Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

Classification Factored Gated Restricted Boltzmann Machine

Ivan Sorokin
ITMO University, Department of Secure Information Technology,
9 Lomonosova str., St. Petersburg, 191002, Russia
i.sorokin@cit.ifmo.ru

Abstract. Factored gated restricted Boltzmann machine is a generative model, which capable to extract the transformation from an image pair. We extend this model by adding discriminative component, which allows directly use this model as a classifier, instead of using the hidden unit responses as features for another learning algorithm. To evaluate the capabilities of this model, we have created a synthetically transformed image pairs and demonstrated that the model is able to determine the velocity of object presented on two consecutive images.

Keywords: Multiplicative interaction, temporal coherence, translational motion, gated Boltzmann machine, supervision learning

1 Introduction

The gated Boltzmann machine is one of the models that uses *multiplicative interactions* [8] for learning the representation, which can be useful to extract the transformation between pairs of *temporally coherent* video frames [12]. Factorized version of this model is presented in [9], where authors train the model on shifts of random dot images and demonstrate that the model is able to identify the different directions correctly. We continue this research by studying the possibility to predict not only directions, but also a shift value. From all types of motion, we chose only translational motion, because it gives a great opportunity to use this model in many vision tasks, such as object tracking or visual odometry [4]. Therefore, the main objective of this work is to create a model that is trained to identify velocity vector in the image coordinate.

Instead of using additional model on top of the *mapping units*, we are adding discriminative component directly to the model. This technique was first applied for restricted Boltzmann machine [6] and since that has become widely used for similar models [11, 10]. In this paper, we are focused on the model that extracts transformation from two consecutive images. Without considering the additional discriminative component, there are several approaches of three-way structure model training [9, 13]. We propose a simple learning algorithm and show that it is not inferior to the existing. Moreover, our learning algorithm takes into account additional label variables and we demonstrate how it effects the training discriminative features. We refer to our model variants as *classification factored gated restricted Boltzmann machine* (cfgRBM).

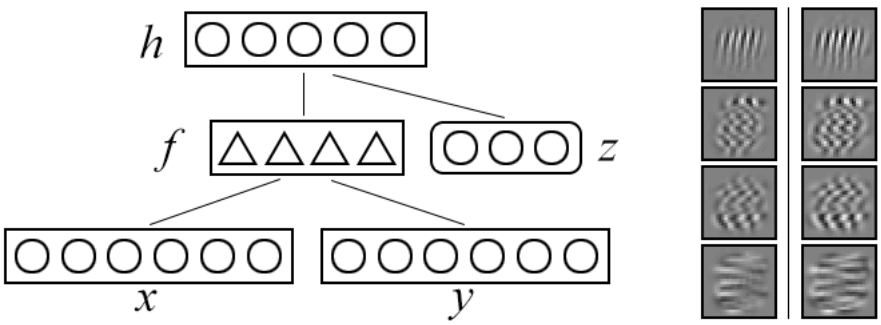

Fig. 1. Left shows the schematic representation of the cfgRBM model. Factorized form of the multiplicative interactions between two visible x, y and hidden h vectors depicted by triangles. The discriminative component is presented as one-hot encoded label vector z. Right shows specially chosen four filter pairs learned on horizontally shifted handwritten digits.

2 The Model

We propose a model (Fig. 1) in which the hidden units h not only captures the relationship between two images x and y, but also interacts with associated label z. The model is defined in the terms of its energy function and the function consists of two basic parts. The first of these is the factored three-way Boltzmann machine [13] and the second is classification restricted Boltzmann machine [5]. Combining these two models we defined expression for the energy function as follows:

$$E(x,y,z,h) = -\sum_f (\sum_i W^x_{if} x_i)(\sum_j W^y_{jf} y_j)(\sum_k W^h_{kf} h_k) - \sum_{kl} h_k V_{kl} z_l \\ - \sum_i a_i x_i - \sum_j b_j y_j - \sum_k c_k h_k - \sum_l d_l z_l \ , \tag{1}$$

where matrices W^x, W^y, W^h has size $I \times F, J \times F$ and $K \times F$ respectively, I and J are equal size of visible units, F - number of factors, K - number of hidden units. The discriminative component is weight matrix V with size $K \times L$ and one-hot encoded label vector z with L classes. Bias terms a, b, c and d associated with two visible, hidden and label vectors respectively. We will assume that the visible vectors are binary, but the model can be defined with real-valued units [1]. Every column $W^x_{\cdot f}$ and $W^y_{\cdot f}$ can be consider as filter pairs (Fig. 1).

