ASSAM UNIVERSITY

Design and Development of Handwritten Mathematical Image Captioning System

A report submitted in partial fulfilment of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted By:

Aishik Das

Registration Number: **20200016764** of **2020-21**

Anupal Saikia

Registration Number: **20200016745** of **2020-21**

Kasturi Sharma

Registration Number: **20210000064** of **2020-21**

Under the guidance of:

Dr. Sourish Dhar

Assistant Professor, Department of Computer Science and Engineering

Triguna Sen School of Technology

Department of Computer Science and Engineering

Assam University, Silchar 788011

April 2025

Declaration of Authorship

We, the undersigned, declare that this report titled "**Design and Development of Handwritten Mathematical Image Captioning System**" and the work presented in this report is our own. We confirm that this work submitted for assessment is our own and is expressed in our own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. To the best of our knowledge and belief, the same report has not been submitted either by us or by any other person for the award of any other degree or diploma of the University or other institute of higher learning.

Candidate: 1. Aishik Das - 20200016764

Candidate: 2. Anupal Saikia - 20200016745

Candidate: 3. Kasturi Sharma - 20210000064

Date:

Place:

Department of Computer Science and Engineering
Assam University, Silchar-788011

Certificate

This is to certify that the report entitled
submitted by
Aishik Das
Registration Number: **20200016764** of **2020-21**
Anupal Saikia
Registration Number: **20200016745** of **2020-21**
Kasturi Sharma
Registration Number: **20210000064** of **2020-21**
to the Department of Computer Science and Engineering, Assam University, Silchar in partial fulfillment of the requirements for the award of the Degree of
Bachelor of Technology
in
Computer Science and Engineering
is a bonafide record of the work carried out them under my supervision. It is further certified that the candidates have complied with all the formalities as per the requirements of Assam University.

Name & Signature (Supervisor)

Prof. Sudipta Roy (Head of the Department)

External Examiner's Name & Signature

Date and Seal

Abstract

In this project, we proposed a novel method for generating textual descriptions and LaTeXnotations for images of handwritten mathematical expressions. Our method is based on a well-annotated dataset which includes images of handwritten mathematical expressions along with their textual descriptions and LaTeXnotations. The given set of data is rather helpful for the model training and testing processes. For our study we developed a fusion model with the help of this model, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). For extracting the features from images, several pretrained CNN models like VGG16, VGG19, ResNet50, InceptionV3 have been employed. These models are commonly used for extracting some important features from the images. Next, to handle the order of text generation issue RNN models were employed. Two variants of RNNs which are popular for their use are the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU). They were employed for learning short-term and long-term dependencies in the sequences of text. In the last part of the model we have used a softmax layer to generate the output. It would be in the form of a text description or in the form of a LaTeXnotation of the handwritten mathematical expression image. Thus, in our study, we have achieved the highest accuracy of 68.9% success rate in generating textual descriptions and 75.7% for LaTeXnotations. This work indicates that a combination of NLP with CV can help in the recognition of handwritten mathematical expressions. This method may be of benefit to the general public in increasing the improvement of the tools used in education as well as making much content on the particular subject accessed and easily understandable by physically disabled persons.

Acknowledgements

This project explores the process of creating text descriptions for images of handwritten mathematical expressions. It's an important step in making math content more accessible and easier to understand. Through a comprehensive analysis of existing research, methodologies, and technologies, this survey explores the advancements in optical character recognition (OCR) techniques, deep learning models, and sequence-to-sequence architectures tailored for converting handwritten mathematical symbols and equations into human-readable text. By examining various approaches and their effectiveness in accurately deciphering intricate handwritten characters, this survey offers insights into the challenges, trends, and future directions in this evolving field.

Contents

List of Figures

List of Tables

Abbreviations

HMER	**H**andwritten **M**athematical **E**xpression **R**ecognition
HMEI	**H**andwritten **M**athematical **E**xpression **I**mages
NLP	**N**atural **L**anguage **P**rocessing
CV	**C**omputer **V**ision
CNN	**C**onvulational **N**eural **N**etwork
VGG	**V**isual **G**eometric **G**roup
RNN	**R**ecurrent **N**eural **N**etwork
LSTM	**L**ong **S**hort **T**erm **M**emory
GRU	**G**ated **R**ecurrent **U**nit
BLEU	**B**i**L**ingual **E**valuation **U**nderstudy
METEOR	**M**etric for **E**valuation of **T**ranslation with **E**xplicit**OR**dering
ROUGE	**R**ecall **O**riented **U**nderstudy for **G**isting **E**valuation

Dedicated to our parents, whose unwavering love and support have been our greatest strength and inspiration.. . .

Chapter 1

Introduction

1.1 What is Image Captioning?

"Image captioning—the task of providing a natural language description of the content within an image" [1]. Image captioning is the process of creating descriptive text for images. This involves analyzing the visual content of an image and producing a natural language description that accurately reflects the content or objects within it. Usually, it begins with the vision analysis, during which objects, actions, and other components are searched and recognized in the image. This is usually done with the help of deep learning approaches, most commonly Convolutional Neural Networks (CNN). Once the object recognition has been done, a language model, which can be either Recurrent Neural Network (RNN) or Transformer, creates a semantically coherent and contextually appropriate caption based on these features [2].

Image Captioning is the intersection of Natural Language Processing and Computer Vision. .In image captioning NLP is used for generating the textual descriptions and CV for interpreting the contents of images [3]. CV assists in the

FIGURE 1.1: A farmer pulling out radish(credit:Getty images

FIGURE 1.2: A man surfing on wave(credit:Wikimedia)

recognition of generic objects and activities within the image and NLP turns this data into comprehensible and meaningful text. Combined, they allow for creation of correct and timely descriptions for images.

1.1.1 History of Image Captioning

The generation of image captions has progressed over time. At the beginning, it was based on the technical captions where the translator selected either prepared in advance or from a list of standard recorded captions. During the early 2000s, machine learning methods appeared frequently and were particularly focused on the retrieval-based methods wherein the image they got was matched with the sentence they had from the predefined set of possible sentences.

The scale of a breakthrough arrived in the 2010s with deep learning's emergence. Convolutional Neural Networks or CNNs began to look into image content and Recurrent Neural Networks or RNNs which include LSTM networks started to generate captions. This was a big improvement which enabled dynamic and more accurate description generation.

Today, Transformer models are dominant and enhance not only the understanding but also the generation of captions. These models utilizing big datasets and powerful algorithms generate rich descriptive text which is very close to the given input images.

1.1.2 Techniques of Image Captioning

There are two main techniques of image captioning techniques [3]:

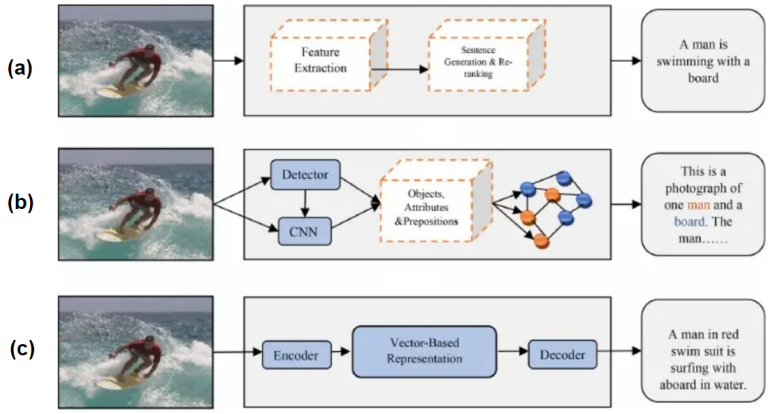

FIGURE 1.3: (a) Retrieval-Based Caption Model (RCM), (b) Template-Based Caption Model (TCM), (c) Deep-Learning-Based Caption. *adapted from* [4]

- **Retrieval and Temporal-based Image Captioning:**This traditional method uses a query image to find matching sentences from a predefined pool to describe the image. The caption can be a single sentence or a combination of sentences found in the pool. Temporal method generates captions by following a strict process that ensures both correct syntax and meaning.
- **Deep Learning-based Image Captioning:** This modern approach uses neural networks to get the label descriptions of the images. The neural networks are trained with images and their captions. Hence, they are capable of generating proper descriptions on the image all alone. This method is used by us for the research.

1.1.3 Application

Image captioning has several valuable use cases:

- **Assisting Visually Impaired Individuals:** For users with some form of vision impairment, it can allow them to get a feel of visual content by describing the picture. This text can explain something in the picture to the users, so, therefore, to enable the users have a feel of the world and what they come across with.
- **Enhancing Visual Question Answering Systems:** Image captioning is essential in systems that are employed in answering questions regarding images. Because it describes the content of the VIS images it strengthens abilities of the foregoing systems to give accurate and relevant answers concerning what an image illustrates.

- **Supporting Language Learners:** For individuals in the process of learning the loan language, the image captioning will be quite useful. It provides context-restricted use of language, by coming up with captions that states descriptions of images which after learning, would assist the learner in knowing how to use the new vocabularies and syntactic patterns in context.

1.2 Problem Domain

Mathematical image captioning is related to converting an image containing mathematical content into a readable and natural language description or into an accurate LaTeXnotation. The reason is to transform the given mathematical expression into an easily understandable form. The types of mathematical image are: a) Printed Mathematical Expression Images b) Handwritten Mathematical Expression Images [4]. We have performed our research work by referring to *Handwritten Mathematical Expression Images* (HMEI) [5] [6].

For an example, we can take the expression "$x^2 + 2x + 1 = 0$". The textual description of the expression is *"The quadratic equation x squared plus two x plus one equals zero."* and the markup is "\[x^2 + 2x + 1 = 0\]".

1.3 Motivation

In recent years, Handwritten Mathematical Expression Recognition (HMER) has emerged as a key research area. Image captioning plays a important role in recognizing and generating descriptions for handwritten mathematical expressions.The primary challenge in handwritten mathematics recognition is ambiguity. This problem arises from the variability in human handwriting because different people have different way to write an alphabet or character. For example, the numeral "1" can be easily confused with the letter "l" due to similar strokes.

Our motivation for this research is to enhance accessibility for visually impaired individuals by converting handwritten math into text and then into audio, making mathematical content more accessible. It also supports the trend of digitized education by transforming handwritten notes into digital formats, streamlining the process. This automation saves time and reduces errors, which is especially valuable in fields like engineering, physics, and data analysis, where quick and

accurate conversion is crucial. Additionally, it assists students in verifying their work and understanding complex expressions more easily .

1.4 Objective

The broader objective of generating textual descriptions and LaTeXnotation for handwritten mathematical expressions is to make mathematical content more accessible and manageable. By converting handwritten math into clear text and LaTeXmarkup, we aim to help people to understand, share, and work with mathematical information. This process not only supports education and research by digitizing and organizing content but also assists in understanding complex expressions and usable for a wider audience [6] [7].

We have three major contributions in our research:

- We have reviewed a total of 24 research papers for the purpose of our research.
- We have created an annotated dataset of handwritten mathematical expressions with corresponding LaTeXnotation and textual descriptions to train and test our model.
- We have developed an end-to-end CNN-RNN-based deep learning model capable of generating both the LaTeXnotation and textual description of a given input handwritten mathematical expression image . This model helps us to interpret handwritten mathematical expressions and images accurately.

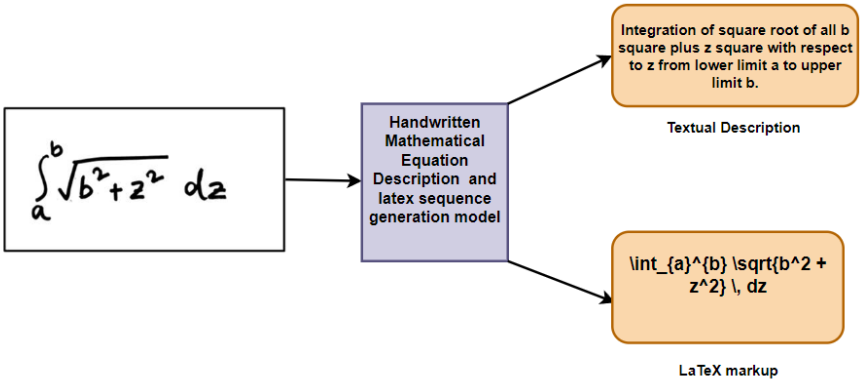

FIGURE 1.4: Mathematical Image Captioning

1.5 Challenges

We have faced two major challenges during the implementation of our work.

- There is currently no standard offline dataset available for training models to generate captions for handwritten mathematical expressions. So we have to create our own dataset for our work.
- One of the major challenges we encountered during implementation was the ambiguity in handwriting or the difficulty in accurately recognizing handwritten text [8]. This issue made it tough to convert handwritten mathematical expressions into clear and precise captions. As an Example,

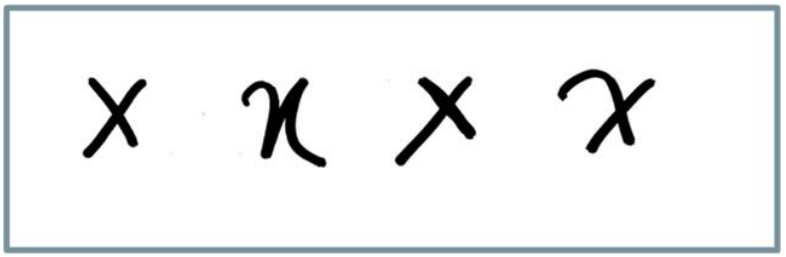

FIGURE 1.5: Ambiguities present in handwritten math symbol

1.6 Structure of Project Report

The rest of the paper is organized as follows: **Chapter-2** represents Literature Review, **Chapter-3** represents the Proposed Methodology , **Chapter-4** discusses the Result and Discussion,finally **Chapter-5** Concludes the report with a mention of the future scope of the present work.

