Sparse Restricted Boltzmann Machines as a model of the Mirror Neuron System

Mark Marijnissen
markmarijnissen@gmail.com
September 22, 2011 (revision)

Bachelor Thesis Artificial Intelligence
Supervised by Sebo Uithol & Pim Haselager
Radboud University Nijmegen

Abstract

I will defend a two-fold hypothesis. First (1), Restricted Boltzmann Machines (RBM) can successfully emulate the human Mirror Neuron System (MNS), by using association. This supports the association hypothesis, which states that Mirror Neurons are a by-product of associating perception with motor codes. Second (2), Sparse Coding is an necessity for Mirror Neurons to emerge from association learning.

Methods: I simulated a dataset with three actions and two goals. Each stimuli has five features; three to indicate the actions, two to indicate the goal. I trained Sparse RBMs of various sparsities and sizes to model the MNS.

Conclusion 1: RBMs prove to be successful in emulating various aspects of MNS behavior. This includes action execution, observation, imitation, goal inference, dealing with missing values (i.e. in the dark) and handling multiple modalities (i.e. integrate vision and proprioception). The performance, strength and certainty of responses of the model in different circumstances is similar to data from experiments.

Conclusion 2: The Mirror Units that emerge are only dependent on the Sparse Coding. They are robust to network size, although size tends to diminish strength of the response. The units are capable of representing one or more causes (i.e. a single action, goal, or both). The optimal sparsity turns out to be $1/N$, where N is the number of distinctions the Unit needs to make.

While Mirror Units emerge, association learning was insufficient to create two distinct populations , as found in the brain. This might be solved with the addition of classification learning.

Keywords: Restricted Boltzmann Machines, Mirror Neuron System, Sparse Coding, Associative Memory.

Table of Contents

1. Introduction

Mirror Neurons have caused much excitement since their discovery in the early 90's. Mirror Neurons are neurons that become active when an action is executed, as well as when that action is observed being done by another. Not only do they link action with perception, they also link self with others. This provides a foundation for theories about empathy, imitation and action understanding. (Heyes, 2010).

There are two competing theories that explain the existence of Mirror Neurons: the adaptation and the association hypothesis. (Heyes, 2010). The first theory argues that Mirror Neurons are the result of natural selection for either action understanding or imitation learning. The second theory views Mirror Neurons as a by-product of association learning. It suggests Mirror Neurons emerge by associating the perception with the execution of an action.

This thesis will provide evidence to support the association hypothesis. However, I will argue that associative learning is not enough for Mirror Neurons to emerge: we need Sparse Coding too. Sparse Coding limits the number of cases a neuron responds to, or limits the number of neurons that are active at any given time. Without Sparse Coding, associations can be learned without creating neurons that respond to a single goal or action.

In order to do this, I have modeled the human Mirror Neuron System (MNS) with a Restricted Boltzmann Machine (RBM). RBMs are undirected belief networks that learn a generative model of the observed data. They can be used as an associative model, because given a partial observation, they can generate the remainder. I used this to associate stimulus (perception) with response (action). This way, RBMs can perform typical MNS tasks such as action execution, observation, imitation and goal inference.

There are more computer models of the MNS based on association. For example, Chaminade et al (2008) have learned viseomotor associations to a robotic hand from self-observation. This could be used to imitate a human hand. However, they did not focus on the occurrence of Mirror Units (which may not have emerged due to the lack of sparse coding). There are a few other MNS models that use the association hypothesis, but none of them investigate sparsity. See Oztop et al (2006) for an excellent review on the different goals and methodologies used to model the MNS. To be clear, this thesis investigates two hypothesis:

1. Restricted Boltzmann Machines can successfully emulate MNS behavior, by functioning as an associative memory.
2. Sparse Coding is a necessity for Mirror Units to emerge from learning associations. In fact, there is an optimal sparsity for Mirror Units.

I will focus on the accuracy of the model to verify the first hypothesis. So, unlike other models, I will focus not only on the performance of MNS. Instead, I also focus on various other aspects, such as the strength of the response under different circumstances. For example: Do Mirror-Units respond with less strength if there is no goal-context, just like real Mirror-Neurons? Note that I use "units" when referring to the computer model, and "neurons" when referring to actual biological neurons.

In order to investigate the second hypothesis, I will experiment with different network sizes and sparsities.

The thesis is structured as follows: First, I will briefly describe relevant properties of Mirror Neurons in chapter 2. Then, I will explain what RBMs are and how they learn in chapter 3. These RBMs will use Sparse Coding, which I describe in chapter 4. The MNS, RBM and Sparse Coding provide the background knowledge required to understand the experiment as described in chapter 5. I will present the results in chapter 6, and draw conclusions in chapter 7.

2. The Mirror Neuron System

In this section, I will explain briefly various properties of Mirror Neurons that I will attempt to model.

Mirror Neurons are found using single cell-recordings in the F5 area of the macaque monkey, which is a part of the premotor cortex. Only 92 of the 532 neurons (17%) that are recorded by Gallese et al (1996) have mirror properties. The other neurons only respond to action execution.

While single-cell recordings are not possible in humans due to the destructive surgery that is required, various indirect evidence suggest the existence of a human Mirror Neuron System (MNS). (Rizzolatti & Craighero, 2004)

Mirror Neurons only respond to *goal-directed* movements. They do not respond to random movements or movements without a goal (i.e. grasping without an object that can be grasped). Note that in this case, "goal" refers to a successful execution of the grasping action: You have an object in your grasp. Experiments were conducted in which the object is hidden behind a screen. When the observing monkeys knows there is an object, half of the Mirror Neurons still fire. A quarter of them even fire at the same strength as with full vision. (Umilta, et al., 2001)

In fact, the actual movement can be completely opposite, as long as the goal is the same. In an experiment, monkeys grasped an object with either pliers or reverse pliers. (Umilta, 2008) The former required the monkey to close the hand to grasp, while the latter required the monkey to open its hand. In spite of this difference, Mirror Neurons were found that respond to both cases.

So Mirror Neurons do not respond exclusively to the visual perception of movement, but they are able to take contextual information into account, such as the object grasped or tool used.

In both action execution and observation, monkeys could see the action being executed. To exclude visual perception as explanation for action execution, monkeys executed actions in a dark room. The Mirror Neurons still fired (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). In other experiments, the monkeys only *heard* action execution, and 15% still responded. Rizzolatti and Craighero (2004) conclude that mirror neurons can respond to multiple modalities and fire as soon as the perception contains enough clues to infer an action.

These clues can be quite general, because Mirror Neurons also respond to perception of other species. This means Mirror Neurons do not respond to direct perceptions, but to more general features extracted from them. It is thought that Mirror Neurons respond if the action is in the action repertoire of the observer. For example, human mirror neurons respond to monkeys that are "talking", but not to a dog that is barking. (Buccino, et al., 2004)

This is congruent with what we know about the brain regions involved in the human MNS. Kilner et al (2007) describe this as follows: The Superior Temporal Sulcus (STS) provides the input for the human MNS. The STS is known to encode body positions and specific actions. The STS is reciprocally connected to area PF of the inferior parietal cortex. The area PF, in turn, is reciprocally connected to the F5 area of the premotor cortex. While the STS is often considered a part of the MNS, the do not fire on action execution. Mirror Neurons are only found in area PF and F5.

Gallese et al (1996) defined several types of mirror neurons according to their response profiles:

Type of Mirror Neuron:	Percentage	N	Response profile (M=Motor, V=Visual)	Lowest common property in motor and visual response profile
Strictly Congruent	31,5%	29	M: Specific grip. V: Specific grip.	Specific grip. (e.g. grasping with precision grip).
Broadly Congruent	60,9%	56		
Group 1	7,6%	7	M: Specific grip. V: Various grips.	Specific action (e.g. grasping with a hand).
Group 2	50%	46	M: Specific hand action. V: Various hand action.	Specific category of actions. (e.g. hand actions)
Group 3	3,2%	3	M: Specific action. V: Various actions.	Specific goals (grasping to eating)
Non-congruent	7,6%	7	M: Various actions. V: Various actions.	Object-related actions.
Total	**100%**	**92**		

Figure 2-1. Different categories of Mirror-Neurons, as defined and found by Gallese et al (1996). Table adapted from Uithol et al (2008).

As you can see, Mirror Neurons often respond to a more broad range of perceptions than actions.

3. Restricted Boltzmann Machines

In this chapter I will explain what RBMs are, how they are trained and how they can be used for classification. With this understanding, I can explain how RBM will be used to model the MNS. Finally, I will give different arguments that support the use of a RBM.

3.1 Restricted Boltzmann Machines

A Restricted Boltzmann Machine (RBM) is an undirected belief network that learns a generative model of the observed data. It consists of two layers. The hidden layer, containing latent variables (y), is used to generate the visual layer, containing observed variables (x). While generation, $p(x|y)$, is learned, the undirected connections also allow recognition, $p(y|x)$. The layers are fully connected, and there are no connections between units of the same layer. The network uses stochastic, binary units with a logistic sigmoid activation function: $1/(1 + \exp(-x))$.

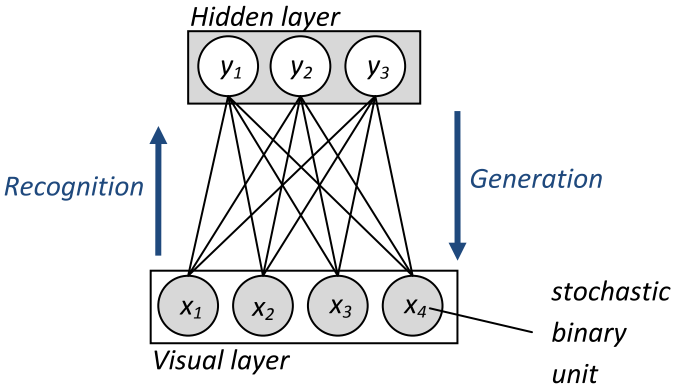

Figure 3-1. A Restricted Boltzmann Machine.
Since there are no connections within the same layer, the activation function can update all units simultaneously.

