

University of Bahrain

College of Science

Postgraduate Programs



A Novel model for Chart-to-Text Generation by utilizing NN models

Thesis submitted in partial fulfilment of the requirements for the degree
of M.Sc. In Big Data Science and Analytics

Prepared by:

Hamad Munir Malik

ID# 20180548

Supervised by:

Prof. Nabil Hewahi

Professor

University of Bahrain

Kingdom of Bahrain

April 2024



Thesis Discussion Committee Decision

Student Name: HAMAD MUNIR MALIK	Student ID: 20180548
Department: Choose an item.	College: Science
Program: M.Sc. in Big Data Science and Analytics	
Thesis Title: A Novel model for Chart-to-Text Generation by utilizing NN models	
Supervisor Name: Prof. Nabil Hewahi	Student GPA: 3.39

The Discussion Committee met on 24-Jun-24 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

PASS

The student is awarded a grade Choose an item. in the thesis subject.

- ☐ Without making any changes to the thesis.
- ☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name

1. **Prof. Nabil Hewahi**
2. **Dr. Riadh Ksantini**
3. **Prof. Hani A. Bani-Salameah**

Signature

(Advisor)

Nabil Hewahi

(Internal Examiner)

P.t

(External Examiner)

Hani

Abstract

This research introduces a novel neural network-based model designed to enhance the automatic generation of textual descriptions for scientific charts, building on the foundational work of "SCICAP: Generating Captions for Scientific Figures." Our proposed model combines advanced Convolutional Neural Network (CNN) architectures for effective feature extraction from scientific charts with multiple Recurrent Neural Network (RNN) layers to generate descriptive, accurate captions. This study is distinguished due to its comparative analysis of different RNN architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), to identify the most effective model for textual caption generation. Central to our methodology is the utilization of CNNs to parse visual input, transforming complex chart images into a rich set of feature vectors. These vectors serve as inputs for the sequential RNN layers, which are tasked with constructing coherent and contextually relevant textual descriptions. The performance of various CNN-RNN combinations is rigorously evaluated using the BLEU-4 (Bilingual Evaluation Understudy) metric, a standard in the natural language processing field for assessing the linguistic quality of machine-generated text against human-written references. Our Research showed that Bi-LSTM performed the best with a BLEU score of 0.3, future researches will be required for more reliable captioning.

Acknowledgements

I extend my heartfelt gratitude to the institution that made my journey possible, the University of Bahrain, for presenting me with the opportunity to explore the vast world of Big Data through their Master's program in Big Data Science and Analytics.

I am indebted to Prof. Nabil Hewahi, whose mentorship and insightful critiques guided me during my research. My sincere thanks are also due to Dr. Sawsan Hilal, for her support throughout the duration of this program. Lastly, my appreciation is extended to all the dedicated faculty members, their collective wisdom has been the guiding light that illuminated my path towards the accomplishment of this academic milestone.

Table of Contents

Defense Committee's Approval.....	A
Abstract	B
Acknowledgements	C
Table of Contents	D
List of Tables.....	G
List of Figures	H
List of Abbreviations.....	I
Chapter One Introduction.....	2
1.1. Introduction	2
1.1.1. The Importance of Figures in Scientific Communication	2
1.1.2. Enhancing Scientific Communication.....	3
1.1.3. Visualizing Chart-to-Text	4
1.1.4. Contribution to the Field	5
1.1.5. Challenges for Chart to Text	6
1.2. Research Problem.....	8
1.3. Research Objectives	8
1.4. Research Significance	9
1.5. Research Scope and Limitations	10
1.6. Theses Questions.....	11
1.7. Theses Structure	12
Chapter Two Literature Review.....	14
2.1. Text and Data Summarization.....	15

2.2. Scientific Applications of Neural Networks	16
2.3. Neural Networks and Image Description	18
2.4. Chart Image Processing and Text Generation.....	19
2.5. Analysis of SCICAP and Justification for New Research	22
Chapter Three Methods and Procedures	25
3.1. Research Design.....	25
3.2. Data Preprocessing.....	27
3.2.1. Tokenization.....	27
3.2.2. Normalization.....	28
3.2.3. Word Embedding	29
3.2.4. JSON Data Format	29
3.3. Implemented Models.....	31
3.3.1. Convolutional Neural Networks (CNNs).....	32
3.3.2. Recurrent Neural Networks (RNNs).....	33
3.3.3. Model Selection Rationale	34
3.4. Performance Metrics	36
Chapter Four Results and Discussion	40
4.1. Data	40
4.1.1. Dataset Characteristics:	40
4.1.2. Data Format.....	43
4.2. Experimentation and Results.....	44
4.2.1. Experimental Setup	44
4.2.2. Hyperparameter Variations	45
4.2.3. Experimental Findings	46

4.3. Example Outputs	50
4.4. Answers to Research Questions	53
Chapter Five Conclusion and Future Work	56
5.1. Conclusion.....	56
5.2. Future Work	57
References	59
الملخص.....	62

List of Tables

Table 1: NN Models contributions to field of scientific figures	5
Table 2: Summary of Text and Data Summarization Researches....	15
Table 3: List of Scientific Applications of Neural Networks	17
Table 4: Comparison of studies for Chart to Text.....	18
Table 5: Summary of Chart to text researches	20
Table 6: Distribution of figure types in the dataset.....	40
Table 7: Number of Figures in Each Subset	42
Table 8: Hyperparameters used in experiments	45
Table 9: Experiment results - comparing different RNN Models....	46

List of Figures

Figure 1: The figure captioning model takes a scientific figure (e.g., a box plot) as input and generate captions that describes the figure.	4
Figure 2: Illustration of research paradigm	26
Figure 3: Jso	30
Figure 3: Graphical representation of implemented models	32
Figure 4: image for the figures without subfigures	41
Figure 5: image for the figures with subfigures	42
Figure 6: Comparison of results with other researches	50
Figure 7: Examples of generated summaries	51

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-Term Memory
NLP	Natural Language Processing
BLEU	Bilingual Evaluation Understudy
ResNet	Residual Neural Network
Adam	Adaptive Moment Estimation
arXiv	Archive (electronic repository of scientific papers)

Chapter One

Introduction

1.1. Introduction

1.2. Research Problem

1.3. Research Objectives

1.4. Research Significance

1.5. Research Scope and Limitations

1.6. Thesis Questions

1.7. Theses Structure

Chapter One

Introduction

1.1. Introduction

In the world of data visualization, the integration of textual descriptions plays an important role in enhancing comprehension and accessibility of charts. With the exponential growth of data in various fields, there is an increasing need for automated methods that can generate accurate and meaningful textual descriptions corresponding to visual data. These descriptions not only aid in understanding the content of charts but also facilitate efficient communication of key findings and insights. Consequently, researchers and practitioners have been exploring innovative approaches to automate the generation of textual descriptions, aiming to bridge the gap between visual data representation and linguistic interpretation

1.1.1. The Importance of Figures in Scientific Communication

Researchers and academics frequently utilize figures and charts to succinctly present complex data and insights in scientific papers. These visual representations, ranging from basic graphs to intricate diagrams, play a crucial role in conveying critical results and explaining complex phenomena. As noted by (Borkin, 2013), the ability of visual information to communicate the essential elements of scientific research efficiently makes it indispensable in scholarly communication.

1.1.2. Enhancing Scientific Communication

The motivation for developing advanced models to automatically generate high-quality captions for scientific figures is primarily to enhance the clarity and effectiveness of scientific communication. Accurate and detailed captions play a crucial role in the understanding of scientific results, as they help contextualize and clarify the information presented in visual formats.

1. **Critical Role of Clear Communication in Science:** Visual data, when accompanied by precise and informative captions, can significantly help in understanding of the complex scientific figures and data. According to (Heimlich, 2016), clear visualizations combined with effective textual descriptions dramatically improve readers' ability to absorb and understand intricate scientific insights.
2. **Improving Readability and Accessibility:** The inclusion of high-quality captions is not just about enhancing understanding but also about increasing the accessibility of scientific materials. Publications that effectively communicate their findings can reach a broader audience, including interdisciplinary researchers and the general public. Ensuring that figures in scientific papers are well-captioned is directly linked to the overall impact and reach of the research, as noted by (Rush, 2016), who emphasize the role of automated systems in improving the accessibility and usability of graphical content in digital scientific publications.
3. **Support for Non-Expert Audiences:** Effective communication in science is also vital for educating non-experts and fostering public understanding of scientific topics. Automated captioning tools can aid in this process by providing clear, concise, and relevant descriptions of

complex figures, as discussed by Visualization of Biomedical Data (O'Donoghue, 2018), These tools can demystify scientific findings, making them more approachable for non-specialists.

By focusing on enhancing scientific communication through the development of an advanced automated figure-captioning model, this research aims to contribute significantly to the field of scientific publishing. Improved captions can lead to better comprehension, wider dissemination, and greater engagement with scientific research, fulfilling the critical need for clarity and precision in scientific narratives.

1.1.3. Visualizing Chart-to-Text

To clearly illustrate the problem addressed by this research, consider the Figure 1 in which a chart image is added as input, figure captioning model converts the image to useful captions describing the chart image:

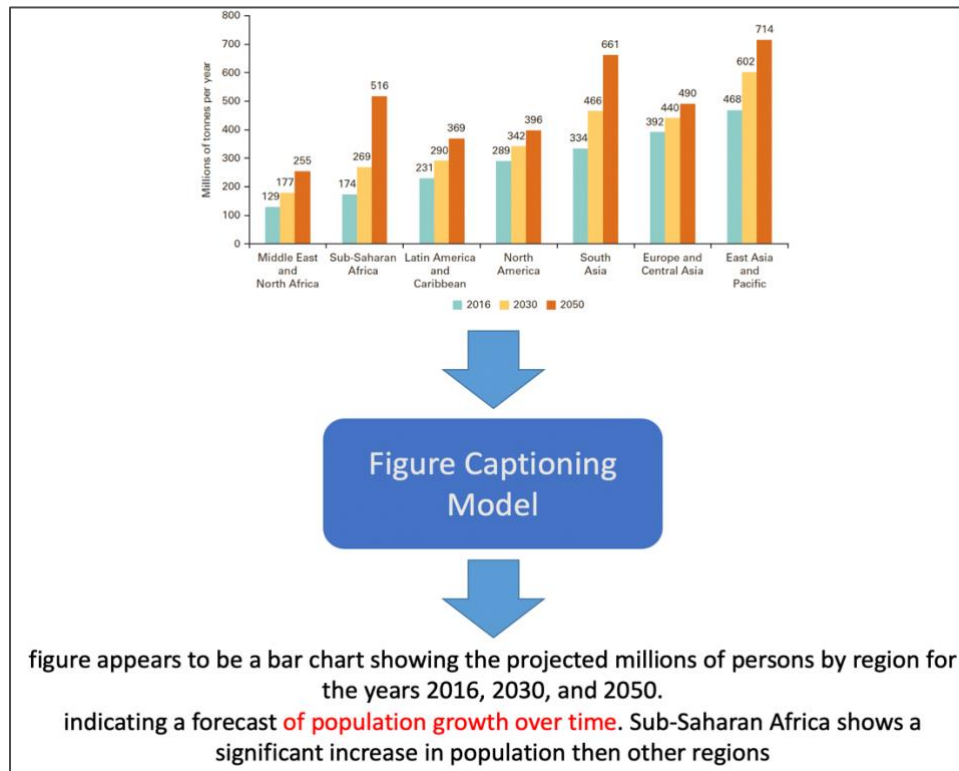


Figure 1: The figure captioning model takes a scientific figure (e.g., a box plot) as input and generate captions that describes the figure.

