International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Review Paper on Enhanced Image Captioning with Deep Learning: Encoder-Decoder and Attention Mechanism

Vikash Kumar Singh[1], Ankita Gandhi[2], Brijesh Vala[3]

[1]Department of Computer Engineering
Parul University
Vadodara, Gujarat
vik439@gmail.com

[2]Department of Computer Engineering
Parul University
Vadodara, Gujarat
ankita.gandhi@paruluniversity.ac.in

[3]Department of Computer Engineering
Parul University
Vadodara, Gujarat
brijesh.vala@paruluniversity.ac.in

Abstract— Image captioning involves the generation of textual descriptions that describe the content within an image. This process finds extensive utility in diverse applications, including the analysis of large, unlabelled image datasets, uncovering concealed patterns to facilitate machine learning applications, guiding self-driving vehicles, and developing software solutions to aid visually impaired individuals. The implementation of image captioning relies heavily on deep learning models, a technological frontier that has simplified the task of generating captions for images. This paper focuses on the utilisation of encoder-decoder model with attention mechanism for image captioning. In classic image captioning model, the words usually describe only a part of the image, however with attention mechanism special attention is given to the low level and high level features of the image. Object detection using attention mechanism has shown to have increased the CIDEr score by 15%. With the use of stable dataset of MSCOCO through keras datasets, it is possible to score more on caption generation and accurate description of image.

Keywords- Image Captioning, encoder-decoder, attention mechanism, TensorFlow datasets, MSCOCO.

I. INTRODUCTION

Image captioning, the task of generating human-like textual descriptions for images, is a rapidly evolving field of research at the intersection of computer vision and natural language processing [5]. It seeks to harness the strengths of these two areas to create models capable of understanding visual content and conveying this understanding in natural language.

In fig. 1 we show a typical example of how an automatic image captioning system can generate captions for images supplied to the model. The system analyses the high level features present in the image and generates appropriate captions relevant to the context.

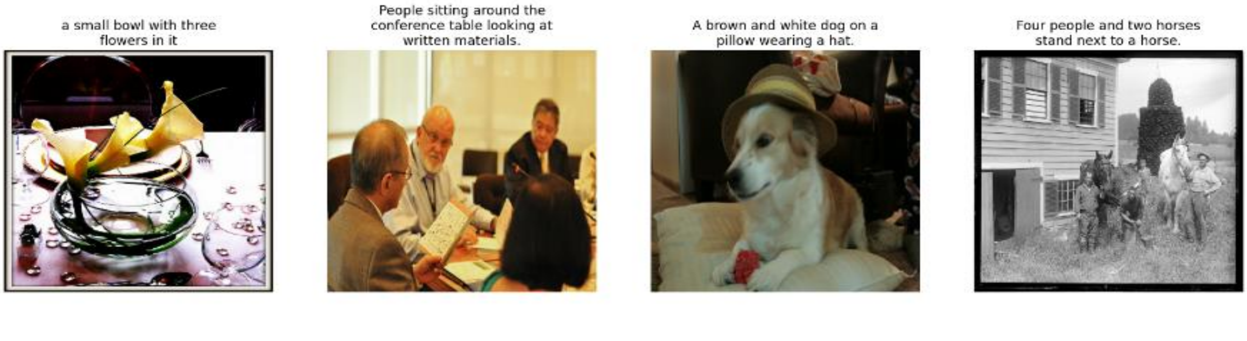
a small bowl with three flowers in it
People sitting around the conference table looking at written materials.
A brown and white dog on a pillow wearing a hat.
Four people and two horses stand next to a horse.

Figure 1. An example of caption generation for images performed by automatic captioning systems.

The significance of image captioning lies in its wide range of applications [4]. For instance, it can be used in social media platforms to automatically generate captions for uploaded images, facilitating content discovery and accessibility. In the field of healthcare, it can assist doctors by providing descriptions for medical images. For visually impaired individuals, image captioning systems can describe the content of images, thereby aiding in their understanding of the visual world. Automated image captioning can also enhance the performance of search engines by providing them with a textual context for images, improving the accuracy of image-related search queries.

