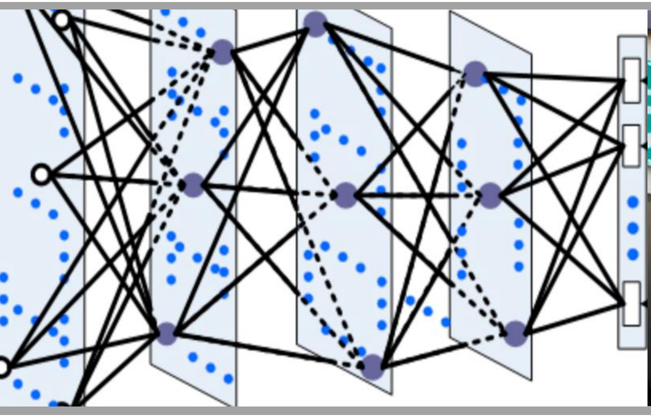
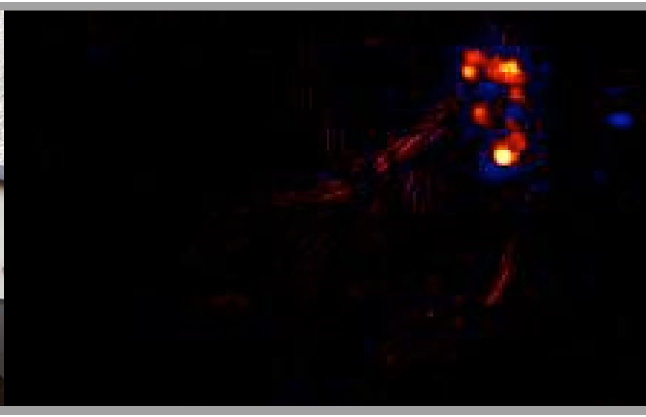
Fraunhofer
Heinrich Hertz Institute

Efficient, Distributed and Interpretable Deep Learning

Dr. Wojciech Samek
Fraunhofer HHI, Machine Learning Group



Workshop “Impact of AI on ICT Infrastructures” 25th April, Xi’an, China

Today’s AI systems

Today’s AI has “superhuman” performance

Most success in image & nlp domain

Key ingredients for the success:

- Huge amounts of training data
- Very deep (black-box) models
- Incredible computing power

Can we also expect such a revolution in ICT ?

Yes, but …

ICT settings are slightly different

Key ingredients for the success:

- Huge amounts of training data
- Very deep (black-box) models
- Incredible computing power

data often distributed
—> distributed learning

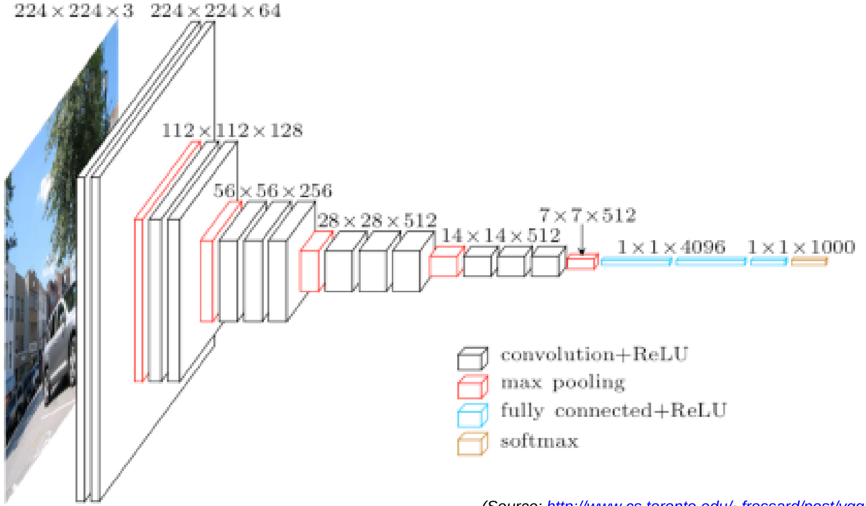
using black-boxes not an option
—> interpretable learning

not available (e.g. mobile devices)
—> efficient learning

Efficient Deep Learning

DNNs are large and energy hungry

DNN with Millions of weight parameters
- large size
- energy-hungry training & inference
- many floating point operations



(Source: http://www.cs.toronto.edu/~frossard/post/vgg16/)

For instance, VGG16
- 16 weight layers
- 138 000 000 parameters
- 553 MB (uncompressed)
- 30940 M float operations (sum+mult) for inference

—> 71 mJ just for the float operations on 45nm CMOS process

DNNs are large and energy hungry

What can we do to bring deep learning to ICT ?