This research aims to develop automatic figure-captioning models that generate high-quality captions for figures and charts in scientific papers.

1.1.4. Contribution to the Field

The advent of neural network (NN) models in the realm of scientific figure captioning marks a significant technological advancement, with implications that extend across various dimensions of scientific communication and data interpretation. These models are not only revolutionizing how data is processed but also enhancing the way information is conveyed in scientific disciplines.

Table 1: NN Models contributions to field of scientific figures

Contribution	Description	Reference
Enhanced Accuracy and Contextuality	NN models improve the precision and context of captions, enhancing understanding of scientific visuals.	(Bengio P. V., 2018)
Scalability and Efficiency	NN models handle large and diverse datasets efficiently, essential in environments where quick dissemination of findings is critical.	(Saxena, 2016)
Video Classifications	Use of deep learning techniques in the field of video processing, specifically focusing on video classification and captioning, applicable to scientific figures with dynamic or complex components.	(Wu, 2017)

Table 1 provides a concise overview of the various ways in which neural network models contribute to the field of scientific figure captioning, highlighting the significant impact they have on enhancing scientific

communication, education, and accessibility. Each contribution is supported by a reference to recent research, demonstrating the current relevance and ongoing development in this area.

In summary, the integration of neural network models into scientific figure captioning represents a pivotal advancement in how scientific information is processed and communicated. As these models continue to evolve, they promise to further bridge the gap between complex scientific data and its comprehension across various audiences, thereby enriching the scientific community and beyond. This ongoing development not only supports current scientific endeavors but also paves the way for future innovations in automated text generation, enhancing both the accuracy and accessibility of scientific communication.

1.1.5. Challenges for Chart to Text

Scientific charts and figures are crucial for representing complex data succinctly and clearly in research publications. They facilitate understanding and analyzing patterns, trends, and correlations among large sets of scientific data. However, the utility of these visual representations is heavily reliant on the accompanying textual captions that explain their context, contents, and implications. This connection between visual data and text is not only fundamental for human readers but also for digital systems that aim to catalog, search, and disseminate scientific knowledge.

In the current digital age, there is an increasing demand for automating the generation of accurate and contextually appropriate textual descriptions for scientific charts. This task presents multiple challenges that stem from the inherent complexity of the charts, the variety of data they represent, and the specialized knowledge required to interpret them

correctly. Existing systems that attempt to automate chart-to-text generation face several significant issues:

1. **Complexity of Content:** Scientific charts often contain complex information with multiple variables and intricate relationships. Current automated systems struggle to decipher and describe these complexities accurately, leading to oversimplified or incorrect captions that can mislead or confuse users.
2. **Contextual Relevance:** The ability to generate captions that are not only correct but also contextually relevant to the specific field of study is another challenge. Many automated systems fail to incorporate domain-specific knowledge, which is crucial for producing meaningful and useful descriptions.
3. **Diversity in Chart Types:** There is a wide range of chart types used in scientific communication, including but not limited to bar graphs, line charts, scatter plots, and pie charts. Each type has unique features and common use cases, and an effective automated system must adapt its analysis and generation techniques to accommodate this diversity.
4. **Accuracy and Precision:** Ensuring the accuracy and precision of the generated text is paramount, as even minor errors can significantly alter the interpretation of scientific data. Existing technologies often struggle with maintaining a high level of accuracy, especially when dealing with data-intensive or complex figures.
5. **Scalability:** The ability to scale the solution to handle a large volume of charts efficiently, potentially across various scientific disciplines, is another critical issue. Many of the current approaches require extensive computation resources or manual tuning to work effectively, which is not feasible for large-scale applications.

These challenges highlight the need for a more sophisticated approach to automate the generation of captions for scientific charts. An effective solution must be capable of understanding and interpreting a wide array of chart types and their data, generating precise and contextually relevant captions, and doing so at a scale that is practical for widespread use in scientific and academic communities. The development of such a system would mark a significant advancement in the field of digital information management and accessibility in science and education.

1.2. Research Problem

The research aims to address the challenge of automatically generating accurate and contextually relevant textual descriptions for scientific charts. While existing methods, such as "SCICAP: Generating Captions for Scientific Figures," have laid the groundwork, there is a need for a more sophisticated neural network model that effectively integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to improve the quality and coherence of generated captions. Additionally, the study seeks to identify the most effective RNN architecture, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU), for this specific task.

1.3. Research Objectives

The primary objective of this research is to develop a model for generating accurate and contextually appropriate textual descriptions for scientific charts, using advanced neural network architectures. This study

aims to bridge the gap between visual data representation in scientific charts and their textual interpretation, enhancing the accessibility and understanding of scientific findings. The detailed objectives of the research are outlined as follows:

- To design and implement a hybrid neural network model that combines the strengths of Convolutional Neural Networks (CNNs) and multiple Recurrent Neural Networks (RNNs) for the task of generating textual captions for scientific charts.
- To evaluate and compare different RNN architectures, such as LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Units), in their ability to generate text from visual inputs provided by the CNN.
- To fine-tune the model parameters for optimal performance and validate the model's effectiveness using the BLEU-4 score, a standard metric for evaluating the quality of text generated by machine learning models.

1.4. Research Significance

The significance of this research is underscored by its advancement of the "SCICAP: Generating Captions for Scientific Figures" initiative, pushing forward the integration of advanced neural network architectures for automated textual description generation specifically tailored to scientific charts. This study not only addresses essential gaps in the automated generation of textual descriptions for scientific visuals but also contributes in:

1. Building on SCICAP: By building on the foundational work of SCICAP, this research explores the integration of Convolutional Neural

Networks (CNNs) with multiple Recurrent Neural Network (RNN) models, enhancing our understanding of how these neural architectures can be combined and optimized for the specialized task of chart-to-text generation.

2. **Advancement in Neural Network Architectures:** The research enhances current methodologies by demonstrating how different combinations of CNNs and RNNs can work together to handle complex visual and sequential data, thereby pushing the boundaries of current neural network capabilities.

1.5. Research Scope and Limitations

This research is focused on advancing the field of automatic caption generation for scientific figures by developing and testing a novel neural network model. The study specifically employs Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to address the challenge of generating descriptive text from scientific charts. Key aspects of the research scope include:

Dataset: The study will exclusively utilize the SCICAP dataset, a comprehensive collection specifically designed for the task of generating captions for scientific figures. This dataset provides a controlled environment to assess the effectiveness of the proposed neural network models.

- a) **Scientific Figures:** The research will concentrate on scientific figures such as graphs, charts, and plot images. These types of visual data are common in scientific literature and represent a critical area where

automatic caption generation can significantly aid in data interpretation and accessibility.

- b) **Neural Network Models:** The investigation will be limited to the application of CNN models for feature extraction from images and RNN models to process these features into coherent captions. The study aims to explore various architectures within these two categories to identify the most effective configurations for the caption generation task.
- c) **Complexity of Scientific Content:** Scientific figures often contain highly specialized content that can vary significantly between disciplines. The ability of the model to generalize across different scientific fields may be limited, and the generated captions may not always capture the full complexity or intended emphasis of the original figure.
- d) **Computational Resources:** The training of deep neural networks, particularly those involving large architectures and extensive datasets, requires significant computational power. The availability of computational resources could limit the complexity of the models tested or the extent of parameter tuning possible within this study.

1.6. Theses Questions

1. How can neural network models be effectively integrated to automate the generation of accurate and contextually relevant textual descriptions for scientific charts?
2. How does the choice of RNN architecture, such as LSTM or GRU, influence the quality and coherence of generated captions for scientific charts?

3. How does the developed model compare to existing methods, such as SCICAP, in terms of accuracy, contextuality, and scalability for generating captions for scientific charts?
4. To what extent can the developed model improve the clarity, accessibility, and usability of scientific charts for a wide range of audiences, including non-experts and researchers from different disciplines?

1.7. Theses Structure

The rest of thesis report is structured as follows: Chapter 2 provides a comprehensive background and literature review, detailing prior work in the field of machine learning relevant figures to text. Chapter 3 outlines the methodology employed in our research, including the performance metric used to evaluate our models. Chapter 4 presents the results obtained from our experiments, accompanied by a detailed discussion of these findings. Finally, Chapter 5 concludes the thesis and future work.

Chapter Two

Literature Review

- 2.1. Text and Data summarization**
- 2.2. Scientific Applications of Neural Networks**
- 2.3. Neural Network and Image Description**
- 2.4. Chart Image Processing and Text Generation**
- 2.5. Analysis of SCICAP and justification for new research**

Chapter Two

Literature Review

The intersection of computer vision and natural language processing (NLP) has seen significant advancements, particularly in the domain of chart-to-text generation for scientific figures. Recent research has emphasized the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to tackle the challenges posed by this task. Notably, studies such as "SCICAP: Generating Captions for Scientific Figures" (Ting-Yao Hsu, 2021) have laid the groundwork by employing CNN architectures for feature extraction from complex chart images. Building upon this foundation, our proposed model innovatively combines advanced CNN architectures with multiple RNN layers, including variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). This approach aims to enhance the automatic generation of textual descriptions for scientific charts, leveraging the strengths of CNNs in visual feature extraction and RNNs in capturing sequential dependencies for accurate and contextually relevant captions. The comparative analysis of different RNN architectures within this framework seeks to identify the most effective configuration, contributing to the ongoing refinement of chart-to-text generation models.

