Hierarchical Attribute CNNs

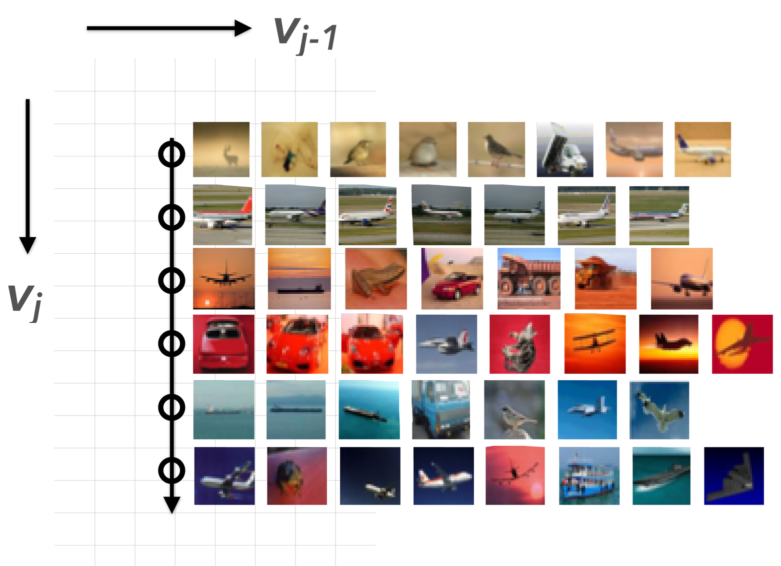
Institute of Informatics, University of Amsterdam
DATA, Département Informatique, Ecole Normale Supérieure/PSL Research University
Jörn-Henrik Jacobsen, Edouard Oyallon, Stéphane Mallat, Arnold W.M. Smeulders
{j.jacobsen,a.w.m.smeulders}@uva.nl, edouard.oyallon@ens.fr

PADL Workshop
ICML 2017 | Sydney

Introduction

- How can we introduce **structure** into deep networks to understand them better?
- *(Mallat, 2016)* proposes to perform high-dim convolutions which organise the channels
- *The Dream:*

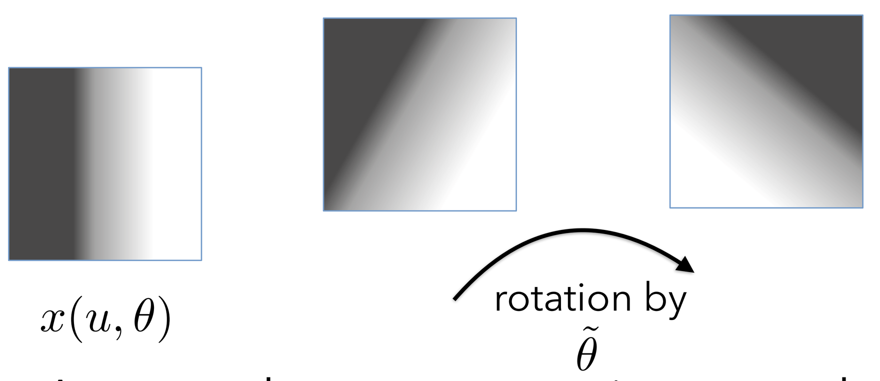
  a) Translations along channels modulate signal attributes
  b) Layers disentangle, increasingly more complex signal attributes with depth



"The Dream"

$x_j(v_{j-1}, v_j)$ channel indices

An Example: *Rotation covariance*



rotation (translation along theta)

$$L_{\tilde{\theta}} x(u,\theta) = x(r_{-\tilde{\theta}} u, \theta - \tilde{\theta})$$

- Assume the representation encodes the rotation (rotation equivariance by L_θ)
- A rotation of the input is a translation of the signal, which is a shift of θ.
- Can we generalise this **beyond** roto-translation?

