concepts of machine learning which are further described in detail in this report. We, in this project, have achieved satisfactory BLEU scores by doing comparative study of different encoders that can be used for the model. Several datasets, including COCO, flicker30k, flicker8k, localized narratives, and SCICAP, are available for training the model; we investigated these datasets and decided to use flicker8k dataset for the project. The Code for the project can be found on GitHub

1 Introduction

Computer Vision and Natural Language processing are evolving fields that have found many applications in the industry today. Automatically describing the contents of an image is a fundamental problem in Artificial Intelligence that connects Computer Vision and Natural Language Processing. Through this project, we studied the working of a generative model that maximizes the likelihood of the target description sentence, for a given input image. The project requires knowledge of Neural Networks - an advanced topic in Statistical Machine Learning(Deep Learning) - which consists of multiple perceptron layers that can carry out powerful/complex tasks in the field of machine learning.

Our image captioning model uses an encoder-decoder design, with the encoder receiving input in the form of abstract image feature vectors. The encoder helps us convert high dimensional image data into compressed vectors that have spatial information encoded in them. The Decoder on the other hand takes this encoded vector and produces sequential output i.e decodes the vector. We have used libraries like Keras, Pandas, Numpy, Glob, CV, among others.

Figure 1 shows block representation of image caption process.

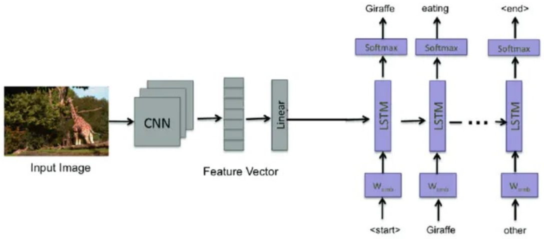


Figure 1

What is CNN?

Convolutional Neural networks are specialized deep neural networks that are capable of extracting spatial relationships in data. CNN is particularly advantageous when working with images due to the high spatial correlation between the pixels of the image. CNN is primarily used for classification, detection, and localization problems in deep learning. This is achieved by convolving filters that are trained to extract specific features and details from the image which further help us to classify the image. It can handle images that have been resized, rotated, translated, and perspective-shifted. For the purpose of this project we used two different CNN models - Resnet50 and Xception for the encoders.

What is LSTM?

Long short term memory, or LSTM, is a form of RNN (recurrent neural network) that is beneficial in resolving problems relating sequence prediction. We can anticipate the following word based on the prior text. By breaking over RNN's short-term memory restrictions, it has distinguished itself from regular RNN as an effective technology. Through the use of a forget gate, the LSTM may carry out necessary information while processing inputs and reject irrelevant information.

Therefore, these architectures were merged to design our model for the image caption generator. Another name for it is the CNN-RNN model. To extract features from the image, CNN is utilized. To formulate a description of the image, have uses LSTM with pretrained CNN models - Xception and ResNet50.

2 Problem Descriptions

The preliminary problem statement of this project is to perform comparative study of Image Captioning using two CNN pre-trained models ResNet50 and Xception, and Identify which model performs better under what conditions. The success of the project will be determined by the BLEU Score indicator. The following objective of the project is to learn and apply the Machine Learning and probability and statistics knowledge to the project.

3 Methodology

To solve the concerned problem of generating image captions we have used multiple neural networks particularly convolutional and recurrent neural networks. There are two main components in the system namely the encoder and the decoder which are neural networks trained in a supervised fashion. Encoders are convolutions neural networks that extract features from an image. We trained different encoder architectures on the data set and selected the best performer according to the metrics used. A Decoder follows LSTM architecture and is trained with the features generated by the encoder and the image captions to generate image captions. Below we have detailed the components we have used in our system.

3.1 Data-set

We have used the Flicker8K data set. It has 8,000 photos with five different captions, each describing the key elements and actions in the image. The pictures were hand-picked to represent a diversity of events and circumstances from six different Flickr groups, and they usually don't feature any famous individuals or places.

3.2 Data Preprocessing

Images and Captions form the input data and we need to format them properly. Images need to be resized and reshaped before being inputted to the encoder. While, text description need to be reformatted before being provided to the decoder model.
Since we are using ResNet-50 and Xception Architecture for the encoder, the input image sizes need to be 224x224 pixels for ResNet-50 and 299x299 pixels for Xception Architecture.

The captions are first cleaned by removing punctuation's, then a dictionary of 3967 words is formed. The captions are then converted to vectors using the dictionary indices.

