CS6301 MACHINE LEARNING – MINI PROJECT

SRIHARI. S T.K.S. ARUNACHALAM

2018103601 2018103616

IMAGE INSCRIPTION AND INTONATOR - A NEURAL NETWORK APPROACH

DESIGN DOCUMENTATION

DATASET USED: MS-COCO Dataset (https://cocodataset.org/#download)

ABSTRACT:

Digital communication technologies have greatly influenced and expanded the way humans interact. The progress of information technology has opened wider opportunities for communication. Social networks have become the modern-day social communities connecting people from different parts of the globe, sharing images and videos on these platforms. By creating virtual communities, digital communication has expanded the scope of communication eliminating barriers. We aim to make further progress in this arena by describing an image in the form of audio to visually impaired people.

A certain section of differently abled people is unfortunately isolated from this world. In-order to combat this issue we have come up with a system that describes an image shown in the form of plain text using an encoder-decoder architecture and is integrated with an end-to-end lexical articulator which produces a vocal description of the given image.

INTRODUCTION:

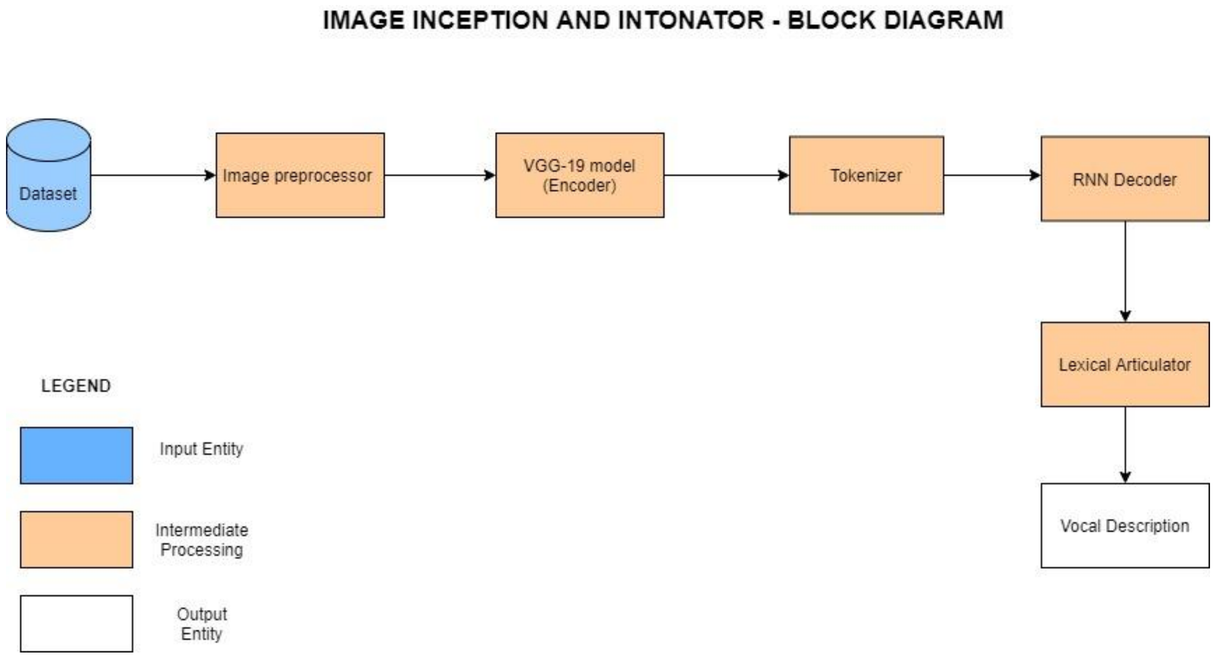
Technology has become an integrated part of our daily lives over the past decades. The efficient processing and association of different multimodal information is a very important research field with a great variety of applications, such as human computer interaction, knowledge discovery, document understanding, etc. Computerized elucidation of an image has been one of the primary goals of computer vision. Not only must description generator models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language.

Hence, to tackle this conundrum we present the development of a novel methodology to extract meaningful information from images, in the form of short descriptions. The results can be further run through a lexical articulator engine to offer full sustainability. This way, a full independent experience could be delivered to visually impaired people.

For an accurate algorithm, it must be fed with a good amount of data. The dataset which we are planning to use for training and testing the model is MS-COCO (Microsoft-Common Objects in Context) which contains more than 80000 labelled images.