1. Design optimized hardware

Qualcomm's deep learning SDK will mean more AI on your smartphone

Chip could bring deep learning to mobile devices

A new MIT computer chip could allow your smartphone to do complex AI tasks

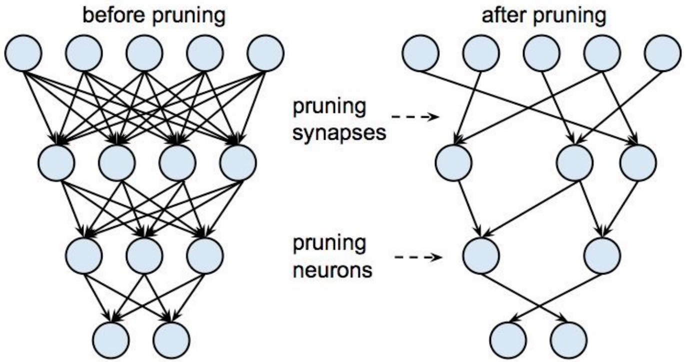
Energy-friendly chip can perform powerful artificial-intelligence tasks

2. Reduce the complexity of the DNN

Popular research topic …

Reducing the complexity of DNNs

1. Network Pruning



2. Weight Quantization

$$\begin{pmatrix} 0 & 4 & 0 & 0 & 0 & 4 & 0 & 4 & 0 & 0 \\ 0 & 2 & 4 & 0 & 0 & 4 & 2 & 0 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 4 & 0 & 0 & 0 & 4 & 0 & 4 \\ 4 & 0 & 0 & 0 & 2 & 0 & 0 & 4 & 2 & 2 \end{pmatrix}$$

3. Efficient Encoding

$$W : [4,4,4,2,4,4,2,2,2,2,2,4,4,4,4,4,2,4,2,2]$$
$$colI : [1,5,7,1,2,5,6,8,0,1,7,2,3,7,9,0,4,7,8,9]$$
$$rowPtr : [0,3,8,11,15,20]$$

Sparse data format
- reduces storage
- fast multiplications

But are compressed DNNs really sparse ?

Quantization leads to low entropy weight matrices with *weight sharing* property.

For such matrices, sparse formats may not be the most efficient ones.

Weight sharing property: Subsets of connections share the same weight value.

$$z_i^l = \sum_j^M w_{ij}^l a_j^{l-1}, \quad \xrightarrow{\text{rewriting trick}} \quad z_i^l = \sum_k w_k^l \sum_{j \in J_{ik}^l} a_j^{l-1}$$

New efficient format for compressed DNNs

$$\begin{pmatrix} 0 & 4 & 0 & 0 & 0 & 4 & 0 & 4 & 0 & 0 \\ 0 & 2 & 4 & 0 & 0 & 4 & 2 & 0 & 2 & 0 \\ 2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 4 & 0 & 0 & 0 & 4 & 0 & 4 \\ 4 & 0 & 0 & 0 & 2 & 0 & 0 & 4 & 2 & 2 \end{pmatrix}$$

more efficient encoding
of low entropy matrices

$$\begin{aligned} W &: [4, 2] \\ colI &: [1,5,7,2,5,1,6,8,0,1,7,2,3,7,9,0,7,4,8,9] \\ wI &: [0,0,1,1,0,0,1] \\ wPtr &: [0,3,5,8,11,15,17,20] \\ rowPtr &: [0,1,3,4,5,7] \end{aligned}$$

