Voice Based Image Caption Generator Using Deep Learning

Second Review Presentation

Abhiram (PKD20IT001)
Akash S Babu(PKD20IT003)
Aravind M(PKD20IT014)
Dhanya P(PKD20IT020)

Guide
Prof. Ebey S Raj

GOVERNMENT ENGINEERING COLLEGE
PALAKKAD, SREEKRISHNAPURA,

Department of Information Technology
Government Engineering College Sreekrishnapuram

June 27, 2023

Outline

1. Introduction
2. Literature Survey
3. Problem Statement
4. Objectives
5. Proposed Method
6. Design
7. Model Development
8. Result
9. Conclusion
10. References

Introduction

- Deep learning is a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher level features from data.
- An image caption generator using deep learning is a powerful technology that can help people with visual impairments to better understand the content of images.

Paper 1:Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data [link...]

- Method for generating captions for images that contain novel object categories without paired training data.
- Deep Compositional Captioning (DCC) is used which consists of two main components: a compositional language model and a compositional visual model.

Paper 1: Deep Compositional Captioning [cont...]

- Compositional language model - trained on a large corpus of text data and is capable of generating syntactically and semantically correct sentences.
- Compositional visual model- trained on a large set of object and scene categories and is capable of recognizing novel object categories in images.

Paper 2:Image Caption Generation Using a Deep Architecture [link...]

- A deep learning-based approach for automatically generating natural language descriptions for images.
- A pre-trained CNN is used to extract visual features from input images.
- RNN is used to generate captions based on the extracted features.
- Gated Recurrent Unit(GRU) is used as the RNN to generate captions.

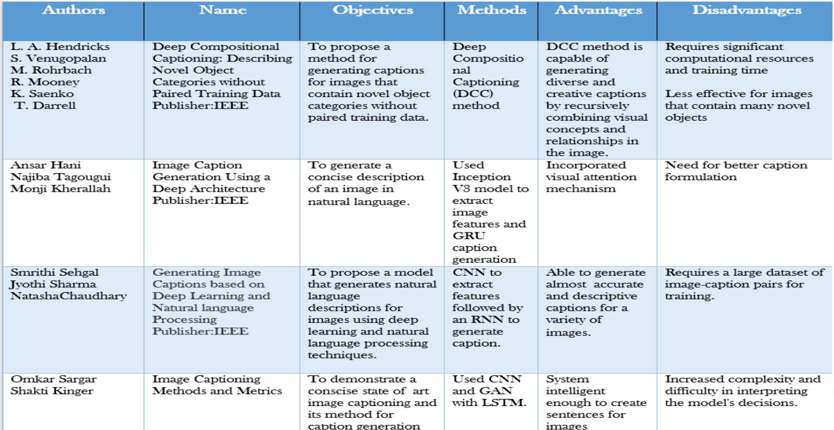
Paper 3:Generating Image Captions based on Deep Learning and Natural language Processing [link...]

- A deep learning-based approach for generating natural language descriptions for images using both convolutional neural networks (CNNs) and natural language processing (NLP) techniques.
- A pre-trained CNN is used to extract visual features from input images.
- An attention-based mechanism is used to selectively focus on important visual features while generating captions.
- A language model based on a recurrent neural network (RNN) is used to generate captions based on the extracted visual features.

Paper 4: Image Captioning Methods and Metrics [link...]

- An overview of the different techniques and evaluation metrics used in image captioning research and to highlight their advantages and disadvantages.
- It uses CNN and GAN with LSTM.
- Highlighted the importance of evaluation metrics in image captioning research and provides insights into the strengths and limitations of different metrics.

Summary of Literature Survey



Authors	Name	Objectives	Methods	Advantages	Disadvantages
L. A. Hendricks S. Venugopalan M. Rohrbach R. Mooney K. Saenko T. Darrell	Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data Publisher:IEEE	To propose a method for generating captions for images that contain novel object categories without paired training data.	Deep Compositional Captioning (DCC) method	DCC method is capable of generating diverse and creative captions by recursively combining visual concepts and relationships in the image.	Requires significant computational resources and training time Less effective for images that contain many novel objects
Ansar Hani Najiba Tagougui Monji Kherallah	Image Caption Generation Using a Deep Architecture Publisher:IEEE	To generate a concise description of an image in natural language.	Used Inception V3 model to extract image features and GRU caption generation	Incorporated visual attention mechanism	Need for better caption formulation
Smrithi Sehgal Jyothi Sharma NatashaChaudhary	Generating Image Captions based on Deep Learning and Natural language Processing Publisher:IEEE	To propose a model that generates natural language descriptions for images using deep learning and natural language processing techniques.	CNN to extract features followed by an RNN to generate caption.	Able to generate almost accurate and descriptive captions for a variety of images.	Requires a large dataset of image-caption pairs for training.
Omkar Sargar Shakti Kinger	Image Captioning Methods and Metrics	To demonstrate a conscise state of art image captioning and its method for caption generation	Used CNN and GAN with LSTM.	System intelligent enough to create sentences for images	Increased complexity and difficulty in interpreting the model's decisions.

