The AAAI 2017 Spring Symposium on
Science of Intelligence: Computational Principles of Natural and Artificial Intelligence
Technical Report SS-17-07

Lifelong Learning of Action Representations with Deep Neural Self-Organization

German I. Parisi, Stefan Wermter
Knowledge Technology Institute, Department of Informatics
University of Hamburg, Germany

Abstract

Lifelong learning is fundamental in autonomous robotics for the incremental acquisition of knowledge through experience. However, most of the current deep neural models for action recognition from videos do not account for lifelong learning, but rather learn a batch of training actions. Consequently, there is the need to design learning systems with the ability to incrementally process available perceptual cues and to adapt their behavioral responses over time. We propose a self-organizing neural network architecture for incrementally learning action sequences from videos. The architecture comprises growing self-organizing networks equipped with recurrent connectivity for dealing with time-varying patterns. We use a set of hierarchically-arranged recurrent networks for the unsupervised learning of action representations with increasingly large spatiotemporal receptive fields. The recurrent dynamics modulating neural growth drive the adaptation of the networks to the non-stationary input distribution during the learning phase. We show how our model accounts for an action classification task with a benchmark dataset also in the case of occasionally missing or incorrect sample labels.

Introduction

Computational models inspired by the hierarchical organization of the visual cortex have become increasingly popular for action recognition from videos, with deep neural network architectures producing state-of-the-art results on a set of benchmark datasets (e.g. (Baccouche et al. 2011; Jain et al. 2015; Jung, Hwang, and Tani 2015)). Typically, visual models using deep learning comprise a set of convolution and pooling layers trained in a hierarchical fashion for yielding action feature representations with an increasing degree of abstraction (Guo et al. 2016). This processing scheme is in agreement with neurophysiological studies supporting the presence of functional hierarchies with increasingly large spatial and temporal receptive fields along cortical pathways (Taylor et al. 2015; Hasson et al. 2008).

The training of deep learning models for action sequences has been proven to be computationally expensive and to require an adequately large number of training samples for the successful learning of spatiotemporal filters. The supervised training procedure comprises two stages: (i) a forward stage in which the input is represented by the current network parameters and the prediction error is used to compute the loss cost from ground-truth sample labels, and (ii) a backward stage which computes the gradients with respect to the parameters and updates them using back-propagation through time. Although different regularization methods have been proposed to boost performance such as parameter sharing and dropout, the training process requires samples to be (correctly) labeled in terms of input-output pairs. Consequently, the question arises whether traditional deep learning models for action recognition can account for real-world learning scenarios, in which the number of training samples may not be sufficiently high and ground-truth labels may be occasionally unavailable or incorrect.

The above-described approaches, as well as other similar hierarchical models, have been designed for learning a batch of training actions, thus implicitly assuming that a training set is available. Ideally, this training set contains all necessary knowledge that can be readily used to predict novel samples in a given domain. However, this training scheme is not suitable for more natural scenarios where an artificial agent should incrementally process a set of perceptual cues as these become available over time. Lifelong learning is considered to be essential for cognitive development and plays a key role in autonomous robotics for the progressive acquisition of knowledge through experience and the development of meaningful internal representations during training sessions (Zhou 1990; Lee 2012).

In this work, we propose a deep neural architecture for incrementally learning action representations from videos. We introduce a recurrent extension of growing self-organizing networks for learning spatiotemporal properties of the input. We implement a hierarchy of recurrent networks that learn unsupervised representations of input sequences with increasing feature complexity. The recurrent dynamics modulating neural growth drive the adaptation of the networks to the non-stationary input distribution during the learning phase. For the purpose of classification, associative connections between visual action representations and action class labels are learned during the training phase. We report on experiments for an action classification task with the Weizmann action benchmark dataset, evaluating our approach also in the case of occasionally absent or incorrect sample labels during training sessions.

