Image Caption Generator

Besharati Ali
Technische Hochschule Deggendorf
Cham, Germany
ali.besharati@stud.th-deg.de

Chaudhary Akash
Technische Hochschule Deggendorf
Cham, Germany
akash.chaudhary@stud.th-deg.de

Samant Vinayak
Technische Hochschule Deggendorf
Cham, Germany
vinayak.samant@stud.th-deg.de

Tom Emil
Technische Hochschule Deggendorf
Cham, Germany
emil.tom@stud.th-deg.de

Abstract— In the digital age, images permeate every aspect of our lives, from social media to personal photo collections. While humans can effortlessly analyze and understand visual content, enabling computers to do the same remains a significant challenge. This report focuses on leveraging advancements in computer vision and natural language processing to develop models that accurately describe images in a way that is meaningful to humans. Through this report, we have created and evaluated two distinct models to achieve effective image captioning. The first model combines EfficientNet for image feature extraction with a Transformer for caption generation. The second model employs VGG16 for feature extraction and Long Short-Term Memory (LSTM) networks for generating captions. Both models were trained on the Flickr8K dataset, which consists of 8,000 images each paired with multiple captions.

Keywords- Image caption generator, CNN + LSTM, CNN+ Transformers, Vit+GPT2

I. INTRODUCTION

In this digital era images are found in every corner such as social media, websites, and even in personal photo collection. Humans can easily understand and analyze an image with their thinking. This report focuses on the task of enabling computers to easily analyze an image by utilizing advancements in computer vision and natural language processing techniques. Our aim is to make a computer to describe an image accurately that makes sense to people. This can help a visually impaired person to understand visual content via verbal descriptions, in social media appropriate captions can be generated for photos and this can improve search engine performance.

This report deals with how we utilized advancements in artificial intelligence for generating meaningful and coherent captions for a wide range of images. The goal of the project is that intelligent models must create meaningful and accurate captions. To achieve this we have created two distinct models and their results were compared to determine the most effective approach. This comparative study will refine our technique and develop the most accurate and reliable image captioning system possible.

As the visual representations on the internet are growing, there is a need to automatically understand and describe an image. Convolutional Neural Networks (CNNs) can extract meaningful features from visual representation. Through successive layers of convolution and pooling, CNN can detect edges, textures, pattern and eventually complex objects and scenes. CNN is highly effective for tasks like object detection, image classification and segmentation. In the case of image captioning Long Short-Term Memory (LSTM) is used to convert extracted visual description to natural language description. The CNN extracted features are fed into LSTM which then predicts the sequence of words which best describe the image. Apart from LSTM, we have used Transformers in our second approach to predict sequence of words from extracted features of CNN.

II. Problem Definition

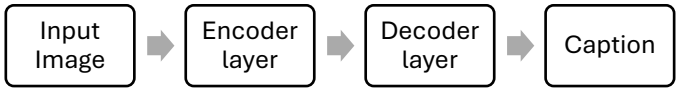
The main challenge is in accurately interpreting visual content and generating suitable, grammatically correct sentences that describe images comprehensively. This involves recognizing what the picture is and generating a caption which describes the image. The task is not only recognizing objects in the image but also framing a caption that conveys the context and details effectively. Traditional computers used pre-defined templates that relied on manually annotated images which is time-consuming and not cost effective. One critical aspect is the

selection of datasets. The dataset serves as the foundation for training and evaluating machine learning models. They need to be diverse, containing a wide range of images containing various scenarios, objects and context. The dataset needs to be annotated with descriptive captions.

Our intelligent system aims to create models which can form accurate, meaningful captions to enhance user experience.

III. System Design and Architecture

To implement Image caption generator, we have created two different model. The first model is built on EfficientNet model along with Transformer. While second model is built on VGG 16 and LSTM layers. The general architecture of both the models are as follows -



a. The purpose of an image captioning model is to produce captions describing an image[1]. These models mainly consist of an image feature extractor model coupled with a Time-Series model that Maps images to captions, usually for the image extraction part a CNN and for the time series part, a time series (recurrent neural network (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM) extended LSTM (xLSTM) or Transformers) will be used.

b. Data Acquisition & Preprocessing -

To create the model, we have used Flickr 8K dataset. It consists of around 8000 images, each paired with five captions. It provides the diverse set of images and annotations for training and evaluating the model. These captions and images are stored in a structured directory, ensuring easy access and management.

