WESTERN MICHIGAN UNIVERSITY

CS-5821 MACHINE LEARNING

PROJECT REPORT
Image Captioning Using Machine Learning Tools

Submitted By:
Harshitha Chollangi (WIN:)
Shrawani Palande (WIN: 813202795)
Swapnil Rajendra Patil (WIN: 453347613)
Shubham Anil Pawar (WIN: 209419404)

Submitted to:

Dr. A. C. Fong

ABSTRACT

In this project, we employ CNN to determine the image's caption. Large datasets and powerful computers are helpful in the development of models that can create captions for images as deep learning techniques advance. In this Python-based project, we will use deep learning methods like CNN and RNN to put this into practice. To understand the context of a picture and deliver it in English, an image caption generator uses computer vision and natural language processing techniques. In this comprehensive work, we strictly adhere to some of the fundamental ideas and methods used in image captioning. For this project, we talk about the Keras library, numpy, and Jupyter notebooks. Additionally, we talk about how CNN and the flickr_dataset are utilized to classify images.

INTRODUCTION

We come across a lot of images every day from several sources, including the internet, news stories, document diagrams, and commercials. These sites include pictures that visitors must interpret for themselves. Although the majority of photographs lack descriptions, most people can still understand them without them. However, if people want automated image captions from the machine, it must be able to understand some kind of captions. Many factors make image captioning crucial. Every image on the internet should include a caption to help with faster and more detailed image searches and indexing.

Since researchers have been working on object identification in photos, it has become obvious that a detailed human-like description is preferable to just listing the names of the items identified. Natural language descriptions will continue to be a problem that needs to be solved as long as robots do not think, speak, or act like people do.

There are several uses for image captioning in a variety of industries, including biomedicine, business, online search, and the military, among others. Social media platforms like Instagram and Facebook, among others, may automatically create captions from photographs.

MOTIVATION

A crucial problem that pertains to both the field of computer vision and the field of natural language processing is the creation of captions for photographs. The capacity of a machine to mimic human abilities to describe visuals with text is already a great advancement in artificial intelligence. The key difficulty in this endeavor is expressing the image's relationships between things in a language that is familiar to humans (like English). Traditionally, computer systems have been using pre-defined templates for generating text descriptions for images. This method, however, falls short of offering the variation needed to provide lexically rich text descriptions. With neural networks' enhanced efficiency, this flaw has been eliminated. Neural networks are frequently used in cutting-edge models to generate captions by taking an image as input and predicting the following lexical unit in the output phrase.

IMAGE CAPTIONING

Process:

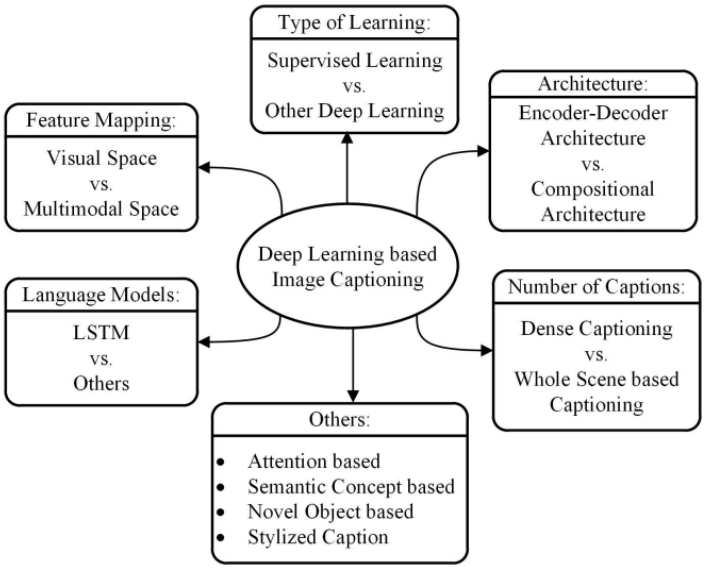
The process of creating a written description of a picture is called image captioning. The captions are produced using both computer vision and natural language processing. Artificial intelligence (AI) research on picture captioning is a hot topic since it involves comprehending images and providing a verbal description for them. Understanding an image requires the ability to find and identify items. Additionally, it must comprehend the sort of scene or location, item attributes, and how those interact. Understanding the language's syntactic structure and semantics are necessary for producing well-formed sentences. Getting picture characteristics is essential for understanding an image. They can be utilized, for instance, for automated picture indexing. Since image indexing is crucial for Content-Based Image Retrieval (CBIR), it may be used in a wide range of contexts, including digital libraries, online searches, biomedicine, business, the military, and education. Social networking sites like Facebook and Twitter have the ability to automatically create descriptions from photographs. The descriptions can include the location we're in (such as a beach or café), what we're wearing, and most importantly, what we're doing.