The history of automated image captioning is marked by several important phases. The initial efforts [3] were simplistic, relying heavily on handcrafted features and template-based generation methods. However, these methods proved inadequate in dealing with the complexity and variability of real-world images.

The advent of deep learning has revolutionized the field of image captioning [6], [18], [19]. Convolutional Neural Networks (CNNs) have become the backbone for extracting visual features, while Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and more recently Transformer [12] models have been utilized for generating the textual descriptions.

Despite these advancements, generating accurate and contextually appropriate image captions is still a challenging problem. The complexity of this task arises from the need to accurately recognize the objects in an image, understand their attributes and interactions, and generate a linguistically correct and meaningful sentence that encapsulates the overall context of the image [17].

This paper aims to address these challenges by using proven and effective architecture such as encoder-decoder, attention-mechanism and transformers for providing a contextually appropriate and accurate solution for automated image captioning. Section 2 covers the related works in the field, while section 3 covers the methodology used to design the system. Section 4 gives a layout of the experiment and results.

II. RELATED WORKS

In one of an early influential paper [1], Alexei A. Efros et al. presents an early approach to image captioning using template-based methods. The author explored the idea of filling in sentence templates with relevant information extracted from images. A probabilistic approach for interpreting images and generating captions have been emphasised in the paper [2] where a graphical model combined visual and contextual information to generate descriptions. Semantic image annotation is a process of labelling or tagging specific objects, regions or features within an image with descriptive and meaningful labels. The paper [3] focuses on semantic image annotation and also touches upon the generation of textual descriptions for images. It presents a probabilistic image model that incorporates both low-level visual features and high-level semantic concepts for generating captions.

One of the earliest papers on deep learning is [6] which has been the most influential and pioneering in the field. The authors introduced a model that aligns fragments of sentences with the region of the image that they are describing. The model used a combination of Convolutional Neural Networks (CNN) for image processing and Recurrent Neural Networks (RNN) for language modelling. Another pioneering work [7] introduces the use of attention mechanisms in image captioning. The authors proposed an attention-based model that can automatically learn to focus on salient parts of the image while generating each word in the caption.

A very important work in the field of image captioning has been in [8] by researchers at Google, who proposed a model that used an encoder-decoder framework for image captioning. The encoder is a deep convolutional network (Inception-V1), and the decoder is a long short-term memory (LSTM) network. Another new approach was proposed in [9] that combined bottom-up (object-level) and top-down (attention-based) mechanisms for image captioning. This approach allowed the model to focus on salient objects in the image and improved the quality of the generated captions.

Neural baby talk model was introduced in 2018 by the paper [11] which aims to generate captions that capture the salient objects or regions in an image while exhibiting a simplified and repetitive language style similar to how young children describe their surroundings. This work proposed a novel approach to image captioning, generating captions that are more like how parents talk to their babies. The model uses object detection to identify key objects in the image and then generates a caption based on these objects.

The 2017 influential paper [20] introduces the Transformer model, which revolutionized various natural language processing tasks. It discusses the potential application of the Transformer model to image captioning and highlights the advantages of self-attention mechanisms in capturing global dependencies and improving caption generation. Semantic attention has always been the popular model in designing an effective image captioning system. In the paper [10], the limitation of traditional attention mechanisms in image captioning is addressed at large. The paper proposes a semantic attention model that incorporates semantic information into the attention mechanism to generate captions that focus on relevant image regions and capture the underlying semantics. The 2014 paper by Kiros et al. [21] explores the integration of visual and textual information in image captioning. It introduces a multimodal neural language model that learns joint representations of images and their corresponding captions, bridging the gap between visual and textual modalities.

In some of the latest works in image captioning, ImageBERT, a cross-modal pre-training model for image captioning was introduced in a pioneering work [13]. It leverages a large-scale weakly supervised image-text dataset to learn joint representations of images and captions, enabling better caption generation. The paper in [14] presents an approach that aligns the cross-modal space between images and captions for image captioning.