Preprocessing involved, resizing the image to a fixed size of 224x224 pixels (for Vgg16) 229x229 (for EffiecientNet) to match it to input requirements and pixel values are normalized to [0,1]. Captions were tokenized into individual words, a vocabulary of all unique words in the captions was built, and words were converted to integer indices based on the vocabulary. Sequences are padded to ensure uniform length for input into the time series network

A. Model 1-

a. Deep Learning Model – EfficientNet

Our main model was EfficientNet which was developed by Google. It achieves state-of-the-art accuracy with less parameters and computational resources, compared to its predecessors. It employs a scaling method to uniformly increase the network's width, depth, and resolution. By balancing these dimensions, EfficientNet models achieve better performance and efficiency compared to traditional CNNs.

b. Deep Learning Model – Transformer

Transformers are a type of neural network architecture primarily used in natural language processing (NLP) tasks. Introduced by Vaswani et al. [2], transformers leverage 3 main mechanisms Positional Encoding, Attention, and Self-attention. **Positional encoding** is to encode each token in a sentence on itself instead of assigning its position to the neural network (the way that RNNs work). **Attention** is to attend to the surrounding words and the context of a token, mostly used in text translation to convey a correct grammar-related translation, respecting the tense, gender of nouns, and word order. **Self-attention** is to process and weigh the relevance of different words in a sentence relative to each other. This allows transformers to capture long-range dependencies and context more effectively than previous models like RNNs or LSTMs. Their parallelizable nature also enables faster training and inference on large datasets.

Summary –

For the first set of caption models and their parameter tuning, we presented 7 different CNN architectures and multiple transformer structures. We trained all of them on the data until the plateau. After the evaluation of various CNN backbones, we evaluated different caption models based on multiple time series sections, changing the number of layers (2,5), the type of activation function (swish, Relu, Sigmoid), and using Dropout layers (30%) as the hyperparameters.

B. Model 2-

a. Overall Architecture-

The architecture of our image caption generator is designed to integrate a Convolutional Neural Network and Recurrent Neural Network to generate coherent and relevant captions for the given image. This is the earliest approach created to perform this task.

The detailed architecture includes following units:

i. Image Feature Extraction (VGG 16) –

Image features represent pieces of information about the content of the image. Generally, image features are of three types. Low level features, mid-level features and High-level features. Low level features consist of edges, corners, textures or simple shapes in the image. While mid-level features consist of objects, patterns and contours. Whereas high-level features are the entire object of significant part of an object.

To extract these image features, we used VGG 16 CNN model. This is a pretrained model and has shown excellent performance in image classification tasks. It comprises of 13 convolution layers and 3 fully connected layers. Each convolution layer uses 3x3 receptive fields. The

model is built on ReLU (Rectified Linear Unit) activation function to introduce non-linearity. In our case we have removed final fully connected layer to get the image feature vector.

ii. Time Series Model (LSTM) –

Long Short-Term Memory, also known as LSTM network, is a type of RNN network capable of learning long-term dependencies.

It generates sequence of words for a given image. LSTM takes image features and previously generated words to predict the next word in the sequence as input.

This sequence generator consists of 3 main set of layers: Embedding layer, LSTM layers and dense layer. The Embedding layer converts words into dense vector of fixed sizes. (In our case it is 1024). This layer captures semantic meanings and relationship between words. LSTM layers consist of LSTM cells with specific number of units. These units process the embedded word vectors and image feature vector. The dense layer produce a probability distribution over the vocabulary for the next word prediction.

iii. Integration –

Image feature extracted from VGG16 are combined with word embeddings of the current caption sequence. The combined vector is then fed into the LSTM layer, which predicts the next word in the sequence. This process is repeated until a complete caption is generated.

b. Summary –

The above-mentioned model was trained using Categorical cross-entropy function, which is used to measure the difference between predicted and actual word while training. We have used Adam optimizer due to its efficient handling of sparse gradients and adaptive learning rates.

To summarize this approach, the entire multimodal network architecture consist of VGG16 as an encoder layer which takes the input image, converts it into feature vector, and sends to LSTM layer or decoder layer to generate captions.

IV. Experimental Evaluation

A. Experimental Evaluation –

In order to evaluate the generated captions, we have used a pretrained model which uses ViT as encoder layer and GPT2 model as decoder layer. This approach is based on the Transformer architecture which enables them to capture long-range dependencies while dealing with sequential data. This method is mentioned in [3] paper.

ViT model is based on transformer architecture. It takes the input image and splits it into fixed sized patches. Then flattens each patch and treats each patch as an input tokens. Tokens are nothing but basic units of input that the model operates on. Tokens are typically discrete elements of the input sequence, which can represent words, sub words, characters depending on the task and model architecture.

GPT2 model is designed to perform natural language processing tasks. It takes sequence of text tokens as input and generates sequence of tokens as outputs. Due to the self-attention mechanism it captures dependencies between tokens in the output sequence.

B. Performance Results

Sets of hyperparameters of model 1, the training accuracy was around 46-48% and the validation accuracy was constant at around 41%.

For the mode 2, we got the training accuracy around 42 % and loss was reduced from 9.53 to 2.54 during the training.