Techniques:

In general, there are two types of approaches that may be utilized for this purpose: (1) traditional machine learning-based techniques, and (2) deep machine learning-based techniques.

Hand-crafted features like Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), the Histogram of Oriented Gradients (HOG), and combinations of similar features are frequently employed in classical machine learning. These methods extract characteristics from the supplied data. After that, in order to categorize an item, they are sent to a classifier like Support Vector Machines (SVM). Since manually created features are task-specific, it is not possible to extract features from a huge and diverse amount of data. Real-world data, including photographs and videos, are very complicated and have a variety of semantic interpretations.

On the other hand, deep machine learning-based algorithms can handle a vast and varied set of photos and videos since features are automatically learnt from training data. For feature learning, Convolutional Neural Networks (CNN) [79] are frequently utilized, while for classification, a classifier like Softmax is employed. Recurrent Neural Networks (RNN) or Long Short-Term Memory Networks (LSTM) are frequently used after CNN to produce captions. Deep learning algorithms do a good job of handling the difficulties and intricacies of picture captioning.



A comprehensive classification of deep learning-based picture captioning.

GOALS

An image caption generator is a type of artificial intelligence (AI) system that can automatically generate a textual description of an image. The output of an image caption generator can be used for a wide range of applications. Here are few of them:

1. Accessibility: Providing image descriptions for people with visual impairments who use screen readers to access digital content.

2. E-commerce: Enhancing product listings with image captions that provide additional information about the product, its features, and its intended use.

3. Social media: Automatically generating captions for images shared on social media platforms to provide context and increase engagement.

4. Content creation: Generating captions for images used in blogs, articles, and other types of digital content to add context and improve search engine optimization (SEO).

5. Art and photography: Automatically generating titles and descriptions for works of art and photographs to enhance their understanding and appreciation.

6. Education: Generating captions for images used in educational materials to provide additional information and support learning.

7. Human-robot interaction: Enhancing the communication between humans and robots by allowing robots to generate captions for images they perceive in their environment.

8. Autonomous vehicles: Providing additional context for the visual data collected by autonomous vehicles, helping to improve their navigation and decision-making capabilities.

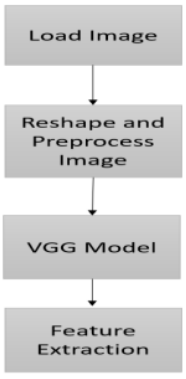
Overall, an image caption generator has the potential to improve the accessibility and understanding of digital content, enhance the user experience on various platforms, and enable new applications in a wide range of fields.

FLICKR8K DATASET

A benchmark dataset for picture to phrase description, the Flickr8k dataset consists of 8000 photographs with five descriptions each that were taken from various Flickr groups. The dataset is more general since it excludes well-known individuals and locations and offers precise descriptions of the objects and activities seen in the photographs. In the dataset, there are 6000 training photos, 1000 development images, and 1000 test sets. The model is general, robust, and less prone to overfitting because to the many descriptions per image and the variety of image categories.

IMAGE DATA PREPARATION

The picture must first be transformed into the appropriate features using a Convolutional Neural Network (CNN) with the VGG-16 model before training an image captioning deep learning model. The 16 weight layers and 33 convolutional layers of the VGG-16 model make it the best choice for feature extraction from pictures. The internal representation of the picture shortly prior to classification is returned as features after the final classification layer has been removed. The resultant feature vector is 1-dimensional with 4096 elements and requires an input picture with dimensions of 224224.



Feature Extraction in images

DATA CLEANING

Flickr8k dataset contains multiple descriptions described for a single image. In the data preparation phase, each image id is taken as key and its corresponding captions are stored as values in a dictionary.

