

# CSCI 6364

## Machine Learning

*Instructor: Sardar H. & Armin M.  
Spring 2023*

*Final Project*

# Image Caption Generator

### **Project Team**

Jahnavi Ramagiri (G46015075)

Mohammed Abdul Irfan (G33655938)

Shubham Jadhav (G30570862)

## Problem Definition

**Problem:** Visually impaired individuals face challenges in accessing and understanding visual content like images.

**Objective:** To develop an image captioning system that can automatically generate accurate and meaningful captions for images to aid visually impaired individuals in understanding the content by recognizing objects, and generate coherent and accurate textual descriptions.

## Dataset

The Flickr 8K dataset is a collection of 8,000 images that are each paired with five different textual descriptions, for a total of 40,000 captions. The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations.

## Data Analysis

Caption Analysis :

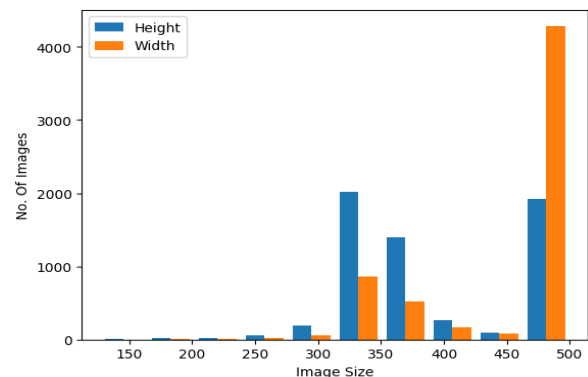
- Unique words vocabulary: 9630.
- Average caption length: 10-12 words.
- Mean caption length: 11.78.
- Standard deviation of caption length: 3.8.
- Maximum caption length: 38.
- The most common words in the captions include "man", "woman", "girl", "water", "black", "white", "boy", and "front". Figure 1 shows the wordcloud of the Flickr8K corpus.

Image Analysis:

- The dataset contains a collection of 8000 images.
- The images in the Flickr 8k dataset cover a wide range of topics and subjects, including landscapes, people, animals, buildings, and more.
- The images in the dataset are of different sizes, with the average image size being around 400x300 pixels.



**Figure 1 : WordCloud for Flickr 8k Dataset**



**Figure 2 : Size Distribution for Flickr 8k Dataset**

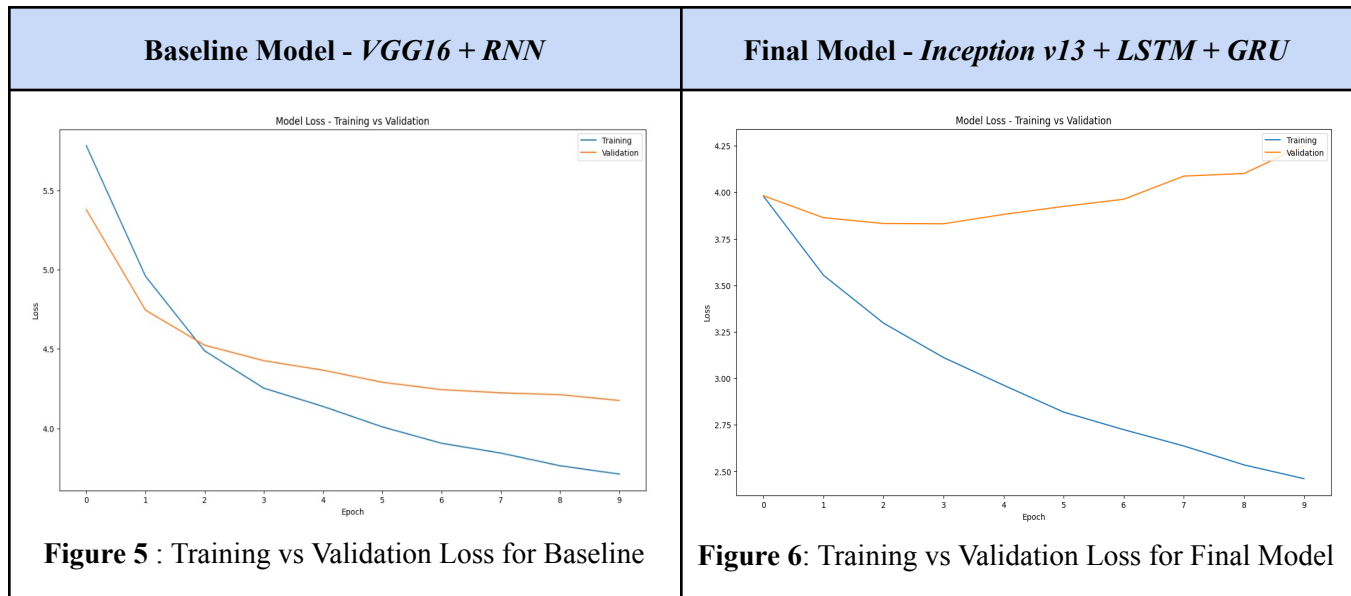
# Experimental Setup

**Model Architectures:** We experimented with five different models for image captioning:

1. VGG16 with RNN (Recurrent Neural Network) - **Baseline**
2. VGG16 with LSTM (Long Short-Term Memory)
3. Inception v13 with LSTM
4. Inception v13 with 2 LSTM layers
5. Inception v13 with LSTM and GRU (Gated Recurrent Unit) - **Final Model**