III. THE METHODOLOGY

In this paper, we use an encoder - decoder model [15] to generate automatic image captioning. We pass images to encoder at first, and it extracts information from the images, and creates some feature vectors. And then the vectors are passed to the decoder, which actually builds captions by generating words one by one.

The basic working of an encoder and decoder model can be explained through the diagram given in Fig. 2. This model was proposed in "Show and Tell - A Neural Image Caption Generator (2015).

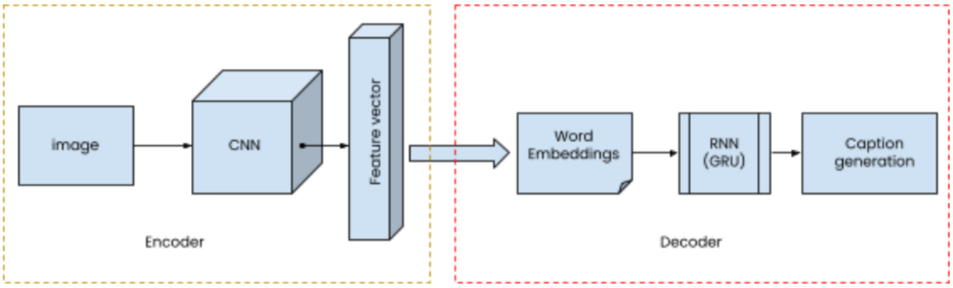

Figure 2. Block Diagram of an Encoder-Decoder model

The encoder is used for extracting meaningful information from the input image [16]. It takes the raw pixel data of the image as input and transforms it into a fixed-size feature vector, which represents the high-level content and context of the image. Convolutional Neural Networks (CNNs) are used as the encoder in image captioning models. CNNs are excellent at extracting features from images due to their ability to capture spatial hierarchies of features. As the image progresses through the encoder layers, it becomes increasingly abstract and condensed, retaining only the most relevant information.

The decoder takes the feature vector produced by the encoder and generates a sequence of words that form a coherent and descriptive caption for the image. Recurrent Neural Networks (RNNs) and Gated Recurrent Unit (GRU) are typically used as the decoder in image captioning models. These networks are capable of handling sequential data and are suitable for generating sequences of words. The decoder begins with an initial input "<start>" and produces one word at a time. It then uses the generated word as input for the next step, continuing until it generates an "<end>" token or reaches a predefined maximum caption length.

The training process of an encoder-decoder model for image captioning involves the following steps:

A. Data Preparation

A dataset consisting of images paired with their corresponding human-generated captions is collected. Each image is preprocessed, and its features are extracted using the encoder (CNN), producing a fixed-size vector. The captions are tokenized into words or sub-words, creating a vocabulary of possible words.

B. Training

During training, the model learns to generate captions by optimizing a loss function that measures the dissimilarity between the generated caption and the ground truth (human-generated) caption. The encoder-decoder model is trained end-to-end, with the encoder's parameters fixed while the decoder's parameters are updated.

C. Inference

To generate captions for new, unseen images, the trained model uses the encoder to extract features from the input image. The decoder then generates a caption word by word, conditioning each word on the previous ones until it produces the "<end>" token or reaches a predefined maximum length.

The reason why encoder-decoder model scores above other models are:

- Compared to template-based approaches, encoder-decoder models are not limited to fixed templates for captions. They generate captions based on the content of the image, allowing for more variability and context-awareness.
- Encoder-decoder models integrate object detection within the model itself, enabling them to describe scenes holistically, including context and relationships between objects.

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

- Attention mechanisms in this project allow to focus on different parts of the image while generating each word, improving the contextual relevance of captions.
- Transformers, which are a type of neural network architecture, can be used as both the encoder and decoder, enabling efficient parallel processing and capturing long-range dependencies in images and text.
- By using this model, we can fuse information from multiple modalities, such as text and images, to generate captions that incorporate diverse information sources.
- Also, for handling complex scenes with multiple objects and relationships, this model can describe the interactions and context effectively.

IV. EXPERIMENTS

In this paper, we use an encoder - decoder model [15] to generate automatic image captioning. We pass images to encoder at first, and it extracts information from the images, and creates some feature vectors. And then the vectors are passed to the decoder, which actually builds captions by generating words one by one.

