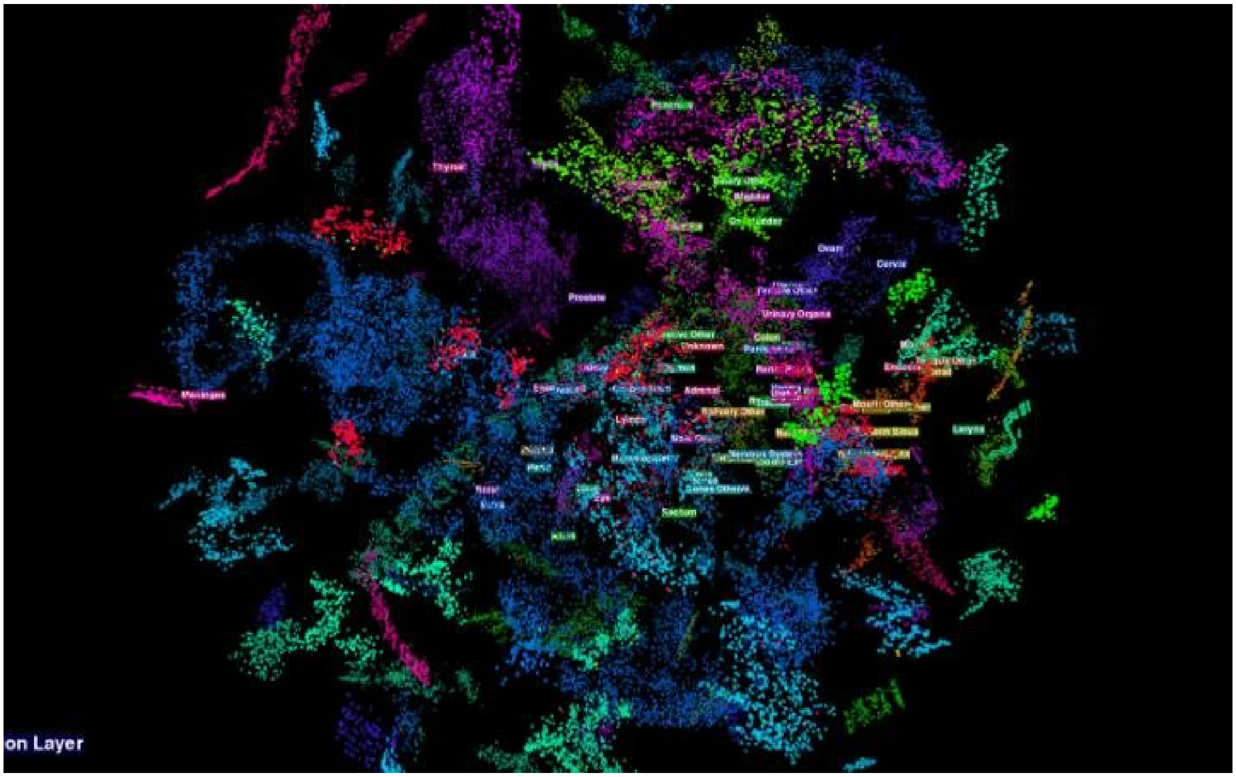
Researchers develop 'multitasking' AI tool to extract cancer data in record time

February 12 2020



The image visualizes how the team's multitask convolutional neural network classifies primary cancer sites. Credit: Hong-Jun Yoon/ORNL

As the second-leading cause of death in the United States, cancer is a public health crisis that afflicts nearly one in two people during their lifetime. Cancer is also an oppressively complex disease. Hundreds of cancer types affecting more than 70 organs have been recorded in the nation's cancer registries—databases of information about individual cancer cases that provide vital statistics to doctors, researchers, and policymakers.

"Population-level cancer surveillance is critical for monitoring the effectiveness of public health initiatives aimed at preventing, detecting, and treating cancer," said Gina Tourassi, director of the Health Data Sciences Institute and the National Center for Computational Sciences at the Department of Energy's Oak Ridge National Laboratory. "Collaborating with the National Cancer Institute, my team is developing advanced artificial intelligence solutions to modernize the national cancer surveillance program by automating the time-consuming data capture effort and providing near real-time cancer reporting."

Through digital cancer registries, scientists can identify trends in cancer diagnoses and treatment responses, which in turn can help guide research dollars and public resources. However, like the disease they track, cancer pathology reports are complex. Variations in notation and language must be interpreted by human cancer registrars trained to analyze the reports.

