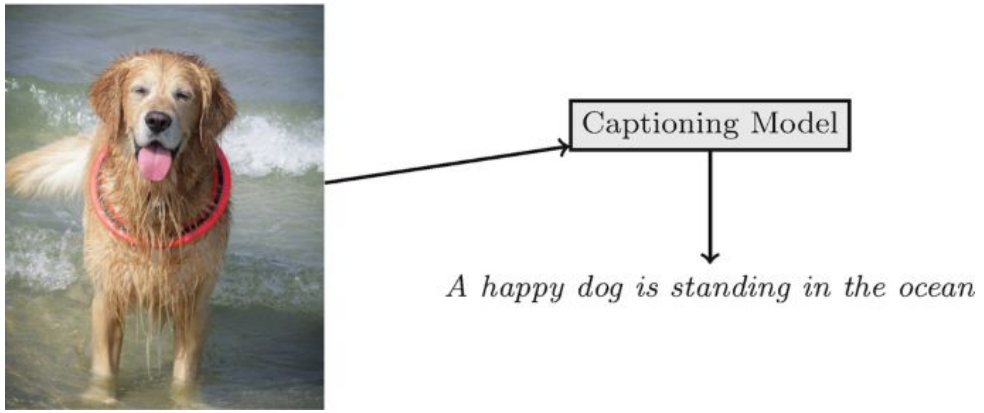
UE20CS302: Machine Intelligence
Project Phase – 1



Project Title : Image captioning
Project ID : 18
Project Team : Jeffrey (651), Ishita(648), Hita(645)

Abstract and Scope

- Well defined problem statement.

  To generate textual descriptions for a given image using the techniques of Natural Language Processing and Computer Vision.

- Abstract and Scope

-> Generating a description of an image is called image captioning. Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. It also needs to generate syntactically and semantically correct sentences.

-> Image captioning has a wide range of applications such as aiding visually and speech impaired people, Storage minimisation .Image captioning is the first step towards video captioning which can be used for human robot interaction , video surveillance etc.

Paper 1

PES UNIVERSITY

Literature Survey

1. A Comprehensive Survey of Deep Learning for Image Captioning:

   Dataset: MS COCO

   This paper by MD. Zakir Hossain, highlights the various deep learning models that can be used for captioning an image based on it's salient features. Traditional machine learning models used handcrafted features which were task specific and extracting features from a large and diverse set of data was not feasible. Moreover, real-world data such as images and video are complex and have different semantic interpretations. Instead, deep learning based techniques automatically learn the feature from the training data and can handle a large and diverse set of images and videos. The authors broadly classify the image captioning techniques in 3 main categories: (1) template-based image captioning, (2) retrieval-based image captioning, and (3) novel image caption generation. Template-based image captioning uses fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled.

Retrieval-based image captioning involves captioning by retrieving a set of captions of other visually similar images in the training dataset. Novel image caption generation generates captions from both visual space and multimodal space. A visual space is generated from the visual content and a caption is generated using a language model. These captions are more accurate than the previous 2 methods hence novel image captioning is highly used. Most deep-learning-based image captioning methods fall into the category of novel caption generation. Some of them are:

Paper 1

Literature Survey

- Visual space based - For explicit mapping from images to descriptions. The image features and the corresponding captions are independently passed to the language decoder.
- Multimodal space based - Incorporates implicit vision and language models. A shared multimodal space is learned from the images and the corresponding caption text which is then fed into the language decoder.
- Supervised learning - Each image comes with a label. The model learns on training images and their captions. The caption generated for test images is compared with the actual label to get the accuracy of the model.
- Dense captioning- Instead of using the whole image at once to generate a caption, this method divides the image into many regions to obtain information of various objects to generate the caption for the entire image.
- Whole scene based- Uses the entire image at once to generate a caption for the image.
- Encoder-decoder architecture based- In this network, global image features are extracted from the hidden activations of CNN and then fed into an LSTM to generate a sequence of words.
- Compositional architecture based- This method uses CNN for feature extraction from images, then understands the visual space. Multiple captions are generated by analyzing from various different subregions in the given image and a similarity model is used to rank the generated captions to find the most suitable one.
- LSTM (Long Short-Term Memory)language model based- Like RNN and used for maintaining information for long periods of time in memory.

and many more.

