Scale- and Translation-Invariant Unsupervised Learning of Hidden Causes Using Spiking Neurons with Selective Visual Attention

Youssef Kashef

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master-Diplom
in Neural Systems and Computation
at
ETH Zürich

7. August 2013

Contents

Abstract

Nessler et al. have demonstrated the ability of a spiking neuronal network governed by spike-timing-dependent-plasticity (STDP) and a stochastic winner-take-all (WTA) circuit to learn and predict causes from visual input. We aim to increase the computational power of the existing network through invariance to translation and scale. The visual system of the brain masters the recognition of objects wherever they appear in the visual scene, regardless of scale, orientation or even with partial occlusions. It achieves this through attention. Therefore, we turn to the pool of literature on modeling visual attention systems inspired from the brain. The architecture of the extended model is composed of the existing recognition module receiving bottom-up input from an attention module. Pre-attentive computations allow the attention module to alter the input window exposed for recognition. Attention is modeled as a network measuring for saliency in a scene by feature extraction with the use of hierarchies. The design and development of this extended model to achieve the required invariance by using processes that approximate their biological counterparts is presented. Emphasis is put on making these approximations through computationally economic implementations. Evaluation of the model is based on its performance and convergence in a set of experiments as well as its computational efficiency. Experiments are constructed to scrutinize the behavior of the model, its ability to converge onto a sight within a scene that enables recognition. Artificial as well as natural images are used to further reveal the capabilities and limitations of our approach. A top-down feedback signal of the recognition module that modulates attention is discussed.

Chapter 1

Introduction

Spike-based Expectation Maximization (SEM) is a model of bayesian modules articulated by Nessler et. al of how the brain analyzes sensory stimuli. The model demostrates the learning of hidden causes in visual stimuli emerging through correlations in a stochasitc soft winner-take-all (WTA) network of spiking neurons. The spiking neurons are activated continuously in the presence of their preferred stimulus [9]. Spike-timing-dependent-plasticity (STDP) in WTA circuits defines the learning method for recognizing the hidden causes in the stimulus. The utilizaton of STDP acts as an approximation of traditional Expectation Maximization [8]. This model forms the basis of the presented work.

More emphasis will be put on how the WTA circuit in the SEM model is constructed and utilized. This WTA circuit constitutes our main building block. It comprises of a feed-forward single layer spiking neural network. The input layer is made out of spiking nodes whose firing activity is governed by a poisson process. External variables undergo a population coding that determines the modality of this poisson process. In the example offered by Nessler et al, the external variables are intensity values, pixel values, form a static visual stimulus, an image. The population coding polarizes the pixel intensity values into binary On-Off states which directly determine the firing probability of the poisson process. A spiking neuron is assigned to encode each state of the population code generated for each pixel. In this case, two spiking nodes per pixel. An On-node and an Off-node. The firing rate of these neurons is proportional to the state of the node in the population code [9].

Selective Attention delivers a strategy to economize computational power and reduce its entropy. Its evolutionary motivation comes for the organism's need to detect prey rapidly. Itti et al. propose a framework for attention involving interactions between bootom-up cues that are stimulus driven and top-down cues that are task-dependent [5]. The bottom-up cues are triggered by a mechanism for static feature detection and possibly also temporal event detection. Top-down attention may originate from predictive mechanisms that

bias selectivity. Top-down cues may also arise from indepdendent motor control [10].

Extending the SEM model with feature-based modules and providing input that is previously filtered by an attention module we demonstrate a reduction in dimensionality, increase in computational efficiency and invariance to scale and translation.

Chapter 2

Methods

2.1 Object Recognition using Spike-based Expectation Maximization

2.1.1 Extending SEM by learning orientations

The original SEM model is made out of an input layer receiving signals from external variables x. Population codes determine the state of a set of spiking neurons y for each node x. The spiking pattern of each node y follows a poisson process. Spikes generated from all y neurons serve as input to a WTA circuit of z neurons with weights w for each neuron. The z neurons form the output layer. The use of notations x, and z is to draw the connection to the Expectation Maximization algorithm that the model approximates. As it tries to maximize $E_p * [log\ \ p(y|\mathbf{w})]$ with $q(z|y) = p(z|y, \mathbf{w}^{old})$ in the E-Step, where $\mathbf{w}^{old}$ is the weight vector for each of the z neurons and replacing w with updated weights for the M-step [9, 8, 4]. Applied to an example, x variables are pixels from a static image of a handwritten digit. Two y neurons fire antognistically, depending on the intensity level of their associated X node (where $X \in x$). Nurons in the z layer produce a firing pattern that is distinct for to the hidden cause. The handwritten digits can be categorized into distinct classes and we see that the SEM can produce a unique firing pattern for each class through unsupervised learning [9].

We will keep the WTA circuit as our main building block. We will also preserve the hebbian learning rule to update the weights of z neurons to use for the M-Step [9]. The population coding for polarizing intensity values to drive the poisson process will remain useful, but the population coding will also be the first entry point for the extension.