For converting still images into natural language text sentences we must first start from the understanding of the context of an image and secondly how this context is expressed into natural language. Thus, for understanding the image a feature extracting convolutional neural network can be used. With the aid of recurrent neural networks, the extracted features can be transformed into a suitable textual description. These results are finally fed to a lexical articulator module which produces the output of the system in the form of speech.

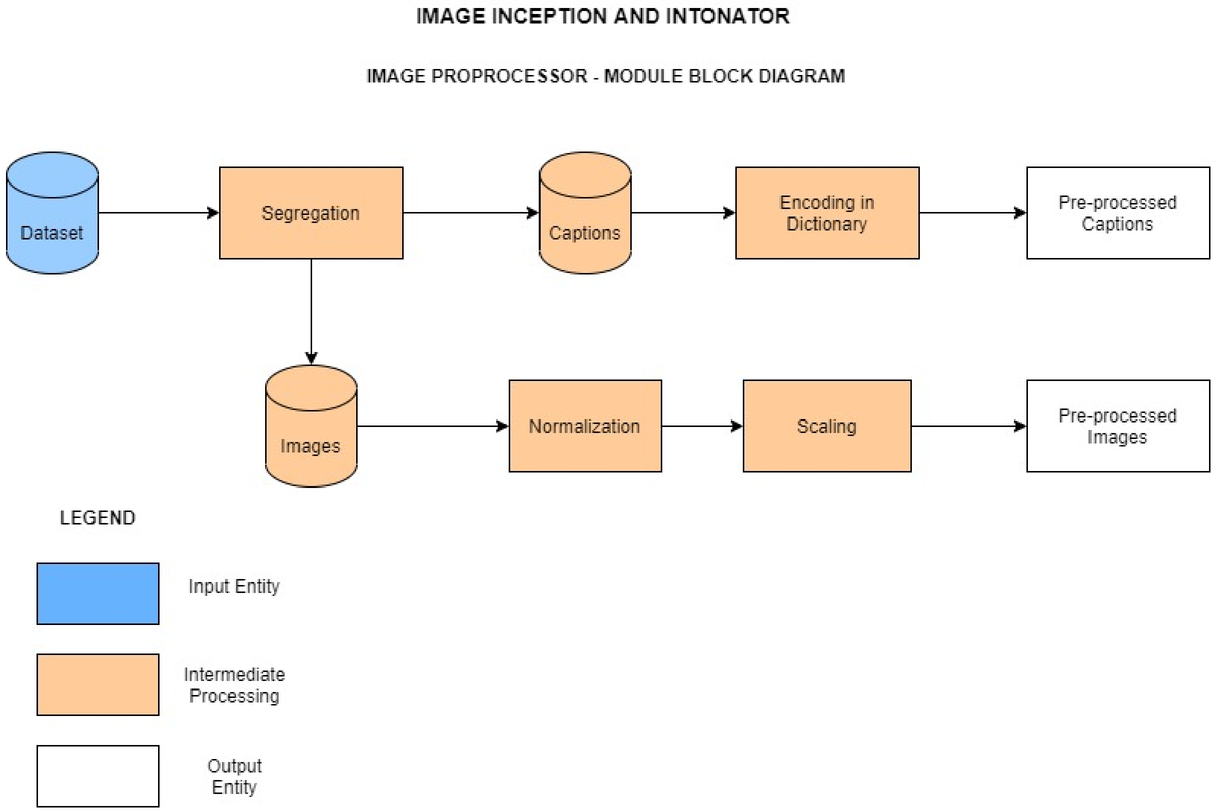
SYSTEM ARCHITECTURE:



The system is modularized into five sections as shown in the block diagram.

Module	Input	Output
Image Pre-processor	MS-COCO Dataset	Pre-processed Images and Captions
VGG-19 Model (Encoder)	Pre-processed input image	Features of the input image
Tokenizer	Captions for the image from the dataset	Sequence of real valued vectors (embeddings)
RNN Decoder	Features of the input image and embeddings from the tokenizer	Textual Description of the image
Lexical Articulator	Textual Description of the image	Vocal Description of the input image

IMAGE PRE-PROCESSOR:

