ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Feature Binding in Deep Convolution Networks with Recurrences, Oscillations, and Top-Down Modulated Dynamics

Martin Mundt, Sebastian Blaes, Thomas Burwick

Frankfurt Institute for Advanced Studies (FIAS), Goethe University Frankfurt
Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany

Abstract. Deep convolution networks are extended with an oscillatory phase dynamics and recurrent couplings that are based on convolution and deconvolution. Moreover, top-down modulation is included that enforces the dynamical selection and grouping of features of the recognized object into assemblies based on temporal coherence. With respect to image processing, it is demonstrated how the combination of these mechanisms allow for the segmentation of the parts of the objects that are relevant for its classification.

1 Introduction

In a recent revival of multi-layered convolutional neural networks deep learning has successfully been shown to form the state-of-the-art in many tasks such as image-based object classification [1]. Comparison with the brain's information processing revealed that DL even competes with the representational performance of primate's IT cortex [2]. It may be expected that superior vision systems may be constructed by extending the deep architecture (DA) obtained from DL with properties that are still unique to the neurophysiological systems. Here, correspondingly, we begin to extend the DA through including two aspects that are essential for the biological workings: recurrences and oscillations. Moreover, we include a top-down modulated dynamics that is inspired by the construction of saliency maps [3] and demonstrate how the combination of these mechanisms allows for the segmentation of parts of an object in a visual scene that are relevant for the classification of this object. This realizes the selection and dynamical binding of features of an object into assemblies (see [4] for an introduction to the concepts of binding and temporal assembly formation) in the context of deep neural architectures.

2 Deep Networks with Recurrences and Oscillations

Basing our approach on a deep learning architecture which was used, for example, by Krishevsky et al. [1], we assume that the network has H hidden layers of convolutional type and two additional final layers that are fully connected. Each hidden layer h has F_h feature maps. The final output is given by a softmax procedure. The output at each layer $\ell = 1, 2, ..., L$ $(L = H + 2)$ is given by

$$y^{\ell}_{a_\ell} = \delta^{\ell}_{a_\ell} \cdot g_\ell \left(x^{\ell}_{a_\ell} \right) \quad \text{with} \quad x^{\ell}_{a_\ell} = \beta^{\ell}_{f_\ell} + \sum_{a_{\ell-1}} \{ W^{(\ell-1)\to\ell} \}^{a_\ell}_{a_{\ell-1}} y^{\ell-1}_{a_{\ell-1}} \ . \qquad (1)$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

where $\ell - 1 = 0$ refers to the input image with f_0 channels. Here, we have chosen a short form for the indices where

$$a_\ell \quad \leftrightarrow \quad (f_\ell, n_\ell, m_\ell) \ , \tag{2}$$

that is, the a_ℓ stand for the index $f_\ell = 1, ..., F_\ell$ of the feature maps ($F_{H+1} = F_{H+2} = 1$) and the indices n_ℓ, m_ℓ that describe the two-dimensional structure of the feature maps (and of the input images). The $\delta^\ell_{a_\ell} \in \{0, 1\}$ describe the results of max-pooling, a kind of local winner-take-all competition. In the following, to keep the equations simple, we assume that the set of indices a_ℓ is reduced to the indices of the units that are selected by the max-pooling procedure.

The β and W in eq. 1 describe biases and coupling weights, respectively. For $\ell = 1, .., H$ the weights are convolutional, i.e., in terms of the detailed indices the weights W may be related to filters K through weight-sharing:

$$\{W^{(\ell-1)\to\ell}\}^{f_\ell m_\ell n_\ell}_{f_{\ell-1} m_{\ell-1} n_{\ell-1}} = K^{f_\ell}_{f_{\ell-1}(m_\ell - m_{\ell-1})(n_\ell - n_{\ell-1})} \tag{3}$$

For $\ell = 1, .., L-1$ the signal functions g_ℓ are defined through $g_\ell(x) = x$ if $x \geq 0$ and $g_\ell(x) = 0$ if $x < 0$ ("ReLU"). The final layer uses a softmax output: $g_L(x) = \exp(x)/X$ where $X = \sum_{p_L} \exp(x^L_{p_L})$; see also the additional remarks in section 3. For the results presented in the next section, we use the implementation and pre-learned weights given by Chatfield et al. [8]; this reference may therefore be consulted for any further details.

