Selective attention improves learning

Antti Yli-Krekola, Jaakko Särelä and Harri Valpola

Department of Biomedical Engineering and Computational Science,
Aalto University, Helsinki, Finland
{antti.yli-krekola,jaakko.sarela,harri.valpola}@tkk.fi
http://www.becs.tkk.fi/en/

Abstract. We demonstrate that selective attention can improve learning. Considerably fewer samples are needed to learn a source separation problem when the inputs are pre-segmented by the proposed model. The model combines biased-competition model for attention with a habituation mechanism which allows the focus of attention to switch from one object to another. The criteria for segmenting objects are estimated from data and are shown to generalise to new objects.

Key words: Selective attention, perceptual learning, segmentation

1 Introduction

Learning task-relevant feature and object representations is a crucial problem for an autonomous agent trying to cope in a real-world environment. Sometimes the problem can be facilitated by collecting data from controlled environments, leading for instance to reduced noise and fewer objects present simultaneously. Such simplifications allow even fairly difficult problems to be solved with the current machine learning methods.

In many situations, however, these controlled environments cannot be provided due to cost, infeasibility of human intervention or other reasons. In those cases, the system should be able to learn feature and object representations autonomously. Furthermore, the learnt representations should be relevant for the tasks the agent faces. For these really difficult cases, machine learning research has provided us with painfully few methods.

The key problem is that the relevant associations and relations are complex and dynamic. As an example, let us consider the interplay between the visual and the motor system in picking up an object. There are many degrees of freedom in the task: the object can be in several places with respect to the hand and the head, the eyes can be viewing in several directions and the hand can be in several orientations, just to name a few. Yet, the autonomous agent should be able to learn the associations that are needed to perform the task of picking up the object. In any particular context of hand, eye and object positions, there exist many simple correlations between the needed motor output and the visual input. However, averaged over all the contexts, the correlations cancel each other out. Thus the agent needs a representational system that can learn and use dynamic

associations and relations that describe the short-lasting correlations between the different modalities.

The best example of a system that has been able to solve the above problems is the human brain. In neuroscience, it is known that attention plays a key role in perceptual learning [1]. The purpose of this paper is to discuss the information processing mechanisms of attention and to show that it can facilitate learning of feature and object representations.

2 Attention and learning

From psychophysical experiments it has become clear that attention plays a significant role in learning. For instance, Ahissar et al. [1] showed that attention guides low-level perceptual learning by focusing the representational capacity (low-level perceptual discriminations) to features that are relevant for the task at hand.

There is experimental evidence to support the idea that attention is realised by a competitive binding process that forms functional networks dynamically [2]. This dynamical binding has been shown to gate the coherence between cortical areas, thereby affecting the associations learnt between these areas [3].

Taken together, it seems plausible that selective attention and the formation of dynamical bindings are the necessary ingredients by which a large learning system can deliver training signals from distant areas, such as from motor cortex to visual cortex [4].

In order to use attention for perceptual learning in machine learning context, it is necessary to 1) implement a model which gives rise to attention, 2) learn the parameters of the model, making attention adaptive, and 3) use it successfully to facilitate perceptual learning. Although each of these three aspects have been studied independently and in pairs, to our knowledge the model presented in this article is the first to combine all three into a functional model.

2.1 Gestalt principles

When we humans see a new object, we may not know its identity but we can nevertheless tell what is part of the object and what is not. In other words, we are able to segment out an object without having seen it before.

In perceptual psychology, the rules of the organisation of perceptual scenes are called Gestalt principles [5]. Psychologists have identified several principles, such as proximity, common fate, similarity, continuity, closure, symmetry and convexity. The Gestalt principle of continuity is illustrated in Figure 1a, where the human visual system groups some of the line segments to form a circle.

