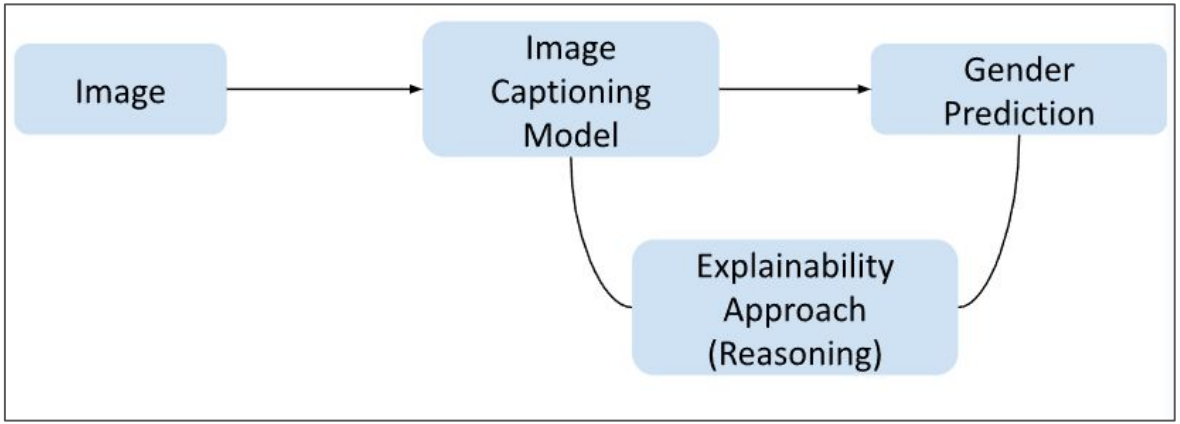
Explainability Approach to understand Gender Bias in Image Captioning

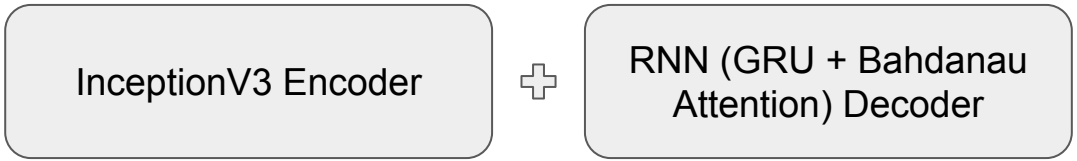
Methodology



Baseline approach : Mapping attention plots using attention weights

Baseline Approach

- The model architecture : [Show, Attend and Tell: Neural Image Caption Generation with Visual Attention]



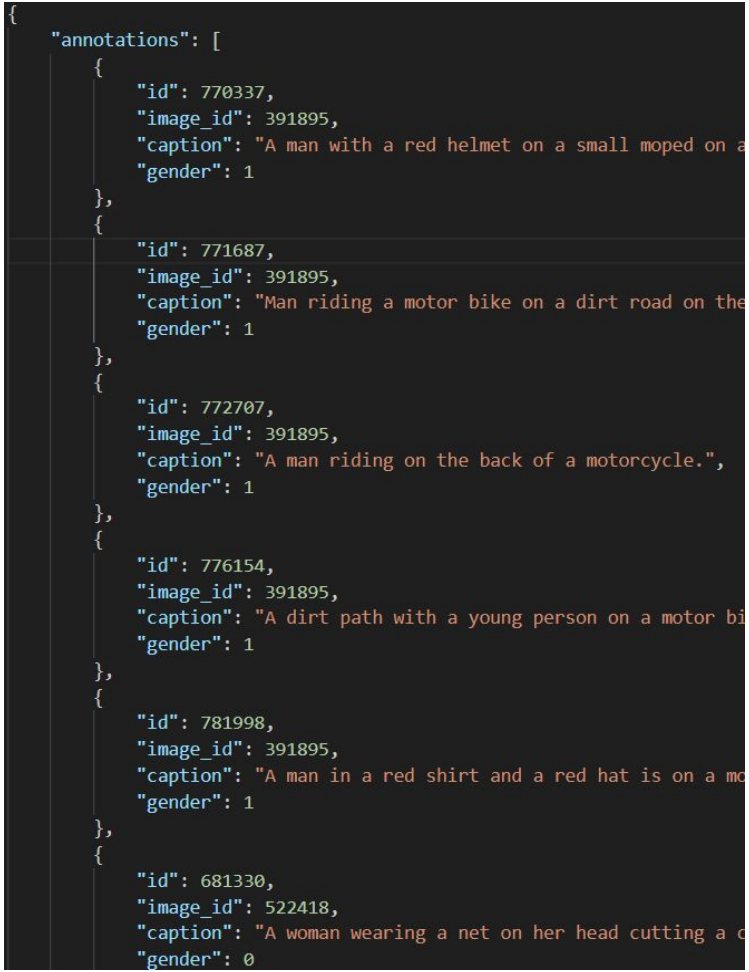
- Training Optimizer : Adam
- Loss: Sparse Categorical Cross entropy
- Dataset : COCOGBv1 (Gender Dataset) [Mitigating Gender Bias In Captioning System]

