CSCE 5215 - MACHINE LEARNING

Project Proposal - Group 1: Image Caption Generator (ICG)

Task Description: The task of this project is to create an image-to-text generator application that combines the techniques of both computer vision and natural language processing to automatically generate textual descriptions for the given images. It involves understanding the elements present in the image and translating it into coherent sentences. The project will leverage the **VGG16 model** which was developed by the University of Oxford. It has 3X3 filters in all convolutional layers and is a 16-layer model. The dataset that we will be using to develop this application is Flicker8k_Dataset available on Kaggle (https://www.kaggle.com/datasets/ming666/flicker8k-dataset).

Expected Challenges: As image captioning is still a rapidly developing field, especially with the recent popularity of AI, we will face several challenges. Firstly, it is important that our data represents a diverse range of data to ensure robust performance, as well as *quality* samples. Without this, our model, as trained as it may become, would be unfit for test data. Another difficult task will be the validation of model outputs, or in other words: how we determine when the model is right and when it is wrong. As this isn't something as simple as something like number detection (where we would have a fixed correct answer), there are multiple ways to describe a scene (a model architecture that can accurately capture both the visual details of the scene and the semantic meaning of the image content), and therefore will either be confirmed by humans or confirmed with a few keywords. The system for this validation may prove to be a great challenge, potentially demanding a considerable investment of time if conducted manually. While additional challenges may arise in the future, our immediate focus should center on addressing these aspects.

Significance of Project: Image caption generator converts a set of pixel data into a human readable form. It is a very significant problem in computer vision that helps the user in scene understanding. It helps in providing additional information about a given image to get a better understanding of the image. Generating a caption to describe the interaction between elements in an image can be very helpful in so many ways. These are a few examples where Image caption generator could be useful, (1) It can help the visually impaired people to listen to the captions and understand them without having to view the image or touching it; (2) It can also help in content recognition and recommendation for similar images; and (3) It can be used to improve the quality of image-based websites and applications.

Related work: "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," which was put forth by Yoshua Bengio, Richard S. Zemel, Kyunghyun Cho,

Aaron Courville, Jimmy Lei Ba, Ryan Kiros, Ruslan Salakhutdinov, and Kelvin Xu in 2016, presents attention-based models for automatically describing image content. Inspired by machine translation and object recognition, this method achieves state-of-the-art results on benchmark datasets such as MS COCO, Flickr8k, and Flickr30k. Enhanced caption quality, attention visualization, and insights into model behavior are among its strengths. Nevertheless, problems are brought about by the attention processes' introduction of training variability and increasing model complexity. To guarantee stability and convergence, managing this complexity necessitates cautious application and training methodologies. "Show, Attend and Tell" is a potential way to significantly improve the interpretation and expression of visual content in natural language, and it marks a substantial development in image captioning overall.

The Neural Image Caption (NIC) model is introduced in "Show and Tell: A Neural Image Caption Generator," a proposal that Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan made in 2015. It addresses the problem of automatically describing images with natural language. NIC provides an end-to-end solution trainable via stochastic gradient descent by fusing recurrent neural networks (RNNs) for language synthesis with deep convolutional neural networks (CNNs) for picture understanding. Higher BLEU scores on several datasets demonstrate its remarkable performance advantages over existing methods due to its inclusion of cutting-edge sub-networks for language and vision models. Even with its advantages, NIC may have drawbacks due to its intricacy and need for large amounts of training data. Its performance could be limited by the quantity and caliber of training datasets, and its architecture could be difficult to deploy and train. Overall, NIC represents a promising advancement in the field of image captioning, though ongoing research is necessary to address its complexities and enhance its robustness across diverse domains and datasets.

Data description: For image captioning we have chosen COCO dataset[3] from the official “cocodataset.org” website, among other datasets such as Flickr30k, Conceptual Captions, NoCaps, VizWiz, ReferItGame, TextCaps. Which has a subset called “COCO Captions” which specifically includes images with five human generated captions for each image, in general the COCO dataset has over 200000 labeled images, 80 object categories, 91 stuff categories and 1.5 million object instances. The data “2017 Train images” is of size 18 GB and has 118287 images, “2017 Val images” is 1GB and has 5000 images, which both have the annotations file “2017 Train/Val annotations” of size 241MB, with train and validation json files of captions, instances and person_keypoints. Considering captions json file, it has the following main keys 'info', 'licenses', 'images', and 'annotations'. And annotations contain a dictionary with 'image_id', 'id', and 'caption'.