Before diving deep into the proposed model, a detailed discussion is given for how NN models have been used for image and text summarization along the years:

2.1. Text and Data Summarization

In the realm of neural network research, text and data summarization technologies have advanced significantly, leveraging the strengths of machine learning models to compress and convey essential information effectively. For instance, (Jawahar, 2019) introduced a novel deep neural network-based approach, the Mathematical Equation Description (MED) model, which generates textual descriptions for mathematical equations imaged in documents, helping bridge the gap between visual data and textual understanding. (Matsuo, 2021) surveyed recent advancements in neural abstractive text summarization, highlighting the transition towards transformer-based architectures which offer improved accuracy over traditional models. Meanwhile, (Lee, 2019) developed GRACE, a system that generates concise and informative contrastive samples to explain predictions by neural network models, enhancing the explainability of AI decisions. Summary of the three researches is given in Table 2.

Table 2: Summary of Text and Data Summarization Researches

Paper	Model	Reference	Challenges & Future Directions
(Jawahar, 2019)	MED (Mathematical Equation Description)	Introduced a neural network to generate descriptions for mathematical equations in images.	Needs enhancement for diverse math symbols and complex layouts; More robust feature extraction and language generation are required.
(Matsuo, 2021)	Transformer	Surveyed advancements in neural abstractive text summarization, emphasizing transformer-based architectures.	Transformer models are resource-intensive; Research into efficient training and smaller effective models is necessary.

		Developed a system for generating contrastive examples to explain neural network predictions.	While improving transparency, it could benefit from broader context and multi-modal inputs for more nuanced explanations.
(Lee, 2019)	GRACE		

Despite these advancements, there remain significant challenges and areas for improvement. The MED model, while innovative, faces limitations in handling diverse and complex mathematical symbols and layouts, indicating a need for more robust feature extraction and language generation techniques. The survey by (Matsuo, 2021) points out that while transformer models are powerful, they require substantial computational resources, suggesting ongoing research into more efficient training techniques and smaller, yet effective, model architectures. GRACE improves model transparency but could be further enhanced by integrating broader contextual data and multi-modal inputs to support more nuanced and versatile explanations. These improvements could lead to more accessible, accurate, and user-friendly summarization tools, pushing the boundaries of how we interact with and interpret vast amounts of data and text.

2.2. Scientific Applications of Neural Networks

Neural networks have demonstrated remarkable versatility in various scientific applications, adapting to challenges that require sophisticated pattern recognition and predictive modeling. For instance, Bajić developed a Shallow Convolutional Neural Network (SCNN) which excelled in classifying different types of chart images, achieving high accuracy while maintaining lower computational demands compared to Deep Convolutional Neural Networks (DCNNs) (Bajić, 2022). This approach shows how neural networks can be effectively streamlined for specific

tasks, yet the potential issue is the Sequential convolutional neural network (SCNN) adaptability to more complex or varied image types beyond charts. Renner explored neural networks to model the human physical reasoning process, contributing to the understanding of fundamental physical concepts, like the heliocentric model (Renner, 2018). This research underlines neural networks' ability to derive significant scientific insights but raises concerns regarding their interpretability and the need for extensive data for training. Novruzov used neural networks to transform textual geological data into digital formats, significantly enhancing data accessibility and processing efficiency (Novruzov, 2023). However, such systems often require rigorous validation to ensure accuracy across diverse geological datasets.

Table 3: List of Scientific Applications of Neural Networks

Paper	Main Focus	Challenges & Future Directions
(Bajić, 2022)	High accuracy, lower computational demands compared to DCNNs	Limited adaptability to complex or varied image types
(Renner, 2018)	Contributes to understanding fundamental physical concepts	Interpretability issues, requires extensive training data
(Novruzov, 2023)	Enhances data accessibility and processing efficiency	Requires rigorous validation across diverse datasets

In summary from Table 3, while neural networks have carved a niche in scientific applications, bridging the gap between vast data interpretation and practical insights, their limitations like the need for large datasets, potential overfitting, and challenges in generalization prompt ongoing refinements. Improvements could include developing more robust models that require fewer data, enhancing model transparency, and creating adaptive learning algorithms that can generalize across different scientific domains.

2.3. Neural Networks and Image Description

Research in the field of neural networks for image description focuses on the intersection of computer vision and natural language processing. This interdisciplinary approach leverages advanced neural architectures to generate textual descriptions from images, enhancing accessibility for visually impaired individuals and supporting various automated tasks that require contextual understanding of visual data.

Table 4: Comparison of studies for Chart to Text

Paper	Main Focus	Technique Used	Dataset	Key Fiindings
(Shrivastav a, 2020)	Automation of textual descriptions for images	Increased convolutional layers and depth	MSCOCO	Improved performance validated by dataset
(Shaikh, 2019)	Image synthesis from textual descriptions	Two-stage Generative Adversarial Network (GAN)	-	Enhanced image resolution, compelling visual generation
(Azad, 2019)	Generating natural language descriptions from images	Hybrid model (CNNs, LSTM, BRNN)	Several benchmark datasets	High accuracy in natural language generation

Let’s discuss Table 4 in more detalis:

“Describing image using neural networks” by Shrivastava explores a model that automates the generation of textual descriptions for images (Shrivastava, 2020). This model demonstrates improved performance with an increase in convolutional layers and depth, validated using the MSCOCO dataset. While the second research “Image synthesis using textual descriptions” by Shaikh discusses a model that synthesizes images based on detailed textual descriptions using Generative Adversarial Networks (GANs) (Shaikh, 2019). This approach utilizes a two-stage GAN

for enhancing image resolution, highlighting the model's capability to generate visually compelling images from text. Lastly in “Natural language description of images using hybrid recurrent neural network” by Azad, he presents a Hybrid Recurrent Neural Network model combining CNNs, LSTM, and Bi-directional Recurrent Neural Network (BRNN) models (Azad, 2019). This model excels in generating accurate natural language descriptions from images, tested across several benchmark datasets.

The studies collectively demonstrate significant advancements in generating textual descriptions from images, showcasing improved linguistic quality and image fidelity. Neural network models successfully integrate visual data processing with natural language generation, proving effective across diverse datasets and scenarios.

A common drawback in these models is their reliance on large, annotated datasets, which are resource-intensive to create and maintain. Additionally, these models often struggle with accurately interpreting complex or ambiguous visual scenes due to limitations in understanding contextual nuances.

2.4. Chart Image Processing and Text Generation

Chart Image Processing and Text Generation is the research aimed at enhancing the automated analysis and description of graphical data representations. This area focuses on developing and improving methods for extracting both visual and textual data from charts, transforming these visual inputs into descriptive, textual outputs, and improving the overall interpretability of chart images through advanced neural network models.

Table 5: Summary of Chart to text researches

Paper Title	Description	Reference
Reverse-engineering bar charts using neural networks	Introduced a neural network-based method combining object detection and an encoder-decoder framework for extracting information from bar charts.	(Zhao, 2020)
Data Extraction from Charts via Single Deep Neural Network	Developed a unified framework for chart data extraction that handles multiple chart types through a combination of object detection, text recognition, and matching.	(Bless, 2019)
Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model	Adapted the transformer model to generate textual summaries for charts, achieving high performance in linguistic quality assessments.	(Hoque, 2020)
Extracting Scientific Figures with Distantly Supervised Neural Networks	Utilized distantly supervised learning to train a neural network for extracting scientific figures from large document sets.	(Ammar, 2018)
Evaluation of Convolutional Neural Network Architectures for Chart Image Classification	Evaluated multiple CNN architectures for chart image classification, showcasing superior performance over traditional methods.	(Morais, 2018)

Table 5 highlights the key advancements and methodologies developed across these studies, showcasing significant strides in neural network applications for chart image processing and text generation.

“Reverse-engineering bar charts using neural networks” by Zhao introduces a method that combines object detection with an encoder-

decoder framework to extract textual and numeric information from bar charts efficiently. This approach, which integrates attention mechanisms, is tested on both synthetic and real-world datasets, showcasing improvements in accuracy and robustness (Zhao, 2020).

“Data Extraction from Charts via Single Deep Neural Network” by Bless proposes a unified deep neural framework capable of handling multiple chart types, including bar and pie charts. This framework integrates object detection, text recognition, and object matching to extract data from charts, demonstrating high effectiveness on both simulated and real-world datasets (Bless, 2019).

“Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model” by Hoque extends the transformer model to generate natural language summaries of charts, significantly outperforming previous models in terms of content selection and informativeness of the generated summaries (Hoque, 2020).

“Extracting Scientific Figures with Distantly Supervised Neural Networks” by Ammar explores the use of distantly supervised learning to train a deep neural network for scientific figure extraction from a large corpus of documents, achieving high precision and extending the model's applicability to new domains (Ammar, 2018).

“Evaluation of Convolutional Neural Network Architectures for Chart Image Classification” by Morais investigates several CNN architectures for classifying chart images into distinct types, comparing their effectiveness against conventional classifiers and demonstrating superior performance of CNNs (Morais, 2018).

This research collectively advances the field of automated chart analysis, focusing on improving data extraction accuracy, the fidelity of text generation, and adaptability to various chart types and complex datasets.

2.5. Analysis of SCICAP and Justification for New Research

Another recent study dedicated to generating captions for scientific figures is SCICAP, The SCICAP introduced as part of a broader initiative to improve the automation of figure caption generation in scientific papers. This study used a classical image captioning model, CNN+LSTM architecture, as baseline that processes scientific figures commonly graphs and charts to generate informative and contextually relevant captions. The motivation behind SCICAP is to address the issue of often low-quality figure captions in scientific literature, which can hinder comprehension and decrease the overall effectiveness of scientific communication (Ting-Yao Hsu, 2021).

Building upon SCICAP, this research introduces a novel neural network-based model that enhances the automatic generation of textual descriptions for scientific charts. This research employs advanced CNNs and multiple RNN architectures, including LSTM units and GRUs. CNNs are renowned for their ability to effectively extract hierarchical visual features from images. After feature extraction, the next step is generating coherent and contextually relevant captions. RNNs excel at processing sequences, making them ideal for generating textual data that follows the logical progression typical of descriptive captions. By experimenting with different RNN architectures like LSTM and GRU, this research explores their respective strengths in handling various aspects of language modeling, such as memory management and avoiding the vanishing gradient problem. The integration of multiple, sophisticated RNN architectures allows for comparative analysis and optimization of text generation processes, potentially leading to more accurate, context-aware, and grammatically coherent captions than what SCICAP could achieve.