3.3 Encoder - Convolution Neural Networks

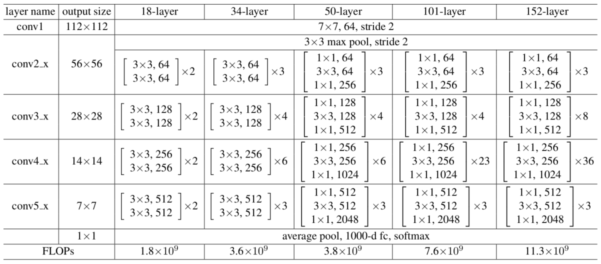
The task of an encoder is to extract important features from an given input digital image and encode them into a vector space suitable for input to the LSTM in the decoding stage. For this task we have chosen ResNet-50 and Xception Architecture. The models are pre-trained.

3.3.1 ResNet-50

ResNet-50 is Convolution Neural Network where ResNet stands for Residual Network and 50 indicates the layer depth of the model. The pretrained network can categorize photos into 1000 different object categories, including several animals, a keyboard, a mouse, and a pencil. The network has therefore acquired rich feature representations for a variety of images. The network accepts images with a resolution of 224 by 224. Convolution Neural Networks face an overwhelming problem of 'Vanish Gradient Descent' where during the course of propagation the value of gradient becomes very small and then hardly any weights are changeds.

ResNet-50 tackles this problem of vanishing gradient descent by introduction the concept of Skip Connection in its architecture. In Skip Connection,a direct connection established between layers and few layers between them are skipped.

Architecture of ResNet50 is as follows:

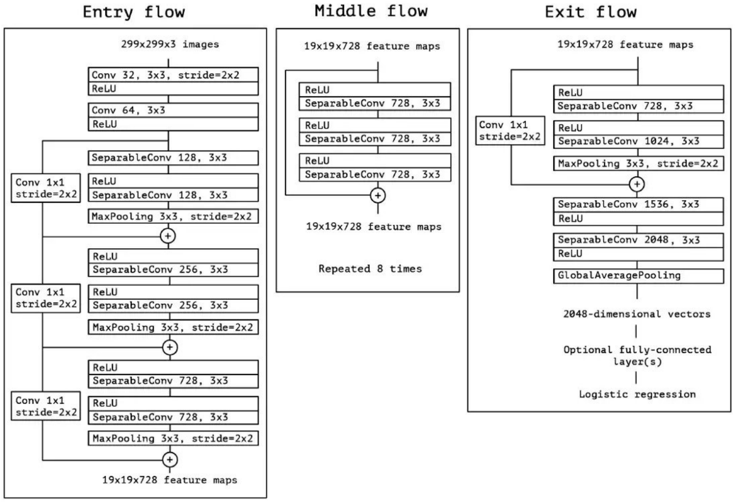


layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

3.3.2 Xception

Xception is a pre-trained Convolutional Neural Network model developed for task classification by Google. Xception stands for Extreme version of Inception, also a Google developed classification Model.
The key architectural feature of Xception is Depthwise Separable Convolutions. Depthwise Separable Convolutions work with both spatial dimension and depth dimension.

Architecture of Xception is as follows:



3.4 Decoder - LSTM

The task of decoder is to generate image captions using from the input of encoded image feature vectors. For decoder we using Recurrent Neural Network specifically LSTM.
LSTM stands for Long Short-Term Memory as the they are capable of long term dependencies.Also LSTM has feedback connections which allow the processing of entire sequence of data.
The Central role in the LSTM architecture is held by a memory cell also known as 'cell state'. Cell State allows the addition and removal of information and is controlled by Gates.
The decoder consists of the image and language model. The image model has Dense 128 unit layer with Relu activation and the LSTM has hidden layers of 128 and 512 units. The activation of the LSTM is softmax. The model is trained with RMSProp as the optimiser for a batch size of 256 for 50 epochs.

3.5 Loss Function

To calculate the loss with the LSTM model we decided to use the Cross Entropy Loss function. Since Cross Entropy loss function computes the loss between the Labels(in this case Descriptions) and the predictions(predicted description).Also Cross Entropy loss is a very effective and popular metric of performance measurement for classification models.
The mathematical representation of Cross Entropy loss is:
$L_{CE} = -\Sigma_{i=1}^{n} t_i log(p_i)$

4 Evaluation

We used two different encoders, ResNet-50 and Xception Architecture, and to numerically measure their performance we used the BLEU Score metric.