A. Dataset

In this model, we have used the MSCOCO dataset, however we have not used it directly. Instead, we have utilized the TensorFlow datasets capability to the COCO captions dataset. This version of COCO dataset consists of images, bounding boxes, labels and captions. The dataset is pre-split into training and test defined by Karpathy and Li (2015) and takes care of data quality issues with the original dataset.

In this model we have used a pretrained InceptionResNetV2 from tf.keras.applications as a feature extractor, so we need to define some of the constants required in the InceptionResNetV2 model.

The InceptionResNetV2 takes (299, 299, 3) images input and returns features in (8, 8, 1536) shape as output. tf.keras.applications is a pretrained model repository in TensorFlow Hub. While TensorFlow Hub hosts models for different modalities including images, texts, audios, etc, the tf.keras.application only hosts popular and stable models for images. tf.keras.applications is more flexible as it contains model metadata which helps us to access and control the model behaviour, while most of the TensorFlow Hub based models contain only compiled saved Models. Due to this, we can get output not only from the final layer of the model, but also from intermediate layers by accessing layer metadata.

B. The CNN Encoder

The CNN encoder extracts feature from an image through a pre-trained model and passes them to a fully connected layer. The features extracted from convolutional layers of InceptionResNetV2 gives a vector (32, 8, 8, 1536). The vectors are reshaped to (32, 64, 1536) and then reduced to a length of ATTENTION_DIM = 32 (here) The reshape is primarily done to make it ready for the attention layer. The attention layer is used in the decoder stage and it utilizes the input image to predict the next word.

C. The Caption Decoder

The caption decoder is the other part of the model which uses an attention mechanism that focuses on different parts of the input image. The purpose of the attention mechanism to selectively focus on parts of the input sequence. The attention takes a sequence of vectors as input for each example and returns an "attention" vector for each example.

Decoder is used to generate predictions for the next output token (Fig. 3). The steps used in decoder are:

- The decoder at first receives the current word tokens in a batch.
- The word tokens are embedded to match the ATTENTION_DIM (Attention dimension)
- The GRU layer keeps track of the word embeddings, and returns GRU outputs and states.
- The attention layer attends over the encoder's output feature by using GRU outputs as a query.
- The attention output and GRU outputs are added which is also called the skip connection, and normalized in the normalized layer.
- It then provides probabilistic predictions for the next token based on the GRU output.

D. Training Model

As we have seen, encoders are used to extract the features of the image, while the decoder is used for generation of captions. So now since we know how the encoders and decoders are defined, these two models are combined into an image model for training. The training model has two inputs - the image input and the text input and an output which is the decoder output.

E. Loss Function

The loss function used in this model is simple cross-entropy. Here we extract the length of the sentence and compute the average of the loss only over the valid sentence part.

F. Tools Used

The model is a software application that comprises digital data for input and output. The essential softwares and tools used for in this project are:

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

- Python 3.11.4
- TensorFlow Keras
- Numpy
- Jupyter Notebook
- GPU NVIDIA Quadro 4000.

