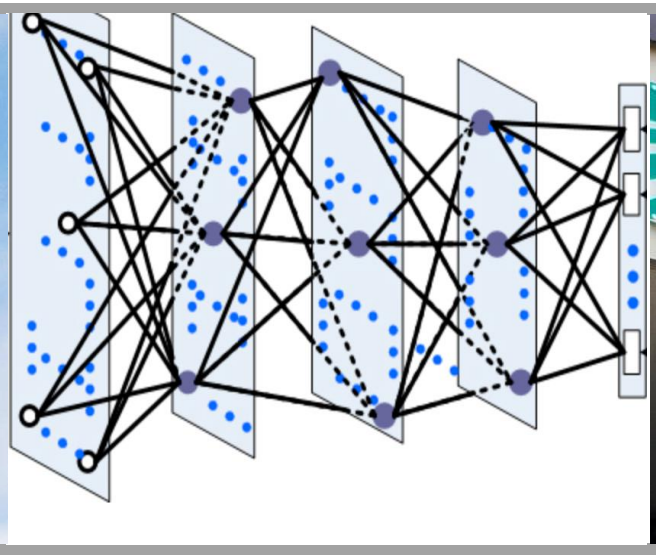
Fraunhofer
Heinrich Hertz Institute

Efficient Deep Learning in Communications

Dr. Wojciech Samek

Fraunhofer HHI, Machine Learning Group



Today's AI Systems

AlphaGo beats Go human champ

Deep Net outperforms humans in image classification

IMAGENET

Dermatologist-level classification of skin cancer with Deep Nets

Revolutionizing Radiology with Deep Learning

DeepStack beats professional poker players

Computer out-plays humans in "doom"

Deep Net beats human at recognizing traffic signs

Today's AI Systems

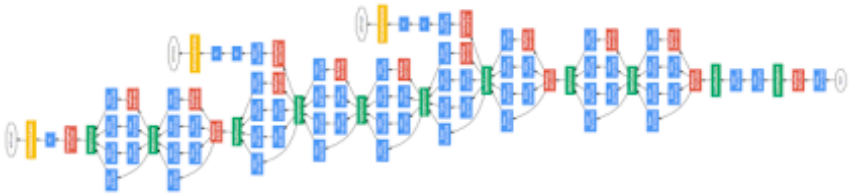
Huge volumes of data

- Millions of labeled examples available

Computing power

- highly parallel processing
- large power consumption (600 Watts per GPU card)

Powerful models



- huge models (up to 137 billion parameters and 1001 layers)
- architectures adapted to images, speech, text …

Communications settings are often different.

ML in Communications

Satellite Communications

Autonomous driving

Smart Data

Smartphones

Internet of Things

5G Networks

Many additional requirements: Small size, efficient execution, low energy consumption …