The above methods are explained in detail in the paper.

Paper 2

PES UNIVERSITY

Literature Survey

2. Show and Tell: MSCOCO Image Captioning Challenge 2015

dataset: MSCOCO

Recent advances in statistical machine translation have shown that, given a powerful sequence model, it is possible to achieve state-ofthe-art results by directly maximizing the probability of the correct translation given an input sentence in an “end-toend” fashion – both for training and inference. These models make use of a recurrent neural network which encodes the variable length input into a fixed dimensional vector, and uses this representation to “decode” it to the desired output sentence.

In his study, Oriol Vinyals presents NIC, an end-to-end neural network system that can automatically view an image and generate a reasonable description in. NIC is based on a CNN that encodes an image into a compact representation, followed by a RNN that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image.

Paper 3

PES UNIVERSITY

Literature Survey

3. Image Captioning: Transforming Objects into Words

dataset: MSCOCO

Summary:

-> The writers have used information about spatial relationships between detected objects such as size and relative position

-> most authors have not used this information

-> the writers of this paper were able to show that the use of spatial relationships by adding an object relation model in the encoder, they were able to increase the scores of almost every metric

-> their model is the same as any other encoder decoder model but they have included the object relation model as well

-> doesn't show a lot of improvement on the METEOR score but shows significant improvement in the CIDEr-D, BLEU-1 and ROUGE-L scores

-> object detection model (input is an image, the output is 2048 dimensional feature vector for each bounding box)

Limitations:

-> geometric attention was incorporated only in the encoder

-> object detection model did not detect all objects

-> some relations were not detected while some were incorrectly detected

PES UNIVERSITY

Literature Survey

4. A Systematic Literature Review on Image Captioning:

Dataset: Flickr30k or MS COCO datasets

This paper is a systematic literature survey of the recent methods providing a brief overview of improvements in image captioning over the last four years. The main focus of the paper is to explain the most common techniques and the biggest challenges in image captioning and to summarize the results from the newest papers (mainly from 2016-2019). Inconsistent comparison of results achieved in image captioning was noticed during this study and hence the awareness of incomplete data collection is raised in this paper. The authors have studied models from different articles that are chiefly based on either Flickr30k or MS COCO datasets.

Most of the models rely on the widespread encoder–decoder framework, which is flexible and effective. Sometimes it is defined as a structure of CNN + RNN. Usually a convolutional neural network (CNN) represents the encoder, and a recurrent neural network (RNN) the decoder. The encoder is the one which “reads” an image—given an input image, it extracts a high-level feature representation. The decoder is the one which generates words—given the image representation from the encoder (encoded image), it generates words to represent the image with a full grammatically and stylistically correct sentence. The distribution of each year’s results based on the 6 main metrics - BLEU-1, BLEU-2, BLEU-3, BLEU-4, CIDEr, METEOR; is beautifully presented. This systematic literature review summarizes all the newest articles and their results in one place in order to prevent good ideas from getting lost and to increase fair competition among the new models created. Future studies should consider if static models are good enough when thinking of long term application or if lifelong learning should be increasingly thought of.

PES UNIVERSITY

Literature Survey

5. Unsupervised Image Captioning

dataset: Open Images Dataset V7

CNNs usually act as an image encoder, while RNNs are naturally suitable for encoding sentences. Due to their different structures and characteristics, the encoders of images and sentences cannot be shared in unsupervised machine translation.

