A Probabilistic Generative Approach To Invariant Visual Inference and Learning

Zhenwen Dai and Jörg Lücke
FIAS, Goethe-Universität Frankfurt am Main, Germany

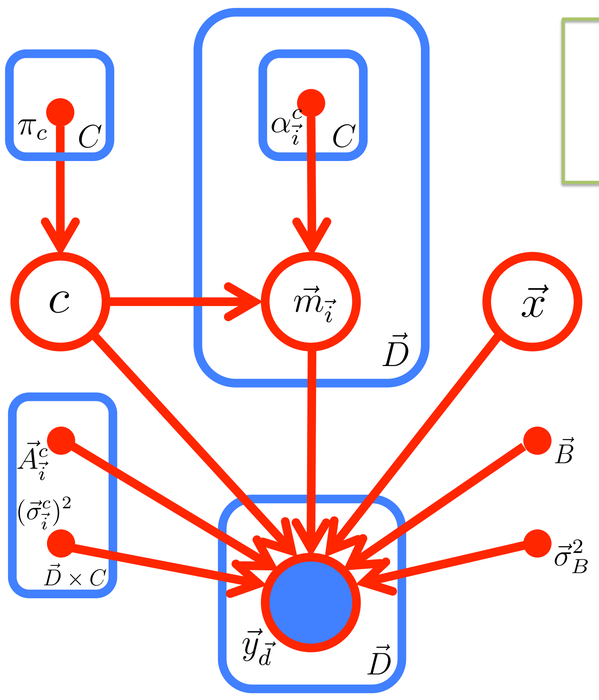
Abstract

We study unsupervised learning in a probabilistic generative approach that explicitly addresses the translation invariance of objects in visual data. The investigated generative model autonomously learns from unlabeled data with object identity and position.

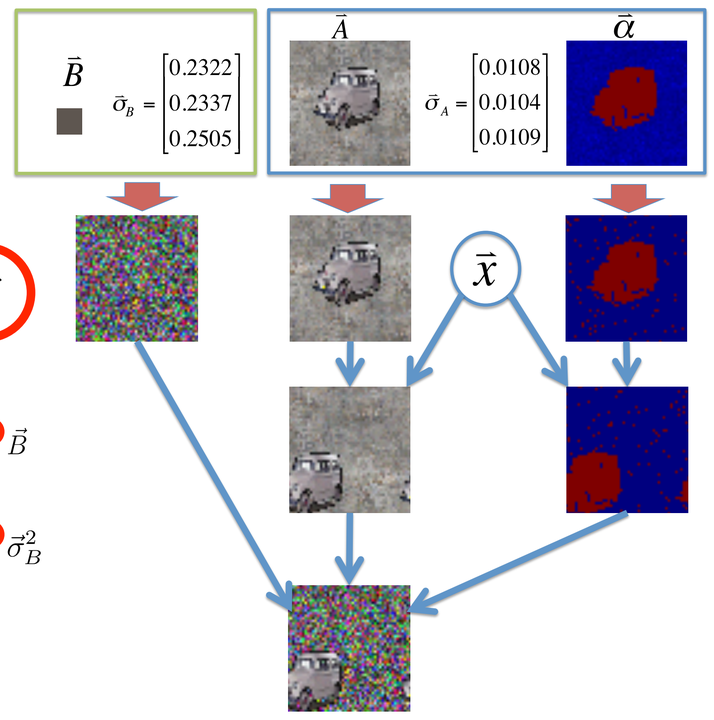
Model features:
- object translation explicitly defined as a hidden variable
- object mask defined as binary hidden variables
- learning multiple objects by a mixture model
- maximum likelihood with exact EM on a GPU cluster

Our algorithm successfully extracts desired objects from both synthetic and real image sequences.