The current encoding of external variables accounts for the intensities of the spatial units, pixels, of the presented stimulus. The encoding of intensities is performed through a population coding by antognistic binary nodes per pixel

that drive a poisson process [9]. Parallel to these intensity encoded nodes, we add a WTA circuit per pixel that determines the preferred orientation of this node relative to its spatial neighbors. This creates an orientation map of the presented stimulus. Whilst counter-intuitive with traditional learning models, SEM benefits from elaborating the dimensionality of WTA's feature space as this increases its resolution for detecting correlations between an output node z and input nodes y on a linear scale. Since SEM aims to reduce dimensionality, it is preferable to describe it as an elaboration of dimensionality. The added dimensions, or nodes, do not carry new information, but rather refine its representation. Recalling the use of using population coding to encode in atagonistic (On-node, Off-node) fashion, thus letting the WTA learn the likelihood of an input node firing and not firing explicitly, as shown by 2.1.

$$p(z = 1|y) \propto y * p(y = 1|z) + (1 - y) * p(y = 0|z) \tag{2.1}$$

where

- z denotes an output node,
- y denotes an input node

As we introduce the orientation map we add additional operands to 2.1 to account for the node's preferred orientation.

$$\begin{aligned} p(z = 1|y_I \cup y_O) \propto & y_I * p(y_I = 1|z) + (1 - y_I) * p(y_I = 0|z) \\ & + y_o * p(y_O = O_p|z) + \sum_{i \neq o_i} (1 - y_O) p(y_O = o_i|z) \end{aligned} \tag{2.2}$$

where

- y_I denotes an input intensity node,
- y_O denotes an orientation input node,
- O denotes the set of orientations available. Orientations can be defined discretly and arbitrarily (e.g. 30, 60,...180 degrees) or they can be learned [9],
- O_p denotes the preferred orientation

We redesign the network with a cascade of hierarchical WTA circuits. The input layer is a matrix of WTA circuits per pixel. Each input WTA circuit decides on the preferred orientation and intensity of its input. We will experiment with configuring the input WTA circuit to only relay intensity, or only orientation as depicted in fig. 2.1, or both information.

The WTA circuits responsible to determine the preferred orientation of all input nodes are activated by convoling the stimulus with a bank of two-dimensional Gabor filters. The filters are defined with different scales and angular orientations from a predefined discrete set. Figure 2.2 provides an illustrative

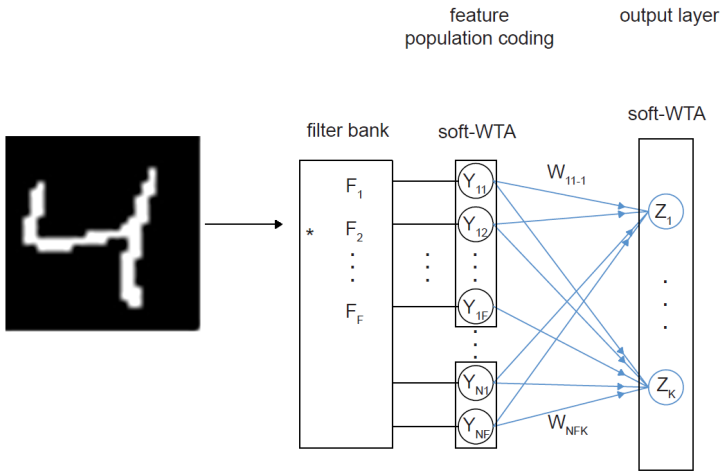

Figure 2.1: SEM extended with a filter bank and an orientation discriminating WTA circuit for learning an orientation map of a visual stimulus.

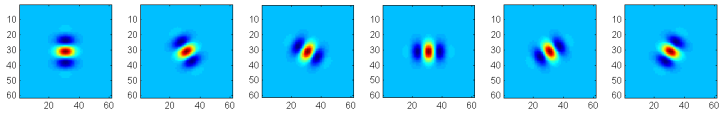

Figure 2.2: Example of Gabor filters (real part). Defined at orientations [0, 150] degrees with increments of 30 degrees.

example of such Gabor filters. By comparing the magnitude of responses between the filters at each pixel we can decide on the pixel's preferred orientation. Talking about the preferred orientation of a single pixel does not actually carry much meaning. The transfomations do yield responses for each pixel but they only become informative in relation to the responses of its neighbors. Examples of filter responses to a binary image stimulus are shown in fig. 2.3

Parameterized Gabor functions are an adequate approximation of simple cells in the primary visual cortex [13]. Daugman demonstrates the construction of a neural network to achieve this transformation [3]. However, this work adopts the traditional systems' approach for defining and applying the filters.