VGG-16
size: 553 MB, acc: 68.73 %,
ops: 30940 M, energy: 71 mJ

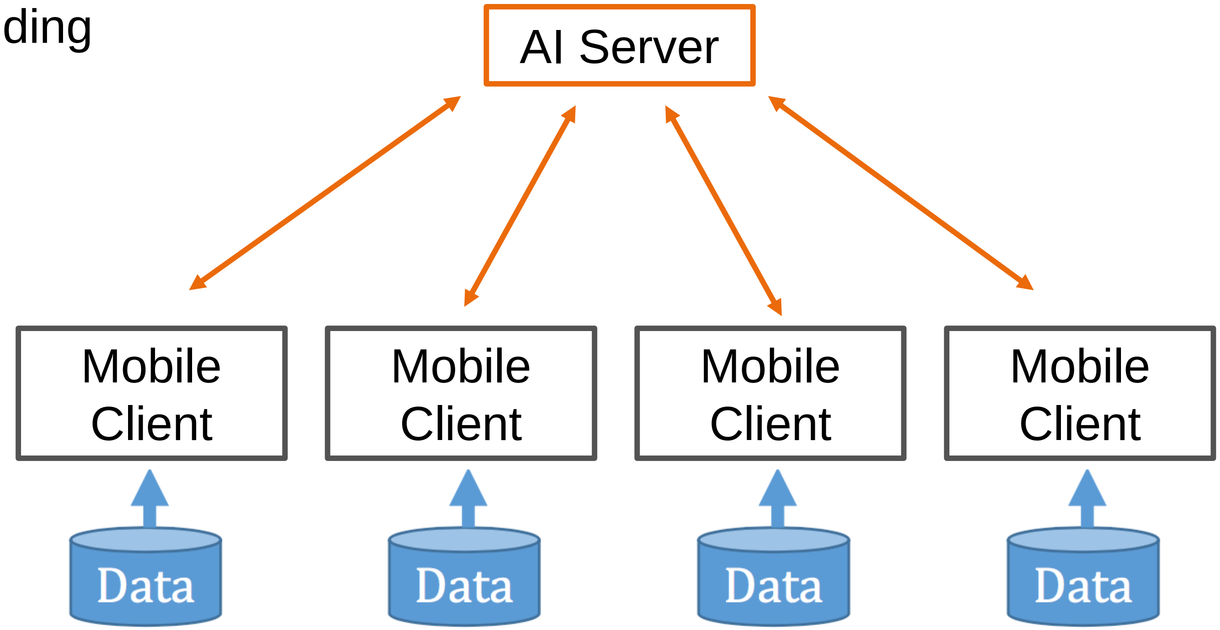
Compression + sparse format
size: 17.8 MB, acc: 68.83 %,
ops: 10081 M, energy: 22 mJ

Compression + Our format
size: 12.8 MB, acc: 68.83 %,
ops: 7225 M, energy: 16 mJ

Distributed Deep Learning

Distributed Training

Our goal
- train a model without sending client data to the server
- minimize communication overhead



Distributed Training

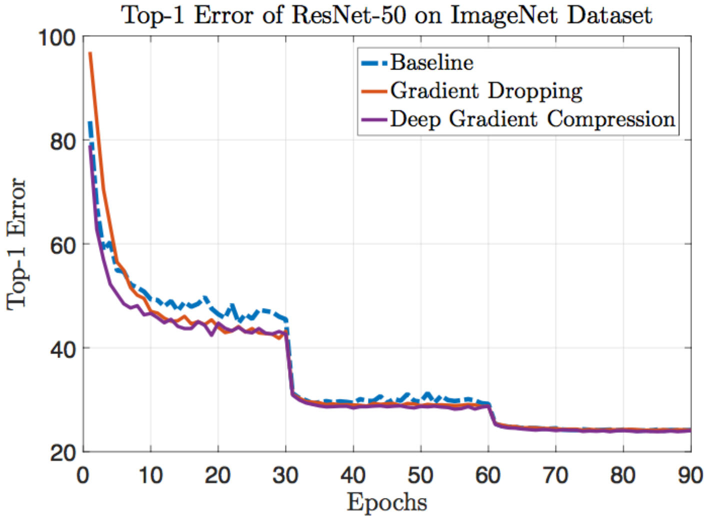
Training algorithm

1. Initialize all clients with the same W
2. Compute weight updates ΔW locally
   and send them to the server
3. Update W and send it to the clients