Chapter 2

Literature review

Wenjun Ke et al. (2024) :

The paper titled "Attention Guidance Mechanism for Handwritten Mathematical Expression Recognition" explores the complex challenges associated with Handwritten Mathematical Expression Recognition (HMER), particularly focusing on the difficulties that arise in converting handwritten mathematical expressions into digital text. This task is particularly challenging due to the intricate and varied layouts of mathematical expressions, which often lead to problems such as over-parsing and under-parsing. Over-parsing occurs when the model divides the expression into too many small components, resulting in an overly fragmented interpretation. Under-parsing, on the other hand, happens when the model fails to break down the expression adequately, missing important elements and thus leading to incomplete recognition. Previous HMER methods have tried to improve the attention mechanism by leveraging historical alignment information—essentially using past context to inform the current parsing process. While this approach helps to some extent, it falls short in addressing under-parsing effectively. The main limitation is that it cannot correct incorrect attention focused on parts of the image that are supposed to be processed in later steps of decoding. This faulty attention leads to the incorporation of future context into the current decoding step, which confuses the alignment process and ultimately degrades the model's performance. To tackle these issues, the authors propose an innovative attention guidance mechanism. This mechanism is designed to explicitly suppress attention weights in areas that are irrelevant at the current step and enhance attention weights in areas that are relevant. By doing so, it ensures that the attention module focuses only on the appropriate parts of the image at each step, thereby preventing the inclusion of unintended context. Depending on the type of guidance applied to the attention

mechanism, the authors developed two complementary approaches:Self-Guidance: This approach coordinates the attention of multiple heads within the attention mechanism.Different attention heads are dedicated to distinct areas of the image while self-attention produce diverse focused attention which is highly necessary for more accuracy of the nets. Neighbor-Guidance: This strategy incorporates attention from the previous and subsequent time point. Recognizing the fact that the distribution of attention from the previous and the subsequent steps is guided by the neighbour, the neighbour-generating mechanism serves the purpose of ensuring a coherent and continuity of attention over the relevant parts of the expression that will enhance the overall alignment and recognition of the expression. Experiments were carried out to test the authors' proposed method, on the datasets considered from CROHME year 2014, 2016 and 2019. The performance was very high; indeed, expression recognition rates of 60 were given by the new method presented. The results were impressive, with the new method achieving expressionrecognition rates of 60.75% for CROHME 2014, 61.81% for CROHME 2016, and 63.30% for CROHME 2019 These results outperform significantly existing state of the art approaches and clearly confirmed the relevance of the attention guidance mechanism.

This paper summarizes a new and efficient method about Handwritten Mathematical Expression Recognition with an attention guidance approach. Here, the relative attention weights are learnt and tuned using self-guidance and neighbor-guidance, which helps to overcome the shortcomings of the prior methods especially for under-parsing. Good enhancements on separate quantitative benchmark results suggest that the usage of this approach can benefit the enhancement of the HMER field with more refined handwritten mathematical expression recognition.

Tang, Jia-Man, et al. (2024) :

The paper titled "Offline Handwritten Mathematical Expression Recognition with Graph Encoder and Transformer Decoder" delves into the persistent challenges of Handwritten Mathematical Expression Recognition (HMER). Despite notable advancements in recent years, driven by deep learning techniques, HMER continues to be difficult due to the complex spatial structures of mathematical expressions and the variability in individual writing styles. Encoder-decoder models with attention mechanisms have significantly improved HMER accuracy by framing the problem as an image-to-sequence generation task, where handwritten expressions are converted into LaTeXcode. However, these models often suffer from low interpretability because they do not explicitly segment the symbols within the

expressions. Explicit segmentation is crucial for post-processing tasks and for enabling human interaction with the system, making it easier to correct errors and understand the recognition process. To tackle these challenges, the authors propose a novel approach called Graph-Encoder-Transformer-Decoder (GETD). This innovative method reformulates the mathematical expression as a graph, thereby enhancing the interpretability and accuracy of the recognition process. The approach involves several key steps:Symbol Detection and Graph Construction: The first step is to detect candidate symbols in the input image using an object detector. These detected symbols are then represented as nodes in a graph, referred to as the symbol graph. The edges of this graph encode the relationships between the symbols, capturing the spatial and structural information of the mathematical expression. Graph Neural Network (GNN) for Spatial Information Aggregation: The spatial information of the symbols and their relationships is aggregated using a Graph Neural Network (GNN). The GNN does forward pass over the symbol graph in order to capture the inherent interaction and dependency patterns in the data. Transformer-Based Decoder for Symbol Identification: These then are decoded using a Transformer-based decoder in a manner that locates the conversion from the graph representation and the relations to the sequence of symbols that comprises the emitted symbols. This decoder quantises the symbols to their classes then it gives the layout of the entire expression and the output is processed in LaTeX. In order to validate the GETD model that has been proposed in this paper, the authors have conducted several experiments on open-source databases. It is evidenced by elements of concern that GETD permits near-optimal work in terms of expression recognition accuracy, and does not worsen most of the previous approaches in the state of the art. Furthermore, identification and using symbols as list members for the partitioning in Graph representation also help in developing detailed interpretations that unveil how the recognition is conducted and enhanced one's handling of the process.

Hence, for the offline handwritten mathematical expression recognition, a Graph-Encoder-Transformer-Decoder is developed in this paper. The previous methods are strengthened by GETD model, which makes use of graph representations, GNN and transformer based de coders that can bring solutions for the problems existent and give a more precise and interpretable model for recognizing handwritten mathematics expressions. For this reason, the differential performance of the enhanced script is revealed with benchmark data sets' comparative results suggesting the applicability of the approach for the advancement of the HMER field and its usage in educative spheres and professions.

Dhruv Sharma et al. (2023) :

The paper entitled "9 - 23 - Evolution of Visual Data Captioning Methods, Datasets, and Evaluation Metrics: A Comprehensive Survey" provides an extensive examination of Automatic Visual Captioning (AVC). AVC involves generating syntactically and semantically correct sentences that describe important objects, attributes, and their relationships within visual data. This field is categorized into two main areas: image captioning and video captioning, each with its own unique challenges and applications. AVC is widely used in numerous fields, including assistance for the visually impaired, enhancing human-robot interactions, improving video surveillance systems, and facilitating better scene understanding. The unprecedented success of deep learning in Computer Vision and Natural Language Processing has significantly advanced research in AVC over the past few years. This survey classifies the state-of-the-art AVC methods based on their conceptual approach to the captioning problem.It separates basics which usually employs retrieval or templated descriptions from a more contemporary deep learning paradigms that harness the strong features of neural networks. The paper also includes a method comparison section in which the author compares the advantages and the disadvantages of each of the methods used. The review also describes social consequences of these methods – one of the aspects that might be quantified by the number of citations of the corresponding papers. Furthermore, the paper discusses the architectures applied in these methods, the datasets on which the methods were tested, and where the readers may find the implementations of these methods on GitHub. Also, the survey provides a clear picture of the benchmark datasets available for image as well as for video captioning. These datasets are very important for training and testing of AVC models. The paper also presents the different evaluation metrics which have been established in an effort to rate the quality of the captions that are produced by the machines. These measures help to make the captions that are produced by AVC systems not only correct but also appropriate and useful. The authors note that more recently, research has shifted to the generation of dense or 'paragraphs' of text and Change Image Captioning (CIC). These areas are particularly interesting because they allow the creation of descriptions which are finer grain and closer to human level of abstraction. While dense captioning comes up with more than one sentence for the image, they are all different capturing different details of the image. While the Change Image Captioning which requires the generation of descriptions that points to changes or differences in the series of images, which is helpful for applications such as Surveillance and Monitoring. Last, the paper considers the further development of the

topic in the sphere of automatic visual caption generation. It suggests that the future advancement of even further sophisticated deep learning methods of attention mechanisms, transformers, multimodal learning and so on will further improve the AVC system development with higher accuracy in the future. However, there is more opportunity to develop AVC with the new technologies such as augmented reality and the other advanced technologies of Robotics to create more smart and creative systems.

Thus, it will provide the systematic and comprehensive overview of the literature regarding the development of the visual data captioning techniques. It starts from the traditional methods and covers up to the current Deep Learning methods for this job; it discusses the pros and cons of them, the social implications of the methods used, the datasets, and the assessment metrics in this field. Therefore, it is possible to regard this survey as the source of information about the state and further development of the AVC technology.

Mirkazemy, Abolfazl et al. (2023) :

The paper with the title "Mathematical Expression Recognition Using a New Deep Neural Model" present a revolutionary method of Mathematical Expression Recognition (MER). The authors have come up with a very rich deep neural model that involves the encoder-decoder transformer that has been integrated with several other pre-and post-processing features. This model is meant to identify images of mathematical formulas and convert them into LaTeX– a highly structured language that is used widely when it comes to writing mathematical texts. As for the two authors' novel idea, they have a pre-processing module to put random margins around the formula images based on the domain knowledge. This step is very important as it produces faster version of the feature maps and keeps all the encoder neurons active during learning which in turn enhances the performance of Neural Nets. Finally, it should be noted the post-processing module that the authors have proposed is also notable. It employs a sliding window approach to obtain position based information from the feature maps. With these additional data, the process of recognition contributes a lot to the enhancement of accuracy. As for the recurrent decoder module, it is intended for connecting the feature maps to the position-based information. It uses a soft attention mechanism; this enables it to better translate the context of the formula into LaTeX. Another positive aspect is that this attention mechanism provides the model's focus on the necessary regions of the input image, which improves the result in terms of accuracy. In addition, the authors have proposed a new RL module in this paper.

This module generate the final output and is involved in the feedback process of the earlier step of the decoder. It allows correcting the mistakes during the further calculations and improving the results of the model as a whole. For the experiment purpose, the proposed model was implemented on the im2LaTeX-100k benchmark dataset. It was found that both pre-processing and post-processing modules and, the RL refinement model positively affected the performance of the model. Furthermore, the proposed model was found to have tested higher accuracy with lesser error percentage than the current advanced methods of MER.

Thus, the current paper presents a brief of an effective and comprehensive model of the process of MER. Based around the deep neural model, four new modules are integrated, which include the novel pre-processing and post processing in addition to the reinforcement learning which enhances the DNM admitting a novel benchmark standard on the recognition in ME. By the means of these experimental modules, the sight is to improve the performance of the model in expressing the high complexity and variations of the handwritten and/or printed mathematical notations. Likewise, the use of reinforcement learning assistance helps in optimizing the repetitive feedback to fine-tune on the outlook of the model as a way of increasing its accuracy. The above approach apart from achieving the state of the art accuracy offers a scalable solution for many extensions in various fields of mathematics as well as notational systems . The presented methodology can be a subject of future work when it is applied to other types of symbolic recognition tasks which can alter the above named areas including automated assessment, electronic heritage, and live learning aids.

Li, Zhe et al. (2023) :

In the paper entitled "Improving Handwritten Mathematical Expression Recognition via Similar Symbol Distinguishing," the authors tackle a critical challenge within the Optical Character Recognition (OCR) community: most of which consist of detecting and recognizing handwritten mathematical expressions (HMER). This task is crucial for converting handwritten mathematical content into digital formats, and it comprises two primary sub-tasks: There are two types of skills which are likely to be used in the course of using the application these are the symbol recognition as well as the structure parsing. Contemporary considerations about HMER tend to be an issue of sequence prediction where LaTeXis given, to be accompanied by separate or joint decodage of the individual constituent signs of the MEs and the structure of the LaTeXsequences. However, existing deep learning-based HMER methods also have several challenges, even though the DL

methods show high accuracy in public benchmarks. A problem is over-sensitivity and misinterpretation of symbols that are similar in appearance, thus restricting this method's applicability to other more complex and real-world contexts. In their paper, the authors propose a comprehensive solution to this problem, addressing it from three distinct aspects:In their paper, the authors propose a comprehensive solution to this problem, addressing it from three distinct aspects: 1) Enhanced Feature Extraction: Hence, the authors enhance the disparities in the feature extraction by proposing path signature features. These are intended for conversion into both the local writing particulars and the global spatial information of symbol hand-writing. The dual focus also helps to improve the distinction between symbols that are quite similar visually, within the model. 2) Contextual Language Model: Here the authors expand on what can be called contextual symbols, which are based on the context of what the simple visual recognition models fail to identify correctly. Incorporating context thus allows the language model to take a more accurate approach when it comes to making the predictions, given that errors arising from similar visuals can be misleading. 3) Dynamic Time Warping (DTW) Algorithm: For the misalignment problems which exist in the currently most used ensemble techniques, the authors present a new DTW based algorithm. For this reason, the presence of this algorithm enhance the comparison of sequences in the sense the symbols are map to the positions correctly in the expression. It also helps in increasing over all the efficiency and accuracy of the recognition procedure. In this way, authors' association of the three enhancements similarly achieves the perfectly high level of corresponding with three of the seven CROHME benchmarks sampled in Competition on Recognition of Online Handwritten Mathematical Expressions. They reported that compared to the earlier related work, they achieved higher accuracies with the aid of which they noticed a significant improvement in the HMER problem.

Therefore, it can be stated that the present paper, to a great extent and, in many-sided manner, outlines a rather sophisticated vision of the potential increase in the effectiveness of the HMER system. Applying the better extraction of the features, the proper context for the identification of the signs and sequences alignment, the authors offer somewhat optimal solution to the problems existing in this field.

Basavaraj Anami et al. (2023) :

The paper titled "An Optimized Neural Network-Based Character Recognition and Relation Finding for Mathematical Expression Images" tackles the complex problem of recognizing Mathematical Expressions (MEs), a task with wide-ranging

practical applications. The challenge stems from the variety in writing styles and forms of MEs, which can complicate the recognition process.Most existing models perform poorly because of noise in the inputs; the images are distorted, and this affects the likelihood of character prediction. To overcome these hurdles, the authors present a Chimp-based Spiking Neural Recognition (CbSNR) that is proposed exclusively for the characters' recognition in handwritten and printed ME images. This is followed by pre-processing the input image datasets with a view of reducing on noise and improve on images recognized by the recognition system. After that, features are extracted from these pre-processed images using tracking function of the Chimp algorithm; the Chimp algorithm tracking function is critical at the feature extraction stage. These images are then separated into individual characters and then recognised in the recognition phase. The position updating function of the Chimp algorithm is also used to predict the position relation between the recognized characters so as to enable it to interpret the entire mathematical expression in a proper manner. The method in the present work is developed and applied using the programming language 'Python' to measure its reliability based on F-score, recall, accuracy, and error rate, precision. These metrics gives a clear account of the performance of the model. To ensure there is a basis for determining the gains of the CbSNR framework, the outcomes are compared to the outcomes of prior studies. The evaluation indicates that the methods give in the proposed CbSNR framework yield the highest character recognition rate than the other methods. This superior performance clearly underpins the technique that is involved in the Chimp-based approach when it comes to addressing the problems that arise when trying to recognize expressions in mathematics.

All in all, the current paper outlines a novel and efficient solution to address ME identification. The prescriptive Chimp-based Spiking Neural Recognition framework hikes character recognition ratio and detailed studies show that the proposed work surpasses existing techniques.