To compare the different models on the fair basis, we passed the same image (Fig 1) through all the models and following are the captions generated by them -

Figure 1

Caption generated from Model 1: ‘a dog is running on the beach’

Model 2: ‘one brown dog is running along the beach carrying a stick in its mouth’

Pre-trained Model: ‘a black and white dog running on the beach’

C. Results for different CNN models –

To achieve a fair comparison between various CNN Backbones we tried ResNet50, VGG19, InceptionV3, ResNet50V2, MobileNetV2, EfficientNetB0, and EfficientNetB7 among which the best metrics were achieved by

EfficientNetB0 and B7. However, the results for these two models were close despite that B7 has a much larger number of parameters (B7 has 66M and B0 has 5.3M parameters) and is heftier and slower in learning in comparison to b0. The results of training are shown in Table 1. Table 1. Training Metrics for various CNN models

CNN model	Epochs trained	%Train Accuracy	% Val Accuracy	% Train Loss	% Val Loss
ResNet 50	16	47.24	47.24	47.24	47.24
VGG 19	16	40.62	40.62	40.62	40.62
Inception V3	16	11.62	11.62	11.62	11.62
ResNet50V2	16	15.21	15.21	15.21	15.21
EfficientNetB0	17	45.46	45.46	45.46	45.46
EfficientNetB7	16	40.15	40.15	40.15	40.15

D. Discussion

Although the learning metrics indicate a good comparison between the performance of different models and are good criteria to check if the models are progressing in the right direction, they will lack in indicating proper caption generation, particularly on new images from outside of the training data set. Therefore, we produced captions for images from within and outside of the Flicker8 dataset and compared the quality of the captions. Subjective evaluation of the generated captions was unanimous with the numerical learning Matrix, which means in that regard EfficientNet B0 was the best, however, generated captions for many images are either unrelated (Fig 2) or inaccurate (Fig 3). Such shortcomings of the models are mostly observed when the input image is from a far distance and has many details instead of one or a few main objects that occupy most of the image.

Figure 2 Caption generation for an image outside of flickr8 dataset.

Figure 1.

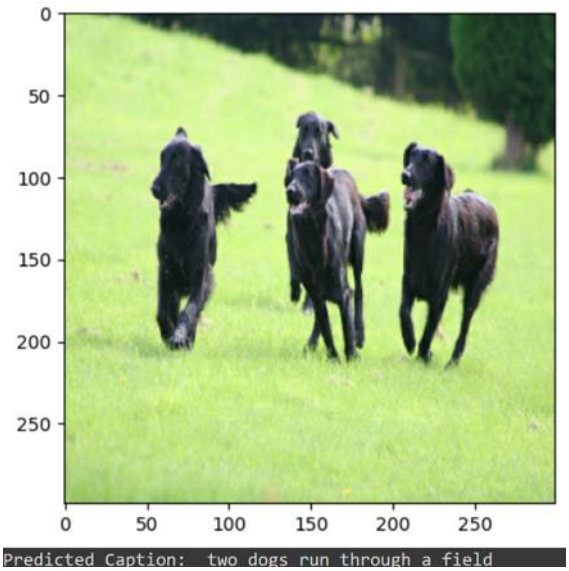


Figure 3 Caption generation for an image from the flickr8 dataset

V. Real-world applications and Deployment

A. Real-world applications -

Image caption generators are becoming increasingly prevalent in various real-world scenarios, revolutionizing the way we interact with visual content. These systems, which

leverage advanced AI techniques, are being deployed across diverse industries and domains.

- *Enhanced Accessibility* -

For visually impaired individuals, image caption generators offer significant value by providing audio descriptions of visual content. This allows users to understand online content, social media posts, and even physical world images through assistive technologies. Integration with screen readers enables seamless access, thereby fostering greater social inclusion and accessibility.

- *Revolutionized Image Search*-

Traditional image search engines often rely on keyword matching, which can lead to irrelevant results. Image caption generators facilitate semantic search by understanding the content of images and retrieving results based on their meaning. This improves search efficiency and user experience, allowing users to find images based on specific objects, actions, or scenes depicted.

- *Streamlined Content Management*-

Businesses and organizations generate vast amounts of digital content, including images. Manually tagging and organizing these images is a time-consuming task. Image caption generators automate this process by generating descriptive captions for image tagging and categorization. This not only saves time and resources but also enhances the searchability and organization of image libraries.

- *Engaging Social Media Content*-

Social media platforms thrive on visual content. Image caption generators can automatically create engaging captions for photos and videos, boosting user interaction and engagement. This functionality enhances the overall user experience on platforms like Instagram and Facebook.

- *Improved Robot Navigation*-

In the realm of robotics, image caption generators are integrated with robot vision systems to provide real-time descriptions of the environment. This aids robots in identifying objects, navigating obstacles, and interacting more effectively with their surroundings, thereby improving their operational efficiency and safety.