As stated in the introduction, motivated by what constitutes neural processing in the brain, we now want to extend the above standard architecture by including recurrences, oscillations, and (in the next section) top-down modulation of the dynamics. The following discussion is intended to be a first step into this direction. Here, we restrict this step to oscillatory dynamics in the first layer. Correspondingly, the recurrence is between the first two hidden layers. (However, the complete network is involved when including the top-down modulations in the next section.) This is sufficient to present some of the fundamental aspects that may also be applicable to a more complete solution.

To describe the oscillatory dynamics in layer 1 and make the relation to earlier oscillatory models particularly obvious, let us choose different notations:

$$a_1 = n \ , \quad a_2 = p \ , \quad y^1_{a_1} = V_n \ , \quad \{W^{1\to 2}\} = \xi \ . \tag{4}$$

For the oscillatory dynamics, we introduce phases θ_n with the dynamics given by

$$\begin{aligned} \frac{d\theta_n}{dt} &= \omega_{0,n} + \omega_{1,n} V_n + \\ &\frac{\kappa}{N} \sum_m h_{nm} \left[\sin(\alpha) \cos(\theta_m - \theta_n) + \cos(\alpha) \sin(\theta_m - \theta_n) \right] V_m \ , \end{aligned} \tag{5}$$

where $\omega_{0,n}$ and $\omega_{1,n}$ are intrinsic and shear frequencies, respectively, N is the size of the filters K^{f_2}, $\kappa \geq 0$ and $0 \leq \alpha \leq \pi/4$ are the coupling constant and

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

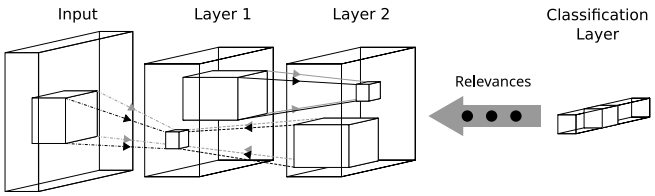

Fig. 1: Architecture of the discussed network type. Two kinds of recurrences are implemented. First, there is dynamical bottom-up (convolution given by eq. 7) and top-down (deconvolution given by eq. 8) processing between hidden layers $\ell = 1$ and $\ell = 2$. Second, there is a top-down modulation of the dynamics as described in section 3 (eq. 9) . The three dots indicate feed-forward (bottom-up) processing through layers $\ell = 3, ..., L-1$, while the arrow indicates the recurrent top-down modulation according to the "relevances" given by eq. 9.

the constant acceleration phase described in [5, 6, 7]. Of the two trigonometric terms, the latter implies a tendency to synchronize the oscillators and therefore describes a binding process, whereas the first term (present only for $\alpha \neq 0$) counteracts this synchronization to avoid a complete synchronization of the system in a manner that introduces a competition for coherence and selective binding [5, 6].

The crucial next step is to define the couplings h_{nm} appropriately. Studies of the learned filters suggest that these filters describe patterns [3, 9] and so it is a natural (though unusual) step to associate a pattern recognizing dynamics with these patterns through identifying

$$h_{nm} = \sum_p \lambda_p \eta_n^p \xi_n^p \ , \tag{6}$$

where λ^p is a quantity weighting these patterns and $\eta_n^p = 0$ if $\xi_n^p = 0$ and $\eta_n^p = \xi_n^p / |\xi_n^p|$ otherwise (the rationale behind using this normalization is given below). The dynamics may then be implemented as a convolution followed by a kind of "normalized deconvolution". With the notations of eqs. 4 the convolution with ξ $(= W^{1\to 2})$ takes the following form, describing responses in layer 2 in terms of quantities R_p, Θ_p that have their analog in the corresponding quantities defined in [5, 6]:

$$R_p \exp(i\Theta_p) = \frac{1}{N} \sum_n \xi_n^p V_n \exp(i\theta_n) \ . \tag{7}$$

It is then obvious that eq. 5 with eq. 6 describes a "normalized deconvolution" resulting in the phase dynamics given by

$$\begin{aligned} \frac{d\theta_n}{dt} &= \omega_{0,n} + \omega_{1,n} V_n + \\ &\kappa \sum_{p=1}^{P} \lambda^p \eta_n^p R^p \left[\sin(\alpha) \cos(\Theta^p - \theta_n) + \cos(\alpha) \sin(\Theta^p - \theta_n) \right] \end{aligned} \tag{8}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

(the notion of "normalized" refers to using η instead of ξ). The notations in this section have been used to make the connection between the deep architecture of equation eq. 1 and the results obtained in [5, 6] as explicit as possible. Incidentally, the reason for using η instead of ξ in eq. 6 is that it allows to apply the pure pattern frequency arguments from [5] also with respect to eq. 8. In the next section, we use the connection to this earlier work and discuss (and extend) the deep oscillatory dynamics described with eqs. 5 and 8.

