[Team 32] ProjF Proposal: Image Captioning using Deep Neural Networks

LAXMI AISHWARYA THELA
Email: lthela@ncsu.edu

CHAITANYA RAJEEV
Email: crajeev@ncsu.edu

SAI SHASHANK NAYABU
Email: snayabu@ncsu.edu

I. MOTIVATION

We want to build a model that can generate meaningful descriptive sentences for a given image. Mimicking Human Intelligence to generate a description of an image by a machine is itself a remarkable step in the field of Artificial Intelligence. Captioning is an interesting problem that has various applications such as accurate and compact information of images shared on social media, helps visually impaired people to better understand the content, and other Natural Language processing applications.

The main challenge of the project is to capture how objects in an image relate to each other and then express them in a natural language. Generally, Computer Systems use pre-defined hardcoded Template-based or retrieval-based image captioning for generating a description for an image, but this traditional approach does not provide sufficient variety for generating lexically rich text descriptions. The increased efficiency of neural networks has suppressed this shortcoming. We will be using dual techniques from Computer Vision to understand the content of the image and then use a language model to generate a meaningful sentence from the words generated.

II. DATASET DESCRIPTION

For this project we plan to train and test our model on two primary datasets, the Microsoft COCO captioned dataset and the Flickr8k and Flickr30k datasets. Both datasets contain images of commonly occurring context with a relevant caption/label for each image. The COCO dataset has around 82,000 images and the flickr8k dataset contains around 8,000 images which are captioned.

III. METHODOLOGY

This section aims to describe the high-level methods and models used to train and build our system.

The model will comprise of two distinct sub-architectures, a CNN architecture for processing images and an RNN architecture for processing and proposing words and phrases. The output of the CNN will be the input of the RNN which will predict a caption for the image. A common mapping for pictures and words will have to be implemented for this.

A. Convolutional Neural Network(CNN):

A Convolutional Neural Network consists of layers of filters, pooling layers and traditional DNN layers towards the tail-end of the architecture. This helps in identifying similar edges or features across multiple locations within the image. The activation function used in CNNs is the Linear Rectified Units function (ReLU). Dropout regularization will also be implemented to prevent overfitting. Many standard architectures including AlexNet, ResNet and VGG will be tried to ascertain the best performing architecture for the problem.[3]

B. Recurrent Neural Network(RNN):

Recurrent Neural Network is a type of Neural Network architecture best used for model sequential data where context from prior data has some correlation to the occurrence of future data. Common examples for sequential modelling including modelling of time series data, language tasks and sound/audio modelling tasks. For the Recurrent Neural Network sub-architecture, we intend to try out a GRU (Gated Recurrent Unit) and an LSTM (Long Short-term memory) architecture.[3]

C. Sentence Generation:

As outlined in the Image Caption Generator using Deep Neural Networks paper [1], the beam search algorithm can be used for sentence generation.

IV. EVALUATION

We intend to report precision using the BLEU (Bilingual Evaluation Understudy) score. It is one of the first and popular metrics to have a high correlation with human judgments and quality. BLEU compares the n-grams of a candidate with n-grams of a reference translation and counts the number of matches [2]. The matches are position-independent and the more the matches, the better is the candidate translation. The unigram scores account for adequacy of translation whereas the longer n-gram scores account for fluency [2]. We will be using 'Show and tell: A neural image caption generator' by Vinyals et al. [1] as our baseline for comparison.

REFERENCES

[1] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555 (2014).

[2] BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu IBM T. J. Watson Research Center Yorktown Heights, NY 10598, USA

[3] Goodfellow, Ian J., Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville and Yoshua Bengio. "Generative Adversarial Nets." NIPS (2014).