First, a language model is trained on the sentence corpus using the adversarial text generation method which generates a sentence conditioned on a given image feature. In an unsupervised setting, the ground-truth caption of a training image is not available. Therefore, adversarial training is employed to generate sentences such that they are indistinguishable from the sentences within the corpus.
Second, in order to ensure that the generated captions contain the visual concepts in the image, the knowledge provided by a visual concept detector is used by the image captioning model. Specifically, a reward will be given when a word, which corresponds to the detected visual concepts in the image, appears in the generated sentence.
Third, to encourage the generated captions to be semantically consistent with the image, the image and sentence are projected into a common latent space. Given a projected image feature, a caption is decoded, which is used to reconstruct the image feature. Similarly, a sentence from the corpus is encoded to the latent space feature and thereafter reconstructs the sentence. By performing bi-directional reconstructions, the generated sentence is forced to closely represent the semantic meaning of the image.
The image captioning model initialisation pipeline is developed to overcome the difficulties in training. The concept words in a sentence are taken as input and used to train a concept-to-sentence model using the sentence corpus only. The visual concept detector is used to recognize the visual concepts present in an image. Integrating these two components together, a pseudo caption is generated for each training image. The pseudo-image-sentence pairs are used to train a caption generation model in the standard supervised manner.

Paper 6 & 7

Literature Survey

6. Image Captioning with Semantic Attention

dataset: MS COCO and Flikr 30k

In this work, the writers propose a novel method for the task of image captioning, which achieves state-of-the-art performance across popular standard benchmarks. Different from previous work, their method combines top-down and bottom-up strategies to extract richer information from an image, and couples them with a RNN that can selectively attend on rich semantic attributes detected from the image. Their method, therefore, exploits not only an overview understanding of input image, but also abundant fine-grain visual semantic aspects. The real power of their model lies in its ability to attend on these aspects and seamlessly fuse global and local information for better caption.

limitations:

-> irrelevant visual attributes may disrupt the model

7.Image Captioning Based on Deep Neural Networks

Dataset: MS COCO

With the development of deep learning, the combination of computer vision and natural language processing has aroused great attention in the past few years. Image captioning is a representative of this field, which makes the computer learn to use one or more sentences to understand the visual content of an image. The meaningful description generation process of high level image semantics requires not only the recognition of the object and the scene, but the ability of analyzing the state, the attributes and the relationship among these objects. Though image captioning is a complicated and difficult task, a lot of researchers have achieved significant improvements. In this paper, we mainly describe three image captioning methods using the deep neural networks: CNN-RNN based, CNN-CNN based and Reinforcement-based framework. This paper introduces the representative work of these three top methods respectively using the COCO dataset, describes the evaluation metrics (BLEU, METEOR, ROUGE, CIDEr, and SPICE) and summarizes the benefits and major challenges.

PES UNIVERSITY

Literature Survey

8. Learning to Evaluate Image Captioning
dataset: COCO and Flickr8 Dataset

This paper highlights the observation that current image captioning models are usually evaluated with automatic metrics instead of human judgments. Commonly used evaluation metrics BLEU, METEOR, ROUGE and CIDEr are mostly based on n-gram overlap and tend to be insensitive to semantic information.
ROUGE - It is essentially a set of metrics for evaluating the automatic summarization of texts. It works by comparing an automatically produced summary or translation against a set of reference summaries (which are human-produced) using precision and recall.

BLEU - The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair. The comparison is made regardless of word order.

METEOR - Metric for Evaluation of Translation with Explicit ORdering is a machine translation evaluation metric, which is calculated based on the harmonic mean of precision and recall, with recall weighted more than precision. It is based on a generalized concept of unigram matching between machine-produced translations and human-produced reference translations. Unigrams can be matched based on their stemmed forms and meanings. Once all generalized unigram matches between the two strings have been found, METEOR computes a score for this matching using a combination of unigram-precision, and unigram-recall.
It captures and works well against the downfalls of BLEU metric.

All above mentioned metrics rely solely on similarity between candidate and reference captions, without taking the image into consideration. On the other hand, the proposed method takes an image feature as input, learns to score candidate captions by training to distinguish positive and negative examples (setup of Turing Test).

