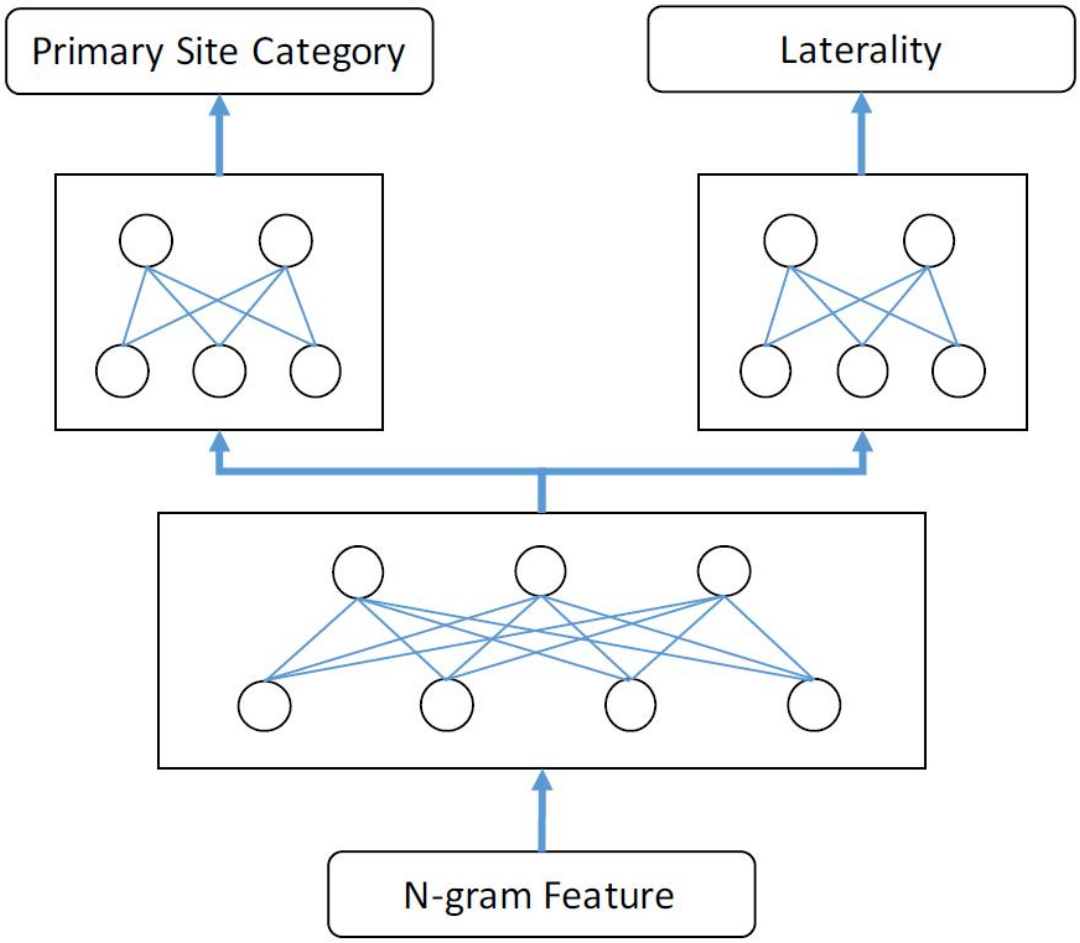
Accelerating cancer research with deep learning

November 9 2016



A representation of a deep learning neural network designed to intelligently extract text-based information from cancer pathology reports. Credit: Oak Ridge National Laboratory

Despite steady progress in detection and treatment in recent decades, cancer remains the second leading cause of death in the United States, cutting short the lives of approximately 500,000 people each year.

To better understand and combat this disease, medical researchers rely on cancer registry programs—a national network of organizations that systematically collect demographic and clinical information related to the diagnosis, treatment, and history of cancer incidence in the United States. The surveillance effort, coordinated by the National Cancer Institute (NCI) and the Centers for Disease Control and Prevention, enables researchers and clinicians to monitor cancer cases at the national, state, and local levels.

Much of this data is drawn from electronic, text-based clinical reports that must be manually curated—a time-intensive process—before it can be used in research. For example, cancer pathology reports, text documents that describe cancerous tissue in detail, must be individually read and annotated by experts before becoming part of a cancer registry. With millions of new reports being produced each year, the information burden continues to grow.

"The manual model is not scalable," said Georgia Tourassi, director of the Health Data Sciences Institute at the US Department of Energy's (DOE's) Oak Ridge National Laboratory (ORNL). "We need to develop new tools that can automate the information-extraction process and truly modernize cancer surveillance in the United States."