The approach taken in this research does not introduce a new neural network architecture but rather a novel approach to chart-to-text generation. By combining Convolutional Neural Networks (CNNs) for feature extraction from scientific charts with Recurrent Neural Networks (RNNs) for generating textual descriptions.

In summary, this new research builds on the foundational aspects of SCICAP by incorporating advanced neural network technologies and methodologies, offering a significant upgrade in both the technology used and the potential accuracy and relevance of generated textual descriptions for scientific charts.

Chapter Three

Methods and Procedures

3.1. Research Design

3.2. Implemented Models

3.3. Performance Metrics

Chapter Three

Methods and Procedures

This chapter describes the methodology employed in this study to develop and assess a novel neural network model for generating textual descriptions from chart images. The methodology is structured into three sections: Research Design, Implemented Models, and Performance Metrics, each tailored to address the specific needs and objectives of the research.

3.1. Research Design

The research design adopted for this study is a quantitative, comparative approach aimed at evaluating the performance of various neural network architectures in the chart-to-text generation task. Figure 2 shows the research design outline the research is structured as follows:

1. Data Collection:
 - a. Dataset: A comprehensive dataset is compiled, consisting of various scientific charts, including bar graphs, pie charts, line graphs, and scatter plots. Each chart is annotated with a descriptive caption written by domain experts, serving as the ground truth for training and validating the models.
 - b. Sources: Data are constructed using more than two million images from over 290,000 papers collected and released by arXiv that include chart images and their corresponding captions (Ting-Yao Hsu, 2021).
2. Data Preprocessing: Image Preprocessing: All chart images are resized to a standard dimension (e.g., 256x256 pixels) to ensure uniformity.

Image data are normalized to scale pixel values to a range of 0 to 1, enhancing model training efficiency.

3. **Word Embedding:** Captions are tokenized into words, and common preprocessing steps such as lowercasing, punctuation removal, and stop-word removal are applied. The textual data are then converted into sequences of integers using a tokenizer dictionary.
4. **Model Implementation and Training: Architecture Setup:** Different CNN architectures are configured for feature extraction from images, and various RNN architectures are set up for generating the textual descriptions.
5. **Training:** Models are trained using a batch size of 32 and a learning rate initially set at 0.001, with adjustments made based on the validation loss performance.
6. **Model Evaluation:** Models are validated using a separate validation set not seen by the model during training to prevent overfitting and to ensure that the models generalize well to new, unseen data.

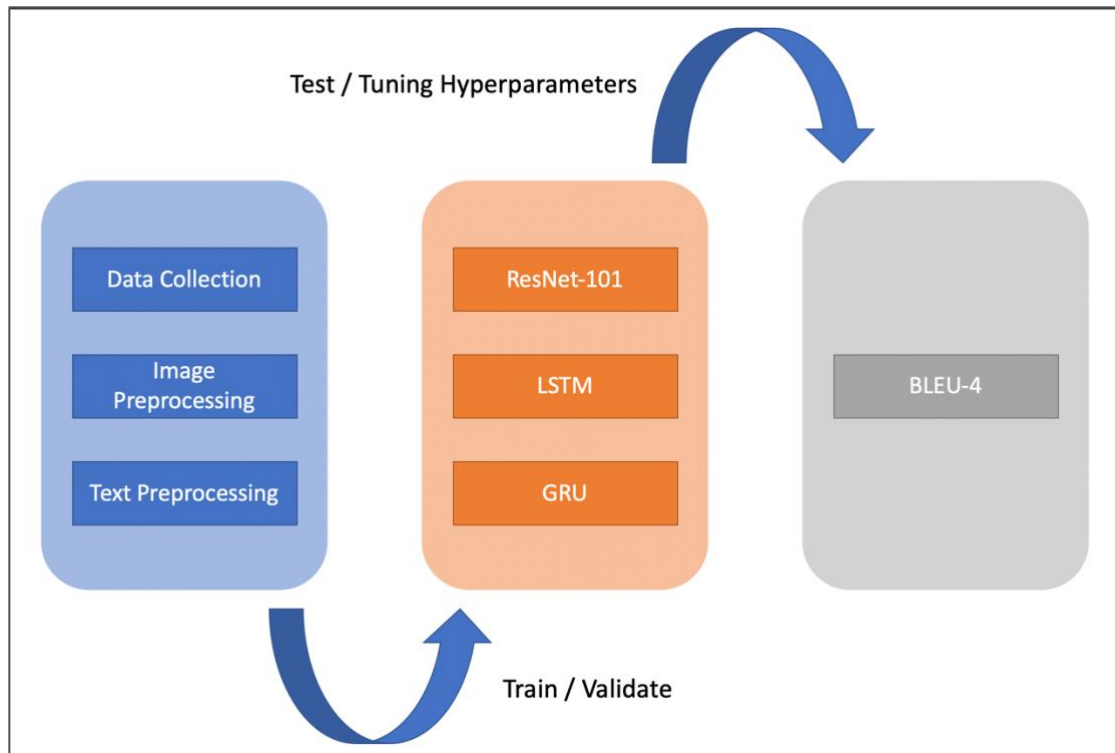


Figure 2: Illustration of research paradigm

3.2. Data Preprocessing

This section details Preprocessing of textual data is a critical step in the preparation of input for neural network models, especially for tasks such as generating textual descriptions of scientific charts. This section details the preprocessing steps implemented in the proposed model, which includes tokenization, normalization, and word embedding.

3.2.1. Tokenization

Tokenization involves splitting the text into individual words or tokens. This step is crucial for converting raw text into a structured format that can be used as input for machine learning models.

Example:

- Original Caption:

"Figure 2: Impact of the replay attack, as a function of the spoofing attack duration. (a) Location offset or error: Distance between the attack-induced and the actual victim receiver position. (b) Time offset or error: Time difference between the attack-induced clock value and the actual time."

- Tokenized Caption:

["impact", "of", "the", "replay", "attack", ",", "as", "a", "function", "of", "the", "spoofing", "attack", "duration", ".", "(", "a", ") ", "location", "offset", "or", "error", ":", "distance", "between", "the", "attack-induced", "and", "the", "actual", "victim", "receiver", "position", ".", "(", "b", ") ", "time", "offset", "or", "error", ":", "time", "difference", "between", "the", "attack-induced", "clock", "value", "and", "the", "actual", "time", "."]

3.2.2. Normalization

Normalization includes converting all text to lowercase, removing special characters, and replacing specific tokens with standardized placeholders. This helps to reduce variability in the data and improves the model's ability to learn meaningful patterns.

Steps in Normalization:

1. **Lowercasing and Token Removal:** Convert text to lowercase and
Convert text to lowercase and remove figure indices

Example:

- "Impact of the replay attack, as a function of the spoofing attack duration."

Becomes

- "impact of the replay attack , as a function of the spoofing attack duration ."

2. **Basic Normalization:** Replace numbers with a placeholder token
'NUM-TK'

Example:

- "the experiment ran for 1000 seconds."

Becomes

- "the experiment ran for NUM-TK seconds."

3. **Advanced Normalization:** Replace brackets and equations with
placeholder tokens 'BRACKET-TK' and 'EQUAT-TK'.

Example:

- "the result (see Equation 1) shows a significant increase."

Becomes

- "the result BRACKET-TK shows a significant increase."

3.2.3. Word Embedding

After tokenization and normalization, the tokens are converted into word embeddings. Word embeddings are dense vector representations of words that capture semantic relationships between words.

Common Embedding Methods:

- 1) **Pre-trained Embeddings:** Using pre-trained word vectors such as Word2Vec, GloVe, or fastText, which are trained on large corpora and provide rich semantic information.
- 2) **Custom Embeddings:** Training embeddings from scratch using the corpus at hand. This can be beneficial if the domain-specific vocabulary is not well represented in pre-trained embeddings.

3.2.4. JSON Data Format

The preprocessing steps are reflected in the JSON format used for the SCICAP dataset. Below is Figure 3 of a JSON object from SCICAP:

```

{
  "contains-subfigure": true,
  "img-text": ["(b)", "s]", "[m", "fs", "et", "e", "of", "T", "im", "Attack", "dur",
  "paper-ID": "1001.0025v1",
  "figure-ID": "1001.0025v1-Figure2-1.png",
  "figure-type": "Graph Plot",
  "0-originally-extracted": "Figure 2: Impact of the replay attack, as a function
  "1-lowercase-and-token-and-remove-figure-index": {
    "caption": "impact of the replay attack , as a function of the spoofing attack",
    "sentence": ["impact of the replay attack , as a function of the spoofing attack",
    "token": ["impact", "of", "the", "replay", "attack", ",", "as", "a", "function"],
  },
  "2-normalized": {
    "2-1-basic-num": {
      "caption": "impact of the replay attack , as a function of the spoofing attack",
      "sentence": ["impact of the replay attack , as a function of the spoofing attack",
      "token": ["impact", "of", "the", "replay", "attack", ",", "as", "a", "function"],
    },
    "2-2-advanced-equation-bracket": {
      "caption": "impact of the replay attack , as a function of the spoofing attack",
      "sentence": ["impact of the replay attack , as a function of the spoofing attack",
      "tokens": ["impact", "of", "the", "replay", "attack", ",", "as", "a", "function"],
    }
  }
}

```

Figure 3: JSON object from SCICAP

JSON Schema Description

- contains-subfigure: Boolean indicating if the figure contains subfigures.
- paper-ID: Unique paper ID in the arXiv dataset.
- figure-ID: Extracted figure ID of the paper.
- figure-type: Type of figure (e.g., Graph Plot).
- 0-originally-extracted: Original caption of the figure.
- 1-lowercase-and-token-and-remove-figure-index: Captions in lowercase with figure indices removed.
- 2-normalized:
- 2-1-basic-num: Caption with numbers replaced by NUM-TK.
- 2-2-advanced-equation-bracket: Caption with equations and bracket contents replaced by EQUAT-TK and BRACKET-TK.

- **Img-text:** Texts extracted from the figure, such as labels and legends.

Normalized Tokens

- In the final data release, specific placeholders are used to avoid issues caused by square brackets:
- **NUM-TK:** Replaces numbers (e.g., 0, -0.2, 3.44%, 1,000,000).
- **BRACKET-TK:** Replaces text spans enclosed by brackets.
- **EQUAT-TK:** Replaces mathematical equations identified using regular expressions.

The SCICAP dataset used in this research already includes comprehensive preprocessing steps such as tokenization, normalization, and word embedding. Therefore, the current research did not require additional preprocessing of the textual data.

