USING CYCLEGANS FOR THE CONVERSION OF HAND-DRAWN CHEMICAL STRUCTURES INTO A DIGITAL FORMAT

Jason Chao
Stuyvesant High School
MehtA+

Alexander Lin
High Technology High School
MehtA+

Wallyson Silva
MehtA+

July 27, 2023

ABSTRACT

Chemical structures are often very complicated and can include many components within them. Hand-drawn chemical structures are even more difficult to comprehend, as there are often imperfections within the structures. Therefore, the conversion of hand-drawn structures into a digital format was the objective. In order to accomplish that, two datasets are used, one with digital structures, and one with hand-drawn structures. Additionally, CycleGANs [1], a machine learning model which conducts image-to-image translation was used to train the data. The model did not produce optimal results, but with more epochs and larger datasets, it could make a better image.

1 Introduction

Chemistry is the base of everything that we know today. It defines how everything is made, and how different molecules can create different items that we use on a daily basis. Everything has its chemical structure, where a tangled web of various elements and bonds come together. Some of those elements and bonds come together to create molecules, which are small structures made of multiple elements that have their distinct possibilities. There are over 350,000 different chemical molecules identified and registered, and many more out there in the world that we don't know of. They all have their own unique chemical structure, which can become very complicated, and one singular error can throw everything off. Drawing chemical structures creates a lot of room for error, and it can be very difficult to identify. Minor errors like drawing a single bond instead of a double bond, or writing a number incorrectly or illegibly, change the entire chemical identity that the molecule has.

Therefore, we make it more manageable for people to be able to identify these structures, as turning them into a digital format makes them devoid of mistakes and thus able to be found and identified.

2 Related Work

Machine Learning has been used in many ways with chemical structures. They have been used for the identification of different molecules, even hand-drawn molecules. Stanford's Bradley Emi created a model to recognize hand-drawn chemical structures. A text recognition model was used to understand the words and numbers and they used a variety of different methods to get results and compared it with original names. Jayampathi Adhikari created a model to turn hand-drawn structures into a 3D visualization and Lewis structure, allowing them to be identified easier. Hayley Weir developed a model to recognize hand-drawn hydrocarbons. We use those to create a new design for handwritten structures for easier and more accurate identification.

3 Methodology

3.1 Dataset

The dataset we used was called DECIMER [2]. It contains 5088 images of distinct hand-drawn molecules. We downloaded it two different formats: PNG and SMILES. SMILES is a machine-readable format that represents chemical structures. This makes it rather useful for our project because we can convert SMILES into a digital format as one of our datasets.

Preprocessing Our preprocessing was made up of multiple steps.

The platform that we used, Google Colab, didn't have many resources available, and would not be able to run a program with a dataset of this scale. So, out of the 5088 available molecules, we only chose the 200 simplest ones. We then converted the SMILES into real images in JPEG format using an external website [3]. After, we resized all the images into squares of 256 by 256 pixels. Finally, we split the two datasets into training and testing data, compiled them into respective Google Drive folders, and converted them into Tensorflow datasets [4] for our model.

3.2 Model

We chose CycleGANs (cycle-Consistent Generative Adversarial Networks) as our machine learning model. It is an image-to-image translation model that does not require paired examples, unlike Pix2Pix and other image translation models. Developed in 2017, CycleGANs is a relatively new model, and more advanced.

CycleGANs operates around three main components: the domains, the generators, and the discriminators. There are two of each. The domains are essentially the datasets of images that we are translating between, where each domain is a dataset. Generators, like their names suggest, "generate" images from one domain type to the other. Discriminators differentiate between images of a domain and the images created by the generator that belong to said domain.

4 Results

The final CycleGANs model ran consisted of 60 images for both the digital and the hand-drawn images. It was much smaller than the original 200 because of RAM limits in Google Colaboratory. The results given were not ideal. The model started converting the images but did not make much progress. The model converted the images to a lot of different colors, which matches the digital images that were fed into the model, but a lot of different parts were cut out from the generated images. Surprisingly, our first set of results seemed as if the model was becoming less accurate; the first image, visually, was closest to the actual result, but subsequent images lost quality.

5 Conclusion and Future Work

Looking back on the project, we were really limited in terms of the resources available to us, as our platform lacked computational power and memory. As a result, our program often ran for hours, before suddenly stopping because it consumed too much memory.

In the future, we could create an interface for people to use this code on, as they could draw a structure and a program would automatically format it for them.

6 Division of Labor

We divided the work as follows:

- Preprocessing: Jason Chao, Alexander Lin
- Model: Jason Chao, Alexander Lin
- Poster: Jason Chao, Alexander Lin
- Paper: Jason Chao, Alexander Lin

7 Acknowledgements

We would like to acknowledge the MehtA+ team for their unwavering support throughout the entire process, especially Ms. Haripriya Mehta, Mr. Bhagirath Mehta, and Ms. Minnie Liang.

References

[1] Cyclegan | tensorflow core.

[2] Decimer—hand-drawn molecule images dataset, Jun 2022.

[3] Convert smiles to image - online tool.

[4] Writing custom datasets | tensorflow datasets.