Since 2014 Tourassi has led a team focused on creating software that can quickly identify valuable information in cancer reports, an ability that would not only save time and worker hours but also potentially reveal overlooked avenues in cancer research. After experimenting with conventional natural-language-processing software, the team's most recent progress has emerged via deep learning, a machine-learning

technique that employs algorithms, big data, and the computing power of GPUs to emulate human learning and intelligence.

Using the Titan supercomputer at the Oak Ridge Leadership Computing Facility, a DOE Office of Science User Facility located at ORNL, Tourassi's team applied deep learning to extract useful information from cancer pathology reports, a foundational element of cancer surveillance. Working with modest datasets, the team obtained preliminary findings that demonstrate deep learning's potential for cancer surveillance.

The continued development and maturation of automated data tools, among the objectives outlined in the White House's Cancer Moonshot initiative, would give medical researchers and policymakers an unprecedented view of the US cancer population at a level of detail typically obtained only for clinical trial patients, historically less than 5 percent of the overall cancer population.

"Today we're making decisions about the effectiveness of treatment based on a very small percentage of cancer patients, who may not be representative of the whole patient population," Tourassi said. "Our work shows deep learning's potential for creating resources that can capture the effectiveness of cancer treatments and diagnostic procedures and give the cancer community a greater understanding of how they perform in real life."

Beauty of the black box

Creating software that can understand not only the meaning of words but also the contextual relationships between them is no simple task. Humans develop these skills through years of back-and-forth interaction and training. For specific tasks, deep learning compresses this process into a matter of hours.

Typically, this context-building is achieved through the training of a neural network, a web of weighted calculations designed to produce informed guesses on how to correctly carry out tasks, such as identifying an image or processing a verbal command. Data fed to a neural network, called inputs, and select feedback give the software a foundation to make decisions based on new data. This algorithmic decision-making process is largely opaque to the programmer, a dynamic akin to a teacher with little direct knowledge of her students' perception of a lesson.

"With deep learning you just throw the document in and say, 'Figure it out,'" Tourassi said. "It's more like a black box, but that's the beauty. We do not impose our own constraints."

GPUs, such as those in Titan, can accelerate this training process by quickly executing many deep-learning calculations simultaneously. In two recent studies, Tourassi's team used accelerators to tune multiple algorithms, comparing results to more traditional methods. Using a dataset composed of 1,976 pathology reports provided by NCI's Surveillance, Epidemiology, and End Results (SEER) Program, Tourassi's team trained a deep-learning algorithm to carry out two different but closely related information-extraction tasks. In the first task the algorithm scanned each report to identify the primary location of the cancer. In the second task the algorithm identified the cancer site's laterality—or on which side of the body the cancer was located.

By setting up a neural network designed to exploit the related information shared by the two tasks, an arrangement known as multitask learning, the team found the algorithm performed substantially better than competing methods.

"Intuitively this makes sense because carrying out the more difficult objective is where learning the context of related tasks becomes beneficial," Tourassi said. "Humans can do this type of learning because

we understand the contextual relationships between words. This is what we're trying to implement with deep learning."

Another study carried out by Tourassi's team used 946 SEER reports on breast and lung cancer to tackle an even more complex challenge: using deep learning to match the cancer's origin to a corresponding topological code, a classification that's even more specific than a cancer's primary site or laterality, with 12 possible answers.

The team tackled this problem by building a convolutional neural network, a deep-learning approach traditionally used for image recognition, and feeding it language from a variety of sources. Text inputs ranged from general (e.g., Google search results) to domain-specific (e.g., medical literature) to highly specialized (e.g., cancer pathology reports). The algorithm then took these inputs and created a mathematical model that drew connections between words, including words shared between unrelated texts.

Comparing this approach to more traditional classifiers, such as a vector space model, the team observed incremental improvement in performance as the network absorbed more cancer-specific text. These preliminary results will help guide Tourassi's team as they scale up deep-learning algorithms to tackle larger datasets and move toward less supervision, meaning the algorithms will make informed decisions with less human intervention.

In 2016 Tourassi's team learned its cancer surveillance project will be developed as part of DOE's Exascale Computing Project, an initiative to develop a computing ecosystem that can support an exascale supercomputer—a machine that can execute a billion billion calculations per second. Though the team has made considerable progress in leveraging deep learning for cancer research, the biggest gains are still to come.

"Focusing on clinical text alone, the value would be tremendous," Tourassi said.

More information: Hong-Jun Yoon, Arvind Ramanathan, and Georgia Tourassi, "Multi-Task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports." International Neural Network Society Conference on Big Data. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-47898-2_21.

Provided by Oak Ridge National Laboratory