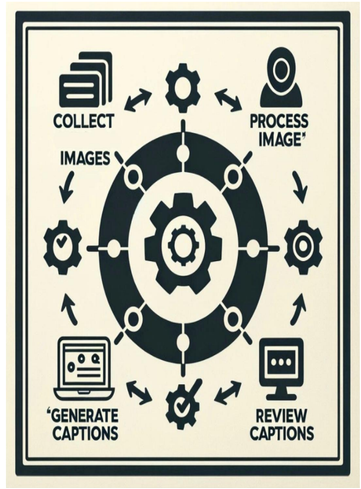
Image Captioning Using CNN And LSTM

Prakyath Davanam, Rithvik Segu, Satwik M Belaldavar

OVERVIEW

- The process of captioning an image involves using computer vision and deep learning to identify the image, analyse its context, and annotate it with relevant captions.
- It involves using datasets provided during model training to classify an image with English keywords. The extraction of features from images is handled using VGG16.
- The LSTM model will receive these extracted features and use them to create the image caption. Two parts of a neural network system are used in this technique. The CNN part focuses on understanding visual input, while the LSTM part is responsible for generating human-like textual descriptions.
- We have used VGG16 model which is pre-trained on ImageNet.
- We are extracting the features from fc2 layer of the VGG16 model.



DATA COLLECTION

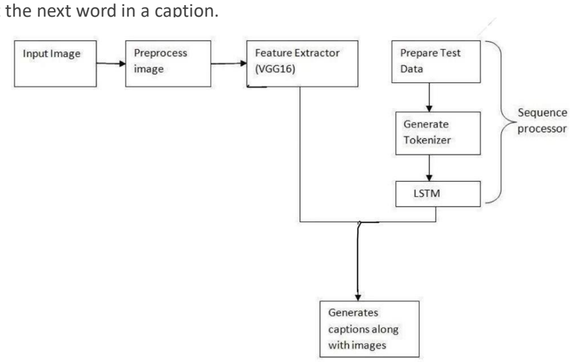
- The project employs the picture captioning research standard Flicker8k dataset. About 8,000 photographs were chosen for their richness and diversity in settings, events, and objects.
- Image Data: Flicker8k pictures range from everyday life to special events. Diversity is needed for the model to label a variety of photos accurately.
- The dataset has five captions for each image. Human annotators write these captions to describe the same image from different angles. Multiple image descriptions help train a strong model that understands and creates natural language phrases.
- Richness of Language: Flicker8k captions use natural and diverse language because there is no predetermined vocabulary or style. This dataset requires the model to grasp and produce many verbal expressions.
- Quality of annotations (picture descriptions) is crucial. Good captions define the image's objects, context, and interactions, helping the model grasp the picture.

DATA PREPERATION

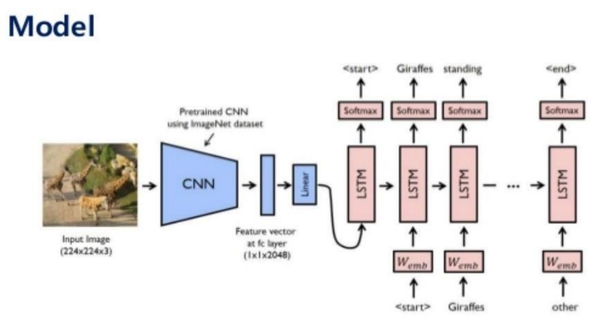
- Pre-trained Convolutional Neural Networks (CNNs) such as VGG-16 are used to extract meaningful characteristics from dataset images. This recognizes image edges, textures, and forms.
- Each image's important visual information is compressed into a vector of features.
- Image captions are cleaned. This usually entails lowercase text, punctuation removal, and spelling correction. To simplify language and remove extraneous variation in the text.
- Text is tokenized into words or tokens. Turning text into a model-friendly format requires this step.
- Create a dictionary of all dataset unique words. To aid neural network processing, this vocabulary assigns each word a unique integer (word indexing).Finally, integer sequences are created from descriptions. Caption words are substituted by vocabulary indexes. The model learns word-image relationships using these sequences.

DATA MODELLING

- Combining CNN and LSTM: The Long Short-Term Memory (LSTM) network and the Convolutional Neural Network (CNN) are combined in this model design. To extract characteristics from images, the pre-trained model VGG-16 is utilized. These characteristics successfully convey the images' visual information.
- Text Generation with LSTM: An LSTM network is fed with the collected features after that. One kind of Recurrent Neural Network (RNN) that works well for sequence creation tasks is LSTM. It creates words for a caption by processing the image's elements one at a time.
- Using the visual attributes and the last generated word sequence, the LSTM learns to predict the next word in a caption.



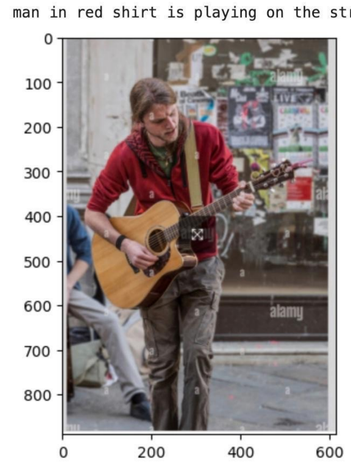
- Word Embeddings: As a component of the LSTM network, word embeddings transform input words into fixed-size, dense vectors. Semantic meaning and word relationships are captured in this representation.
- Training the Model: The model is exposed to pictures and the captions that go with them. It picks up on connecting certain visual patterns in the photos with pertinent language in the captions.
- Optimization and Backpropagation: Algorithms for optimization, such as the Adam optimizer, are used to train the model. In order to help the model make better predictions, the loss function—typically categorical cross-entropy—measures the discrepancy between the actual and predicted captions.



INTERPRETATION AND RESULTS

The Interpretation phase entails examining and comprehending the model's outputs, with a primary focus on the model's ability to produce precise and pertinent captions for images:

- Caption Generation: New, unseen photos are captioned using the learned model. This entails passing the CNN-extracted picture features into the LSTM network, which produces a string of words for the caption.
- Quality Assessment using BLEU ratings: BLEU (Bilingual Evaluation Understudy) ratings are used to quantitatively assess the model's performance. These scores evaluate the quality of the machine-generated captions based on standards such as grammatical correctness and word choice precision, by comparing them to a set of reference captions.
- Practical Implications: Recognizing the model's strengths and weaknesses is made easier by knowing how well it performs. For example, if the model has trouble with specific kinds of images or situations, this knowledge might direct future enhancements.
- The image captioning model demonstrates moderate accuracy, with a BLEU-1 score of 55.43%, indicating a decent ability to match individual words in captions. However, its performance declines with longer phrases, as evidenced by lower BLEU-2, BLEU-3, and BLEU-4 scores of 29.37%, 16.7%, and 9.2% respectively. These results suggest proficiency in basic captioning but highlight the need for improvement in constructing more complex and coherent phrases.



REAL-WORLD APPLICATION

- Content Moderation: Automated systems can use image captioning to understand and flag inappropriate content on websites and digital platforms.
- Healthcare: In medical imaging, captioning can help pre-diagnose images, providing quick, preliminary assessments for further review by a professional.

REFERENCES

1. Cross Validated (stackexchange.com)
2. BUILDING AN IMAGE CAPTIONING SYSTEM USING CNNs AND LSTMs https://www.iijmets.com/uploadedfiles/papeí/volume2/issue_6_june_2020/1703/1628083053.pdf
3. Reference Based LSTM for Image Captioning https://ojs.aaai.oíg/index.php/AAAI/aíticle/view/11198