Contribution ID: **17** Type: **not specified**

Squeezing Deep Learning onto a Phone

Friday, 26 February 2016 12:00 (25 minutes)

Deep learning has recently emerged as one of the most promising techniques for classification, with breakthrough results in fields such as image recognition and natural language processing. However, deep learning calls for a tremendous amount of resources, chiefly in the training phase, but also during the inference phase. This may not be an issue when "Google-scale" computing facilities are available, but it hampers the applicability of deep learning in several fields where computing power, or memory capacity, or energy, are constrained. The talk will focus on one such field: mobile computing, where it is clear that client-side machine learning algorithms will play a key role in the next generation of applications, but nontrivial progress is required to tailor such algorithms to a limited computing or power budget. The talk will present preliminary observations and measurements taken from an image recognition application with convolutional neural networks. As HPC is also being infiltrated by mobile technologies, with ARM processors expected to appear in the Green500 list any time soon, such observations acquire a more general significance.

Presenter: FANTOZZI, Carlo (UNIPD)