Multi-dimensional Convolution

- ***u:*** native spatial dimensions
- ***v_j:*** signal attributes (discriminative channels)
- Operators ***W***: convolution along spatial and channel dimensions
- Convolution with filter k corresponds to:

$$W_j x_{j-1} = x_{j-1} \star^{u,v_1,\ldots,v_{j-1}} k_j(u, v_1, \ldots, v_{j-1}) = \sum_{(\tilde{u},\tilde{v})} x_{j-1}(\tilde{u}, \tilde{v}_1, \ldots, \tilde{v}_{j-1}) k_j(u-\tilde{u}, v_1-\tilde{v}_1, \ldots, v_{j-1}-\tilde{v}_{j-1})$$

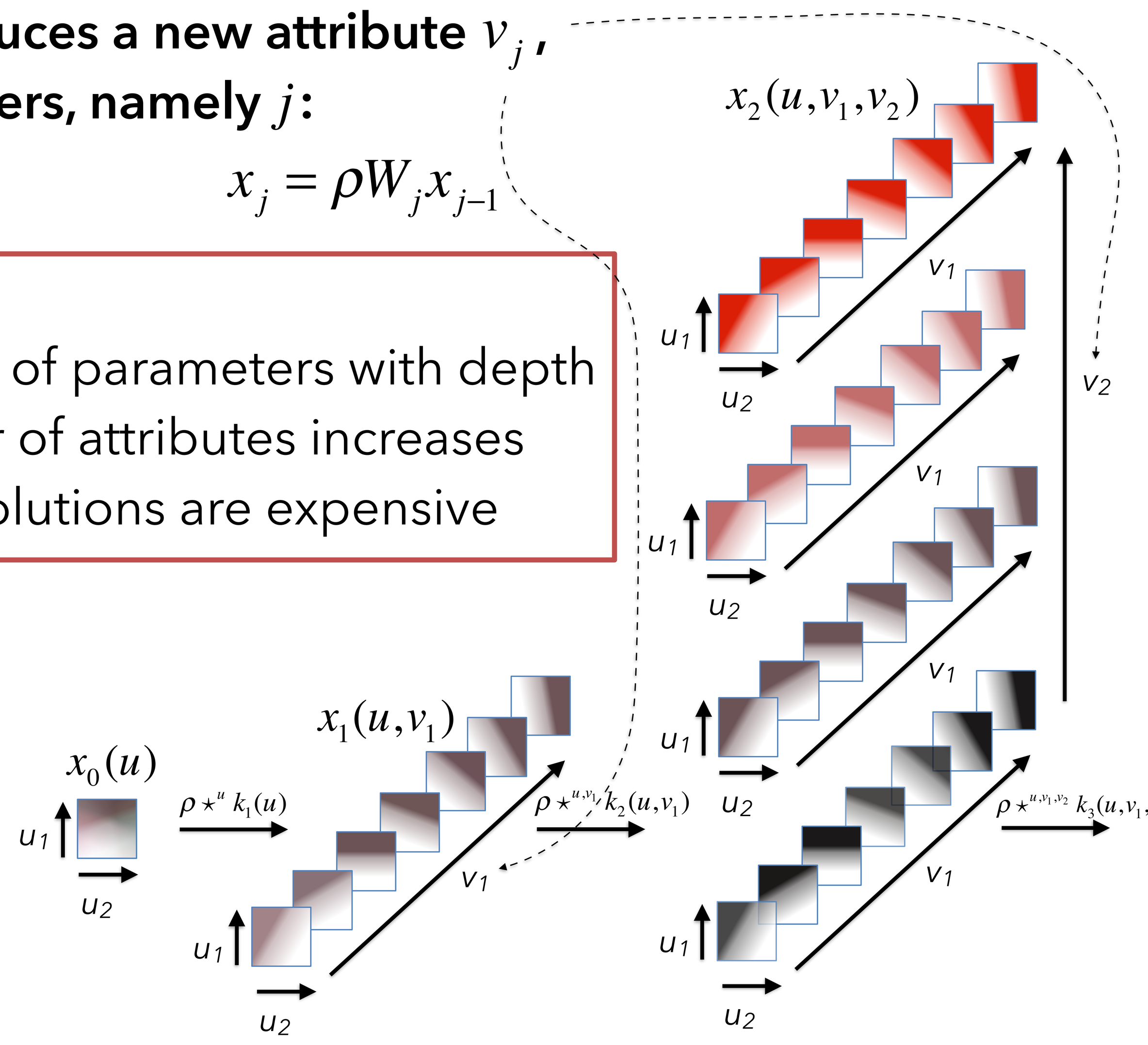
- Covariant with translations along $(v_1, \ldots, v_j)$: $LWx = WLx$