iterate

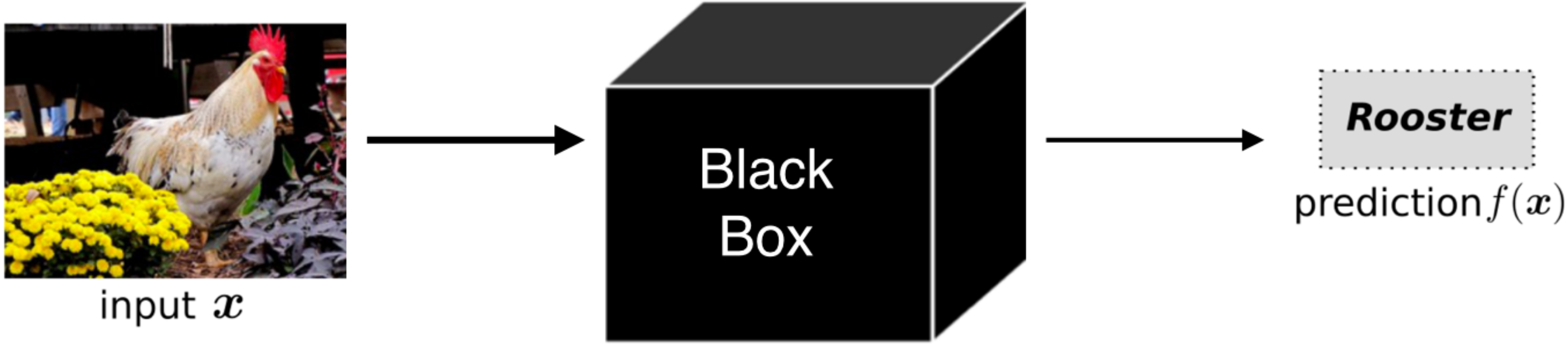
It even works if gradient is highly sparsified (99.9 %)
(see Lin et al. 2018)

We have very promising extension of this approach.



Interpretable Deep Learning

Can we trust these black boxes ?



verify system

legal aspects

learn new strategies

understand weaknesses

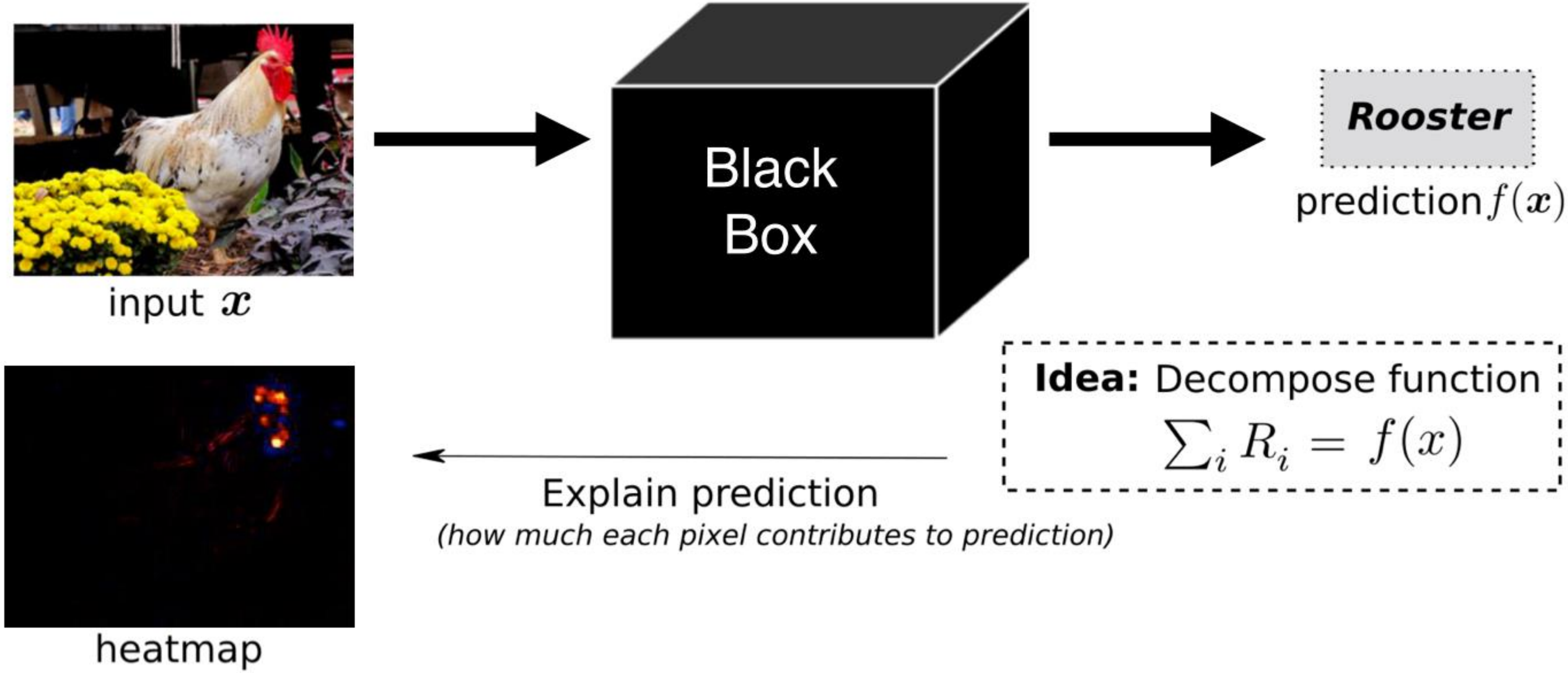
Can we trust these black boxes ?