What makes the Gestalt principles interesting in the current context is that they can be learnt from data. In neural terms, the Gestalt principles can be implemented by giving positive connections between certain neurons in one area and some other neurons in an adjacent area. Learning the connections can be based on simple correlations found in the data. For example, features responding

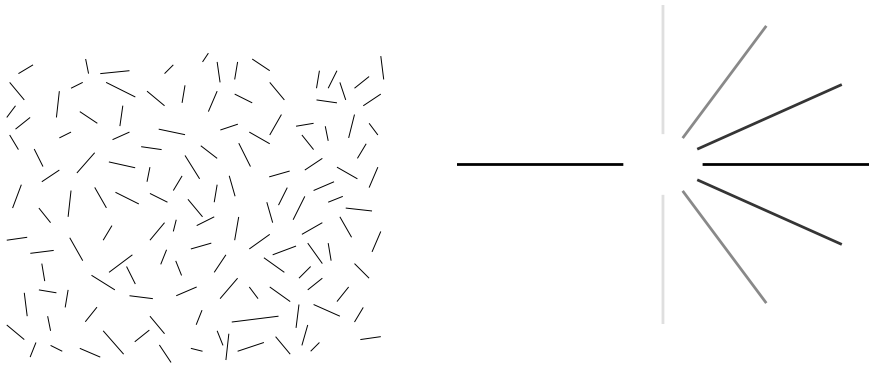

Fig. 1. a) Because of the Gestalt principles, a circle is perceived rather than some other grouping of the lines. b) Gestalt grouping of the neurons. The lines are features coded by different neurons. The shades of gray illustrate the connection strengths between the neurons on the right and the neuron on the left, darker meaning stronger. The lateral connections are stronger when the Gestalt principle is better fulfilled.

to lines of certain orientation in one part of the visual field are more probably co-activated with features of similar orientation in some other part of the visual field. This mechanism is illustrated in Fig. 1b.

These "neural" Gestalt rules can be learnt from the data and they operate on the level on individual feature-coding neurons. The principle is therefore applicable to any modality and also between modalities unlike, for example, many segmentation procedures that make use of the spatial structure of visual images. Moreover, the neural Gestalt rules can be learnt locally and in parallel. In the visual domain this means that the local correlations found in familiar objects generalise to new objects which have different overall shapes but nevertheless obey the same local correlations.

2.2 Biased-competition model for attention

Contextual (predominantly top-down) biasing of local lateral competition had been proposed as a model of covert attention in humans [6]. Usher and Nierbur [7] then suggested a computational model for biased competition that has been shown to replicate many attentional phenomena, for instance both bottom-up and top-down aspects of attention [8].

Deco and Rolls [8] also showed that it is possible to learn the weights for contextual biasing by the mechanism outlined in Sec. 2.1. In other words, the neural Gestalt rules can be applied in a relatively straight-forward manner to implement selective attention.

2.3 Suggested model

One shortcoming of the previously suggested biased-competition models is that they converge to a representation of one of the objects present in the inputs and then will not switch attention to other objects unless the input changes. This is in contrast with human covert attention which keeps switching between salient objects even when the stimulus does not change.

Models with changing attention usually have some kind of habituation mechanism which assures that attention will not get stuck with one object (e.g., [9]). Habituation means that active neurons gradually get "tired", thereby decreasing the stability of the currently active population of neurons. After the support for a population erodes, another population of recovered neurons takes over and the original tired population starts recovering.

Taken together the model has four key mechanisms:

1. Bottom-up input which mostly determines the activation level of the neuron,
2. Contextual (lateral or top-down) input which reflects learnt Gestalt principles,
3. Local competition which is biased by the contextual input and
4. Habituation which ensures that the winning population gradually gets tired and makes room for the winning population.

A more detailed description of the implementation is given in Sec. 3. However, it should be emphasised that the exact details of these mechanisms are not important although they of course need to fit together.

2.4 Relation to previous work

Several systems have been suggested that segment objects and represent them sequentially. Many of them are based on weakly coupled oscillators or other related mechanisms (e.g., [10–13]). Biased competition has the added benefit that it not only groups objects but can also select among them. This will be important when scaling up the system.

There are only a few examples of tackling the problem of using attention to improve learning. Selective attention was used for improving learning by Walther et al. [14] but their selective attention specialised in the visual domain and did not use learnable Gestalt rules which could be applied in any modality and even across modalities. Learning associations between different features has also shown to improve with attention by Kruschke [15], but his model has an external teacher controlling the attention.