Ch Premanvitha et al. (2022) :

The paper entitled "IMAGE CAPTION GENERATOR USING DEEP LEARNING - Convolutional Neural Network, Recurrent Neural Network, (Bilingual Evaluation Understudy) BLEU score, Long Short Time Memory" delves into the critical role that computer vision plays in our society, with applications spanning numerous fields. This paper specifically focuses on the facet of visual recognition known as image captioning. While the generation of language descriptions for visual data has a long history in the context of videos, there has been a recent

shift towards generating natural text descriptions for still images. This shift has been largely enabled by significant advancements in object detection technology, which have made the task of scene description in images more accessible and accurate. The project described in this paper aimed to harness these advancements by training convolutional neural networks (CNNs) with hundreds of hyperparameters and applying them to a vast dataset of images. For this purpose, architectures such as ResNet and VGG were used for this task since they are used extensively in image classification tasks. The essence of the project was to integrate the output of the aforementioned CNN based image classifiers with a recurrent neural network (RNN) to make captions for the images. The integration of CNNs and RNNs leverages the strengths of both types of neural networks: CNNs are good at putting out spatial features from the images while RNNs and LSTM units are good at putting out coherent sequences such as text. The paper also shows the specific details of how the CNN extracts the features of the image and how the RNN then takes the output and generates a sequence of words that will comprise of the caption. Over the generated captions, the BLEU score is used to evaluate the performance since it compares it with the reference caption.

All in all, this paper offers one of the most comprehensive investigations of an image captioning system based on deep learning. Using highly developed CNN architectures like ResNet and VGG trained with large volumes of image data accompanied by the textual descriptions of the images generated by RNNs the authors reveal how a highly comprehensive approach is developed to accurately and descriptively comment on visuals. The detailed description of the model components and the assessment make it possible to appreciate the advancements and show the prospects and difficulties of the development of the automatic image captioning that can be rather inspiring for further researches.

Bian, Xiaohang et al. (2022) :

In the paper entitled "Handwritten Mathematical Expression Recognition via Attention Aggregation-Based Bi-Directional Mutual Learning," the authors take a problem of converting the pictures of handwritten mathematical expressions to LaTeXsequences automatically. At present, encoder-decoder models based on attention mechanisms are among the most used for this purpose. Such models are generally encouraged to produce target sequences from left-to-right (L2R) rather than right-to-left (R2L) which indeed are lost in these models. In order to overcome this drawback, the authors earlier proposed an elegant strategy known as the Attention Aggregation-Based Bi-Directional Mutual Learning Network or ABM

for short. It comprises an encoder component common to both of them and two different inverse decoders, L2R and R2L. The two decoders are improved through the procedure called mutual distillation in which there is one to one transmission of knowledge at each step of training. This process fully utilize the first order information based on the two inverse directions so as to enhance the performance of the model. Also, to rightly treat symbols of different scales, the authors incorporate an Attention Aggregation Module (AAM). This module comprises the multi-scale coverage attentions so that the model will understand symbols of any size to attend to. In the inference stage, the model that has embraced the L2R and R2L training separately, then it employs merely the L2R branch. This approach retains the original quantity of parameters and the speed of making inferences, which is desirable. In particular, numerous experiments confirm the efficiency of the discussed approach. The ABM network achieves recognition accuracies of 56.85% on the CROHME 2014 dataset, 52.92% on the CROHME 2016 dataset, and 53.96% on the CROHME 2019 dataset without the need for data augmentation or model ensembling. These results substantially outperform state-of-the-art methods, highlighting the potential of this innovative approach in the field of handwritten mathematical expression recognition.

All in all, this paper introduces a new and efficient approach for recognizing Handwritten Mathematical Expressions. At the same time, the unique proposed ABM network includes bi-directional mutual learning and attention aggregation, which places the proposed network at the top levels of recognition accuracy in the field.

Bhalekar et al. (2022) :

The paper entitled "The New Dataset MITWPU-1K for Image Caption and Object Recognition Tasks" has discussed about the new and creative dataset. In the image captioning area, scholars employ many pretrained datasets like MS COCO, Flickr, Pascal VOC and so on for the modeling that is designed to describe image content automatically. Nevertheless, these existing datasets have a major drawback: no textual description inside the image that could significantly enhance the accuracy of the produced descriptions. In order to fill this gap, the authors provide a new MITWPU-1K dataset that comprises image, text and their captions. Most of the picture were captured around MIT World Peace University (MITWPU) campus in India and therefore the data this set offers is rich in both context and content. This dataset is designed to be used in both object detection and image captioning so it will be useful for researchers in both of those fields. This paper describes the careful procedure of building this new dataset in detail. It started

with the analysis of what are currently available in terms of dataset models and what the state-of-the-art in this area is. This review made a positive contribution to the generation of the MITWPU-1K dataset which was develop to overcome the problems that were encountered in previous set by incorporating textual information within the images. Besides the methodology of creating a new dataset, the paper also puts forward a sequence convolutional method aiming at the identification of objects in the new MITWPU-1K dataset. This model enhances the object detection precision through the usage of the enhanced data information obtained from the new dataset. An effective analysis of the structuring and the creation of the dataset, which involved the authors is presented with mention of the major challenges and lessons learnt. They take care to note that the textual data should be integrated into the image datasets and show specifically how this can cause the generated descriptions to be significantly more accurate.

In conclusion, this paper is guide, in which all stages are described to obtain a new dataset for the tasks of object recognition and image captioning. When working on this very paper, the-presented work underlines the significance of the textual information to improve the degree of description The paper also reveals multiple considerations while constructing the MITWPU-1K dataset.

Yibo Zhang et al. (2022) :

The authors, in the paper titled "Combining CNN and Transformer as Encoder to Improve End-to-End Handwritten Mathematical Expression Recognition Accuracy," explore a rapidly developing subfield of AED models for improving HMER. Taking into consideration the recent achievements of Transformers in computer vision and the several attempts to combine Transformers with Convolutional Neural Networks (CNNs), this paper presents three novel approaches to enhance the performance of AED-based HMER models using these technologies. The first of which is called the Tandem way, which it makes use of a CNN for feature extraction, after which the sequence moves to the next step. These features are then passed to a Transformer encoder layer, which captures global dependencies, thus enhancing the feature representation by means of both local and global information. The second method, known as the "Parallel way," turns an imagery input into raw image patches and feeds them to a Transformer encoder branch. The final feature representation is then created by concatenating the output of this branch with the output from the CNN as both CNN and Transformer are beneficial in parallel. The third one, called Mixing way, is totally different from previous ways

since it integrates MHSA with CNNs in a more direct way; it replaces the convolutional layer in the final stage of the CNN with MHSA. These three methods were assessed and compared with the help of evaluation on the CROHME benchmark dataset and based on it we utilised the 2016 and 2019 datasets for evaluation. The Tandem method demonstrated expression recognition rates (ExpRate) of 54.85% on CROHME 2016 and 58.56% on CROHME2019. The Parallel method achieved ExpRates of 55.63% on CROHME 2016 and 57.39% on CROHME 2019. Meanwhil the Mixing method obtained ExpRates of 53.93% on CROHME 2016 and 55.64% on CROHME 2019. These results provide a comprehensive comparison of the effectiveness of each method.

To sum up, the author shows that the Tandem and Parallel are superior to the Mixing whereas the tested Tandem and Parallel give comparable performance. From this, the researchers can deduce that the integration of the CNN and the Transformer design, especially through the Tandem and the Parallel approaches, improves the performance of AED-based HMER models efficiently. Therefore, these results demonstrate a combination of CNN and Transformer architectures can significantly improve the accuracy of handwritten mathematical expressions with more trends for the next development of this area. Future studies may also look into the possibility of designing other integration technique to enhance the utilisation of these integrated models. In the same vein, the proposed methods could benefit from experiments with a larger variety of data quantities so as to explore further their applicability and the degree of their generality. Further enhancements can be made exploring other configurations of hyperparameters and modifications of the network architecture. Interdisciplinary connections may reveal new uses and approaches which were not earlier considered in the field. The challenges of making the models efficient and deployable for practical purposes will also be important in realizable practical application.

Khanh-Ngoc et al. (2021) :

The paper entitled "Handwritten Mathematical Expression Recognition: There is titled, 'An Approach on Data Augmentation' describes a new technique that can be used for the improvement of dataset that recognizes the handwritten MEs. The authors suggest that in order to produce ME images from the available CROHME dataset, action will be taken at two levels. The first proposed method is based on the transformation of the ME images acquired with the CROHME dataset through geometric transformations. The second method creates new ME images on the basis of the ME images which are derived from dictionary of character patterns

obtained from the CROHME dataset. These newly generated images adhere to the rules of mathematical notation, thus expanding the dataset significantly. The main contribution of this paper is the introduction of a much larger dataset for the handwritten mathematical expression recognition problem compared to the original CROHME dataset. To evaluate the effectiveness of this augmented dataset, the authors employ a sequential system comprising two modules: the Single Shot MultiBox Detector (SSD) for object detection and DRACULAE for parsing SSD's output into LaTeXstrings. Their focus is on improving the performance of the detector. They trained and evaluated the system on the CROHME 2013 training set, both with and without their generated dataset, to assess the impact of their data augmentation approach. The experimental results indicate that the detector achieves a mean Average Precision (mAP) of 52.57% when using the augmented dataset, compared to 36.98% without it. This significant improvement demonstrates the efficacy of their generative approach in enhancing the performance of handwritten mathematical expression recognition systems.

In conclusion, this paper details a successful strategy for augmenting the dataset used in handwritten mathematical expression recognition, highlighting the substantial benefits of incorporating geometrically transformed and newly generated ME images into the training process.

Zhelezniakov et al. (2021) :

The paper entitled "Online Handwritten Mathematical Expression Recognition and Applications: A Survey" explores the important role of handwritten mathematical expressions in various domains like education, engineering, and science. The widespread availability of powerful touch-screen devices, coupled with the recent advances in deep neural networks as high-quality sequence recognition models, has led to the broad adoption of online recognition of handwritten mathematical expressions. The paper emphasizes the need for a deeper understanding and improvement of these technologies to address the challenges posed by the extensive use of distance learning and remote work, especially due to the global pandemic. This survey delineates the state-of-the-art recognition methods and examines the user experience in pen-centric applications for handling handwritten mathematical expressions. The recognition methods are categorized into different classes, with detailed descriptions of their merits and limitations. Special attention is given to end-to-end approaches based on encoder-decoder architecture and multi-modal input. The paper also reviews evaluation protocols and open benchmark datasets, providing a comparison of recognition performance based on results from open

competitions. Also, the paper provides examples of how handwritten mathematical expression recognition can be applied across quite a number of fields and on different platforms. The specificity of this survey is its focus on the interaction between the UI design and the working recognition approaches, so as to facilitate potential researchers optimize it for application. Last but not the least, this paper proposes a prospective on the future possible work in the field of HMSER and its applications as there is still considerable scope for development and growth.

In conclusion, it is possible to point out that the present paper gives a rather vast and profound idea of the state of current progress and prospective developments in the field of online handwritten mathematical expression recognition, their technologies, and applications.

Chi, Xueke et al. (2021) :

In the paper titled "Handwritten Mathematical Expression Recognition with Self-Attention" the authors discuss the progress in attention based encoder decoder models used for recognising handwritten mathematical expressions in the past years. However, the best models become a target of a critical problem referred to as attention drift very often. This issue arises because the local attention mechanisms, which are often based on the Recurrent Neural Networks (RNNs), can get easily overwhelmed because the encoding features are often similar. In other words, the model does not sustain precise attention when it is in processing sequences; hence, there are problems with recognition. Towards this end, the authors put forward a new encoder-decoder model that embodies the self-attention mechanisms in their work. Different from the classic pattern that depends much on the local focus, this creative approach collects the global information from the feature map. This global information is then, fused with local information derived from CNNs to serve as two different sets of features. Having both global and local information present during the recognition of the signal allows the model to have a more precise focus during the entire recognition process and, therefore, will not be a victim of attention drift. The authors rigorously tested their proposed model on two widely recognized datasets: those of the CROHME2014 and CROHME2016 competition in reading handwritten mathematical expressions. These are standard datasets in the field of HMER, supplying a strong basis to further assess the efficiency of the novel models. The results of the experiments performed during the course of the work were quite encouraging. The proposed model has had recognition accuracies of only with the official training data and without including the body posture data set. The former reached up to 98.51% in the CROHME2014 and the latter

50.74% in the CROHME2016. These results can be seen to be much better than those presented in the benchmark studies, proving the efficiency of the module for learning self-attention employed in the encoder-decoder model.

The significant increase of the recognition accuracy confirms that incorporating self-attention into the encoder-decoder architecture may be considered as one of the most effective ways to address the challenges related to the utilization of conventional types of attention. This way the model is able to keep global contextual information in addition to the local features and thus improving the model's likelihood to distinguish handwritten mathematical expressions accurately. Such a development entails a major advance in the area to make a way for enhancing the level of development of technology that is used in the recognition of handwriting tests especially where the tests involve mathematical expressions. That the analyzed approach positively affected the model performance on standard datasets hints at its broad applicability since accurate interpretation of the given type of handwritten notation can be crucial for a vast number of applications.

Herdade, Simao et al. (2019):

In the paper entitled "Image Captioning: Transforming Objects into Words" by Herdade, Simao, et al. (2019), the authors explore advancements in the field of image captioning. They discuss how models in this domain typically follow an encoder-decoder architecture. Within this framework, the encoder is responsible for processing abstract feature vectors that are extracted from images, serving as the input for the subsequent stages.A particularly effective approach highlighted in their work involves using feature vectors derived from region proposals identified by an object detector. However, Herdade et al. extend this method further by proposing a new model called the Object Relation Transformer. This adds to the conventional method the mechanism termed geometric attention that makes the spatial relationships of the detected objects explicit. In other words, the feature extraction of their model means not only the detection of objects in the image but also their position relative to each other. To validate the presented model the researchers performed experiments that were aimed at investigating the model's performance in terms of producing accurate and descriptive captions. The Object Relation Transformer was tested on the MS-COCO dataset which is widespread in the field of image captioning. The results of their experiments were significant. The model yields higher accuracy rates than the existing methods in all the standard measures that are usually employed. This superior performance

shows that spatial relations should be integrated whenever possible, as done by geometric attention, to boost the quality of image captions.

Therefore, the enhancement of the spatial relationship between objects as suggested by the Object Relation Transformer can be regarded as a giant leap for the field of image captioning. The study of Herdade, Simao, et al. (2019) thus shows that the caption generation task benefits from using the objects and their relative positions, and contributes to the development of the state of the art in this field.

