How can Representation Learning deal with the problem of nuisance variability in Computer Vision?

Nikolaos Karianakis, Stefano Soatto
UCLA, Computer Science

Chen Zhou, Yizhou Wang
Peking University, Computer Science

Overview

- The problem of *nuisance variability* is very acute in Computer Vision [1], where even the same object can yield a large variety of images depending on viewpoint, illumination, partial occlusion etc.
- Many representation and deep learning architectures [2, 3, 4] have shown the ability to learn the intrinsic (class) variability despite the presence of significant nuisance variability.
- We are interested in testing the hypothesis that a representation learning architecture is able to train away nuisance variability, which is present in images, owing to changes of viewpoint and illumination, noise, and miscellaneous defects.
- We design an architecture based on the Gated Restricted Boltzmann Machine and we challenge it in Computer Vision problems such as:
  - Occlusion Detection
  - Image Segmentation
  - Depth Estimation

The Model

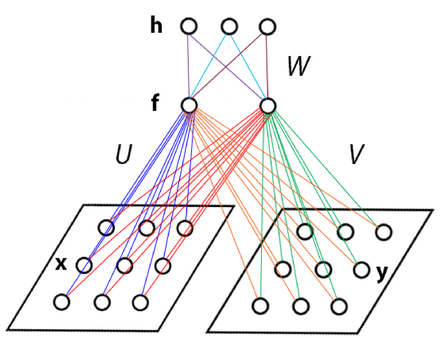
- Our generic one-layer learning architecture is based on the Gated Restricted Boltzmann machine.
- We use a factorized similarity measure [2] among two binary images x, y and the model's hidden layer h:

$$S(\mathbf{x},\mathbf{y},\mathbf{h}) = \sum_{f=1}^{F}((\sum_{i=1}^{I} u_{if}x_i)(\sum_{j=1}^{J} v_{jf}y_j)(\sum_{k=1}^{K} w_{kf}h_k))$$

- The joint probabilty distribution has energy function:

$$E(\mathbf{x},\mathbf{y},\mathbf{h};\theta) = -\sum_{f=1}^{F}(\sum_{i=1}^{I} u_{if}x_i)(\sum_{j=1}^{J} v_{jf}y_j)(\sum_{k=1}^{K} w_{kf}h_k) - \sum_{i=1}^{I} a_i x_i - \sum_{j=1}^{J} b_j y_j - \sum_{k=1}^{K} c_k h_k$$

- Training with *3-way Contrastive Divergence* [4] provides a mechanism that recognize and is able to eliminate small affine transformations and lightning changes between images (nuisance-invariant comparison).

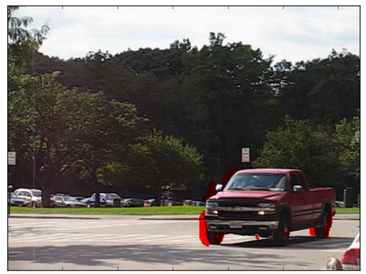


Gated Restricted Boltzmann machine (Gated RBM)

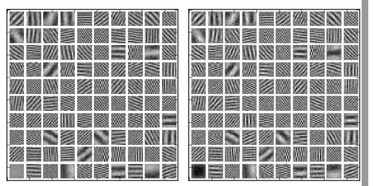
Occlusion Detection

- Occlusion detection is the binary classification task of determining the *co-visibility* from different images (e.g. two sequential video frames) of the same underlying scene.
- The model is trained with training image pairs related by affine transformations, shifts and rotations, scale and illumination variation. The first factors intend to deal with different vantage points where these images are captured from, while the latter one with different lightning conditions.
- The trained model is tested on sequential video frames from Middlebury, Berkeley Moseg and UCL Optical Flow datasets. Comparisons between corresponding patches in the two frames yield the occluded areas, after thresholding a semi-metric based on the log-likelihood of their joint probability.

