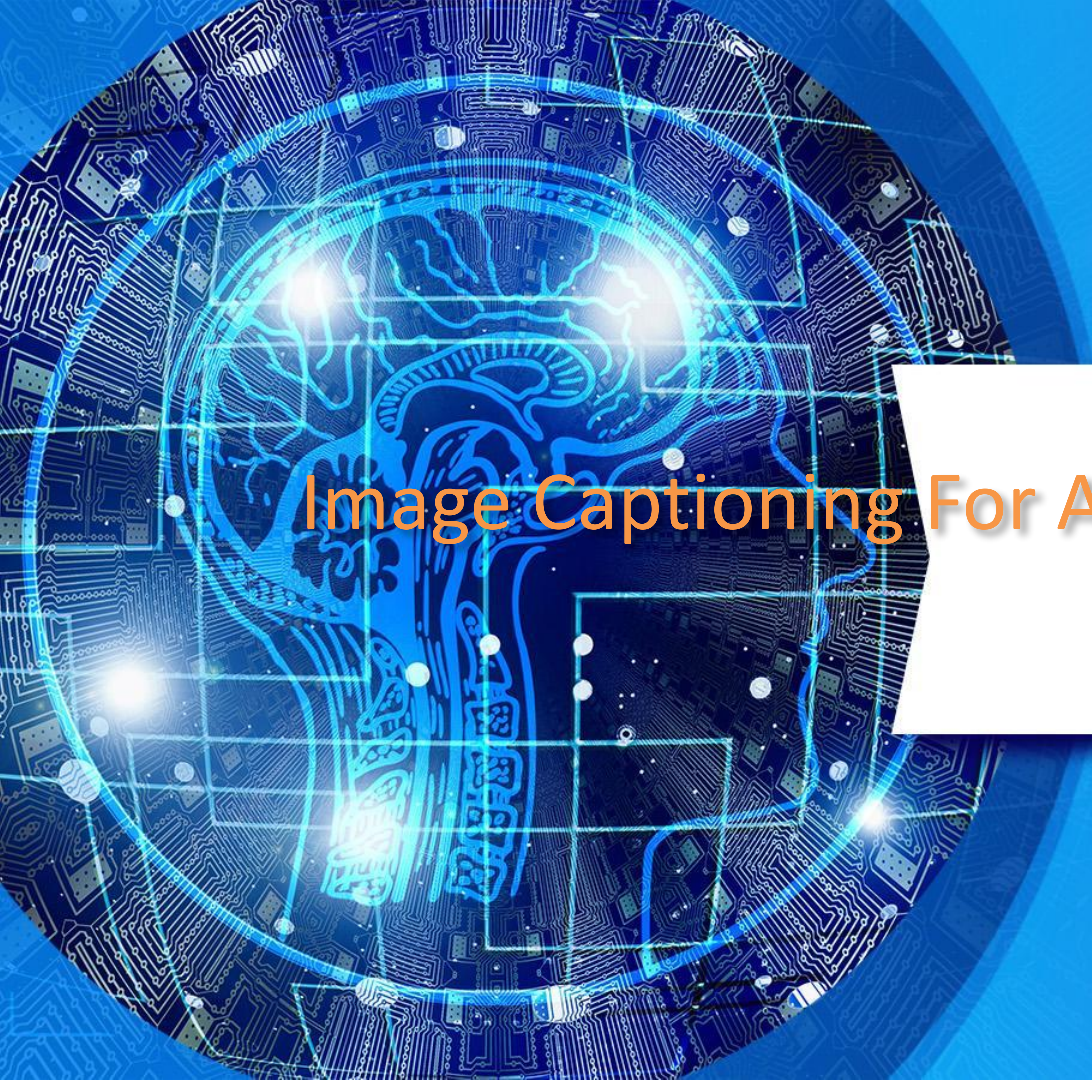


Image Captioning For Academic & Scientific Plots and Figures

Bhaskar Bharat
Ekanki Agarwal
Viraj Kadam
Prerna Mahajan

Problem Statement

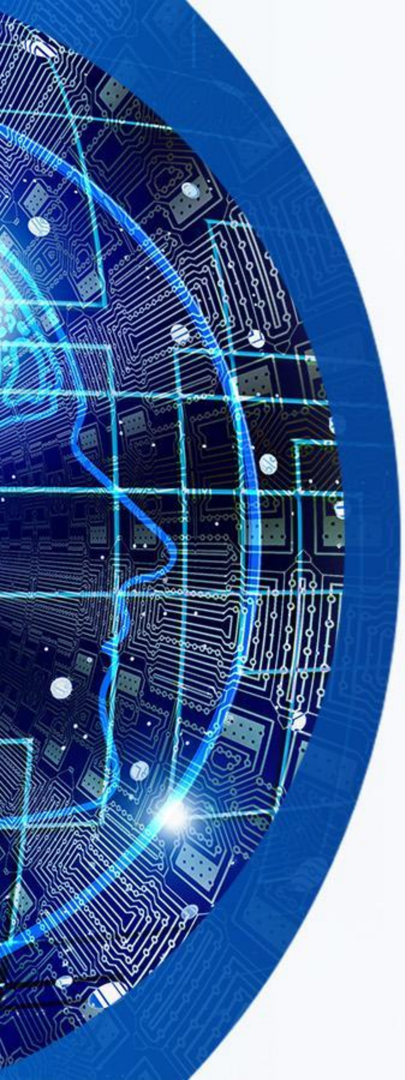
Objective: Automation of image captioning of scientific figures and graphs used in academia.

Motivation: In academia, we use figures and graphs to communicate rich, complex information. The captions of these figures are critical to convey effective messages. Hence, an