Mondal et al. (2019) :

In the paper entitled "Textual Description for Mathematical Equations," the authors tackle the challenging task of interpreting mathematical expressions or equations from document images. This problem is particularly difficult due to the vast variability in mathematical symbols and expressions, which makes consistent and accurate recognition a complex endeavor. In turn the authors offer a new method that can be viewed as a subproblem of generating textual descriptions of the internal semantics of such equations. In line with the tradition of the natural image captioning problem in computer vision, they introduce the Mathematical Equation Description (MED), a model that is entirely new to the present problem. MED is a fresh deep neural network that is fully end-to-end learnable for producing textual description for reading mathematical equation images. It consists of two main components: The proposed model is a CNN and a RNN where the latter includes an attention mechanism. The CNN is inherited as an encoder which was with the intention of capturing the most basic features of the input images in the form of mathematical equations. These extracted features are then fed into the RNN where with the help of an attention function generates natural language descriptions of the input images. The kind of design used can allow the model to focus on the various parts of the picture to generate the text, as people do when reading equations. Since there is presently no data of image of mathematical equations and text of the same equations available the authors had to compile two new sets for these experiments. These datasets were an essential when was buying and testing the MED model. On the same note, the authors took it a notched further to conduct a real world experiment for the very purposes of effecting hypothesis. In this experiment students were to write the equations correctly without looking at the raw equations which were in fact used to train the MED model but only according to the text descriptions generated by the med model. This was rather promising and semi-structured as students were able to write most of the equations properly with little regard to the figures.

Therefore, within the framework of the paper, they introduced a new method for reading MEIs in the documents' images through textual descriptions supported by deep learning models. This is good because having CNN and RNN elements in the core make up of the MED model as described above is a sound method of thinking about and analysing mathematical equations in an interrelated manner. In conclusion, it has been ascertained that implied results of the experiments enlighten the practical usability of the MED model in education as the result, this contribution might be viewed as one of the salient achievements in the field. Apart from this it also assists in mitigation of a major problem, it reveals new avenues for enhancement of presenting and consuming mathematics in different situations.

Liu, Shuang, et al. (2018) :

The paper that was reviewed is titled "Image Captioning Based on Deep Neural Networks" and the paper explores the emerging literature that deals with image captioning, a sub-discipline of computer vision and NLP that has been made possible by the availability of deep learning frameworks. However, in the last few years, image captioning emerged to be a central topic in this domain, that aims at allowing the creation of English textual descriptions by computers that reflect the content of images. This task is inherently ambiguous to some extent because it has to involve not only the recognition of the objects and scenes, but the attributes, states, and neighbour relations of these objects as well. The paper also provides insight about how the generation of Descriptions from images entail higher-level semantics understanding which is not easy. What this implies is that, over and above the identification of objects within an image, the system needs to go further and be able to understand their context and how these objects relate to one another to come up with a meaningful story. Nonetheless, as observed from the various difficulties inherent in image captioning, much has been achieved in the recent past. The authors focus on three prominent methods that leverage deep neural networks for this task: CNN-RNN Based Methods: These are the approaches that use CNNs to extract the image features and RNNs to generate sequences. The CNN captures the spatial hierarchies of an image, while the RNN generates a sequence of words to form a sentence that describes the image. CNN-CNN Based Methods: In these approaches, both image feature extraction and sentence generation are handled by CNNs. These methods aim to streamline the process by utilizing the strengths of CNNs in capturing spatial features for both tasks. Reinforcement Learning-Based Frameworks: These frameworks apply reinforcement learning to improve the quality of the generated captions. By treating the caption generation as a sequential decision-making process, reinforcement

learning optimizes the model based on reward signals that reflect the accuracy and relevance of the captions. The paper also introduces representative works for each of these methods, showcasing the advancements and innovations in the field. Additionally, it discusses the evaluation metrics used to assess the performance of image captioning models, such as BLEU, METEOR, and CIDEr, which measure the similarity between generated captions and ground truth descriptions.

In conclusion, this paper provides an in-depth overview of the current state of image captioning using deep neural networks. It highlights the progress made, the innovative methods developed, and the remaining challenges. By summarizing the benefits and addressing the major obstacles, the paper contributes valuable insights into the continuous efforts to enhance the accuracy and coherence of automated image descriptions. This comprehensive review not only underscores the complexity of image captioning but also celebrates the significant achievements and potential future directions in this exciting field.

Jianshu et al. (2017) :

The paper entitled "A GRU-Based Encoder-Decoder Approach with Attention for Online Handwritten Mathematical Expression Recognition" presents the novel end-to-end approach to recognize the OHMER with the help of encoder-decoder approach with an attention mechanism. Resolving the issue of how to translate handwritten mathematics and the occurrences within a mathematical context accurately can be a rather challenging task, and this is a problem that the study hopes to address. In this approach, in particular the so called two-dimensional ink trajectory information of handwritten expressions, which depicts the path of the pen or stylus, is encoded by means of a recurrent neural network based on a gate recurrent unit (GRU RNN). This calls for a GRU RNN which is very useful when it comes to dealing with sequences and capturing the temporal sequences in this case the movements of the handwriting input. Subsequent to encoding phase, the decoding phase is performed using another GRU RNN which incorporates a coverage based attention model. The emphasis is made on the attention mechanism for this framework because it enables the model to attend to particular regions of the input sequence when generating each symbol of the output. Such selective focus is possible and allows the decoder to handle a great number of details and refinements which are present in handwritten text and mathematical calculations in particular. Ordinarily, when attention is being applied, it might be possible for one or more of these segments to be completely overlooked; the coverage-based

aspect of the attention model takes care of this problem. Using the proposed approach, recognition of symbol, as well as its structural analysis is possible at the same time. This dual capability allows the system to not only recognize the characters but it's spatial form in relation to other mathematical information and to logically arrange and interpret them collectively in order to produce a continuous character sequence in LaTeX. LaTeXis one of the most popular tools for typesetting the mathematical and scientific texts and as far as it is useful to have an ability to produce LaTeXcode out of handwritten input. The foundation of the proposed work was tested in experiments performed on CROHME 2014 competition task which is a benchmark database for HMER. The results were quite encouraging, owing to the fact that the new method developed for the work obtained an expression recognition accuracy of 52.43% which is far greater than the performance of the other methods existing during that period. Sulima and Zamubalov performed an experiment and achieved such a result using only the training dataset provided by the official MCE, which indicates the effectiveness and stability of the approach. In addition, the research also contains the plots of the relations between the input trajectories of the handwriting and the output LaTeXsequences. Such visualizations, made possible by the employment of the attention mechanism, give clear illustrations of how the model handles and decodes the input data to provide its output, and thus gives very practical application and realization of the proposed method.

On the whole, this paper describes a very efficient GRU-based encoder-decoder system with attention for processing digitalised online handwritten mathematical expressions. Introduction of a coverage-based attention model as a component brings in a robust performance of determining the multifaceted handwriting input, writing it with precision in LaTeXformat. The higher accuracy of the method on the CROHME 2014 dataset and the ability to visually analyse the alignments between the input and output provide a basis for the further usage of the technique in learning and working environments where fast and accurate recognition of handwritten mathematical texts is crucial.

Zhang et al. (2017) :

The end-to-end model of online handwritten mathematical expressions described in the paper entitled “A GRU-Based Encoder-Decoder Approach with Attention for Online Handwritten Mathematical Expression Recognition. ” This innovative approach employs the encoder-decoder model that uses the attention mechanism to provide higher accuracy of recognition. This is done by using Gated

Recurrent Unit-based Recurrent Neural Network (GRU-RNN) in encoding the two-dimensional ink trajectory information of handwriting expressions. This encoded information retains the characteristics of the handwritten input and conveys all of its specifics as well as the variability of the traces made by the writing instrument. Subsequent to the encoding phase, the decoding is done with the second GRU-RNN that is endowed with a coverage-based attentiveness model. This attention model has a very important function to emphasize on the decoding process about the parts of the input only. It is a logical strategy to manage the symbolic recognition in parallel with the structural analysis, which in the end, produces the decoding of the input into the sequence of the characters in the LaTeXlanguage. Of all the output formats, this format is most useful since it enables the identified mathematical equations to be embedded onto digital documents for further use. The authors tested their proposed approach on the CROHME task of the CROHME 2014 competition which is a standard benchmark for OHMER. The results were incredible, the method has surpassed the existing methods of the state art and has got the accuracy of expression recognition of 52. 43 using for training only the official dataset. This gives evidence of the efficiency of the proposed GRU-based encoder-decoder framework with attention in the recognition of HMEs. However, the paper also supports the usage of the developed method by visualizing the alignments of the input trajectories of handwritten expressions with the output LaTeXsequences. These are due to the attention mechanism the visualizations give elucidation on how the model decodes and handles the handwritten inputs where it is believed to be capable of selectively paying attention to parts of the input during decoding.

Lastly, it is a detailed comprehensive paper that provides a concise and efficient way at recognizing the online handwritten mathematical expressions through the utilization of the GRU-RNNs and attention models in their sophisticated manner. Applying such advanced techniques, the authors have elaborated the highly effective tool that allows improving the identification of the handwritten mathematical expressions to the extent that can further be a ground for more accurate and faster documentation of the mathematical contents.

Jianshu et al. (2017) :

The paper entitled "Watch, Attend and Parse: The research paper titled "An End-to-End Neural Network-Based Approach to Handwritten Mathematical Expression Recognition" seeks to deal with a very elaborate problem of identifying

HMEs. Algorithm learning is especially challenging because handwriting is not always uniform and clear, and since most mathematical expressions are plane. The authors present a new WAP which learning from WAP, aims at recognising HMEs in two dimensional layout and then converts these to one dimensional character sequences in LaTeXform. In contrast with other approaches, in the WAP model questions are free from problems such as the segmentation of symbols, and there is no need for prior definition of an expression grammar. Instead, it handles symbol recognition and structural analysis using two main components: a watcher and a parser" The SUS model clearly depicts four different activities of a user or a human when operating a system. The watcher operates on HME images and it has been discussed in detail elsewhere through a convolutional neural network (CNN) encoder. Outputs of the parser are the LaTeXsequences being generated by an RNN decoder with the aid of an attention mechanism. The attention mechanism contributes significantly to the auto-learning process of the correspondence between the input expressions and the LaTeXsequence of the output. The applicability of the proposed approach is tested on the dataset of CROHME international competition. On the CROHME 2014 dataset, the expression recognition accuracy of the WAP model was found as 46.55%. C-SM has median accuracy of 44.55% on the IAM Handwriting database and on the CROHME 2016 database. These results are higher than state of the art methods indicating that the proposed WAP approach is efficient and accurate.

All in all, the current work presents a new approach for recognizing handwritten mathematical expressions that is free from the challenges of this activity based on the deep learning methods. It also proposes interesting enhancements regarding the WAP model where most obvious limitations are also bypassed and where they achieve high performances on benchmarks data sets. Therefore, when working in connection with the attention mechanisms, combined with efficient sequential record, the WAP model can rather effectively interpret various notations of the mathematics. It has been observed across the section that no matter what kind of data is utilized, this method remains very steady and hence deputies how universally applicable this method must certainly be. In addition, the structure of the model also allows the model to work faster. Therefore, the model may be readily applied to real-time issues. As for the further research, one can then continue with the integration of the WAP model with other other kinds of sophisticated techniques to fine-tune the observed rates of the performance. Besides, the establishment of a new generation of more general and diverse data sources will be the essential condition for the further development of this field. These are the grounds

for the future practices in EdTech, intelligent assessing instruments and the future of designs and creations of digital materials that will smoothen the future AI in the symbolic representations and their understandings.

Dai Nguyen et al. (2016) :

In the paper entitled "Recognition of Online Handwritten Math Symbols Using Deep Neural Networks," the authors present an innovative approach that leverages deep learning to recognize online handwritten mathematical symbols. Recent advancements in various deep learning architectures, such as Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) RNNs, have revolutionized fields like computer vision, speech recognition, and natural language processing, consistently outperforming state-of-the-art methods across a range of tasks. In this study, the authors specifically employ max-out-based CNNs and Bidirectional LSTM (BLSTM) networks. The max-out-based CNNs are applied to image patterns created from online patterns, while the BLSTM networks are applied to the original online patterns. By combining these two approaches, the authors aim to enhance the recognition accuracy of handwritten mathematical symbols. The proposed methods are rigorously compared with traditional recognition methods, specifically Markov Random Fields (MRFs) and Modified Quadratic Discriminant Functions (MQDFs). These comparisons are conducted through recognition experiments on the CROHME (Competition on Recognition of Online Handwritten Mathematical Expressions) database. The paper includes detailed analysis and explanation of the results, highlighting the superior performance of the deep learning approaches over the traditional methods.

Overall, this paper demonstrates the potential of deep learning architectures, particularly CNNs and BLSTMs, in significantly improving the recognition of online handwritten mathematical symbols. This advancement not only sets a new benchmark in the field but also paves the way for further research and development in the application of deep learning to handwritten mathematical expression recognition.

Giovanni Yoko et al. (2014) :

In the paper 'Extracting Textual Descriptions of Mathematical Expressions in Scientific Papers', the authors explore the importance of mathematics within different branches of science. They put forward that by connecting papers with the help of mathematical expressions, the topography of scholarship and the understanding of the information provided can be expanded and improved. This research work

has the following broad objectives: To determine a complete guide on how to annotate natural language descriptions of mathematical expressions and to define a complete guide on how to detect natural language descriptions of mathematical expressions. This, in turn, enhances the semantic content of scientific publications and contributes to better connectedness of papers. The researchers start their process of labeling and evaluating the descriptions of the mathematical expressions by being very thorough and selective in terms of the types of textual spans that they choose to employ. These spans are fixed context windows, appositions, minimal noun phrases and noun phrases. After this manual annotation phase is over, the researchers proceed to creating a procedure for automatically generating such descriptions. They cast this extraction problem as a more specific form of binary classification. In this task, for each mathematical expression, there are possible description candidates, and they are divide into correct and incorrect. To this end, the researchers use Support Vector Machines (SVMs) with diverse features to accomplish this classification. From the experimental outcomes they found that the best model is SVM which uses all the noun phrases. This model reaches an F1-score of 62.25% that is so good in light of the fact that a portion of the information in the set of tests consisted of text, which is notably higher than the baseline method that obtains 41.47% of F1-score using the nearest noun. This significant improvement speaks of the efficiency of the strategy they proposed for identifying and labelling NL descriptions of MEs in scientific texts.

