Demystifying Sparse Rectified Auto-Encoders

Kien Tran
Department of Computer Science
Faculty of Information Technology
Vietnam University of Science - HCM
ttkien@fit.hcmus.edu.vn

Bac Le
Department of Computer Science
Faculty of Information Technology
Vietnam University of Science - HCM
lhbac@fit.hcmus.edu.vn

ABSTRACT

Sparse Auto-Encoders can learn features similar to Sparse Coding, but the training can be done efficiently via the back-propagation algorithm as well as the features can be computed quickly for a new input. However, in practice, it is not easy to get Sparse Auto-Encoders working; there are two things that need investigating: sparsity constraint and weight constraint. In this paper, we try to understand the problem of training Sparse Auto-Encoders with L1-norm sparsity penalty, and propose a modified version of Stochastic Gradient Descent algorithm, called Sleep-Wake Stochastic Gradient Descent (SW-SGD), to solve this problem. Here, we focus on Sparse Auto-Encoders with rectified linear units in the hidden layer, called Sparse Rectified Auto-Encoders (SRAEs), because such units compute fast and can produce true sparsity (exact zeros). In addition, we propose a new reasonable way to constrain SRAEs' weights. Experiments on MNIST dataset show that the proposed weight constraint and SW-SGD help SRAEs successfully learn meaningful features that give excellent performance on classification task compared to other Auto-Encoder variants.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*connectionism and neural nets, concept learning, parameter learning*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*representation, data structures, and transforms*; I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement—*feature representation*

General Terms

Algorithms, Design, Experimentation

Keywords

unsupervised feature learning, deep learning, sparse coding, sparse auto-encoders, rectified linear units

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SoICT'13, December 05 - 06 2013, Danang, Viet Nam
Copyright 2013 ACM 978-1-4503-2454-0/13/12$15.00.
http://dx.doi.org/10.1145/2542050.2542065.

1. INTRODUCTION

Recently, unsupervised feature learning and deep learning have attracted a lot of interest from various fields such as computer vision, audio processing, text processing, and so on. The idea is that instead of designing features manually, one lets the learning algorithms automatically learn features from unlabeled data; and deep learning means learning multiple levels of features with increasing abstraction. Auto-Encoders (AEs) and Restricted Boltzmann Machines (RBMs) are two main groups of algorithms that have been used in unsupervised feature learning and deep learning [1]. AEs belong to the non-probabilistic group while RBMs belong to the probabilistic group. One big disadvantage of RBMs compared to AEs is that the objective function of RBMs is intractable. For this reason, here we will focus on the study of AEs.

Several criteria have been proposed to guide AEs to learn useful representation. They include: sparsity criterion [6], denoising criterion [14], and contraction criterion [13, 12]. Among them, sparsity is an interesting and promising one (here sparsity means forcing the majority of elements of the feature vector to be zeros). The first reason is that it has the inspiration from biology. In the brain, there is a very small fraction of neurons active simultaneously. Sparsity was first introduced in Sparse Coding and interestingly, it helped learn features similar to the primary visual cortex [11]. AEs with sparsity criterion, called Sparse Auto-Encoders (SAEs), can learn features much like Sparse Coding, but unlike Sparse Coding, the training can be done efficiently via the back-propagation algorithm, and with a new input, the features can be computed quickly. Secondly, sparsity can help learn high-level features - concepts. The intuitive justification is that there are only a few concepts per example; therefore, sparsity can help learn a dictionary of concepts and each example will be explained just by a small number of concepts. Thirdly, sparsity can potentially help speed up the training of SAEs. With each example, in the forward propagation phase, there is only a small fraction of neurons active; and hence, in the backward propagation phase, there is only a small fraction of parameters (corresponding to active neurons) updated. This point can be made use of to speed up the training process. It is important because if the training is fast, the model can be scaled up (i.e. increase the number of features); in unsupervised feature learning and deep learning, large-scale is a key factor to get good performance [4, 7].

