Molecular and Machine Learning Integration Utilizing AutoML

Georgia Gwinnett College
Lawrenceville, GA 30043

Authors: Nick Tyner, Dana Doghaimat, Shiv Patel, Luis Vazquez
Mentors: Evelyn R. Brannock, Robert Lutz, Derek Behmke

Abstract:

Machine learning (ML) and artificial intelligence (AI) can provide innovative technological tools to aid in chemistry learning. This research team will combine these two technologies to create and train an automated machine learning (AutoML) tool that can differentiate between unknown molecules.

Introduction:

Chemistry is a complex subject where traditional teaching methods, such as lectures and textbook learning, are insufficient. A challenge that students face is being able to correctly identify molecular structures. Typically, structures are built using physical model kits or computer applications, but students oftentimes do not have the expertise to assess the accuracy of these models. AI is the application of algorithms and mathematical models to solve common problems. ML is a subset of AI that learns from data to identify patterns and make decisions. ML can be used to recognize the image of a molecular structure and predict said structure.

ML Impact on Our World:

ML are creating tools that dramatically improve the quality of education. The use of AI and ML in educational settings is expected to increase by as much as 47.5% by the year 2021 (1). An example of ML in education is a question, answer based web application named Brainly. Brainly allows the user, typically an educator, to create a model that consists of frequently asked student questions. Brainly utilizes ML algorithms to filter through the asked question and present the verified answer automatically (2). In another research article, ML is being used to perform image recognition on pictures of various professional workers (3). Thousands of images of firefighters, police officers and chefs are organized into a dataset known as IdenProf (3). The IdenProf dataset is used to create a ML model that uses image recognition to predict a person's occupation based on what he or she is wearing in a given photograph (3). Based on our research and despite tremendous growth, ML still have not been integrated into chemistry classrooms.

Google AutoML:

This research team selected Google AutoML Vision as the primary ML engine for this project after obtaining a grant from Google. Google AutoML Vision is an automated machine learning service for image recognition, offered by Google Cloud Platform (GCP), using state-of-the-art transfer learning technology (4). Currently in Beta, AutoML's ML platform permits users to provide labeled images to train a machine operated system for categorization and image recognition projects. AutoML allows users to utilize an interface to create a data repository. The model will begin training and testing, and the user will refine the results. Developers will then deploy the resulting model. A fully functional ML model can then be inserted into a variety of applications as an image recognition solution for end users. (View Figure 1).

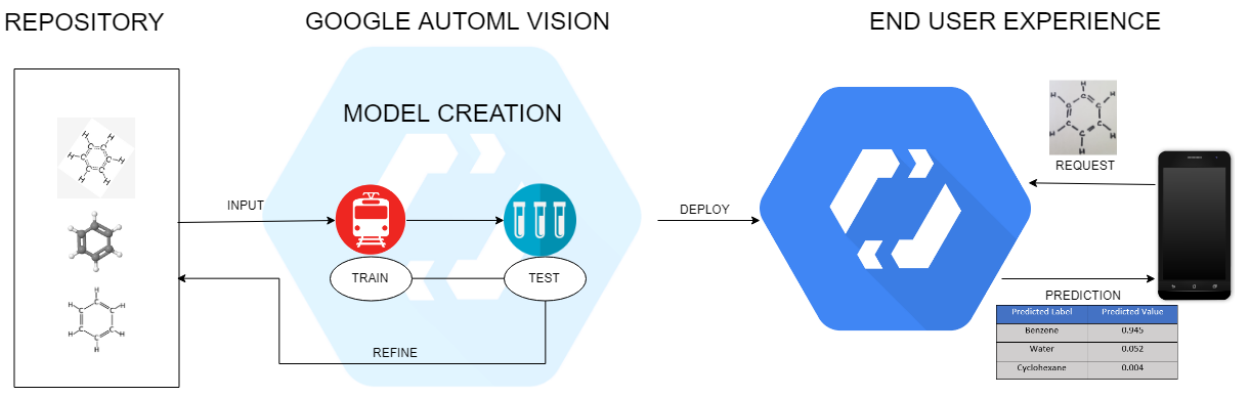

Figure 1: Process flow for identifying unknown images through AutoML.

Users can either evaluate the trained model by viewing the performance metrics offered by AutoML, or by testing the model against a unseen image. The model is trained using many correctly labeled images and is evaluated based on metrics such as precision and recall. Precision and recall both evaluate the percentage of true positive results that the model returns. The precision score penalizes false positive values, and the recall score penalizes false negative values. AutoML uses both scores to report the accuracy of a trained model. (See Figure 2).

$$\textbf{Precision} = \frac{\#\text{ True Postives}}{\#\text{ True Positives} + \#\text{ False Positives}}$$

$$\textbf{Recall} = \frac{\#\text{ True Positives}}{\#\text{ True Positives} + \#\text{ False Negatives}}$$

Figure 2: Formulas for Precision and Recall, which are used in the AutoML training process.

Project Plan

Utilizing AutoML, this research team has already constructed an image recognition model for differentiating between water and benzene. The research team is currently working on integrating additional molecules, such as cyclohexane and aspirin, into an AutoML dataset. The team is collecting over 1500 images for each molecule and organizing the data into a repository on GCP. 2D style hand-drawn images, 2D style computer generated images (CGI) and 3D style CGI are to be used in the repository. The organized dataset will be trained and evaluated in AutoML. Once trained, the image recognition model will be able to recognize never-before-seen images of all labels from the trained dataset. The research team's goal is to create a model that achieves a precision rate greater than or equal to 95 percent for all labels. The team will repeat this process in the future as more molecules are integrated into the model. Ultimately, the research team wants to provide a robust, operationalized model for use as the image recognition solution in an augmented reality (AR) molecule viewing app.

References:

1. Roy, Kunal, et al. “A Primer on QSAR/QSPR Modeling - Fundamental Concepts | Kunal Roy | Springer.” Www.springer.com, Springer India, www.springer.com/us/book/9783319172804.

2. Lynch, Matthew. “5 Examples of Artificial Intelligence in the Classroom.” The Tech Edvocate, 24 July 2018, www.thetechedvocate.org/5-examples-artificial-intelligence-classroom/.

3. Olafenwa, Moses. “Train Image Recognition AI with 5 Lines of Code.” Towards Data Science, Towards Data Science, 20 July 2018, towardsdatascience.com/train-image-recognition-ai-with-5-lines-of-code-8ed0bdd8d9ba.

4. “SMART: Facial Recognition for Molecular Structures.” UC San Diego Jacobs School of Engineering, jacobsschool.ucsd.edu/news/news_releases/release.sfe?id=2353.

5. “Preparing Your Training Data | Cloud AutoML Vision | Google Cloud.” Google, Google, cloud.google.com/vision/automl/docs/prepare.