Problem Statement

- To design an image caption generator using deep learning techniques to help visually challenged individuals to understand the content of images.

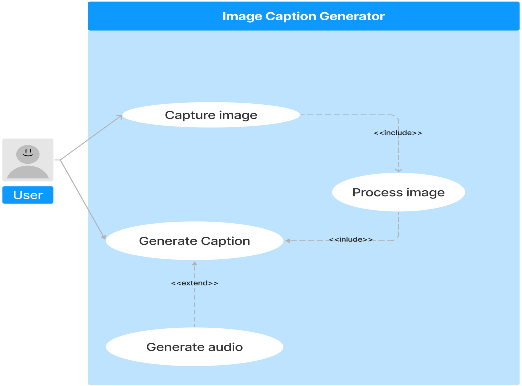
Objectives

- To build an image caption generator that takes an image as input and generates a description of the image in natural language and convert it into audio.
- To improve the performance of already existing image caption generator by using more suitable neural network.

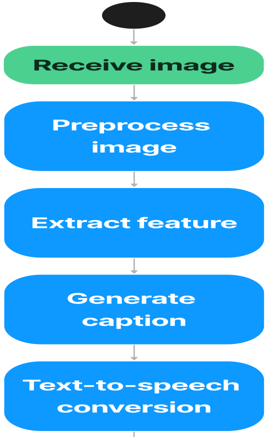
Proposed Method

- To build an image caption generator using Convolutional Neural Network(CNN) and Long Short Term Memory(LSTM).
- We use pretrained ResNet50 model for image feature extraction.
- LSTM is used for caption generation.
- We plan to use flickr8K dataset to train the model.

Use Case Diagram



Activity Diagram



Receive image

Preprocess image

Extract feature

Generate caption

Text-to-speech conversion

Hardware Requirements

- CPU: A modern multicore processor, such as an Intel Core i5 or i7, is recommended for training and deploying machine learning models.
- RAM: A minimum of 8GB of RAM is recommended, although more may be required for larger datasets.
- GPU: A powerful graphics processing unit (GPU) can significantly speed up training and inference of deep learning models. Nvidia GPUs are the most commonly used for machine learning tasks.
- Storage: A large amount of storage is required to store and process large datasets. Solid-state drives (SSDs) are recommended for fast access to data.

Software Requirements

- Development Tools: IDEs such as PyCharm, Jupyter Notebook or google colab are commonly used for machine learning development.
- Libraries: Libraries such as NumPy, Pandas, keras, tqdm, tensorflow and Matplotlib are commonly used for data manipulation, analysis, and visualization.
- Installed Anaconda packages for runnning the project.
- Installed flask web framework for connecting the model to the frontend

Model Architecture

- Image feature extractor
- Text processor
- Output predictor

Image Feature Extractor

- Image size to be passed to feature extractor is 224x224x3.
- Model uses ResNet50 pretrained on ImageNet dataset where the features of the image are extracted just before the last layer of classification.
- Another dense layer is added and converted to get a vector of length 2048.

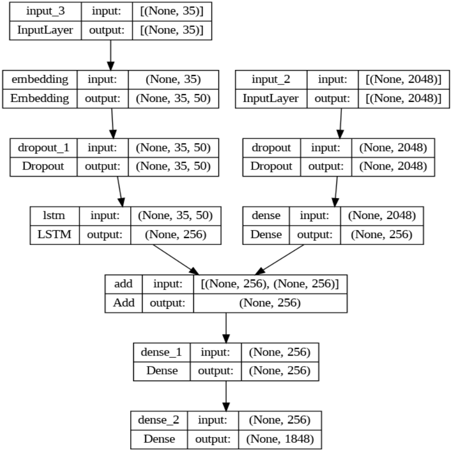
Text Processor

- Layer contains the word embedding layer to encode text data.
- Long Short-Term Memory (LSTM) in RNN is added with 256 memory units.

Output Predictor

- Output vector from both the image feature extractor and the text processor are of same length and a decoder merges both the vectors using an addition operation.
- This is then fed into two dense layers.
- The first layer is of length 256 and the second layer makes a prediction of the most probable next word in the caption.