Despite above advantages, it is not easy to get SAEs working in practice. To make SAEs work, there are two things

that need investigating: sparsity constraint and weight constraint. Although L1-norm is a natural (because it is used in Sparse Coding) and simple (in case the feature vector has positive values, it is just simply sum of them) way to constrain sparsity, it is not often used in SAEs for reasons that remain to be understood [1]. Instead of L1-norm, people often constrain sparsity in SAEs by pushing the average output of a hidden neuron (e.g. over a minibatch) to a fixed target (close to zero) [6, 4, 3]. But this fixed target adds one more hyper-parameter to the list of SAEs' hyper-parameters which already has many ones. As a result, the process of tuning hyper-parameters will become more tedious and more time-consuming. Regarding weight constraint, many different ways were used in the literature. [3, 14, 13, 12] tied the weights of encoder and decoder together. [6, 4] used weight decay; this way even adds one more hyper-parameter. [15] constrained the weights of decoder to have unit norm. However, it is not clear which way should be used as well as why weights should be constrained like those.

Two questions remain to be answered: (i) why is L1-norm sparsity penalty not often used in SAEs?; (ii) is there a better and more reasonable way to constrain SAEs' weights? In this paper, we try to understand the problem of training SAEs with L1-norm sparsity penalty. Then, we propose a modified version of Stochastic Gradient Descent algorithm (SGD), called Sleep-Wake Stochastic Gradient Descent (SW-SGD), to remedy this problem. Here we focus on SAEs with rectified linear units (ReLUs) in the hidden layer because such units compute fast and can produce true sparsity (exact zeros) [10, 5, 15]. We call these Sparse Rectified Auto-Encoders (SRAEs). Furthermore, we propose a new reasonable way to constrain SRAEs' weights. With these two ingredients, our proposed weight constraint and SW-SGD, our experiments show that SRAEs can successfully learn meaningful features that give excellent classification performance on MNIST dataset compared to other Auto-Encoder variants.

The rest of the paper is organized as follows. We start by reviewing Sparse Coding and Sparse Auto-Encoders (SAEs) to see advantages of SAEs compared to Sparse Coding. Then, Section 3 presents Sparse Rectified Auto-Encoders (SRAEs): Subsection 3.1 explains the problem of training SRAEs with L1-norm sparsity penalty and describes our remedy for this problem; Subsection 3.2 presents our proposed weight constraint for SRAEs. Experiment and analysis are shown in Section 4 followed by the conclusion in Section 5.

2. REVIEW OF SPARSE CODING AND SPARSE AUTO-ENCODERS

2.1 Sparse Coding

Sparse Coding was first introduced in neuroscience to model the primary visual cortex [11]. The goal is to find an over-complete set of basic vectors so that each input can be explained just by a small number of basis vectors (i.e. the feature vector is sparse). Specifically, given the unlabeled data $\{x^{(1)}, \dots, x^{(N)}\}$ with $x^{(n)} \in \mathbb{R}^D$, Sparse Coding solves the following optimization problem:

$$\begin{aligned} \underset{\phi, a}{\text{minimize}} \quad & \sum_{n=1}^{N} \left(\|x^{(n)} - \sum_{k=1}^{K} a_k^{(n)} \phi^{(k)}\|_2^2 + \lambda \|a^{(n)}\|_1 \right) \\ \text{subject to} \quad & \|\phi^{(k)}\|_2^2 = 1, \forall k = 1, \dots, K \end{aligned} \quad (1)$$

Here, the optimization variables are *the basis vectors* $\phi = \{\phi^{(1)}, \dots, \phi^{(K)}\}$ with each $\phi^{(k)} \in \mathbb{R}^D$, and *the coefficient vectors* (the feature vectors) $a = \{a^{(1)}, \dots, a^{(N)}\}$ with each $a^{(n)} \in \mathbb{R}^K$; $a_k^{(n)}$ is the coefficient of basic $\phi^{(k)}$ for input $x^{(n)}$. With this optimization problem, we want to learn a representation having the following properties:

- Preserving information about the input (by minimizing the reconstruction error).
- Being sparse (by minimizing the L1-norm of the feature vector).

λ is the hyper-parameter controlling the trade-off between reconstruction error and sparsity penalty.

The problem (1) can be solved by iteratively optimizing over a and ϕ alternately while holding the other set of variables fixed [9]. However, this process often takes a long time to converge. Furthermore, after training, to find the feature vector for a new input, we still have to do optimization (with fixed ϕ).