Raw text has to be transformed into a format that can be used by machine learning or deep learning models. Before using the text for the project, the following text cleaning procedures are carried out: • Eliminating punctuation. • Elimination of numbers. • Elimination of words with only one word. • Lowercase characters are converted to uppercase ones. Stop words are included in the text data because removing them will make it more difficult to create the grammatically correct caption that is required for this project.

Original Captions	Captions after data cleaning
Two people are at the edge of a lake, facing the water and the city skyline.	two people are at the edge of lake facing the water and the city skyline
A little girl rides in a child 's swing.	little girl rides in child swing
Two boys posing in blue shirts and khaki shorts.	two boys posing in blue shirts and khaki shorts

EXTRACTING THE FEATURE VECTOR FROM ALL IMAGES

To minimize the need for recurrent feature extraction during the training or testing of a deep learning model, the term "caching the feature vector" refers to preserving the extracted features of an image in a different file or database. Caching the feature vector, in the context of the Flickr8k dataset, refers to extracting the picture features using the VGG-16 model just once and storing them in a file or database as opposed to extracting the features for each training or testing instance each time the model is run.

This may be helpful because feature extraction can take a while, especially with huge datasets. We can speed up model training and improve model performance by caching the feature vectors. Furthermore, caching can lower the possibility of memory problems, particularly when handling big datasets. Overall, caching the feature vector can enhance the deep learning model's effectiveness and performance.

TOKENIZING THE VOCABULARY

Because English words are not understood by computers, we must express them using numbers. Therefore, we will assign a distinct index value to each word in the lexicon. We may build tokens from our vocabulary using the tokenizer function provided by the Keras framework.

Creating Data Generator:

Let's first have a look at what our model's input and output will entail. We must give the model input and output for training if we want to turn this task into a supervised learning activity. Our model has to be trained on 6000 photos, each of which has a 2048-length feature vector and a caption that is likewise represented as a number. Because it is impossible to store this much data in memory for 6000 photos, we will use a generator approach that will produce batches.

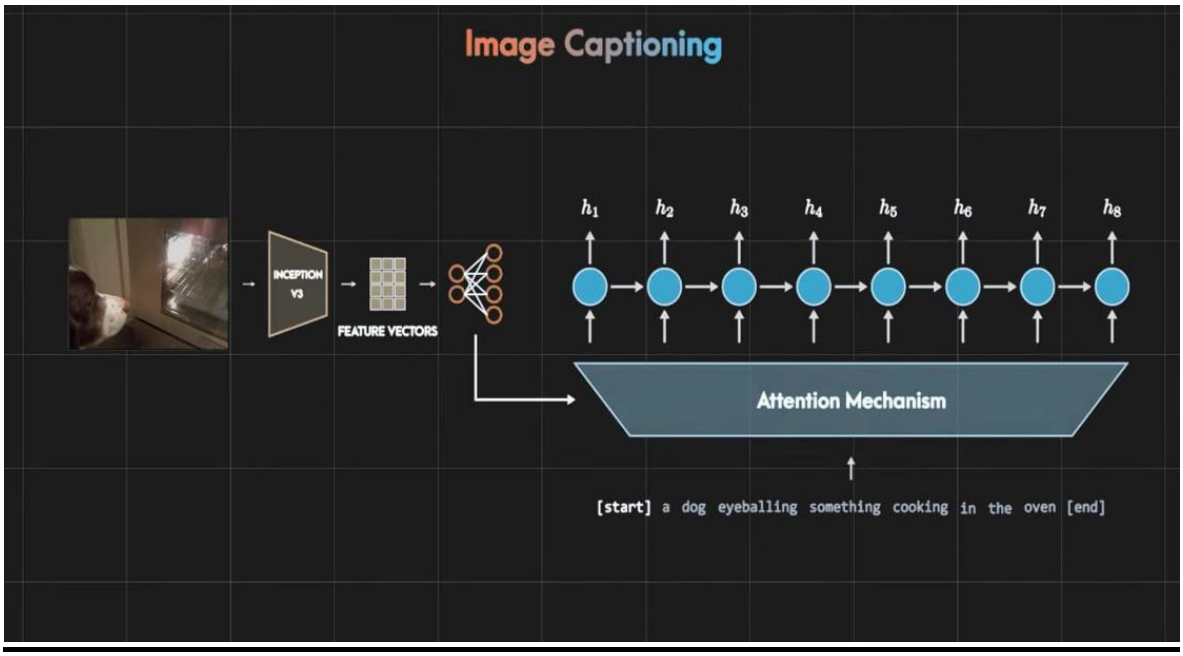
DEFINING THE CNN AND RNN MODEL

We'll use the Keras Model from the Functional API to specify the model's organizational structure. It will be divided into three main sections:

- Feature Extractor: With a thick layer, we can shrink the 2048 node feature that was retrieved from the picture to 256 nodes.
- Sequence Processor: The textual input will be handled by an embedding layer, then by an LSTM layer.
- Decoder: We will process by the dense layer after integrating the output from the two previously mentioned layers in order to arrive at the final forecast. The number of nodes in the top layer will be the same as the amount of our vocabulary.

ARCHITECTURE

The architecture of our project has been displayed below:



RESULTS

- **[start] parents are pushing little children in red car carts [end]**
- **[start] two girls in red costumes play together [end]**

CONCLUSION

We have reviewed deep learning-based picture captioning methods in this project, along with a taxonomy, assessment criteria, datasets, and possible future research lines. Even though a lot of progress has been made, a reliable approach for creating high-quality captions for almost all photographs is still a way off. With the creation of cutting-edge deep learning network topologies, the topic of automatic picture captioning will continue to be a research hotspot.

The text file also contains the captions for the almost 8000 photographs that make up the Flickr_8k dataset that we utilized. Although deep learning-based image captioning techniques have made significant strides in recent years, a reliable technique that can provide captions of a high caliber for almost all photos has not yet been developed. Automatic picture captioning will continue to be a popular study topic for some time to come with the introduction of innovative deep learning network designs. With more people using social media every day and the majority of them posting images, the potential for image captioning is quite broad in the future. Consequently, this effort will significantly assist them.

LIMITATIONS

A helpful framework for learning to map from photos to human-level image descriptions is provided by the neural image caption generator. The algorithm learns to extract pertinent semantic information from visual attributes by training on a large number of image-caption pairings.

We may train the image encoder as a component of the caption generation model to enhance the production of picture captions. This perfects the encoder so that it can produce captions more effectively. The produced captions might, however, nonetheless be unoriginal and uninteresting.

FUTURE SCOPE

- Improving caption accuracy: Despite the significant progress made in recent years, there is still room for improvement in caption accuracy. As new data and techniques emerge, it will be possible to train more accurate models that can better capture the complex relationships between images and captions.
- Multimodal understanding: Current image captioning models primarily rely on visual features to generate captions. However, integrating other modalities such as audio, text, and even haptic feedback can help improve the quality of captions and make them more comprehensive.
- Real-time captioning: Currently, image captioning models require some time to generate captions. In the future, researchers can work towards developing models that can generate captions in real-time. This will be useful in situations where immediate captions are necessary, such as for live events or video conferencing.

- Multi-language captioning: Currently, most image captioning models generate captions in a single language. As the need for multilingual support grows, developing models that can generate captions in multiple languages will become increasingly important.
- Explaining the reasoning behind captions: Understanding how image captioning models generate captions is a challenging problem. In the future, researchers can work towards developing models that can provide explanations for why they generated a particular caption. This will improve the interpretability of image captioning models and allow them to be used in a wider range of applications.

REFERENCES

- *Natural Language Processing (CSE 490U):* Neural Language Models

  Noah Smith (2017) - University of Washington (nasmith@cs.washington.edu)

  https://courses.cs.washington.edu/courses/cse490u/17wi/slides/lm-neural-slides.pdf

- *Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin.* (2003). A neural probabilistic language model. Journal of machine learning research 3, Feb, 1137–1155.
- A Comprehensive Survey of Deep Learning for Image Captioning
  MD. ZAKIR HOSSAIN, Murdoch University, Australia
  FERDOUS SOHEL, Murdoch University, Australia
  MOHD FAIRUZ SHIRATUDDIN, Murdoch University, Australia
  HAMID LAGA, Murdoch University, Australia
- Step by Step Guide to Build Image Caption Generator using Deep Learning
  https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/
- Using Machine Learning to Generate Image Captions
  https://towardsdatascience.com/using-machine-learning-to-generate-captions-for-images-f9a5797f31d6
- *Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing.* (2018).
  Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.