idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft

Jana Schlütter

13.07.2020

http://idw-online.de/de/news750914

Forschungsergebnisse, Forschungsprojekte
Biologie, Ernährung / Gesundheit / Pflege, Informationstechnik, Medizin
überregional

Janggu macht Deep Learning zum Kinderspiel

Forscher*innen des MDC haben eine neues Softwareanwendung entwickelt, mit der sich Deep Learning für Genomik-Studien optimal und einfach nutzen lässt: Janggu stellen die Forschenden nun erstmals im Fachjournal Nature Communications vor.

Stellen Sie sich folgendes Szenario vor: Um das Abendessen zubereiten zu können, müssen Sie erst die Küche passend für das jeweilige Rezept umbauen. Die Vorbereitung würde deutlich mehr Zeit in Anspruch nehmen als das eigentliche Kochen. Bislang brauchten Bioinformatiker*innen für die Analyse genomischer Daten ähnlich lange. Bevor sie überhaupt mit ihrer Analyse beginnen konnten, investierten sie zunächst viel Zeit in die Formatierung und Aufbereitung riesiger Datensätze, die in Deep-Learning-Modelle integriert werden.

Um diesen Prozess zu straffen, haben Forschende des Max-Delbrück-Centrums für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC) eine universelle Programmiersoftware entwickelt, das eine Vielzahl genomischer Daten in das für die Analyse durch Deep-Learning-Modelle erforderliche Format konvertiert. „Bislang nahmen die technischen Aspekte viel Zeit in Anspruch – Zeit, die dann für die biologischen Fragestellungen fehlt, die wir beantworten wollen“, sagt Dr. Wolfgang Kopp, Wissenschaftler in der Forschungsgruppe „Bioinformatics and Omics Data Science“ am Berliner Institut für Medizinische Systembiologie (BIMSB) des MDC und Erstautor der Studie. „Janggu soll einen Teil dieses technischen Aufwands tilgen. Das Softwarepaket möchten wir so vielen Menschen wie möglich zugänglich machen.“

Ein besonderer Name für eine universelle Lösung

Janggu ist nach einer traditionellen koreanischen Trommel benannt, deren Form an eine auf der Seite liegende Sanduhr erinnert. Die beiden großen Teile der Sanduhr stehen für die Bereiche, auf die sich Janggu konzentriert: die Aufbereitung genomischer Daten sowie die Ergebnisvisualisierung und Modellauswertung. Das schmale Verbindungsstück in der Mitte stellt einen Platzhalter für ein beliebiges Deep-Learning-Modell dar.

Deep-Learning-Modelle beinhalten Algorithmen, die riesige Datenmengen verarbeiten und dabei wichtige Merkmale oder Muster erkennen. Obwohl Deep Learning eine sehr leistungsfähige Methode ist, kommt sie in der Genomik bislang nur eingeschränkt zum Einsatz. Die meisten veröffentlichten Modelle sind auf bestimmte Datentypen angewiesen und können nur eine spezifische Frage beantworten. Um Daten auszutauschen oder hinzuzufügen, muss man oft wieder bei null anfangen – ein immenser Programmieraufwand.

Janggu konvertiert verschiedene Genomik-Datentypen in ein universelles Format. So können die Daten in jedes Modell – ob Deep Learning oder maschinelles Lernen – eingebunden werden, das die gängige Programmiersprache Python verwendet. „Das Besondere an unserem Ansatz ist, dass man für ein Deep-Learning-Problem jeden genomischen Datensatz verwenden kann – wir können mit jedem Format arbeiten. Die Möglichkeiten sind endlos“, sagt Dr. Altuna Akalin, Leiter der Forschungsgruppe „Bioinformatics and Omics Data Science“.

Trennung als Schlüsselaspekt

Akalins Forschungsgruppe hat aber noch eine andere Aufgabe: Das Team entwickelt neue Softwareanwendungen für Maschinelles Lernen und will diese bei Forschungsfragen in der Biologie und Medizin einsetzen. Bei ihren eigenen Forschungsprojekten waren die Wissenschaftler*innen oft frustriert, dass die Formatierung der Daten so viel Zeit in Anspruch nimmt. Sie erkannten, dass ein Teil des Problems darin bestand, dass für jedes Deep-Learning-Modell eine Aufbereitung der Daten nötig war. Durch die Trennung von Datenextraktion und -formatierung von der Analyse lassen sich Datenabschnitte viel einfacher austauschen, kombinieren und wiederverwenden. Das ist etwa so, als hätte man alle Küchenutensilien und Zutaten bereits zur Hand, um ein neues Rezept auszuprobieren.

