Image Captioning using Machine Learning

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

September 24, 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 implement an algorithm that can accurately describe the images using machine learning, making them suitable for a wide range of applications, from social media to help those with visual impairments. By doing this, the project will study and learn the concepts of machine learning. To accomplish this, we want to achieve respectable results for several success indicators, such as Bleu, CIDR, METEOR, and SPICE. Several datasets, including COCO, flicker30k, flicker8k, localized narratives, and SCICAP, are available for training the model; we want to investigate these datasets and pick the one that is balanced and yields the best results. The Code for the project can be found on GitHub

1 Group

Our group of five students has a range of technical backgrounds. Software developers Amey and Gaurav have three years of professional experience, while Apoorv and Ninad have one year of professional work experience. Varad recently received a Bachelor of Engineering degree with honors in Machine Learning and Artificial Intelligence. We have all studied different facets of Machine Learning and have all worked on projects involving various components of Deep Learning and Machine Learning. Since we all share the desire to learn more and apply what we have learned about theoretical machine learning to practical applications, we have come together as a group.
Contact Information for each of the team members is as below:

1. Amey Sadanand Bhilegaonkar: amey.bhilegaonkar@asu.edu
2. Gaurav Gourang Hoskote: ghoskote@asu.edu
3. Ninad Vijaysinh Nale: nnale@asu.edu
4. Apoorv Suresh Kakade: akakade@asu.edu
5. Varad Vijay Deshmukh: vdeshmu4@asu.edu

2 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 intend on studying 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.

The majority of image captioning models use an encoder-decoder design, with the encoder receiving input in the form of abstract image feature vectors. We will get an understanding of CNN and LSTM concepts through this project, which will produce the captions. We will be using libraries like TensorFlow, Keras, Pillow, Numpy, and TQDM, among others.

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.

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, we will merge these architectures 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, we will use LSTM with CNN data with hand with the help of the pre-trained model Xception

The overall purpose of this project is to introduce some basic and advanced machine learning concepts and algorithms while also using statistics and probability to categorize words and assign them to accurate image descriptions.

3 Background

Most of us on the team already have a rudimentary grasp of some Machine Learning and Deep Learning principles due to our shared backgrounds, but this project will hone our knowledge and educate us a lot about the application and practicality of these techniques.

The following lists the references and technical materials.

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

4 Measures of Success

The evaluation of image captioning models is generally performed using metrics such as BLEU, METEOR, ROUGE or CIDEr, all

of which mainly measure the word overlap between generated and reference captions. The recently proposed SPICE measures the similarity of scene graphs constructed from the candidate and reference sentence, and shows better correlation with human judgments.

Baseline: The project can be called a success if we are able to get a decent score for success indicators like Bleu, CIDEr, METEOR, etc. **Stretch**: The project can be stretched to further analysis of better indicators for measurement of success and exploring more complex models to achieve better accuracy.

5 Preliminary plan

Our plan to complete this would revolve around the 7 weeks till Nov 14, 2022. **Proposed Plan**

(a) Week 1 - Explore and Research about different Machine Learning Models and Datasets, try to find the best datasets with respect to balance of data points, available power of GPU/hardware to train the model etc.

(b) Week 2 - Refine Dataset and Start working on finalized model.

(c) Week 3 - Continue working on building and training model. Start working on Success measure indicator module.

(d) Week 4 - Verify training with validation sets and test on the test dataset.

(e) Week 5 - Start Preparing project reports. Tune hyper-parameters if required.

(f) Week 6 - Enhance module to meet stretch goals.

(g) Week 7 - Resolve Errors and keep working on stretch goals. Finalize project reports.

(h) Week 8 - Nov 14, 2022 - Be ready for the presentation.

Proposed Work Distribution Amongst the team

(a) Amey Bhilegaonkar, Ninad Nale, Varad Deshmukh - To work on Refinement of Data, Data Pre-processing, Driver code creation, Versioning and Documentation as well as Success measuring Indicator Modules for model Testing.

(b) Gaurav Hoskote, Apoorv Kakade - To work on Model Building, Training.

(c) Everyone - To contribute on Project proposal.

This is just a rough allocation of tasks and they might be shuffled in future as per requirements of the project.