Graphical Model



A Generation Example



Introduction

Inference and learning from visual data is a challenging task because of noise and the data's ambiguity. The most advanced vision systems to date are the sensory visual circuitries of higher vertebrates. To understand and to rebuild biological systems, they have been modeled using approaches from artificial intelligence, artificial neural networks, and probabilistic models. In terms of how the problem of invariant recognition is approached, these models can coarsely be grouped into two classes: models that passively treat invariances (e.g., [1,2]) and models that actively address the typical transformation invariances of object identities (e.g., [3-8]). The former approaches are often feed-forward while the latter approaches are usually recurrent.
In this work we study a probabilistic generative approach that explicitly addresses the translation invariance of objects in visual data. Object location is modeled using an explicit hidden variable while the object itself is encoded by a specific spatial combination of features.

The Generative Model

Variables:

$\vec{x} \in \{1, \cdots, D_1\} \times \{1, \cdots, D_2\}$	position variable
$c \in \{1, \cdots, C\}$	class variable
$\vec{m} \in \{0,1\}^{D_1 \times D_2}$	mask variables
$\vec{y} = \{\vec{y}_{\vec{d}}\}_{\vec{d}=\vec{1},\ldots,\vec{d}}$	observed datum

Parameters:

$\vec{D} = [D_1, D_2]^T$	image feature resolution
$\pi_c \in (0,1)$	mixture parameters
$\vec{\alpha}^c \in (0,1)^{D_1 \times D_2}$	mask prior parameters
$\vec{A}^c, (\vec{\sigma}^c)^2$	mean and variance for object *c*
$\vec{B}, \vec{\sigma}_B^2$	mean and variance for background
$\Theta = (\pi_1, \vec{\alpha}^1, \vec{A}^1, (\vec{\sigma}^1)^2, \cdots, \pi_C, \vec{\alpha}^C, \vec{A}^C, (\vec{\sigma}^C)^2, \vec{B}, \vec{\sigma}_B^2)$	the parameter set

$$p(\vec{x}|\Theta) = \frac{1}{D_1 D_2}$$

$$p(c|\Theta) = \pi_c$$

$$p(m_{\vec{i}}|c,\Theta) = (\alpha_{\vec{i}}^c)^{m_{\vec{i}}}(1-\alpha_{\vec{i}}^c)^{(1-m_{\vec{i}})}$$

$$p(\vec{y}|\vec{x},\vec{m},c,\Theta) = \prod_{\vec{d}=\vec{1}}^{\vec{D}} \mathcal{N}(\vec{y}_{\vec{d}};\vec{\mu}_{\vec{d}},\Sigma_{\vec{d}})$$

$$\Sigma_{\vec{d}} = \begin{pmatrix} (\vec{\sigma}_{\vec{d}}^{(1)})^2 & & & \\ & (\vec{\sigma}_{\vec{d}}^{(2)})^2 & & \\ & & \ddots & \\ & & & (\vec{\sigma}_{\vec{d}}^{(F)})^2 \end{pmatrix}, \vec{\sigma}_{\vec{d}}^2 = \begin{pmatrix} (\vec{\sigma}_{\vec{d}}^{(1)})^2 \\ (\vec{\sigma}_{\vec{d}}^{(2)})^2 \\ \vdots \\ (\vec{\sigma}_{\vec{d}}^{(F)})^2 \end{pmatrix}$$

$$\vec{\mu}_{\vec{d}} = \vec{A}^c_{(\vec{d}-\vec{x})} m_{(\vec{d}-\vec{x})} + \vec{B}(1 - m_{(\vec{d}-\vec{x})})$$

$$\vec{\sigma}_{\vec{d}}^2 = (\vec{\sigma}^c_{(\vec{d}-\vec{x})})^2 m_{(\vec{d}-\vec{x})} + \vec{\sigma}_B^2(1 - m_{(\vec{d}-\vec{x})})$$

(Note that $\vec{1} = [1,1]^T$)

The Learning Algorithm

The learning algorithm aims to maximize the data likelihood. The following update rule is derived by using exact EM.