PES UNIVERSITY

Literature Survey

9. Image Captioning with Deep Bidirectional LSTMs

dataset: flikr8k, flikr 30k and MSCOCO

summary:

This work presents an end-to-end trainable deep bidirectional LSTM (Long-Short Term Memory) model for image captioning. Their model builds on a deep convolutional neural network (CNN) and two separate LSTM networks. It is capable of learning long term visual-language interactions by making use of history and future context information at high level semantic space. Two novel deep bidirectional variant models, in which they increase the depth of nonlinearity transition in different way, are proposed to learn hierarchical visual-language embeddings.

They further designed deep bidirectional LSTM architectures to embed image and sentence at high semantic space for learning visual-language models. They also qualitatively visualized internal states of proposed model to understand how multimodal LSTM generates word at consecutive time steps. The effectiveness, generality and robustness of proposed models were evaluated on numerous datasets. Their models achieve highly state-of-the-art results on both generation and retrieval tasks.

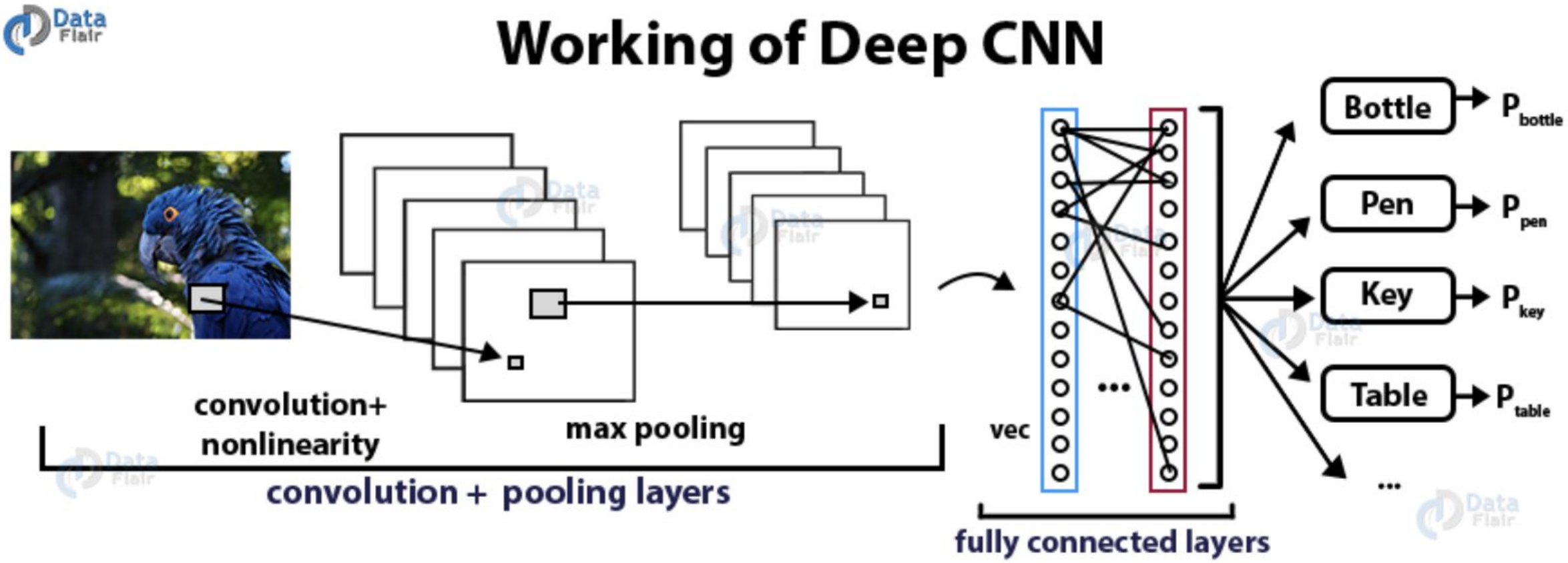
What is the design approach followed? And Why?
Benefits of this approach & are there any drawbacks?
Alternate design approaches, if any.

