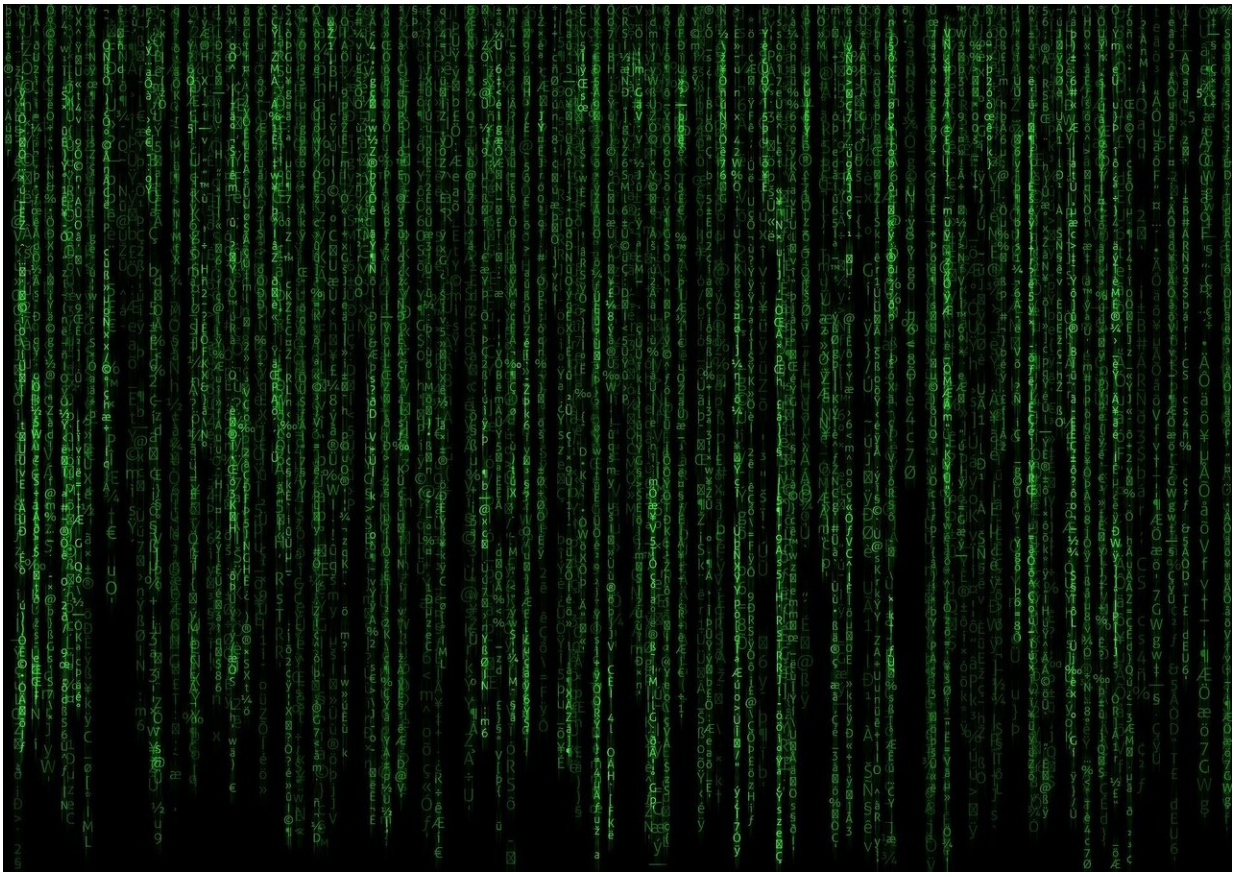
New way of studying genomics makes deep learning a breeze

July 13 2020



Credit: Pixabay/CC0 Public Domain

Researchers from the Max Delbrück Center for Molecular Medicine have developed a new tool that makes it easier to maximize the power of

deep learning for studying genomics. They describe the new approach, Janggu, in the journal *Nature Communications*.

Imagine that before you could make dinner, you first had to rebuild the kitchen, specifically designed for each recipe. You'd spend way more time on preparation, than actually cooking. For computational biologists, it's been a similar time-consuming process for analyzing genomics data. Before they can even begin their analysis, they spend a lot of valuable time formatting and preparing huge data sets to feed into deep learning models.

To streamline this process, researchers from MDC developed a universal programming tool that converts a wide variety of genomics data into the required format for analysis by deep learning models. "Before, you ended up wasting a lot of time on the technical aspect, rather than focusing on the biological question you were trying to answer," says Dr. Wolfgang Kopp, a scientist in the Bioinformatics and Omics Data Science research group at MDC's Berlin Institute of Medical Systems Biology (BIMSB), and first author of the paper. "With Janggu, we are aiming to relieve some of that technical burden and make it accessible to as many people as possible."

Unique name, universal solution

Janggu is named after a traditional Korean drum shaped like an hourglass turned on its side. The two large sections of the hourglass represent the areas Janggu is focused: pre-processing of genomics data, results visualization and model evaluation. The narrow connector in the middle represents a placeholder for any type of deep learning model researchers wish to use.

Deep learning models involve algorithms sorting through massive amounts data and finding relevant features or patterns. While deep

learning is a very powerful tool, its use in genomics has been limited. Most published models tend to only work with fixed types of data, able to answer only one specific question. Swapping out or adding new data often requires starting over from scratch and extensive programming efforts.

Janggu converts different genomics data types into a universal format that can be plugged into any machine learning or deep learning model that uses python, a widely-used programming language.

"What makes our approach special is that you can easily use any genomic data set for your deep learning problem, anything goes in any format," Dr. Altuna Akalin, who heads the Bioinformatics and Omics Data Science research group.

Separation is key

Akalin's research group has a dual mission: developing new machine learning tools, and using them to investigate questions in biology and medicine. During their own research efforts, they were continually frustrated by how much time was spent formatting data. They realized part of the problem was each deep learning model included its own data pre-processing. By separating the data extraction and formatting from the analysis, it provides a much easier way to interchange, combine or reuse sections of data. It's kind of like having all the kitchen tools and ingredients at your fingertips ready to try out a new recipe.

"The difficulty was finding the right balance between flexibility and usability," Kopp says. "If it is too flexible, people will be drowned in different options and it will be difficult to get started."

Kopp has prepared several tutorials to help others begin using Janggu, along with example datasets and case studies. The *Nature*

Communications paper demonstrates Janggu's versatility in handling very large volumes of data, combining data streams, and answering different types of questions, such as predicting binding sites from DNA sequences and/or chromatin accessibility, as well as for classification and regression tasks.

Endless applications

While most of Janggu's benefit is on the front end, the researchers wanted to provide a complete solution for deep learning. Janggu also includes visualization of results after the deep learning analysis, and evaluates what the model has learned. Notably, the team incorporated "higher-order sequence encoding" into the package, which allows to capture correlations between neighboring nucleotides. This helped to increase accuracy of some analyses. By making deep learning easier and more user-friendly, Janggu helps throw open the door to answering all kinds of biological questions.

"One of the most interesting applications is predicting the effect of mutations on gene regulation," Akalin says. "This is exciting because now we can start understanding individual genomes, for instance, we can pinpoint genetic variants that cause regulatory changes, or we can interpret regulatory mutations occurring in tumors."

More information: Wolfgang Kopp et al. Deep learning for genomics using Janggu, *Nature Communications* (2020). DOI: 10.1038/s41467-020-17155-y

Provided by Max Delbrück Center for Molecular Medicine