All in all, this paper demonstrates the possibility to utilize the state-of-art machine learning algorithms in order to improve the semantic capabilities of scientific papers and provide exact correspondences between the mathematical formulas and textual explanations. The approaches and procedures suggested by the authors of this paper offer substantive basis for the further study and advancement of this area and improve the quality of the inclusion and utilization of mathematics in the scientific texts.

Lin, Tsung-Yi, et al. (2014) :

The paper entitled "Microsoft COCO: In the paper named "Common Objects in Context" a new dataset is introduced for object recognition within the framework of scene understanding is described. This grand objective is achieved when such typical pictures are purchased so that they depict intricate and accurate scenes which have normal objects in their settings. Concerning the images that emerge from these objects, all are accompanied by per-instance segmentation that can

give detailed location and contains much context. The database might be regarded as rather big: there were ninety-one different kinds of objects in it and all of them can be identified easily; for example, even a child is capable of doing this. In total, the current dataset looked at so far contains 2. Labeled into 5 million instances, spread over 328 thousand images, it is among the largest repositories that researchers can use to investigate object recognition. Microsoft COCO was developed with a lot of assistance from crowd workers who used specific interfaces for category localization, instance recognition or instance annotation. In this way, it was ensured that only high quality and great accuracy labeled instances were obtained. The paper is not only the analysis of the presented MS COCO dataset but also a statistical analysis and comparison of the MS COCO with other known datasets PASCAL, IMAGE, INTEN, SUN. Such comparisons will also enable to highlight relative specificities of the COCO dataset in terms of size, diversification, and contexts to stress the stakes it can provide in comparison with other comparable datasets. Furthermore, applying a Deformable Parts Model the authors give a benchmark performance on the two established measures of bounding box and segmentation detection. It is beneficial, when carrying out future research, as the given set of results can contribute to improving the object detection and segmentation techniques that have been presented. However, regarding the substance of the paper and our interests as researchers intending to expand on this work, the presentation of these baseline analyses is not masking complexity as the feature of novelty; rather, each of these analyses illustrates the value of the dataset, as every given analysis here allows the subsequent research to pose different questions and build frameworks beyond the existing one.

All in all, this paper presents an essential resource for object recognition by stressing the context on understanding scenes and the current framework of a promising dataset for further development of object recognition techniques. Due to its ability of encouraging future work and innovation in object recognition, it can be considered as the basis for current and future work in the field of computer vision.

Awal et al. (2010) :

The paper with the title “The Problem of Handwritten Mathematical Expression Recognition Evaluation,” points to several main concerns with respect to the recognition of handwritten mathematical expressions. They begin by discussing the problem of setting up ground truth for a set of handwritten mathematical expressions. Ground truth entails the establishment of a reference database that

is accurate with which to train and test the recognition systems. This step is significant as the ground truth dataset mainly affects the efficiency and recognitive ability of the developing and testing recognition systems. Once the ground truth dataset is established, the authors move on to the next major issue: for instance comparing the performance of various recognition systems. Benchmarking entails comparison of performance of a variety of systems with a view of identifying efficiency of each in the recognition of handwritten mathematical expressions. In the authors' opinion, to reach this aim, it is necessary to identify concrete performance indicators, describe possibilities for assessing these indicators. All these performance indicators are carefully crafted for the express purpose of depicting the true performance of each recognition system as would be expected. In standardizing these performance measures, the authors mention that the intent is to have a set of measures that are universally acceptable so that evaluation or comparison of various strategies on MER can be done. This way of assessment is standardized to provide the reliability and objectivity for the development of the R and D in this subject area. As these metrics indicate about the systems, the clearer and more consistent roadmap to compare them is advantageous to provoke progress and development into the recognition of handwritten mathematical expressions.

Therefore, this paper aims at providing an insight on the evaluation issues of the HMER systems. However, through developing the presented ground truth database and offering the performance metrics, the authors present a clear foundation for the evaluation and comparison of recognition systems. The outlined approach not only increases the credibility of the assessments of performance but also contributes to the further development of the research in this sphere.

Deng, Jia, et al. (2009) :

The paper entitled "ImageNet: A Large-Scale Hierarchical Image Database" addresses the significant challenge of harnessing and organizing the explosion of image data available on the Internet. This abundance of image data holds the potential to foster the development of more sophisticated and robust models and algorithms for indexing, retrieving, organizing, and interacting with images and multimedia data. However, the question of how to effectively harness and organize such vast amounts of data remains a critical issue. To tackle this problem, the authors introduce "ImageNet," a large-scale ontology of images constructed upon the backbone of the WordNet structure. ImageNet aims to populate the majority of the 80,000 synsets in WordNet with an average of 500-1000 clean, full-resolution images per

synset. This ambitious endeavor will result in tens of millions of annotated images organized according to the semantic hierarchy of WordNet. The paper offers a thorough characterization of the current state of ImageNet, the specifics of which are as follows: 12 subtrees, 5247 synsets, and approximately 3. 2 million images. The authors show proves that the ImageNet is orders of magnitude larger in scale, is way more diverse and is much more accurate than the existing image data sets. The construction of a database at such a large scale is a mammoth task and the authors, therefore, outline their data collection strategy using Amazon Mechanical Turk. Lastly, the paper illustrates the utility of ImageNet through three straightforward applications: Including object recognition, image classification as well as auto partitioning objects. The following applications demonstrate high practical value of ImageNet and generalizable potential of the data source that can stimulate the development of the field of Computer Vision.

Thus, the main idea of this paper is to introduce ImageNet to the computer vision community as a valuable tool that would bring changes to the field, to demonstrate its feasibility in terms of its scale, diversity, and accuracy, to underline its significance for the further investigation.

Chapter 3

Proposed Methodology

3.1 Workflow

We have explained this chapter in five parts. The workflow diagram of our research is given below.

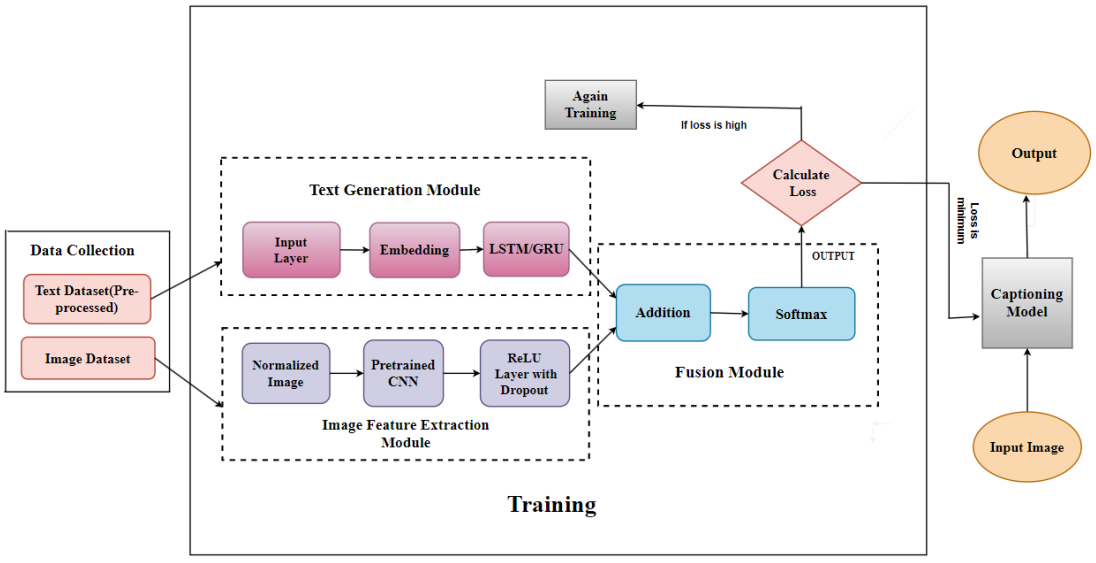

FIGURE 3.1: Workflow Model

3.2 Dataset Creation

In deep learning, the dataset is crucial because it serves as the foundation for training models. A well-curated dataset allows the model to learn patterns and make accurate predictions. If the dataset is diverse and representative of the problem

domain, the model can generalize better to new, unseen data. Conversely, poor-quality or biased datasets can lead to inaccurate or biased models. Essentially, the quality and quantity of the dataset directly impact the performance and reliability of deep learning models.

The dataset creation process is divided into two parts: an image dataset and a text dataset.

3.2.1 Image Dataset

In our study, we have created an annotated dataset specifically tailored to handwritten mathematical expressions to support our research. This initiative was driven by the lack of a suitable offline standard dataset that meets our unique requirements.This process was commenced with the fine process of writing different mathematical expressions on paper. Following this, high resolution pictures of these expressions were captured, where each picture was in the JPG format. The last dataset is actually a set of $2,000$ images handpicked for the purpose of the research and that span various mathematical ideas and concepts. Some of the topics include limits, Integration, Trigonometry, Differential Equations and Polynomial Expressions. All the images are manually pre-processed to meet the uniformity and consistency in the dataset by resizing them to the dimension of $7680 * 4329$. A large and uniform image size is useful especially when training batch by batch and evaluating the model. It is noted that the mathematical expressions are written by black or blue colour ink. The choice of the ink colours was deliberate and this was done to make sure that the contrast was high and visibility was good. This is very important in mathematical image processing and recognition.

The whole set of such images included in the dataset amounts to $2,000$, but their total volume does not exceed $1.6GB$. For this reason, this comprehensive dataset has been very useful when training the model since it was easy for us to deduce the performance of the model. In this manner, we have offered enhancements to our work and our model of using high-quality images that positively impacted the field through the delivery of a significant work in the area of recognition and interpretation of the handwritten mathematical expressions.

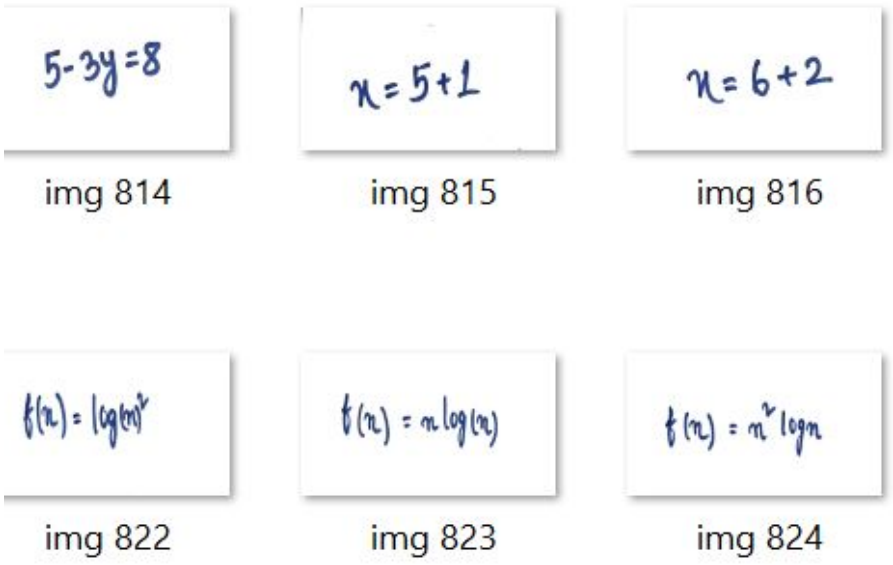

FIGURE 3.2: Image Dataset

3.2.2 Text Dataset

The text dataset is of two types. a) Textual Descriptions of handwritten Mathematical images, b) LaTeXmarkup.

- We have a total of 6,000 textual descriptions corresponding to our 2,000 images. This means that every image is described by three different captions (texts) at the same time. This way, the generation of multiple captions for the same image enriches and diversifies the dataset and offers a efficient language description of every mathematical expression. The size of the caption dataset is 167KB. This well-organized caption dataset is quite important to our study since it allows us to train the model and conduct detailed analysis.

```
img38.jpg: function g of y equals to one.
img38.jpg: this equation represents function g of y equals to one.
img38.jpg: this equation shows function g of y is equal to one.
img39.jpg: a plus b plus c plus d plus e.
img39.jpg: this expression represents a plus b plus c plus d plus e.
img39.jpg: this expression defines addition of a, b, c, d and e.
img40.jpg: a multiplied by c equals to one.
img40.jpg: this expression represents a multiplied by c equals to one.
img40.jpg: the equation shows the multiplication of a and c which becomes one.
img41.jpg: six a square.
img41.jpg: the expression represents six a square.
img41.jpg: the expression denotes square of a multiplied by six.
img42.jpg: y square is equal to x cube divided by two multiplied by a minus x whole.
img42.jpg: this equation represents y square equal to x cube divided by two multiplied
img42.jpg: x cube divided by two multiplied by a minus x whole equals to y square.
```

FIGURE 3.3: Textual Description Dataset

- We have 2000 LaTeXnotations which correspond to the 2000 images in our dataset and each image is associated with one unique LaTeXnotation. For each mathematical expression image, we have a corresponding LaTeXrepresentation. This way, the dataset is useful for tasks that require accurate markup and formatting of mathematical expressions. This is very essential in mathematical image captioning and similar tasks.

```
img502.jpg: \log_x\left(\frac{1}{16}\right) = -2
img503.jpg: y = \log_{3}(x)
img504.jpg: y = \log_{3}(|x|)
img505.jpg: y = 5 - \log_{4}(x)
img506.jpg: y = \log_{0.5}\left(\frac{x}{2}\right)
img507.jpg: \log(x) - \log(x + 6) = -1
img508.jpg: \frac{2}{3} \cdot \log_{7}(z) - 2
img509.jpg: \log_{5}(8) - \log_{5}(t)
img510.jpg: 2 \cdot \log_{2}(x) + 4 \cdot \log_{2}(y)
img511.jpg: \log_{2}(5x) - 4 = 2
img512.jpg: \log_{3}(x) = 2
img513.jpg: \log(x) + 2 \cdot \log(x + 1)
img514.jpg: \log_{3}(18) + \log_{3}\left(\frac{3}{2}\right)
img515.jpg: \log(e^2) + \log(e^{-2})
img516.jpg: \log(20) + \log(50)
img517.jpg: \log(200) + \log(5) - \log(100)
img518.jpg: \log_{3}(5x) - 4 \cdot \log_{3}(x)
```

FIGURE 3.4: LaTeXMarkup Dataset

3.3 Data Pre-processing

Data pre-processing is essential in any image captioning model that one wishes to develop. For text, it includes processes such as conversion of the text to tokens, conversion of all the characters in the text from upper case to lower case and the elimination of special characters so as to make the text clean. These steps help the model to learn from the provided captions and generate good output.