automated system to generate informative, high-quality captions for these could potentially provide assistance for visually impaired in study of these advanced materials.

Scope: The problem scope involves 2 stages:

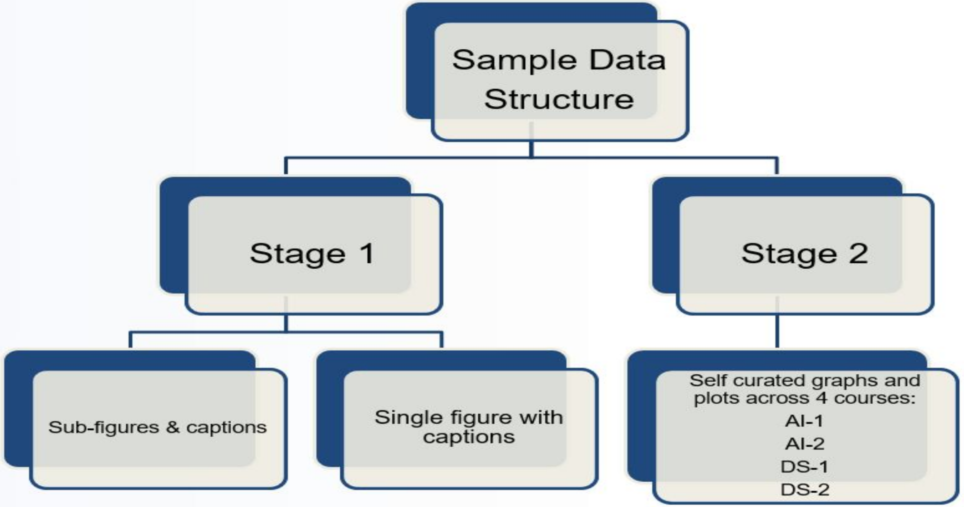
- STAGE 1- Generation of Image captioning system modelled using SCICAP scientific dataset

- STAGE 2- Using transfer learning and fine tuning the above trained model on self curated dataset of figures and graphs based on Univ.AI course