„Die Schwierigkeit bestand darin, das richtige Gleichgewicht zwischen Flexibilität und Benutzerfreundlichkeit zu finden“, sagt Kopp. „Bei zu viel Flexibilität hätten die Benutzerinnen und Benutzer zu viele Optionen, was sie überfordern würde und es wäre schwierig, überhaupt einen Anfang zu finden.“

Kopp hat mehrere Tutorials sowie Beispieldatensätze und Fallstudien vorbereitet, die Benutzer*innen im Umgang mit Janggu unterstützen sollen. Die Veröffentlichung in Nature Communications zeigt, wie anpassungsfähig Janggu ist – im Umgang mit sehr großen Datenmengen, bei der Kombination von Datenströmen und bei der Beantwortung verschiedener Fragestellungen, z. B. bei der Vorhersage von Bindungsstellen aus DNA-Sequenzen, der Chromatin-Zugänglichkeit und der Klassifizierung und Regression.

Unbegrenzte Anwendungsmöglichkeiten

Die Vorzüge von Janggu zeigen sich vor allem in der Datenaufbereitung. Dennoch wollten die Forschenden eine Komplettlösung für Deep Learning anbieten. Janggu ermöglicht auch eine Ergebnisvisualisierung nach der Deep-Learning-Analyse und wertet aus, was das Modell gelernt hat. Bemerkenswert ist, dass das Team eine „übergeordnete Sequenzkodierung“ in das Programm integriert hat, die es erlaubt, Zusammenhänge zwischen benachbarten Nukleotiden zu erfassen. So konnte die Genauigkeit einiger Analysen erhöht werden. Janggu macht Deep Learning einfacher und benutzerfreundlicher und trägt dazu bei, verschiedenste biologische Fragestellungen zu beantworten.

„Eine der interessantesten Anwendungen ist die Prognose der Auswirkung von Mutationen auf die Genregulation“, sagt Akalin. „Das ist wirklich spannend, weil wir so einzelne Genome besser verstehen können. Wir sind beispielsweise in der Lage, genetische Varianten aufzuspüren, die die Genregulation beeinflussen und wir können regulatorische Mutationen in Tumoren interpretieren.“

Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC)

Das Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC) wurde 1992 in Berlin gegründet. Benannt wurde es nach dem deutsch-amerikanischen Physiker Max Delbrück, der 1969 mit dem Nobelpreis für Physiologie und Medizin ausgezeichnet wurde. Das MDC untersucht molekulare Mechanismen, um die Entstehung von Krankheiten zu verstehen und sie so besser und wirksamer diagnostizieren, verhindern und bekämpfen zu können. Dabei arbeitet das MDC mit der Charité – Universitätsmedizin Berlin und dem Berlin Institute of Health (BIH) sowie mit nationalen Partnern wie dem Deutschen Zentrum für Herz-Kreislaufforschung e. V. und zahlreichen internationalen Forschungseinrichtungen zusammen. Über 1.600 Mitarbeiter und Gäste aus fast 60 Ländern arbeiten am MDC, knapp 1.300 davon in der wissenschaftlichen Forschung. Das MDC wird zu 90 % vom Bundesministerium für Bildung und Forschung und zu 10 % vom Land Berlin finanziert und ist Mitglied der Helmholtz-Gemeinschaft Deutscher Forschungszentren.

wissenschaftliche Ansprechpartner:

Dr. Altuna Akalin
Laborleiter „Bioinformatics and Omics Data Science“
Berliner Institut für Medizinische Systembiologie (BIMSB)
Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC)
Altuna.Akalin@mdc-berlin.de

Originalpublikation:

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

URL zur Pressemitteilung: https://www.mdc-berlin.de/bioinformatics Forschungsgruppe von Altuna Akalin

URL zur Pressemitteilung: https://www.mdc-berlin.de/de/news/press/maui Pressemitteilung „Deep Learning erkennt molekulare Muster von Krebs“

• Die Wissenschaftler Altuna Akalin (links) und Wolfgang Kopp (rechts) aus der Arbeitsgruppe „Bioinformatics and Omics Data Science“.
Felix Petermann
MDC