2.1.2 Extending SEM by learning hidden features

We have seen the computational power of the SEM model as an unsupervised method for identifying hidden causes. So far the hidden causes have been used synonymously with predefined classes (e.g. numerical digits from the MNIST database [7]). We extend the SEM model in a way that breaks this assumption. We insert an additional layer, a WTA circuit, responsible for learning hidden causes that depict abstract features of the object we attempt to detect

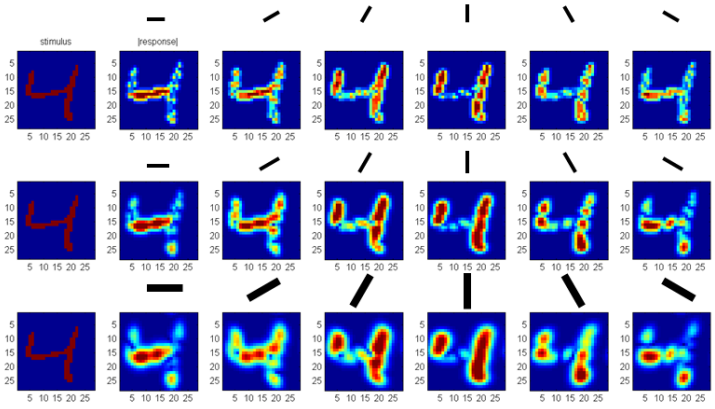

Figure 2.3: Example of filter response magnitudes when convoling a binary image with Gabor filters defined at different scales and orientations.

and recognize. This feature layer will contribute to the bottom-up learning as we expose it to the low-level input and and have it drive the WTA circuit already encountered in the original SEM architecture. With this additional feature-WTA circuit introduced, we no longer require presentation of the entire stimulus but will restrict stimulus presentation to subregions within the space of a stimulus. These subregions may represent salient regions within a stimulus. The definition and method of selecting these subregions will be discussed in more detail as we discusss the object detection module.

2.2 Object Detection Using Visual Attention

2.2.1 Achieving invariance

Itti et. al anchor their bottom-up computational model of attention as a saliency search within a visual scene. They demonstrate how attention is achieved in an image based environment. The image is evaluated for conspicuities in features such as illumination, color, texture or other. The feature extraction is pre-attentive. A spatial map of each feature at different spatial scales is unified into a single conspicuity map for this feature. The conspicutiy maps are combined linearly into a saliency map of the image. The saliency map represents a reconsiliation of the pre-attentive features as the magnitudies need to be normalized before we can can combine them. The saliency map seeds the search to locate visual objects in a scene. The objects can later be processed for recognition with less computational overheads [6, 5].

The above saliency mechansim defines the distribution from which we sample windows of attention. The windows of attention are forwarded to the recognition module as a continuous feed of input. Itti et al. emphasize that their

model does not cover any top-down attention components, yet they address the importance of preventing the 'focus of attention from immediately returning to the previously-attended location' [6]. This is expected to happen in a purely bottom-up method as the saliency value remains constant. Therefore, we need to bias the model to attend to the second-most-likely salient location. Since we are using windows of attention as input to a recognition system, inhibition of return (IOR) enables such input to be more diverse. The proposed mechanism for IOR, adopted fro Itti et al is the convolution of the saliency map with a Off-centre-On-urround kernel that inhibits the centre located on top of the previous window of attention and enhances the saliency of its surroundings. Itti et al. point out that the effect of IOR should remain transient and need only be maintained temporarily [6].

Olshausen et al.'s earlier attempt to formalize a framework for attention includes the concept of bottom-up flow of information before attempting recognition but also mention the presence of a top-down mechansim intercepting the signal. This mechanism arises from "control-neurons" that dynamically modulate synapses, or weights, for routing a selected spatial window from lower levels in the visual system to higher ones **??**. They describe how scale-invariance is achieved through magnification of the window of attention to fit a "canonical reference window". This concept is computationally equivalent to resizing an image via interpolation leading to a blurred appearance of the attended window inside the reference frame. Their control units come into play as they can shift the window of attention up or down scale levels or translate within a single plane, scale. This conscious control represents behavioral control of attention that can arise from motor control. This motor control can be voluntary or not, however the authors are more focused on the case of involuntary control. To demonstrate the employment of the same control neurons as part of a closed-loop autonomous system, Olshousen et al. describe a strategy that governs these control units. This strategy is made out of:

- Low-pass filter the visual scene into blobs.
- Point the reference window to the location of the blob, adapting to its topography (i.e. size and location).
- Pass original input represented by the blob to the recognition module.
- Assess match of the presented object with previously learned objects. Take action according to recognition results (e.g. learn as a new object, discard it, perform a task, etc.) [10]

The authors provide an elaborate derivation and description of this model. We see them taking the initial steps of what we learn from Itti's salient-based model in addition to the notion of turning the open-loop system into a closed-loop one through the feedback signal of modulatory control neurons whose activiy is a relay of the response of a recognition module. However we encounter some gaps

in how some of the intermediate steps of modulation are defined, as these may depend on application or the recognition module itself, how it operates, how it responds.

A computational model of top-down attention is proposed by Baluch et al. The top-down model is described as a mechanism that influences the stimulus drive generated from the familiar bottom-up mechanism. The influences come in the form of feature bias, spatial bias, context or a task being carried out. These same influences can generate an attention field independent of the stimulus drive. The attention field and modulated stimulus drive are multiplied and normalized to yield a response to apply detection on. An analogous top-down attention model is presented that involves a learned approach. A learner is presented with bottom-up derived features to predict top-down saliency. Top-down saliency is multipled by bottom-saiency to form a unified priority map over the visual space [1]. This model provides a concise interaction of bottom-up and top-down mechanisms, thus bridging bridging Itti et al's with Olshausen et al's frameworks.