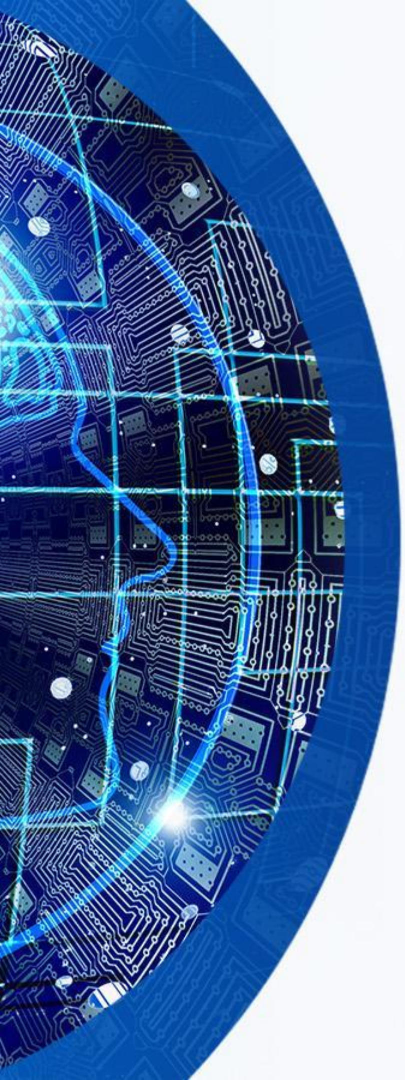
Data: Deep Dive

- **Source:** SCICAP from arXiv, self curated images on Univ.AI course
- **Input Data:** Academic pictures and captions



Sample Data Structure
Stage 1
Stage 2
Sub-figures & captions
Single figure with captions
Self curated graphs and plots across 4 courses:
AI-1
AI-2
DS-1
DS-2

- **Sample size:**
  - Stage 1: >2M images & captions
  - Stage 2: 100 images and captions

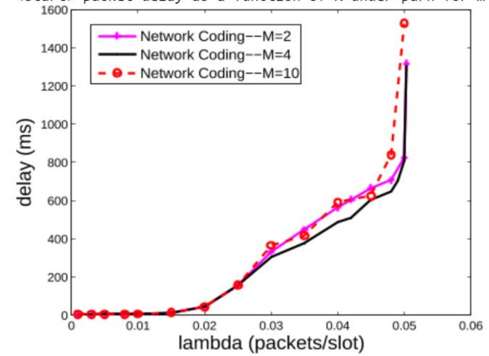


Data: Deep Dive

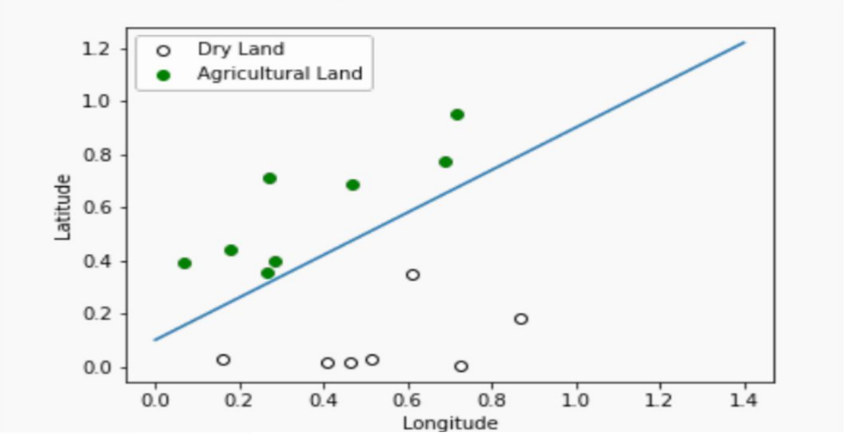
For the purpose of model development, training was carried out only using images with single figure (that is images with no-subfigures) with captions. As an example from the dataset:

SCICAP data

<start> packet delay as a function of λ under parn for m > 0 in the wireless network under 2-hop interference model with network coding <end>



<start> decision boundary classifying agricultural and dry land using latitute and longitude <end>



Univ.AI data

General Approach of Modelling

Modelling Approach: The following modelling approach was adopted in sequence across the two stages

Stage 1: Image features are extracted using pre-trained Inception v3 CNN architectures followed by using sequential modeling techniques like RNNs and attention modules.