3.3.1 Text Data Preprocessing

Text dataset preprocessing prepares the textual descriptions and LaTeXnotations associated with mathematical expressions for model training and evaluation. The key steps of text data preprocessing are:

- **Noise Removal:** In text data, there are often unwanted characters or symbols that can clutter the input and make it difficult for the model to learn

effectively.The removal of such components simply means that the quality of text data is considerably improved from the start. This process of cleaning the text helps to filter out unwanted information from the text that can have no role in the training. Therefore, the model can learn the genuine patterns. This clearer input text in the training process results in better and more credible training results thus improving the model.

- **Unwanted Space Removal:** A lot of white spaces in text data may cause confusion and lead to the provision of wrong information during training of the model. If there are multiple spaces between words, or around any characters, it could interfere with the model's capacity to interpret the text. In doing this we eliminate unnecessary blanks, thus making our data more standardized and easy to work with. This modification assists the model to have a better understanding on how the text is structured and what it conveys.

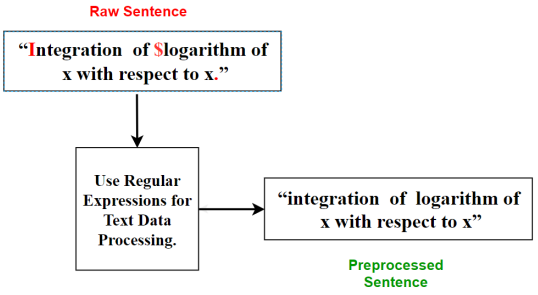

FIGURE 3.5: Example of text pre-processing

- **Tokenization:** Tokenization is the process applied in the Natural Language Processing and the machine learning to transform a string of text into tokens. This preliminary step is important in order to pre-process the text data in order to be used in other mathematical formulas and models for analysis. Tokenization is a process of splitting the text into pieces as small as a single character or as large as a whole possible, a word, if needed for the certain task.
  - *Example:* As an example we have taken a sentence from the preprocessed captions in.
  - *Text Sentence:* " \[x = r \cdot \cos (\theta) \] ".
  - *Tokens:* "\", "[", "x", "=", "r", "cdot", "cos", "(", "theta" , ")", "]".

3.4 Image Feature Extraction Module

In generating the textual descriptions and the LaTeXnotations of the handwritten mathematical expressions, feature extraction is very efficient in analyzing the image data In this research, we use the pre trained convolutional neural network (CNN) based feature extraction [9]. To identify the most effective model for our dataset, we experimented with four different pre-trained CNNs: Its use in the four types of architectures of Neural Networks, namely VGG-16, VGG-19, Inception-V3, and ResNet 50. These models have been chosen due to their proven performance in various image recognition tasks.

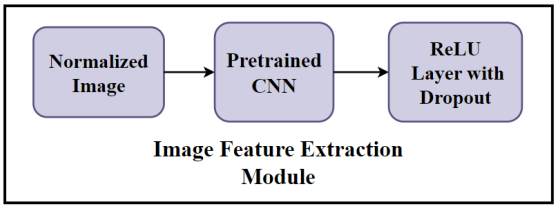

FIGURE 3.6: Image Feature Extraction Module

Each CNN model is used to process the images and extract high-level features that capture important patterns and details in the handwritten mathematical expressions [10].Here we have demonstrated how features by are extracted by pre-trained CNN from the input image [11].

- **Image Representation and Normalization:** After converting the image into a 2-D matrix of pixel values ranging from 0 to 255, the next step is to normalize these values between 0 and 1. This normalization is done using min-max normalization, which involves dividing each pixel value by 255. By doing this, we scale down the values so that they fall within the range of 0 to 1. This step is important because it helps to standardize the data, making it easier for the model to process and learn from the images.

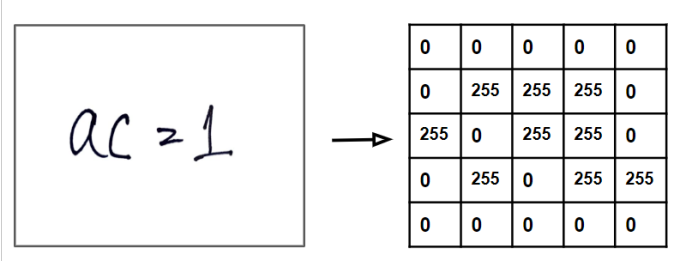

FIGURE 3.7: Representation of Input Image

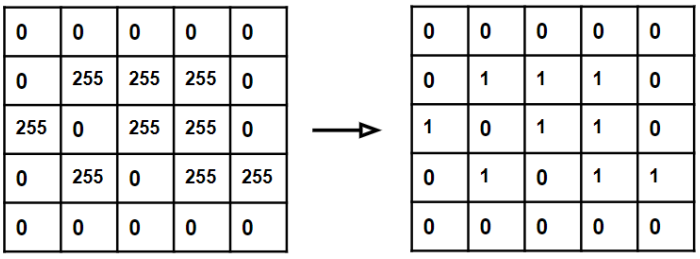

FIGURE 3.8: Normalize the image using min-max normalization

- **Convolution Operation:** Convolution is the process of integrating two functions and generates a third function, which illustrates the alteration of one of the functions by the other one. In the context of Convolutional Neural Networks (CNNs), convolution is used to merge two sets of information: input image and a filter or kernel [11]. The filter moves over the image and at every position, the filter multiplies each element of the filter with the corresponding input values and sums the outputs to give a single output. This process assists in arriving at crucial features in the input image include edges, textures and patterns. As the subsequent layers apply several filters, a CNN can learn features in different layers, allowing it to address a series of aspects of the visual input efficiently [12].

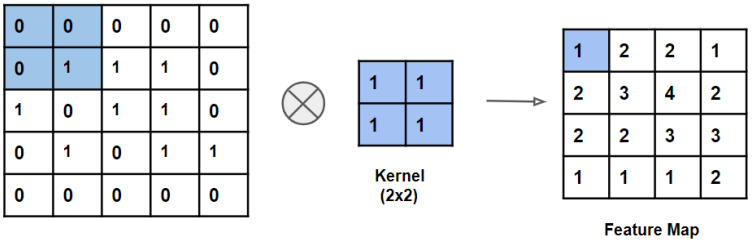

FIGURE 3.9: Convolution operation

- **Pooling:** Pooling layers are the essential components of Convolution Neural Networks (CNNs) that are instrumental in the down-sampling of the feature maps' dimensionality. This dimensionality reduction reduces the number of parameters that the network has to learn and the amount of computations that has to be done, thus the network's efficiency is increased. There are two main types of pooling:There are two main types of pooling:
  - *Max pooling*: In max pooling, the input feature map is divided into smaller parts and the maximum value from each part is selected to form the output feature map. This method helps in capturing the most important features while reducing the spatial dimensions of the data.

– *Average pooling:* In average pooling, the input feature map is also divided into smaller regions, but instead of selecting the maximum value, the average value of each region is calculated to form the output feature map. This technique smooths the data and retains the average presence of features.

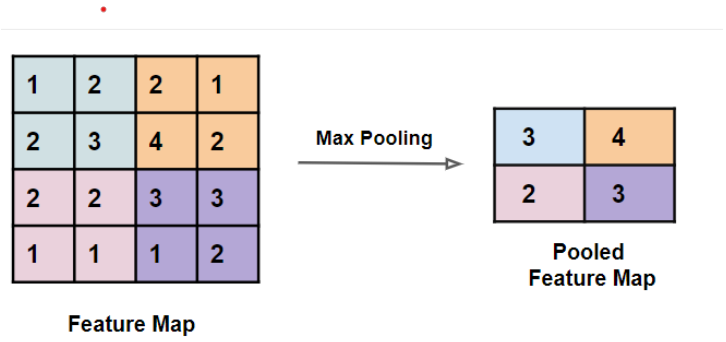

FIGURE 3.10: Max Pooling

Both max pooling and average pooling help in downsampling the feature maps, which reduces the complexity of the model and helps prevent overfitting, while preserving important information.

- **Flattening**:After finishing the previous two steps, we are supposed to have a pooled feature map. As the name of this step implies, we have flattened our pooled feature map into a column ($1 - D$ array).

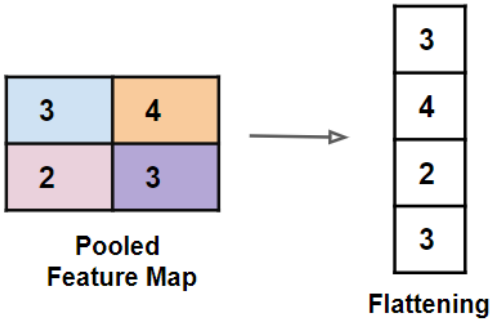

FIGURE 3.11: Flattening the pooled feature map

- **Fully Connected Layer:** A Fully Connected (FC) layer, also known as a dense layer, is a type of layer used in artificial neural networks where each neuron or node from the previous layer is connected to neuron of the current layer. It is called fully connected because of this complete linkage between the layers. Features are extracted from the previous layer of the last layer of fully connected layer.

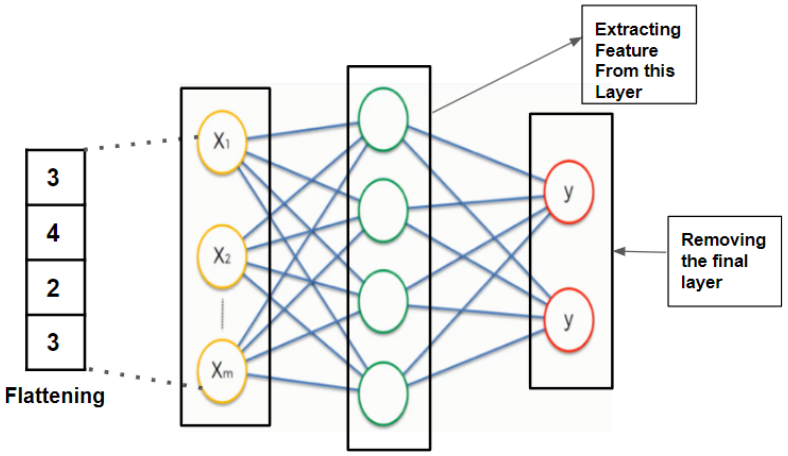

FIGURE 3.12: Extract the feature from Fully Connected Layer

The following steps describe the processes in Image Feature Extraction Module:

- Feature extraction is a crucial step in processing image data for our research, which aims to generate textual descriptions and LaTeXnotations for handwritten mathematical expression images. To extract meaningful features from these images, we employ a pre-trained convolutional neural network (CNN).
- Different types of pre-trained CNNs are deployed because these models have been trained on large datasets like ImageNet and are adept at extracting important information from images. Utilizing these pre-trained networks leverages their ability to recognize complex patterns and features that are useful for our specific task.
- For our purposes, we eliminate the pre-trained CNN's last layer, which is typically a classification layer used for identifying objects in images. This modification allows us to repurpose the network for feature extraction, focusing on capturing the essential characteristics of the images without constraining them to predefined classes.
- The first step is the image preparation of handwritten mathematical expressions that will be used as inputs for the CNN model. To pass from the pre-trained CNNs , images are scaling down the to fit input dimensions of the pre-trained CNN as well as standardizing the pixel values in the images.
- Once the images are prepared, they are passed through the convolutional and pooling layers of a pre-trained CNNs. The final classification layer is removed from the CNN's. This helps in capturing the learned features of the images, fine details and patterns from the penultimate layer of the CNN's.

- The output of this feature extraction process is the feature vectors from the penultimate layer of the network. All these vectors contain such important and elaborate features about the patterns and information concerning the handwritten mathematical expressions.
- Next we add a linear transformation layer(ReLU layer) to adjust the dimensionality of the feature vectors. This transformation ensures that the feature vectors are compatible with the subsequent layers in the fusion module, facilitating seamless integration and improving the overall performance of the image captioning system.
- For a further termination of overfitting, we add a dropout layer after the linear transformation layer known as ReLU layer. This dropout layer puts a percentage of neurons momentarily to sleep as a method of training, which keeps the model from getting too familiar to the training data. Since dropout grasps the situation, it eliminates overfitting hence allowing the model to perform well on new data it has not seen before.

Firstly, a pre-trained Convolution Neural Network (CNN) is used to extract features from the handwritten images in our model. Let I represents the input image. The input image I is fed into a pre-trained CNN, denoted as $\text{CNN}_{\text{pre-trained}}$. This CNN is typically pre-trained on large image datasets like ImageNet, which allows it to capture complex features of the given images. The output of the CNN is a high-dimensional feature map $F = \text{CNN}_{\text{pre-trained}}(I)$. The dimension of the pre-extracted feature vector is a.

Secondly, the high-level features F extracted by the pre-trained CNN are passed through a Dense (fully connected) layer. This layer transforms the features into a suitable format for further processing by the Addition Layer in the Fusion Model. Let W_d and b_d denote the weights and biases of the Dense layer, respectively. The output of the Dense layer is :

$$f_d = \text{ReLU}(W_d \cdot F + b_d) \tag{3.1}$$

ReLU is the Rectified Linear Unit activation function. This output f_d serves as a high-level representation of the input images.The dimension of f_d is m. Then a dropout layer is with a regularization is added to remove the overfitting issue. f_r represents the feature vector after regularization. This vector is passed to the addition layer.

3.5 Text Generation Module

The Text Generation Module processes the textual descriptions and the LaTeXnotation associated with the images.This model preserves the semantics of the textual data and finds the relationship between the words and characters.

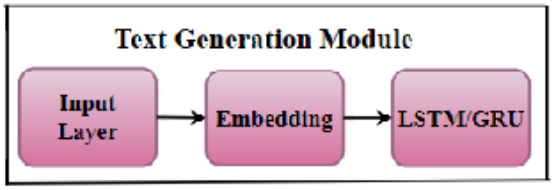

FIGURE 3.13: Components of Text Generation Module

This module consists of three steps. These steps are:

3.5.1 Handling Text Sequences:

Handling text sequences involves converting the words and characters of the vocabulary into sequences of integers. Each word or character is assigned a unique integer identifier, which allows the text to be transformed into a numerical format suitable for model processing. This transformation is very important because the model has to work with textual data in some sort of structured manner. In order to structure the text, each vocabulary item is given a unique integer and thus the input is made more intelligible to the model to learn from. This numerical representation is used for subsequent processing operations such as embedding and sequence modeling relevant for text analysis and generation.

3.5.2 Embedding Layer

Word Embedding is a technique of mapping words to vectors of a multi dimensional space such that words having similar meaning and similar context are mapped to similar vector. Here in this work, the embedding technique that we have used is *Word2vec*.