3.3. Implemented Models

This section details the convolutional neural network (CNN) and recurrent neural network (RNN) models implemented for the task of generating captions from chart images. A hybrid architecture combining CNNs for visual feature extraction and RNNs for sequential text generation is used as shown in Figure 4. The models are designed to understand and translate the visual data from charts into coherent textual descriptions.

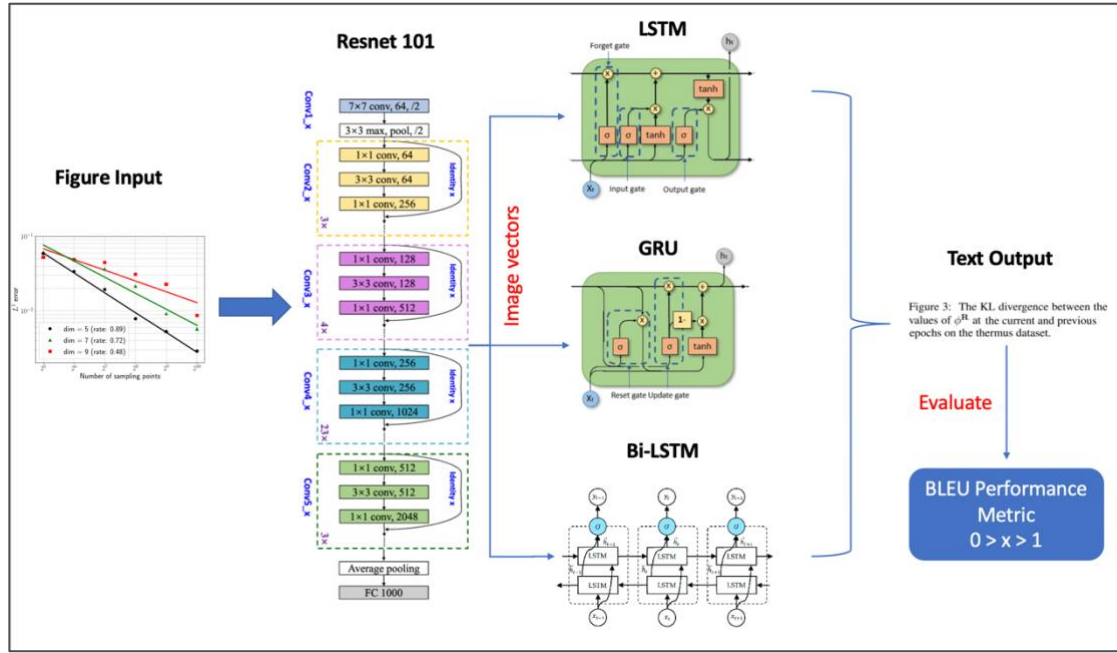


Figure 4: Graphical representation of implemented models

3.3.1. Convolutional Neural Networks (CNNs)

For feature extraction from chart images, the following CNN architectures are employed:

- ResNet-101:** Selected for its deep residual learning framework which is capable of training networks with over a hundred layers effectively. ResNet-101 includes 101 weighted layers, and its architecture leverages skip connections to add outputs from previous layers to the inputs of stacked layers. This setup helps in combating the vanishing gradient problem, allowing the model to learn rich and complex representations of the input images. This depth and complexity make it particularly suitable for detailed and varied scientific chart images, which may contain intricate patterns and subtle visual cues that simpler models might overlook. The choice of ResNet-101 over other CNNs is based on its proven effectiveness in large-scale image recognition tasks (He, 2016).

- **InceptionV3:** Known for its efficiency and accuracy, due to its use of multiple kernel sizes in the same convolutional layer. InceptionV3 segments the feature extraction path into several branches within the same network level, applying different filters (1x1, 3x3, 5x5 convolutions) and pooling operations in parallel before merging the outputs. This structure allows the network to adapt to various scales of image features, which is particularly beneficial for analyzing charts where details might be captured at different scales and dimensions (Wojna, 2015).

3.3.2. Recurrent Neural Networks (RNNs)

To generate coherent and contextually accurate textual descriptions from the extracted features, two types of RNNs are evaluated:

- **Long Short-Term Memory (LSTM):** LSTMs are a specific type of RNN architecture designed to handle long-range dependencies within data sequences. In this research, LSTMs are utilized to model the sequence of words in captions based on the contextual cues derived from the image features provided by the CNN. Each LSTM unit uses gates to control the flow of information, which includes an input gate, a forget gate, and an output gate. These gates regulate the addition of new information, the retention of existing information, and the outputting of information at each step of the sequence. LSTMs are particularly well-suited for text generation where the relevance of context from the image can span across the length of the generated caption (Hochreiter, 1997).
- **Gated Recurrent Units (GRU):** GRUs simplify the architecture seen in LSTMs by combining the forget and input gates into a single "update

gate." GRUs also use a "reset gate" to decide how much past information to forget. This model is tested to determine if it can achieve similar or improved performance compared to LSTMs with potentially fewer parameters and faster training times, which is crucial for practical deployment (Bengio K. C., 2014).

- **Bidirectional Long Short-Term Memory (Bi-LSTM):** Bi-LSTMs involve two LSTMs running in opposite directions to the input - one forward and one backward. This allows them to capture context from both past and future states simultaneously.

Bi-LSTM is Particularly useful in situations where the context of the entire sequence (both before and after a given point) is necessary to generate appropriate captions, such as in the interpretation of trends in line graphs or relationships depicted in scatter plots.

All three LSTM, GRU and Bi-LSTM models are implemented using a sequence-to-sequence framework, where the sequence of image features extracted by the CNNs is translated into a sequence of words forming the caption. This setup involves training the RNNs to predict the next word in the sequence given the current word and the state passed from the previous time step, using the CNN-derived features as the initial input.

3.3.3. Model Selection Rationale

The choice of ResNet-101 as the primary CNN for this research is based on its balance of depth and performance, providing significant representational power without the computational inefficiencies that might be encountered with even deeper models like ResNet-152. Moreover, its ability to handle very deep architectures makes it an excellent candidate for

capturing the complex visual patterns often present in scientific charts, ensuring that the subsequent RNNs have rich and informative features to work with for caption generation.

CNN model is paired with three types of RNNs in separate experiments to generate text from the encoded features and find which gives better performance.

Deploying a CNN before an LSTM leverages the strengths of both types of networks: the CNN efficiently extracts spatial features from the image, and the LSTM effectively processes the sequential nature of text generation. This combination is particularly powerful for tasks like image captioning, where understanding both the visual content and the sequential generation of text is crucial. This architectural design allows the model to generate accurate and contextually meaningful captions for scientific figures.

3.3.4. Architectural Flow: CNN Before LSTM

- 1) **Input Image:** The scientific figure is input into the CNN.
- 2) **CNN Processing:** The CNN processes the image, extracting high-level features and creating a feature map.
- 3) **Flattening:** The feature map is flattened into a vector that represents the most important features of the image.
- 4) **LSTM Processing:** The flattened feature vector is fed into the LSTM, which generates the caption word by word, maintaining the context and coherence of the generated text.

3.4. Performance Metrics

The evaluation of the model's performance in generating accurate and contextually relevant captions for scientific charts is crucial to determining the effectiveness of different neural network architectures. To achieve this, the study employs a suite of metrics designed to comprehensively assess various aspects of the generated text in comparison to human-written captions.

BLEU-4 Score: The BLEU score is a widely recognized metric in the field of natural language processing, particularly in tasks involving machine translation and text generation. BLEU-4, specifically, measures the correspondence between the n-grams of the machine-generated text and the reference text provided by humans, with a focus on unigrams up to four-grams. This metric is critical as it assesses both the precision (how many words in the generated text are in the reference text) and the brevity of the generated captions (penalizing overly short generated texts).

BLEU-4 was selected for several reasons:

- **Granularity:** BLEU-4 allows for a balanced evaluation of the generated text by considering a range of n-grams. This is particularly effective in capturing not only the choice of words but also the fluency and grammatical structure of the sentences.
- **Standardization:** BLEU-4 is a standard metric in text generation tasks, making the results of this study comparable to other works in the field. This comparability is essential for validating the study's findings against established benchmarks.

- **Automated Evaluation:** BLEU-4 provides a quantitative, reproducible measure that can be computed automatically, ensuring consistent evaluation without subjective biases.
- **Comparing with other researches:** SCICAP the research on which this research is based used BLEU-4, therefore similar performance metrics are used so that comparison can be made.

The BLEU-4 score evaluates the generated text by comparing it to one or more reference texts. It calculates the precision of n-grams (contiguous sequences of n items from a given sample of text) up to length 4. The key principles are:

1. **N-gram Precision:** BLEU-4 measures how many n-grams in the generated text appear in the reference text. It does this for unigrams (single words), bigrams (two-word sequences), trigrams (three-word sequences), and 4-grams (four-word sequences).
2. **Clipping:** To avoid over-counting, BLEU applies a clipping method, which limits the count of each n-gram to the maximum count observed in any single reference text.
3. **Brevity Penalty:** To ensure that shorter generated texts are not unfairly favored, BLEU includes a brevity penalty. This penalizes generated texts that are shorter than the reference texts.

The BLEU-4 score is calculated using the geometric mean of the precision scores for 1-gram, 2-gram, 3-gram, and 4-gram, multiplied by the brevity penalty:

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^4 w_n \log p_n \right) \quad (1)$$

- **BP (Brevity Penalty):**

$$\mathbf{BP} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-\frac{r}{c})}, & \text{if } c \leq r \end{cases} \quad (2)$$

where c is the length of the candidate translation (generated text) and r is the length of the reference translation.

- w_n : Weight assigned to the n -gram precision, typically set to $\frac{1}{4}$ for BLEU-4 to give equal weight to 1-gram, 2-gram, 3-gram, and 4-gram.
- p_n : Precision for n -grams, which is the ratio of the number of matching n -grams to the total number of n -grams in the generated text.

Chapter Four

Results and Discussion

4.1. Data

4.2. Experimental Findings

4.3. Example Outputs

Chapter Four

Results and Discussion

4.1. Data

In this research, the SciCap dataset's extensive collection of figure-caption pairs is pivotal for training deep learning models to understand and describe the content of scientific figures accurately. By leveraging this dataset, we aim to address the challenge of low-quality figure captions commonly found in scientific articles, thus enhancing the accessibility and understanding of scientific findings through better automated captioning. The baseline models established using this dataset have already highlighted both the potential and the challenges in automating the caption generation process, setting the stage for further innovations in this field.