ML in Communications

Distributed setting

Large nonstationarity

Restricted ressources

Communications costs

Interoperability

Security & privacy

Interpretability

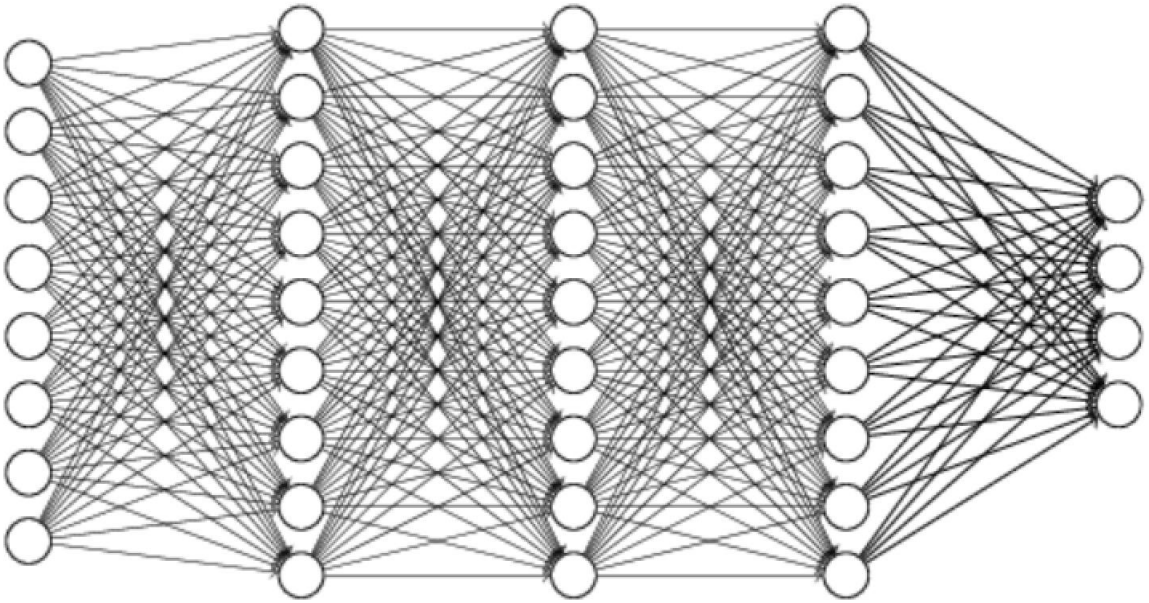
Trustworthiness

…

We need ML techniques which are adapted to communications

But it's not only the algorithms, also:
- protocols
- data formats
- frameworks
- mechanisms
- …

Problem 1: Restricted ressources



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

Many recent work on compressing neural networks by weight quantization.

Problem 1: Restricted resources

$$\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & w_{2,3} & \cdots & w_{2,n} \\ w_{3,1} & w_{3,2} & w_{3,3} & \cdots & w_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & w_{n,3} & \cdots & w_{n,n} \end{bmatrix}$$

quantization

(rate-distortion theory)

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

compressed sparse row format
- reduces storage
- fast multiplications

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

can we do better ?

Problem 1: Restricted resources

RD-theory based weight quantization does not necessarily lead to sparse matrices.

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

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

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

Problem 1: Restricted resources

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

more efficient format
than CSR

$$W : [4, 2]$$
$$colI : [1, 5, 7, 2, 5, 1, 6, 8, 0, 1, 7, 2, 3, 7, 9, 0, 7, 4, 8, 9]$$
$$wI : [0, 0, 1, 1, 0, 0, 1]$$
$$wPtr : [0, 3, 5, 8, 11, 15, 17, 20]$$
$$rowPtr : [0, 1, 3, 4, 5, 7]$$

Problem 1: Restricted resources

iphone8 25 kJ

VGG-16

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

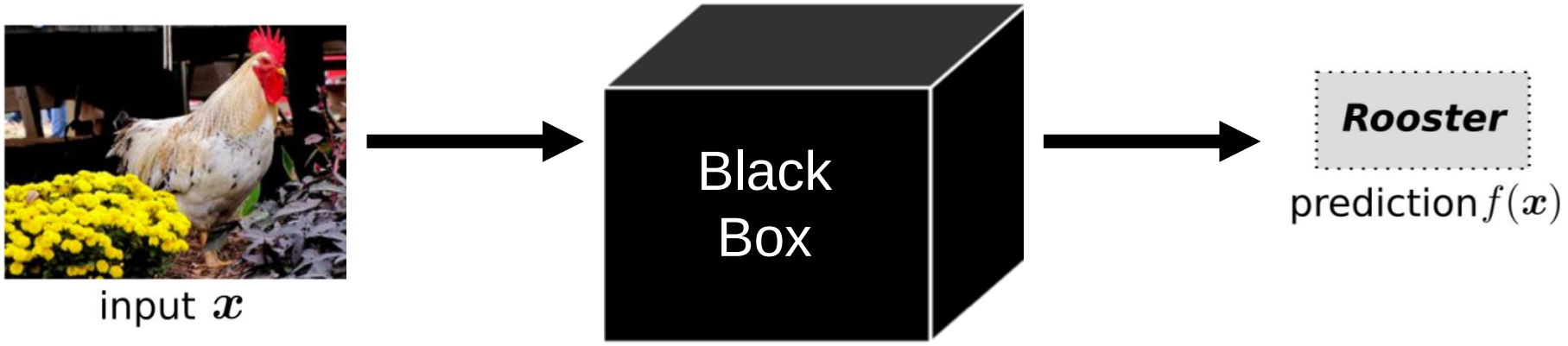
State-of-the-art compression + sparse format