Proposed Method

In line with previous work on the self-organizing neural integration of action representations (Parisi, Weber, and Wermter 2015), the proposed deep architecture is composed of two distinct processing streams for pose and motion features, and their subsequent integration in the superior temporal sulcus (STS).

The two-pathway architecture is illustrated in Fig. 1. Each layer in the hierarchy comprises a recurrent Growing When Required (GWR) network (Marsland, Shapiro, and Nehmzow 2002) and a pooling mechanism for learning action features with increasingly large spatiotemporal receptive fields. In the last layer, we extend the proposed recurrent GWR to learn associative connections between visual representations and symbolic labels for the purpose of action classification.

Recurrent Neural Self-Organization

The traditional GWR network (Marsland, Shapiro, and Nehmzow 2002) creates new neurons whenever the activity of trained neurons is smaller than a given threshold. As a criterion for neural growth, the training algorithm considers the amount of network activation at time t computed as a function of the distance between the current input $\mathbf{x}(t)$ and its best-matching neuron $\mathbf{w}_b$:

$$a(t) = \exp(-\|\mathbf{x}(t) - \mathbf{w}_b\|). \quad (1)$$

Additionally, the algorithm considers the number of times that neurons have fired so that recently created neurons are properly trained before creating new ones. The network implements a firing counter $\eta \in [0, 1]$ used to express how frequently a neuron has fired based on a simplified model of how the efficacy of an habituating synapse reduces over time (Stanley 1976).

We introduce a temporal GWR network that equips each neuron with an arbitrary number of context descriptor (Strickert and Hammer 2005) to increase the memory depth and temporal resolution in the spirit of a Gamma memory model (de Vries and Príncipe 1992). A similar approach has been previously applied to self-organizing neural networks, showing good results in nonlinear time series analysis (Estévez and Hernández 2011; Estévez and Vergara 2012). For dealing with temporal processing, the activation of the network at time t is as follows:

$$d_i(t) = \alpha_w \cdot \|\mathbf{x}(t) - \mathbf{w}_i\|^2 + \sum_{k=1}^{K} \alpha_k \cdot \|\mathbf{C}(t) - \mathbf{c}_k^i\|^2, \quad (2)$$

$$\mathbf{C}(t) = \beta \cdot \mathbf{w}_{I(t-1)} + (1 - \beta) \cdot \mathbf{c}_{I(t-1)}, \quad (3)$$

for each $k = 1, ..., K$, where $\alpha, \beta \in (0; 1)$ are constant values that modulate the influence of the current input and the past activations, $\mathbf{c}_k^i$ is the kth context descriptor of neuron i, and $I(t-1)$ is the index of the best-matching neuron from the previous timestep. Since the definition of the context descriptors is recursive, setting $\alpha_w > \alpha_1 > ... > \alpha_K > 0$ reduces the propagation of errors from early filter stages to higher-order contexts for other competitive networks (Estévez and Hernández 2011; Estévez and Vergara 2012).

The training is carried out by adapting the weight and the context vector of the best-matching neurons and its neighbors towards the current input according to:

$$\Delta\mathbf{w}_i = \epsilon_i \cdot \eta_i \cdot (\mathbf{x}(t) - \mathbf{w}_i), \quad (4)$$

$$\Delta\mathbf{c}_k^i = \epsilon_i \cdot \eta_i \cdot (\mathbf{C}_k(t) - \mathbf{c}_k^i), \quad (5)$$

where ϵ_i is the learning rate that modulates neural update.

Lifelong Learning

GWR networks are inherently incremental and adapt their topology to the distribution of the input. The two parameters modulating the growth rate of the networks are the activation threshold and the firing rate threshold, with the former having stronger influence. Along the hierarchical flow, the recurrent dynamics modulating neural growth drive the adaptation of the networks to the non-stationary input distribution during the learning phase. Therefore, lifelong learning of action representations in the self-organizing hierarchy can be achieved by yielding prediction-driven neural dynamics, so that each higher-level network will learn to predict input sequences from lower-level networks. At each iteration, a new neuron will be added when

$$a(t) = \exp(-d_i(t)) < a_T, \quad (6)$$

where a_T is the activation threshold and $d_i(t)$ is given by Eq. 2. The activation threshold a_T sets the maximum discrepancy (distance) between the input sequence and its best-matching neuron in the higher-level network.