To train the model, we also need to define the joint probability distribution over three vectors:

$$p(x,y,z) = \frac{\sum_h exp(-E(x,y,z,h))}{\sum_{x,y,z,h} exp(-E(x,y,z,h))} \ , \tag{2}$$

where the numerator is summing over all possible hidden vectors and denominator is partition function which cannot be efficiently approximated.

2.1 Inference

The inference task of proposed model is defined as the problem of classifying the motion between two related images. In order to choose the most probable label under this model, we must compute conditional distribution $p(z|x, y)$. We have adapted the calculations from the case of single input units [5] for the case of three-way interaction. As a result, for reasonable numbers of labels L, this conditional distribution can be also computed exactly and efficiently, by writing it as follows:

$$p(z_l = 1 \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{exp(d_l) \prod_k (1 + exp(o_{kl}(\boldsymbol{x}, \boldsymbol{y})))}{\sum_{l^*} exp(d_{l^*}) \prod_k (1 + exp(o_{kl^*}(\boldsymbol{x}, \boldsymbol{y})))} \ , \tag{3}$$

where

$$o_{kl}(\boldsymbol{x}, \boldsymbol{y}) = c_k + V_{kl} + \sum_f W^h_{kf} (W^{x\top}_{\cdot f} \boldsymbol{x})(W^{y\top}_{\cdot f} \boldsymbol{y}) \tag{4}$$

is an input to k hidden unit received from images x, y and estimated label l.

2.2 Learning

In order to train a cfgRBM to solve a classification problem, we need to learn the model parameters $\Theta = (W^x, W^y, W^h, V, a, b, c, d)$. Given a training set $\mathcal{D}_{train} = \{(x^\alpha, y^\alpha, z^\alpha)\}$ and a predefined joint distribution (2) between three variables, the model can be trained by minimizing the negative log-likelihood:

$$\mathcal{L}_{gen}(\mathcal{D}_{train}) = - \sum_{a=1}^{|\mathcal{D}_{train}|} \log p(x^\alpha, y^\alpha, z^\alpha) \ . \tag{5}$$

In order to minimize this function the gradient for any cfgRBM parameters $\theta \in \Theta$ can be written as follows:

$$-\mathrm{E}_{h|x^\alpha, y^\alpha, z^\alpha} \left[\frac{\partial E(x^\alpha, y^\alpha, z^\alpha, h)}{\partial \theta} \right] + \mathrm{E}_{x,y,z,h} \left[\frac{\partial E(x, y, z, h)}{\partial \theta} \right] \ , \tag{6}$$

where subscript of the expectation denotes the distribution for variables. There exists a learning rule [2], called "Contrastive Divergence", which can be used to approximate this gradient. Taking this rule into consideration we proposed the Algorithm 1 for the training of cfgRBM model. The main difference from the other approaches for training three-way interaction is in symmetrically sample vectors $\boldsymbol{x}, \boldsymbol{y}$ in the negative phase. Detailed information about the partial derivatives with respect to the model parameters can be obtained from [9, 5].

In case of factored three-way interactions the calculation of the gradient (6) involves numerical instabilities. Especially when using a large input vectors. To avoid this we also use a norm constraint on columns of matrics W^x and W^y. It is a common approach to stabilizing learning. For example, the same recommendations are given by [3] for method "Adaptive Subspace Self-Organizing Map" to learn invariant properties of moving input patterns.

