Contribution ID: **191** Type: **Poster + Flashtalk**

Predicting the trainability of deep neural networks with reconstruction entropy

An important challenge in machine learning is to predict the initial conditions under which a given neural network will be trainable. We present a method for predicting the trainable regime in parameter space for deep feedforward neural networks (DNNs) based on reconstructing the input from subsequent activation layers via a cascade of single-layer auxiliary networks. We show that a single epoch of training of the shallow cascade networks is sufficient to predict the trainability of the deep feedforward network on a range of datasets (MNIST, CIFAR10, FashionMNIST, and white noise), thereby providing a significant reduction in overall training time. We achieve this by computing the relative entropy between reconstructed images and the original inputs, and show that this probe of information loss is sensitive to the phase behaviour of the network. We further demonstrate that this method generalizes to residual neural networks (ResNets) and convolutional neural networks (CNNs). Moreover, our method illustrates the network's decision making process by displaying the changes performed on the input data at each layer, which we demonstrate for both a DNN trained on MNIST and the vgg16 CNN trained on the ImageNet dataset. Our results provide a technique for significantly accelerating the training of large neural networks.

AI keywords

explainable AI, hyperparameter optimization, entropy

Primary authors: Prof. ERDMENGER, Johanna (University of Wuerzburg); Prof. JEFFERSON, Ro (Utrecht University); THURN, Yanick (Deutsch)

Presenter: THURN, Yanick (Deutsch)

Track Classification: Explainability & Theory