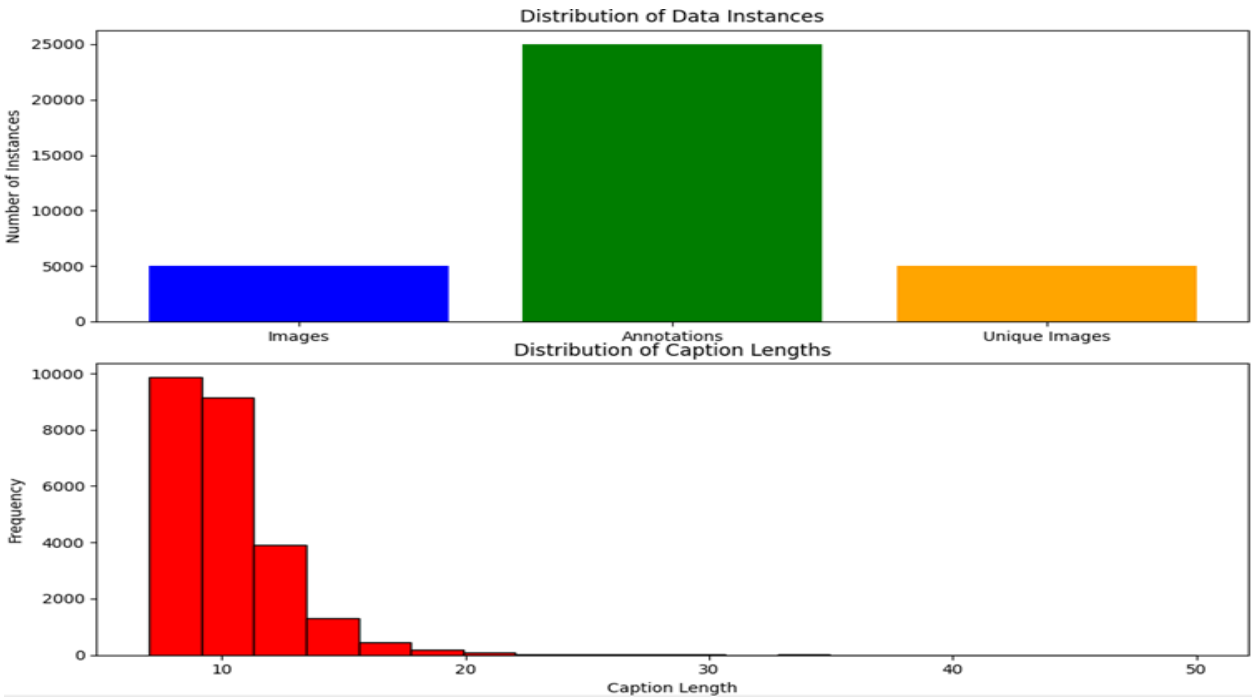


Figure 1: shows us the distribution of data instances and caption lengths.

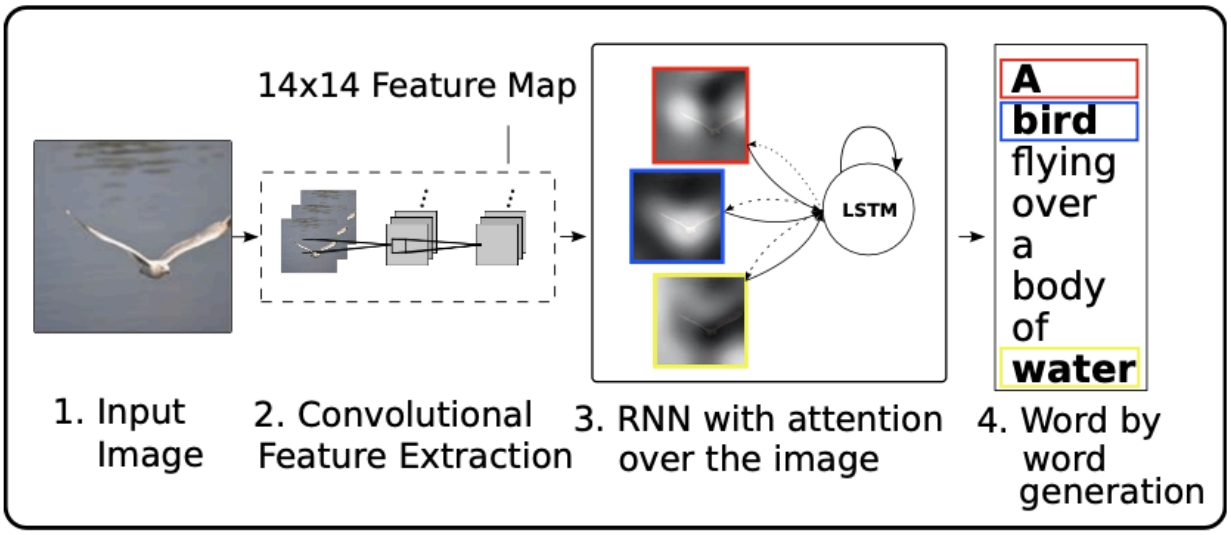
Here, "Frequency" represents the number of captions that fall within a particular range of caption lengths and "number of instances" refers to different aspects of the COCO captions validation dataset such as the total number of images in the dataset, total number of caption annotations in the dataset and the number of unique images with at least one associated caption. The bar plot represents these quantities, indicating the count of each aspect being considered.

[3] Tsung-Yi Lin et al., 2014. Microsoft COCO: Common Objects in Context. CoRR, abs/1405.0312. Available at: http://arxiv.org/abs/1405.0312

Proposed approach:

Our project will involve two concurrent tasks: integrating **Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs).**

CNNs extract features from images efficiently, making them well-suited for image processing tasks. CNNs can be used to encode images and extract pertinent visual information for image captioning. The matching captions are generated word by word by an RNN, such as a Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM), when these features are provided to it.



Limitation: CNN-RNN models might have trouble producing captions that are both contextually correct and detailed in semantic sense, particularly for complicated photos with lots of different items or situations. Additionally, they could provide bland captions that are unable to convey subtle nuances in the picture or lack originality.

Transformer-based Models: Transformer designs have demonstrated impressive performance in a variety of natural language processing tasks, including language production. Examples of these architectures are the well-known BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) models.

Token embeddings that represent the captions can be paired with encoded image attributes in the context of image captioning. After processing this combined data, the transformer model creates captions that make sense and are contextually relevant.

Limitation: Developing big transformer-based models may involve high processing costs and a significant volume of data. Additionally, in contrast to CNN-RNN models, producing captions using transformers may occasionally produce outputs that are excessively verbose or lack diversity since transformers value coherence and fluency over inventiveness.

References:

[1] https://arxiv.org/pdf/1411.4555v2.pdf

[2] https://arxiv.org/pdf/1502.03044v3.pdf