Summaries for “Convolutional Learning of Spatio-temporal Features”

Weituo Hao 109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author introduces a model that learns latent representation of image sequences from pairs of successive images. Experiments on NORB dataset show their model can extract the transformation and low-level motion features in a multi-stage architecture for action recognition.

Main idea

First the writer introduces the gated Restricted Boltzmann Machine (GRBM) which differs from other conditional RBM architecture in that inputs change the effective models. By defining an energy function that captures third-order interactions among three types of binary stochastic variables, the energy function can be rewritten as follows:

$$E\left(\mathbf{y}, \mathbf{z}; \mathbf{x}\right) = -\sum_{ijk} W_{ijk} x_i y_j z_k - \sum_{k} b_k z_k - \sum_{j} c_j y_j$$

where W_{ijk} are the components of a parameter tensor, and W is learned. Then the energy of any joint configuration {**y**, **x**; **z**} can be converted to a conditional probability by normalizing:

$$p(z_k = 1|\mathbf{x}, \mathbf{y}) = \sigma(\sum_{ij} W_{ijk} x_i y_j + b_k)$$

Thus given an input-output pair of image pairs, {x, y}, it follows

$$p(z_k = 1|\mathbf{x}, \mathbf{y}) = \sigma(\sum_{ij} W_{ijk} x_i y_j + b_k)$$

where $\sigma(z) = 1/(1 + \exp{(-z)})$ is the logistic.

Based on GRBM, the paper further lists how to organize convolutional GRBM. Note that the author makes such two assumptions to give the energy function of convGRBM.

1) The input and output images are the same dimensions, $N_x = N_y$
2) The filter dimensions in the input and the output are the same, $N_w^x = N_w^y$
So the energy function can be written as:

$$E\left(\mathbf{y}, \mathbf{z}; \mathbf{x}\right) = -\sum_{k=1}^{K} \sum_{m,n=1}^{N_z} \sum_{r,s=1}^{N_w^y} z_{m,n}^k \gamma(\mathbf{x})_{r,s,m,n}^k y_{m+r-1,n+s-1} - \sum_{k=1}^{K} b_k \sum_{m,n=1}^{N_z} z_{m,n}^k - c \sum_{i,j=1}^{N_y} y_{i,j}$$

For the pooling layer, the author adopts the probabilistic max pooling method that means once a feature in a block is on, then the pooling unit must be on.

Experimental results

The writer first evaluates the third-order RBMs on Small NORB dataset. A convGRBM with real-valued outputs and 20 binary feature maps. The filter dimensions were set as $N_w^x = N_w^y = 9$. By visualizing the flow implicit in the hidden units in figure 1.the author claims that the transformation encoded by the feature maps can be much richer than what is expressed by optical flow alone.

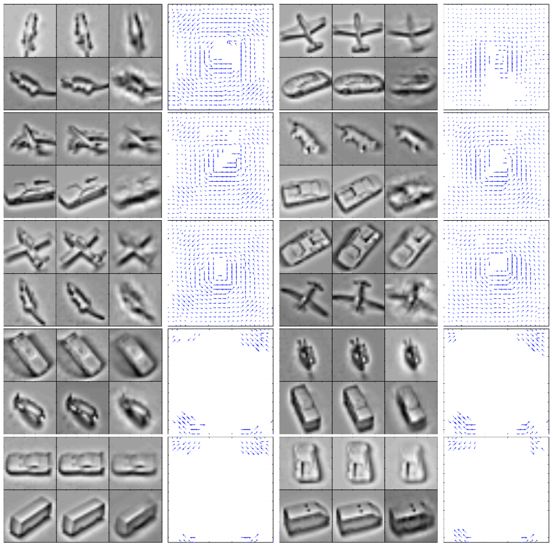
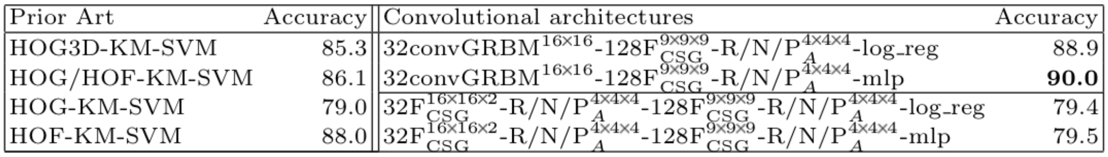


Figure 1 Image Analogies.

For experiments on KTH dataset, model based on a convGRBM with $N_z = 32$ feature maps and a pooling layer factor C=4 with filter size $N_w^x = N_w^y = 16$ is trained. The author claims to achieve the best result by giving such compare table



Prior Art	Accuracy	Convolutional architectures	Accuracy
HOG3D-KM-SVM	85.3	$32\text{convGRBM}^{16\times16}\text{-}128\text{F}_{\text{CSG}}^{9\times9\times9}\text{-R/N/P}_{A}^{4\times4\times4}\text{-log\_reg}$	88.9
HOG/HOF-KM-SVM	86.1	$32\text{convGRBM}^{16\times16}\text{-}128\text{F}_{\text{CSG}}^{9\times9\times9}\text{-R/N/P}_{A}^{4\times4\times4}\text{-mlp}$	**90.0**
HOG-KM-SVM	79.0	$32\text{F}_{\text{CSG}}^{16\times16\times2}\text{-R/N/P}_{A}^{4\times4\times4}\text{-}128\text{F}_{\text{CSG}}^{9\times9\times9}\text{-R/N/P}_{A}^{4\times4\times4}\text{-log\_reg}$	79.4
HOF-KM-SVM	88.0	$32\text{F}_{\text{CSG}}^{16\times16\times2}\text{-R/N/P}_{A}^{4\times4\times4}\text{-}128\text{F}_{\text{CSG}}^{9\times9\times9}\text{-R/N/P}_{A}^{4\times4\times4}\text{-mlp}$	79.5

Table 1 Comparison with previous work

For experiments on Hollywood2 dataset, model based on a convGRBM with $N_z = 32$ feature maps and a pooling layer factor C=4 with filter size $N_w^x = N_w^y = 16$ is trained. The author claims to achieve average precision of 46.6% that is a little below to the best published result 47.4% but better than most previous work

Conclusion

The author proposes convolutional gated RBM based on gated RBMs to learn represent optical flow and performed image analogies in a controlled context and achieve competitive performance.

