CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday – August 16, 2019

Improving Simple Models with Confidence Profiles

In this paper, we propose a new method called ProfWeight for transferring information from a pre-trained deep neural network that has a high test accuracy to asimpler interpretable model or a very shallow network of low complexity and apriori low test accuracy. We are motivated by applications in interpretability andmodel deployment in severely memory constrained environments (like sensors).Our method uses linear probes to generate confidence scores through flattenedintermediate representations. Our transfer method involves a theoretically justified weighting of samples during the training of the simple model using confidence scores of these intermediate layers. The value of our method is first demonstrated on CIFAR-10, where our weighting method significantly improves (3-4%) networks with only a fraction of the number of Resnet blocks of a complex Resnet model. We further demonstrate operationally significant results on a real manufacturing problem, where we dramatically increase the test accuracy of a CART model (the domain standard) by roughly 13%.