A Project report on

Visual Content Captioning

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

S. Deepthi
(20H51A0551)

S. BhagyaShree
(20H51A05J7)

L. Jatin
(20H51A05P2)

Under the esteemed guidance of

Mr. T. Upender
(Assistant Professor)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase I report entitled **“Visual Content Captioning"** being submitted by S. Deepthi (20H51A0551), S. BhagyaShree (20H51A05J7), L. Jatin (20H51A05P2) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mr. T. Upender
Assistant Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr. T. Upender, Assistant Professor**, Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Ghanta Devadasu**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

S. Deepthi	20H51A0551
S. BhagyaShree	20H51A05J7
L. Jatin	20H51A05P2

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

In the modern era, visual content captioning has become one of the most widely required tools. Moreover, there are inbuilt applications that generate and provide a caption for a certain image, all these things are done with the help of deep neural network models. The process of generating a description of an image is called visual content captioning. It requires recognizing the important objects, their attributes, and the relationships among the objects in an image. It generates syntactically and semantically correct sentences.In this report, we present a deep learning model to describe images and generate captions using Computer Vision and Machine translation. This report aims to extract features of different objects found in an image, recognize the relationships between those objects and generate captions.

The dataset used is Flickr8k, it consists of 8000 images and for every image, there are 5 captions. The 5 captions for a single image helps in understanding all the various possible scenarios, the programming language used was Python3. For evaluation of the performance of the described model we can use BLEU scores. Through the scores, one can apart the generated captions as good captions and bad captions.This report will also elaborate on the functions and structure of the various Neural networks involved. Generating image captions is an important aspect of Computer Vision and Natural language processing. We will be implementing the visual content caption generator using CNN (Convolutional Neural Networks) and Transformer. The image is passed to the CNN model to extract features. The image features are passed to the Transformer encoder to encode them into a vector representation. Then Transformer decoder is used to generate the caption from the encoded image features. The caption is decoded and predicted caption is printed. Main applications of this model include usage in virtual assistants, for image indexing, for social media, for visually impaired people, recommendations in editing applications and much more.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1.Problem Statement

In today's data-driven world, this model bridges gap between natural language processing as well as computer vision, the convergence of these two has given rise to the fascinating field of Visual Content Captioning . To develop a deep learning model that can automatically generate captions for images, using the pre-trained image captioning model provided . The model should be able to generate captions that are accurate, informative, and grammatically correct. The captions should also be relevant to the content of the image. It will be evaluated on a held-out test set of images. The evaluation metrics will include accuracy, fluency, and relevance. At its core, with encoder-decoder architecture this project harnesses the power of Convolutional Neural Networks (CNNs) and Transformers networks to automatically generate descriptive text for images. By employing CNNs for image feature extraction and Transformers for vector representation and language generation, this technology holds the potential to revolutionize content accessibility, image indexing, and enrich the human-computer interaction experience. In this endeavour, we embark on a journey to explore the capabilities and applications of this fusion of neural networks, aiming to bridge the gap between visual data and human understanding.Some of the major applications of the application are self-driving cars wherein it could describe the scene around the car, secondly could be an aid to the people who are blind as it could guide them in every way by converting scene to caption and then to audio, CCTV cameras where the alarms could be raised if any malicious activity is observed while describing the scene, recommendations in editing, social media posts, and many more.

1.2.Research Objective

The research objective of our project is to develop a model that can automatically generate captions for images. This is a challenging task, as it requires the model to understand the content of the image and to generate a caption that is both informative and engaging.

Here are some research objectives that we could consider for our project :

- **Improve the quality of the generated captions:** This could involve using a different pre-trained model, training your own model on a larger dataset, or using different techniques to generate the captions.
- **Make the model more interpretable:** This would allow you to understand how the model is generating the captions and to identify any potential biases in the model.

- **Make the model more efficient:** This could involve optimizing the code or using a different model architecture.

We could also consider combining these research objectives to create a more comprehensive project . For example, we could train a new captioning model on a dataset of images and their corresponding captions from a specific domain, such as social media or news articles. We could then evaluate the performance of the model on a held-out test set and compare it to the performance of a pre-trained model. Finally, deploy the model to a production environment and use it to generate captions for real-world data.