3 Experiments

In this section, we use artificially generated data to demonstrate that it is possible and useful to combine attentional mechanisms and learning of feature representations in a single scheme. The Gestalt principles are first learnt on one data

set. The resulting lateral connections are then used for segmenting new objects. We show that this greatly improves learning at the next stage, for which we used FastICA [16]. The MATLAB scripts for performing the experiments can be downloaded at `http://www.lce.hut.fi/research/eas/compneuro/repository/attention_learning.zip`.

3.1 The data

We generated artificial data which had "objects" analogous to closed contours. For instance, the closed contour in Fig. 1a (circle) consists of 12 line segments which follow the local Gestalt rule of continuation. Our objects had five 100-dimensional patches (analogous to line segments) that were connected cyclically as shown in Fig. 2. Each object had one active element on each patch. In other words, an object was a 500-dimensional binary vector with five ones and 495 zeros.

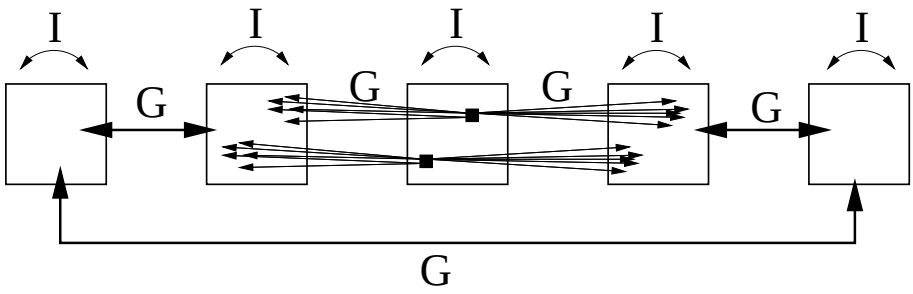

Fig. 2. The structure of the data is an idealisation of the Gestalt rules for closed contours (Fig. 1). Each of the five patches consists of 100 elements. According to the Gestalt rules (G), each element has five permissible neighbours in the adjacent patch. The model structure is similar, with local inhibitory connection (I) and excitatory lateral connections (G).

The objects were generated as follows. First the Gestalt rules (G in Fig. 2), which hold for all objects, were chosen randomly. Each element had five randomly selected permissible neighbours in both the adjacent patches. The five active elements of each object were selected in stages: 1) select one of 100 elements on the first patch, 2) select one of the five permissible elements (out of 100) on the second patch, 3) repeat for all the patches and finally 4) accept or reject the object depending on whether the element on the last patch is a permissible neighbour of the selected element on the first patch. On average there are 3,125 different objects that fulfil our continuity rules. The exact number depends on the Gestalt rules which were randomised.

We used these objects to generate noisy data which follow a linear independent component analysis (ICA [17]) model. Each 500-dimensional sample vector was a sum of five randomly selected objects and additive binary noise with 25

ones and 475 zeros. A noisy sample vector together with the five constituent objects are depicted in Fig. 3a.

3.2 Learning the Gestalt principles

We selected 20 objects which were reserved as "new objects" for the testing phase. We generated a data set with 10,000 samples using the remaining objects (3,105 objects on average). The lateral connections where then set to the values corresponding to the covariances between the input elements. Note that it would be difficult to learn reliably any correlations between 3,105 objects from such a small data set but it is perfectly feasible to estimate the correlations of the constituent elements. The estimated covariances are noisy but good enough for the next stage.

3.3 Biased-competition model with habituation

As explained in Sec. 2.3, the biased-competition model used for segmenting data has four mechanisms: 1) bottom-up inputs drive the activations, 2) contextual input, which biases 3) local competition, and 4) habituation. The structure of the model (Fig. 2) reflects the structure of the data: there are five areas (each with 100 neurons) laterally connected by the weights learnt with the procedure explained in the previous section. Local inhibition operates within each individual area and is denoted by I in the figure.

