Seneca Polytechnic College

AUGUST 2023

Bridging the Visual Divide: A Deep Learning Approach to Image Captioning for the Visually Impaired

Sagar Shrestha

Table of Contents

Abstract

Being able to automatically describe the content of an image using properly formed English sentences is a challenging task, but it could have great impact by helping visually impaired people better understand their surroundings. Most modern mobile phones can capture photographs, making it possible for the visually impaired to take images of their environments. These images can be used for generating text or captions which further can be read aloud and be easier for the visually impaired persons. In this white paper, I have researched a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models to generate vocabulary describing the images and a Long Short-Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. This research bridges the gap between visually impaired people and technology. I have discussed the obstacles faced and how the system is prepared to close the gap between them.

Keywords: Image Captioning, Deep Learning

Problem Statement

Visual content, such as images and photographs, plays a crucial role in conveying information and emotions. However, visually impaired individuals often face challenges in understanding the content of images due to their reliance on non-visual means of perception. To address this issue, this research aims to develop an advanced deep learning-based solution that can generate precise and descriptive captions for images. By providing textual descriptions of the visual content, this solution intends to empower visually impaired individuals to perceive and comprehend images effectively.

Introduction

Every day, we encounter many images from various sources such as the internet, news, articles, documents, diagrams, and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have an altered text in it, but humans can largely understand them without their detailed captions. However, machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. For example, they can be used for automatic image indexing. Image indexing is important or content-based image retrieval (CBIR) and therefore, it can be applied to many areas including biomedicine, commerce, educations, and web form images.

What is Image Captioning?

Image captioning is a computer vision and natural language processing task that involves generating textual descriptions or captions for images. The goal of image captioning is to develop algorithms or models that can analyze the content of an image and generate accurate and descriptive captions that describe what is happening in the image.

Figure 1: Motivation figure illustrating extraction of simple natural language description from visual data (Benn)

Image captioning combines techniques from both computer vision and natural language processing fields. It often uses deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to perform image analysis or transformer models for caption generation CNN processes the image to extract visual features, which then they provide a language model for loading text.

Background and Motivation

Recently, a survey revealed that approximately 55% of visually impaired individuals' own smartphones, which they access using accessibility features such as Talkback for Android and Voiceover for iPhones (Wagner). While text-based data is easily accessible and readable for them, the challenge lies in comprehending images. Without altered text accompanying images or descriptive explanations, visually impaired individuals miss the content conveyed through visuals.

Objective of Research

The objective of this research is to develop a deep learning-based approach for image captioning specifically designed to bridge the visual divide and enhance accessibility for individuals with visual impairments. The research aims to leverage state-of-the-art techniques in computer vision and natural language processing to automatically generate accurate and descriptive captions for images, enabling visually impaired individuals to perceive and understand visual content through text-based descriptions.

Obstacles for visually impaired person in image

Visually impaired individuals face several obstacles when it comes to accessing and understanding images. These obstacles include:

i. Limited availability of alternative image descriptions:
While some images may have alternative text descriptions (alt text), these descriptions are not always comprehensive or accurately capture the full content of the image. The lack of standardized and detailed alternative descriptions further hinders their understanding.

ii. Lack of visual information
Images are primarily visual content, which poses a challenge for visually impaired individuals who cannot perceive or interpret visual cues. Without alternative means of accessing image content, they miss out on valuable visual information.

How does Image Captioning work?

There are basically two main approaches to Image Captioning: bottom-up and top-down.

Bottom-Up Approach

Bottom-up approaches generate items observed in an image, and then attempt to combine the items identified into a caption.

Let us say we have an image of a beach scene with people, umbrellas, and palm trees. The bottom-up approach will first detect and identify these individual items in the image. It might identify people in the image, detect umbrellas, and recognize palm trees as separate entities. Then, the model will combine this information to generate a caption, such as "People enjoying the beach under umbrellas with palm trees in the background."

Top-Down Approach

Top-down approaches attempt to generate a semantic representation of an image that is then decoded into a caption using various architectures, such as recurrent neural networks.