1.3 Project Scope and Limitations

The project scope are :

- **Image Caption Generation:** The primary goal of an Image Caption Generator is to automatically generate descriptive and contextually meaningful textual captions for a given image.
- **Multimodal Integration:** Image Caption Generators typically combine computer vision techniques (e.g., Convolutional Neural Networks or CNNs) for image analysis and natural language processing techniques (e.g., Recurrent Neural Networks or RNNs with LSTM,Transformer) for text generation.
- **Applicability:** These systems can find applications in various domains, including content enrichment, image indexing, accessibility for visually impaired individuals, and enhancing human-computer interaction.
- **Language Support:** Depending on the design, the generator may be designed to produce captions in a specific language or support multiple languages.
- **Scalability:** A well-designed Visual Content Caption Generator should be scalable, capable of handling a wide range of image types and large datasets.

The limitations are :

- **Object Recognition Errors:** Accuracy in recognizing objects and scenes within images is contingent on the quality of training data and model architecture, leading to occasional recognition errors.

- **Ambiguity Handling:** Handling ambiguous images or scenarios where multiple valid captions are possible can be challenging for image caption generators.

- **Generalization Limitations:** The system may struggle to describe concepts, objects, or scenes it hasn't encountered during training, highlighting limitations in generalization.

- **Lack of Common Sense Reasoning:** Image caption generators often lack common-sense reasoning abilities, potentially leading to captions that are factually incorrect or implausible.

- **Evaluation Metric Limitations:** Common evaluation metrics like BLEU and METEOR may not fully capture the quality of generated captions or align with human judgment.

- **Resource Intensity:** Complex image caption generators can be computationally intensive during both training and inference, demanding significant computing resources.

CHAPTER 2

BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1 Show and Tell

2.1.1. Introduction

Show and Tell is a simple but effective image captioning model. It consists of two main components: a convolutional neural network (CNN) and a recurrent neural network (RNN). The CNN is used to extract features from the image, and the RNN is used to generate the caption.

The CNN first downsamples the image to a smaller size. This helps to reduce the computational complexity of the model and also makes it more robust to noise in the image.
The CNN then extracts features from the image at different levels of abstraction. The lower-level features represent the basic visual features of the image, such as edges and corners. The higher-level features represent more complex concepts, such as objects and scenes.
The extracted features are then passed to the RNN. The RNN generates the caption one word at a time. At each time step, the RNN takes the current hidden state and the previous word as input and generates the next word as output. Show and Tell is a simple and effective image captioning model. It is easy to implement and can be trained with a relatively small dataset. However, Show and Tell can generate inaccurate or repetitive captions, and it struggles to capture long-range dependencies in the image and the caption.

The training process for Show and Tell is as follows:

1. A batch of images and their corresponding captions is loaded.
2. The CNN is used to extract features from the images.
3. The features are passed to the RNN.
4. The RNN generates the captions.
5. The captions are compared to the ground truth captions and a loss is calculated.
6. The parameters of the CNN and RNN are updated to minimize the loss.

2.1.2. Merits, Demerits and Challenges

- **Merits:**

  1. Simple to implement
  2. Can be trained with a relatively small dataset

- **Demerits:**
  1. Can generate inaccurate or repetitive captions
  2. Struggles to capture long-range dependencies in the image and the caption

- **Challenges:**
  1. Generating captions for complex images
  2. Generating captions in multiple languages

2.1.3.Implementation

- Can be implemented using a variety of deep learning libraries, such as PyTorch and TensorFlow
- Pre-trained models are available online