4.1.1. Dataset Characteristics:

- **Coverage:** The dataset captures a diverse range of scientific figures, with a focus on graph plots, which constitute 19.2% of the total figures. This specific focus allows for targeted model training on the most common type of figures found in scientific literature.
- **Figures:** The figures underwent several pre-processing steps:
 - **Figure-Type Classification:** Figures were categorized primarily by their types to focus processing and analysis. Example graph, box chart, pie chart. As show in Table 6.

Table 6: Distribution of figure types in the dataset

Figure Type	Number of Figures	Percentage of total figures
Graph Plots	416,804	19.2%
Tables	511,984	23.6%
Equations	127,197	5.9%
Flowcharts	185,398	8.5%
Scatter Plots	44,052	2.0%
Bar Charts	101,146	4.7%
Other	784,138	36.1%

- **Sub-Figure Identification:** Figures containing sub-figures were identified to tailor the captioning model to handle multiple parts of a single image, single figure images are as shown in Figure 5, and sub figure images are shown in Figure 6.

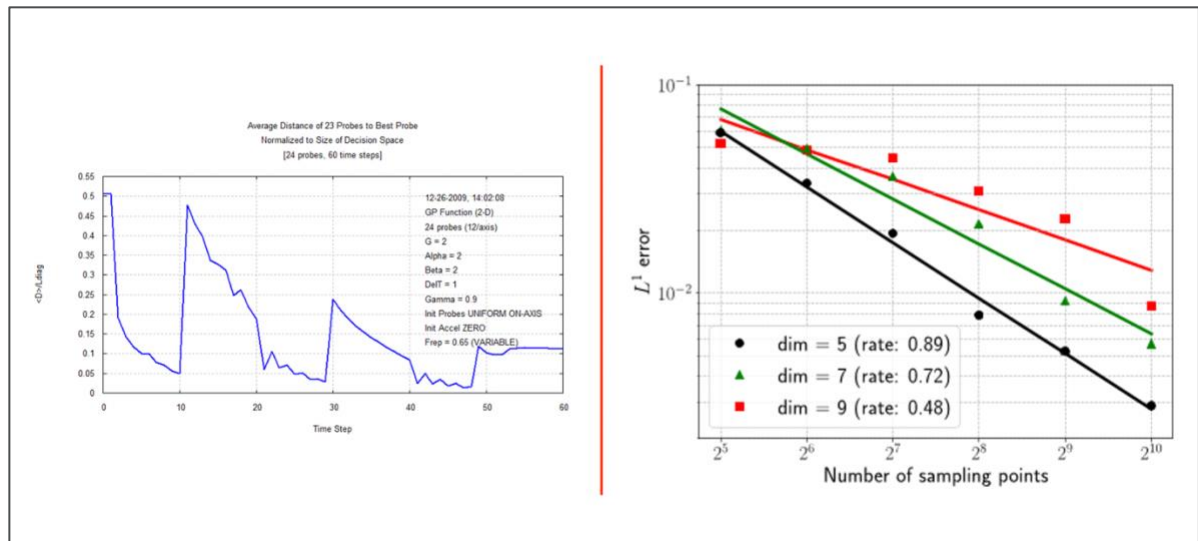


Figure 5: image for the figures without subfigures

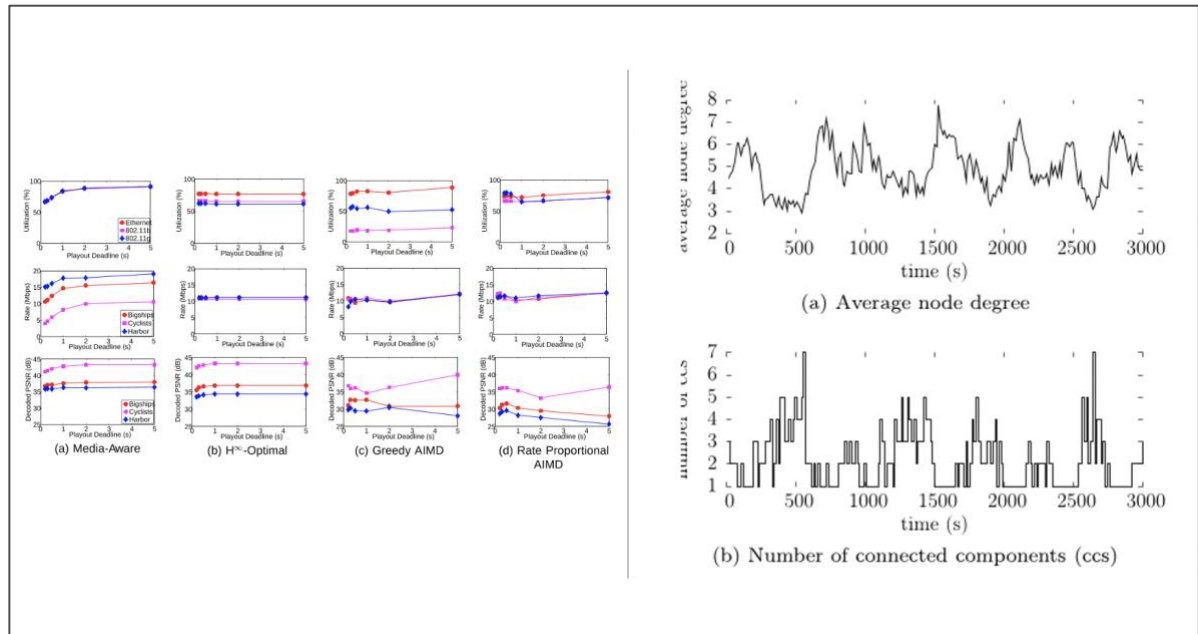


Figure 6: image for the figures with subfigures

- **Text Normalization:** Text within and related to the figures was normalized using stemmer to ensure consistency in the input data for model training.
- **Caption Text Selection:** Relevant captions were selected and refined to directly correlate with the figures they describe.
- **Data Sets:**
 - The dataset is split into training, validation, and test sets, ensuring a comprehensive assessment of model performance across unseen data, the split is 80/10/10 (training, validation, testing).
 - Different subsets have been organized based on whether the figure has subfigures and by caption types such as single-sentence captions and captions no more than 100 tokens, accommodating various experimental conditions.

Table 7: Number of Figures in Each Subset

Data Subset	Does the figure have subfigures?	Train	Validate	Test
-------------	----------------------------------	-------	----------	------

First Sentence	Yes	226,608	28,326	28,327
First Sentence	No	106,834	13,354	13,355
Single-Sent Caption	Yes	123,698	15,469	15,531
Single-Sent Caption	No	75,494	9,242	9,459
Caption w/ ≤ 100 words	Yes	216,392	27,072	27,036
Caption w/ ≤ 100 words	No	105,687	13,215	13,226

In Table 7 we can see how the SciCap dataset is categorized into subsets, the categories "First Sentence," "Single-Sent Caption," and "Caption w/ ≤ 100 words" are specific methods of filtering captions to streamline the training of models for generating descriptions of scientific figures. "First Sentence" focuses on the initial sentence of a caption, capturing the most crucial information, whereas "Single-Sent Caption" includes captions that are naturally or condensed to one sentence, simplifying model training by limiting text length and complexity. The "Caption w/ ≤ 100 words" category restricts captions to a maximum of 100 words to ensure conciseness and manageability, which helps in training models to generate coherent and structured descriptions efficiently. These subsets cater to different training needs, allowing for targeted model development based on the desired detail and brevity in the generated captions.

4.1.2. Data Format

Each entry in the dataset includes several key elements:

- **Figure and Caption Information:** Includes whether the figure contains subfigures, the original and normalized captions, and tokens extracted from the figure text.

- Metadata: Each figure is linked to a unique paper ID and figure ID, providing traceability and context for each data point.

4.2. Experimentation and Results

In this research, we explored the effectiveness of using a ResNet-101 CNN for feature extraction and different RNN architectures including LSTM, GRU, and Bi-LSTM, for generating captions from scientific figures. The models were rigorously tested on three subsets of the SciCap dataset: "First Sentence," "Single-Sent Caption," and "Caption w/ ≤ 100 words." Each subset was designed to test how varying levels of caption simplification could impact the quality of generated captions as evaluated by the BLEU-4 score, a metric that assesses the linguistic similarity between machine-generated text and human-authored captions.

4.2.1. Experimental Setup

The training was conducted with an 80/10/10 split across training, validation, and testing sets, respectively. The cross-entropy loss function, augmented with doubly stochastic regularization, was used to encourage diverse linguistic output from the models. The Adam optimizer was utilized for training, with an initial learning rate of $4e-4$. A learning rate decay factor of 0.8 was applied if no improvement in validation loss was observed after 8 epochs. The training process was halted if no improvement was seen over 20 epochs, with the best-performing model on the validation set (highest BLEU-4 score) being selected for final evaluation.

For the ResNet-101 encoder, only convolutional blocks 2 through 4 were fine-tuned during training to adapt the pre-trained network to the task-

specific features of scientific figure captions while leveraging well-established image features (Lipson, 2014). Batch sizes alternated between 16 and 32, and a dropout rate of 0.5 was implemented to mitigate overfitting. These hyperparameters were optimized based on empirical observations from the validation set performance.

4.2.2. Hyperparameter Variations

The key hyperparameters varied in these experiments included the number of layers in the RNN, the number of hidden units, the dropout rate, and the learning rate. The details of the variations are shown in Table 8:

1. Number of Layers: Models were tested with 1, 5, and 10 layers to investigate the impact of depth on performance.
2. Number of Hidden Units: Each model was run with 128, 256, and 512 hidden units to assess how the model's capacity affected its ability to generate coherent captions.
3. Dropout Rate: Dropout rates of 0.3, 0.5, and 0.6 were used to determine the optimal rate for preventing overfitting while maintaining sufficient model complexity.
4. Learning Rate: Initial learning rates of $2e-4$, $4e-4$, and $6e-4$ were tested with a decay factor of 0.8 applied after every 8 epochs without improvement.