2.2 Sparse Auto-Encoders

An Auto-Encoder (AE) is a feed-forward neural network with two layers. The first layer, called *encoder*, maps the input x to the hidden representation a: $a = f\left(W^{(e)}x + b^{(e)}\right)$ where $f(\cdot)$ is some activation function (e.g. sigmoid), $W^{(e)}$ and $b^{(e)}$ are parameters of the encoder. The second layer, called *decoder*, then tries to reconstruct the input from the hidden representation a: $\hat{x} = W^{(d)}a + b^{(d)}$ where $\hat{x}$ is the reconstructed input, $W^{(d)}$ and $b^{(d)}$ are parameters of the decoder. In this way, we hope that the hidden representation can capture the structure of the input.

In Sparse Auto-Encoders (SAEs), besides reconstruction error, we also constrain the representation to be sparse (i.e. with a input, there are only a few hidden neurons active). Specifically, given the unlabeled data $\{x^{(1)}, \dots, x^{(N)}\}$ with $x^{(n)} \in \mathbb{R}^D$, SAEs minimize the following objective function:

$$J(W^{(e)}, b^{(e)}, W^{(d)}, b^{(d)}) = \sum_{n=1}^{N} \|x^{(n)} - \hat{x}^{(n)}\|_2^2 + \lambda s(a^{(n)}) \quad (2)$$

where: $a^{(n)} = f\left(W^{(e)}x^{(n)} + b^{(e)}\right)$; $\hat{x}^{(n)} = W^{(d)}a^{(n)} + b^{(d)}$; $s(\cdot)$ is some function that encourages the feature vector $a^{(n)}$ to be sparse; and λ is the hyper-parameter controlling the trade-off between reconstruction error and sparsity penalty.

Similar to Sparse Coding, SAEs aim at learning a representation that both preserves information about the input and is sparse. The difference between them is that SAEs have an explicit parametric encoder, while Sparse Coding has an implicit non-parametric encoder. This point helps training SAEs be more efficient than Sparse Coding; it can be done via the back-propagation algorithm. In addition, with a new input, SAEs can compute the corresponding feature vector very quickly just by one step.

3. SPARSE RECTIFIED AUTO-ENCODERS

The typical activation functions have been used in neural networks are the sigmoid function and the tanh function. Recently, a new activation function which have been found to work very well is the rectified linear function [10, 5, 15]:

$f(x) = \max(0, x)$. Units with such activation function are called rectified linear units (ReLUs).

ReLUs fit well with SAEs because such units naturally produce a sparse feature vector. Unlike logistic units that give small positive values when the input is not aligned with the filters (the incoming weight vectors of hidden units), ReLUs often give exact zeros. Furthermore, ReLUs compute faster than logistic or tanh units because they do not involve exponentiation and division; they just have to compute the max operation. Finally, ReLUs can potentially help jointly train multi-layers of features (instead of training layer by layer in greedy fashion) because ReLUs have been used to train supervised deep networks successfully [5, 15]. Therefore, here we will focus on SAEs with ReLUs (in the hidden layer). We call them Sparse Rectified Auto-Encoders (SRAEs).

3.1 Sparsity Constraint in SRAEs

The typical way that have been used to constrain sparsity in Sparse Auto-Encoders (SAEs) is pushing the average output $\bar{a}_j$ of hidden neuron j (over a minibatch) to some fixed target ρ (a value close to zero) [6, 4, 3]. In case the hidden neuron's output $\in [0, 1]$ (e.g. sigmoid unit), this can be done through the Kullback-Leibler (KL) divergence: $\sum_j \mathrm{KL}(\rho \| \bar{a}_j) = \sum_j \rho \log \frac{\rho}{\bar{a}_j} + (1-\rho) \log \frac{(1-\rho)}{(1-\bar{a}_j)}$. In case using ReLUs, the squared error can be used: $\sum_j (\bar{a}_j - \rho)^2$. Note that this way does not directly encourage the feature vector (corresponding to a example) to be sparse, but encourage the values of a feature (the outputs of a hidden neuron) over examples to be sparse. It, however, indirectly leads to a sparse feature vector because the reconstruction error tends to make learned features differ from each other; therefore, with an example, if some feature is active (having a non-zero value), the majority of the rest will be inactive (having a zero value).

