CIS-490H Edge Computing

With Dr. Zheng Song

Paper Review: Week 8

"Distributed Deep Neural Networks over the Cloud, the Edge and End Devices"

Student: Demetrius Johnson

22 February 2023

1. Summary.

(1) Motivation

The primary motivation for the *Distributed Deep Neural Networks over the Cloud, the Edge and End Devices* is to provide a framework for distributing a deep neural network across the hierarchy of computational structures such as end devices, the edge (fog) and the cloud. With the emergence of so many IoT devices that are heterogeneous but that collect the same kinds of data – although the data representation itself might be heterogenous – it is natural that artificial intelligence would take advantage of the data generated to produce deep neural networks that can give high accuracy when using sensor data for classification. Then, with the distributed systems paradigm, edge computing has also made a way for processing data closer to the edge, especially the processing power required to run a neural network model to yield a classification output. Thus, the motivation is to fuse sensor data by conveniently utilizing distributive nature of the structures of neural networks and of the internet device-edge-cloud connection paradigms.

(2) Contribution

The researcher provide a the framework for a distributed deep neural network, which they call a DDNN. They include newer deep neural network methodologies that they also supported in their previous work, such as the Binarized Deep Neural Network, which can save memory and be deployed on end devices while still maintain a high enough degree of inference accuracy. Part of the framework they provide includes a model that they train jointly on the cloud, so that not only are all layers of the model trained on the cloud, but sections of the model that will be deployed to end or edge devices are trained on the cloud as well so that the parts of the model can simply be deployed to the respective devices that will not have to worry about training their model – the devices can simply use the already-trained model for inference. Lastly, they provide a way to aggregate output in order to make a more informed inference or a decision to send data to the next level in the device-edge(fog)-cloud hierarchy for further processing.

(3) Methodology and/or argument

For their implementation, they use only the cloud-end device model. They use six end devices, and one aggregator where the outputs from the end devices can go to be analyzed before a decision is made to either exit the neural network (and thus confirm that the analysis is good enough to make an inference) or continue along to remaining layers in the neural network (and thus send the output to the cloud). They also deploy their DNN on the cloud to be trained jointly with partial parts of the model that will be deployed to the end devices. For the data collection, they use a multi-view multi-camera environment in order to demonstrate how their distributed neural network could be used to fuse sensor data through aggregating the neural network outputs of each end device at the aggregator device. For all models including the complete model in the cloud, Binarized Neural Networks were used in order to simplify outputs.

(4) Conclusion

Overall, the *DDNN* system showed its ability to fuse sensor data from heterogenous devices, as well as greatly reduce the network traffic that cloud-only deep neural network processing demands. For future work, they propose to not use Binarized Neural Networks in order to simplify output on the cloud, since

the cloud is not as restricted on memory use and processing power. This could help make the overall DDNN more accurate.

2. Critique.

There is one severe issue of explanation which I think this paper greatly failed. The authors did a very poor job at describing the aggregator device and just exactly what it does and where it is along the internet network. They mention it as part of several schema, but it took me some time and even now I have to make some assumptions about where exactly the aggregator device sits on the internet network – is it in the cloud? Is it attached to the local network of the end devices? Is it technically an edge server that does minimal analysis and not including neural network computation? Those are the questions that the author fails to answer and only does a little bit of a better explanation of the aggregator during the methodology section. However, the design schema portion was really well written until they mention the aggregator and never formally described the device and where it was located.

3. Synthesis.

The idea of a distributed neural network that incorporates a distributed internet network is a very powerful concept that seems to naturally fit together. I think that jointly training the full model with partial parts of the model all on the cloud is an excellent idea. However, this sort of distributed network seems to eventually require that it incorporates the ability of the distributed system to also train the model instead of only using it. I say this only because some applications may see large environmental changes that would potentially require a model to re-trained or trained more to maintain a required level of accuracy. Of course the model can be retrained on the cloud, but then that requires downtime; so although I know things are already so complex even with just distributed a trained model, I predict some day in the near future we need to talk more about distributed models that can both use the model to infer but also that can train the model.