A mechanism used to control the durability of information in a network is the connection age. When a neuron is fired, the age of the connections from the neuron to its neighbours is set to 0, while the age of the rest of the connections is increased by a value of 1. At each iteration, old connections of neurons that have not fired for a while and neurons without connections are deleted. Removing a neuron from the network means that the information learned by that unit is permanently forgotten. Therefore, a convenient maximum age of connections μ_{max} must be set so that the network removes neurons that are no longer fired while avoiding *catastrophic forgetting*, i.e. forgetting previously learned representations during the process of learning new ones.

Pooling Layers

Typically, computational models with deep architectures obtain invariant responses by alternating layers of feature detectors and nonlinear pooling neurons using the maximum (MAX) operation, which has been shown to achieve higher feature specificity and more robust invariance with respect to linear summation. Although robust invariance to translation has been obtained via MAX and average pooling, the MAX operator has shown faster convergence and improved generalization (Scherer, Müller, and Behnke 2010). In our architecture, we implemented MAX pooling over the multi-dimensional neuron weights.

Action Classification

The aim of classification is to predict action labels from unseen action samples. For this purpose, the last network of

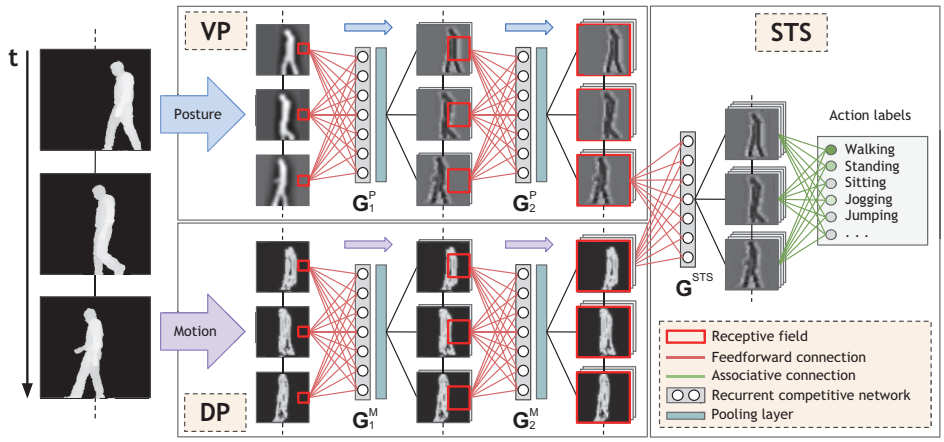

Figure 1: Diagram of our deep neural architecture with recurrent GWR networks for action recognition. Posture and motion action cues are processed separately in the ventral (VP) and the dorsal pathway (DP) respectively. At the STS stage, the recurrent network learns associative connections between prototype action representations and symbolic labels.

the hierarchy is equipped with an associative learning mechanism to map sample labels to prototype neurons representing action segments. During the training phase, neurons in the G^{STS} network can be assigned a label l (with l from a set L of label classes) or remain unlabeled.

This labeling mechanism yields neurons associated to most frequent labels, thus also handling situations in which sample labels may be occasionally missing or incorrect. To predict the label λ of a novel sample $\tilde{\mathbf{x}}_t$ after the training is completed, we return the label class with the highest value of the associative matrix for the best-matching neuron b of $\tilde{\mathbf{x}}_t$ according to Eq. 2.