Note at this point that the strategy of using static amplitudes but dynamical phases as a first step towards a more complete dynamics was also chosen by Finger and König [11].

3 Top-Down Modulated Dynamics and Segmentation

Due to the relation with the earlier work on oscillatory pattern recognition made explicit in the foregoing section, we may now apply several of the results obtained in [5, 6] (see also [7]) and expect that a competition for coherence together with binding based on temporal coherence arises with the phase dynamics given by eq. 8; see these references for introductions to the properties of oscillatory networks needed for the following discussion. Here, we want to demonstrate that this competition allows for the segmentation of object parts that are relevant for a classification.

Although we include oscillatory dynamics only at hidden layer $\ell = 1$, we need to use the full depth (that is, all layers) of the network for the following top-down modification of the dynamics. As each "pattern" p (where p now rather refers to features and positions; see section 2) is strengthened in the competition for coherence if its weight λ_p is increased, we allow for a top-down modification of the dynamics in eq. 8 by choosing the following values that are determined by top-down processing (in a backpropagation-like manner) from the final layer:

$$\lambda_p(I) = Z^{-1} \cdot \mathrm{abs}\left(\frac{\partial x^L_{a_L}}{\partial x^2_p}\right)_{\mathrm{inputimage}=I} \quad (\text{"relevances"}) \; . \tag{9}$$

(The final step in processing back to layer 1 is completed with eq. 8.) Note that the final two layers, $\ell = H + 1$ and $\ell = H + 2 = L$, are fully connected and do not reflect the two-dimensional geometry of the images [1, 8]. We may, however, keep the notation of section 2 by setting $n_L = m_L = 1$ and letting $f_L = 1, ..., M$ index the M different classification categories. The particular f_L used for the a_L in eq. 9 is the one that achieves the highest score $y^L_{a_L}$ when processing the inputimage I. Thus, the chosen f_L is the index of the obtained classification. The normalization is given by $Z = \max_p\{\mathrm{abs}(\partial x^L_{a_L}/\partial x^2_p)\}$.

Adapting an argument given by Simonyan et al. [3] for the construction of saliency maps it may be seen that the gradient term in eq. 9 serves to quantify the strength of the "patterns" p (that is, features f_2 present at positions n_2, m_2) in the competition for coherence according to their relevance for the obtained classification. In the next section, we demonstrate that this top-down modulation results in a competition for coherence that lets the phase dynamics

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

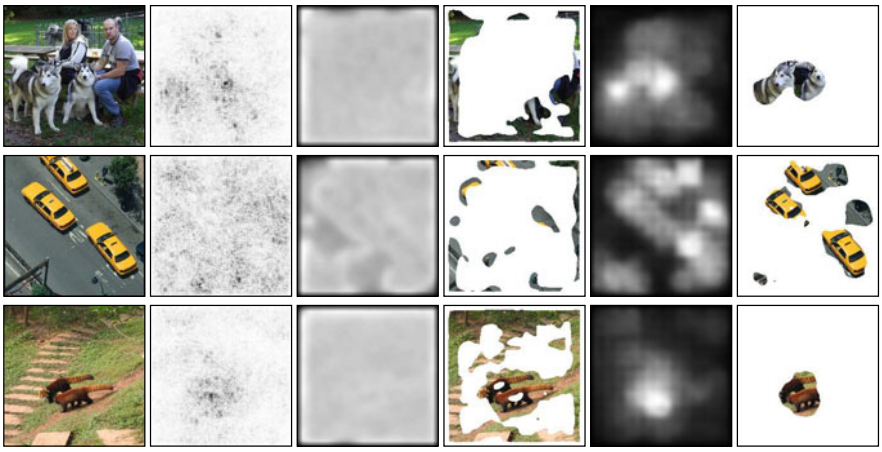

Fig. 2: (A) Input images and (B) class specific saliency maps [3] obtained for the classes "Alaskan malamute", "taxi" and "red panda". (C and D) No top-down modulation, $\lambda_p = 1$ for all p. (E and F) With top-down modulation, that is, λ_p given by eq. 9. (C to F) Notice that these panels are obtained by projecting the phase values in layer 1 back to the pixels of the input layer (including an appropriate unpooling procedure). (C and E) the phase maps as described in section 4. (D and F) The resulting segmentations. In case (F) the top-down modulation implies a binding of features into assemblies, corresponding to parts of the objects that are relevant for the classification.

imply a segmentation of (relevant parts of the) objects that are recognized by the network (that is, classified in the final layer).