This way, however, adds one more hyper-parameter (the fixed target ρ) to the list of SAEs' hyper-parameters which already has many ones (the trade-off parameter λ, the number of features, learning rate, minibatch size, and so on). As a result, the process of tuning hyper-parameters will become more annoying and more time-consuming. Why do not use L1-norm? It is natural because L1-norm is used in Sparse Coding. In addition, it doesn't have any extra hyper-parameter. It is also very simple; in case using ReLUs, it is just the sum of elements of the feature vector a. In the following section, we will explain the problem of training SAEs, in particular SRAEs, with L1-norm.

3.1.1 The Difficulty of Training SRAEs with L1-norm

The problem of training SAEs with L1-norm is that during the optimization process, L1-norm can drive the incoming weight vector of a hidden neurons to the state in which the hidden neuron is always inactive (produce zero with all examples in the dataset). And once the incoming weight vector has been in such a state, it will be stuck there forever and never get updated; the outgoing weight vector of this hidden neuron will also never get updated. Formally, let's consider a hidden neuron j which have a weight $W^{(e)}_{ji}$ connecting to an input neuron i and a weight $W^{(d)}_{kj}$ connecting to an output neuron k. The gradients of the objective function J in equation (2) (with the sparsity function $s(\cdot) = \|\cdot\|_1$) with respect to $W^{(e)}_{ji}$ and $W^{(d)}_{kj}$ are:

$$\frac{\partial J}{\partial W^{(d)}_{kj}} = \sum_{n=1}^{N} 2(\hat{x}^{(n)}_k - x^{(n)}_k) a^{(n)}_j \tag{3}$$

$$\frac{\partial J}{\partial W^{(e)}_{ji}} = \sum_{n=1}^{N} (\epsilon^{(n)}_j + \lambda) f'(a^{(n)}_j) x^{(n)}_i \tag{4}$$

where:

- $x^{(n)}_k$ and $\hat{x}^{(n)}_k$ are respectively the k^{th} element of the input vector $x^{(n)}$ and the reconstructed input vector $\hat{x}^{(n)}$.
- $a^{(n)}_j$ is the j^{th} element of the feature vector $a^{(n)}$.
- $\epsilon^{(n)}_j$ is the "error" that the hidden neuron j receives from the output layer (corresponding to the input $x^{(n)}$).

From equations (3) and (4), one can easily see that, during the optimization, if once the hidden neuron j has been in the state having a_j equal zero with all examples, the gradients $\frac{\partial J}{\partial W^{(d)}_{kj}}$ and $\frac{\partial J}{\partial W^{(e)}_{ji}}$ will be zeros with all examples (in case $f(\cdot)$ is the rectified linear function, the derivative $f'(0)$ equals 0) and the weights of this neuron will never get updated anymore. We call such neurons *"sleep" neurons*. Especially, the "easy to get exact zeros" property of ReLUs can make this problem easier to happen during the optimization.

The above problem may explain why people often don't use L1-norm in SAEs but instead, push the average output of a hidden neuron to a fixed target close to zero (but not zero!); this way may prevent the hidden neuron from the situation in which it is inactive for all examples and then never get updated. With sigmoid units, the KL divergence can be used and the average output cannot be zero because if so, the KL divergence will give an infinite penalty. With ReLUs, the KL divergence cannot be used because the outputs of ReLUs are not in $[0, 1]$. The squared error can be used instead but we found experimentally that the "sleep" neuron problem still happens. It is because with a zero average output, unlike the KL divergence, the squared error still gives a very small penalty. See Figure 1 for a comparison of them with the fixed target ρ of 0.1.

Although using L1-norm, Sparse Coding clearly doesn't have this problem because the encoder of Sparse Coding is implicit.

3.1.2 Sleep-Wake Stochastic Gradient Descent

To remedy the problem of training SRAEs with L1-norm, we propose a modified version of Stochastic Gradient Descent algorithm (SGD), called Sleep-Wake Stochastic Gradient Descent (SW-SGD). The idea is that during each epoch of SGD, we track the average outputs of hidden neurons. Then, after each epoch, we check if there are any "sleep" neurons (having the average output equal zero), and we will "wake-up" them by simply re-initializing their incoming weight vectors (including the biases). Despite its simplicity, our experiments showed that this strategy can help SRAEs successfully learn meaningful features without any "sleep" features.