In the top-down approach, the model might create a semantic representation that understands the overall scene, such as "a sunny beach with people relaxing under umbrellas." The RNN then decodes this representation to generate a caption like "People are enjoying a sunny day at the beach, lounging under colorful umbrellas."

The latter approach follows in the footsteps of recent advances in statistical machine translation, and the state-of-the-art models mostly adopt the top-down approach. (Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney)

This paper shows the top-down image generation models mentioned above. A deep convolutional neural network is used to generate a vectorized representation of an image that is then fed into a Long-Short-Term Memory (LSTM) network, which then generates captions. *Figure 2* provides the broad framework of the approach.

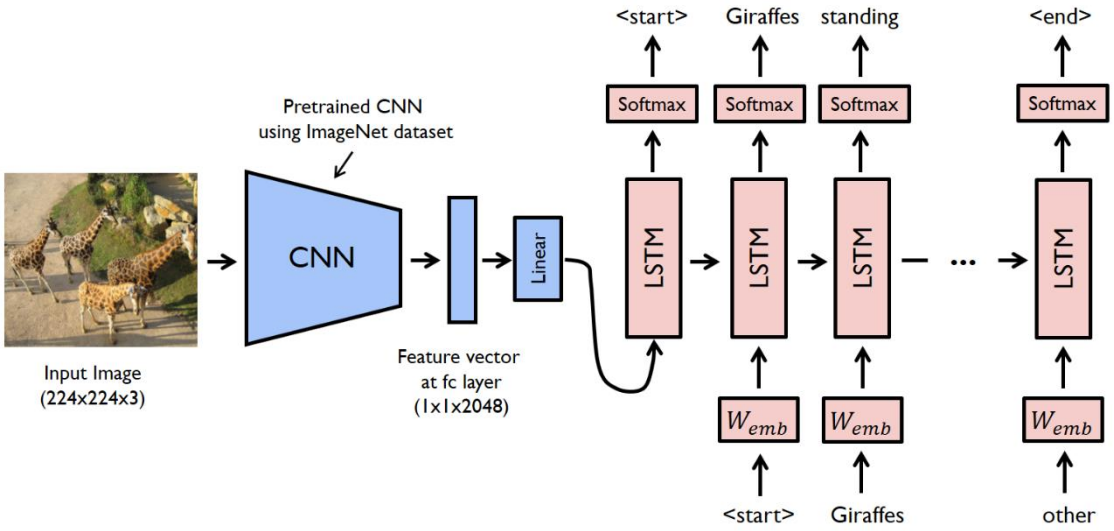

Figure 2: Automatic Image Captioning using Deep Learning (Top-Down Approach) (Shaikh)

CNN-LSTM architecture uses a deep convolution neural network to create a semantic representation of an image, which we then decode using a LSTM network. All LSTMs share the same parameter. The vectorized image representation is fed into the network, followed by a special start of sentence token. The hidden state produced is then used by the LSTM to predict/generate the caption for the given image.

Algorithms and Implementations

Deep Learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the- art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) contains at least one convolutional layer and at least one completely associated layer as in a standard multilayer neural system. The architecture of a CNN is intended to exploit the 2D structure of an information picture. This is accomplished with nearby associations and tied weights took after by some type of pooling which brings about interpretation invariant highlights (Kabir). Another advantage of CNNs is that they are less demanding to prepare and have numerous less parameters than completely associated systems with a similar number of concealed units.