3.2 Training

RBM learn a generative model using the Contrastive Divergence algorithm. This works as follows: the visual layer is activated using a case from the dataset. This case is reconstructed by alternating recognition and generation (alternating Gibbs sampling). This can be repeated multiple times and is known as the number of steps. The correlation between visual and hidden units is calculated, and the difference between reconstruction and the original data is used a learning signal.

$$\Delta w_{ij} = \varepsilon\left(\langle x_i^0, y_j^0 \rangle - \langle x_i^1, y_j^1 \rangle\right)$$

Where ε is the learning rate and $\langle \cdot,\cdot \rangle$ denotes the correlation. The superscript denotes the step – 0 is the original case from the dataset. Learning is often done in batches of several cases at once.

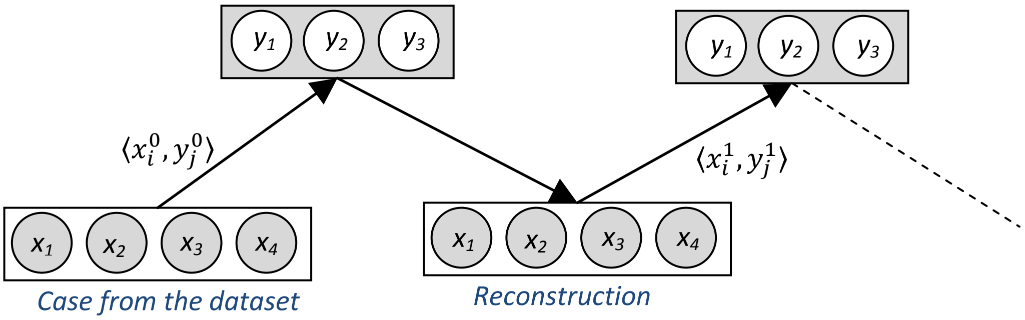

Figure 3-2. The Contrastive Divergence algorithm. Arrows represent recognition and generation.

Hinton, Osindero & Teh (2006) explain Contrastive Divergence approximates maximizing the log probability of the data. This the same as minimizing the Kullback-Leibler divergence, $KL(P^0||P_\theta^\infty)$, between the distribution of the data, P^0, and the equilibrium distribution defined by the model, P_θ^∞. Since only a few steps of alternating Gibbs sampling are used, Contrastive Divergence only approximates minimizing this divergence. This can be understood as ignoring the higher layers of an infinite belief net with tied weights.

While it is important to understand Contrastive Divergence approximates maximum likelihood, the exact details are not relevant for this thesis. For more information, see Hinton (2010) and Hinton, Osindero & Teh (2006).

3.3 Classification

After the model has been trained, recognition and generation can be used to fill in missing values of the visual layer. First, recognition activates the hidden layer using known values. Then, generation fills in missing values of the visual layer. This can be exploited for classification if the visual layer contains both stimulus and response. The model can generate the missing response when given a stimulus. In other words, the RBM uses the bidirectional connections to function as an associative memory, using a partial observation (stimulus) to reconstruct the remainder (response).

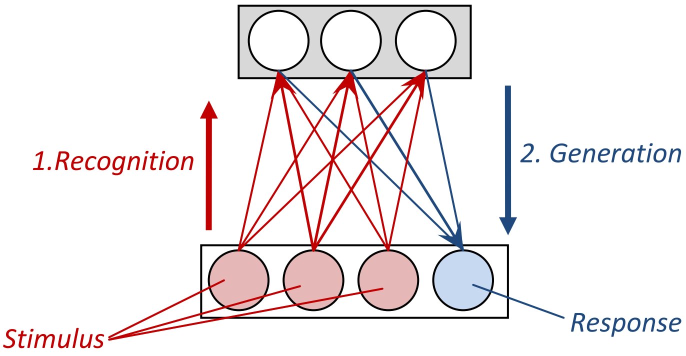

Figure 3-3. Using a RBM for the classification of a response, given a stimulus.

Before the model can be used for classification, the model is trained using stimulus-response data. In general, Contrastive Divergence learns to generate the observed data in unsupervised fashion.

However, using the generative model for classification makes learning supervised, since the observed data needs to contain the correct response.

3.4 Restricted Boltzmann Machines as the Mirror Neuron System

Classification can be used to model most tasks of the MNS. These tasks are: action observation, action execution, goal inference and imitation. Only the stimulus and response will differ between these tasks.

The stimulus will be a perception of either the “self”, or an “other”, as to distinguish action execution from imitation. Action execution and goal inference are combined efficiently in a single response. So, the response contains both an action and a goal. This can be done without any harm because there are no connections between units of the same layer. The stimuli and responses are further explained in the methods section.

Finally, action observation does not require a response, so it will only consist of recognition: using a stimulus to activate the hidden layer.

MNS function	RBM function	Stimulus	Response
Action execution + Goal inference	Classification	Self	Action + Goal
Imitation	Classification	Other	Action + Goal
Action observation	Recognition	Other	-

Table 3-1.

Now, we can map different areas of the MNS map to different units of the RBM. The STS provides the input and corresponds with the stimulus of the visual layer. The STS is reciprocally connected to area PF, so this is the hidden layer in the model. The area PF is connected to the premotor cortex, which is represented by the response units in the visual layer. This is shown in the table below:

MNS part	Function	RBM
STS	Sensory input	Visual layer (stimulus units)
Parietal MN (Area PF)	Broca’s area	Hidden layer (all units)
Frontal MN (Area F5)	Premotor cortex	Visual layer (response units)

Table 3-2.Mapping of brain regions to units of the RBM

There are some other good reasons for using RBMs:

First, RBM can handle multi-modal stimuli. Units are independent of each other, since there are no connections between units of the same layer. This allows the units to encode different modalities without probems. Also, there are no modality specific heuristics to boost performance, so a stimuli can contain proprioception, visuals and audio. The RBM will still function if one modality is missing, since performance degrades gracefully with the number of missing values.

Second, RBMs are just basic building blocks. They can be stacked on top of each other to create Deep Belief Networks. They can be extended to model time-series (Sutskever, Hinton & Taylor, 2009) or to feature three-way interactions (Taylor & Hinton, 2009). They can also be modified to work with real-valued units (with Gaussian noise).

Finally, RBMs fall under Bayesian models, which have been used extensively to model the brain (Vilares & Kording, 2011; Knill & Pouget, 2004; Friston, 2003). However, full Bayesian models are computationally intractable and therefore biologically implausible. By using only an approximation, we have not only fast inference and learning, but also a more plausible algorithm.

4. Sparse Coding

The previous section explained how a RBM can model the tasks of the MNS. Initial pilot studies confirmed that RBMs can indeed be used for these tasks. Unfortunately, almost no Mirror Units were found in the hidden layer. Learning a stimulus-response association is not sufficient for Mirror Units to emerge. We need Sparse Coding too.

4.1 Why Sparse Coding?

The problem was that units where highly unselective – and selectivity for a single action or a goal is a prerequisite for Mirror Units. As it turns out, this was caused by the distributed representation that RBMs.

A distributed representation means that given a case, many units in the hidden layer are activated. This might seem unrelated to the selectivity of a unit, but the concept of *sparse coding* provides the missing link. Willmore & Tolhurst (2001) have identified two kinds of sparsity:

1. **Population sparsity:** Given any case, there are few units active.
2. **Lifetime sparsity:** A unit is only responds to a few cases.

The two kinds are related, but not necessarily the same. We need lifetime sparsity, to create selective units. However, the distributed representation is the opposite of population sparsity (and therefore lifetime sparsity). Their relation is explained in the figure below:

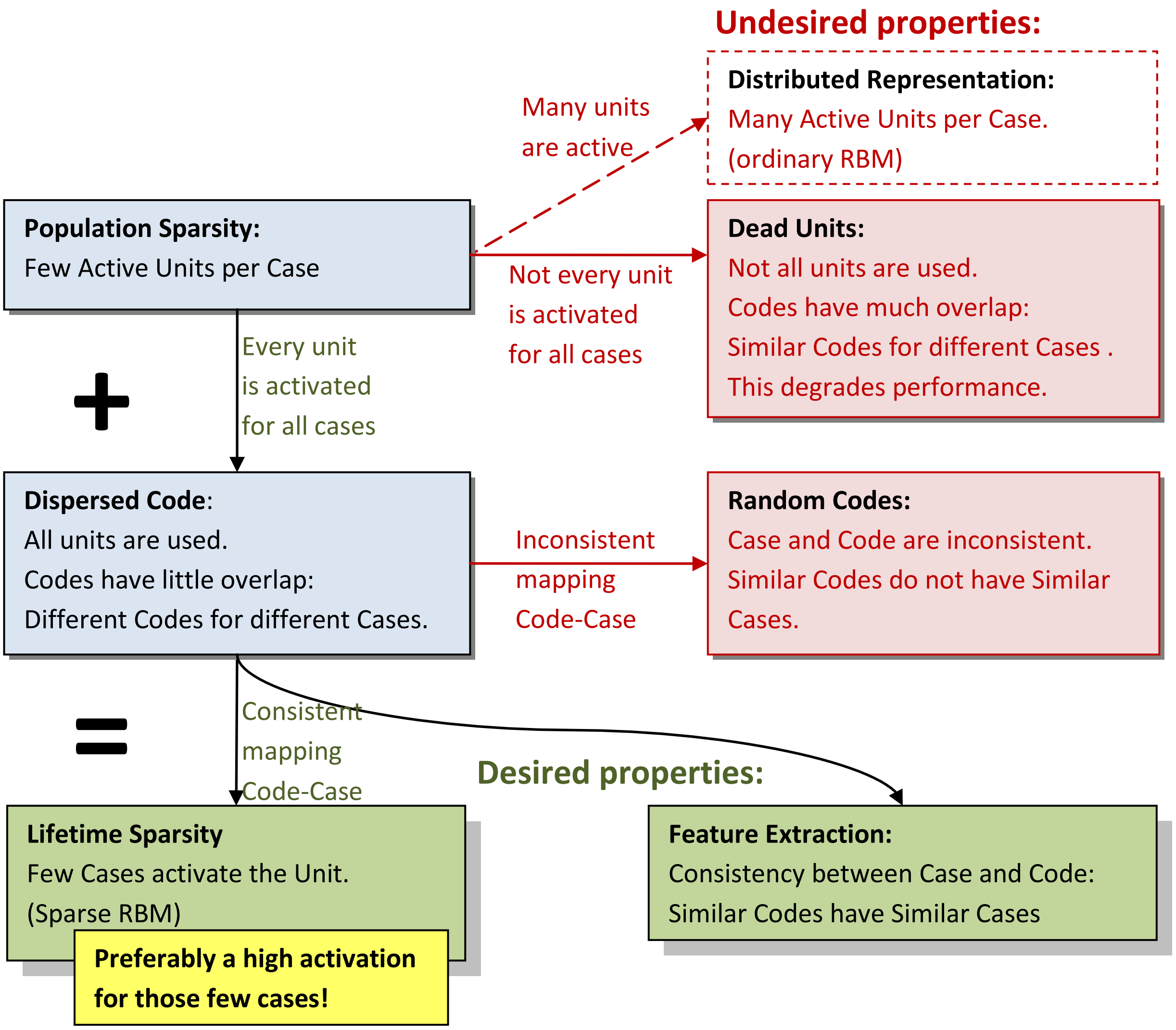

Figure 4-1. Population and lifetime sparsity are connected trough learning. Learning disperses code and prevents dead units. Learning also extracts features, which causes the unit respond to cases from a single action, rather to a random selection of cases.

Population sparsity is often expressed as the percentage of active units given a single case. If we multiply this number with the number of units (the layer size), we get the *description length*. The description length is a term borrowed from information theory, and is the number of units needed to describe a message. Units encode features, and generation is done by combining the features from activated units. Therefore, if we decrease the description length, we force a case to be described in less features. This causes the features to become more prototype-like. On the other hand, if we have a large description length, each unit will tend to encode only a fragment of the resulting reconstruction. More features (larger networks) and less population sparsity (more active units) increase discription length, and therefore directly improve generation accuracy. However, for discriminative purposes such as classification, the network might actually benefit from prototype-like features. This is because prototype-like features will likely only fire on a single class, as opposed to a fragment-feature, which can be used in many classes.

I have implemented sparsity as suggested by Hinton (2010). Here, a unit is considered sparse if the mean activation probability over all cases is low. Since learning drives activation probabilities towards one and zero, a unit is forced to respond to only a few cases (**lifetime sparsity**), rather than responding to all cases with a low probability.

p The sparsity target: the desired probability that a unit is active.
q The estimated actual probability that a unit is active.

q of batch t is calculated as follows: $q_t = \lambda\, q_{t-1} + (1-\lambda)\, q_{batch}$

q_{batch} The average activation probability that a unit is active for the current batch.
λ Decay parameter.

The cross entropy between the desired and actual distributions of being active is used as a penalty measure: $-p \log q - (1-p)\log(1-q)$. This has the derivative of $q - p$ and is scaled by a meta-parameter called sparsity-cost. So, sparsity adds three meta-parameters to the model:

1. The sparsity target p,
2. the decay rate λ,
3. and the sparsity cost, $cost$.

When we add the sparsity penalty and momentum to the learning rule, the new weight update formula becomes:

$$\Delta w_{ij}^{t} = (Momentum * \Delta w_{ij}^{t-1}) + \varepsilon * (\langle \mathrm{x}_\mathrm{i}^0, \mathrm{y}_\mathrm{j}^0 \rangle - \langle \mathrm{x}_\mathrm{i}^1, \mathrm{y}_\mathrm{j}^1 \rangle + cost * (q_t - p))$$

Momentum is added to speed up learning and avoid becoming stuck in local minima. It is a familiar trick used in most neural network applications.

4.2 Sparse Coding and Mirror Neurons

To understand how sparse coding encourages Mirror Units, we need to define the perfect Mirror Unit. A Mirror Unit needs to respond both strong and selective to a single action (or goal):

$$p(unit = 1 \mid action = x) = 1 \;\wedge\; p(unit = 1 \mid action \neq x) = 0$$

This property can be quantified as *Unit-Contrast*:

$$UnitContrast = p(unit = 1 \mid action = x) - p(unit = 1 \mid action \neq x)$$

This combines the *Strength* caused by learning with the *Selectivity* caused by the sparsity penalty:

$$Strength = (p(unit = 1) \geq 0{,}5) - (p(unit = 1) < 0{,}5)$$

$$Selectivity = p(action = x \mid unit = 1)$$

Note the discrepancy in definitions: Lifetime sparsity reduces the number of cases given a unit is active: $p(action = x \mid unit = 1)$, while Mirror Units must active a unit, given a (specific) few cases: $p(unit = 1 \mid action = x)$. When the sparsity-target becomes too low, the unit will fail to respond to all instances of a single action. $p(action = x \mid unit = 1)$ will be maximized, but $p(unit = 1 \mid action = x)$ will not. This will reduce the Unit-Contrast. Therefore, there is a lower buond on the sparsity-target.

Lifetime sparsity results in population sparsity with Sparse RBMs, because when every unit only responds to a few cases, then each case can only activate a few units. Population sparse coding is argued to be beneficial for classification tasks, since it easier to find a linear separation when only few units are active. (Ranzato, Poultney, Chopra, & LeCunn, 2006).

The brain does not maximize lifetime sparsity, but sparse coding is plausible because it requires less energy (Lennie, 2003). Indeed, neurons in the primary visual cortex (V1) are known to have high lifetime sparsity (Willmore, Mazer, & Gallant, 2011).

To summarize, there are three reasons to use sparse coding. First, it causes units to be more selective, which is required for Mirror Units. Second, it is argued it increases classification performance. Third, it is biologically plausible because it is more energy efficient.

5. Methods

5.1 The simulated world

I simulated a world in which the model can execute three different actions. It works in stimulus-response fashion, so it can only use the current world state to select the correct action. It has no memory of past world states, nor is it able to predict the future.

The model pursues one of the two goals by continuously executing the correct action. In the context of Mirror Neurons, goals can be interpreted in two fundamentally different ways. It depends on whether the goal can be inferred from the action alone, or if we need any context.

In the first interpretation, the goal is interpreted as an action that is higher in the action-hierarchy. The goal can be described as a collection of actions. For example, "to grasp" is the goal of a "whole hand grasp" and a "precision grasp". Since a goal is *less* specific than an action, the action is sufficient to infer the goal.

In the second interpretation, an action can serve different goals, depending on the context of the situation. For example, you can use the "whole hand grasp" to both move and eat an apple. Action and goal are two independent *affordances* of the stimulus. The term affordance is used here to describe how suited an object is for a specific action (i.e. grasping), but also to describe how suited an object is to reach a goal (i.e. eating). In this interpretation, goal inference without additional context is only possible if the action is used exclusively for that single goal.

My interpretation of goals and action is a very specific one, one that is not common in Mirror Neuron literature. In essence, I argue that goal inference and action execution are responses to a stimuli. This stimuli contains information about the action (to be executed) and the goal (that is being pursued). So, I make no distinction in the perception of goal-clues and action-clues.

Therefore, It is misleading to describe these goals as a "*desired world state*" or "*intention*". It would be more accurate to describe goals in terms of their observable clues. It is not that a different intention triggers a different mirror neurons, it is a different observation. And while observations and intentions might be related, they are not always the same. Therefore, it would be more accurate to describe a Mirror Neuron as responding to "grasping eatable objects" rather than "grasping in order to eat". It emphasizes the observable context (eatable object), rather than the intention (to eat). While unconventional, most MNS experiments ignore the middle man of what is observed and conclude Mirror Neurons fire based on intention.

I used the second interpretation in my experiments.

Since an action can serve different goals, goal inference is an additional classification task. Action and goal are two separate affordances of the stimuli that need to be classified. Some models define the goal as a predetermined, desired world state, and finish when they reach it. Other models never finish and go on indefinitely. I used the latter, in which the goal is more a direction than a destination.

The figure below shows an interpretation of the simulated world. Every combination of action and goal is possible, so there are six different stimuli.

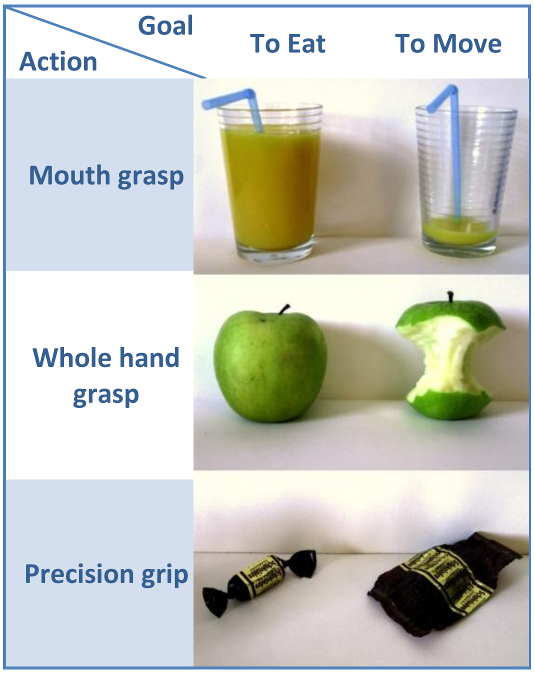

Table 5-1. An interpretation of 3 actions and 2 goals. Every combination is possible, so there are six different stimuli. Action and goal are two affordances of the stimuli. The kind of object determines the action, and the context determines the goal.