In processing captions or LaTeXnotations, these sequences of integers have to be of fixed size. However, if the size of a caption or notation is less than the required size, zeros are added at the end of the sequence to match the maximum length. This is referred as *padding*. Padding makes sure that all the sequences are of equal length as it helps during processing batches of text data. Padding helps in

avoiding the skewing of the sequences to the length of the shortest sequence by standardizing the length of the sequences. With the help of the embedding matrix embedding sequence is generated and passed to the RNN layer.

3.5.3 RNN Layer

Subsequently to the embedding layer, the sequence of embeddings is passed through the RNN layers, namely LSTM and GRU [13]. These layers are expected to learn long-range dependencies and context in the text data. They operate by keeping a sequence of hidden states, which hold information of the previous time steps of the sequence [14]. This capability is important to capture the relational context in the text as well as the nuances that enable the model to provide better captions. In the case of sequences, LSTM and GRU layers are quite effective since they have the ability to remember information for long time and at the same time forgot what is not useful [15]. As the sequence of embeddings passes through these layers a series of hidden states are produced that contains the learned context and the relations within text [16].

The hidden states which come from the RNN layers are then transferred to an addition layer for more computation to yield the caption. Here we have discussed two types of RNN briefly:

3.5.4 Long Short Term Memory (LSTM)

LSTM, short for Long Short Term Memory networks are a type of recurrent neural network architecture which addresses the problem known as vanishing gradient through capturing long-term dependency in sequence data [17] [18]. These gates are given as follow:

- **Input Gate:** The input gate controls the flow of information into the cell state. It takes the current input and the previous hidden state as inputs. It passes these inputs through a sigmoid activation function, producing values between 0 and 1. The output of the sigmoid gate is multiplied element-wise with the candidate cell state (explained below) to determine which information to keep or discard.
- **Forget Gate:** The forget gate decides what information to discard from the cell state. It takes the current input and the previous hidden state as inputs.

Similar to the input gate, it passes these inputs through a sigmoid activation function, producing values between 0 and 1. The output of the sigmoid gate is multiplied element-wise with the previous cell state, determining which information in the cell state to keep or forget.

- **Cell State:** The cell state represents the memory of the LSTM unit. It is updated using the input gate, forget gate, and a new candidate value (explained below) The previous cell state is multiplied element-wise with the forget gate output, and the input gate output is multiplied element-wise with the new candidate value. These two results are added together to produce the new cell state.
- **Candidate Cell State:** The candidate cell state represents the new candidate values to be added to the cell state. It is calculated based on the current input and the previous hidden state. The calculation involves passing and through a tanh activation function, which squashes the values between -1 and 1. The output of the tanh activation function is then multiplied by the input gate output, allowing the LSTM to decide how much of the candidate value to add to the cell state.
- **Output Gate**: The output gate defines what the next hidden state is going to be based depending on the new cell state and the current input. It is derived from the present input and the previous hidden state as inputs. These inputs are fed into a sigmoid activation function. The current cell state is passed through a tanh function so that the output lies between -1 to 1. The output from the tanh activation function is then multiplied with the sigmoid gate output element-wisely. This then generates the next hidden state.

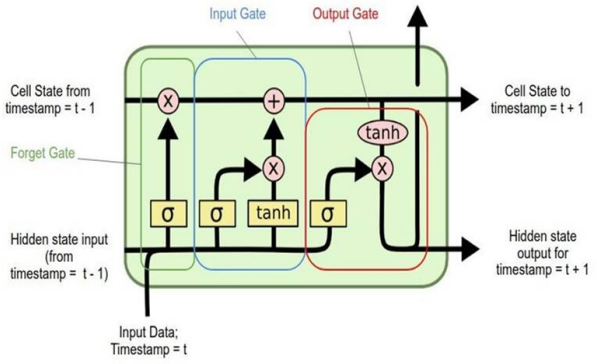

FIGURE 3.14: Long short term memory(Credit:Analytics Vidhya)

3.5.5 Gated Recurrent Unit (GRU)

A Gated Recurrent Unit (GRU) is a category of Recurrent Neural Network (RNN) and it is used in handling sequential data and it helps to overcome the vanishing gradient problem of traditional RNN [19]. It features two gates: the update gate and the reset gate which enables the information to pass through and retain the information of the sequence. Comparing to LSTMs, GRUs are less complicated and take less time for computations as they have fewer parameters; thus, they are preferred for time series prediction, NLP, or speech recognition. The gates of GRU are described below:

- **Reset Gate:** The reset gate is used to allow the GRU to select which portion of the previous information it should forget. On receiving new information, the reset gate verifies the old information and determines if any of it can be discarded. This way it is made easier to forget all the past info and focus on the new info as it is only new info that is stored as new knowledge.
- **Update Gate:** The update gate determines the quantities of old information that need to be retained and the new information that need to be added. The update gate assists the GRU in controlling the flow of information between the short and long term memory. Thus, the GRU is more accurate in decision-making processes.

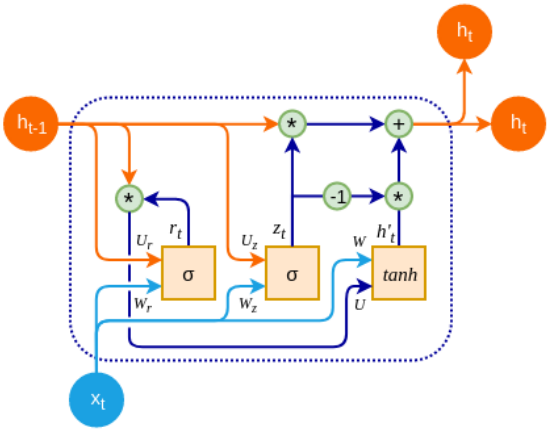

FIGURE 3.15: Gated recurrent unit(credit:OREILLY)

Let $T = \{t_1, t_2, \ldots, t_n\}$ represent the integer sequence (which represents tokenized words and characters) after preprocessing, where t_i denotes the i-th token in the

sequence. This sequence consists of integer representing parts of the handwritten mathematical expressions. The text sequence T is passed through an Embedding layer, which converts each token t_i into a dense vector representation e_i of dimension m. Let E denote the embedding matrix of dimension $m * a$,where a is the vocabulary size. The embedding for the i-th token is $e_i = E[t_i]$.

The embeddings $\{e_1, e_2, \ldots, e_n\}$ are then fed into a Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) layer. These recurrent layers are designed to capture the sequential dependencies and context within the text data. Let h_i denote the hidden state at time step i. The hidden states are computed as.

$$h_i = \text{LSTM/GRU}(e_i, h_{i-1}) \tag{3.2}$$

for i=1,2,. . . ,n h_i is calculated.

This h_i of dimension m is passed in the addition layer of the Fusion Module for further processing.

3.6 Fusion Module

This module is used to merge the output that comes from the feature extraction model and the sequential model out generate the output based on the probability distribution over vocabulary.

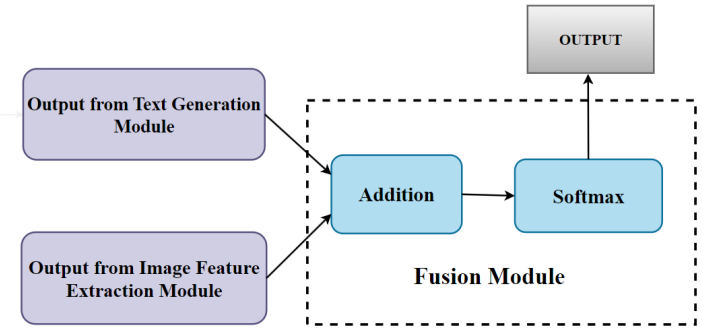

FIGURE 3.16: Components of fusion module

This model consists of two components:

3.6.1 Addition Layer

The output from the Image Feature Extraction Module is added element-wise to a Text Generation Module for addition. Any two corresponding elements from the

two outputs, their values will be added. Similar to an attention mechanism, this layer assists the model in paying more weights for significant features by jointly combining information of these two [20]. The model combines features learned from the image with context or information obtained by passing sequences of words through a sequential model(by adding outputs together). This allows a better overall understanding and consequent performance for the mathematical image captioning or recognition based tasks.

3.6.2 Softmax Layer

The linear transformation produces a sequence of vectors from the output Of Addition Layer in Softmax layer [21]. This is called as the logit vector, or in other words raw prediction scores for each vocabulary token. This logit vector sequence has the same length as the size of your vocabulary, and contains a unique value associated to each character or word [22]. Then we apply the softmax function on a sequence of logit vectors . After applying the softmax function to these raw scores, they will be converted into a probablity distribution for every score; i.e., each of them assigned with some probability value between 0 and 1. The sum of all these probabilities is equal to 1. This is the distribution of how much a word in each vocabulary is related to the input data. In each step, the token with highest probability is selected as output. Therefore, the model now picks a single character/word from vocabulary naturally based on input features and context provided by earlier layers. This will be repeated for each and every step so that the model can generate a sequence of tokens to make into final output.

The final hidden state from the LSTM/GRU layer, h_n, and the feature representation from the Dense layer, f_r, are added element-wise to form a combined feature vector c of dimension m. The combined feature vector c is transformed to logit vector sequence z where $z = [z_1 z_2 .... z_a]$ and a is vocabulary size.This z vector sequence is passed in the Softmax layer which generates a probability distribution over the logit vector sequence which represents the tokens present in the vocabulary. The token with the highest probability is reflected as output. The output of the Softmax layer is :

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{a} e^{z_j}} \tag{3.3}$$

where z_i represents the probability of ith token in the vocabulary. In this equation, $\mathbf{z} = [z_1, z_2, \ldots, z_n]$ is the input vector sequence to the softmax function.Probability

is calculated over every vector in the sequence. The predicted token (represented by logit vector) which has an higher probability among all the tokens presents in the vocabulary, is out from this final layer as output.

Chapter 4

Results and Discussion

This section shows the results of experiments using different pre-trained Convolutional Neural Networks (CNNs) combined with Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The results are displayed in a tabular format.

The model is trained with different combinations to determine which configuration yields the best results. Following evaluation measures as defined in are used to compare the performance of different combinations [23]

- **BLEU Score:** The Bilingual Evaluation Understudy (BLEU) score is a widely recognized metric for evaluating the quality of text generated by machines. Initially designed for assessing the quality of machine-translated text, the BLEU score has been adapted for various natural language processing tasks, including the evaluation of image captions generated by models. This metric works by comparing the generated text to one or more reference texts, calculating the precision of n-grams (contiguous sequences of n items) in the candidate text relative to the reference texts. Higher BLEU scores indicate closer similarity to human-generated text, thereby reflecting higher quality in the machine-generated captions. Due to its ability to provide a quantifiable measure of text quality, the BLEU score has become a standard tool for assessing the performance of image captioning models [24].

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

Here, BP is the brevity penalty. p_n is the precision for n-grams of size n. w_n is the weight assigned to the n-gram precision (usually, $w_n = \frac{1}{N}$ for uniform weighting).N is the maximum n-gram size considered (typically 4).

- **METEOR Score:**The Metric for Evaluation of Translation with Explicit Ordering (METEOR) score is a tool used to measure the quality of text generated by machines. It compares the machine-generated text to human-written reference texts. Unlike simpler methods that only look for exact word matches, METEOR also considers synonyms, word stems (words with the same root), and paraphrasing. It also checks the word order to see if the text makes sense. By looking at these different aspects, METEOR provides a more accurate and reliable measure of text quality, making it useful for tasks like image captioning and machine translation. This helps ensure the generated text is closer to what a human might write.

$$\text{METEOR} = F_{mean} \cdot (1 - Penalty) \tag{4.2}$$

  F_{mean} is the harmonic mean of precision (P) and recall (R). *Penalty* is the fragmentation penalty.

- **ROUGE Score:** The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score is a set of metrics used to measure the quality of text generated by machines, like image captions. It does this by comparing the generated text to human-written reference texts. The main focus of ROUGE is on recall, which means it checks how well the generated text includes important parts of the reference text. This helps determine how accurately the machine-generated text captures the essential information.

$$\text{ROUGE} = \frac{\text{No of matching n-grams}}{\text{Total no of n-grams in reference text}} \tag{4.3}$$

We have divided our result section mainly into two parts. First is result of textual description generation of handwritten mathematical expression and second is LaTeXnotation generation of handwritten mathematical expression.

4.1 Textual Description Generation

In this section, we have demonstrated three tables. The first table shows the accuracy matrix, the second table presents the BLEU score, METEOR score, and ROUGE score combinations with LSTM, and the third table shows same matrices combination with GRU for LaTeXnotation generation.

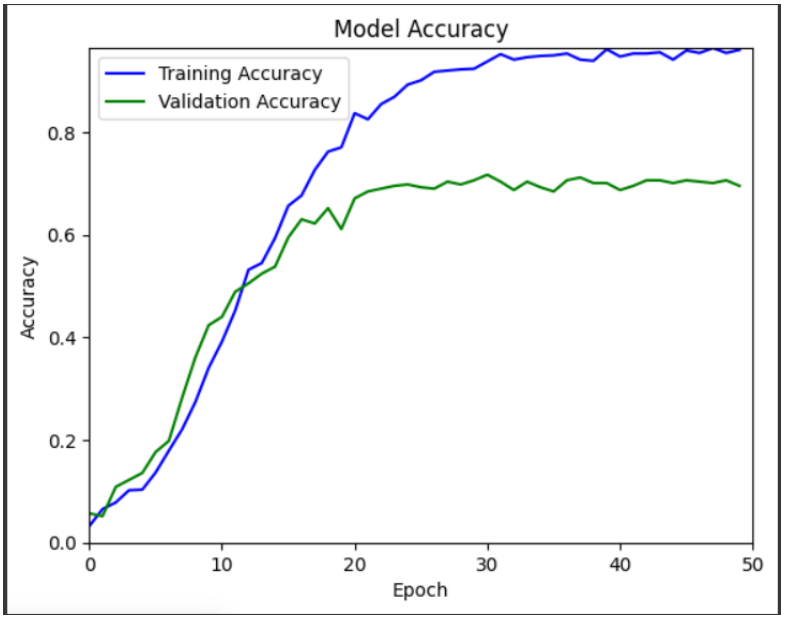

FIGURE 4.1: Accuracy vs Epoch

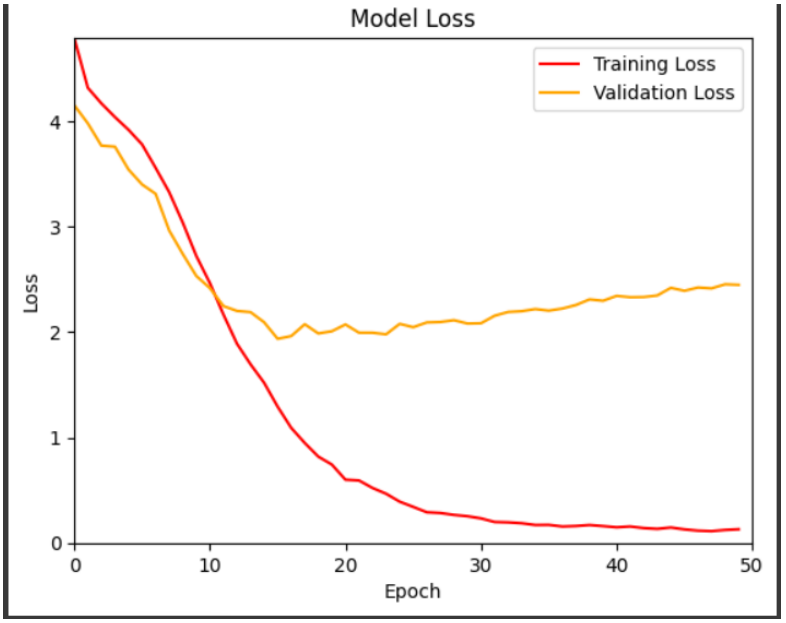

FIGURE 4.2: Loss(units) vs Epoch

FIGURE 4.2 indicates the accuracy of our model with different pretrained CNN and RNN combinations. When we used *LSTM* as the RNN, we achieved the highest accuracy of *67.3%* by using *VGG19* as the pretrained CNN. If we used *GRU* as the RNN, we achieved the maximum accuracy of *68.9%* by using *InceptionV3* as the pretrained CNN.