One of us has previously shown that biased competition is fully compatible with competitive learning which can learn meaningful features from bottom-up inputs [18]. Here we simplified the situation by assuming that the bottom-up inputs are already the input features to be represented. The neurons thus get bottom-up activations $\mathbf{x}$ which are simply the data samples.

Contextual lateral input from previous activations $\mathbf{y}(t-1)$ modulates the bottom-up activations as follows:

$$y_i^*(t) = \left[\left(g_i(t) + \alpha \mathbf{a}_i \mathbf{y}(t-1) \right) x_i \right]_+ \ , \tag{1}$$

where $\mathbf{a}_i$ is a row vector of lateral connections implementing the estimated Gestalt rules and $\alpha = 0.1$. The term $g_i(t)$ is a gain which implements the habituation and will be explained shortly. The activations $y_i^*(t)$ are restricted to be positive.

After this, lateral competition selects the final activations

$$y_i(t) = \left[y_i^*(t) - I_{\text{area}} \right]_+ \ , \tag{2}$$

where I_{area} is a function of $y_i^*(t)$ within an area. All the neurons in one area have the same I_{area}. This inhibitory term is adapted with a fast time-constant such that the target sparseness would be reached. We use the following sparseness measure for a local activation pattern $\mathbf{y}_{\text{area}}$:

$$s(\mathbf{y}_{\text{area}}) = \frac{1}{||\mathbf{y}_{\text{area}}||} \sum_{i \ \in \ area} y_i \ . \tag{3}$$

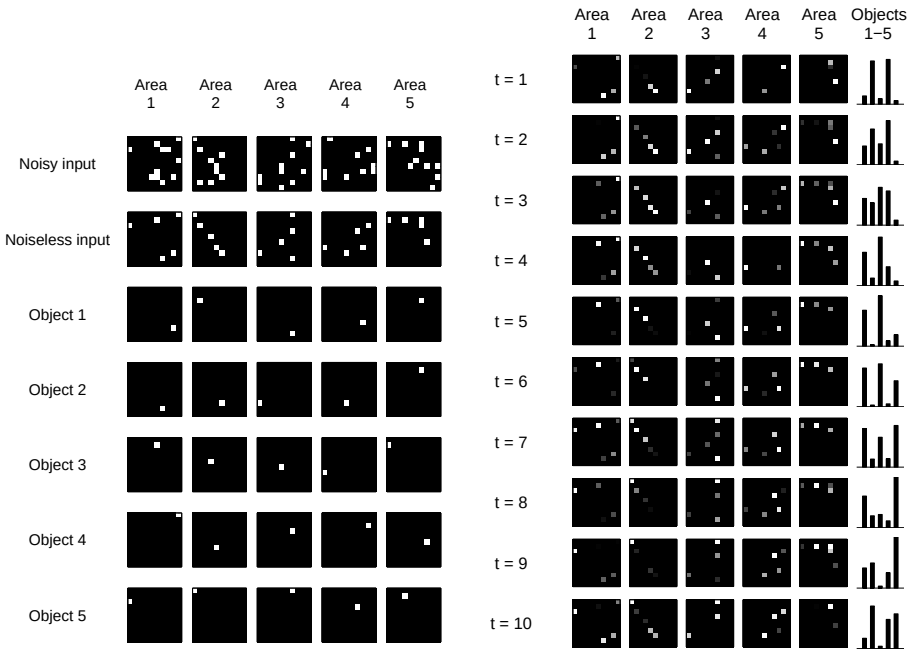

Fig. 3. a) A sample input, on the top row. The next row is the same input without noise. This noiseless input is used to produce the segmentation on the right (b). This input consists of the five shown objects. b) An example of segmentation of the noiseless input on the left (a). The 10 consequent time steps are taken after 50 steps after introducing the input. Two examples: In the first row, the 2nd and the 4th objects are seen. The 3rd object is growing from $t = 1$ until $t = 4$, and then starts to disappear.

On each time step, the local inhibition I_{area} is adjusted to make the pattern closer to the right sparseness level. We chose it to be the sparseness of the vector in which there are three ones and 97 zeros.