We can derive three categories of Mirror Units from these stimuli:

1. Goal-specific: Units that respond to a single goal, no matter which action.
2. Action-specific: Units that respond to a single action, no matter which goal.
3. Action+Goal-specific: Units that respond to a single stimulus.

5.2 Encoding the Stimulus-Response

Next, we need to translate these six stimuli into a binary activation of the visual layer. The stimuli has five affordances: two goals and three actions. Note that my interpretation of goals and actions are somewhat unconventional, as I argue that both goals and actions can be reduced to observable clues, when modeling the MNS with a stimulus-response mechanism. These affordances are:

1. Eatable
2. Moveable.
3. Mouth graspable
4. Whole hand graspable
5. Precision grip-able

These five affordances are encoded using 10 units each. These 10 units are split up into two modalities of 5 units. The modalities can be interpreted as vision and proprioception data. I will assume an 1:1 mapping between vision and proprioception. In other words, actions that look the same, feel the same.

The response consists of the 5 units – two encode a goal, and three encode an action.

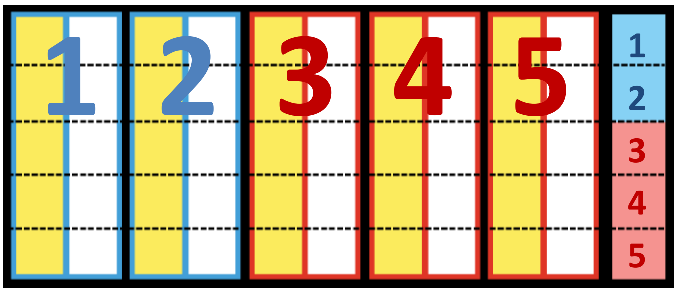

Figure 5-1. The visual layer. There are five affordances that consist of 10 units each. Two for the goals (blue), and three for the actions (red). Yellow indicates the visual modality, white indicates proprioception. The response is located on the right. The top two units encode the goal, while the bottom three encode the action.

When encoding the stimuli, we can safely ignore the shape of the activation pattern. This is because there are no connections between units of the visual layer, and the model does not consider the location of the unit in any way. Therefore, we can look at each unit individually. Each unit can be activated by one or more affordances. For example, a unit could respond to both a goal- and an action affordance. In that case, we would need six groups to cover every stimulus. Instead, I used the most basic encoding, in which each unit is activated by a single affordance.

I used a normal distribution to randomly activate the 5 units of a one modality. I scaled the normal distribution such that there are 3 units active on average. I used two different distributions to distinguish self-perception from perception of another individual. This is shown in the figure below:

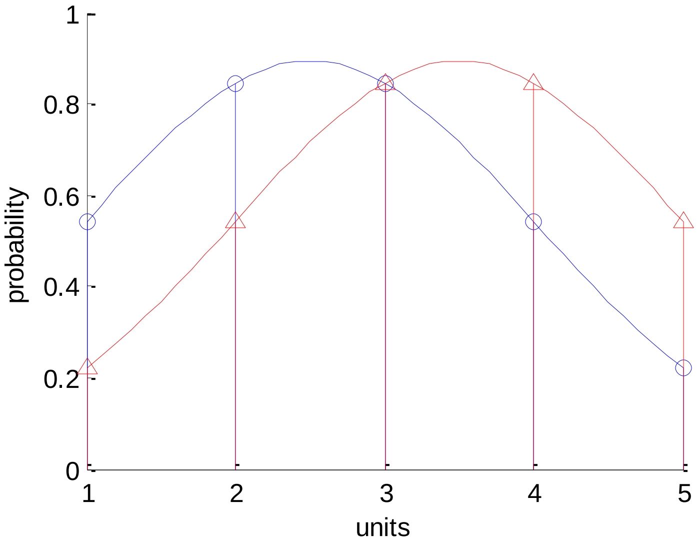

Figure 5-2. Stimuli are generated by activating a modality according to a normal distribution. The blue line represents self-perception, while the red line represents perception of another individual.

Since the model is trained only on self-perception, it needs to generalize in order to perform action observation or imitation. On top of that, perception of another individual lacks proprioception. This requires the model to deal with missing values.

5.3 Generating datasets

I generated several datasets to test behavior and performance of the model. These are summarized in the table and figure below. Starting from the default case of action *execution* (1), we modify the stimuli in various ways. We create the *imitation* dataset (2) by using a different distribution (*Other*, 3) and deleting proprioception (*Missing*, 4). Then, we also delete the goal-affordances (*No goal*, 5) to see if goal-specific Mirror Units stop responding. Furthermore, we combine execution with imitation (*Both*, 6) to see if execution performance degrades when also observing other actions. Finally, the *mixed* (7) dataset contains 50% cases from execution, and 50% cases from imitation. The mixed dataset is used to test the model as a whole.

	MNS task	Dataset name	Stimulus		Modality	
			Self	Other	Vision	Proprioception
1.	**Action Execution:**	Execution:	+	-	+	+
2.	**Imitation:**	Imitation:	-	+	+	-
3.	*- Other stimuli:*	*Other:*	-	+	+	+
4.	*- Missing proprioception:*	*Missing:*	+	-	+	-
5.	**Imitation without goal context:**	No Goal:	-	+	+	-
6.	**Execution while also observing:**	Both:	+	+	+	+/-
7.	**50% Execution, 50% Imitation**	Mixed:				

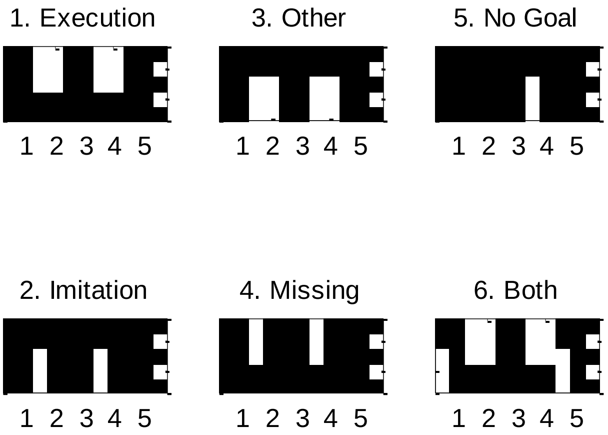

Figure 5-3. A typical stimulus and response from different datasets.

5.5 Hypothesis

The different datasets will be used to test the different hypothesis:

A Restricted Boltzmann Machine can successfully function as a Mirror Neuron System

Action Execution & Observation

The RBM will perform good (>95%) on selecting the correct goal and action, when executing actions or imitating. Imitation will perform worse than execution, because the network needs to generalize and deal with missing values. Generalizing will be harder than dealing with missing values, because vision and proprioception have an 1:1 mapping.

Action Execution while Observing

Slower reaction time and increased errors are reflected as less certainty and lower performance compared to normal action execution.

There is an optimal size and sparsity for Mirror Units

Training causes Mirror Units

I hypothesize that the contrast of a trained network is higher than of a random initialized network.

Small networks have more Mirror Units

A small network has a short description length. A single unit encodes a more prototype stimulus compared to longer description lengths. Units in long description lengths contain only a partial description of a stimulus. This makes it harder to find Mirror Units, because the unit needs to describe a distinctive feature for all cases of a single stimuli, action or goal.

Each category of Mirror Units has an optimal sparsity.

Sparsity encourages units to respond to fewer cases. This allows units to respond to a single stimulus, action or goal. However, when a unit responds to too few cases, it does not respond to *all* cases of that single stimuli, action or goal anymore. Responding to “too few cases” happens sooner for a single goal (50% of all cases), than for a single action (33% of all cases) than for single stimuli (17% of all cases). Therefore, the optimal sparsity-target varies between categories.

The Strength of the response changes with different stimuli

Goal-specific Mirror Units represent a distinctive goal and are unspecific for an action. These units will not be activated if the stimulus has no goal and only a specific action. The strength of the goal-specific Mirror Unit is less when observing actions without a goal, compared to observing actions with a goal.

Also, the Strength of the Mirror-Unit response will be lower for imitation than for execution, because the network is trained on action execution.

5.4 Training

I order to investigate the influence of size and sparsity on the model, I trained 171 models with various sizes and sparsity-targets. I used the same parameters on all models:

Parameter	Value
Visual units	55 units
Hidden units	From 20 to 200 units, with increments of 10 units (19 different values).
Learning rate (ε)	0,02
Momentum	0,9; but 0,5 in the first 5 epochs.
Sparsity-Cost	0,1
Sparsity-Target	From 0,1 to 0,5; with increments 0,05 (9 different values).
Sparsity-Decay (λ)	0,95
Batch-size	10
Dataset-size	100
Epochs	Depends on the stopping criteria, but max 200 epochs.

In order to make comparisons between models as accurate as possible, I tried to control confounding variables with a stopping criteria. Training stops when these four criteria are satisfied:

1. The sparsity-target has been reached with 1% accuracy.
2. Action execution selects more than 98% of the time the correct action and goal.
3. The Network-Strength is more than 80%.
4. The total sum-of-squared reconstruction error of one epoch is less than 1000.