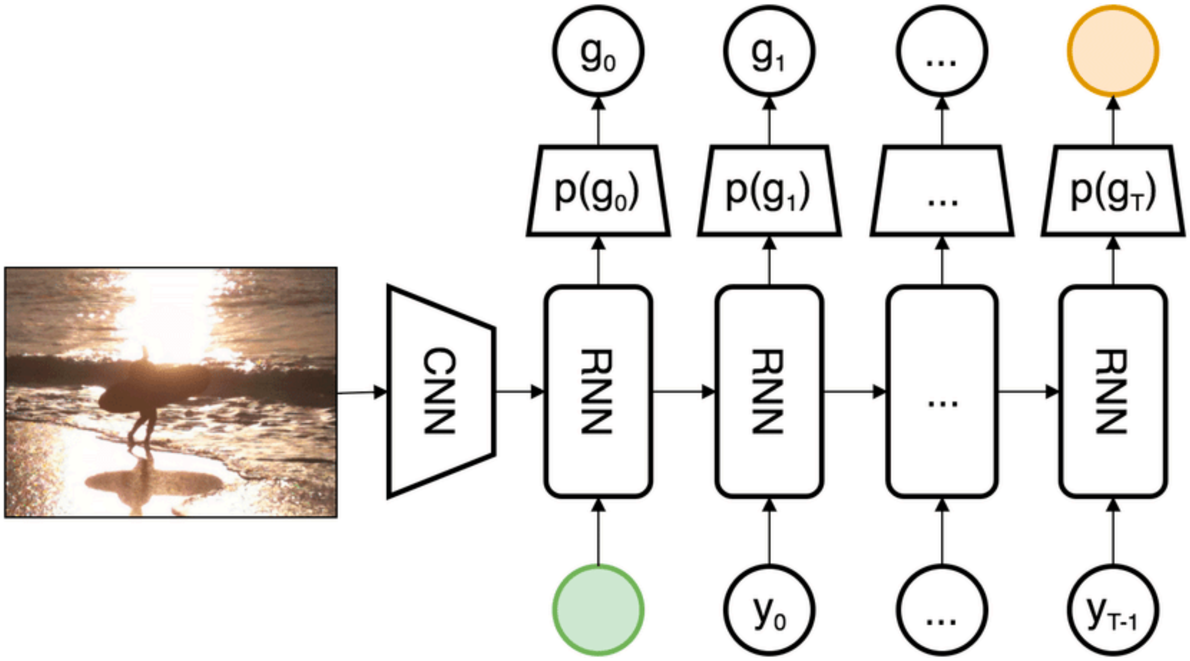


Fig:2.1.1

2.2 Attend and Tell

2.2.1. Introduction

Attend and Tell is an extension of the Show and Tell model. It introduces an attention mechanism that allows the model to focus on different parts of the image as it generates the caption.

The attention mechanism works as follows:

1. The model is given a set of features extracted from the image by the CNN.
2. The model then generates a set of attention weights. The attention weights represent how important each feature is to the current word being generated.
3. The attention weights are then used to compute a weighted sum of the features.
4. The weighted sum of the features is then passed to the RNN to generate the next word in the caption.

The training process for Attend and Tell is similar to the training process for Show and Tell. However, the attention mechanism is added to the RNN.

Attend and Tell improves the accuracy of Show and Tell by using an attention mechanism. However, Attend and Tell is more complex to implement than Show and Tell, and it requires a larger dataset to train.

2.2.2. Merits, Demerits and Challenges

- **Merits:**

  1. Improves the accuracy of Show and Tell by using an attention mechanism
  2. Can capture long-range dependencies in the image and the caption

- **Demerits:**

  1. More complex to implement than Show and Tell
  2. Requires a larger dataset to train

- **Challenges:**

  1. Generating captions for complex images
  2. Generating captions in multiple languages

2.2.3.Implementation

- Can be implemented using a variety of deep learning libraries, such as PyTorch and TensorFlow
- Pre-trained models are available online