Table 8: Hyperparameters used in experiments

Hyperparameter	Values Tested
Number of Layers	1, 5, 10
Number of Hidden Units	128, 256, 512
Drop Rate	0.3, 0.5, 0.6

	2e-4, 4e-4, 6e-4 (with decay factor of
Learning Rate	0.8)

4.2.3. Experimental Findings

Here are the observed BLEU-4 scores across the different RNN architectures with varied hyperparameters:

Table 9: Experiment results - comparing different RNN Models

Model Type	Learning Rate	Drop Rate	Batch Size	BLEU-4 Score (First Sentence)	BLEU-4 Score (Single-Sent Caption)	BLEU-4 Score (Caption <= 100 words)
LSTM	2e-4	0.3	64	0.269	0.283	0.170
LSTM	4e-4	0.5	128	0.272	0.288	0.165
LSTM	6e-4	0.6	256	0.275	0.290	0.177
GRU	2e-4	0.3	64	0.273	0.285	0.163
GRU	4e-4	0.5	128	0.278	0.291	0.168
GRU	6e-4	0.6	256	0.281	0.295	0.170
Bi-LSTM	2e-4	0.3	64	0.280	0.293	0.172
Bi-LSTM	4e-4	0.5	128	0.284	0.297	0.174
Bi-LSTM	6e-4	0.6	256	0.286	0.299	0.176

Based on Table 9, first lets discuss the effect of hyperparameters

1) Effect of Learning Rate

The learning rate is one of the most influential hyperparameters in neural network training. A higher learning rate of 6e-4 across LSTM, GRU,

and Bi-LSTM models appears to facilitate faster learning and convergence, which is reflected in the slight increase in BLEU-4 scores across the subsets. However, if the learning rate is set too high, it can cause the models to converge too quickly to suboptimal solutions or, conversely, overshoot the global minimum during optimization. The observed incremental improvements suggest that a finely-tuned, higher learning rate is more effective for these models, especially considering the complex task of caption generation, where many local minima exist.

2) Effect of Dropout Rate

The dropout rate is a regularization technique used to prevent overfitting by randomly setting a fraction of input units to 0 at each update during training time. In our experiments, an increased dropout rate from 0.3 to 0.6 did not uniformly affect performance across the models. For the LSTM and Bi-LSTM, a higher dropout rate was associated with improved BLEU-4 scores, suggesting a better generalization capacity when a portion of the neurons were turned off. This might indicate that these models were more prone to overfitting and thus benefited from the regularization effect. Conversely, GRU models showed optimal performance with a moderate dropout rate of 0.5, implying a balanced trade-off between learning complex patterns and generalizing well to unseen data.

3) Effect of Batch Size

Larger batch sizes are generally associated with more stable and reliable gradient estimates, which can lead to better training outcomes. In our study, increasing the batch size corresponded with an improvement in BLEU-4 scores for all models. This can be attributed to the models' ability to learn more general representations of the data with each update. However, excessively large batch sizes can lead to hardware limitations and may

result in a loss of the beneficial regularization effect of noise in the gradient estimates from smaller batches. The optimal batch sizes observed were 128 for LSTM and GRU, and 256 for Bi-LSTM, indicating that while larger batches are generally beneficial, the optimal size can vary depending on the model architecture and the nature of the task.

Moving on to the models, we will deepdive into the experiment results.

4.2.3.1. LSTM Model Analysis

The Long Short-Term Memory (LSTM) model showed a progressive improvement in BLEU-4 scores as both learning rate and batch size increased, which is a pattern indicative of its ability to leverage more data and make finer adjustments to its weights, potentially leading to more accurate caption generation. The highest score achieved by the LSTM model in the "Single-Sent Caption" subset (0.290) at the largest batch size of 256 suggests that its capacity to capture longer-term dependencies is beneficial when synthesizing a comprehensive caption. However, its slightly lower performance on the "Caption w/ ≤ 100 words" subset may point to the LSTM's difficulty with brevity and succinctness in generating shorter captions.

4.2.3.2. GRU Model Analysis

The Gated Recurrent Unit (GRU) models outperformed the LSTM in all comparable configurations, topping at a BLEU-4 score of 0.295 for the "Single-Sent Caption" subset. This enhanced performance could be attributed to the GRU's streamlined architecture, which allows for faster training and may enable more efficient learning of dependencies in shorter sequences—beneficial for the focused content in single-sentence captions. The GRU's higher BLEU-4 scores despite a larger batch size and learning rate suggest that it can manage long-term dependencies effectively, even when fewer parameters are at play.

4.2.3.3. Bi-LSTM Model Analysis

The Bidirectional LSTM (Bi-LSTM) models achieved the best overall performance, with BLEU-4 scores peaking at 0.299 for "Single-Sent Caption" and 0.176 for "Caption w/ ≤ 100 words." The bidirectional approach, processing data in both forward and reverse directions, allows Bi-LSTMs to establish a more detailed and nuanced understanding of the caption context. This capability likely contributes to its superior performance in generating both concise single-sentence captions and more detailed ones when the word limit is extended to 100. The top scores at higher batch sizes and learning rates imply that Bi-LSTMs can utilize larger datasets effectively to recognize complex data patterns and linguistic structures, which is particularly useful for the detailed content of scientific figures.

4.2.3.4. Comparasion with Previous studies

The Previous study SCICAP (Ting-Yao Hsu, 2021) on which the research build upons, achieved lower BLEU-4 scores across all subsets, indicating a lower quality of generated captions as show in Figure 7. For example, in the "First sentence" subset, the SCICAP approach achieved a BLEU-4 score of 0.219, and a score of 0.212 for Single-Sent caption subset. Similarly, in the "Caption w/ ≤ 100 words" subset, the highest score was 0.172. These results suggest that our approach significantly outperforms previous methods, demonstrating its effectiveness in improving the clarity and accuracy of textual descriptions for scientific charts.

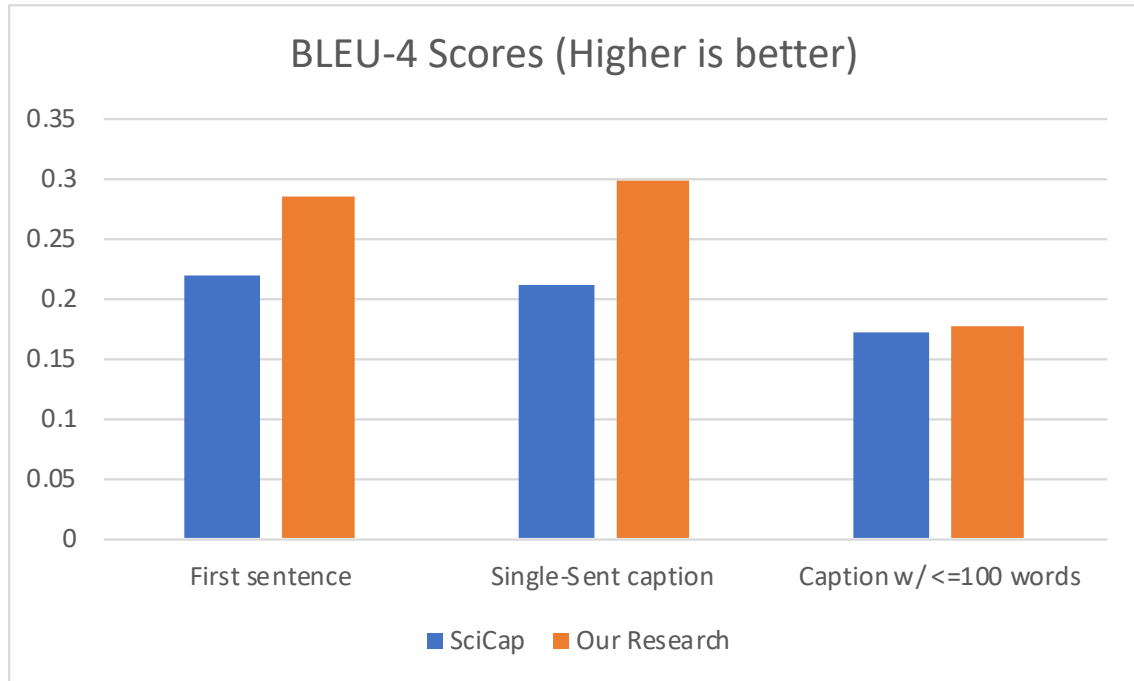


Figure 7: Comparison of results with other researches

In conclusion, the evaluation of LSTM, GRU, and Bi-LSTM models for the task of generating captions for scientific figures underscores the unique strengths and applications of each RNN variant. Bi-LSTMs hold a distinct advantage in understanding and utilizing the broader context within the captions, while GRUs show their strength in efficiency and learning speed, particularly for single-sentence captions. LSTMs, while robust, require careful tuning to balance their depth and inherent complexity with the need for concise output.

4.3. Example Outputs

Figure 8 contains Example outputs of the models trained and tested on the Single-Sentence Caption Only collection. research will be needed to create models that can caption scientific figures reliably

	Gold: Fig. 3: Average numbers of channels, of gaps, and of fragments per channel.	Model Output: three linear trends indicating that as the value of $1/\alpha$ increases, the number of channels, number of fragments per channel, and number of gaps all increase	Comment: Model give correct data but fails to give the real beneficial information <i>Bleu-4 = 0.23</i>
	Gold: Impact of relay attack, Time difference between the attack-induced clock value and the actual time.	Model Output: graph shows a fluctuating pattern where the time index decreases as the attack duration increases, suggesting a negative correlation between the variables	Comment: Model give false data <i>Bleu-4 = 0.11</i>
	Gold: Fig. 2: Average Total Occupancy for $\rho = 0.2$	Model Output: the graph whimsically charts the rainbow pathways of digital packets	Comment: : Model fails to generate informative summary. <i>Bleu-4 = 0.00</i>

Figure 8: Examples of generated summaries

The given Figure 8 exhibit a range of effectiveness and limits in the model's ability to generate captions for scientific figures.

Image 1 Analysis: The caption generated for the first figure ("Fig. 3. Average numbers of channels, of gaps, and of fragments per channel.") appears to correctly identify and summarize key aspects of the chart, noting the average counts of various elements per channel. The model output suggests a linear trend related to the variable $1/\alpha$, which seems to match the data trends visible in the graph. The comment notes the model provides correct data but might miss some beneficial information. This is reflected in the BLEU-4 score of 0.23, indicating that while the generated caption aligns with the visual data to a certain extent, it falls short of delivering a comprehensive summary.

Image 2 Analysis: For the second figure ("Fig. 2: Average Total Occupancy for $\rho = 0.2$ "), the model's output does not align with the graph's displayed trend, instead offering a misleading narrative of a fluctuating pattern and a negative correlation. The provided comment highlights that the model output contains false data. This discrepancy results in a BLEU-4 score of 0.11, underscoring the model's struggle to accurately interpret and verbalize the data trends in this instance.