Baseline Approach

Dataset Format:

Gender Label : "woman," "man," "woman & man" (if woman and man are included in a single picture) and "discard" (no human appears in the image or gender is indistinguishable)

- For training, considered only “woman”, “man”, “woman & man”



Baseline Approach

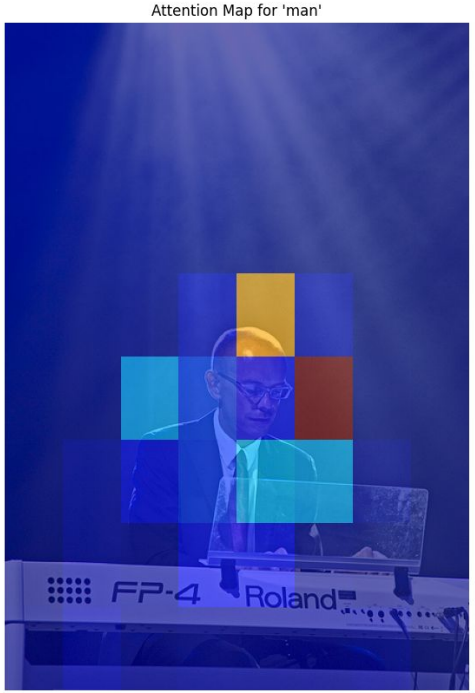
1. Preprocess Images accordingly to match Inception v3 model input requirements.
    a. Resize images to 299x299
    b. Normalize pixels between -1 to 1
2. Extract Features using Inceptionv3 model & cache them.
3. Tokenize the captions, define vocabulary size, create word-index mappings, pad all sequences to be the same length as the longest one.
4. Create train, validation & test datasets.
5. Train the model on train dataset, validate using validation dataset while training.
6. Create attention plot using the attention weights from the decoder module for Gender words generation in caption : “man” , “woman”

Baseline Approach

Attention plot Generation for Gender word generation in captions:

True Caption: a pianist in a suit and glasses playing a keyboard

Predicted Caption: a man in black and a speech

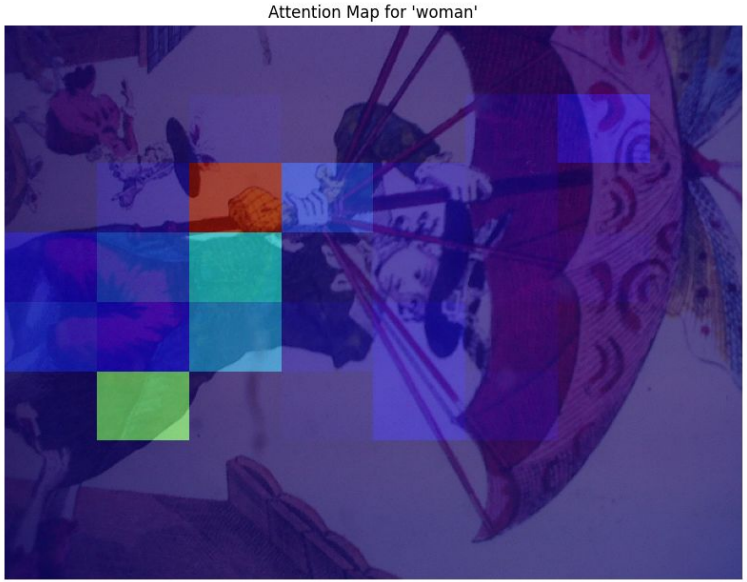


Baseline Approach

Attention plot Generation for Gender word generation in captions:

True Caption: a drawing of a child holding an umbrella

Predicted Caption: a woman holding an open umbrella

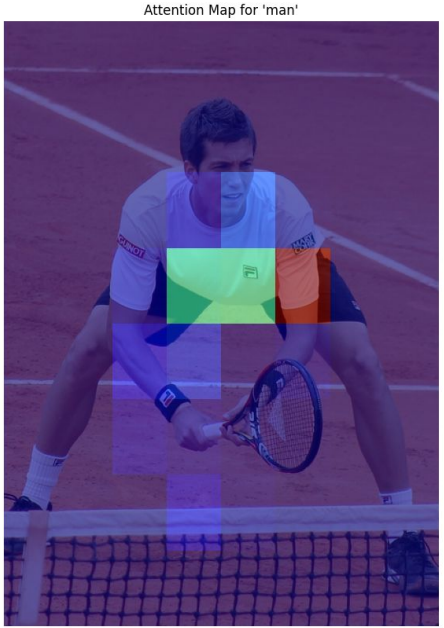


Baseline Approach

Attention plot Generation for Gender word generation in captions:

True Caption: a man in a white shirt and black pants holds a tennis racket

Predicted Caption: a man is playing a game



Baseline Approach

Attention plot Generation for Gender word generation in captions:

True Caption: a woman holding food in her hand while sitting at a table

Predicted Caption: a woman standing next to a dinner in front of a lap