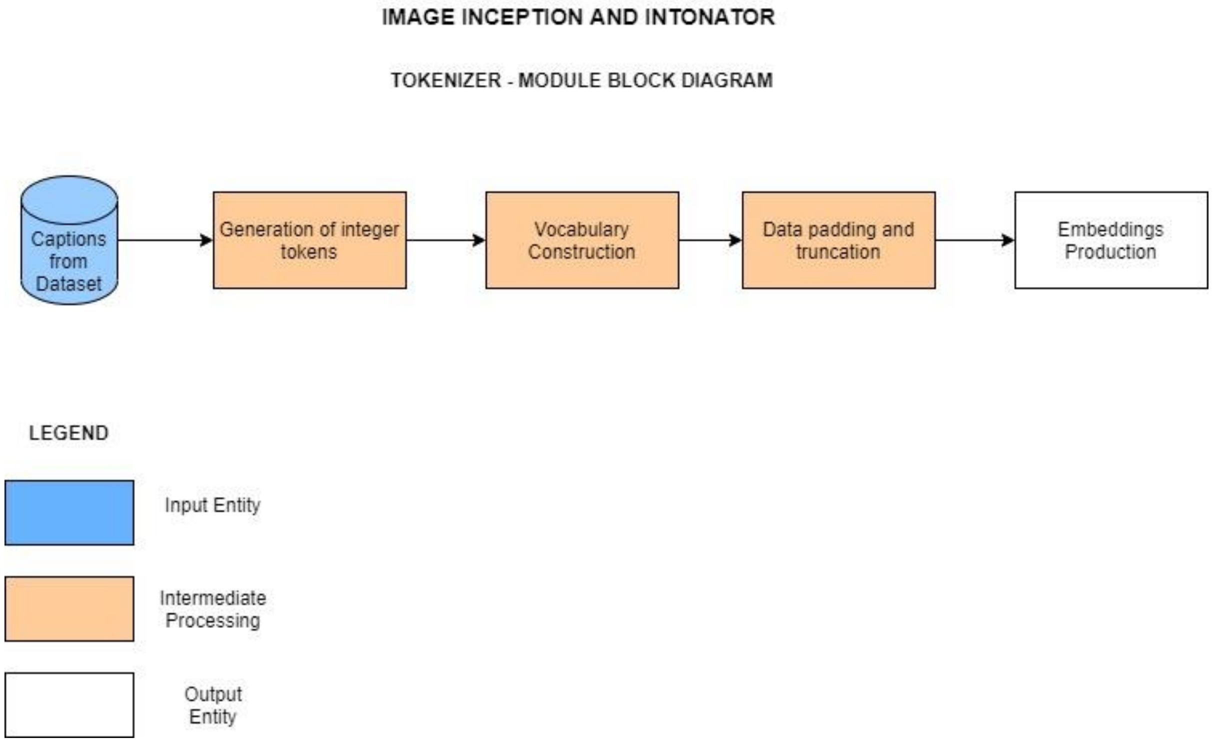


The MS-COCO dataset which consists of both images and captions together is loaded. The images and the corresponding captions are then segregated and stored in two separate files. The images then undergo normalization followed by scaling to finish the pre-processing. On the other hand, the captions are encoded in a dictionary and are thus pre-processed so that it could be used by the tokenizer.

VGG-19 MODEL (ENCODER):

The pre-processed image is fed into the VGG-19 model which consists of 19 layers (16 convolution layers, 3 Fully connected layers, 5 MaxPool layers and 1 SoftMax layer). The VGG-19 model was trained on the ImageNet challenge (ILSVRC) 1000-class classification task. This model is used for the purpose of encoding the image and later using the features for image elucidation. The network takes a (224, 224, 3) RGB image as the input. The output of the penultimate layer i.e. the 3rd fully connected layer is fed as the input for the next module.

TOKENIZER:



Neural Networks cannot work directly on text-data. We use a two-step process to convert text into numbers that can be used in a neural network. Before we start processing the text, the beginning and end of each text-sequence are marked in-order to keep track of the captions. The first step is to convert text-words into integer-tokens. We pad all the token-sequences with zeros so they all have the same length and can be input to the recurrent neural network. The second step is to convert integer-tokens into vectors of floating-point numbers using an embedding-layer. This is necessary because the integer-tokens may take on values between 0 and the vocabulary size. But the RNN cannot work on values in such a wide range. This sequence of real valued vectors is now passed onto the next module for further processing.