Experiments and Results

For evaluating our architecture, we used the Weizmann dataset (Gorelick et al. 2005) containing 90 low-resolution (180×144) sequences with 10 actions performed by 9 subjects. The actions are *walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack*, and *skip*. Sequences are sampled at 180×144 with a static background and are about 3 seconds long. For our experiments, we used aligned foreground body shapes by background subtraction included in the dataset. To be consistent with other evaluation schemes in the literature, we evaluated our approach by performing *leave-one-out* cross-validation, i.e., 8 subjects were used for training and the remaining one for testing. This procedure was repeated for all 9 permutations and the results were averaged.

Training Parameters

We used sequences of actions at 10 frames per second with resized images of 30×30 pixels containing only the segmented body silhouette. In the first layer, we select image patches of 3×3 pixels for both posture and motion sequences, i.e. a total of 100 patches for each input image. The patches from posture and motion images are fed into the recurrent networks G_1^{P} and G_1^{M} respectively, both comprising a recurrent GWR with 1 context descriptor ($K = 1$). In the second layer, the input is represented by the pooled activation from the first two networks, yielding a 10×10 representation for each processed frame. From this representation, we compute 2×2 patches that are fed into $\mathrm{G}_2^{\mathrm{P}}$ and $\mathrm{G}_2^{\mathrm{M}}$ with 3 context descriptors ($K = 3$) each. In the third layer, the pooled activation from the pose and the motion pathways is concatenated, producing 10×5 matrices for each frame that we use to train $\mathrm{G}^{\mathrm{STS}}$ with $K = 5$. If we consider the hierarchical processing of the architecture, the last network yields neurons that respond to the latest 10 frames, which corresponds to 1 second of video. We assign decreasing values to α_i according to the function $\mathbf{p}_\alpha = [(\alpha_i / \sum_i \alpha_i) : \alpha_i = (1/(K+1)) - \exp(-(i+2))]$.

Evaluation

Results for the recognition of 10-frame snippets are shown in Table 1. Our experiments yielded an overall accuracy of 98.6%, which is a very competitive result with respect to the state of the art of 99.64% reported by (Gorelick et al. 2005). Our results outperform the overall accuracy reported by (Jung, Hwang, and Tani 2015) with three different deep learning models: convolutional neural network (CNN, 92.9%), multiple spatiotemporal scales neural network (MSTNN, 95.3%), and 3D CNN (96.2%). However, a direct comparison of the above-described methods with ours is hindered by the fact that they differ in the type of input and number of frames per sequence used during the training and the test phase.

An additional experiment consisted of decreasing the percentage of available and correct labels (from 100% to 0%) from randomly chosen samples. The average accuracy with different percentages of omitted and incorrect labels over 10 runs is displayed in Fig. 2. Classification performance over 80% was obtained for at least 40% available labels, whereas in the case of incorrect labels, we obtained an overall performance under 40% for less than 50% correct labels.

Table 1: Results on the Weizmann dataset for 10-frame snippets. Results from (Jung, Hwang, and Tani 2015) with 3 different models: 1) CNN, 2) MSTNN, and 3) 3D-CNN.

	Accuracy (%)
(Gorelick et al. 2005)	99.64
(Schindler and Van Gool 2008)	99.6
Our approach	**98.6**
(Jung, Hwang, and Tani 2015)	92.9[1], 95.3[2], 96.2[3]

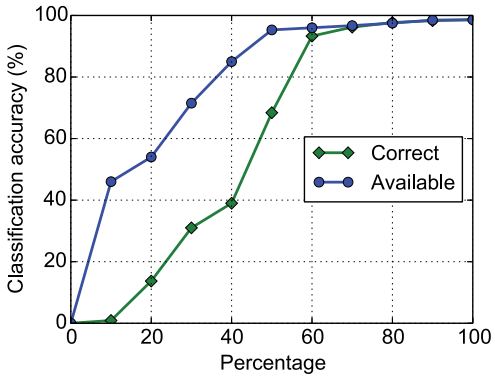

Figure 2: Average classification accuracy on the Weizmann dataset over 10 runs for a decreasing percentage of available and correct action labels.