Model



input_3 | input: | [(None, 35)]
InputLayer | output: | [(None, 35)]

embedding | input: | (None, 35)
Embedding | output: | (None, 35, 50)

input_2 | input: | [(None, 2048)]
InputLayer | output: | [(None, 2048)]

dropout_1 | input: | (None, 35, 50)
Dropout | output: | (None, 35, 50)

dropout | input: | (None, 2048)
Dropout | output: | (None, 2048)

lstm | input: | (None, 35, 50)
LSTM | output: | (None, 256)

dense | input: | (None, 2048)
Dense | output: | (None, 256)

add | input: | [(None, 256), (None, 256)]
Add | output: | (None, 256)

dense_1 | input: | (None, 256)
Dense | output: | (None, 256)

dense_2 | input: | (None, 256)
Dense | output: | (None, 1848)

Result

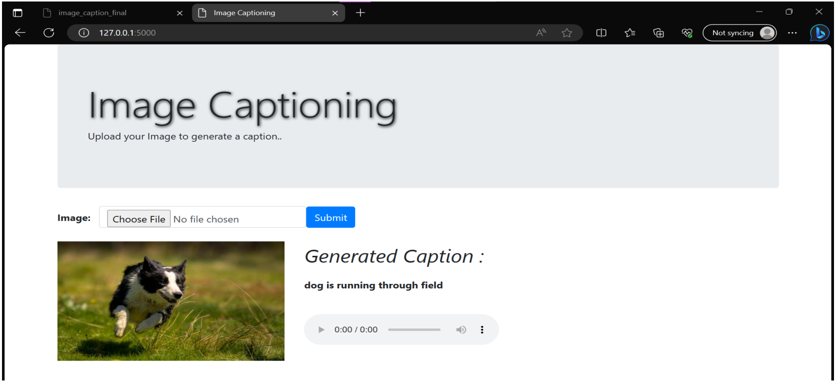


Image Captioning

Upload your Image to generate a caption..

Image:

Generated Caption :

dog is running through field

Result

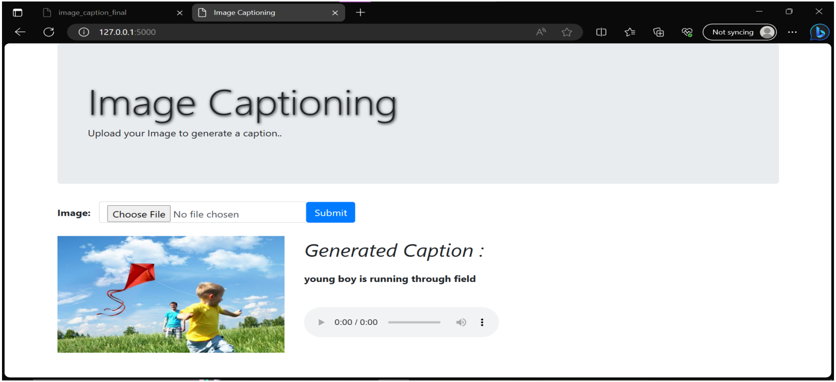


Image Captioning

Upload your Image to generate a caption..

Image: Choose File No file chosen Submit

Generated Caption :

young boy is running through field

Result

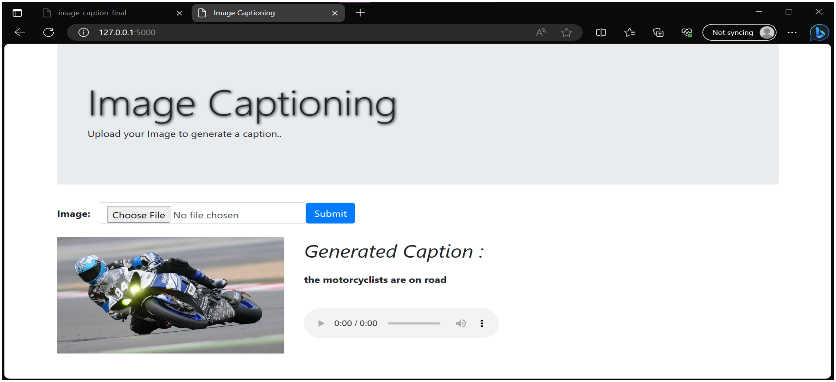


Image Captioning

Upload your Image to generate a caption..

Image: Choose File No file chosen Submit

Generated Caption :

the motorcyclists are on road

Conclusion

- Implemented the 'Merge' model architecture.
- This is then fed into two dense layers.
- Resnet50 network worked better as expected.
- Using larger datasets such as MS-COCO,Flickr 30k can yield a better robust model.
- We completed 187 epochs with an accuracy of 50 percent.
- Increasing the number of epochs could increase the accuracy and decrease the loss.

References

- L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1-10, doi: 10.1109/CVPR.2016.8.
- A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.

References

- S. Sehgal, J. Sharma and N. Chaudhary, "Generating Image Captions based on Deep Learning and Natural language Processing," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 165-169, doi: 10.1109/ICRITO48877.2020.9197977.
- O. Sargar and S. Kinger, "Image Captioning Methods and Metrics," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 522-526, doi: 10.1109/ESCI50559.2021.9396839.

Questions

Thank You