The SEM model requires some extensions to achieve scale- and translation invariance. Scale invariance can be achieved through the use of intermediate layers that learn features of visual objects that are more abstract than pixel intensity values. We can predefine these features in the form of orientation maps. It is also possible to let the SEM model learn abstract features driven by the stimuli. The features can be pre-extracted using decimated versions of the orginal visual stimulus, thus reusing existing neurons in the network as if they were with enlarged receptive fields. We expect that these abstract features can pick up simple shapes such as oriented bars as well as more complex shapes. The network wiring will emerge from the features extracted by the data and the interactions within the stochastic processes within the SEM. The model will become invariant to XY-plane translations through the use of selective visual attenion. Bottom-up saliency-based attention is employed following the model proposed by Itti, Koch and Niebur [6]. The introduction of a top-down component can be made by using the spikes of the output layer as a feed-back signal to modulates attention.

2.2.2 Feature-population coding

Towards invariance to translation we employ Itti et al.'s bottom-up saliency-based attention mechanism. It serves the purpose of locating regions within a visual scene to present to the network. Population coding will be used for encoding an analog and/or binary state of a neuron, y neurons, into spike trains. The state of these neurons will be derived from pixel intensity values, binary, and analog filter responses when convolving salient regions with filters from a predefined filter bank. Population coding derived from filter responses will be referred to as feature population coding. The original simple population coding

and feature population coding generate input to the new network iluustrted in fig. 2.4.

We refer to the original population coding as 'simple' because it operates directly on the bi-state external variable and only generates spikes that elaborate on these two states. The feature popluation coding investigates the use of a soft-WTA circuit to determine the state of the spiking neuron in a population 2.1. The population represents a set of feature states and the soft-WTA circuit determines the dominant state(s). In this case the population represents the response of a node after convolving the stimulus with a bank of Gabor filters. The WTA circuit draws the preferred orientation from the stochastic process whose distribution is defined by the analog response to each filter. The use of winner-take-all circuits correlates with the HMAX framework [13, 12]. Employment of the soft-max operator allows for a representation of intermedia features (i.e. an angular orientation that lies between two states in the predefined set). It also provides good mitigation against local maxima during the learning process.

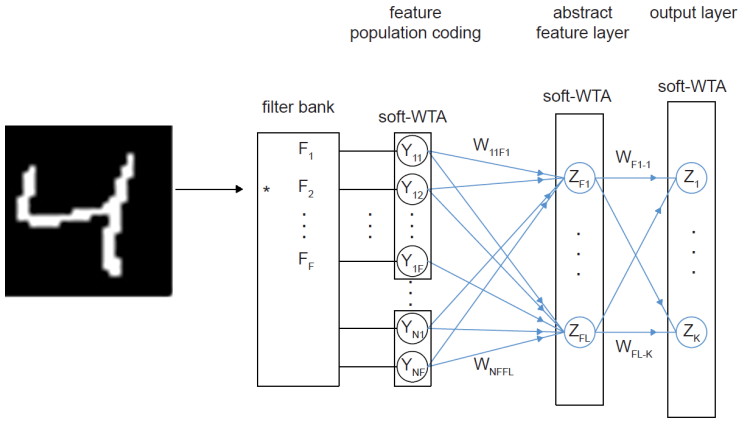

Figure 2.4: The network with which SEM is extended. The network is composed of multiple layers, namely a layer of spiking y neurons that are stimulated by intensities and filter responses of the presented salient image region, a layer of z neurons inside a WTA circuit that learn the patterns presented to them which will be referred to as the f layer, or f neurons, and finally the WTA circuit of z neurons that learn the spiking pattern of the f layer in a familiar and equivalent WTA fashion.

The bottom-up attention mechansim employed is a varied implementation of Itti et al.'s model. Their model defines a salient-based mechanism to evaluate a visual scene on location worth attending to. The commonalities between this intepretation and the original model are:

- the use feature maps based on intensity and orientation,
- detecting conspicutity from center-surround differences,

- normalizing feature conspicuity maps to combine them into a single saliency map.

As for differences, unlike the original model, this intepretation:

- use feature population coding, Gabor filter transformations with soft-WTA circuits, instead of linear filters,
- omits color feautres, focusing only on binary image stimuli,

2.2.3 Learning abstract features

The attention mechanism provides a continuous feed of subregions in the original stimulus space that may contain patterns worth learning for recognizing objects. This can happen whether the attention mechanism is a purely stimulus-driven or modulated by top-down cues. We are willing to restrict the input space used for training and testing the recognition system to patterns that qualify as salient. The SEM model treats such salient windows as its external variables x. Population coding, simple and feature-population coding, produces spike trains that represent the intensities in the pixels and their the orientation map of such a window. A layer of f neurons interact within a soft-WTA circuit. These f neurons employ the same learning rule. Their WTA-circuit is essentially the building block brought over from the original SEM model. This layer is described as the abstract feature layer, or F-layer, as it learns to disriminate the abstract features provided by the salient window input. The spikes emerging from the F-layer activate the z neurons in the output layer. The output layer remains responsible for learning the hidden causes. Except that it now disriminates such causes in the distribution of abstract feautres, of far more reduced dimnensionality.