Habituation was implemented as follows. The gains are adjusted on each time step with a slower time-constant than the inhibition. The updates try to match the average activity with the original input: $E\{y_i\} \approx x_i$. On each time step, g_i is increased (decreased) a little if the moving average of y_i is below (above) x_i.

3.4 Results

An example of the segmentation dynamics is shown in Figure 3b. In the segmented representations, individual objects can be seen to appear and disappear more or less coherently. Note that for the sake of visual clarity, the segmentation dynamics is shown for the noiseless input from Figure 3a although all data used in the learning experiments contained noise.

The success of the segmentation was measured by separating new objects with the model. Recall that the lateral weights were estimated from data which lacked the 20 objects reserved for testing. These previously unseen objects were used for generating new samples, again with five objects added together with noise.

The biased-competition model with habituation segmented each original input sample into many new samples. First we let the network converge for 100 time steps and then we used the following 30 samples as inputs to FastICA. Each original sample was therefore expanded into 30 segmented samples.

We measured the accuracy of separation by a modified Amari index (for the original, see [19]):

$$a(C) = \frac{1}{N} \sum_{i}^{N} \left(\sum_{j}^{N} \frac{C_{ij}^2}{\max_k C_{kj}^2} - 1 \right) , \tag{4}$$

where C_{ij} corresponds to the ith separated signal using the jth object as the input. The Amari index is a standard way of measuring separation success.

ICA was done with different numbers of samples to both the original samples and the segmented samples generated by the proposed model. We used FastICA 2.5 package [16] with deflatory estimation and pow3 non-linearity, which in this case was more robust than the usually recommended tanh-nonlinearity. Because of local minima, different initialisations give different results. We used 30 different initialisations for each number of samples, and for each object, chose the component that gives the smallest Amari index.

The results are shown in Figure 4. ICA for the original non-segmented data needs about a hundred times as much samples as does ICA for the segmented data. For fine-tuning though, the non-segmented case seems to be better. The segmentation gives rough guesses about what the objects could be, but can also sometimes break them, and move the fixed points of the FastICA algorithm. The segmented case Amari index saturated to 1 milliAmaris at about 200 samples. The non-segmented case got better results with $N > 7000$.

4 Discussion

In this paper we demonstrated that selective attention can improve learning. We concentrated on showing that, with pre-segmentation, considerably fewer samples are needed to learn meaningful features. The segmentation was based on lateral connections whose strengths were estimated from another data set. The setting thus mimicked a situation where local Gestalt rules have already been learnt from past experience, allowing new objects to be segmented and thus greatly reducing the number of samples needed for learning about new objects. In this paper the learning task was chosen to be independent component analysis but reduced learning time should generalise to other types of associative learning as well due to reduced amount of distractors.

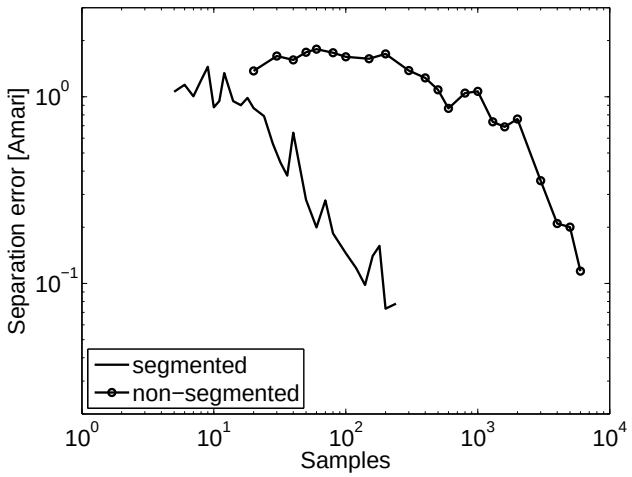

Fig. 4. Separation results with segmented and original data are shown. The separation quality is measured with a modified Amari index which measures the deviation of the unmixing matrix from optimal. Segmenting the data reduces the number of samples needed for reaching a given value of Amari index by roughly a factor of 100.