Is the way error is measured a satisfying specification of the problem?

Are we measuring the error on the true data distribution?

$$\min_{f \in \mathcal{F}} \int_{\mathbf{x},y} \|f(\mathbf{x}) - y\|^2 \, dp(\mathbf{x}, y)$$

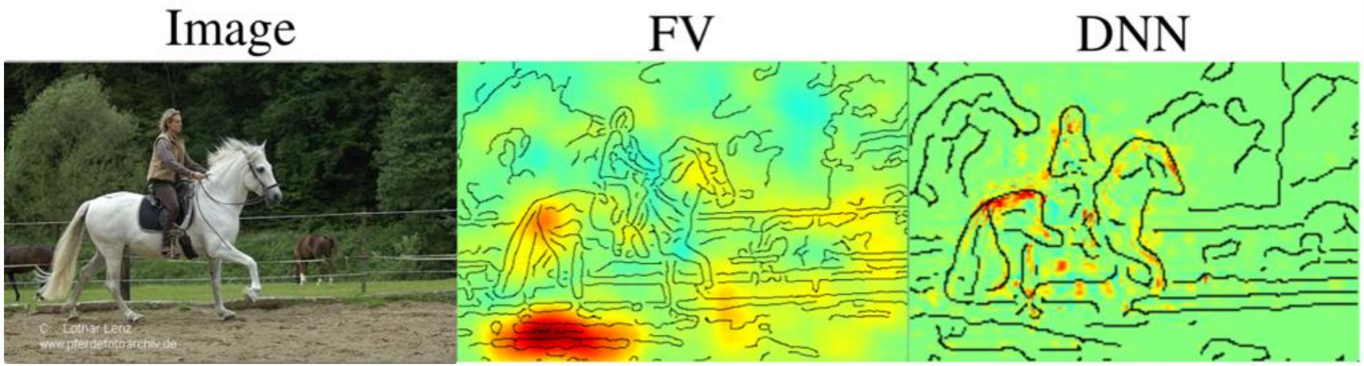
Can we trust these black boxes ?



Opening the black box

Test error for various classes:

	aeroplane	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	**chair**	**cow**	**diningtable**	**dog**	**horse**	**motorbike**
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%



(Lapuschkin et al., 2016)

Upcoming tutorials on interpretability

CVPR 2018

MICCAI 2018 Granada SPAIN

ICIP 2018

Thank you for your attention

Questions ???

Contact Information:

Wojciech Samek
Fraunhofer HHI, Machine Learning Group
Einsteinufer 37, 10587 Berlin, Germany
Mail: wojciech.samek@hhi.fraunhofer.de

More information: http://iphome.hhi.de/samek