4 Results

The following results (figure 2) were based on the pre-trained model given by Chadfield et al. [8] on the ILSVRC-12 data set (VGG-model); see this reference for details. We also confirmed the functioning of the described mechanisms with other networks and data sets: we trained a network as given by [1] (AlexNet), also on the ILSVRC-12 data set, and two networks that were inspired by the architecture of [1] on the CIFAR-10 dataset. All simulations used two Nvidia 970-GTX graphics processing units and Torch7 [10].

Given the three example images shown in figure 2A, the static amplitude values and phase dynamics is given by eqs. 1 and 5, respectively, where the latter is equivalent to a convolution followed by a deconvolution described with, corrrespondingly, eqs. 7 and 8. The parameters are chosen $\omega_{0,n} = \omega_{1,n} = 1$ for all n, $\kappa = 30$, $\alpha = \pi/4$, and $N = 5 \cdot 5 \cdot 64 = 1600$ is the size of the filters K^{f_2}. An Euler discretization is used with a time step $dt = 0.01$. For the simulation 900 time steps are computed. Panels C to F of figure 2 refer to the situation at

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

the end of the simulated time period.

As a result of the competition for coherence, a binding process like the one described in [6] occurs. We find that the winning "patterns" (understood as described in section 2) phase-lock to each other at higher phase-velocity compared to the non-winning units of the network. In that respect, we find the top-down modulation to be crucial; compare the panels C and D with the panels E and F in figure 2. We adopted a simple procedure to read out these assemblies. Starting with the initial phases of the θ_n randomly distributed between 0 and 2π, we determined their values at the end of the simulated time period (without applying the modulo 2π operation). The higher phase-velocity was indicated by larger values of these final phase values; these are displayed in figure 2C and 2E. The segmentation is then achieved through introducing a threshold of 60% of the maximal phase and keep only pixels that have values above this threshold.

In summary, over the course of the dynamic system's time evolution the described oscillatory dynamics gives rise to communication and propagation of information in a bottom-up and top-down fashion, resulting in lateral binding of features into assemblies based on temporal coherence. Going beyond the present discussion, future work may aim at a more complete machinery by extending the recurrent oscillatory dynamics to the complete set of layers.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems 25*, 1097, 2012.

[2] Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J., Deep neural networks rival the representation of primate IT cortex for core visual object recognition. In *PLoS Computational Biology*, 10(12):e1003963.

[3] K. Simonyan, A. Vedaldi and A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[4] T. Burwick, The binding problem, *Wiley Interdisciplinary Reviews: Cognitive Science*, 5.3: 305-315, 2014.

[5] T. Burwick, Temporal coding: assembly formation through constructive interference, *Neural Computation 20*, 7:1796, 2008.

[6] T. Burwick, Assemblies as phase-locked pattern sets that collectively win the competition for coherence, *Artificial Neural Networks-ICANN 2008*, Springer, 617–626, 2008.

[7] S. Blaes and T. Burwick, Attentional bias through oscillatory coherence between excitatory activity and inhibitory minima, *Neural Computation 27*, 7:1405, 2015.

[8] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, A., Attentional, Return of the devil in the details: delving deep into convolutional nets, *British Machine Vision Conference 2014 (BMVC)*, 2014.

[9] M.D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, *Computer Vision–ECCV 2014*, Springer International Publishing, 818-833, 2014.

[10] R. Collobert, K. Kavukcuoglu, and C. Farabet, Torch7: A matlablike environment for machine learning. In *BigLearn, Neural Information Processing Systems* (NIPS), no. EPFL-CONF-192376, 2011.

[11] H. Finger and P. König, Phase synchrony facilitates binding and segmentation of natural images in a coupled neural oscillator network, *Front. Comput. Neurosci.*, 7:195, 2014.