Stage 2: Transfer learning using the above created model on the Univ.AI dataset

Performance Metrics: Training data was split into train and validation datasets in the ratio 80% to 20%. Loss and BLEU-4 scores were the performance metrics on test data for final model selection.

General Approach of Modelling

1
- *Preprocess the images using SOTA like Inception V3*
- *Extracting the features obtained from above and feeding into feature extractor*

2
- *Preprocess the text captions*
- *Tokenizing the text captions into integer sequences and creating a word vocabulary. These captions are used by the decoder/sequence processor.*

3
- *Using Encoder+Decoder model architectures.*
- *Two different types of architectures were explored given as follows:*

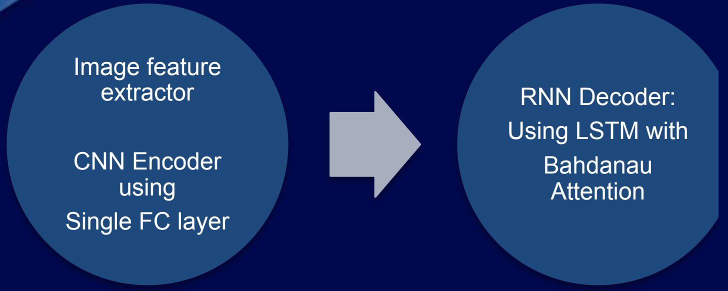


Image feature extractor

CNN Encoder using Single FC layer

RNN Decoder: Using LSTM with Bahdanau Attention

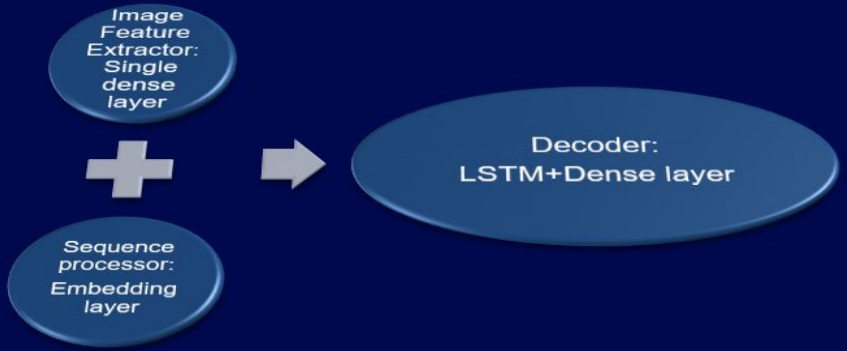
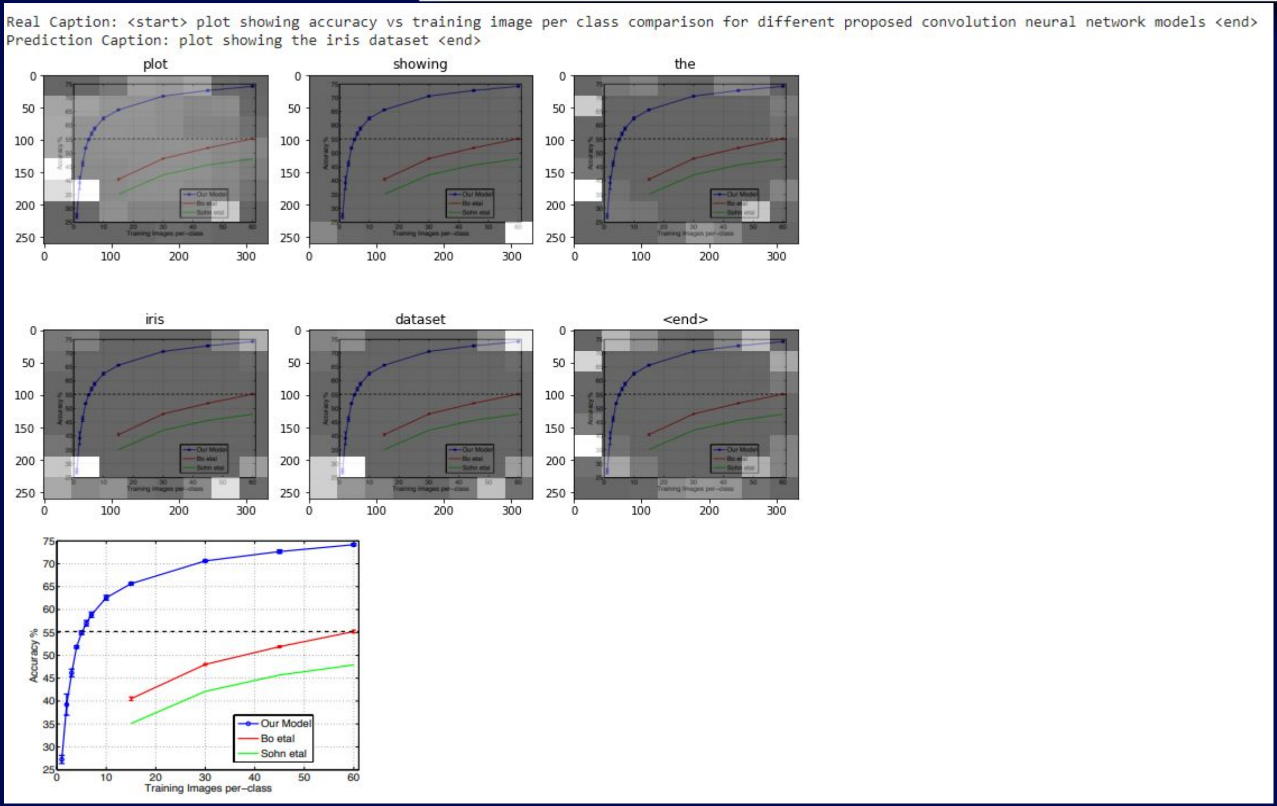