To better leverage cancer data for research, scientists at ORNL are developing an artificial intelligence-based natural language processing tool to improve information extraction from textual pathology reports. The project is part of a DOE-National Cancer Institute collaboration known as the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) that is accelerating research by merging cancer data with advanced data analysis and high-performance computing.

As DOE's largest Office of Science laboratory, ORNL houses unique computing resources to tackle this challenge—including the world's most powerful supercomputer for AI and a secure data environment for processing protected information such as health data. Through its

Surveillance, Epidemiology, and End Results (SEER) Program, NCI receives data from cancer registries, such as the Louisiana Tumor Registry, which includes diagnosis and pathology information for individual cases of cancerous tumors.

"Manually extracting information is costly, time consuming, and error prone, so we are developing an AI-based tool," said Mohammed Alawad, research scientist in the ORNL Computing and Computational Sciences Directorate and lead author of a paper published in the *Journal of the American Medical Informatics Association* on the results of the team's AI tool.

In a first for cancer pathology reports, the team developed a multitask convolutional neural network, or CNN—a deep learning model that learns to perform tasks, such as identifying key words in a body of text, by processing language as a two-dimensional numerical dataset.

"We use a common technique called word embedding, which represents each word as a sequence of numerical values," Alawad said.

Words that have a semantic relationship—or that together convey meaning—are close to each other in dimensional space as vectors (values that have magnitude and direction). This textual data is inputted into the neural network and filtered through network layers according to parameters that find connections within the data. These parameters are then increasingly honed as more and more data is processed.

Although some single-task CNN models are already being used to comb through pathology reports, each model can extract only one characteristic from the range of information in the reports. For example, a single-task CNN may be trained to extract just the primary cancer site, outputting the organ where the cancer was detected such as lungs, prostate, bladder, or others. But extracting information on the histological grade, or growth of cancer cells, would require training a separate deep learning model.

The research team scaled efficiency by developing a network that can complete multiple tasks in roughly the same amount of time as a single-task CNN. The team's neural network simultaneously extracts information for five characteristics: primary site (the body organ), laterality (right or left organ, if applicable), behavior, histological type (cell type), and histological grade (how quickly the cancer cells are growing or spreading).

The team's multitask CNN completed and outperformed a single-task CNN for all five tasks within the same amount of time—making it five times as fast. However, Alawad said, "It's not so much that it's five times as fast. It's that it's n-times as fast. If we had n different tasks, then it would take one-nth of the time per task."

The team's key to success was the development of a CNN architecture that enables layers to share information across tasks without draining efficiency or undercutting performance.

"It's efficiency in computing and efficiency in performance," Alawad said. "If we use single-task models, then we need to develop a separate model per task. However, with multitask learning, we only need to develop one model—but developing this one model, figuring out the architecture, was computationally time consuming. We needed a supercomputer for model development."

To build an efficient multitask CNN, they called on the world's most powerful and smartest supercomputer—the 200-petaflop Summit supercomputer at ORNL, which has over 27,600 deep learning-optimized GPUs.

The team started by developing two types of multitask CNN architectures—a common machine learning method known as hard parameter sharing and a method that has shown some success with image classification known as cross-stitch. Hard parameter sharing uses the same few parameters across all tasks, whereas cross-stitch uses more parameters fragmented between tasks, resulting in outputs that must be "stitched" together.

To train and test the multitask CNNs with real health data, the team used ORNL's secure data environment and over 95,000 pathology reports from the Louisiana Tumor Registry. They compared their CNNs to three other established AI models, including a single-task CNN.

"In addition to offering HPC and scientific computing resources, ORNL has a place to train and store secure data—all of these together are very important," Alawad said.

During testing they found that the hard parameter sharing multitask model outperformed the four other models (including the cross-stitch multitask model) and increased efficiency by reducing computing time and energy consumption. Compared with the single-task CNN and conventional AI models, the hard sharing parameter multitask CNN completed the challenge in a fraction of the time and most accurately classified each of the five cancer characteristics.

"The next step is to launch a large-scale user study where the technology will be deployed across cancer registries to identify the most effective ways of integration in the registries' workflows. The goal is not to replace the human but rather augment the human," Tourassi said.

More information: Mohammed Alawad et al, Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks, *Journal of the American*

Medical Informatics Association (2019). DOI: 10.1093/jamia/ocz153

Provided by Oak Ridge National Laboratory

Citation: Researchers develop 'multitasking' AI tool to extract cancer data in record time (2020, February 12) retrieved 2 May 2024 from https://medicalxpress.com/news/2020-02-multitasking-ai-tool-cancer.html