However, training will always stop after 200 epochs, even when some criteria are left unsatisfied. It turns out that low sparsity-targets and small networks have trouble reducing the error (see “a” in figure below). High sparsity-targets and large sizes have trouble reaching enough strength (see “d” in figure below). This is why the strength criterion is quite low at 80%. Since some networks are just unable to reach enough strength, demanding higher strength would only increase epochs for nothing. The sparsity and performance targets are easily reached in all models. This is shown in the figure below:

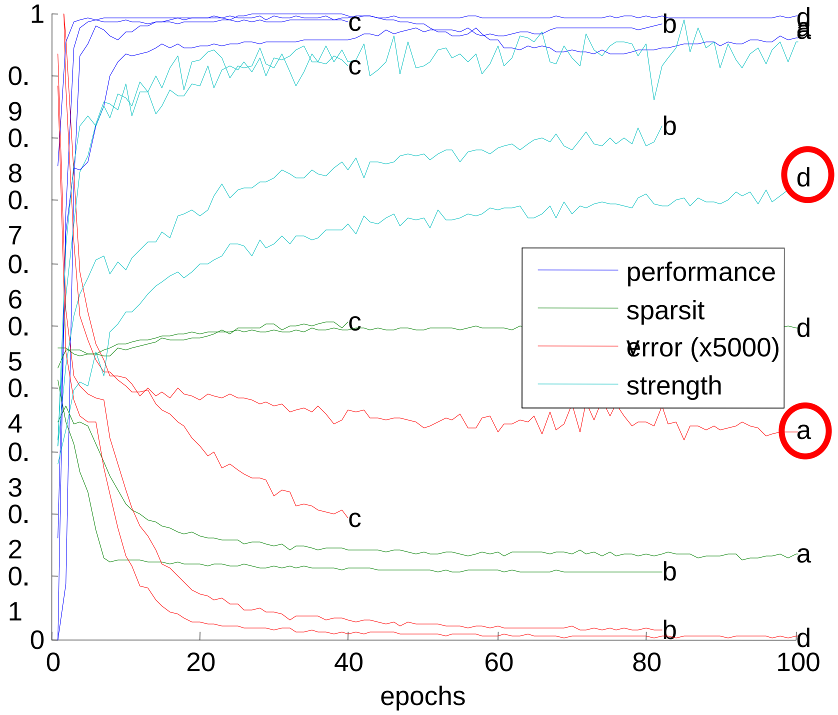

Figure 5-4.Stopping criteria for different models. Sparsity (green line) and performance (dark blue line) criteria are easily satisfied. Attaining enough strength (cyan line) is difficult for “d”, while reducing the error (red line) is difficult for “a”.
a) Size: 20; Sparsity-target: 0,10.
b) Size: 20; Sparsity-target: 0,50.
c) Size: 200; Sparsity-target: 0,10.
d) Size: 200; Sparsity-target: 0,50.

I used a control condition to investigate whether training causes the Mirror Units to emerge. In this control condition, the 171 models were not trained, but initialized with random weights. These random weights were set using a normal distribution typical for the trained models ($\mu = 0;\ \sigma = 1$).

6. Results

6.1 Performance of action execution & imitation

Performance is measured as the percentage of correct responses.

After the model generates the response, each unit is rounded towards either 0 or 1. If this matches the target response, the response is considered correct. Performance is averaged over every case from the dataset, and over every network.

Significance and confidence intervals are calculated using a Repeated-Measures ANOVA, which corrects for confounding variables such as number of epochs and strength.

All performances are good (>95%), except for action execution while observing (61%) and the control condition (3,4%). Imitation performance is lower than action execution performance. Note that contrary to what I predicted, "missing" has a lower performance than "other".

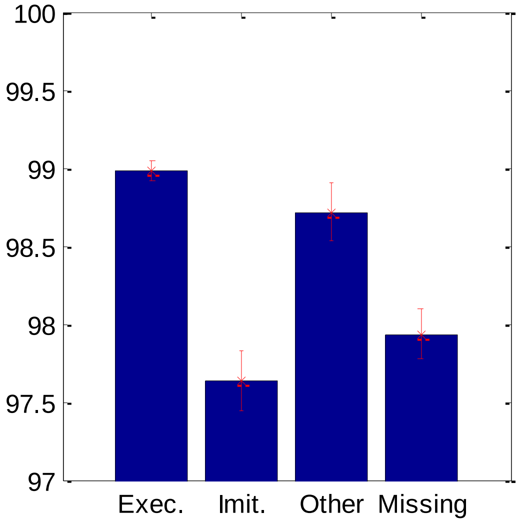

Figure 6-1. Percentage of correct responses for different datasets.
The red lines indicate the 95% confidence interval. Not every dataset is shown.

Dataset	Percentage Correct	Std. Error	95% Confidence interval	
Action Execution	98,978	0,033	98,914	99,043
Imitation	97,635	0,098	97,441	97,828
- Other Stimuli	98,713	0,093	98,528	98,897
- Missing Proprioception	97,923	0,081	97,773	98,091
Both Execution & Observation	60,752	0,742	59,287	62,218
Control Condition (Mixed)	3,551	0,313	2,934	4,169

Table 6-1.p < 0,000.

6.2 Certainty of action execution & imitation

Certainty is measured based on how much a response is rounded towards either zero or one.

$$Certainty = 1 - \frac{1}{N}\sum_{i=1}^{N} abs(x_i - round(x_i))$$

Where x is the generated response, and i indexes the units in the response. In this case, there are 5 (N) units. The certainty is averaged over every case in the dataset, and over every model.

The hypothesis predicted less certainty for action execution while observing. It also predicts the models would be little less certain imitating than executing actions. In the figure below, you see the average certainty of all models. Again, a Repeated-Measures ANOVA is used to determine confidence intervals and significance of the results.

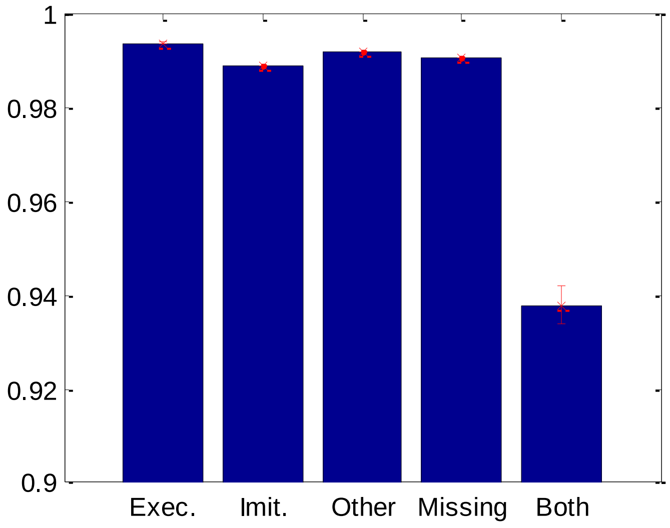

Figure 6-2. Certainty of performance on different datasets.
The red lines indicate the 95% confidence interval.

Dataset	Certainty	Std. Error	95% Confidence interval	
Action Execution	0,994	0,000	0,994	0,995
Imitation	0,989	0,000	0,989	0,989
- Other Stimuli	0,992	0,000	0,992	0,992
- Missing Proprioception	0,991	0,000	0,991	0,992
Both Execution & Observation	0,938	0,002	0,935	0,941
Control (Mixed)	0,909	0,003	0,904	0,915

Table 6-2. p < 0,000.

6.3 Network-Contrast

We need a measure to investigate how size and sparsity influence the occurrence of Mirror Units. First, we need measure how Mirror-Unit-like a single unit is: Unit-Contrast.

$$UnitContrast(unit) = p(unit = 1 \mid category = x) - p(unit = 1 \mid category \neq x)$$

Where $p(\cdot)$ is calculated over the entire dataset, $category \in \{stimuli, actions, goals\}$ and $x \in category$.

This captures the requirement to respond both strong and selectively to a single stimulus, action or goal. We will now extend this measure to an entire network. Since a few Mirror Units in a network is enough, we cannot simply average Unit-Contrast over all units. So, I have plotted the contrast of all units, sorted from high to low, for networks of various sizes and sparsities:

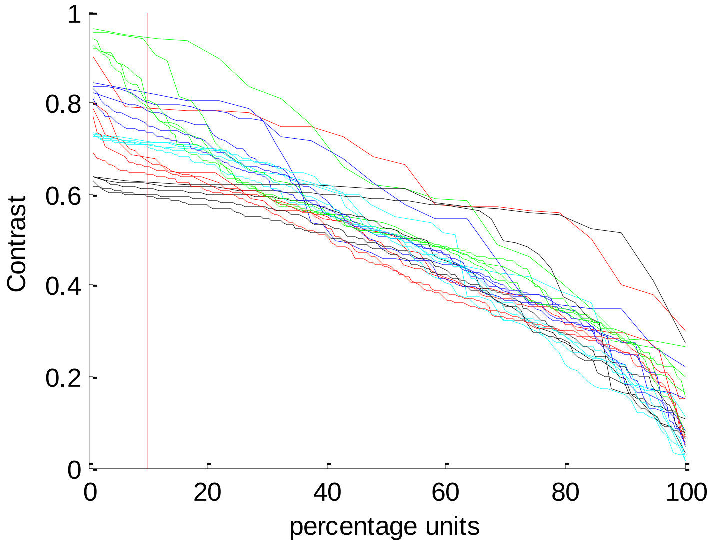

Figure 6-3. Goal-Contrast per unit, sorted from high to low, for networks of various sizes and sparsities.
***Sparsity**: 0,1 (red) – 0,2 (green) – 0.3 (blue) – 0.4 (cyan) – 0.5 (black)*
***Sizes**: 20 – 50 – 100 – 150 – 200 (same color)*

We could count the number of units that exceed a certain threshold as “Mirror Units”. This would be a very imprecise measure for small networks since it is limited to whole numbers. Also, there is no sudden drop in Unit-Contrast which justifies a specific threshold. Instead, I have defined Network-Contrast as the mean Unit-Contrast of the best 10%.

$$NetworkContrast = \frac{1}{N}\sum_{i=1}^{N} UnitContrast(unit_i)\text{ , where } i \in \{10\% \text{ best units}\}$$

This gives a more precise measurement. It will not be misleading because the contrast decreases in a linear fashion. I used a percentage instead of a fixed number, because I consider a large network with two Mirror-Units worse than a small network with two Mirror-Units.

