**MIT CSAIL Alliances | LabDemo_NeuralNetworks_LucasLiebenwein_06/18/2020**

Neural networks are a type of AI system that can be used as a powerful tool to predict patterns in data, images, text, or video. However, neural networks tend to be extremely large, making them inefficient to train and deploy on small devices like robots and phones. Thus neural networks usually require expensive cloud infrastructure, like GPUs. Graduate students at MIT CSAIL, Lucas Liebenwein and Cenk Baykal and several other colleagues, aim to take a large neural network and reduce the amount of parts into a smaller architecture, allowing it to be deployed onto smaller devices like robots and phones.

Our work is also backed up by our accompanying mathematical theory that enables you to predict how much pruning can be done. In some sense, it opens up the black box of deep learning by analyzing each network component individually. To the best of our knowledge, actually, this is the first time that there's rigorous theory for understanding how pruning affects neural networks, and specifically how pruning affects the accuracy of neural networks.

The algorithm they're using is most useful for scenarios for real-time computing, where you want the most accurate predictions within the smallest time frame possible. With the smaller architectures, you're able to design cheaper devices/infrastructures and make more efficient energy chips.

We don't need to rely on an internet connection or the cloud in order to produce an answer. Instead, we can guarantee execute the neural network on the device itself. That makes it much more fault tolerant. In addition, by keeping the data security on the device itself, you can alleviate many of the privacy concerns surrounding AI, because otherwise the data would have to be sent to the cloud, where it might be processed unencrypted, or third parties could gain access to it.

For robot systems, in particular, sending data to the cloud is not a viable option. Robots need to interact with their environment. They need to react to changes in their environment. And they need to do so with minimum time delay possible.

So sending data to the cloud could cause a delay in the system, which could cause catastrophic consequences in turn. So by using our algorithms, now we can instead directly execute the neural network on the robot and use the neural network for tasks that might otherwise be too computationally expensive for the robot system to run.

There is still so much to learn from neural networks. And Lucas and his group are hoping to contribute tools that can improve our high level understanding of them.

[MUSIC PLAYING]

[CLICK]

[DING]