Algorithm 1 Symmetric training update of the cfgRBM model

Require: training triplet $(x^\alpha, y^\alpha, z^\alpha)$ and learning rate λ

Notation
$a \leftarrow b$ means a is set to value b
$a \sim p$ means a is sampled from p

Positive phase
$x^0 \leftarrow x^\alpha$, $y^0 \leftarrow y^\alpha$, $z^0 \leftarrow z^\alpha$
$h_k^0 \leftarrow sigm(o_{kl^0}(x^0, y^0))$

Sample
$\hat{h} \sim p(h|x^0, y^0, z^0)$

Negative phase
$\boldsymbol{x^1} \sim p(x|y^0, \hat{h})$, $\boldsymbol{y^1} \sim p(y|x^0, \hat{h})$, $\boldsymbol{z^1} \sim p(z|\hat{h})$
$h_k^1 \leftarrow sigm(o_{kl^1}(\boldsymbol{x^1}, \boldsymbol{y^1}))$

# Update
for $\theta \in \Theta$ **do**
$\theta \leftarrow \theta - \lambda \left(\frac{\partial E(x^0, y^0, z^0, h^0)}{\partial \theta} - \frac{\partial E(x^1, y^1, z^1, h^1)}{\partial \theta} \right)$
end for

3 Experiments

The main goal of this research is to build a model that is capable to extract translational motion from two related images. Therefore, we created a synthetic data consisting of image pairs in which the second image is horizontally and relatively shifted towards the first. We take MNIST dataset[1] and randomly choose a shift value in the range [-3 3] for each image. As a result we get 7 possible labels for $60,000$ training and $10,000$ test image pairs of relatively shifted handwritten digits. All the models in the following experiments have 200 factors and 100 hidden units. For detailed information about learning parameters we refer to our implementation[2] of the models.

In the first experiment (Fig. 2), we compare different learning strategies for the cfgRBM model. The first learning method is taken from [9], where authors described a conditional model. The second method is proposed in [13], where authors define the joint distribution for an image pair. The results show that Algorithm 1 in the end of learning has the lowest classification and reconstruction test error. It is also interesting to note that there are different delays before filters become specialized in their frequency and phase-shift characteristics.

In the second experiment (Fig. 3), we compare hidden units activities of models with and without a discriminative component. In the first case we trained a model completely unsupervised without any labeled information. In the second case cfgRBM model was trained using Algorithm 1. The results show that

[1] `http://yann.lecun.com/exdb/mnist/`
[2] `https://cit.ifmo.ru/~sorokin/cfgRBM/`

descriminative component has a strong effect on hidden features. In addition, we also demonstrate an effect on the hidden units in the case with wrong label information.

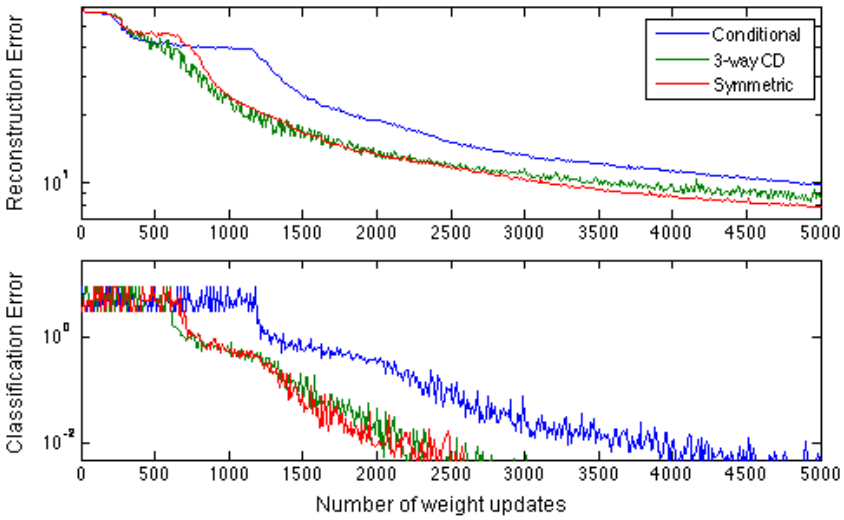

Fig. 2. Three learning stratagies for cfgRBM model. The reconstruction was calculated only for y visible units. In both graphs the error value obtained on test set and the ordinate is scaled logarithmically. Best view in color.