3.2 Weight Constraint in SRAEs

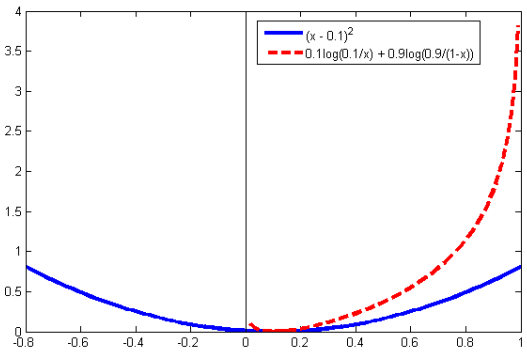


Figure 1: Comparison of KL divergence to squared error with the fixed target ρ of 0.1. When the average output of a hidden neuron is zero, KL divergence gives an infinite penalty while squared error still gives a very small penalty.

Besides sparsity constraint, weight constraint is also a key ingredient to get SAEs working. There are several ways have been used to constrain SAEs' weights:

- **Tied weights**: the weights of encoder and decoder are tied together ($W^{(d)} = (W^{(e)})^T$) [3]. This way was also used in other Auto-Encoder variants such as Denoising Auto-Encoders and Contractive Auto-Encoders [14, 13, 12]. Note that all [3, 14, 13, 12] used sigmoid units in the hidden layer. There is a trivial descent direction of SAEs' objective function in which the hidden neuron's output a_j is scaled down (by scaling down the incoming weight vector of this hidden neuron) and the outgoing weight vector of this hidden neuron is scaled up by some large constant; as a result, the sparsity penalty can decrease arbitrary while the reconstruction error is unchanged. Tied weights can help prevent from this trivial direction, but it is not clear what is going on when the encoder's weights and the decoder's weights are tied together, especially in case using sigmoid units.

- $W^{(d)}$ **norm constraint**: [15] constrained the basis vectors of the decoder (the outgoing weight vectors of hidden neurons) to have unit norm. This constraint is similar to Sparse Coding and also helps prevent from the scale problem. But how about the encoder's weights? For example, to be fair between features, the incoming weight vectors of hidden neurons should have the same norm.

- **Weight decay**: weights of the encoder and decoder are kept small by penalizing the sum of squares of them [6, 4]. As two previous ways, this way prevents SAEs from the scale problem too. It can be interpreted as a "soft" way to constrain the norms of the incoming weights vector of hidden units to be approximately equal to each other and the norms of the outgoing weight vectors of hidden units to be approximately equal to each other. However, this way introduces one more hyper-parameter; it's annoying.

3.2.1 *Our Proposed Weight Constraint for SRAEs*

In this section, we propose a reasonable way to constrain SRAEs' weights. It also doesn't introduce any extra hyper-parameter. Concretely, our way consists of two constraints:

- First, we tie the encoder's weights and the decoder's weights together: $W^{(d)} = \left(W^{(e)}\right)^T$

- Second, we also constrain the incoming weight vectors as well as the outgoing weight vectors of hidden units to have unit norm.

With an example x, if one just pays attention to non-zero rectified linear units, the whole system is a linear system. Therefore, with two above constraints, the encoder will project linearly the input vector x onto a few normalized basis vectors (in the whole set of normalized basis vectors) corresponding to non-zero hidden units; and then, the decoder will reconstruct the input vector from these basis vectors: $\hat{x} = W^T W x$ where x is a column vector and rows of W corresponds to normalized basis vectors selected by ReLUs (here, we just ignore the biases for simplicity). In other words, with above constraints, SRAEs will learn a set of normalized basis vectors such that different inputs can be explained by different small subsets of basis vectors (by projecting linearly the input onto the subset of basis vectors selected by ReLUs and then reconstructing the input from this subset).