Figure 2.6 illustrates the wiring of a spiking neural network responsible for recognizing hidden causes in visual objects. The causes are represented by spiking patterns of the output-layer neurons that discriminate the spiking pattern of preceeding neurons in an intermediate layer. Neurons of the intermediate layer respond to the presence of abstract features present in a series of windows of attention. With the presentation of a stimulus a detection module employing selective visual attention generates this series of windows of attention which are encoded into spike trains through population coding. Population coding may be 'simple', based on pixel intensity values, more complex by evaluating the response of the sub image to filters, or both.

Here it becomes necessary to point out a predicament. Attention windows arrive at the recognition module serially. This does not cause any problems for the f neurons as the scope of their operation is within a single window of attention. However, in the case z neurons in the output layer need a larger scope as their response needs to be associated with the original stimulus. We may at first look at it as a problem of detecting a temporal pattern. Panchev et. al demonstrate the detection of temporal patterns using spiking neurons.

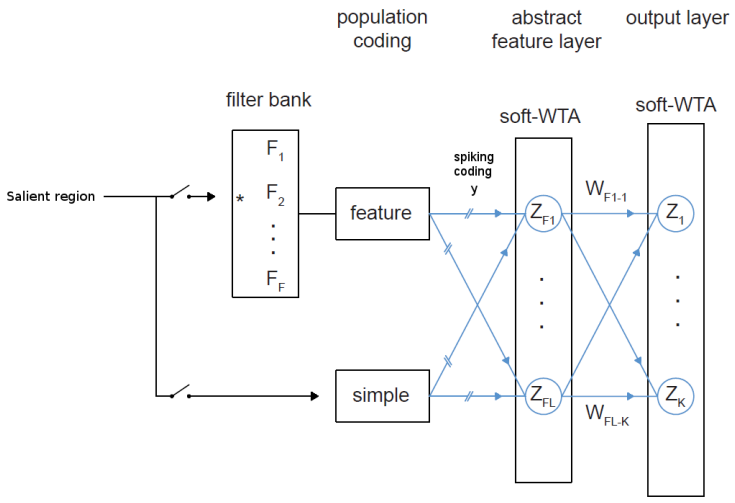

Figure 2.5: The architecture of the recognitiion module of the extended SEM model. The input is restricted to salient regions provided by the detection module via selective attention. The windows of attention is passed on to population coding, wheter simple bi-state coding of pixel intensity values or convlution with Gabor filters and WTA-circuit that discriminates the preferred orientation of the external variables. We will determine if the use of both population coding schemes are necessary or if only one is sufficient. Spikes generated from the population coding activate a soft-WTA circuit in the abstract feature layer, the F-layer. F-layer neurons learn patterns present in the window of attention. The neurons in the output layer learn hidden causes represented by the spiking distribution of the f-layer.

There model is employed in the context of training robots a natural language of instructions. The learning mechanism models dendritic and somatic membrane potentials, thus not restricting itself to only timing information through STDP but also temporal integration of amplitude information. They key takeaway of their work is in achieving a temporal "delay" mechanism and to encode the spatio-temporal structure of the delay mechanism into spikes and have neurons learn such spiking patterns [11]. A similar solution to detecting spatio-temporal patterns is demonstrated by Byrnes et. al. Their work abstracts the spatio-temporal patterns as a temporal sequence of symbols. The spiking patterns of neuron reflect the detection of a symbol suceeding the detection of a previous symbol. The preceeding symbol primes the model and this biases the model when the most recent symbol is introduced [2]. This state machine logic aligns well with the problem at hand. To draw an analogy with the model of Byrnes et al. The windows of attention would be represented by symbols and the stimulus would be encoded by the temporal sequence of such symbols. However, our model may not necessarily benefit from learning suc explicit temporal pat-

terns of windows of attention. The presence of symbols is sufficient, yet their temporal order may be negligible. It may suffice to treat the the sequence of symbols ABC the same as the sequence CBA, ACB or any permutation of this sequence. Therefore, a solution to this problem is likely to lie in the use of a state machine that models memory and preserves the state of the output layer for different windows of attention belonging to the same stimulus, while remaining invariant to the sequence of such states. We learn that markov chains present a possible approach for maintaining temporal states. Other works dealing with the detection of spatio-temporal patterns also teach us that this may be achieved through the use of self-excitatory synapses for neurons in the output layer. The different ways of approaching this are quite intriguing. We opt for a simple time-delay mechanism:
Time-delay components are needed to reconcile attention windows arriving at the recognition module serially. For associating the response of z neurons to a stimulus we employ a time-delay component in the network. The time-delay component is represented by associating f neurons with multiple WTA circuits at the same time. The WTA circuits differ in the timing at which they inhibit the f neurons. The number of WTA circuits matches the number of windows of attention sampled from the same stimulus. Effectively this results in superimposing f-layer spike trains for each window of attention assicated with the same stimulus. Furthermore, this time-delay mechanism effectively discards the temporal information of the sequence.