RNN DECODER:

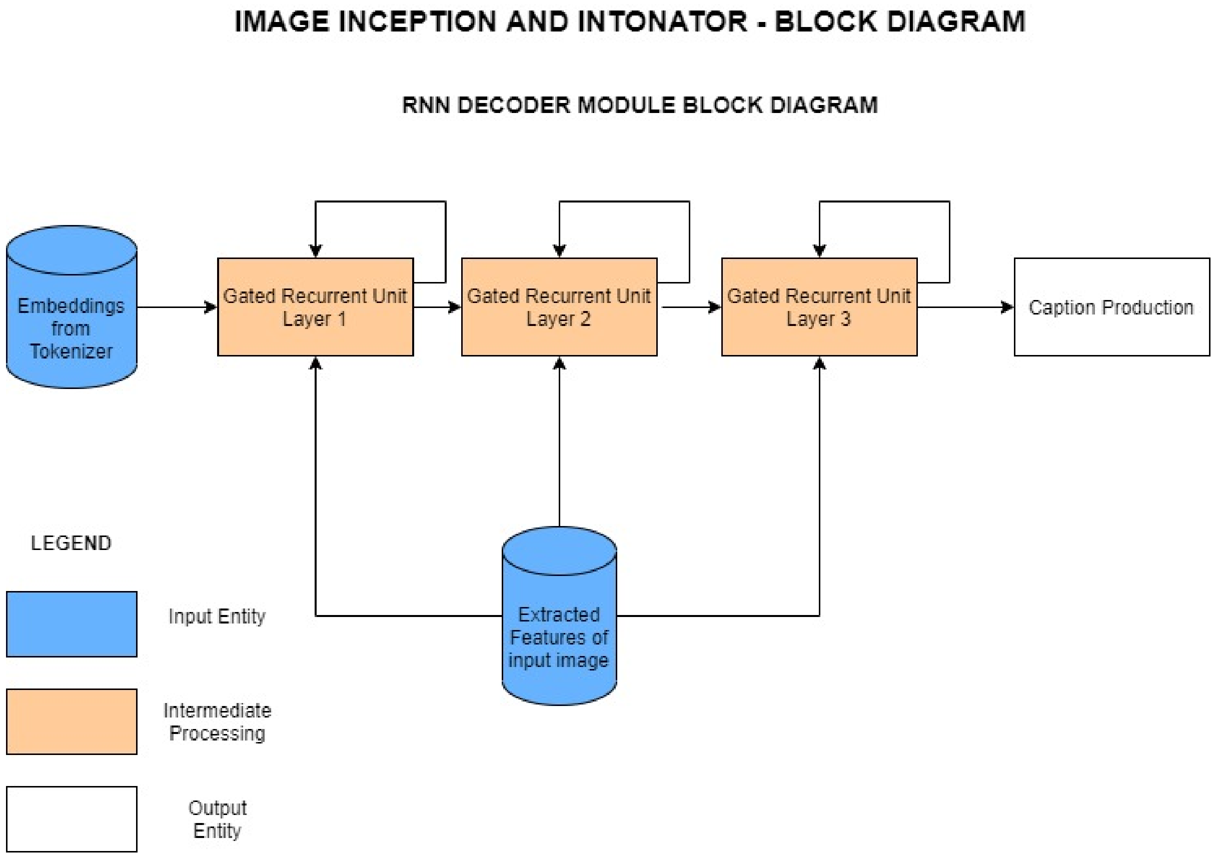


IMAGE INCEPTION AND INTONATOR - BLOCK DIAGRAM
RNN DECODER MODULE BLOCK DIAGRAM
Embeddings from Tokenizer
Gated Recurrent Unit Layer 1
Gated Recurrent Unit Layer 2
Gated Recurrent Unit Layer 3
Caption Production
Extracted Features of input image
LEGEND
Input Entity
Intermediate Processing
Output Entity

Recurrent neural networks are used for decoding the extracted features from the image. Gated Recurrent Unit (GRU) is a sophisticated recurrent unit used to capture dependencies of various time scales, process memories of sequential data by storing previous inputs in the internal state of networks and plan from the history of previous inputs to target vectors in principle. GRU consists of Update and reset gates, these gates are responsible for regulating the information to be kept or discarded at each time step.

Reset Gate:

$$gate_{reset} = \sigma(W_{input_{reset}} \cdot x_t + W_{hidden_{reset}} \cdot h_{t-1})$$

$$r = tanh(gate_{reset} \odot (W_{h_1} \cdot h_{t-1}) + W_{x_1} \cdot x_t)$$

Update Gate:

$$gate_{update} = \sigma(W_{input_{update}} \cdot x_t + W_{hidden_{update}} \cdot h_{t-1})$$

$$u = gate_{update} \odot h_{t-1}$$

Combining outputs:

$$\underline{h_t = r \odot (1 - gate_{update}) + u}$$

LEXICAL ARTICULATOR:

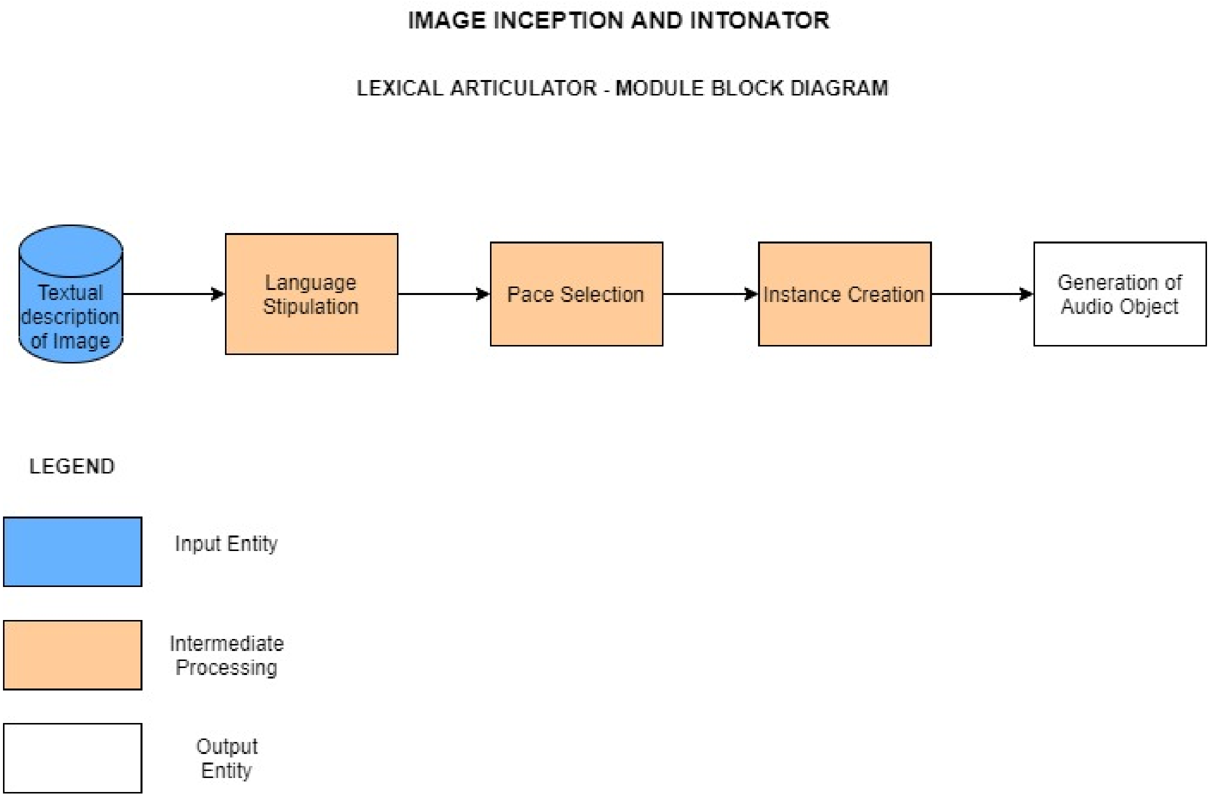


IMAGE INCEPTION AND INTONATOR
LEXICAL ARTICULATOR - MODULE BLOCK DIAGRAM
Textual description of Image
Language Stipulation
Pace Selection
Instance Creation
Generation of Audio Object
LEGEND
Input Entity
Intermediate Processing
Output Entity

The final module of this system brings into play a state-of-the-art Lexical Articulator. Being fed with the textual description of images from the decoder, the language to be translated into is stipulated. Further we specify the pace of the vocal description we wish to obtain. Next, we create an mp3 instance of the description using the gtts package. The terminal step in this module is the creation of an audio object which produces the required vocal output and enables us to play the articulated description.

NOVELTY:

- Inclusion of a new module which converts image descriptions into speech using a state-of-the-art Lexical Articulator which provides access of this system to a wider section of the society.
- Improvement in the precision of the output by using the VGG-19 model.
- Implementation of parallel processing of the most computationally expensive sections using Map Reduce programming which could significantly reduce the overall computational duration.

REFERENCES:

- A. Puscasiu, A. Fanca, D. Gota and H. Valean, "Automated image captioning," 2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Cluj-Napoca, Romania, 2020, pp. 1-6, doi: 10.1109/AQTR49680.2020.9129930.
- 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.
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention: Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015.
- N. Bourbakis, "Automatic Image-to-Text-to-Voice Conversion for Interactively Locating Objects in Home Environments," 2008 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, OH, USA, 2008, pp. 49-55, doi: 10.1109/ICTAI.2008.123.
- Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, "Image Captioning with Semantic Attention," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4651-4659, doi: 10.1109/CVPR.2016.503.