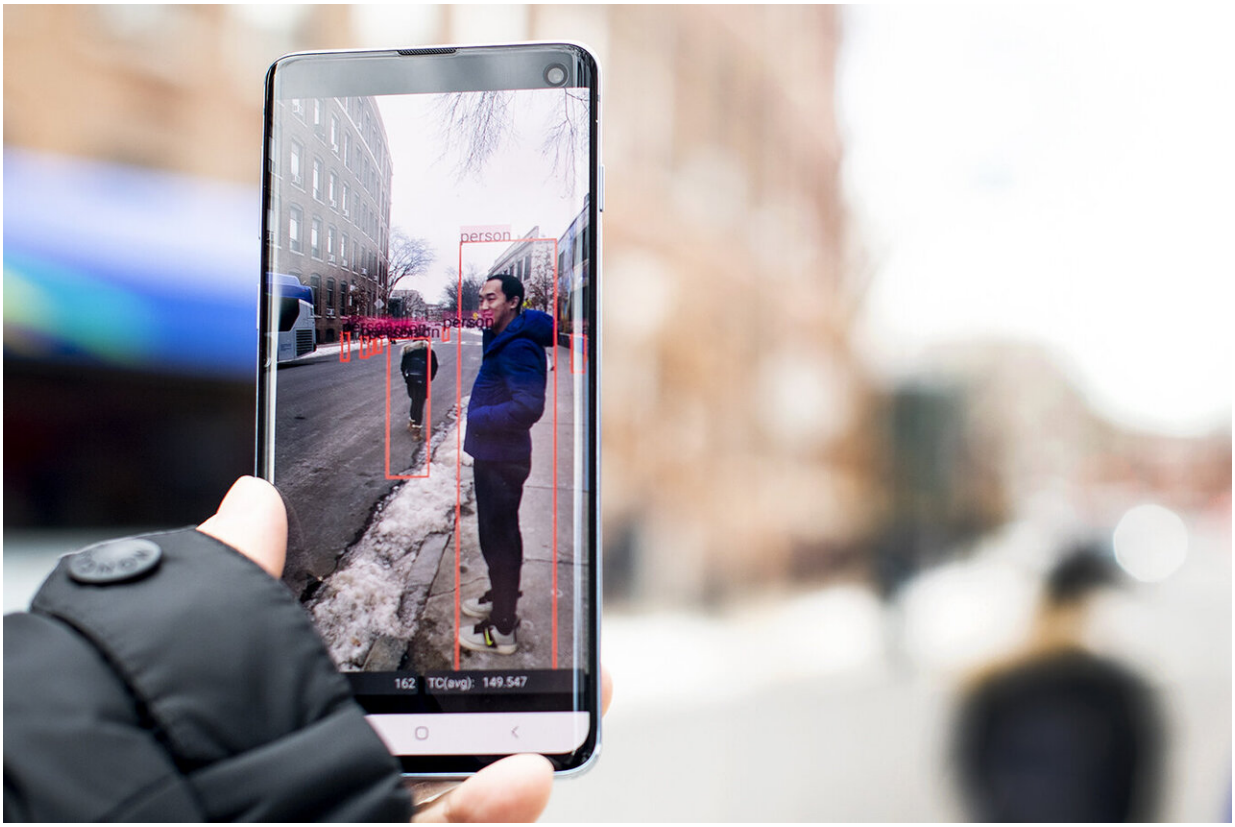
Deep neural networks are coming to your phone

January 29 2020, by Laura Castañón



Yanzhi Wang, assistant professor of electrical and computer engineering, has devised a way to run deep neural networks on mobile devices like the average cell phone. Credit: Ruby Wallau/Northeastern University

How does a self-driving car tell a person apart from a traffic cone? How

does Spotify choose songs for my "Discover Weekly" playlist? Why is Gmail's spam filter so effective?

The answer is a type of artificial intelligence known as deep neural networks. These networks are very good at recognizing and classifying data, but they tend to take a lot of computing power and memory to run—too much to work quickly on something like your average smartphone.

Now researchers at Northeastern have demonstrated a way to run deep neural networks on a smartphone or similar system. Using their method, the networks can execute tasks up to 56 times faster than demonstrated in previous work, without losing accuracy. They will be presenting their work at a conference on artificial intelligence next month in New York.

"It is difficult for people to achieve the real-time execution of neural networks on a smartphone or these kinds of mobile devices," says Yanzhi Wang, an assistant professor of electrical and computer engineering at Northeastern. "But we can make most deep learning applications work in real-time."

Typically, a mobile device needs to be connected to the internet to have access to a deep neural network. The phone collects data, but the processing is done on remote servers—this is why you can't talk to Siri when your iPhone is in airplane mode.

Wang and his colleagues have devised a way to both reduce the size of the neural network model and automatically generate code to run it more efficiently. This work could allow deep neural networks to be implemented in off-the-shelf devices that may not have consistent internet access. And that has uses far beyond hands-free communication with your phone.

"There are so many things that need intelligence," Wang says. "Medical devices, wearable devices, sensors, smart cameras. All of these, they need something enhancing recognition, segmentation, tracking, surveillance, and so many things, but currently they're limited."

Yanzhi Wang is an assistant professor of electrical and computer engineering. Credit: Ruby Wallau/Northeastern University

Artificial intelligence is already being used to improve medical technology in hospitals. There are plenty of opportunities for it to expand the uses of wearable devices as well, potentially providing guidance for disabled individuals or alerting patients and doctors to changes in heart rhythm or other concerns. But imagine missing an alert

about a potential heart attack because you were on the subway and didn't have service.

"For many of the medical device applications, we cannot assume that this kind of device is always connected to the internet," Wang says. "And in connecting to the internet, there is always a significant delay. Everything needs to be computed and sent back."

When Wang says "a significant delay," he is talking about fractions of a second. But that is enough to make a difference.

"For self-driving cars, all the data needs to be sent to a cloud data center, then there is a transmission delay sending it back," Wang says. "Maybe .1 seconds. And this .1 seconds may cause damage."

Eliminating that fraction of a second delay might save lives.

Wang also notes that deep neural networks can raise privacy concerns, because personal information is shared over the cloud in order for these networks to function. Processing data locally, without sending it off to distant servers, could make people more comfortable using devices powered by artificial intelligence.

"Previously, people believed that deep learning needed dedicated chips, or could only be run on servers over the cloud," Wang says. "This kind of assumption of knowledge limits the application of the deep learning. We cannot always rely on the cloud. We need to make local, smart decisions."

More information: PCONV: The Missing but Desirable Sparsity in DNN Weight Pruning for Real-time Execution on Mobile Devices. arXiv:1909.05073v3 [cs.LG]: arxiv.org/abs/1909.05073

Provided by Northeastern University

Citation: Deep neural networks are coming to your phone (2020, January 29) retrieved 30 January 2023 from https://techxplore.com/news/2020-01-deep-neural-networks.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.