We can now explain the flow of information inside the model illustrated by **??** the following way: Given a stimulus the detection module provides a series of attention window in serial fashion. Population coding represents each window in spike trains that excite learners in the intermediate f-layer. For each window of attention a WTA circuit with a delay component inhibits f neurons. This leads to the superimposition of f spike trains due to different windows of attention of the same stimulus. The superimposed f-layer spike trains provde excitatory input to z neurons that compete in a soft-WTA circuit and form the output layer of the recognition module.

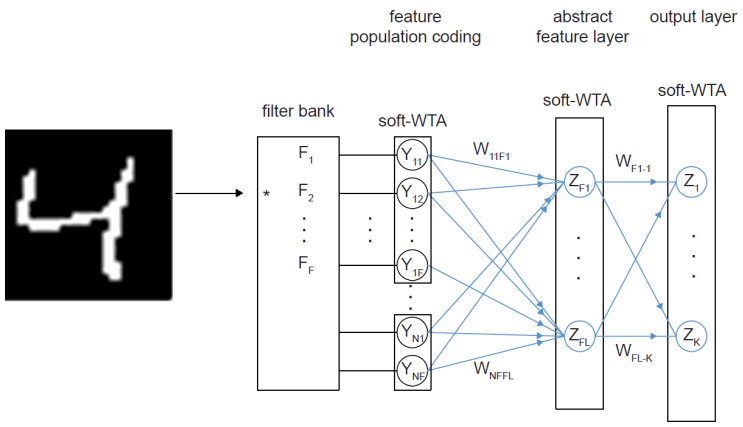

Figure 2.6: The network of the recognition module extending the SEM model. Input is provided by the attention mechanism in the form of windows of attention. Population coding encodes analog intensity or filter response values into spike trains which activate a soft-WTA circuit. Spikes generated from the f neurons of the soft-WTA circuit in the abstract feature layer excite z neurons in the output layer. z neurons in the abstract feature layer are responsible for learning and predicting the abstract pattern in the window of attention. z neurons in the output layer are responsible for learning spiking patterns of the preceeding layer associated with hidden causes in the orignal stimulus.

Chapter 3

Results

3.1 Exploiting inherit noise of a stochastic process

The first extension to the SEM network was to enrich the representation of the external variables x. Instead or in a addition to bi-state pixel values we convolve the image input with a set of Gabor filters. A WTA-circuit is appplied to the response of each pixel to the Gabor filters to determine its preferred orientation through a stochasitc process. The use of the stochastic process is not as noise-free as applying $argmax(y_i)$, where y_i denotes the magnitude responses of pixel i. The $argmax$ operator would even align well with the HMAX paradigm presented in the works of Riesenhuber, Poggio and Serre**????????**. However we have found that the noise inherit in the stochastic process mitigates the computational pitfall of low magnitdue response vectors very well. To elaborate, parts of the stimulus with very low intensity values, will results in low magnitude response values for all filters in the filter bank. However the insignificant response vector for a pixel still yields a result from the $argmax$ operator. It will actually yield a near constant result. To avoid learners from learning this constant pattern that does not carry any true information, we need a mechanism to handle such low response values. The problem is not really in the response vector of all-low magnitudes. We can generalize the problem to pixel response vectors with high uniformity. We can still align this model with the HMAX model if we resort a stochastic process. The stochastic process is able to resolve high-uniformity, high entropy, distribution of pixel response vectors through its inherit noise. If a learner relies on such response vectors as weights, the spike trains coding the response vector will reflect the uniformity and the learning rule will yield accordingly uniform weight distributions for the corresponding pixel, leaving such weights computationally useless. We find that the weight vector associated with this pixel is as uniform as the response, thus with the use of a soft-WTA the contribution of this vector of weights for this pixel relative to the global weight vector of the learner is merely constant. Thus the membrane

potential of this neuron will not experience fluctuations due to these weights but rather weights whose local vectors have low entropy. Figure 3.1 depicts the distribution of weights associated with different orientations after a learning process involving handwritten digits as stimuli. We find that outer areas with generally low intensity activity are high regardless of orientation.

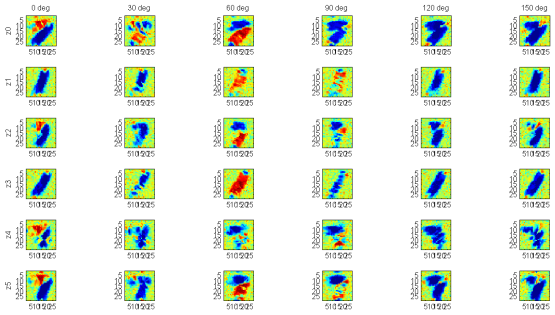

Figure 3.1: Example of z neuron weight distributions for different orientations when using a hard max operator such as $argmax$ without accoutning for the uniformity of a pixel's response. Each row of subplots is reserved for a learner Zi, while each column is associated for an orientation. This is from an experiment involving handwritten variants of the digits 1 and 7 **??**. Blue colored regions have low weight values that yellow, orange and red, represent strong weights.

