



School of Information Technology and
Engineering at the
ADA University



School of Engineering and Applied Science
at the
George Washington University

OPTICAL STRUCTURE RECOGNITION OF CHEMICAL IMAGES USING
TRANSFORMERS

A Thesis

Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Fidan Musazade

April, 2022




THESIS ACCEPTANCE

This Thesis by: Fidan Musazade

Entitled: *Optical Structure Recognition of Chemical Images Using Transformers*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Dr. Jamaladdin Hasanov		28.04.2022
(Adviser)		(Date)
Dr. Abzatdin Adamov		28.04.2022
(Program Director)		(Date)
Dr. Sencer Yeralan		28.04.2022
(Dean)		(Date)

ABSTRACT

The problem of optical chemical structure recognition has been tackled by various researchers using both rule-based and machine learning approaches. However, it still does not have a viable solution that would produce the end-to-end pipeline with high enough accuracy. The approaches tried in this research include implementation of the concept of Transformers to solve this problem as well as image manipulation tactics. The research is focused around applying attention mechanism used in Transformer architecture and Transfer Learning to arrive at results with low Levenshtein Distance, which is a measure of difference between the actual and predicted label for chemical images. The label for images in the study is InChI. Several setups, including Vision Transformers in combination with Vanilla Decoders, as well as EfficientNetV2 backbone with Transformer Encoder and Decoder have been tried. The study suggests that using EfficientNetV2 in couple with Transformer architecture produces best results for the chemical images in Bristol-Myers Squibb dataset published in 2021 electronically. Additionally, resizing with padding instead of stretching produces significantly better results due prevention of information loss. Background and foreground inversion appears to improve the results as well. As a suggestion, further work is recommended to increase the number of epochs and generalize the results for the full dataset instead of a sample used in the study.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
LIST OF FIGURES	iv
LIST OF TABLES	v
LIST OF ABBREVIATIONS	vi
1 Introduction	1
1.1 <i>Definition of Problem</i>	1
1.2 <i>Objective of the Study</i>	3
1.3 <i>Significance of the Problem</i>	3
1.4 <i>Review of Significant Research</i>	4
1.4.1 <i>Rule-Based Systems</i>	4
1.4.2 <i>Machine Learning-Based Systems</i>	6
1.4.3 <i>Current Gaps</i>	7
1.5 <i>Assumptions and Limitations</i>	7
2 METHODOLOGY	9
2.1 <i>Data Source</i>	9
2.1.1 <i>PubChem</i>	9
2.1.2 <i>Bristol Myers Squibb</i>	10
2.2 <i>Chemical Identifiers</i>	11
2.2.1 <i>SMILES</i>	11
2.2.2 <i>InChI</i>	12
2.2.3 <i>Data Preprocessing</i>	13
2.3 <i>Possible Modeling Approaches</i>	15
2.4 <i>Attention Mechanism</i>	15
2.5 <i>Transformers</i>	17
2.6 <i>Vision Transformers</i>	17

2.7	<i>Transfer Learning</i>	19
2.8	<i>Combination of Transfer Learning and Transformers</i>	20
2.9	<i>Evaluation Metric</i>	21
3	Research Results and Analysis of Results	23
3.1	<i>Vision Transformers and Transformers</i>	23
3.1.1	Random Dataset.....	23
3.1.2	Square Images.....	27
3.1.3	Summary Results of Vision Transformers.....	30
3.2	<i>Simplification for Vision Transformers</i>	30
3.2.1	Experiment 1.....	31
3.3	<i>ElasticNetV2 and Transformers</i>	33
3.3.1	Experiment 1.....	33
3.3.2	Experiment 2.....	35
3.3.3	Experiment 3.....	36
3.3.4	Summary Results of EfficientNetV2 and Transformers.....	37
3.4	<i>Analysis of Wrong Predictions</i>	37
4	Summary and Conclusions	39
4.1	<i>Suggested Approach</i>	40
4.2	<i>Future Work</i>	41
	REFERENCES	42

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to Dr. Jamaladdin Hasanov, who has been a great source of motivation and valuable advice during the research efforts.

LIST OF FIGURES

Figure 1. 2D and 3D Images of Aspirin in PubChem.....	10
Figure 2. Train (left) and test (right) images from BMS dataset	11
Figure 3. Train (left) and test (right) images from BMS dataset after preprocessing	14
Figure 4. Scaled dot-product attention [27].....	16
Figure 5. Encoder input (left) and output (right)	18
Figure 6. Example validation image – Experiments 1 and 2.....	23
Figure 7. Encoder Output	24
Figure 8. Example validation image – Experiments 3 and 4.....	26
Figure 9. Example training image – Experiment 1 on Square Images	28
Figure 10. Validation Levenshtein Distance - Experiment 1 on Square Images.....	28
Figure 11. Validation Levenshtein Distance - Experiment 2 on Square Images.....	29
Figure 12. Validation Levenshtein Distance - Experiment 3 on Square Images.....	30
Figure 13. Formula Prediction Training Image and Label	31
Figure 14. Validation Levenshtein Distance - Experiment 1 using EfficientNetV2	35
Figure 15. Validation Levenshtein Distance - Experiment 1 using EfficientNetV2	36
Figure 16. Misclassified Images	38

LIST OF TABLES

Table 1. Training and Validation Statistics - ViT Experiments	25
Table 2. Training and Validation Statistics - ViT Experiments on Square Images.....	27
Table 3. Training and Validation Statistics - Formula Prediction Transformers Experiment 1	31
Table 4. Training and Validation Statistics - Formula Prediction Transformers Experiment 2.....	32
Table 5. Training and Validation Statistics - EfficientNetV2 Experiment 1.....	34
Table 6. Training and Validation Statistics - EfficientNetV2 Experiment 2.....	35
Table 7. Training and Validation Statistics - EfficientNetV2 Experiment 3.....	36