E-step:

$$p_\Theta^{(n)}(m_{\vec{i}}|\vec{x},c) = p(m_{\vec{i}}|\vec{x},c,\vec{y}^{(n)},\Theta) \qquad p_\Theta^{(n)}(\vec{x},c) = p(\vec{x},c|\vec{y}^{(n)},\Theta)$$

M-step:

$$\pi_c = \frac{1}{N}\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c)$$

$$\alpha_{\vec{i}}^c = \frac{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) p_\Theta^{(n)}(m_{\vec{i}}=1|\vec{x},c)}{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c)}$$

$$\vec{A}_{\vec{i}}^c = \frac{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) p_\Theta^{(n)}(m_{\vec{i}}=1|\vec{x},c) \vec{y}^{(n)}_{(\vec{i}+\vec{x})}}{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) p_\Theta^{(n)}(m_{\vec{i}}=1|\vec{x},c)}$$

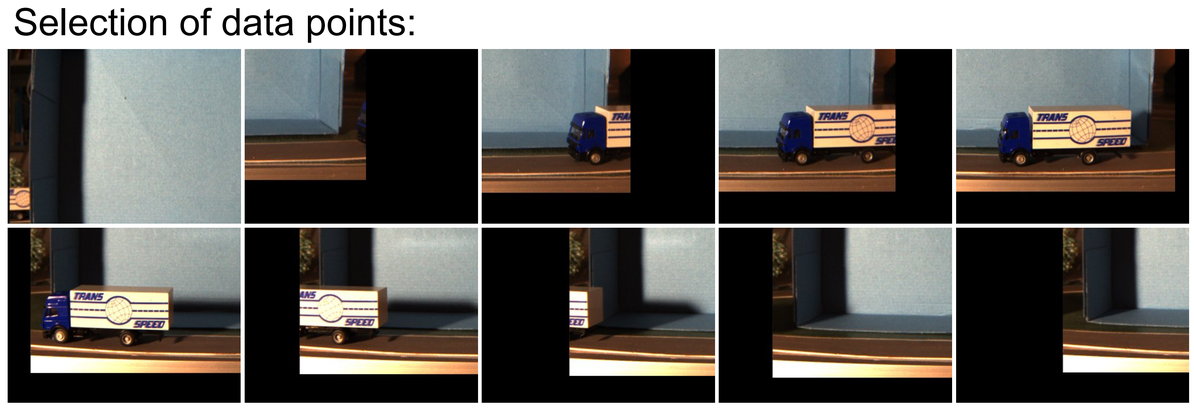
$$\vec{B} = \frac{\sum_n \sum_c \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) \sum_{\vec{i}} p_\Theta^{(n)}(m_{\vec{i}}=0|\vec{x},c) \vec{y}^{(n)}_{(\vec{i}+\vec{x})}}{\sum_n \sum_c \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) \sum_{\vec{i}} p_\Theta^{(n)}(m_{\vec{i}}=0|\vec{x},c)}$$

$$(\vec{\sigma}_{\vec{i}}^c)^2 = \frac{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) p_\Theta^{(n)}(m_{\vec{i}}=1|\vec{x},c) (\vec{y}^{(n)}_{(\vec{i}+\vec{x})} - \vec{A}_{\vec{i}}^c)^2}{\sum_n \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) p_\Theta^{(n)}(m_{\vec{i}}=1|\vec{x},c)}$$