Baseline Model - VGG16 + RNN	Final Model - Inception v13 + LSTM + GRU
<p><b>Figure 3: Baseline Architecture</b></p>	<p><b>Figure 4: Final Model Architecture</b></p>
<p><b>Image Processing using VGG16</b></p> <ul style="list-style-type: none"> <li>• Input Size: 224</li> <li>• Feature Embedding Size: 4096</li> </ul> <p><b>Text Processing using RNN</b></p> <ul style="list-style-type: none"> <li>• Glove Embedding Dimension: 50d</li> </ul> <p><b>Dense Layers</b></p> <ul style="list-style-type: none"> <li>• No of Layers: 1</li> <li>• Dense Layer Dimension: 32</li> </ul>	<p><b>Image Processing using Inception v3</b></p> <ul style="list-style-type: none"> <li>• Input Size: 299</li> <li>• Feature Embedding Size: 2048</li> </ul> <p><b>Text Processing using RNN</b></p> <ul style="list-style-type: none"> <li>• Glove Embedding Dimension: 200d</li> </ul> <p><b>Dense Layers</b></p> <ul style="list-style-type: none"> <li>• No of Layers: 2</li> <li>• Dense Layer Dimension: 256</li> </ul>

**Model Training:** The models were trained for 10 epochs, with a batch size of 32. We used the Adam optimizer with a learning rate of 0.001 and a categorical cross-entropy loss function to optimize the models.




The two graphs represent comparison of two models training and validation loss curves during the training process. As training progresses for the final model, the training loss decreases at a faster rate than the validation loss, resulting in an increasing gap between the two curves, which may suggest overfitting. The baseline model shows the training loss initially being higher than the validation loss until the 2nd epoch. After the 2nd epoch, the training loss continues to decrease, while the validation loss remains almost constant, indicating a possible learning plateau for the validation dataset.

**Evaluation Metric:** We evaluated the quality of generated captions using the BLEU (*Bilingual Evaluation Understudy*) score. The BLEU score compares generated text to one or more reference texts and provides a score between 0 and 1, with a higher score indicating better alignment with the reference text(s). BLEU scores were calculated for n-grams of length 1 to 4 (*BLEU-1, BLEU-2, BLEU-3, and BLEU-4*).

Image	Captions	Parameters	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG16	RNN	866,129	0.39	0.30	0.12	0.05
VGG16	LSTM	4,691,715	0.40	0.31	0.14	0.06
Inception v3	LSTM	5,084,931	0.41	0.25	0.14	0.07
Inception v3	LSTM, LSTM	5,610,243	0.45	0.27	0.16	0.09
Inception v3	LSTM, GRU	5,545,475	0.48	0.29	0.17	0.09

**Table 1 :** Experiment Results for different model architectures.

Our results show that the Inception v13 architecture with LSTM and GRU layers achieved the best performance in terms of BLEU scores of **0.48**.

 <p><b>Figure 7:</b> Sample image from the Flickr 8K dataset</p>	<p><b>Reference Captions:</b></p> <ol style="list-style-type: none"> <li>1. brown dog running</li> <li>2. brown dog running over grass</li> <li>3. brown dog with its front paws off the ground on grassy surface near red and purple flowers</li> <li>4. dog runs across grassy lawn near some flowers</li> <li>5. yellow dog is playing in grassy area near flowers</li> </ol>
<p><b>Baseline Model - <i>VGG16 + RNN</i></b></p> <p><b>Predicted Caption:</b> two dogs are running in the grass</p> <p><b>BLEU Score:</b> 0.49</p>	<p><b>Final Model - <i>Inception v13 + LSTM + GRU</i></b></p> <p><b>Predicted Caption:</b> brown dog is running through the grass</p> <p><b>BLEU Score:</b> 0.74</p>

## Error Analysis

Error analysis to examine the differences between the models and their impact on the generated captions:

1. Inception's efficient handling of images by capturing both local and global features, whereas VGG focuses on depth with small kernels.
2. The combination of LSTM and GRU layers in capturing complex relationships between image features and generated captions, compared to RNNs, which suffer from the vanishing gradient problem.
3. The higher complexity of the Inception-LSTM-GRU model, allowing it to capture intricate patterns, while being mindful of potential overfitting and the need for regularization.

## Future Scope

1. Enhanced Accuracy: Continued advancements in deep learning techniques, such as transformer-based models like GPT-3, will likely lead to increased accuracy and descriptiveness in generated captions
2. Multimodal Approaches: Exploring multimodal approaches that combine image features with other modalities such as text, audio, or user interactions can lead to more comprehensive and context-aware captions.
3. Attention Mechanisms: Investigating advanced attention mechanisms, such as self-attention or transformer-based attention, can enhance the model's ability to focus on relevant image regions and generate more precise and detailed captions.

## Conclusion

In conclusion, the Inception v3 architecture-based LSTM and GRU layers-based image captioning system displayed superior performance, achieving higher BLEU scores in comparison to other models. Improved accuracy was made possible by capturing complicated correlations between image features and captions. Future improvements in Inception v3's fine-tuning and research into multimodal methods have the potential to substantially improve the precision and descriptiveness of picture captioning systems.