Ultimate tensorization: compressing convolutional and FC layers alike

Timur Garipov[1] Dmitry Podoprikhin[1,2]
Alexander Novikov[3] Dmitry Vetrov[2,3]

[1]Moscow State University,
Moscow, Russia

Yandex

[2]Yandex LLC,
Moscow, Russia

[3]Higher School of Economics,
Moscow, Russia

Why compress convolutions?

Large neural networks:

- are expensive to download to smartphones;
- drain battery.

Why compress convolutions?

Large neural networks:

- are expensive to download to smartphones;
- drain battery.

We focus on compressing convolutional layers because:

- several modern architectures (Inception, ResNet) lack fully-connected layers;
- we can already compress fully-connected layers to move the bottleneck into convolutional layers[1].

[1] Novikov et al. *Tensorizing Neural Networks*, NIPS-15

How to compress convolutions?