Design Approach

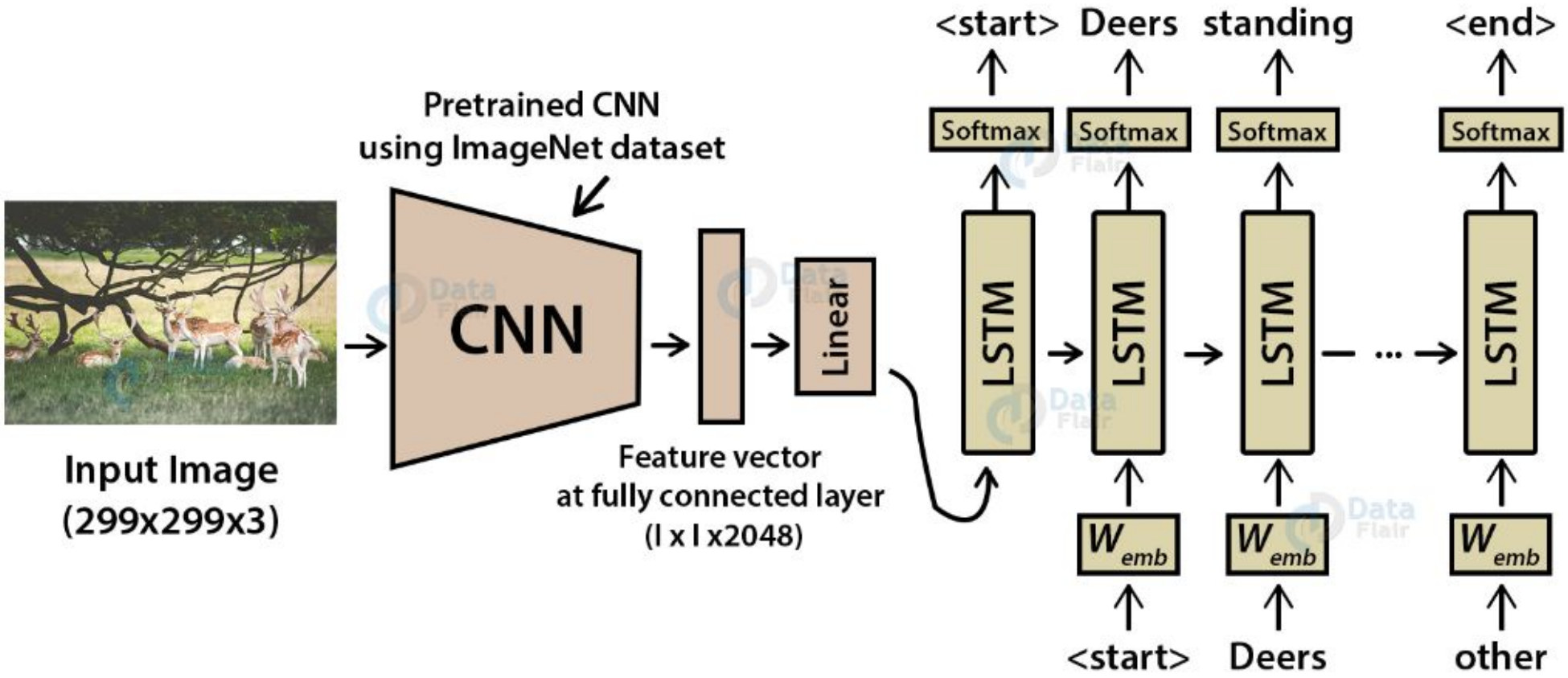
- A CNN and LSTM model is used in build a working model of Image caption generator by implementing CNN with LSTM.
- The image features will be extracted from Xception which is a CNN model trained on the imagenet dataset and then we feed the features into the LSTM model which will be responsible for generating the image captions.
- <u>Alternate design approaches:</u>

A Midge system based on maximum likelihood estimation, which directly learns the visual detector and language model from the image description dataset. First analyze the image, detect the object, and then generate a caption. Words are detected by applying a convolutional neural network (CNN) to the image area and integrating the information with MIL. The structure of the sentence is then trained directly from the caption to minimize the priori assumptions about the sentence structure. Finally, it turns an image caption generation problem into an optimization problem and searches for the most likely sentence.