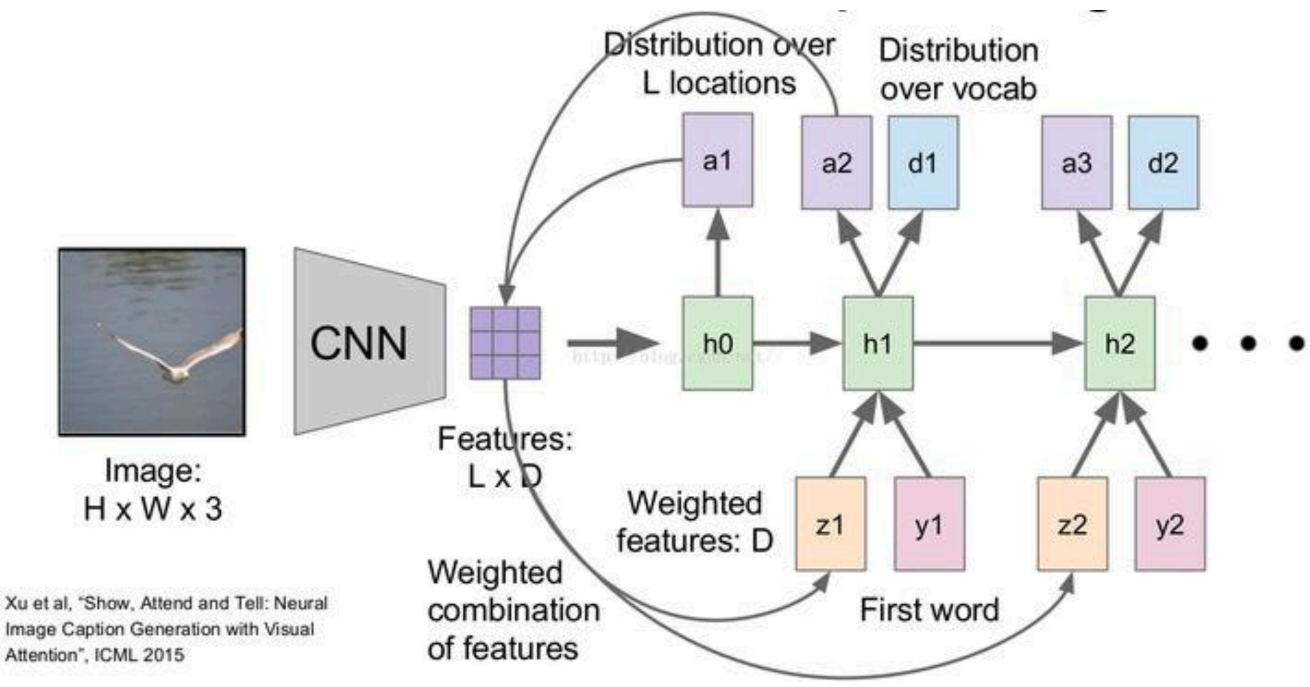


Fig:2.1.2

2.3 Transformer

2.3.1. Introduction

- The Transformer is a neural network architecture that was originally developed for machine translation. However, the Transformer has also been shown to be effective for image captioning.
- The Transformer consists of two main components: an encoder and a decoder. The encoder is used to encode the image into a sequence of hidden states. The decoder is used to decode the hidden states into the caption.
- The encoder consists of a stack of self-attention layers. Self-attention is a mechanism that allows the model to learn long-range dependencies in the image.
- The decoder consists of a stack of decoder layers. Each decoder layer consists of a self-attention layer and a masked attention layer. The masked attention layer prevents the model from attending to future words in the caption.
- The training process for the Transformer is similar to the training process for Show and Tell and Attend and Tell. However, the Transformer architecture is more complex than the Show and Tell and Attend and Tell architectures.
- The Transformer has achieved state-of-the-art results on a number of image captioning benchmarks. However, the Transformer is more complex to implement than Show and Tell and Attend and Tell, and it requires a very large dataset to train.

2.3.2. Merits, Demerits and Challenges

- **Merits:**

    1. Achieved state-of-the-art results on a number of image captioning benchmarks
    2. Can capture long-range dependencies in the image and the caption

- **Demerits:**

    1. More complex to implement than Attend and Tell
    2. Requires a very large dataset to train

- **Challenges:**

  1. Generating captions for complex images
  2. Generating captions in multiple languages

2.3.3.Implementation

- Can be implemented using a variety of deep learning libraries, such as PyTorch and TensorFlow
- Pre-trained models are available online