Image 3 Analysis: The third example is notably less successful. The gold standard caption describes the figure as showing the "Average Total Occupancy for $\rho = 0.2$," but the model's output diverges significantly from this description, whimsically referring to "rainbow pathways of digital packets," which is not only inaccurate but also nonsensical in the given context. This results in a BLEU-4 score of 0.00, as the generated caption does not provide any informative value with respect to the actual content of the figure.

Across the examples, the model's varying performance highlights the challenges inherent in automated caption generation for scientific figures. When the model can recognize and summarize the data trends accurately, as partially seen in the first figure, it can produce somewhat relevant captions. However, when the model fails to grasp the underlying patterns or the context of the data, as in the second and third figures, the resulting captions can be misleading or entirely incorrect.

4.4. Answers to Research Questions

1) How can neural network models be effectively integrated to automate the generation of accurate and contextually relevant textual descriptions for scientific charts?

To automate the generation of accurate and contextually relevant textual descriptions for scientific charts, neural network models can be effectively integrated by combining Convolutional Neural Networks (CNNs) for feature extraction from charts with Recurrent Neural Networks (RNNs) for generating captions. CNNs extract visual features from charts, which are then fed into RNNs to generate coherent and contextually relevant text descriptions. The integration of these neural network architectures allows for the extraction of meaningful information from visual data and the generation of descriptive captions that enhance the understanding of scientific charts.

2) How does the choice of RNN architecture, such as LSTM or GRU, influence the quality and coherence of generated captions for scientific charts?

The choice of RNN architecture, such as LSTM or GRU, influences the quality and coherence of generated captions for scientific charts. LSTM and GRU are both capable of capturing long-term dependencies in sequential data, which is important for generating accurate captions. However, LSTM has been found to perform better in capturing longer-term dependencies, making it more suitable for generating detailed captions. On the other hand, GRU, with its simpler architecture, may perform better in scenarios where efficiency and speed are prioritized, such as generating shorter captions or processing large volumes of data.

3) How does the developed model compare to existing methods, such as SCICAP, in terms of accuracy, contextuality, and scalability for generating captions for scientific charts?

The developed model compares favorably to existing methods, such as SCICAP, in terms of accuracy, contextuality, and scalability for generating captions for scientific charts. The LSTM model achieved a BLEU-4 score of 0.290, outperforming the highest BLEU-4 score of 0.219 obtained by SCICAP. Additionally, the Bi-LSTM model achieved the best overall performance, with a BLEU-4 score of 0.299 for "Single-Sent Caption," indicating a significant improvement over existing methods in generating accurate and contextually relevant captions for scientific charts.

4) To what extent can the developed model improve the clarity, accessibility, and usability of scientific charts for a wide range of audiences, including non-experts and researchers from different disciplines?

The developed model improved the clarity, accessibility, and usability of scientific charts for a wide range of audiences, including non-experts and researchers from different discipline. however, while showing promising results, is not perfect and requires further research to enhance its performance. Despite achieving higher BLEU-4 scores compared to existing methods, such as SCICAP, there are still areas for improvement, particularly in generating concise and contextually relevant captions for scientific charts. Further research is needed to refine the model's architecture, optimize hyperparameters, and explore more approaches to address the remaining challenges in automated caption generation for scientific charts.

Chapter Five

Conclusion and Future Work

5.1. Conclusion

5.2. Future Work

Chapter Five

Conclusion and Future Work

5.1. Conclusion

The research embarked on a quest to significantly enhance the process of generating textual descriptions for scientific charts, an endeavour rooted in the work of "SCICAP: Generating Captions for Scientific Figures". At the heart of this endeavor lay a sophisticated neural network-based model that strategically melds the capabilities of advanced Convolutional Neural Networks (CNNs) and a suite of Recurrent Neural Network (RNN) architectures, namely Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Bidirectional LSTM (Bi-LSTM). This confluence of technologies was orchestrated to distill the complex visual information contained within scientific charts into coherent, descriptive, and contextually precise captions.

Our research methodically utilized CNNs, particularly the ResNet-101 architecture, for their prowess in dissecting visual inputs into rich, detailed feature vectors, effectively capturing the nuances of scientific charts. These vectors then flowed into the RNN layers, where the LSTM, GRU, and Bi-LSTM models vied in their tasks of weaving these threads of data into textual tapestries – captions that aim to match the descriptive quality of human-written references.

The comparative analysis of the RNN architectures was a cornerstone of this study, shedding light on the strengths and limitations of each model.

The Bi-LSTM emerged as a frontrunner, showcasing an unparalleled ability to incorporate contextual nuances from both preceding and succeeding sequence elements, which translated to superior BLEU-4 scores across various subsets of the SciCap dataset. This finding underscores the pivotal role of context comprehension in the caption generation for scientific figures.

The research builds upon existing architectures in a unique way. While the individual components of the model are not new, the integration and optimization of these components for the specific task of chart-to-text generation represent a novel approach that aims to improve the accuracy and quality of generated captions. This approach is defended by its ability to achieve superior performance compared to existing methods, as demonstrated by the experimental results, and its potential to advance the field of automated caption generation for scientific charts.

Moreover, this research highlighted the challenges inherent in such complex tasks. The BLEU-4 scores, while the best within the Bi-LSTM models, remained below the threshold of 0.3, revealing the steep ascent still to be climbed to reach the zenith of human-like captioning by machines. The LSTM and GRU models, while trailing, offered insights into the trade-offs between model complexity and performance, drawing attention to the nuanced dynamics of model training and optimization.

5.2. Future Work

Future work in the field of automated caption generation for scientific figures should prioritize the incorporation of advanced neural architectures

like transformers that bring attention mechanisms and large-scale pre-training advantages. Exploring multimodal learning by integrating visual data with textual content from related scientific papers could provide richer contextual cues, enhancing caption informativeness and accuracy. Enhancing datasets with a diverse range of disciplines, figure types, and hierarchical captions will be vital for training more robust models. Moreover, improving the interpretability of models and developing more nuanced evaluation metrics will contribute to a deeper understanding of model performance and trust in automated systems. Lastly, optimizing models for computational efficiency will enable practical applications, allowing deployment in various real-time, resource-constrained environments, thus democratizing access to scientific knowledge and aiding in the efficient communication of complex information.

References

- Borkin, M. A. (2013). What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, 2306-2315.
- Ting-Yao Hsu, C. L.-H. (2021). *SciCap: Generating Captions for Scientific Figures*. Association for Computational Linguistics.
- Rush, H. S. (2016). LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics*, 667-676.
- O'Donoghue, S. I.-H. (2018). Visualization of Biomedical Data. *Annual Review of Biomedical Data Science*, 275-304.
- Heimlich, K. B. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 198-213.
- Bengio, P. V. (2018). *International Conference on Learning Representations*.
- Saxena, A. J. (2016). Structural-RNN: Deep Learning on Spatio-Temporal Graphs.
- Wu, Z. a.-G. (2017). *Deep learning for video classification and captioning*. Association for Computing Machinery and Morgan & Claypool.
- Jawahar, A. M. (2019). Textual Description for Mathematical Equations. *2019 International Conference on Document Analysis and Recognition (ICDAR)*.
- Matsuo, A. S. (2021). A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access*, 13248-13265.
- Lee, T. L. (2019). GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction.

Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

- Bajić, F. a. (2022). A Multi-Purpose Shallow Convolutional Neural Network for Chart Images. *Sensors*.
- Renner, R. I. (2018). Discovering physical concepts with neural networks. *Physical review letters*.
- Novruzov, Y. E. (2023). Digital core: neural network recognition of textual geological and geophysical information. *Oil and Gas Studies*.
- Shrivastava, A. K. (2020). Describing Image Using Neural Networks. 605-612.
- Shaikh, H. (2019). Image Synthesis using Textual Descriptions. *International Journal for Research in Applied Science and Engineering Technology*.
- Azad, M. A. (2019). Natural language description of images using hybrid recurrent neural network. *International Journal of Electrical and Computer Engineering (IJECE)*.
- Zhao, F. Z. (2020). Reverse-engineering bar charts using neural networks. *Journal of Visualization*.
- Bless, X. L. (2019). Data Extraction from Charts via Single Deep Neural Network. *ArXiv*.
- Hoque, J. O. (2020). Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. 138-147.
- Ammar, N. S. (2018). Extracting Scientific Figures with Distantly Supervised Neural Networks. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*.
- Morais, P. C. (2018). Evaluation of Convolutional Neural Network Architectures for Chart Image Classification. *2018 International Joint Conference on Neural Networks (IJCNN)*.

- He, K. a. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wojna, C. S. (2015). Rethinking the Inception Architecture for Computer Vision.
- Hochreiter, S. a. (1997). Long Short-term Memory. *Neural computation*.
- Bengio, K. C. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.
- Lipson, J. Y. (2014). How transferable are features in deep neural networks?

المخلص

يقدم هذا البحث نموذجاً جديداً قائماً على الشبكة العصبية مصمماً لتعزيز التوليد التلقائي للأوصاف النصية للمخططات العلمية، بناءً على العمل التأسيسي لـ "SCICAP" لإنشاء تسميات توضيحية للشخصيات العلمية". يجمع نموذجنا المقترح بين معماريات الشبكة العصبية التلافيفية المتقدمة (CNN) لاستخراج الميزات الفعالة من المخططات العلمية مع طبقات متعددة من الشبكة العصبية المتكررة (RNN) لإنشاء تسميات توضيحية وصفية ودقيقة. تتميز هذه الدراسة بسبب تحليلها المقارن لمعماريات RNN المختلفة ، مثل الذاكرة طويلة المدى (LSTM) والوحدات المتكررة ذات البوابات (GRU) ، لتحديد النموذج الأكثر فعالية لإنشاء التسمية التوضيحية النصية. من الأمور الأساسية في منهجيتنا استخدام مقياس BLEU-4 ، وهو معيار في مجال معالجة اللغة الطبيعية لتقييم الجودة اللغوية للنص الذي تم إنشاؤه آلياً مقابل المراجع المكتوبة بشرياً.

أظهر بحثنا أن أداء Bi-LSTM هو الأفضل بدرجة 0.3 BLEU ، وستكون هناك حاجة إلى أبحاث مستقبلية للحصول على تسميات توضيحية أكثر موثوقية.

جامعة البحرين

برامج الدراسات العليا

كلية العلوم



نموذج جديد لإنشاء مخطط إلى نص من خلال استخدام نماذج الشبكات العصبية

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

حمد منير ملك

٢٠١٨٠٥٤٨

إشراف

د. نبيل حويحي
أستاذ

جامعة البحرين

مملكة البحرين
أبريل 2024