If we picked the same weight from each orientation for the same learner, we'll find the distribution of this local weight vector to be fairly uniform. A local weight vector belonging to a pixel in an active region, will have some very low weight values for most orientations, and very high wiehgt values for fewer. Figure 3.2 accomplishes this. It generates a local weight vector for each pixel, calculates its entropy to visualize the non-uniformity of the weights associated with each pixel in the XY-plane. Stimuli that activate weights belonging to a local weight vector with high netropy are a mere DC component in the resulting membrane potential of a neuron. While changing the input that invokes activity in weights belonging to low-entropy local distribution, will yield significant changes in that same neuron's membrane potential. **??** provides a more compact visualizations of the weight distribution by indicating the dominant weight, or preferred orientation associated with each external variable x for different neurons z.

3.2 Measuring Saliency using feature-population coding

Saliency is measured with the use of feature maps, one for intensity features and a second for orientation features. The intesnity feature map is generated

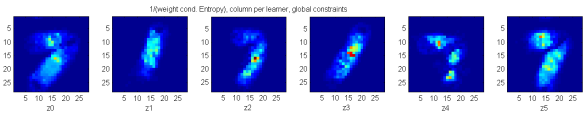

Figure 3.2: Entropy of local weight vectors representing orientations associated with the same pixel. The intensity values in the image are anti-proportional to the entropy of the pixel's weight vector. This is from an experiment involving handwritten variants of the digits 1 and 7 **??**. Bright intensities are associated with low-entropy, strongly disriminating weight vectors, while low intensities are assoicated with uniform weight vectors whose contributions are merely DC components to the neuron's membrane potential. Each image represents this non-uniformtiy distribution of different learners.

by convolving the stimulus with a Difference of Gaussians for achieving centre-surround operations. Normal localization is applied on the filter response to empahsize conspicuities and a global normalization is applied on the entire feature map for later reconciling this itenstiy contrast map with other feature maps. The same stimulus is convolved with bank of Gabor filters. For each pixel the filter responses for that pixel are taken as input to a soft-WTA circuit. The soft-WTA circuit determines the preferred orientation of the pixel. Performing this for all pixels we generate an a map of the preferred orientation for each pixel, an orientation map. In order to detect conspicuities in these orientations we measure the distance of a pixel's preferred orientation relative to its neighbors and use the local variance of these diatance measures as a measure of concpicuity in orientation. The feauture map is then normalized globally and then masked for nodes with low magnitude responses to all filters before its use alongside the intensity contrast map. The intensity contrast and orientation conspicuity maps are combined linearly into a single saliency map. Nodes in the saliency map with high response-magnitude indicate either high contrast, orientation conspicuity or both. The saliency map serves as a two-dimensional distribution from which we draw coordinates of locations worth attending to. Figure 3.4 depicts examples of saliency maps evaluated for static images.

The selection of the attention window is determined through a stochasitc process. This non-deterministic process may mitigate the problem of locking onto the same region during successive sampling events, thus increasing the chance of delivering more diverse information to the recognition system.

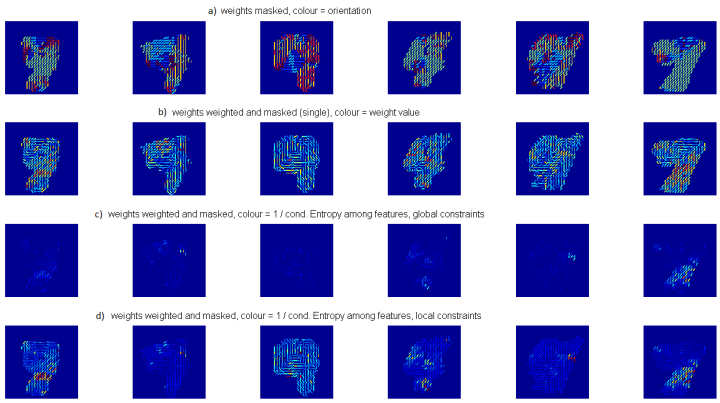

Figure 3.3: Visualizing weights of z neurons after learning orientation maps of handwritten digits. Data used: a subset of handwritten digits from the MNIST database (classes 7 and 4) **??**, no overlap between training and test sets. Each column of images belongs to a z neuron. The weight weight vectors have been reduced to two dimensional matricies by finding the dominant weight for each pixel. The preferred orientation of each pixel is determined via $argmax(\mathbf{w}_x)$ where X refers to the pixel and w is the weight vector associated with this pixel. The length of w is proportional to the number of possible orientations defined. Regions of the image with very low intensity activity have been masked and appear plain while color represents **a)** the preferred angular orientation of a pixel, **b)** the magnitude of the weight of the preferred orientation, **c)** the non-uniformity of a weight vector w_x normalized across all learners and **d)** the non-uniformity of a weight vector w_x normalized across the image.

3.3 The f-layer

The injection of an intermediate layer that actvates z neurons in the output layer provides a reduction of dimensionality for these neurons and exploits the input provided from the attention mechanism. This layer is mainly concerned with learning abstract geometrical shapes based on low level features such as intensities and orientation maps.