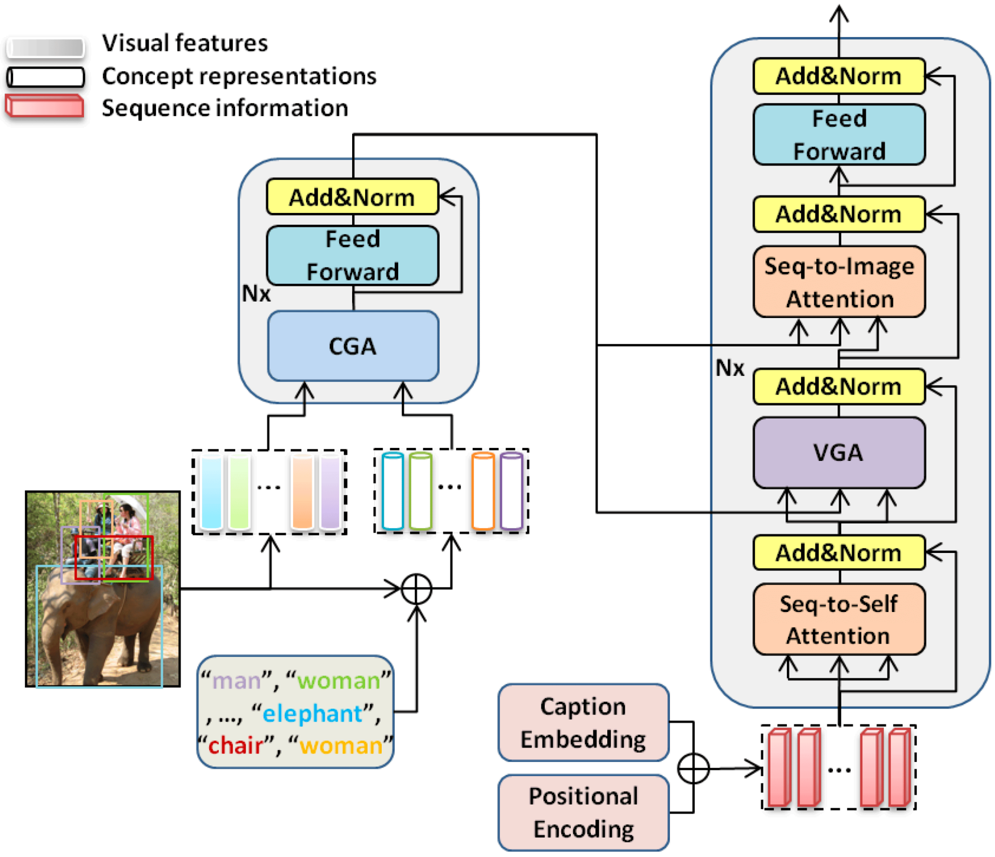


Fig:2.1.3

2.4 Captioning with Bidirectional Reasoning

2.4.1. Introduction

Captioning with Bidirectional Reasoning is an image captioning solution that uses bidirectional reasoning to consider the context of the entire image when generating captions. This can help to generate more informative and accurate captions.
The model works as follows:

1. The image is passed to a CNN to extract features.
2. The features are passed to an RNN to generate a sequence of hidden states.
3. The hidden states are passed to a bidirectional attention mechanism.
4. The bidirectional attention mechanism generates a sequence of attention weights.
5. The attention weights are used to compute a weighted sum of the hidden states.
6. The weighted sum of the hidden states is passed to another RNN to generate the caption.

The training process for Captioning with Bidirectional Reasoning is similar to the training process for Show and Tell. However, the bidirectional attention mechanism is added to the RNN.

Captioning with Bidirectional Reasoning can generate more informative and accurate captions than Show and Tell, as it can consider the context of the entire image. However, Captioning with Bidirectional Reasoning is more complex to implement than Show and Tell, and it requires a larger dataset to train.