Despite the attempt to fix the average Strength during training, there is still much difference:

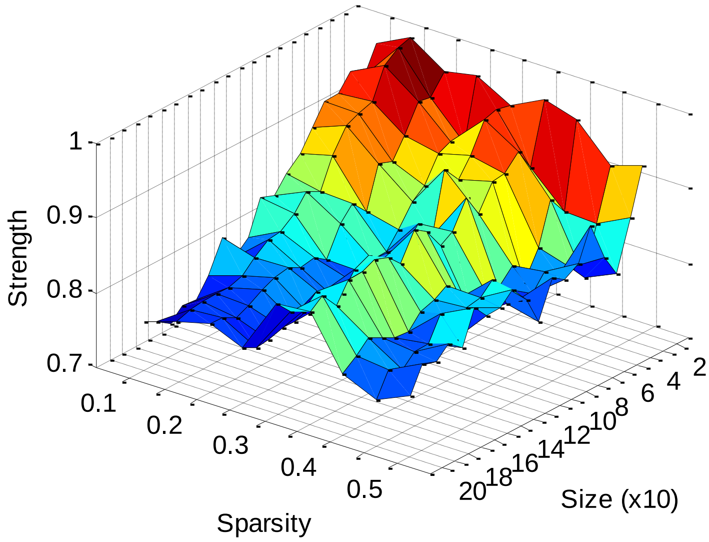

Figure 6-4. Strength for each network.

Since contrast takes both strength and selectivity into account, stronger networks are biased to get a higher Network-Contrast. To remove this bias, I have defined the *adjusted Network-Contrast* as:

$$NetworkContrast * \frac{1}{N}\sum_{i=1}^{N} p(unit_i = 1)\text{, for every } i \text{ whose } p(unit_i = 1) > 0{,}5$$

Basically, I multiplied the Network-Contrast with the average “on” probability. However, unlike the strength measure, I did not subtract the average “off” probability.

6.4 Effect of Training on Network-Contrast

To test whether training causes Mirror Units, I have compared the contrast of trained networks with untrained networks. I used the mixed dataset, so the Network-Contrast is based on both action execution as well as observation. Note these results are averaged over all models, including models which may have a far from ideal size or sparsity.

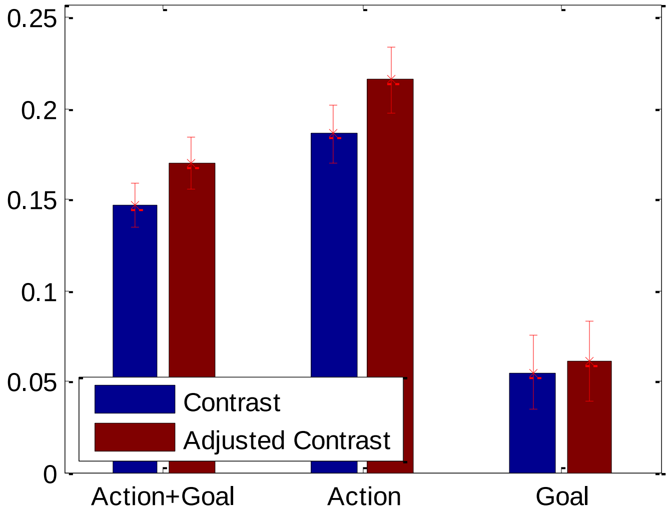

Figure 6-5. Increase in Network-Contrast of trained networks compared to the control condition.

Category		Increase	Std. error	95% Confidence Interval	
Action+Goal	Contrast:	+0,147	0,006	0,135	0,158
	Adjusted Contrast:	+0,170	0,007	0,157	0,182
Action	Contrast:	+0,186	0,008	0,170	0,201
	Adjusted Contrast:	+0,216	0,009	0,198	0,234
Goals	Contrast:	+0,055	0,010	0,036	0,075
	Adjusted Contrast:	+0,061	0,011	0,039	0,083

Table 6-3. Increase in Network-Contrast of trained networks compared to the control condition. p < 0,000.

6.5 Unit-Contrast

Training causes an increase in Mirror Units. However, the models are only trained on action execution, not on imitation. Units that respond strong and selective on execution might not respond likewise on imitation. To quantify this effect, I calculated the correlation of Unit-Contrast between “Execution” and other datasets:

Category	Imitation	Other	Missing
Action+Goal	0,7286	0,7392	0,9162
Action	0,8602	0,8712	0,9594
Goal	0,8531	0,8557	0,9651

Table 6-4. Correlation of unit-contrast between the “Execution” dataset and other datasets.

Note how the correlation with “Missing” is *much* higher than with the “Other” dataset.

6.6 Effect of Size & Sparsity on Network-Contrast

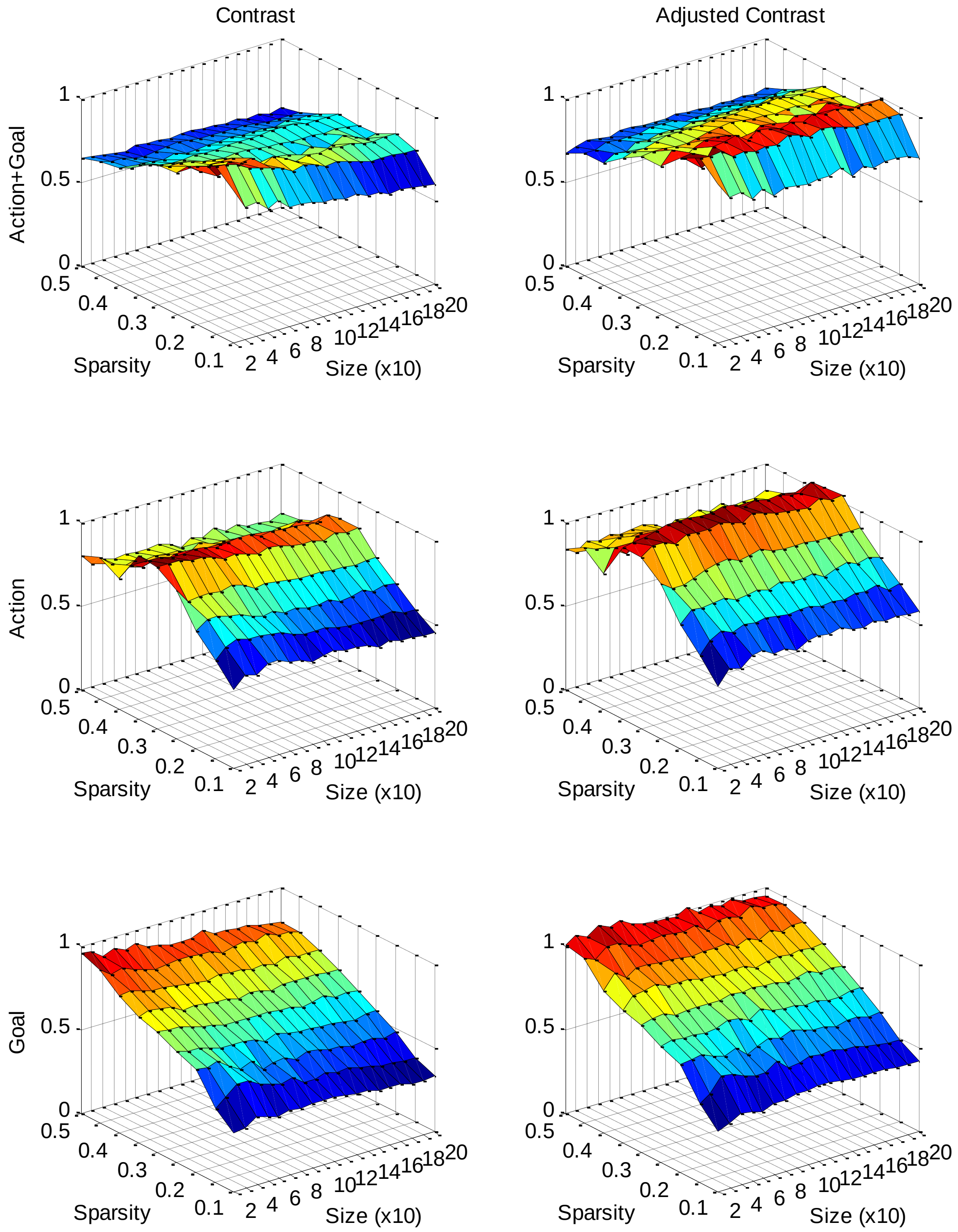

Figure 6-6. Network-Contrast and the Adjusted Network-Contrast for each category and every trained network.

Category	Sparsity	Size	Contrast
Action+Goal	0,15	20	0,9969
Action	0,35	20	0,9712
Goal	0,50	20	0,9529

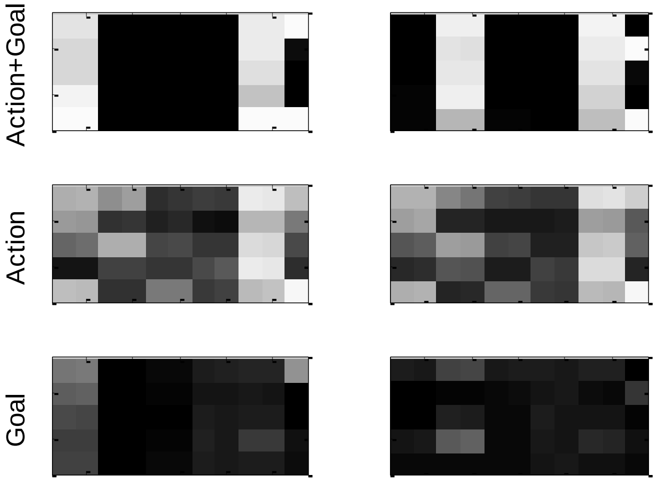

Figure 6-7. The best Mirror Units for each category.
Note how "action" tries to respond to every goal and a single action, and how "goal" tries to respond to every action and a single goal.