Multiscale Hierarchical CNNs

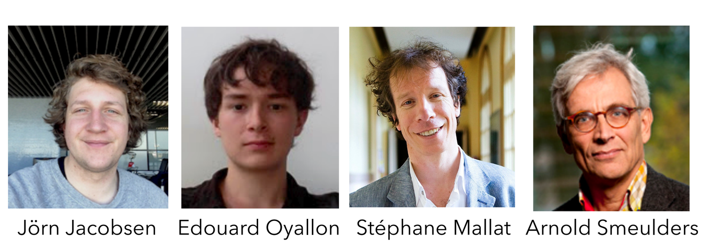
- Class of CNN where one-dimensional channel index is replaced by multi-dimensional vector of attributes: $v = (v_1, \ldots, v_j)$.
- Output of layer ***j*** is represented by: $x_j(u_1, u_2, v_1, \ldots, v_j)$
- All linear operators W_j are convolutions over (u,v), introduced above; **each convolution introduces a new attribute** v_j, **the index of the new filters, namely** j:

$$x_j = \rho W_j x_{j-1}$$

- Issues:
  - Exponential increase of parameters with depth
  - Large layers, number of attributes increases
  - N-dimensional convolutions are expensive



Authors



Hierarchical Attribute CNNs

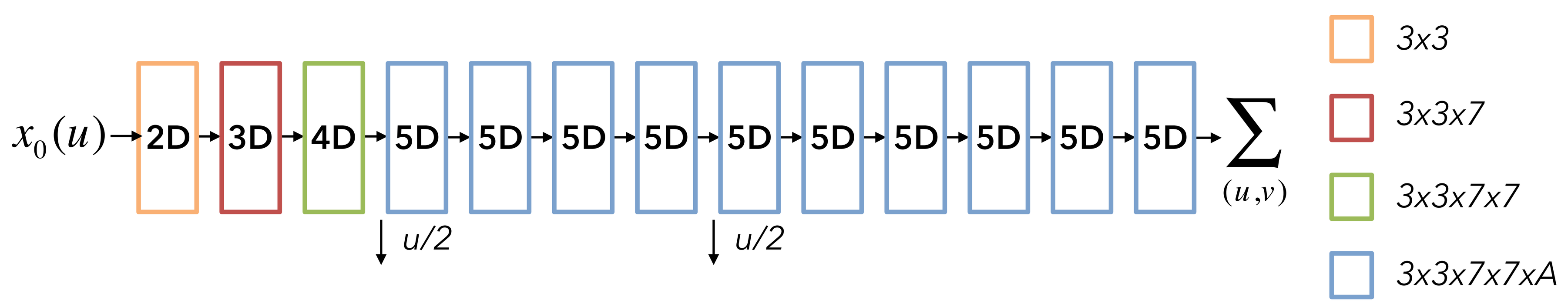
- Overcomes implementation **limitations** of Multiscale Hierarchical CNNs and introduces **increasing** invariance along attributes
- Core idea: Eliminate dependency to all attributes, but last three:

$$x_j(v_{j-2}, v_{j-1}, v_j) \xrightarrow{W_j} x_{j+1}(v_{j+1}, v_j, v_{j-1})$$

will be removed via an averaging in the linear operator; new channel

- New attributes are created, convolved along twice and eliminated

- ***Good performance requires regularisation and careful implementation***
- Architecture:

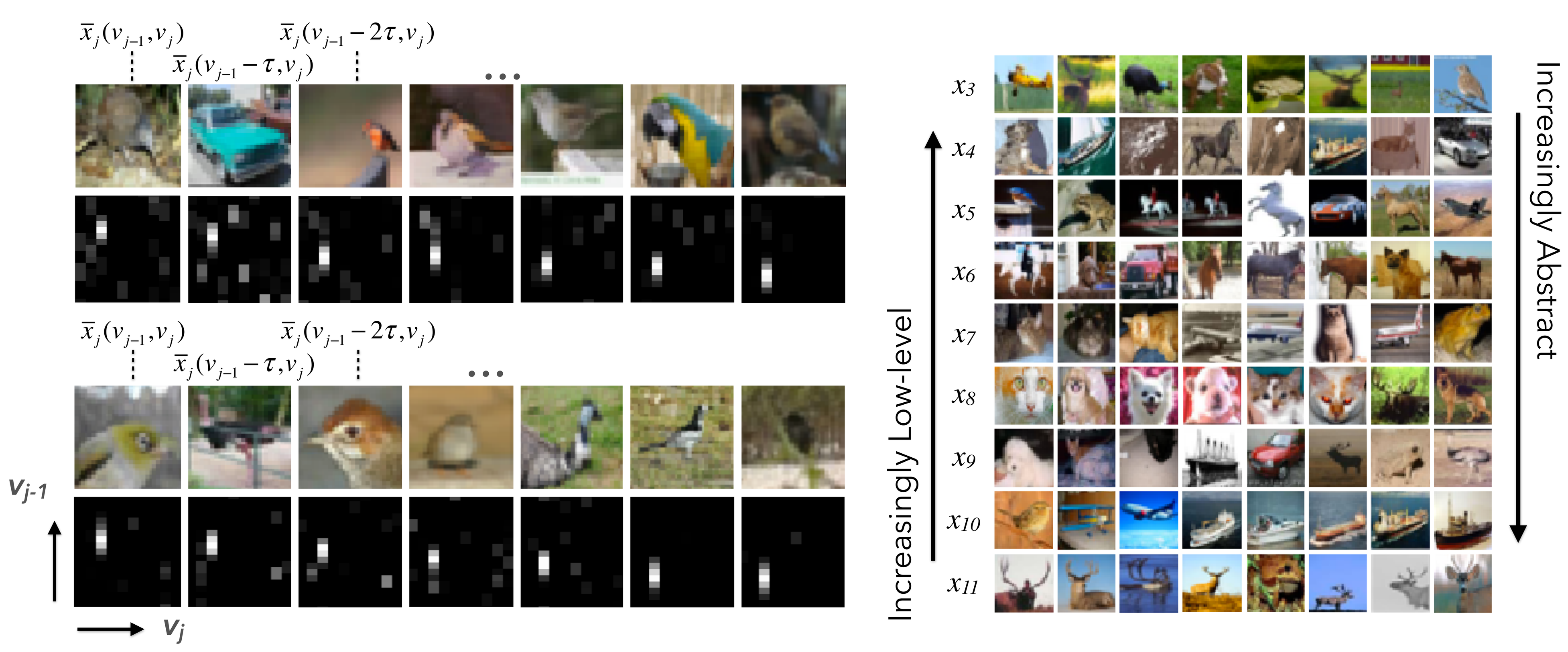


Numerical Results

- Hierarchical Attribute CNNs dramatically reduce trainable parameters
  –> ***Organisation is effective!!***
- Improved classification performance vs. #Parameters trade-off / complexity of training, compared to FitNet or teacher student methods
- Cifar-10 classification performance on par with comparable CNNs

Model	HCNN	HCNN(+)	All-CNN	ResNet-20	NiN
% Acc	91.23	92.3	92.75	91.25	91.20
#params	0.1M	0.3M	1.3M	0.3M	1.0M

- Investigating translations ***L***:



Conclusions

- Highly structured: They achieve good classification performance with much less parameters than comparable CNNs
- Hierarchical Attribute CNNs give a framework to study deep network representations
- But attributes are hard to interpret!! Is there a theoretical limitation, are we missing an idea or is there something deeper?

Code

- https://github.com/jhjacobsen/HierarchicalCNN