Pretrained CNN	VGG16	VGG19	ResNet50	InceptionV3
Validation Accuracy(%)	68.5	67.3	65.5	**66.7**
BLEU-1 score	0.512	0.502	**0.542**	0.511
BLEU-2 score	0.274	0.271	**0.278**	0.272
BLEU-3 score	0.184	0.181	**0.189**	0.182
BLEU-4 score	0.084	0.080	**0.088**	0.212
METEOR score	0.225	0.215	**0.254**	0.212
ROUGE score	0.313	0.307	**0.338**	0.317

TABLE 4.1: Evaluation matrix with LSTM(RNN)

Pretrained CNN	VGG16	VGG19	ResNet50	InceptionV3
Validation Accuracy(%)	68.2	68.1	66.9	**68.9**
BLEU-1 score	0.517	0.508	**0.547**	0.514
BLEU-2 score	0.277	0.273	**0.287**	0.274
BLEU-3 score	0.187	0.181	**0.192**	0.183
BLEU-4 score	0.082	0.076	**0.089**	0.080
METEOR score	0.223	0.215	**0.256**	0.213
ROUGE score	0.313	0.305	**0.337**	0.318

TABLE 4.2: Evaluation matrix with GRU(RNN)

TABLE 4.1 indicates the evaluation metrics of our model. We tested four different combinations with pretrained CNN and LSTM. We achieved a maximum BLEU-1 score of 0.542, METEOR score of 0.254, and ROUGE score of 0.338 by using ResNet50 with LSTM.

TABLE 4.2 indicates the evaluation metrics by using GRU instead of LSTM. We achieved a maximum BLEU-1 score of 0.547, METEOR score of 0.256, and ROUGE score of 0.337 by using ResNet50 with GRU.

4.2 LaTeXMarkup Generation

In this section, we have demonstrated three tables. The first table shows the accuracy matrix, the second table presents the BLEU score, METEOR score, and ROUGE score combinations with LSTM, and the third one shows the same matrices combination with GRU for LaTeXnotation generation.

FIGURE 4.3 indicates the accuracy of our model with different pretrained CNN and RNN combinations. When we used *LSTM* as the RNN, we achieved the highest accuracy of *74.5%* by using *InceptionV3* as the pretrained CNN. If we used *GRU* as the RNN, we achieved the maximum accuracy of *75.7%* by using *InceptionV3* as the pretrained CNN.

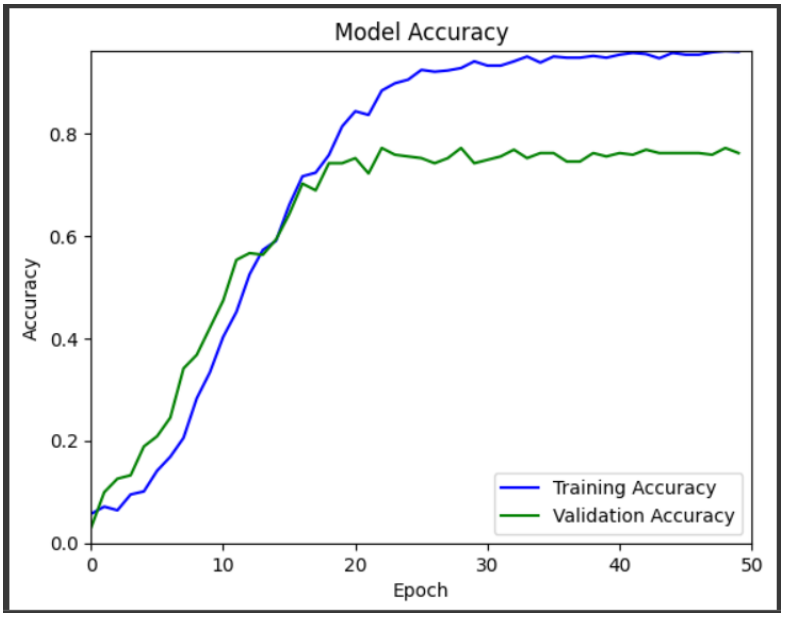

FIGURE 4.3: Accuracy vs Epoch(LaTeX)

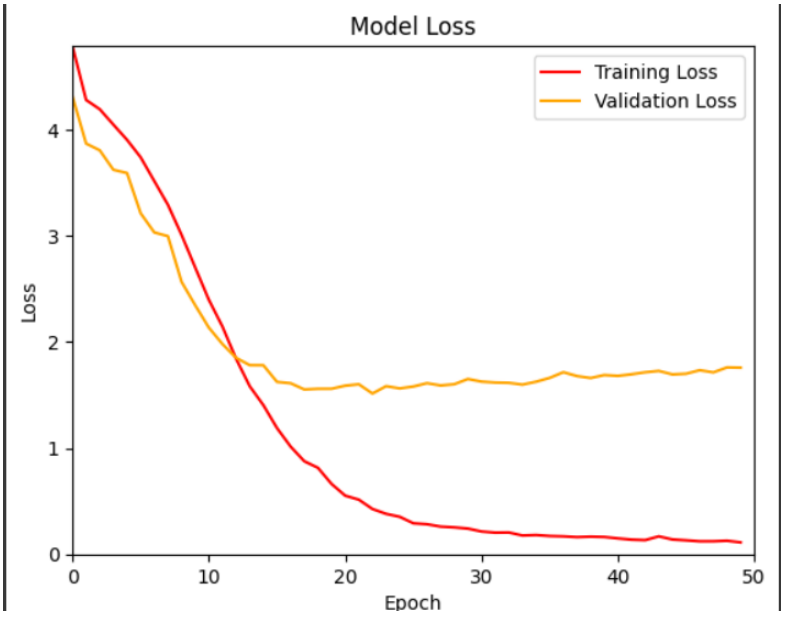

FIGURE 4.4: Loss (units) vs Epoch(LaTeX)

Pretrained CNN	VGG16	VGG19	ResNet50	InceptionV3
Validation Accuracy(%)	69.2	68.7	73.6	**74.5**
BLEU-1 score	0.531	0.511	0.543	**0.548**
BLEU-2 score	0.275	0.273	0.276	**0.279**
BLEU-3 score	0.184	0.179	0.186	**0.188**
BLEU-4 score	0.057	0.083	0.089	**0.091**
METEOR score	0.231	0.232	0.243	**0.248**
ROUGE score	0.313	0.314	0.326	**0.337**

Table 4.3: Evaluation Matrix with LSTM(RNN)

Pretrained CNN	VGG16	VGG19	ResNet50	InceptionV3
Validation Accuracy(%)	70.2	68.9	73.2	**75.7**
BLEU-1 score	0.532	0.519	0.542	**0.549**
BLEU-2 score	0.274	0.272	0.277	**0.281**
BLEU-3 score	0.183	0.180	0.187	**0.189**
BLEU-4 score	0.085	0.081	0.086	**0.091**
METEOR score	0.231	0.225	0.240	**0.246**
ROUGE score	0.301	0.304	0.321	**0.339**

Table 4.4: Evaluation Matrix with GRU(RNN)

TABLE 4.3 indicates the evaluation metrics of our model. We tested four different combinations with pretrained CNN and LSTM. We achieved a maximum BLEU-1 score of 0.549, METEOR score of 0.248, and ROUGE score of 0.337 by using InceptionV3 with LSTM.

TABLE 4.4 indicates the evaluation metrics by using GRU instead of LSTM. We achieved a maximum BLEU-1 score of 0.549, METEOR score of 0.246, and ROUGE score of 0.339 by using InceptionV3 with GRU.

Chapter 5

Conclusion and Future work

In this project, we developed an advanced model to generate text descriptions and LaTeXnotations for handwritten math expressions. Our methodology harnessed several pretrained Convolutional Neural Networks (CNNs) such as VGG16, VGG19, ResNet50, and InceptionV3, combined with Recurrent Neural Networks (RNNs) like LSTM and GRU. Rigorous experimentation showed that integrating these CNNs with LSTM and GRU architectures significantly improved performance metrics, including BLEU, METEOR, and ROUGE scores. These metrics measure the accuracy and quality of the generated descriptions and notations. Our model achieved a maximum accuracy of 68.9% for text descriptions and 75.7% for LaTeXnotations, demonstrating its effectiveness and robustness.

The implications of our findings suggest numerous opportunities for future research and development. One promising direction is exploring more advanced architectures, such as Transformer models. Integrating Transformer models could potentially enhance our model's performance further. Additionally, the emerging field of quantum CNNs presents another intriguing avenue. Quantum CNNs leverage quantum computing principles to process information in ways classical computers cannot, potentially leading to unprecedented improvements in our model's capabilities. Another crucial aspect for future improvement is diversifying our dataset. Incorporating a broader range of handwriting samples can significantly enhance the model's precision and reliability. Our primary objective moving forward is to optimize the current model by fine-tuning its parameters and incorporating additional layers to capture more complex features of handwritten mathematical expressions.

References

[1] Herdade, Simao, et al. "Image captioning: Transforming objects into words." Advances in neural information processing systems 32 (2019).

[2] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.

[3] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[4] Sharma, Dhruv, Chhavi Dhiman, and Dinesh Kumar. "Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey." Expert Systems with Applications 221 (2023): 119773.

[5] Li, Zhe, et al. "Improving handwritten mathematical expression recognition via similar symbol distinguishing." IEEE Transactions on Multimedia (2023).

[6] Kristianto, Giovanni Yoko, and Akiko Aizawa. "Extracting textual descriptions of mathematical expressions in scientific papers." D-Lib Magazine 20.11 (2014): 9.

[7] Awal, Ahmad-Montaser, Harold Mouchère, and Christian Viard-Gaudin. "The problem of handwritten mathematical expression recognition evaluation." 2010 12th International Conference on Frontiers in Handwriting Recognition. IEEE, 2010.

[8] Bian, Xiaohang, et al. "Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 1. 2022.

[9] Dai Nguyen, Hai, Anh Duc Le, and Masaki Nakagawa. "Recognition of online handwritten math symbols using deep neural networks." IEICE TRANSACTIONS on Information and Systems 99.12 (2016): 3110-3118..

[10] Mondal, Ajoy, and C. V. Jawahar. "Textual description for mathematical equations." 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

[11] Zhang, Jianshu, et al. "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition." Pattern Recognition 71 (2017): 196-206.

[12] HN, Sharada, Basavaraj Anami, and Shridhar Allagi. "An optimized neural Network-based character recognition and relation finding for mathematical expression images." Multimedia Tools and Applications (2023): 1-23.

[13] Bhalekar, Madhuri, and Mangesh Bedekar. "The New Dataset MITWPU-1K for Object Recognition and Image Captioning Tasks." Engineering, Technology Applied Science Research 12, no. 4 (2022): 8803-8808.

[14] Zhang, Zhang, and Yibo Zhang. "Combining CNN and Transformer as Encoder to Improve End-to-End Handwritten Mathematical Expression Recognition Accuracy." International Conference on Frontiers in Handwriting Recognition. Cham: Springer International Publishing, 2022.

[15] Tang, Jia-Man, et al. "Offline handwritten mathematical expression recognition with graph encoder and transformer decoder." Pattern Recognition 148 (2024): 110155.

[16] Liu, Shuang, et al. "Image captioning based on deep neural networks." MATEC web of conferences. Vol. 232. EDP Sciences, 2018.

[17] Zhang, Jianshu, Jun Du, and Lirong Dai. "A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition." 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. IEEE, 2017.

[18] Zhelezniakov, Dmytro, Viktor Zaytsev, and Olga Radyvonenko. "Online handwritten mathematical expression recognition and applications: A survey." IEEE Access 9 (2021): 38352-38373.

[19] Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar. "Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey."Expert Systems With Applications 221 (2023) 119773.

[20] Bui, Khanh-Ngoc, and Thanh-Sach Le. "Handwritten Mathematical Expression Recognition: An approach on data augmentation." 2021 15th International Conference on Advanced Computing and Applications (ACOMP). IEEE, 2021.

[21] Mirkazemy, Abolfazl, et al. "Mathematical expression recognition using a new deep neural model." Neural Networks 167 (2023): 865-874.

[22] Liu, Yutian, Wenjun Ke, and Jianguo Wei. "Attention Guidance Mechanism for Handwritten Mathematical Expression Recognition." arXiv preprint arXiv:2403.01756 (2024).

[23] Chi, Xueke, et al. "Handwritten mathematical expression recognition with self-attention." Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence. 2021.

[24] Sneha, Ch Premanvitha, B Shanmukh, Biradavolu Chaduvula, Kavitha. (2022). IMAGE CAPTION GENERATOR USING DEEP LEARNING-Convolutional Neural Network, Recurrent Neural Network, (Bilingual Evaluation Understudy)BLEU score,Long Short Time Memory.