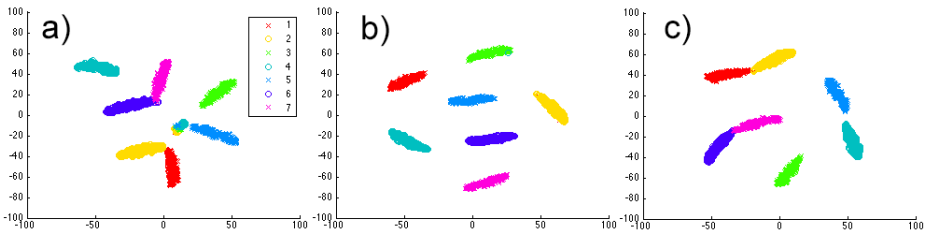

Fig. 3. Hidden units activations. For every test sample, activation of 100 hidden units projected to 2D coordinates using t-SNE [7]. a) model trained without discriminative component. b) model extendet with additional labeled units. c) exactly the same model as in the case (b), but labels of classes {-3,-2} and {2,3} are deliberetly combined.

4 Conclusion

In this paper, we incorporate supervision learning for factored gated restricted Boltzmann machine model. Our results show that proposed model is capable to identify the velocity of the object presented on two consecutive images. In the future work we plan to apply this model for videos which may be represented as a temporally ordered sequence of images. Particularly, the ability to extract translational motion will be useful for tracking tasks.

References

1. Fischer, A., Igel, C.: Training restricted Boltzmann machines: an introduction. Pattern Recognition 47(1), pp. 25–39 (2014)
2. Hinton, G.: Training products of experts by minimizing contrastive divergence. Neural computation 14, pp. 1771-1800 (2002)
3. Kohonen, T.: The adaptive-subspace som (assom) and its use for the implementation of invariant feature detection. In: Proc. ICANN95, Int. Conf. on Artificial Neural Networks, pp. 3-10 (1995)
4. Konda, K., Memisevic, R.: Learning visual odometry with a convolutional network. International Conference on Computer Vision Theory and Applications. (2015)
5. Larochelle, H., Mandel, M., Pascanu, R., Bengio, Y.: Learning algorithms for the classification restricted boltzmann machine. Journal of Machine Learning Research 13, pp. 643-669 (2012)
6. Larochelle, H., Bengio, Y.: Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th international conference on Machine learning, pp. 536–543 (2008)
7. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9, pp. 2579–2605 (2008)
8. Memisevic, R.: Learning to relate images. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, pp. 1829-1846 (2013)
9. Memisevic, R., Hinton, G.E.: Learning to represent spatial transformations with factored higher-order boltzmann machines. Neural Computation 22, pp. 1473-1492 (2010)
10. Reed, S., Sohn, K., Zhang, Y., Lee, H.: Learning to disentangle factors of variation with manifold interaction. In: Proceedings of the 31st International Conference on Machine Learning, pp. 1431-1439 (2014)
11. Sohn, K., Zhou, G., Lee, C., Lee, H.: Learning and selecting features jointly with point-wise gated boltzmann machines. In: Proceedings of The 30th International Conference on Machine Learning, pp. 217-225 (2013)
12. Srivastava, N.: Unsupervised Learning of Visual Representations using Videos. Department of Computer Science, University of Toronto. Technical Report. (2015) Retrived from `http://www.cs.toronto.edu/~nitish/depth_oral.pdf`
13. Susskind, J., Memisevic, R., Hinton, G., Pollefeys, M.: Modeling the joint density of two images under a variety of transformations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2793-2800 (2011)