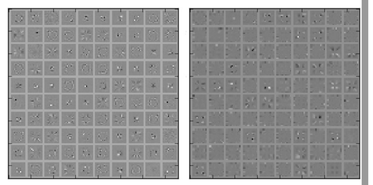
Baseline Algorithm with average intensities patchwise



Nuisance-Invariant Occlusion Detection



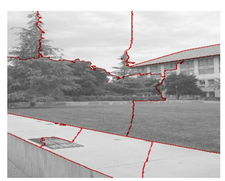
Learned features of shifted image pairs

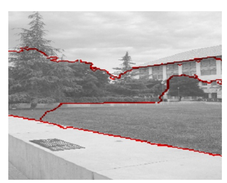


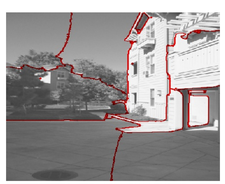
Learned features of scaled images pairs

Image Segmentation

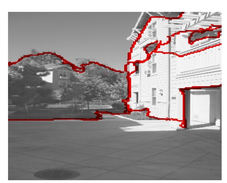
- The Gated RBM is trained with pairs of random images related with affine transformations, different illumination and scale. This way, the model can recognize patches coming from the same object.
- Comparisons of the neighboring patches in a single image reveal the "semantic" boundaries, that is separating curves between *different* objects in the scene. Not meaningful edges, such as the tiles and the texture within the clouds and the grass, are successfully disregarded by our algorithm.
- The Gated RBM's "similarity" map among all the patches that cover the image drives the Normalized-Cuts algorithm and combined yield a semantically meaningful segmentation.



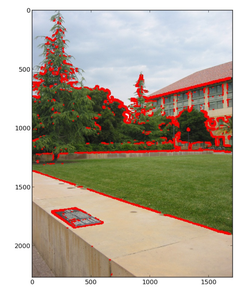




Normalized-Cuts segmentation



Gated RBM-driven Normalized-Cuts segmentation



Semantic and Nuisance-Invariant boundary detection

Conclusions

- Our architecture based on the Gated Restricted Boltzmann Machine can satisfactorily eliminate nuisance variability and support different Computer Vision tasks.
- In the occlusion detection setting, currently, we do not outperform algorithms which are specifically engineered for the task, such as the ones that use the optical flow. However, the results are decent. In later work we show that a systematically designed architecture based on the Gated RBM and superpixels provide a state-of-the-art method.
- The image segmentation algorithm provides *semantic* (edges that are not object boundaries are not detected) and *invariant* (learns the intrinsic variability over many nuisance factors) results.

Further Directions

- Another Computer Vision problem that we currently tackle is depth estimation of the 3D scene in a single image. The Gated RBM is trained twice with pairs of identical images with different scale and blur (zoom-in and zoom-out training) and then the joint likelihood of different pairs of patches is considered according to the two models.
- Depth relationships between many pairs of patches are extracted within the image and then all together can be incorporated in a global optimization problem, so that the depth map is estimated. Occlusion cues can assist on extracting depth relationships.
- In another pipeline the depth relationships between pairs of patches along with the values of the learned features feed the Ranked SVM algorithm, which in turn gives the depth map. The Ranked SVM is initially trained with weak texture and appearance features in different spatial zones of the training images.

Acknowledgements

I would like to thank Prof. Jason Cong and JRI for providing me the opportunity to participate in this exchange program; I also wish to thank Prof. Yizhou Wang (PKU) and Prof. Stefano Soatto (UCLA) for their constant support and guidance, and the graduate student Chen Zhou for our collaboration in the depth estimation project.

References:

[1] G. Sundaramoorthi, P. Petersen, V. S. Varadarajan, and S. Soatto. “On the set of images modulo viewpoint and contrast changes”, CVPR, 2009.

[2] R. Memisevic and G. E. Hinton. "Learning to Represent Spatial Transformations with Factored Higher Order Boltzmann Machines", Journal of Neural Computation, 2010.

[3] M. A. Ranzato, A. Krizhevsky, and G. E. Hinton. "Factored 3-Way Restricted Boltzmann Machines For Modeling Natural Images", Journal of Machine Learning Research, 2010.

[4] J. Susskind, G. E. Hinton, R. Memisevic, and M. Pollefeys. "Modeling the joint density of two images under a variety of transformations", CVPR, 2011.

Joint Research Institute in Science and Engineering by Peking University and UCLA

email: nikos.karianakis@gmail.com