Image Feature Extractor: Single dense layer

Sequence processor: Embedding layer

Decoder: LSTM+Dense layer

Visual Attention Plots



```
Real Caption: <start> plot showing accuracy vs training image per class comparison for different proposed convolution neural network models <end>
Prediction Caption: plot showing the iris dataset <end>
```

Main Results And Inferences:

Encoder-Decoder with Visual Attention
Loss: 2.307683

CNN+LSTM without Attention
Loss: 1.346560

Transfer Learning on Univ.Ai data
Using attention model: Loss: 1.089520
Without attention: Loss: 0.167286

- *The captions generated from Visual Attention model were better than the captions generated by the CNN+LSTM model.*
- *In general, the generated captions are not coherent and are not semantically matching with the original ones.*
- *The plots in SCICAP Dataset do not contain very distinguishing features but have captions very different than the others, therefore the performance of our models has suffered because of this inherent problem with the dataset.*
- *The results achieved on Univ.AI (Test Dataset) are also not promising, this may be due to the size of the dataset used.*
- *Attention plots were used to visualize positions in the image corresponding to segments of words in the generated captions/description.*

Future Work Recommendations For Model Improvement

A. Better data collection strategies and creation of synthetic data for model development purposes for creation of robust models. Currently, the dataset used for modelling consists of captions which explain the objective of the corresponding images rather than describing the images in detail. This leads to poor performance of trained models as the learnt image features do not go well with the given captions. For future, captions explaining the figures with greater detail should be used for modelling.

B. Due to time and computational resource constraints, the current models could not be trained for sufficient number of epochs as well as more complex architectures could not be adopted that could be well explored for future developments. For example, VisualTransformers architectures could be explored.

C. Also, we trained an Embedding layer for the tokens. Instead, pre-trained embeddings such as Word2Vec, ELMo etc. can be explored as well.

D. Due to small size of curated dataset, not much fine tuning could be performed using developed models which is an additional area for the scope of improvement.

Thank You