The second constraint, however, cannot be enforced by gradient-based methods. To overcome this problem, we change the forward propagation formula of SRAEs as follows:

$$\hat{x} = (\hat{W}^{(e)})^T \max(0, \hat{W}^{(e)} x + b^{(e)}) + b^{(d)} \qquad (5)$$

where $\hat{W}^{(e)}$ is a row-normalized matrix of $W^{(e)}$ (each row of $W^{(e)}$ corresponds to an unnormalized basis vector). Here, the learned parameters are still $W^{(e)}$, $b^{(e)}$, and $b^{(d)}$. In this way, gradient-based methods can be used as usual.

Finally, the first constraint, tied weights, also helps save about haft of memory compared to untied weights. It will be beneficial when using GPU (for parallel computing).

4. EXPERIMENTS

4.1 Setup

We experimented on the MNIST dataset which composes of grayscale images (28×28 pixels) of 10 hand-written digits (from 0 to 9) [8]. Figure 2 shows some examples of this dataset. The images were preprocessed by scaling to $[0, 1]$. We used the usual split: $50,000$ examples for training, $10,000$ examples for validation, and $10,000$ examples for test.

We conducted all experiments using the Python Theano library [2], which allows for quick development and easy use of GPU (for parallel computing). We used a single NVIDIA GTX 560 GPU.

After the unsupervised feature learning phase, we evaluated the learned features by feeding them to a softmax regression and measuring the classification error. Concretely, given the training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ where $x^{(i)} \in \mathbb{R}^D$ is the image vector and $y^{(i)} \in \{0, \ldots, 9\}$ is the class label, we fed $x^{(i)}$ to the trained Auto-Encoder (the Auto-Encoder was trained on the unlabeled data $\{x^{(1)}, \ldots, x^{(N)}\}$)

Figure 2: Some examples of MNIST dataset

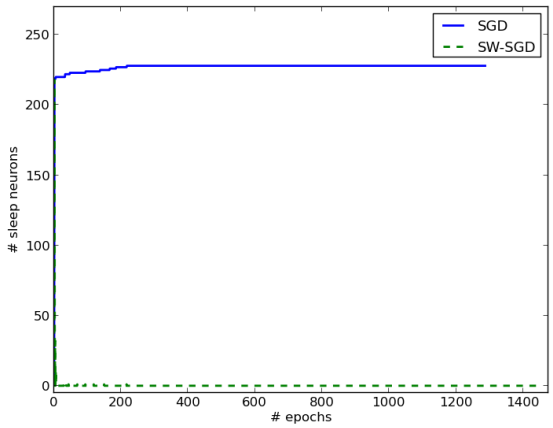


Figure 3: The number of "sleep" hidden neurons of SRAEs during the optimization process with SGD and SW-SGD. The optimization of SGD ended up with 228/1000 "sleep" neurons while SW-SGD ended up without any "sleep" neurons. (These two optimizations terminated after different number of epochs because of the early stopping strategy.)

to get the corresponding feature vector $f^{(i)}$; by this way, we got the new training set $\{(f^{(1)}, y^{(1)}), \ldots, (f^{(N)}, y^{(N)})\}$. Then, we used this new training set to train a softmax regression. With a test example x, we first used the trained Auto-Encoder to compute the feature vector f; then, we fed f to the trained softmax regression to get the class prediction.

In both unsupervised and supervised phase, we used Stochastic Gradient Descent as the optimization algorithm with mini-batch size 100 and early stopping (in the unsupervised phase, we stopped the optimization based on the objective value on the validation set; in the supervised phase, we based on the classification error on the validation set). In all experiments, we used SRAEs with 1000 hidden units, a trade-off parameter λ of 0.25, an unsupervised learning rate of 0.05, and a supervised learning rate of 1.

4.2 SGD versus SW-SGD

To see the problem of training SRAEs with L1-norm sparsity penalty and the effect of our "sleep-wake" strategy, we compared training SRAEs with ordinary Stochastic Gradient Descent (SGD) and our modified version, Sleep-Wake Stochastic Gradient Descent (SW-SGD). In this experiment, we used our proposed weight constraint (tied weights + $W^{(e)}$ norm constraint + $W^{(d)}$ norm constraint).