4.1 BLEU Score

A statistic for comparing a generated sentence to a reference sentence is the Bilingual Evaluation Understudy Score, or BLEU. A score of 1.0 indicates a perfect match, whereas a score of 0.0 indicates a perfect mismatch. For assessing the predictions produced by automatic machine translation systems, the score was created. It is not ideal, but it does have five excellent advantages:

1. It is quick and inexpensive to calculate.
2. It is easy to understand.
3. It is language independent.
4. It correlates highly with human evaluation.
5. It has been widely adopted.

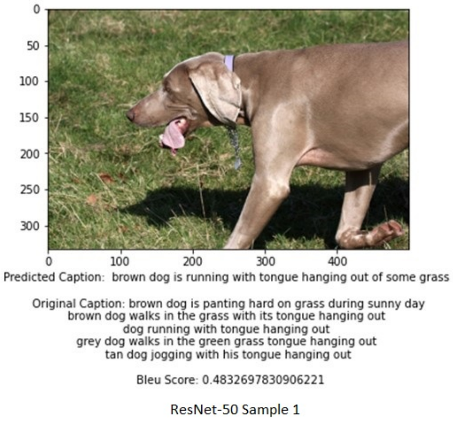
$BLEU = BP \exp(\Sigma_{n=1}^{N} w_n log p_n)$
BP = Brevity Penalty
N = Number of n-grams, like unigram, bigram, 3-gram,4-gram
w_n = Weight of each modified precision
p_n = Modified Precision

4.2 Model Performance - BLEU Score

We have calculated the BLEU score for two models, ResNet-50 and Xception. For Xception we achieved a BLEU score of 0.59 and for ResNet-50 we got an BLEU score of 0.059. ResNet-50 gave an significantly low score compared to Xception.

5 Results

Below are the output image sample with model generated captions for ResNet-50 and Xception



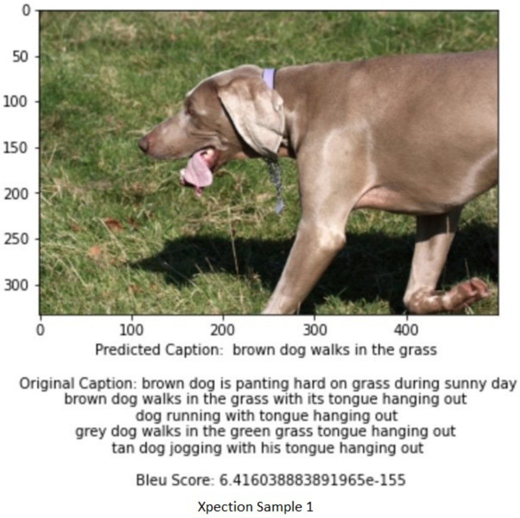
Predicted Caption: brown dog is running with tongue hanging out of some grass

Original Caption: brown dog is panting hard on grass during sunny day
brown dog walks in the grass with its tongue hanging out
dog running with tongue hanging out
grey dog walks in the green grass tongue hanging out
tan dog jogging with his tongue hanging out

Bleu Score: 0.4832697830906221

ResNet-50 Sample 1

Figure 1: ResNet-50 Dog Image



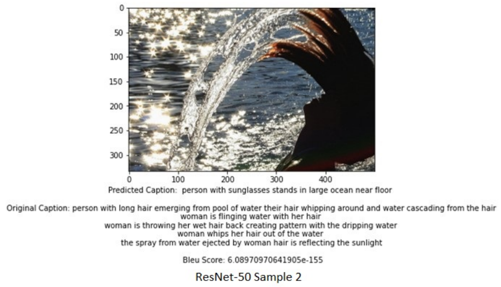
Predicted Caption: brown dog walks in the grass

Original Caption: brown dog is panting hard on grass during sunny day
brown dog walks in the grass with its tongue hanging out
dog running with tongue hanging out
grey dog walks in the green grass tongue hanging out
tan dog jogging with his tongue hanging out

Bleu Score: 6.416038883891965e-155

Xpection Sample 1

Figure 2: Xception Dog Image



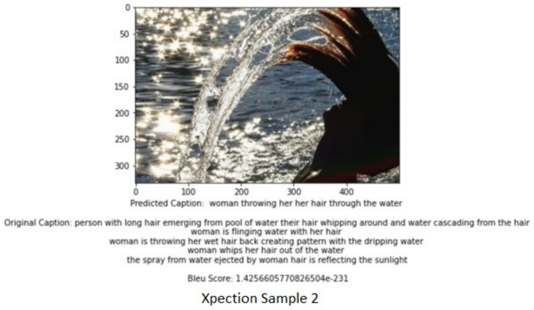
Predicted Caption: person with sunglasses stands in large ocean near floor