LIST OF ABBREVIATIONS

Abbreviation	Explanation
API	Application Programming Interface
BMS	Bristol Myers Squibb
CAS-RN	Chemical Abstract Service Registry Number
CNN	Convolutional Neural Networks
DECIMER	Deep lEarning for Chemical ImagE Recognition
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
IBM	International Business Machines Corporation
ICMDT	Image Captioning Model based on Deep TNT
InChI	International Chemical Identifier
LD	Levenshtein Distance
MLP	Multi-Layer Perceptron
NLP	Natural Language Processing
OCR	Optical Character Recognition
OCSR	Optical Chemical Structure Recognition
R-CNN	Region-Based Convolutional Neural Network
REST	Representational State Transfer
RNN	Recurrent Neural Networks
SMILES	Simplified Molecular-Input Line-Entry System
SOAP	Simple Object Access Protocol
TNT	Transformer in Transformer
TPU	Tensor Processing Unit
UNII	Unique Ingredient Identifier
ViT	Vision Transformers

1 INTRODUCTION

During the last few decades, significant changes have been made to the presence of scientific documents in the world. Most research entities have switched to electronic databases to preserve the existing literature. The same trend is true for chemistry, where even though many documents with images of molecules exist in various books and printed publications, newer findings are kept in electronic format in the databases. However, most of the historical research is still in the paper format and thus this may lead to some of the documents being lost which could result in loss of knowledge in the field of chemistry. Since digitization in the field of chemistry is rather new, there is a vast amount of work to be done in this area.

The emergence of scanners made it possible to scan existing documents and preserve them in a digital format, however, the documents are still required to be manually parsed to extract data related to individual chemical compounds. A particular hardship in this area is the extraction of chemical identities from molecular drawings, which are mostly drawn by the chemists using paper and pen [18]. With the current level of technological advancements, it should be possible to translate these images into some form of chemical identifiers to be able to collect and store data in an easily searchable format.

The decades of scientific knowledge are present in the raw paper format, which makes it extremely hard to correctly store and extract information when needed. It is obvious that with such old methods as keeping books in libraries without any digital record, most of the knowledge may either be lost or never used because of the complex search process needed to be done. Considering the importance of this digitization, the researchers should focus on the ways to transform the existing data from the paper and pen format to some digital database of easily recognizable structure.

1.1 Definition of Problem

The problem of having a digital version of chemical images and their annotations is quite significant both in academic research and in the field of education. Since most of the currently existing literature is present in the paper format, molecular structures and their text annotations are also in the raw image and textual format inside articles and books [19]. Currently, these molecular images are extracted manually by chemists. Existing tools for extraction are not perfect and have a significant level of error embedded inside the systems.

The new documents have machine-readable parts embedded, which include InChI (International Chemical Identifier), SMILES (Simplified Molecular-Input Line-Entry System), UNII (Unique Ingredient Identifier), CAS RN (Chemical Abstract Service Registry Number), and other possible encoding schemes. These labels help distinguish the molecular images and have a common identifier to make searches and obtain the required visualization. The existence of such tools would accelerate the research process for chemists, provide more learning opportunities in the education sector, and provide the community with more opportunities to mine and obtain necessary information from the chemical data present.

The choice of the identifier to focus on and translate the images to does not have that much importance, since the purpose is to have digital versions of chemical images and not to focus on the

choice of identifier. The motivation behind these identification systems is to have one common notation for all the chemical molecules. To put it simply, one molecule should have one name, which is believed to accelerate the process and make navigation through information faster and more efficient. According to the International Union of Pure and Applied Chemistry, the existence of such common labeling techniques provides the following benefits:

- Enable finding chemical compounds using text-based search engines.
- Help interaction between various databases.
- Uniting data collected through different standards and norms.
- Maintain chemical inventory.
- Help store non-duplicate data due to different drawing styles [5].

The ones most used in the current research are InChI and SMILES. It is important to note that the aim of InChI is not to compete with SMILES or other chemical notation but to help standardize the molecules by adding one additional variable of InChI. The fact that there exists such a common standard that is easily translated into machine language, makes it possible to further develop the research and build a tool that can automatically assign InChI labels to the scanned images from chemical proceedings. There also exist tools that do the same using SMILES notation, however for the purpose of this research the notation does not make any significant difference.

The complexity of this problem may be demonstrated by looking at how InChI labels are constructed by humans. The label consists of layers, each of which provides information about the molecule. Overall, there are six types of layers that can be present in an InChI label. These include:

- Main layer: this layer includes the empirical formula of the element according to Hill convention. Hill convention means starting with carbon, then hydrogen, and then all other elements in alphabetical order.
- Charge layer: provides information about the net charge of the molecule.
- Stereochemical layer: "double bond stereochemistry and tetrahedral stereochemistry" information is presented in this layer [12].
- Isotopic layer: isotopes contained in the molecule are described.
- Fixed-H layer: optional layer.
- Reconnected layer: includes "coordination compounds and organometallics" [11].

The way these layers are presented in the identifier is presented in Figure 1.

The InChI itself is produced after several steps, such as normalization, canonicalization, serialization, and hashing. Each of these processes help produce a clean label. For example, normalization helps remove useless information, canonicalization is concerned with a process of creating a unique number label for each atom, while serialization serves the purpose of generating a string of textual and numeric characters. More details about InChI and how it is derived is presented in the methodology section.

The problem has been tackled by various researchers during the past two decades and several solutions have been implemented to solve the problem. However, they did not use the most contemporary methods that can be implemented to improve the performance of the algorithm. With