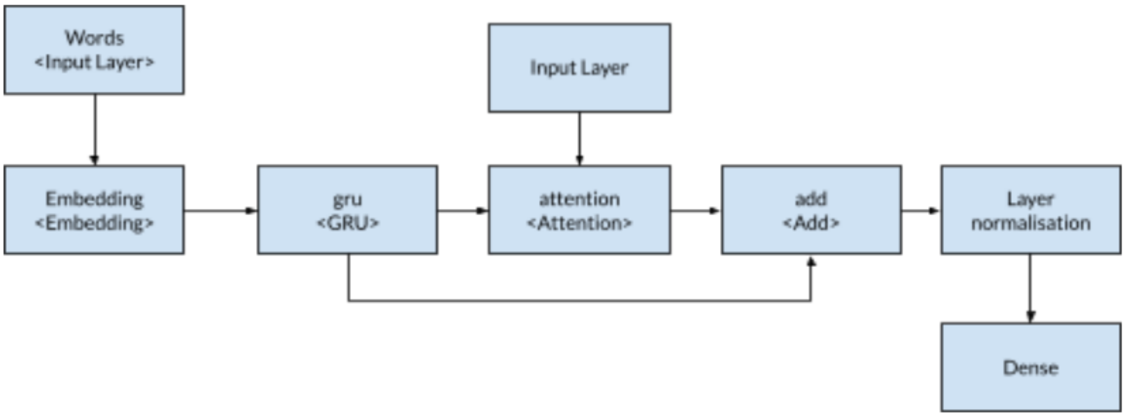

Figure 3. Working of a Decoder Model with Attention Mechanism

V. CONCLUSION AND FUTURE WORK

In this paper we have proposed an image captioning system using a deep learning model. The dataset used is MSCOCO from the TensorFlow dataset. It was basically used to give a more stable image and caption correlation. InceptionResNetV2 has been used as a feature extractor. The model works on the encoder - decoder principle which has proved to yield quality results. The decoder uses a combination of GRU and attention mechanism to accurately generate the captions. This architecture is used for extracting the image features and these image features are given as input to Long Short Term Memory units and captions are generated with the help of vocabulary generated during the training process.

This model has higher accuracy compared to CNN-RNN and VGG Model The model was tested on the MS COCO and Flickr30k datasets [22], and the performance has been compared to performance in similar works. The model with VGG feature extractor scheme has raised the CIDEr score by 15.04% [22]. The InceptionResNetV2 used in our model is the most advanced form of feature extractor, used in creating BERT models will definitely give better results.

Also, instead of using MSCOCO dataset directly, we will use more refined and stable form of the same from TensorFlow dataset, which will again enhance the image captioning accuracy and generation.

The objective of this paper is to create a model, which more accurate and precise than the existing models in generating image captions. The proposed model is capable of giving the desired results.

REFERENCES

[1] “Generating Descriptions for Images” by Alexei A. Efros and Thomas K. Leung (1999).

[2] “A Probabilistic Framework for Semantic Image Interpretation” by Antonio Torralba, Kevin P. Murphy, and William T. Freeman (2003)

[3] “Semantic Image Annotation and Retrieval using Probabilistic Image Models” by Li, L., and Wang, J. Z. (2003)

[4] “A Comparative Study of Image Captioning Approaches” by Yang, Y., and Teo, C. L. (2011)

[5] “Simulating the Human for Automatic Image Captioning” by Hodosh, M., Young, P., and Hockenmaier, J. (2013)

[6] “Deep Visual-Semantic Alignments for Generating Image Descriptions” by Karpathy & Fei-Fei, 2015

[7] “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” by Xu et al., 2015

[8] “Show and Tell: A Neural Image Caption Generator by Vinyals et al., 2015

[9] “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering” Anderson et al., 2018

[10] “Image Captioning with Semantic Attention” by You et al., 2016

[11] “Neural Baby Talk” by Lu et al., 2018

[12] “Meshed-Memory Transformer for Image Captioning” by Lei Ji, Jiajun Tang, Juanzi Li, et al. (2020)

[13] “ImageBERT: Cross-Modal Pre-training with Large-scale Weak-supervised Image-Text Data” by Xiaoyi Zha, Xinyue Cheng, Yue Liao, and Jianlong Fu (2021)

[14] “Aligning Cross-modal Space for Image Captioning” by Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen (2021)

[15] "Show and Tell: A Neural Image Caption Generator" by Vinyals et al. (2015)

[16] "Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy and Fei-Fei (2015)

[17] "Self-Critical Sequence Training for Image Captioning" by Rennie et al. (2017)

[18] "DenseCap: Fully Convolutional Localization Networks for Dense Captioning" by Johnson et al. (2016)

[19] "Image Captioning with Hierarchical Reinforcement Learning" by Ren et al. (2017)

[20] "Attention Is All You Need" by Vaswani et al. (2017)

[21] "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models" by Kiros et al. (2014)

[22] "Image Captioning Model using Attention Mechanism and Object Features to mimic Human Image Understanding", Muhammad Abdelhadie Al-Malla, Assef Jafar, Nada Ghneim (2022)