In the reported experiment, the segmentation principles were learnt offline. In actual use, it would be more useful to learn the object representations and the segmentation principles at the same time in a feedforward-feedback loop. This would, for instance, allow selective attention to guide the learning by discarding some structure in the data and focusing the representational capacity to relevant features. However, it will also be necessary to take into account the danger of run-away learning of self-induced correlations. Similar problems arise in learning any non-directed graphs, such as Markov fields [20]. A popular solution is to have two separate learning stages: one driven by the input and another, sleep-like, driven by expectations. The idea is to forget the unwanted representations during the sleep stage. When learning and forgetting balance each other, the learnt weights have captured the statistics of the input.

The model proposed in this article is based on biased-competition model which has been shown to be able to implement attention in large hierarchical networks. We have previously shown that the model is compatible with competitive learning and thus can learn meaningful bottom-up features under the guidance of selective attention [18]. In this paper, we added a mechanism for habituation which allows the focus of attention to change from one object to another and then showed that the resulting segmentation greatly improves associative learning. We believe that this work provides a fruitful starting point for future efforts in building a representational system flexible and powerful enough for an autonomous agent to survive in a complex real-world environment.

References

1. Ahissar, M., Hochstein, S.: Attentional control of early perceptual learning. Proceedings of the National Academy of Sciences **90** (1993) 5718–5722
2. Reynolds, J.H., Chelazzi, L.: Attentional modulation of visual processing. Annual Review of Neuroscience **27** (2004) 611–47
3. Miltner, W.H.R., Braun, C., Arnold, M., Witte, H., Taub, E.: Coherence of gamma-band EEG activity as a basis for associative learning. Nature **397**(6718) (1999) 434–436
4. Särelä, J., Valpola, H.: Denoising source separation: a novel approach to ICA and feature extraction using denoising and Hebbian learning. In: Correlation Learning Workshop in the Eighteenth Canadian Conference on Artificial Intelligence, Victoria, Canada (May 2005) 45–56
5. Todorovic, D.: Gestalt principles. Scholarpedia **3**(12) (2008) 5345
6. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Review of Neuroscience **18** (1995) 193–222
7. Usher, M., Niebur, E.: Modeling the temporal dynamics of it neurons in visual search: A mechanism for top-down selective attention. Journal of cognitive neuroscience **8** (1996) 311–327
8. Deco, G., Rolls, E.T.: A neurodynamical cortical model of visual attention and invariant object recognition. Vision research **44** (2004) 621–642
9. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40** (2000) 1489–1506
10. Wang, D.L., Terman, D.: Locally excitatory globally inhibitory oscillator networks. IEEE Trans. Neural Net. **6** (1995) 283–286
11. Choe, Y., Miikkulainen, R.: Self-organization and segmentation in a laterally connected orientation map of spiking neurons. Neurocomputing **21**(1-3) (1998) 139 – 158
12. Weng, S., Wersing, H., Steil, J., Ritter, H.: Learning lateral interactions for feature binding and sensory segmentation from prototypic basis functions. IEEE Transactions Neural Networks **17**(4) (2006) 843–862
13. Lessmann, M., Würtz, R.P.: Image segmentation by a network of cortical macrocolumns with learned connection weights. In: Proceedings of Biologically Inspired Cooperative Computing (BICC). Springer Verlag (2008)
14. Walther, D., Rutishauser, U., Koch, C., Perona, P.: Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding **100** (2005) 41–63
15. Kruschke, J.K.: Toward a unified model of attention in associative learning. Journal of Mathematical Psychology **45** (2001) 812–863
16. FastICA: The FastICA MATLAB package. (1998) Available at `http://www.cis.hut.fi/projects/ica/fastica/`.
17. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley (2001)
18. Yli-Krekola, A.: A bio-inspired computational model of covert attention and learning. Master's thesis, Helsinki University of Technology, Finland (2007)
19. Amari, S., Cichocki, A., Yang, H.: A new learning algorithm for blind source separation. In: Advances in Neural Information Processing 8 (Proc. NIPS'95). MIT Press, Cambridge, MA (1996) 757–763
20. Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for boltzmann machines. Cognitive Science **9**(1) (1985) 147 – 169