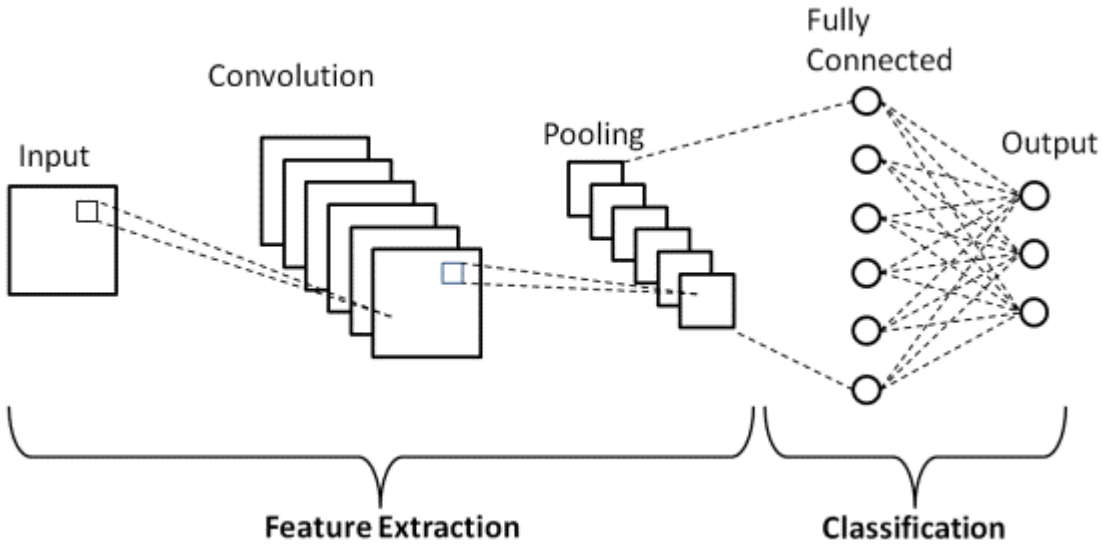

Figure 3: Outline of CNN (Balaji)

Architecture of CNN

CNN is basically used for image recognition, video analysis system, natural language processing and many more. In CNN, input layer, convolutional layer, polling layer, fully connected layer and output layer exist.

Layer	Description
Input Layer	Receives the raw pixel values of the input image, serves as the network's entry point.
Convolutional Layer	Extracts visual features through convolution operations, learning local patterns.
Pooling Layer	Reduces spatial dimensions of the feature maps, focusing on important information.
Fully Connected Layer	Transforms the high-level features into a compact representation for caption generation.
Output Layer	Generates the caption based on the learned visual features and encoded image content.

Table 1: Roles of Layers in CNN

Recurrent Neural Network (RNN)

A recurrent neural network is a class of simulated neural system where the pieces of text generated from the CNN are processed and complete semantic sentences are formed. RNN basically utilizes language demonstrating and creating content, machine translation, speech recognition, generating image description. In RNN, process sequences are different type like as one to one, one to many, many to one, many to many existing. (Susmita Das, Amara Tariq, Thiago Santos, Sai Sandeep Kantareddy & Imon Banerjee)

Long Short-Term Memory (LSTM)

Long Short-Term Memory is a part of RNN trained to memorize the term and data for the longer period. The data generated from the RNN are basically equipped and looped for the multiple of times so that the generated output remains stable.

Conclusion and Limitations

Conclusion

In conclusion, this paper presents a pioneering approach to address the visual divide and enhance accessibility for the visually impaired through image captioning using deep learning techniques.

Limitation

- The language model requires huge computing resources to train, which as a result, if the less data is trained might give the less accuracy.

- Managing the length of generated captions is challenging. The model might produce overly verbose or excessively concise descriptions, impacting the user experience.
- High-quality, well-lit images are essential for accurate object detection; blurry or low-quality images may result in inaccurate outputs.

References

Balaji, Sai. *TensorFlow*. 8 Aug 2020. <https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697>.

Benn, Matt. *Unsplash*. 24 Oct 2022. <https://unsplash.com/photos/WbgyHlntKs8>.

Kabir, Md. Humayun. "Convolutional Neural Network." *Research Gate* (2021): 1-3.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney. *The Computer Vision Foundation*. 2018. <https://openaccess.thecvf.com/content_cvpr_2018/CameraReady/1163.pdf>.

Shaikh, JalFaizy. *Analytics Vidhya*. 2 April 2018. <https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/>.

Susmita Das, Amara Tariq, Thiago Santos, Sai Sandeep Kantareddy & Imon Banerjee. "Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research." *Machine Learning for Brain Disorders* (2012): 117-138.

Wagner, Lise. *Inclusive City Maker*. n.d. <https://www.inclusivecitymaker.com/the-smartphone-a-revolution-for-the-blind-and-visually-impaired/>.