Original Caption: person with long hair emerging from pool of water their hair whipping around and water cascading from the hair
woman is flinging water with her hair
woman is throwing her wet hair back creating pattern with the dripping water
woman whips her hair out of the water
the spray from water ejected by woman hair is reflecting the sunlight

Bleu Score: 6.08970970641905e-155

ResNet-50 Sample 2

Figure 3: ResNet-50 Woman Image



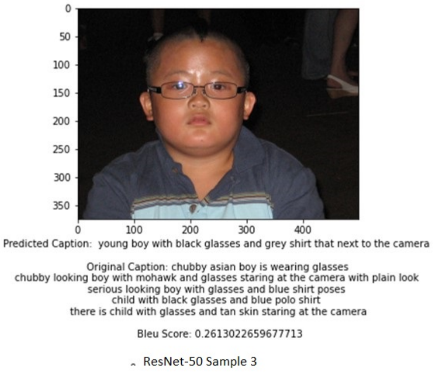
Predicted Caption: woman throwing her her hair through the water

Original Caption: person with long hair emerging from pool of water their hair whipping around and water cascading from the hair
woman is flinging water with her hair
woman is throwing her wet hair back creating pattern with the dripping water
woman whips her hair out of the water
the spray from water ejected by woman hair is reflecting the sunlight

Bleu Score: 1.4256605770826504e-231

Xpection Sample 2

Figure 4: Xception Woman Image



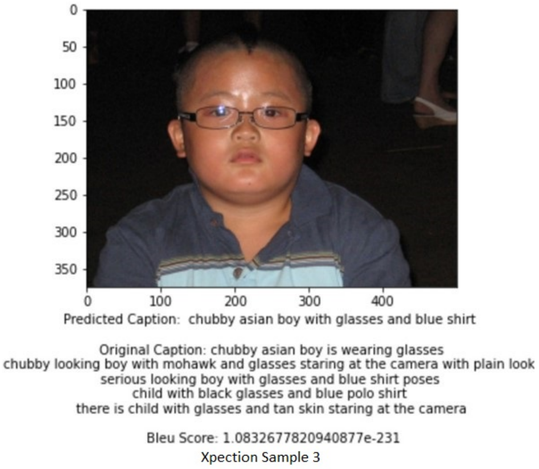
Predicted Caption: young boy with black glasses and grey shirt that next to the camera

Original Caption: chubby asian boy is wearing glasses
chubby looking boy with mohawk and glasses staring at the camera with plain look
serious looking boy with glasses and blue shirt poses
child with black glasses and blue polo shirt
there is child with glasses and tan skin staring at the camera

Bleu Score: 0.2613022659677713

ResNet-50 Sample 3

Figure 5: ResNet-50 Kid Image



Predicted Caption: chubby asian boy with glasses and blue shirt

Original Caption: chubby asian boy is wearing glasses
chubby looking boy with mohawk and glasses staring at the camera with plain look
serious looking boy with glasses and blue shirt poses
child with black glasses and blue polo shirt
there is child with glasses and tan skin staring at the camera

Bleu Score: 1.0832677820940877e-231

Xpection Sample 3

Figure 6: Xception Kid Image

6 Conclusion

Task of Image Captioning has come a long way but still can get significantly be better with the help of the ongoing research in the field of image feature extraction, classification and natural language sentence generation. We used pre-trained CNN models on machine with 12 GB of RAM which gave us the results but having a strong hardware and computing power, we could have clearly achieve better results in terms of BLEU score and training set number. We can also make use of larger data-sets like Flicker30k with 30,000 images, to refine the model training and increase accuracy. Overall learning from the project was how both of the Xception and ResNet50 model works, what are advantages and disadvantages of both of these models, how a fusion of CNN - a Image feature / information retrieval model and LSTM - a RNN mainly used for linguistic / text based features can be used to solve a complex task such as Image Captioning, which involves feature extraction and fed it as the sequential output.

7 References

1. A Guide to Image Captioning - Medium Blog

2. Image Captioning - : Transforming Objects into Words - Paper at Yahoo Research by Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares

3. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections - Paper at Yahoo Research

by Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares

4. Image Captioning: Transforming Objects into Words - Paper at Alibaba Group by Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan†, Bin Bi†, Jiabo Ye, Hehong Chen,Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, Luo Si, DAMO Academy
5. Image Captioning Datasets - Papers With Code Article
6. Image captioning with visual attention - Blog by TensorFlow
7. Learning to Evaluate Image Captioning - Paper by Cornell University