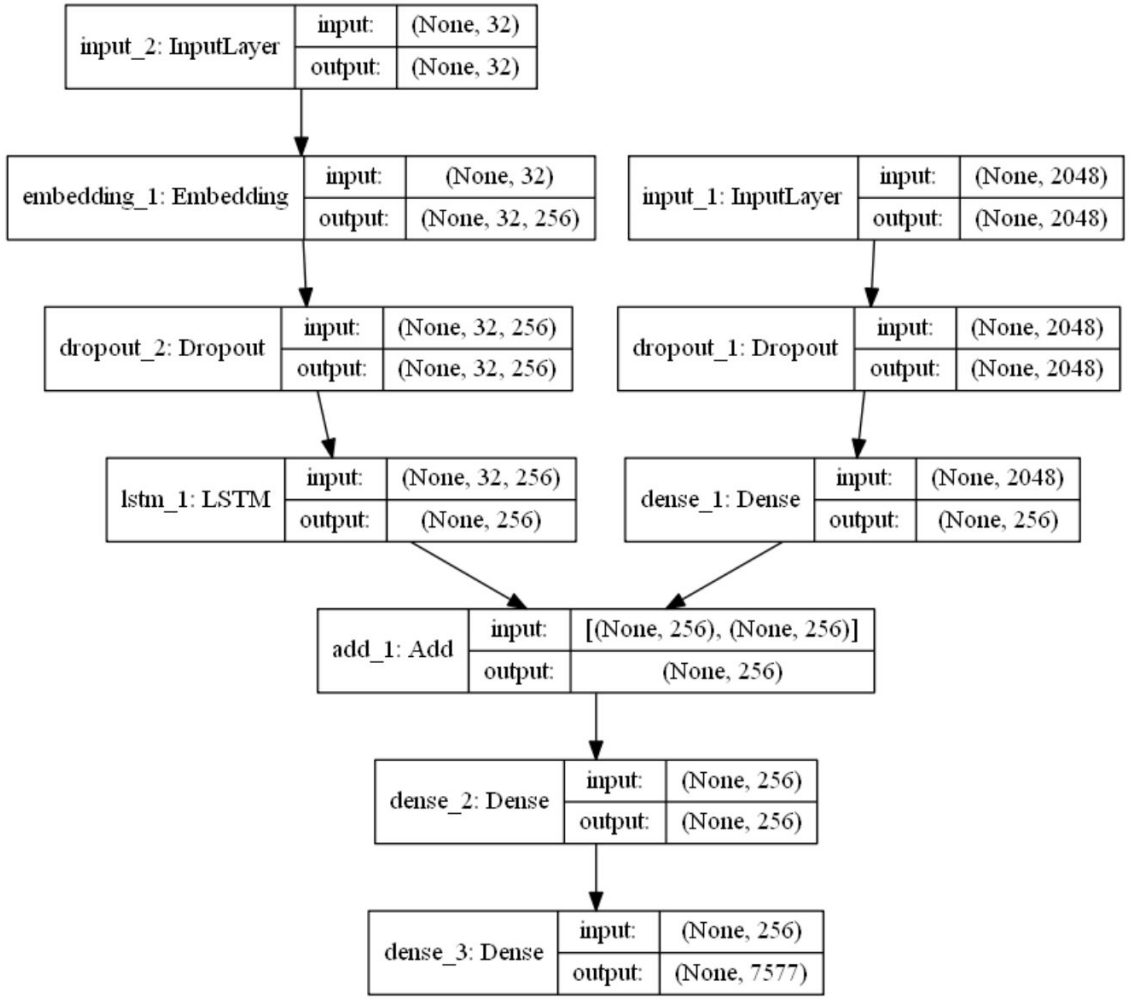
Architecture



Model - Image Caption Generator

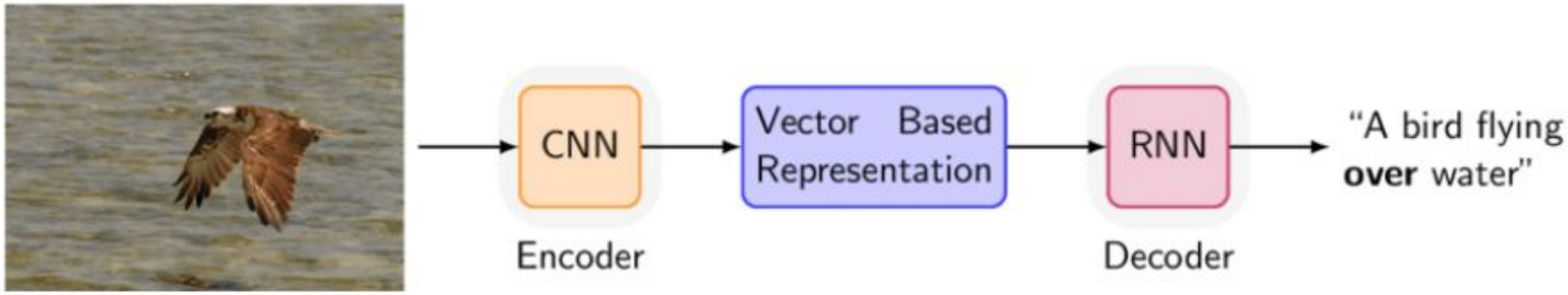


Architecture



Architecture

Provide high-level design view of the system.



Target Users:

- Visually impaired people
- People well versed with the English language
- Kids

Applications

- Helping the blind visualise a scene
- First step to video captioning
- Learning tool for kids

Technologies Used

- Python: Most popular language that suits for ML
- TensorFlow and Keras: For preprocessing input, loading images, tokenizer, embedder, dropout, load models - LSTM
- Xception:to use CNN Pre trained on ImageNet.
- Pillow: for basic image processing functions.
- NumPy: storing pixels as integers in an array, expanding dimensions.
- Pandas: Used for data manipulation and storing and retrieving files.
- Pickle: storing, saving and loading model.
- Matplotlib: Plot graphs.
- Google Colab: Easy to configure interactive environment to share and run code, with free GPU access.
- Google Drive: To make large storage dataset available amongst team.

Project Demo

- Demonstrate the Implemented work if available.

Project Progress

What is the project progress so far?

- Problem statement has been defined
- Literature review has been done
- Implementation of Image captioning using basic encoder-decoder model has been done
- Qualitative testing

References

1. Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning." *ACM Computing Surveys (CsUR)* 51, no. 6 (2019): 1-36.
2. Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 4 (2016): 652-663.
3. Herdade, Simao, Armin Kappeler, Kofi Boakye, and Joao Soares. "Image captioning: Transforming objects into words." *Advances in Neural Information Processing Systems* 32 (2019).
4. Staniūtė, Raimonda, and Dmitrij Šešok. "A systematic literature review on image captioning." *Applied Sciences* 9, no. 10 (2019): 2024.
5. Feng, Yang, Lin Ma, Wei Liu, and Jiebo Luo. "Unsupervised image captioning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4125-4134. 2019.
6. You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651-4659. 2016.
7. Liu, Shuang, Liang Bai, Yanli Hu, and Haoran Wang. "Image captioning based on deep neural networks." In *MATEC Web of Conferences*, vol. 232, p. 01052. EDP Sciences, 2018.
8. Cui, Yin, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. "Learning to evaluate image captioning." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5804-5812. 2018.
9. . Wang, Cheng, Haojin Yang, Christian Bartz, and Christoph Meinel. "Image captioning with deep bidirectional LSTMs." In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 988-997. 2016.

IEEE Format


Thank
You