size: 17.8 MB, acc: 68.83 %, ops: 10081 M, energy: 22 mJ

State-of-the-art compression + WS format

size: 12.8 MB, acc: 68.83 %, ops: 7225 M, energy: 16 mJ

Problem 2: Interpretability



verify system

legal aspects

understand weaknesses

learn new strategies

Problem 2: Interpretability

input x

Black Box

Rooster

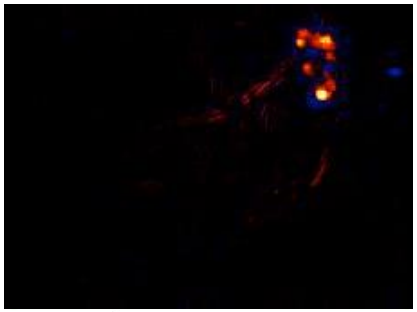
prediction $f(x)$

Idea: Decompose function

$$\sum_i R_i = f(x)$$

Explain prediction

(how much each pixel contributes to prediction)

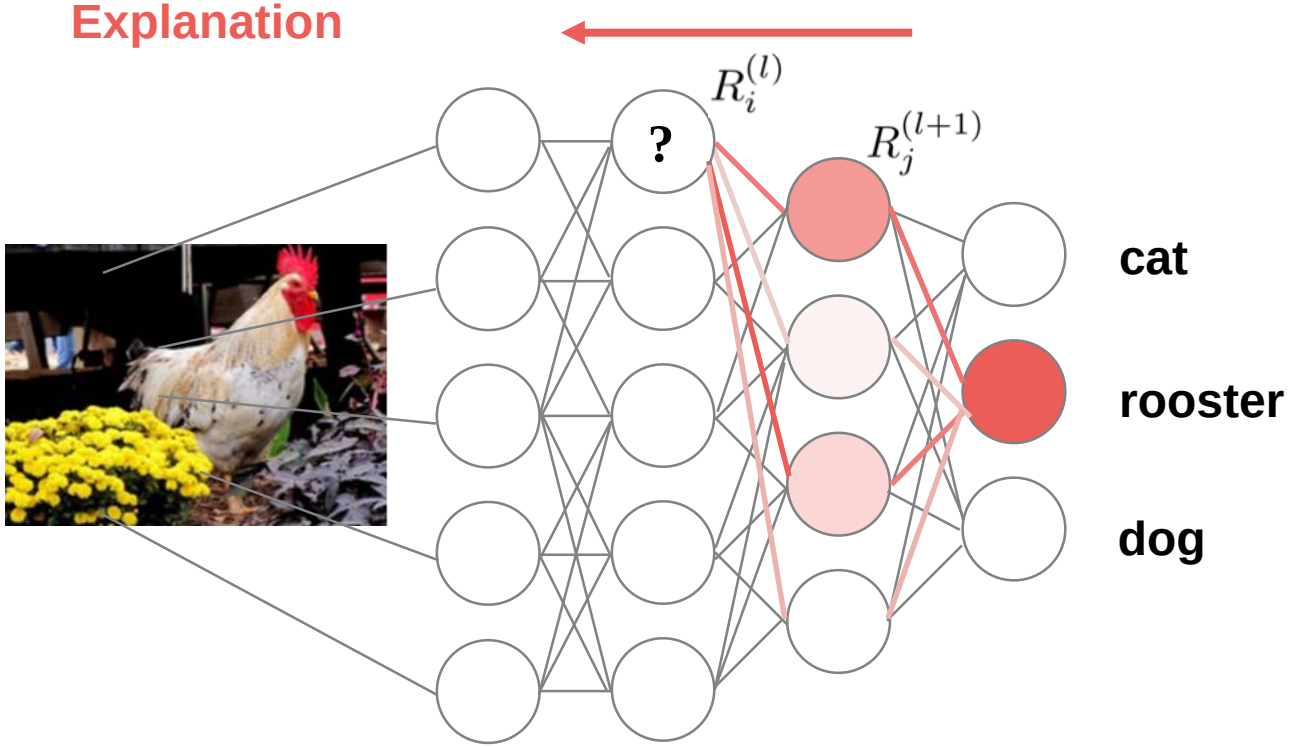


heatmap