Figure 3 shows the number of "sleep" hidden neurons of SRAEs during the optimization process with SGD and with SW-SGD. The problem of training SRAEs with L1-norm sparsity penalty is that during the optimization, L1 penalty can push the incoming weight vectors of hidden neurons to "sleep" states (meaning that the corresponding hidden neurons always give zero outputs with all examples in the dataset) and then, they will never get updated anymore; as can be seen from the figure, with ordinary SGD, the number of "sleep" neurons increased during the optimization, especially during the first epochs when the optimization had not stable yet. The SGD optimization finally ended up with 228/1000 "sleep" neurons. This problem of L1 penalty can be remedied by our simple "sleep-wake" strategy; the SW-SGD optimization ended up without any "sleep" neurons.

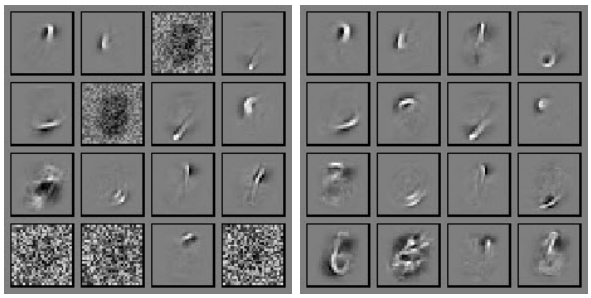
Figure 4 visualizes some example filters (the incoming weight vectors of hidden neurons) learned by SGD and SW-SGD. With SGD, there are five "sleep" filters; they look meaningless. With SW-SGD, there are not any "sleep" filters; all of them look meaningful, like "pen stroke" detectors.

Making use of all filters, SW-SGD achieved better training unsupervised objective value and better test classification performance (with softmax regression) than SGD (Table 1).

4.3 Our Proposed Weight Constraint versus Other Weight Constraints

In this second experiment, we compared our proposed weight constraint for SRAEs to other weight constraints that are possible to be applied to SRAEs. Concretely, we considered the following weight constraints:

- $W^{(d)}$ **norm constraint**: the outgoing weight vectors of hidden units (the columns of $W^{(d)}$) are constrained to have unit norm.



(a) SGD (b) SW-SGD

Figure 4: Figure (a) shows example filters learned by SGD; one can recognize there are five "sleep" filters looking meaningless. Figure (b) shows example filters learned by SW-SGD; all filters look meaningful, like "pen stroke" detectors.

Table 1: Unsupervised objective value on the training set and classification error (with softmax regression) on the test set when training SRAEs with SGD and with SW-SGD

	SGD	SW-SGD
Train Unsupervised Objective Value	9.84	**9.48**
Test Classification Error (%)	1.70	**1.62**

- $W^{(e)}$ **&** $W^{(d)}$ **norm constraint**: both the incoming and outgoing weight vectors of hidden units (the rows of $W^{(e)}$ and the columns of $W^{(d)}$ respectively) are constrained to have unit norm.
- **Tied weights**: the encoder's weights and the decoder's weights are tied together ($W^{(d)} = (W^{(e)})^T$).

Our weight constraint combines both $W^{(e)}$ **&** $W^{(d)}$ **norm constraint** and **tied weights**. In this experiment, we used SW-SGD to train SRAEs. As can be seen from Table 2, our weight constraint gave the best test classification performance (with softmax regression). In the last column, we also show the (approximate) training time per epoch of SRAEs with these different weight constraints (because of the early stopping strategy, the training processes of SAREs with different weight constraints can terminate after different number of epochs; therefore, it will be more accurate to compare them in term of the training time per epoch rather than the total training time). Weight constraints sorted from lowest to highest training time per epoch are: tied weights (2 seconds), $W^{(d)}$ norm constraint (3 seconds), our weight constraint (4 seconds), and $W^{(e)}$ & $W^{(d)}$ norm constraint (5 seconds). This order is reasonable because:

- In tied weights, SRAE doesn't have to do normalization in the forward propagation phase.
- In $W^{(d)}$ norm constraint, SRAE's decoder has to do normalization in the forward propagation phase; and because of this, in the back-propagation phase, the computation of derivatives with respect to the decoder's parameters will also become more expensive than usual.
- In our weight constraint, although we have to do normalization in both the encoder and decoder, we just have to compute the encoder's normalized weights and use them for the decoder thanks to the tied weights constraint. Its epoch time is higher than $W^{(d)}$ norm constraint above because in the back-propagation phase, the computation of derivatives with respect to both the encoder's parameters and the decoder's parameters is more expensive than usual.
- In $W^{(e)}$ & $W^{(d)}$ norm constraint, the training time per epoch is highest because SRAE has to do normalization in the encoder and decoder separately and the computation of derivatives with respect to both the encoder's parameters and the decoder's parameters is more expensive than usual.