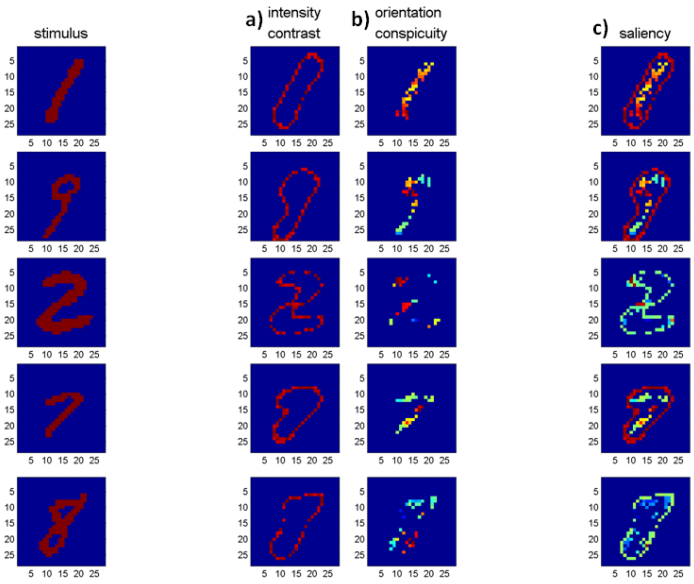

Figure 3.4: Examples of saliency maps generated from handwritten digits from the MNIST database**??**, where columns depict **a)** intensity contrast; response to convovling stimuli with a Difference of Gaussians kernel, **b)** orientation conspicuity; local variance of a pixel's preferred orientation relative to neighboring pixels, **c)** stimulus map of input stimuli; linear combination of intensity contrast and orientation conspicuity.

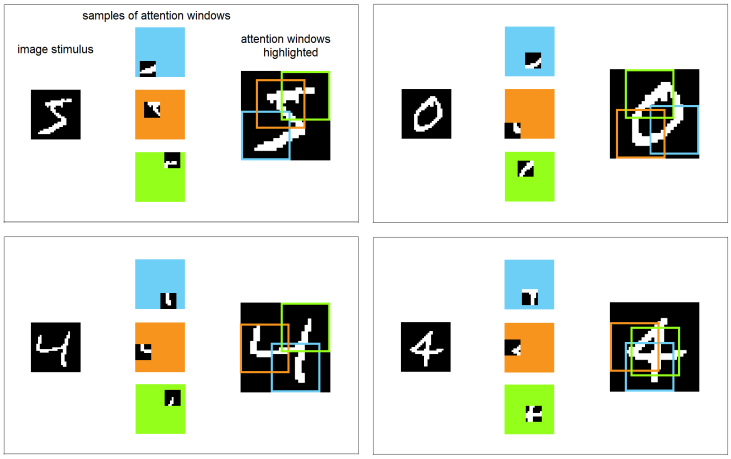

Figure 3.5: Examples of attention windows determined by sampling from a two-dimensional saliency distribution.

Chapter 4

Discussion

Chapter 5

Conclusion

That's all folks!

Chapter 6

Acknowledgments

Michael My family: Sahra, father My friends Malte Alf Matthew Cook INI

Appendix A

The First Appendix

The `\appendix` command should be used only once. Subsequent appendices can be created using the Chapter command.

Appendix B

The Second Appendix

Some text for the second Appendix.

Bibliography

[1] Farhan Baluch and Laurent Itti. Mechanisms of top-down attention. *Trends in neurosciences*, 34(4):210–24, April 2011.

[2] Sean Byrnes, Anthony N Burkitt, David B Grayden, and Hamish Meffin. Spiking neuron model for temporal sequence recognition. *Neural computation*, 22(1):61–93, January 2010.

[3] John G Daugman. Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression. *IEEE Transactions On Acousticsm Speechm and Signal Processing*, 36(7):1169–1179, 1988.

[4] Stefan Habenschuss, Helmut Puhr, and Wolfgang Maass. Emergence of optimal decoding of population codes through STDP. *Neural computation*, 25(6):1371–407, June 2013.

[5] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, March 2001.

[6] Laurent Itti, Christof Koch, and Ernst Niebur. Short papers from the April, 1998 Social Capital Conference at Michigan State University. *Journal of Socio-Economics*, 29(6):579–586, November 2000.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the ...*, 86(11):2278–2324, 1998.

[8] Bernhard Nessler, Michael Pfeiffer, Lars Buesing, and Wolfgang Maass. Bayesian Computation Emerges in Generic Cortical Microcircuits through Spike-Timing-Dependent Plasticity. *PLoS computational biology*, 9(4):e1003037, April 2013.

[9] Bernhard Nessler, Michael Pfeiffer, and Wolfgang Maass. STDP enables spiking neurons to detect hidden causes of their inputs. *In Proc. of NIPS 2009: Advances in Neural Information Processing Systems. MIT Press*, 22:1357–1365, 2010.

[10] Bruno A. Olshausen, Charles H. Anderson, and David C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition

based on dynamic routing of information. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 13(11):4700–19, November 1993.

[11] Christo Panchev and Stefan Wermter. Temporal Sequence Detection with Spiking Neurons : Towards Recognizing Robot Language. *Connection Science*, 18(1):1–22, 2006.

[12] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–25, November 1999.

[13] Thomas Serre and Maximilian Riesenhuber. Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model , and Implications for Invariant Object Recognition in Cortex. *Methods*, (July), 2004.