Theoretical interpretation: (Deep) Taylor decomposition of neural network

Problem 2: Interpretability

Explanation



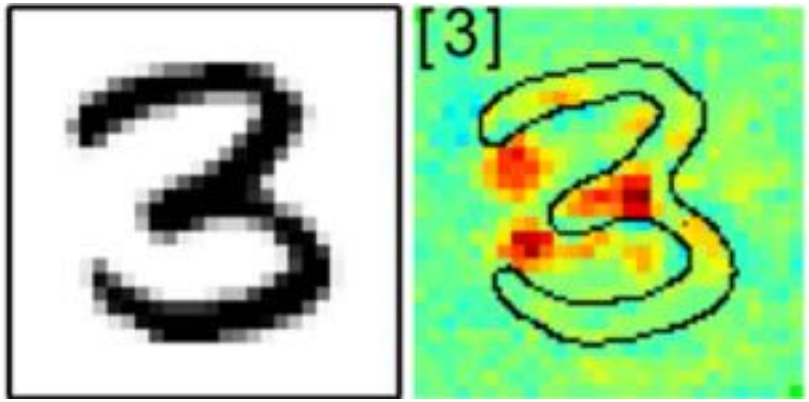
Simple LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \frac{x_i \cdot w_{ij}}{\sum_{i'} x_{i'} \cdot w_{i'j}} R_j^{(l+1)}$$

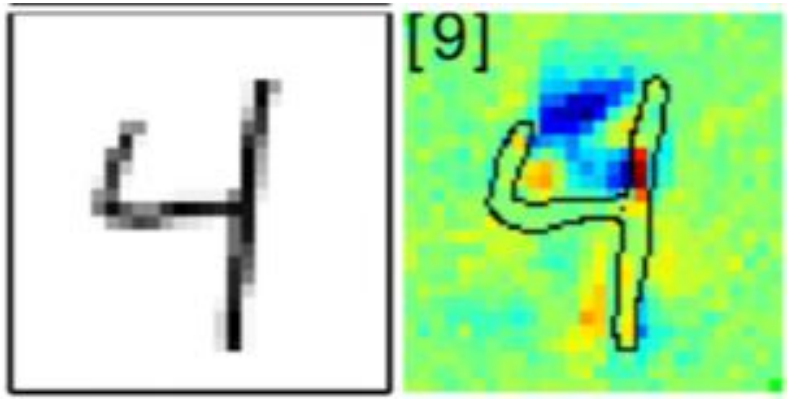
Every neuron gets its "share" of the redistributed relevance

Problem 2: Interpretability

what speaks for / against classification as “3”



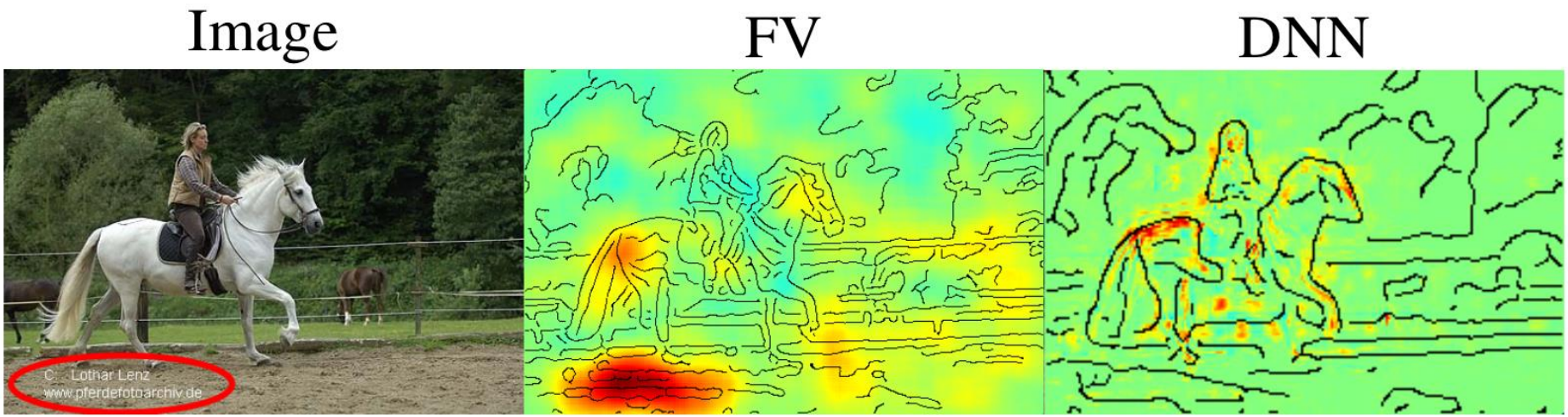
what speaks for / against classification as “9”