The optimal sparsity seems to be $1/N$, where N is the number of stimuli (6), actions (3) or goals (2). This is shown in the figure below:

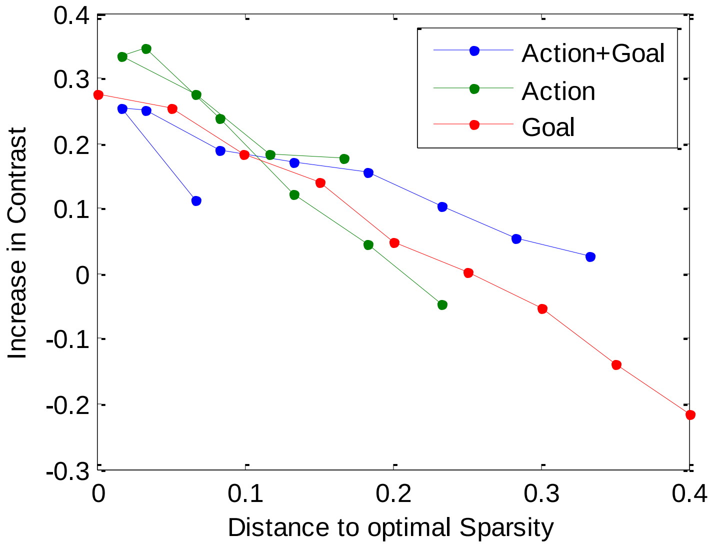

Figure 6-8. Increase in contrast compared to the control condition. The distance to the optimal sparsity is shown on the x-axis. The optimal sparsity is 0,16 for Action+Goal, 0,33 for Action and 0,50 for Goal.

Note that when we take sparsity into account, there is little difference between categories left. There seems to be a linear relationship. To test this, I performed a MANCOVA for each category. Distance and Size are covariates. Network-Contrast and adjusted Network-Contrast are the dependent variables. The results are shown below:

Model: Intercept + α_1*Distance + α_2 * Size

Parameters		Intercept	Distance (α_1)	Size (α_2)
Action+Goal	Contrast:	0,896	-0,633	0,00
	Adjusted Contrast:	0,986	-0,794	0,00(1)
Action	Contrast:	0,990	-1,760	0,00
	Adjusted Contrast:	1,105	-2,140	0,00(2)
Goals	Contrast:	0,953	-1,264	0,00
	Adjusted Contrast:	1,054	-1,449	0,00

Table 6-5. Parameters of the MANCOVA model.

Partial Eta^2		Intercept	Distance	Size	Model Total (Adjusted R^2)
Action+Goal	Contrast:	0,981	0,672	0,334	0,715
	Adjusted Contrast:	0,980	0,722	0,000(1)	0,719
Action	Contrast:	0,988	0,907	0,374	0,911
	Adjusted Contrast:	0,990	0,931	0,000(2)	0,930
Goals	Contrast:	0,996	0,980	0,634	0,980
	Adjusted Contrast:	0,994	0,973	0,048	0,973

Table 6-6.Proportion explained variance by the MANCOVA. p<0,001, except for 1 (0,787) and 2 (0,929).

Note how size has no influence, and is even insignificant for the adjusted Network-Contrast.

6.7 Strength of the Mirror Unit response

I calculated the mean Strength of the 10% best goal-specific Mirror Unit. I averaged over all networks.

The hypothesis are confirmed: Goal-specific Mirror Units respond much weaker to the “No goal” dataset than to action execution. The response is even weaker compared to the control condition. Also, as predicted, the response on imitation is slightly weaker compared to action execution.

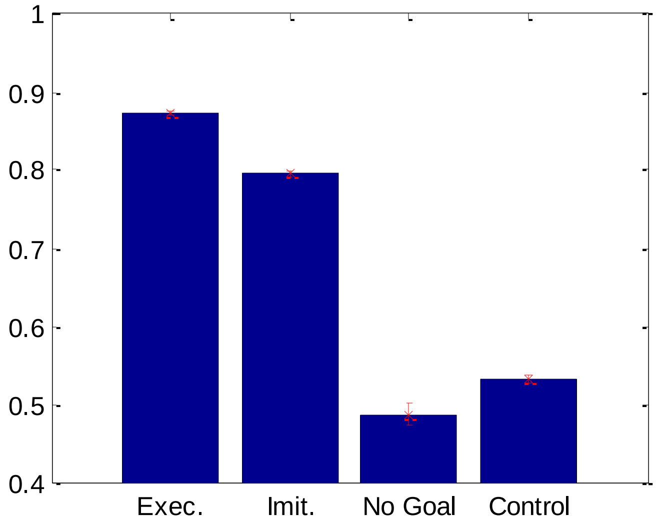

Figure 6-9. Mean Strength of the 10% best goal-specific Mirror Units, averaged over all models.

Dataset	Strength	Std. Error	95 % Confidence interval	
Action Execution	0,873	0,001	0,870	0,875
Imitation	0,795	0,002	0,791	0,799
No Goal	0,487	0,007	0,473	0,501
Control (No Goal)	0,531	0,003	0,526	0,536

Table 6-7.p < 0,01.

7. Conclusion & Discussion

The results confirm the hypothesis and show that indeed, a Restricted Boltzmann Machine can successfully emulate MNS behavior. Also, units with Mirror-Neuron like properties emerge in the hidden layer. In the following section, I will discuss these results.

Restricted Boltzmann Machines as Mirror Neuron System

There are three ways in which the RBM emulates the MNS:

First, the performance of the model is quite good (>97%). Indeed, imitation performs worse (-1,343%) than execution because the model needs to generalize to unseen data. Surprisingly, the model had more difficulty dealing with missing proprioception than with generalizing. This was unexpected since the modalities are identical to each other. This difference cannot be explained by less net input from the missing dataset, since both datasets had very certain responses. Perhaps generalization was too easy. The limits of generalization capabilities could be explored in further research. Also, the model has a good performance after only a few epochs – which suggests that these networks have potential to learn more and complex stimuli.

Second, the certainty is less when performing action execution while observing. The certainty only drops a little (-5,6%) while the performance becomes unacceptable (60,75%). While this validates that observation interferes with execution, experiments did not show such drastic decrease in performance. This difference might be explained by lack of attention regulation in the model. Attention regulation could be simulated by attenuating the activation caused by observing. This can be explored in further research. As expected, certainty is a little less (-0,5%) on imitation compared to execution. This mimics the slightly reduced response found by Gallese et al (1996).

Third, goal-specific Mirror Units responded with less strength (-38,6%) when the goal-affordance was absent in the stimuli. This effect is learned, because the response was even less compared to the control condition (-4,6%).

Note that I interpreted goals as “observable context” and not as “collection of actions”. Essentially, there is no difference between goals and actions, because they both are affordances of the stimuli. This means a similar reduction in strength will be found with Action-specific Mirror Units when we only provide the goal affordances.

We can also use the other interpretation in my experiments. Goal-specific Mirror Units represent a collection of three actions, since Goal-Specific Mirror-Units do not *ignore* the action-affordances, but respond to all of them (within the context of a goal-affordance). We could also look for Action-specific Mirror-Units that respond to two actions and both goals. However, with only three actions, the action hierarchy is too small to make any sensible claims.

Mirror Units

Researchers have found two fundamentally different neurons in the MNS: canonical neurons and Mirror Neurons. Thus, Units should score either high or low on Unit-Contrast, and it should be easy to set a threshold to distinguish the two populations. However, the experiment produced a different result.

When sorting units from high to low contrast, it showed a slow monotonic decrease instead of a sudden steep descent. This suggests that association learning from self-observation is not sufficient to explain the occurrence of Mirror Neurons in the brain. Unlike learning algorithms designed for classification, Contrastive Divergence does not propagate errors from the response back to the stimuli. However, it is plausible such reinforcement learning happens in the brain. This might cause the distinct functionality. Fortunately, RBMs can combine Contrastive Divergence learning with ordinary classification learning such as backpropagation. Further research should investigate if this provides more accurate emergence of Mirror Units.

Nevertheless, it is certain that training causes Mirror Units to emerge in the experiments. The average Network-Contrast over all models has increased for all categories. This average includes models which have far from the optimal sparsity. Goal-specific Mirror-Units have more models which are further from optimal sparsity than Action-specific Mirror-Units do. This fully explains the difference between categories. Indeed, Figure 6-8 shows that increase in contrast is similar between categories if you take distance to optimal sparsity into account.

This is interesting, because Stimuli-specific Mirror Units are fundamentally different from a Action- or Goal-specific Mirror Units. There are three differences.

First, the Stimuli-specific Mirror-Units specify the entire visual layer, while the Action- and Goal-specific Mirror-Units need to be highly specific on certain parts, while remaining highly unspecific on other parts. This can be seen in the figure below. I generated the visual layer from a single Mirror-Unit in the hidden layer. Stimuli-specific Mirror-Units are specific everywhere. Action-specific Mirror-Units try to define a specific action, while responding to every goal. Goal-specific Mirror-Units specify a single goal, while covering as many actions as they can.

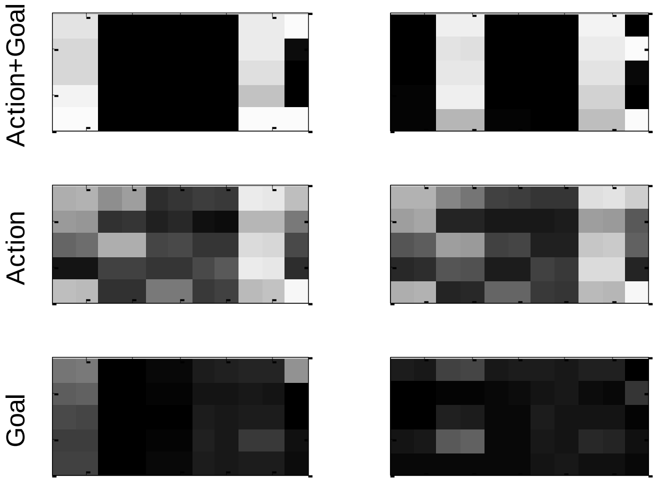

Figure 7-1. Generation of the visual layer when activating the best Mirror-Units from all models. The size of every model was 20, and the sparsity was 0,15 (Action+Goal), 0,35 (Action) and 0,50 (Goal).