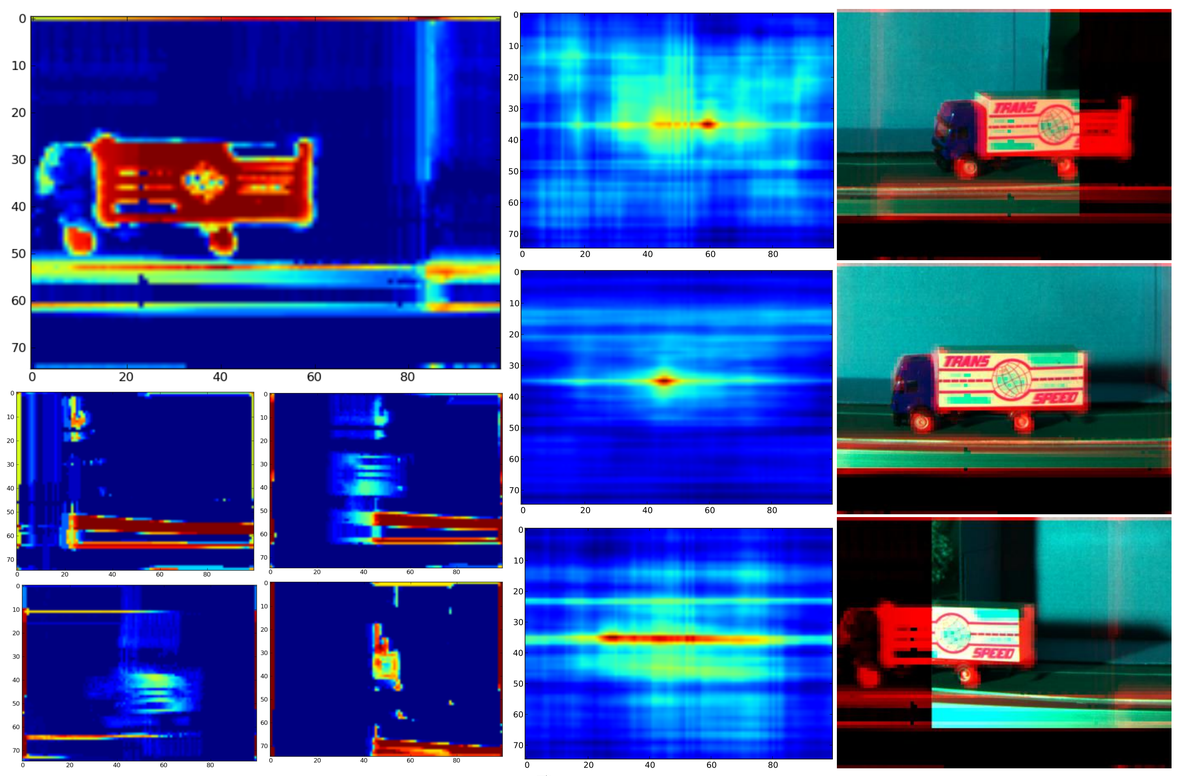
$$\vec{\sigma}_B^2 = \frac{\sum_n \sum_c \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) \sum_{\vec{i}} p_\Theta^{(n)}(m_{\vec{i}}=0|\vec{x},c) (\vec{y}^{(n)}_{(\vec{i}+\vec{x})} - \vec{B})^2}{\sum_n \sum_c \sum_{\vec{x}} p_\Theta^{(n)}(\vec{x},c) \sum_{\vec{i}} p_\Theta^{(n)}(m_{\vec{i}}=0|\vec{x},c)}$$

Numerical Experiments

Selection of data points:



object mask | $\log p(\vec{x}|c=0,\vec{y},\Theta)$ | MAP*



* In red channel it is the mask at the MAP $\vec{x}$ position.

Discussion

We studied unsupervised learning with translation invariance in visual data. Future work aims at:
- learning more objects by using approximate EM
- learning multiple objects per scene with occlusion

Reference

[1] Y. LeCun, F.J. Huang, and L. Bottou, *Proc. of IEEE CVPR*, 97–104, 2004.
[2] M. Riesenhuber, and T. Poggio, *Nature Neuroscience,* 211: 1019–1025, 1999.
[3] B.A. Olshausen, C.H. Anderson, and D.C. Van Essen, *Journal of Neuroscience,* 13: 4700–4719, 1993.
[4] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, *IEEE Trans. Computers,* 42:300-311, 1993.
[5] J. Lücke, C. Keck, and C. von der Malsburg, *Neural Computation,* 20: 2441–2463, 2008.
[6] G. Hinton, *Proc. of IJCAI*, 2, 683–685, 1981.
[7] N. Jojic and B. Frey, *Proc. of CVPR,* 199-206, 2001.
[8] C.K.I. Williams and M.K. Titsias, *Neural Computation,* 16: 1039-1062, 2004.

Acknowledgements. We gratefully acknowledge funding by the German Research Foundation (DFG) in the project LU 1196/4-1 and by the German Federal Ministry of Education and Research (BMBF) in the project 01GQ0840 (BFNT Frankfurt). Furthermore, we gratefully acknowledge support by the Frankfurt Center for Scientic Computing and by the VSI of Frankfurt University for recording visual sequences.