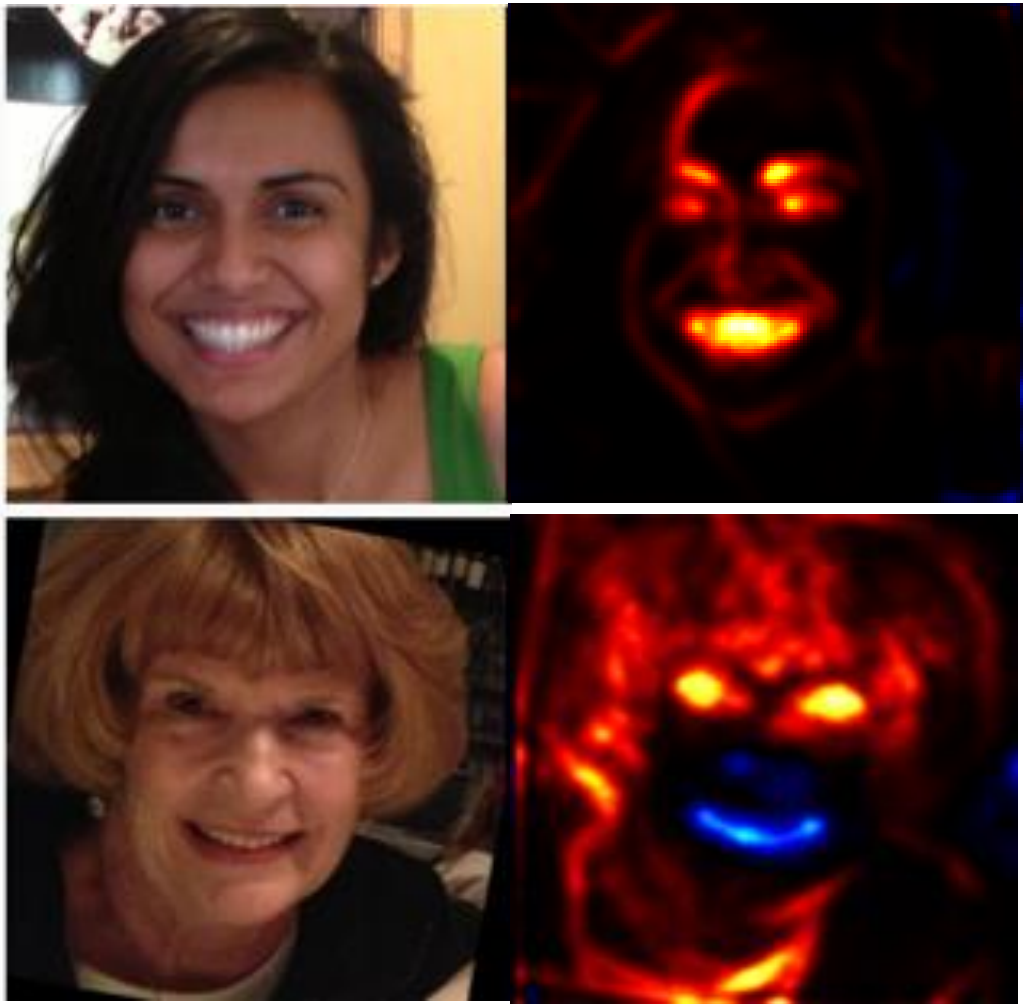
Problem 2: Interpretability

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%



Problem 2: Interpretability

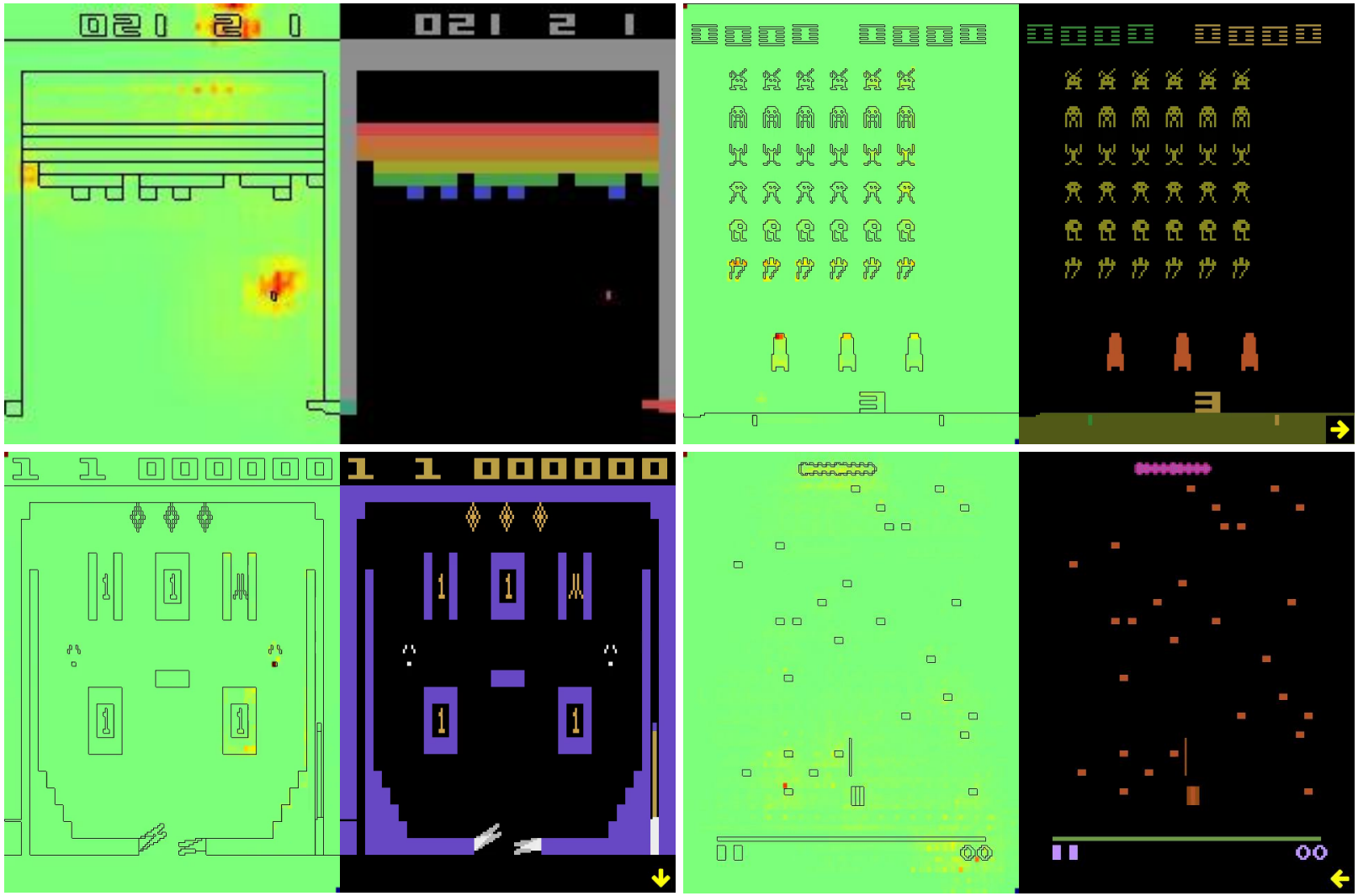


Predictions

25-32 years old

60+ years old

Problem 2: Interpretability



Conclusion

Bringing ML to communications comes with new challenges

AI systems may behave differently than expected

Need for best practices & recommendations (protocols, formats, …)

Thank you for your attention

Questions ???

All our papers available on:

http://iphome.hhi.de/samek

Acknowledgement

Simon Wiedemann (HHI)
Klaus-Robert Müller (TUB)
Grégoire Montavon (TUB)
Sebastian Lapuschkin (HHI)
Leila Arras (HHI)
…