Summaries for “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis”

Weituo Hao   109241801
Stony Brook University
Computer Science Department
Advanced Computer Vision

Contribution

The author introduces an extension of the Independent Subspace Analysis algorithm to learn invariant spatio-temporal features from unlabeled video data. Combing with deep learning techniques, the writer claims to achieve classification results superior to all previous published results on different datasets. Especially on the challenging dataset Hollywood2 and YouTube action datasets the author achieves 53.3% and 75.8 % respectively, which are nearly 5% better than the best published results.

Main idea

The first part is about the independent subspace analysis. Given an input pattern x^t, the activation of each second layer unit is $p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik}(\sum_{j=1}^{n} W_{kj}x_j^t)^2}$. ISA learns parameters W through finding sparse feature representations in the second layer by solving:

$$\begin{array}{ll} \underset{W}{\text{minimize}} & \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^t; W, V), \\ \text{subject to} & WW^T = \mathbf{I} \end{array}$$

where $\{x^t\}_{t=1}^{T}$ are weighted input examples. The following figure shows the two-layer architecture of ISA

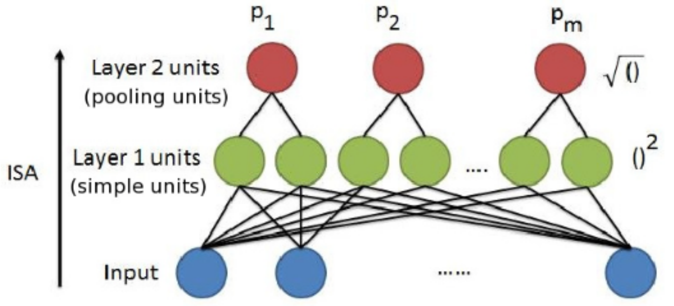


Figure 1 The neural network architecture of an ISA network

This ISA network has an invariant property that very suitable to recognition tasks. And that is the reason why ISA perform much better than other simpler methods such as ICA and sparse coding claimed by the author.

Based on the basic structure of ISA, the author presents the key idea of this paper: First train the ISA algorithm on small input patches and then take the learned network and convolve with a larger region of the input image. The combined responses of the convolution step are then given as input to the next layer that is also implemented by another ISA algorithm with PCA as a prepossessing step. Similar to the first layer, we use PCA to whiten the data and reduce their dimensions such that the next layer of the ISA algorithm only works with low dimensional inputs. Figure 2 shows this idea.

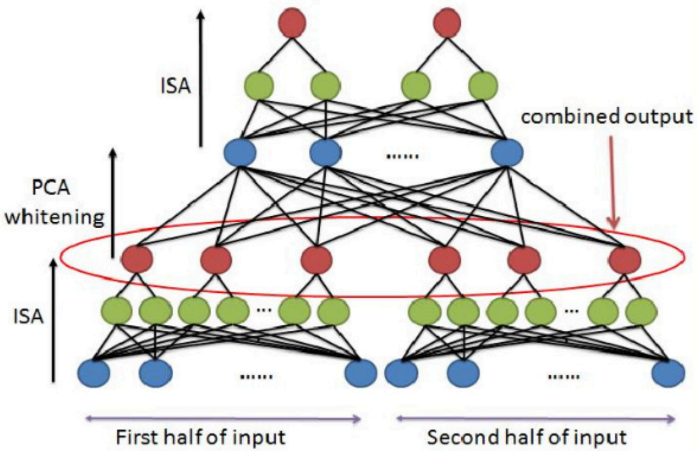


Figure 2 Stacked Convolutional ISA network

The input is a sequence of image patches and will be flattened into a vector. Here the author states that since batch gradient descent does not need any tweaking with learning rate and the convergence criterion, it is very easy to train such neural network.

Then the author gives a visualization of the first and second layer. By varying the velocity and plotting the response of the features with respect to the changes, the author concludes that the neurons are highly sensitive to the change of velocity that is a good property used to detect actions in movies. The second layer indicates complex shapes and invariances suitable for detecting high-level structures.

Experimental results

To verify the proposed method, four datasets KTH, UCF sports actions, Hollywood2 and YouTube action are used. All experiments are based on such model. The inputs to the first layer are of size 16*16 and 10, to second layer are of size 20*20 and 14. And total 200,000 video blocks sampled from the training set of each dataset.

For KTH dataset, the accuracy results of method with dense sampling and norm-thresholding are 91.4% and 93.9%, respectively. For Hollywood2 dataset, the mean average precision is 53.3%. For UCF sports actions dataset, the accuracy is 86.5%. For YouTube action dataset, the accuracy is 75.8%. Note that by removing the second layer, each result will drop significantly, so the author claims the usefulness of the second layer. In addition, the author also mentions their method can be used to learn features for classification on widely-available unlabeled video data.

Conclusion

This paper presents us a method by scaling up the independent subspace analysis. Experiments on the four challenging datasets show the method outperforms many state-of-art methods.