CAPSTONE PROJECT (CDS)

Image captioning using a CNN and a Transformer

Brief problem statement: -Image captioning is the task of generating natural language descriptions for images, typically using deep learning techniques. Despite recent advances in this field, there are still several challenges, some of which are enumerated hereunder: -

- ☑ **Handling complex scenes and objects**:- Need for advanced & sophisticated models for capturing rich and detailed information in images.
- ☑ **Language variability and ambiguity**: - Posing a challenge in generating accurate and informative captions.
- ☑ **Evaluating caption quality**:- Developing better evaluation metrics that can assess the semantic coherence, relevance, and informativeness of captions remains a challenge.
- ☑ **Real-time captioning**: - Developing efficient and lightweight models that can generate captions in real-time is an ongoing challenge.

Background information: -

- ∞ **Domain Information**: - The central idea is to identify key features of an image and create meaningful sentences that describe the image.
- ∞ **Problem Description and analysis:-** Creation of a description of an image in a natural language involves several steps when implemented for machines.

At present, an end-to-end Image Captioning (IC) model has two sections- one that extracts image features and the second that will generate a caption using those features. To extract the image features, a pre-trained Convolutional Neural Network (CNN) is employed. These features are then passed through a Recurrent Neural Network (RNN) leveraging advanced features such as Long Short-Term Memory (LSTM) which helps generate a caption.

Possible Applications: -

- ✓ Recommendations for Editing Applications
- ✓ Assistance for the Visually Impaired
- ✓ Media and Publishing Houses
- ✓ Searching images based on Textual Description
- ✓ Organizing & classifying Images
- ✓ Social Media Posts

Motivation for selection of the project: -

- The main challenge of mimicking the human ability to provide descriptions is to capture how objects relate to each other in the image and to express them in a natural language (like English).
- Using pre-defined templates for generating text descriptions does not provide the sufficient variety required for generating lexically rich text descriptions.

Detailed dataset description and dataset source:- We are planning to use the following datasets:-

a. **FLICKR 8K Dataset** :- https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip:- Link contains over 8,000 images, taken from six different Flickr groups, which are then paired with five different captions available in a separate Text file. All these captions are used to provide accurate descriptions of the possible features, salient entities, and other events. These captions are manually described, and images are carefully chosen to avoid any publically known personality. This dataset is considered as a benchmark for Image Captioning & Processing.

b. **Flickr30K** (having 31K+ Images along with Captions) (**https://www.kaggle.com/datasets/eeshawn/flickr30k?select=flickr30k_images**)

c. **MS COCO Dataset** -For further tuning, we will use the MSCOCO dataset, containing 300k+ images, along with multiple sets of Captions. (**https: https://cocodataset.org/#download**)

Current benchmark: -

☆ **MS COCO dataset:** Combines a Vision Transformer (ViT) and a Contrastive Language-Image Pre-Training (CLIP) model. **The model achieves a CIDEr score of 144.5 on the test set**.

☆ **Flickr30k dataset**: Combines a ViT and a Language Pre-Training (LPT) model. **The model achieves a CIDEr score of 49.8 on the test se**t.

☆ **Conceptual Captions dataset**: Combines a CNN and a Transformer. **The model achieves a CIDEr score of 41.9 on the test set.**

Proposed Plan:

☼ **Approaches**: -

i. Extract the features with CNN from the image using pre-trained models such as InceptionV3.
ii. Preprocessing of captions
iii. Sequential data preparation.
iv. Train the Model with image features & sequential data with RNN leveraging LSTM.

☼ **Stages with defined deliverables**: -

i. 1st Stage → EDA and finalize Data to use and propose top 2 algorithms to use (1st week of April)
ii. 2nd Stage → Model Development & testing (2nd week of April)
iii. 3rd Stage → Refine the model & evaluating the Performance (2nd & 3rd week of April)
iv. 4th Stage → deployment & UI Development (2nd week to 4th week of April)

☼ **Methodology**

i. Packages and tools: - Tensorflow, Keras, AWS/MS Azure, Python
ii. Algorithms under exploration: -

⌖ InIT injecting, PreInjecting, ParInjecting, Feature Vector of Image from CNN to RNN along with a sequence of words from caption.(Ref: Where to put the Image in an Image Caption Generator Marc Tanti (1), Albert Gatt (1), Kenneth P. Camilleri (1) ((1) University of Malta) https://arxiv.org/pdf/1703.09137.pdf)

⌖ Transformer with Self-Attention.

⌖ Pre-processing image with FFT before feeding into RNN. (**Ref: An Attentive Fourier-Augmented Image-Captioning Transformer by Raymond Ian Osolo 1,2ORCID, Zhan Yang 2,3,*ORCID and Jun Long 2,3 https://www.mdpi.com/2076-3417/11/18/8354**)

Predict the object, action, and environment of the caption separately and form a unified Single Image caption.

iii. Metrics:- We will be using BLEU-(1,2,3,4) score or any of standard metrics like CIDEr, METEOR, ROUGE-L.

iv. Deployment plan:- Set up a public website where the user can drag and drop an image to get the captions or a mobile App. Deployed in Cloud (Azure or AWS).

Preliminary Exploratory Data Analysis: -

We have explored.

(a) **Data Distribution on Dataset**: -

(i) Finding an imbalanced class of a particular image class over the other has, that would affect the model accuracy.
(ii) If the data set has fewer caption lengths the model is bound to generate smaller captions, whereas a longer one would be more meaningful.
(iii) Histogram or a Bar chart can be used to visualize the same.

(b) **Word Frequency Analysis**: -

(i) It helps to identify rare and least frequent images and remove them to reduce the vocabulary and improve the quality of captions.
(ii) If there are many instances of a word or phrase the model may learn to associate those words with images and improve the caption quality.
(iii) **A scatter plot or a line chart** can help us to identify the same.

(c) **Image Characteristics: -**

(i) If the image is complex where there are too many objects and detailing, the model would struggle to generate a justifying caption for that complex image.
(ii) The quality of the image i.e., if the image is blurry or poorly lit, it would be difficult to identify and predict.
(iii)Histogram can be used to identify the color density.

Expected outcomes.

1. Trained Model with adequate accuracy (BLEU Score, CIDEr Score, etc.)
2. Generalizability & Explainability
3. Given an Image, the model predicts meaningful captions.
4. Ability to train the model with optimal computation resources.

Project demonstration strategy

Team introduction → Purpose - Brief about the project background, motivation, & importance in real-time scenarios→ Brief introduction about captioning model →Methodology and Architecture → Exploratory Data Analysis and Data Pre-processing → Brief description on CNN, RNN and Loss function→ Model Training→ Model performance via BELU score→ Model deployment→ Live Demo→ Conclusion and Future scope

Proposed timeline of project stage executions: -

Sl. No.	Key Deliverables	Week1	Week2	Week3	Week4
1	Data Understanding, EDA and data pre-processing				
2	Determine important features, build model and assess the model				
3	Evaluate the results, review processes and determine measures for performance				
4	Model deployment and UI development				

Team members' names

Sridhar Sowmiyanarayanan; Soumya Sonakumeru; Shivendu; Karthick; Manjula Tanawade; Hareeshwar Channapragada; Manoranjan Sahu; Sindhu; Parul Kumar Sharma; Sandeep Gubbi

Designated team coordinator's name(s)

Manoranjan Sahu, Sridhar Sowmiyanarayanan, Manjula Tanawade, Shivendu