2.4.2. Merits, Demerits and Challenges

- **Merits:**
  Can generate more informative and accurate captions by considering the context of the entire image

- **Demerits:**
  1. More complex to implement than other solutions
  2. Requires a larger dataset to train

- **Challenges:**
  1. Generating captions for complex images
  2. Generating captions in multiple languages

2.3.3.Implementation

- Can be implemented using a variety of deep learning libraries, such as PyTorch and TensorFlow
- Pre-trained models are available online

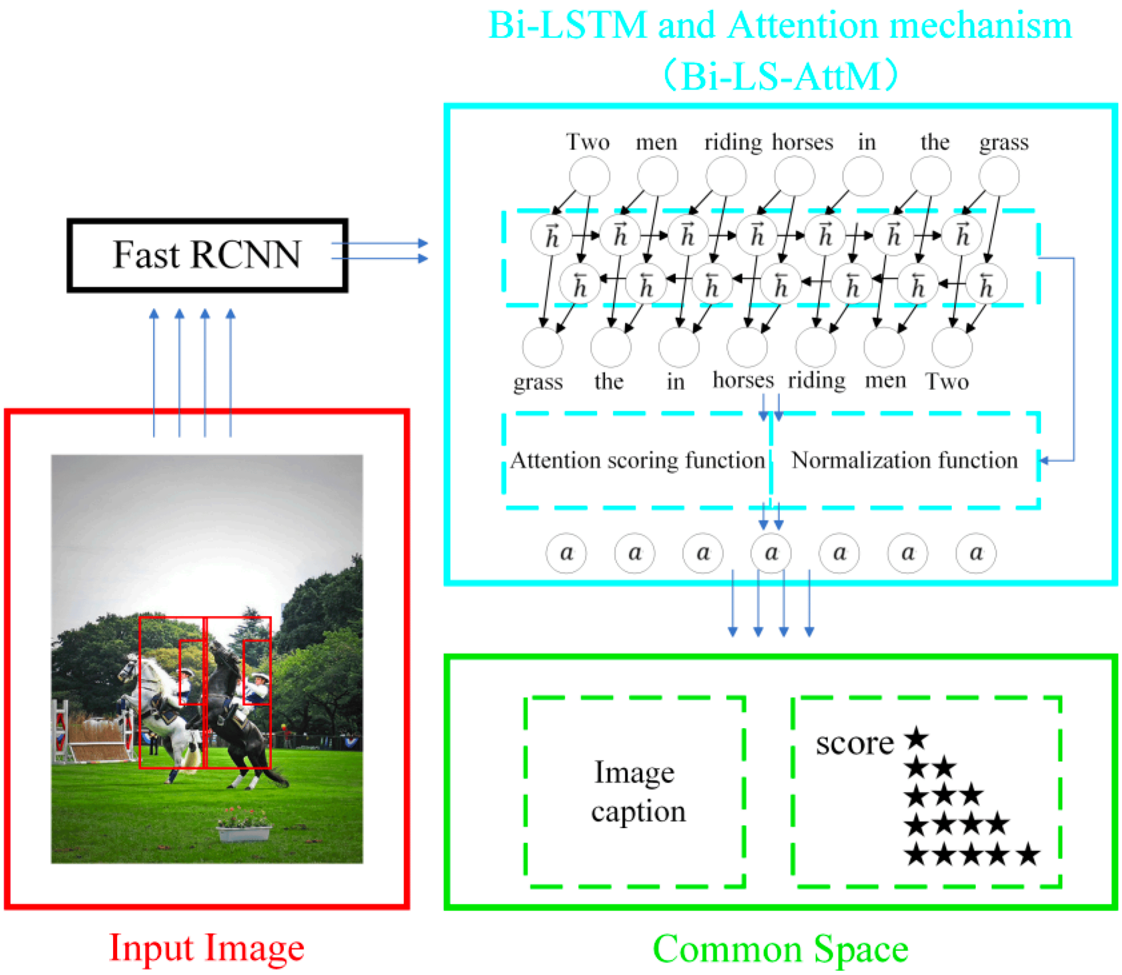

Bi-LSTM and Attention mechanism
（Bi-LS-AttM）
Two men riding horses in the grass
grass the in horses riding men Two
Attention scoring function
Normalization function
a a a a a a a
Fast RCNN
Input Image
Image caption
score
Common Space

Fig:2.1.4

CHAPTER 3
RESULTS AND DISCUSSION

CHAPTER 3
RESULTS AND DISCUSSION

Creating a performance matrix for an visual content caption generator using a combination of Convolutional Neural Networks (CNNs) and Transformers is an essential step in evaluating and improving your model. A performance matrix is typically a set of metrics and evaluation techniques used to measure the quality of generated captions. Here are some common performance metrics for image caption generators:

1. BLEU Score:

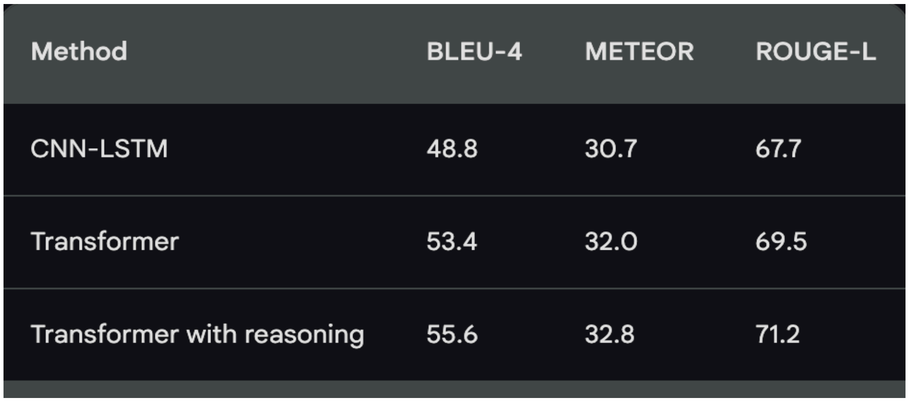
- BLEU (Bilingual Evaluation Understudy) is a commonly used metric to measure the quality of machine-generated text, such as image captions.
- It calculates the precision of n-grams (usually 1-4) in the generated captions compared to reference captions.
- Higher BLEU scores indicate better quality captions.

2. ROUGE Score:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the similarity between generated captions and reference captions.
- Like BLEU, ROUGE considers various n-gram overlapping measures.
- ROUGE scores help assess the fluency and coherence of the generated captions.

3. METEOR Score:

- METEOR (Metric for Evaluation of Translation with Explicit ORdering) is another metric that measures the quality of machine-generated text.
- It considers precision, recall, stemming, synonymy, and word order.
- METEOR is considered more robust than BLEU for certain applications.



Method	BLEU-4	METEOR	ROUGE-L
CNN-LSTM	48.8	30.7	67.7
Transformer	53.4	32.0	69.5
Transformer with reasoning	55.6	32.8	71.2

Fig:3.1

CHAPTER 4
CONCLUSION

CHAPTER 4
CONCLUSION

In conclusion, our Visual Content Captioning represents a significant step toward bridging the gap between computer vision and natural language processing.Visual Content Captioning using CNN and transformer are a promising new approach and is a powerful new technique that has the potential to revolutionize the way we interact with digital images.They have the potential to overcome the shortcomings of existing image caption generators and to generate more accurate, informative, and creative captions, by combining the strengths of CNNs and transformer models, they can capture long-range dependencies in images, understand the relationships between objects in images, and generate captions that are both accurate and informative.

Here are some additional benefits of using CNN and transformer for image caption generation:

- **End-to-end training:** CNN and transformer models can be trained end-to-end, which means that they can learn to extract visual features and generate captions in a single step. This is more efficient than previous approaches, which typically involved training separate models for visual feature extraction and caption generation.
- **Scalability:** CNN and transformer models can be scaled to large datasets, which makes them suitable for training on large-scale image captioning datasets.
- **Flexibility:** CNN and transformer models can be adapted to a variety of image captioning tasks, such as generating captions for social media images, e-commerce product images, and news article images.

One of the most exciting applications of image caption generators using CNN and transformer is in the field of accessibility.

REFERENCES

REFERENCES

1. HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020)
2. B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, (international Journal of Advanced Science and Technology- 2020)
3. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, “A Comprehensive Survey of Deep Learning for Image Captioning” ,(ACM-2019)
4. Rehab Alahmadi, Chung Hyuk Park, and James Hahn, “Sequence-tosequence image caption generator”, (ICMV-2018)
5. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”,(CVPR 1, 2- 2015)
6. Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma Sharma, “Visual Image Caption Generator Using Deep Learning”, (ICAST-2019)
7. Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode, “Camera2Caption: A Real-Time Image