# Image Captioning with Visual Attention

CS626- Endterm Project discussion

Yash Garg, 200110039
Muskan Bhutra, 200040085
Naman Agarwal, 19d180017

# Problem Statement

**Develop an advanced image captioning model using a modified version of the "Show, Attend and Tell" approach, by modifying the decoder by a 2-layer Transformer-decoder.**

# Related works

- **"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" paper.**

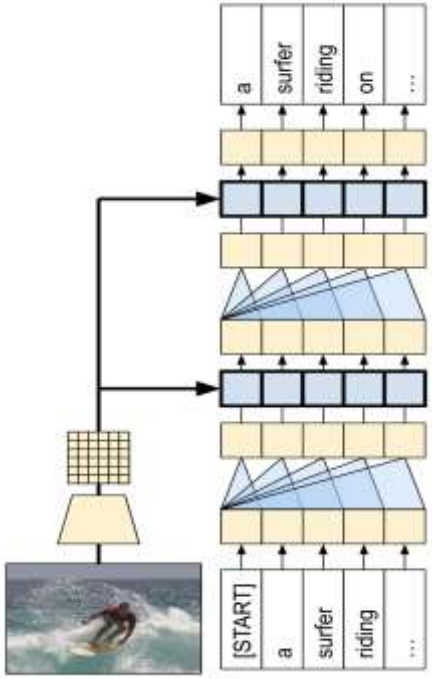
# Dataset (s)

- Utilized the **Flickr8k** dataset for training and evaluation.
- Contains **8,000** images, each paired with **5** reference captions for diversity.
- A widely used dataset in the field of image captioning and language understanding.
- Images cover a variety of scenes, objects, and activities, providing a diverse training set for the model.

# Architecture Plan

- **Original "Show, Attend and Tell" Model:**
  - Utilized a single-layer decoder with attention mechanisms.
  - Used LSTM (Long Short-Term Memory) units to process input sequences and generate captions.
  - The attention mechanism allowed the model to focus on specific image regions while generating each word of the caption.

- **Modified Model with 2-layer Transformer-decoder:**
- Each layer includes:
  - Causal Self-Attention Layer: Enables output locations to attend to the output generated so far, preventing future information leakage.
  - Cross Attention Layer: Allows output locations to attend to the input image, capturing visual information relevant to generating accurate captions.
  - Feed Forward Network Layer: Processes each output location independently, further refining the generated captions.




The above is a 2-layer Transformer model

- Replaced the single-layer decoder with a 2-layer Transformer-decoder:Capitalized on the transformer's ability to capture long-range dependencies and learn contextual relationships.
- Improved the model's ability to handle complex language patterns and image-caption relationships.

# Metrics used and Result

- **BLEU-1 Score:** Evaluated the quality of generated captions by comparing them to reference captions using the BLEU-1 metric.

**Results:**

- Achieved a BLEU-1 score of 0.4 on the modified 2-layer Transformer-decoder model.
- This indicates that the model captures a significant portion of the unigram overlaps between generated and reference captions, demonstrating its effectiveness in generating relevant and accurate image captions.

# Conclusion

- Successfully extended the "Show, Attend and Tell" model using a 2-layer Transformer-decoder architecture and evaluated it on the challenging Flickr8k dataset.
- Demonstrated the improved caption generation capability through the achievement of a BLEU-1 score of 0.4
- The enhanced model holds promise for various applications in the field of image captioning and natural language processing.