Second, Stimuli-specific Mirror-Units combine multiple sources of variation (i.e. action and goal), whereas Action or Goal-specific units represent only once source of variation.

The effects of these two differences are indistinguishable in my experiments, because units only respond to a single affordance. Covering more sources of variation equals covering more units. We can only distinguish these effects when a unit responds to multiple sources of variation (i.e. it encodes both an action and a goal affordance). This can be a subject for further research.

Third, generalization is much more difficult for Stimuli-specific Mirror Units, because they are more specific. This is shown using the correlation Unit-Contrast between execution and observation. Stimuli-specific Mirror Units have a correlation of 0,7286, while Action- and Goal-specific Mirror Units have a correlation of 0,8602 and 0,8531, respectively. One would expect this to decrease Network-Contrast score – but it does not.

In spite of these three differences, the average increase in contrast it similar between different categories, if we take the distance to optimal sparsity into account. The ability of representing only parts of the stimuli, represent one or multiple causes, and robust generalization make association learning a good method for creating powerful Mirror Units. However, their emergence does not match the natural occurrence in the brain. This may be alleviated when we mix association learning with reinforcement or classification learning.

Optimal Sparsity

Results showed that the optimal sparsity is $1/N$, where N is the number distinctions within a category. This is the result of both learning and the sparsity penalty.

The sparsity penalty reduces the number of cases a unit responds to. As mentioned earlier, this number needs to be enough to cover all instances, but not so much that it includes irrelevant cases. For example, an Action-specific Mirror-Unit needs to respond to 1/3 of all cases, because there are three actions that are equally distributed in the dataset. Learning ensures these few cases fit to one and the same action.

In my experiments, the structure of the data (i.e. three different actions) matches the occurrence in the dataset (i.e. each action is 1/3 of the dataset). This allows learning and the sparsity penalty to work in synergy. Indeed, machine learning techniques often balance the dataset to improve results. Nevertheless, real-life observations might not be balanced, so results in real-life applications may vary. Of course, there are solutions to balance real-life observations, such as using some kind of attention regulation mechanism to weigh observations.

Since the sparsity penalty opposes learning, the sparsity-target cannot be set too low. This limits the number of distinctions, and therefore the action-repertoire of Mirror-Units. Alternative learning algorithm exists which allow for lower sparsity-targets, such as Sparse Encoding Symmetric Machines (Ranzato, Boureau, & LeCun). However, we can also limit the connections such that each hidden unit only encodes a small part of the visual layer. This way, each unit only makes a limited number of distinctions within its area of interest (i.e. only hand-actions); while the network as a whole can make many distinctions (i.e. actions from the entire body). This solution is biologically plausible as the neurons in the premotor cortex specialize in specific body regions. (source). In essence, the current network only models a few neurons from each brain region.

Size

Surprisingly, the size has almost no influence. It only affects the occurrence of Mirror Units because larger networks tend to have weaker strength . This goes hand in hand with the description length: Large description lengths allow weaker activation of each individual unit. As mentioned earlier, such cases are described as the combination of many subtle filters; whereas short description lengths describe a case by a few strong prototypes.

Summary

RBMs have been proven successful in emulating various aspects of MNS behavior. This includes action execution, observation, imitation, goal inference, dealing with missing values (i.e. in the dark) and handling multiple modalities. The performance, strength and certainty of responses of the model in different circumstances is similar to data from experiments. This provides evidence that the MNS can be understood as a belief network that functions through association. While training causes Mirror Units to emerge, associative learning is insufficient to explain the occurrence of Mirror Units, since it does not cause two distinct populations.

Nevertheless, the Mirror-Units that emerge are quite powerful. Their emergence is robust to network size. They can represent one or more underlying causes (i.e. an action, a goal or both). They can respond to a region of choice in the visual layer (e.g. specific for the entire visual layer, or specific for only a single affordance).

In fact, the only factor that influences Mirror Units is the sparsity-target. The optimal sparsity-target for Mirror-Units depends on the number of distinctions a unit needs to make, which in turn depends on the structure of the observed data.

Therefore, research should not only consider how the model works (through association) or how the model learns (Bayesian), but also how units are encoded (sparse) – since this turns out to be essential for Mirror Neurons.

Further research

I already mentioned various directions for future research. I will briefly summarize the most important below:

1. When we add more actions to the model: How are Mirror-Units distributed across the levels of the action-hierarchy?
2. Will the addition of classification learning algorithms make the emergence or Mirror-Units better resemble the natural occurrence of Mirror Neurons?
3. What is the effect of stimuli encoding on the occurrence of Mirror-Units? Will this change if a single unit encodes multiple affordances?
4. Can we use local connectivity between the visual and hidden layer to extend the model with more actions in a biologically plausible way?

However, the worst limitation is the inability to model time-series. Fortunately, RBMs can be extended in various ways to model time-series: Recurrent Temporal RBMs, Temporal Convolution Machines and Conditional RBMs are all viable options. These also support real-valued units, which can be used to model actual movement trajectories.

Also, I assumed an 1:1 mapping between proprioception and visual data. Changing this mapping can explain more behavior of the MNS. For example, Gallese et al (1996) reported that Mirror Neurons respond to a more broad range of actions on observation than on execution, while this was not the case in my experiments.

Finally, the “missing” dataset can be interpreted as the experiments in the dark room. Here, a small percentage of the original visual Mirror Neurons also fired when the monkey only heard the action. Since *counting* Mirror Units was an imprecise measure, I did not investigate this effect in my experiments. However, this could be subject for future research, especially if we manage to set a decent threshold for Mirror Units, and if there is no 1:1 mapping between sound and vision.

References

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural Circuits Involved in the Recognition of Actions Performed by Non-Conspecifics: An fMRI Study. *Journal of Cognitive Neuroscience , 16* (1), 114-126.

Chaminade, T., Oztop, E., Cheng, G., & Kawata, M. (2008). From self-observation to imitation: Visuomotor association on a robotic hand. *Brain Research Bulletin* (75), 775-784.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks* (16), 1325-1352.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain* (119), 593-609.

Heyes, C. (2010). Where do Mirror Neurons come from? *Neuroscience and Biobehavioral reviews* (34), 575-583.

Hinton, G. (2010, August 2). *A Practical Guide to Training Restricted Boltzmann Machines.* Retrieved August 20, 2011, from Geoffrey Hinton: http://www.cs.toronto.edu/~hinton/absps/guideTR.pdf

Hinton, G., & Salakhutdinov, R. (2006, July 28). Reducing the dimensionality of data with neural networks. *Science , 313* (5786), pp. 504-507.

Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A fast learnign algorithm for deep belief nets. *Neural Computation* (18), 1527-1554.

Kilner, J. M., Friston, K. J., & Firth, C. D. (2007). Predictive Coding: an account of the mirror neuron system. *Cogn Process* (8), 159-166.

Knill, D. C., & Pouget, A. (2004). The Bayesian Brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences , 27* (12), 713-719.

Lennie, P. (2003). The Cost of Cortical Computation. *Current Biology , 13* (6), 493-497.

Ranzato, M. A., Boureau, Y., & LeCun, Y. (n.d.). Sparse Feature Learning for Deep Belief Networks.

Ranzato, M. A., Poultney, C., Chopra, S., & LeCunn, Y. (2006). Efficient Learning of Sparse Representations with an Energy-Based Model. *Advances in Neural Information Processing Systems (NIPS 2006).* MIT Press.

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience* (27), 169-192.

Sutskever, I., Hinton, G., & Taylor, G. (2009). The Recurrent Temporal Boltzmann Machine. *Advances in Neural Information* .

Taylor, G. W., & Hinton, G. E. (2009). Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style. *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, (pp. 1025-1032).

Umilta, M. A. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Science , 105* (6), 2203-2213.

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing: A neurophysiological study. *Neuron , 31* (1), 155-165.

Vilares, I., & Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Acadamy of Sciences , 2011*, 23-39.

Willmore, B. D., Mazer, J. A., & Gallant, J. L. (2011). Sparse coding in striate and extrastriate visual cortex. *J Neurophysiol* (105), 2907-2919.

Willmore, B., & Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network: Computation in naural systems* (12), 255-270.

Note for supervisors:

A self-reflection on my thesis. These are some points that could be improved, in my opinion:

1. The abstract only covers the hypothesis and conclusions, not the methods.

2. Regarding claims about empathy, imitation and action understanding (page 2): I only refer to Heyes (2010) instead of specific studies.

3. Comparison to other MNS models can be more precise (page 2), as I now just refer to Oztop et al (2006)

4. I do not refer to any experiments which mention action execution while observing. I recall that Oztop et al (2006) mentions that observation interferes with execution, but I am not sure. Also, I recall this evidence is based on fMRI experiments; in which the reaction time was slower when observing a different action than on execution. Could this be the study with hand images and numbers?

5. Chapter 2 lacks a short conclusion/summary, and is quite incoherent in general. Also, the links between different parts of the thesis can be improved: especially introduction (1), the MNS (2) and the conclusion (7).

6. The sparsity chapter is supposed to introduce the reader to description length, and how it influences the representation. This is never done, while the hypothesis (page 17) *does* mention description length. This might confuse the reader.

7. The conclusion is too long – although I do make conclusions, they do not stand out, but rather get lost in the text. To compensate for this, I repeat various conclusions multiple times – but this is repetitive and long winding.

8. The conclusion should preferably say more about the MNS in general – it seems to stay stuck in my model. I tried to salvage this by writing: “*Therefore, research should not only consider how the model works (through association) or how the model learns (Bayesian), but also how units are encoded (sparse) – since this turns out to be essential for Mirror Neurons.*”