- *Case Study: Microsoft's Seeing AI*

Microsoft's Seeing AI application uses image captioning to assist visually impaired individuals by providing real-time verbal descriptions of objects, people, and scenes captured by the user's smartphone camera. This application has significantly enhanced the independence and quality of life for visually impaired users by granting them greater access to visual information.

- *Case Study: Google Photos*

Google Photos leverages image captioning to organize and categorize users' photo libraries automatically. By generating descriptive tags and captions, it enables efficient search and retrieval of images. This enhances the user experience and management of large photo collections.

B. Deployment -

Despite the numerous benefits, deploying image caption generators in real-world settings presents several challenges and limitations that need to be addressed for optimal performance and user satisfaction.

- *Data Quality and Diversity*-

High-quality, diverse datasets are essential for training effective image caption generators. Inadequate or biased data can result in poor performance and biased outputs. It is crucial to use large, diverse datasets and continuously update them to include new and varied images. Data augmentation techniques can also be employed to enhance dataset diversity.

- *Model Complexity and Computational Resources*-

Advanced models, especially those combining Convolutional Neural Networks (CNNs) and transformers, require substantial computational power for training and inference. Optimizing models for efficiency and deploying them on scalable cloud-based infrastructure with hardware accelerators like GPUs and TPUs can manage these computational demands effectively.

- *Integration with Existing Systems*

Seamlessly integrating image caption generators with existing applications and platforms can be complex. Developing modular APIs and interfaces facilitates easy integration. Collaboration with stakeholders ensures that the system meets user requirements and integrates smoothly with existing workflows.

- *Ethical Considerations and Bias*

AI systems can inadvertently perpetuate biases present in the training data, leading to unfair or biased outputs. Implementing bias detection and mitigation strategies, and regularly auditing the system's outputs, is essential to ensure fairness and inclusivity.

- *Performance, Accuracy, and Reliability*

Evaluating the performance, accuracy, and reliability of image caption generators in practical settings is crucial. Research and case studies provide valuable insights into their effectiveness.

VI. Conclusion and Future Directions

A. Summary of Key Findings

The development and deployment of image caption generators have led to significant advancements and practical

applications. These systems offer numerous benefits, including enhanced accessibility, improved content management, and enriched user engagement. However, challenges such as data quality, computational demands, and ethical considerations must be addressed to maximize their effectiveness.

B. *Future Research Directions*

The future of image caption generation holds immense promise, with several exciting advancements on the horizon:

i. *Multimodal Captioning* -

Integrating additional modalities, such as audio descriptions or object detection, can provide a richer and more comprehensive understanding of image content. This multimodal approach enhances the accuracy and relevance of generated captions.

ii. *Domain-Specific Captioning* -

Developing models tailored to specific domains, such as medical imaging or art history, can lead to more specialized and accurate captions. This opens new avenues for applying image captioning technology in niche areas, catering to specific industry needs.

iii. *Explainable AI*

Enhancing the explainability of image caption generators improves user trust and transparency. By unveiling the reasoning behind generated captions, users can better understand and trust the system's outputs. This is particularly important in applications where explainability is paramount, such as medical image analysis.

iv. *Lifelong Learning*

The ability of image caption generators to continuously learn and improve over time will be crucial. Techniques like continual learning allow models to adapt to new data without forgetting previously learned information. This ensures that image caption generators remain relevant and effective as new visual content emerges.

v. *Broader Impact: The Future of AI*

Image caption generation is emblematic of a broader trend in AI: the convergence of computer vision and natural language processing. As these fields continue to evolve, we can expect more powerful and sophisticated systems that not only understand visual content but also interact with the world in a more human-like way.

vi. *Potential Areas for Improvement*

Continued research and development are essential to enhance the accuracy, robustness, and versatility of image caption generators. Areas such as human-robot collaboration, augmented reality (AR) and virtual reality (VR), and education and training present significant opportunities for the application of this technology.

C. *Conclusion* -

Through this project, e have implemented Image Caption Generator using various two different methods. One is different CNN architectures with Transformers and the second with Vgg16 with LSTM. In the end we have compared their results with the existing trained Vit + GPT2 model.

REFERENCES

[1] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Feb. 2015, [Online].

[2] A. Vaswani, "Attention Is All You Need," Long Beach, CA, USA, 31st Conference on Neural Information Processing

[3] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

[4] Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing

[5] Python Based Project to Build Image Caption Generator With CNN & LSTM

[6] Show and Tell: A Neural Image Caption Generator

Appendix

1. Code Repository – https://drive.google.com/drive/folders/14UkVcArRmDGC9YYUNk-yfynuFg_68SMO?usp=sharing

2. Working Videos – https://drive.google.com/file/d/1uUwTj_1UsOVy0YXlFrrg1oCFaZgzO0UP/view?usp=drive_link