Contribution ID: **11** Type: **not specified**

Google DeepMind: Compressing neural networks

Monday 15 April 2019 15:05 (45 minutes)

Empirically it has been observed numerous times that trained neural networks often have high degrees of parameter-redundancy. It remains an open theoretical question why this parameter redundancy cannot be reduced before training by using smaller neural networks. On the other hand, the recent scientific literature reports a plethora of practical methods to "compress" neural networks during or after training with (almost) arbitrarily small sacrifices to task performance while significantly reducing the computational demands of a neural network model. This talk gives an overview over the field of neural network compression methods and then introduces one family of approaches based on Bayesian neural networks. Some of the appealing theoretical properties of Bayesian approaches to neural network compression are discussed and practical implementations for modern neural network training are sketched. The talk concludes with discussing difficulties with neural network compression in practice and an outlook towards exploiting noise and redundancy in the data for more compute-efficient neural networks.

Preferred contribution length

Primary author: Dr GENEWEIN, Tim (DeepMind)

Presenter: Dr GENEWEIN, Tim (DeepMind)

Session Classification: Industry talks and panel