Conclusion

We proposed a deep neural architecture with a hierarchy of recurrent GWR networks for incrementally learning action features with increasing complexity of representation. Visual representations obtained through unsupervised learning are incrementally associated to symbolic action labels for the purpose of classification. This is achieved through an associative mechanism in the recurrent GWR that attaches labels to prototype neurons based on their frequency. Lifelong learning was proposed in terms of prediction-driven neural dynamics in a self-organizing hierarchy. We have not considered additional important principles that play a role in lifelong learning such as the influence of reward-driven motivational and attentional functions (Ivanov et al. 2012), which will be subject of future research.

Experiments on the Weizmann action benchmark show competitive performance with the state of the art in different evaluation schemes. Additional experiments showed that our learning architecture can also handle situations in which the number of available and correct ground-truth labels is decreased during the training phase.

Acknowledgments

This research was partially supported by the DFG (Deutsche Forschungsgemeinschaft) for the project Crossmodal Learning TRR-169 / A5.

References

Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; and Baskurt, A. 2011. Sequential deep learning for human action recognition. In *Human Behavior Understanding (HBU)*, 29–39. Springer Berlin Heidelberg.

de Vries, B., and Príncipe, J. C. 1992. The gamma model–a new neural model for temporal processing. *Neural Networks* 5(4):565–576.

Estévez, P. A., and Hernández, R. 2011. Gamma-filter self-organizing neural networks for time series analysis. In *WSOM*, 151–159.

Estévez, P. A., and Vergara, J. R. 2012. Nonlinear time series analysis by using gamma growing neural gas. In *WSOM*, 205–214.

Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; and Basri, R. 2005. Actions as space-time shapes. In *ICCV*, 1395–1402.

Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; and Lew, M. S. 2016. Deep learning for visual understanding: A review. *Neurocomputing* 187:27 – 48.

Hasson, U.; Yang, E.; Vallines, I.; Heeger, D. J.; and Rubin, N. 2008. A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience* 28(10):2539–2550.

Ivanov, I.; Liu, X.; Clerkin, S.; Schulz, K.; Friston, K.; Newcorn, J. H.; and Fan, J. 2012. Effects of motivation on reward and attentional networks: an fMRI study. *Brain and Behavior* 2(6):741–753.

Jain, A.; Tompson, J.; LeCun, Y.; and Bregler, C. 2015. Modeep: A deep learning framework using motion features for human pose estimation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 302–315. Cham: Springer International Publishing.

Jung, M.; Hwang, J.; and Tani, J. 2015. Self-organization of spatio-temporal hierarchy via learning of dynamic visual image patterns on action sequences. *PLoS ONE* 10(7):e0131214.

Lee, J. 2012. Encyclopedia of the sciences of learning. 887–893. Boston, MA: Springer US.

Marsland, S.; Shapiro, J.; and Nehmzow, U. 2002. A self-organising network that grows when required. *Neural Networks* 15(8-9):1041–1058.

Parisi, G. I.; Weber, C.; and Wermter, S. 2015. Self-organizing neural integration of pose-motion features for human action recognition. *Frontiers in Neurorobotics* 9(3).

Scherer, D.; Müller, A.; and Behnke, S. 2010. Evaluation of pooling operations in convolutional architectures for object recognition. In *ICANN*, 92–101. Berlin, Heidelberg: Springer-Verlag.

Schindler, K., and Van Gool, L. J. 2008. Action snippets: How many frames does human action recognition require? In *CVPR*. IEEE Computer Society.

Stanley, J. C. 1976. Computer simulation of a model of habituation. *Nature* 261:146–148.

Strickert, M., and Hammer, B. 2005. Merge SOM for temporal data. *Neurocomputing* 64.

Taylor, P.; Hobbs, J. N.; Burroni, J.; and Siegelmann, H. T. 2015. The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports* 5(18112).

Zhou, H. H. 1990. Csm: A computational model of cumulative learning. *Machine Learning* 5(4):383–406.