Although the training time per epoch of our weight constraint is pretty high compared to other weigh constraints, it's still fast (thanks to the use of GPU). Its total training time is roughly 2.5 hours.

Table 2: Comparison of our weight constraint to other possible weight constraints. Our weight constraint gave the best classification performance (with softmax regression) on the test set. The last column shows the training time per epoch (roughly) of SRAEs with these different weight constraints.

Weight Constraint	Test Error (%)	Epoch Time (sec)
$W^{(d)}$ norm constraint	3.28	3
$W^{(e)}$ & $W^{(d)}$ norm constraint	2.51	5
Tied weights	2.04	2
Our weight constraint	**1.62**	4

Table 3: Comparison of SRAEs (with our weight constraint and SW-SGD) to other Auto-Encoder variants, including: Denoising Auto-Encoders (DAEs), Contractive Auto-Encoders (CAEs), and Higher Order Contractive Auto-Encoders (HCAEs), in term of classification error (with softmax regression) on the test set

Feature Learning Algorithm	Test Error (%)
DAEs [12]	2.05
CAEs [12]	1.82
SRAEs	1.62
HCAEs [12]	**1.20**

4.4 SRAEs versus Other Auto-Encoder Variants

Finally, we also compared SRAEs (with our weight constraint and SW-SGD) to other Auto-Encoder variants, including:

- **Denoising Auto-Encoders (DAEs)** [14]: want to learn robust features by making the input corrupted and trying to reconstruct the "clean" input from this corrupted version.
- **Contractive Auto-Encoders (CAEs)** [13]: want to learn features robust to small changes of the input by besides the reconstruction error, penalizing the Frobenius norm of the Jacobian of the feature vector with respect to the input vector.
- **Higher Order Auto-Encoders (HCAEs) [12]**: are the extension of CAEs; besides the reconstruction error and the Jacobian norm, HCAEs also penalize the approximated Hessian norm.

Table 3 compares the test classification performance of SRAEs to these Auto-Encoder variants. Note that with DAEs, CAEs, and HCAEs, [12] used 1000 hidden units, the sigmoid activation function in the hidden and output layer, the cross-entropy reconstruction error, and tied weights. Our SRAEs were better in term of test classification performance than DAEs and CAEs but worse than HCAEs. However, HCAEs are more complicated than our SRAEs with many hyper-parameters which need to be tuned.

5. CONCLUSION

In this paper, we have investigated SRAEs and in particular, two key ingredients to get SRAEs working: spar-

sity constraint and weight constraint. We have tried to understand the optimization problem when training SRAEs with L1-norm sparsity penalty and proposed a simple modified version of SGD, called SW-SGD, to remedy this problem. We have also proposed a reasonable weight constraint for SRAEs. Our experiments on the MNIST dataset have shown that our weight constraint and SW-SGD work well with SRAEs and can help SRAEs learn meaningful features that give excellent classification performance compared to other Auto-Encoder variants.

Our future work will include:

- Making use of sparsity to speed up the training.
- Unsupervised deep learning: SRAEs can be used to learn multiple layers of representation in greedy fashion but the interesting question is how to jointly learn multiple layers of representation?

6. REFERENCES

[1] Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012.

[2] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[3] A. Coates. *Demystifying Unsupervised Feature Learning*. PhD thesis, Stanford University, 2012.

[4] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[5] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pages 315–323, 2011.

[6] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646–654, 2009.

[7] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building high-level features using large scale unsupervised learning. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 81–88, New York, NY, USA, July 2012. Omnipress.

[8] Y. LeCun. The MNIST database. `http://yann.lecun.com/exdb/mnist/`.

[9] H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[10] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[11] B. A. Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[12] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot. Higher order contractive auto-encoder. *Machine Learning and Knowledge Discovery in Databases*, pages 645–660, 2011.

[13] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 833–840, 2011.

[14] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[15] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, et al. On rectified linear units for speech processing. ICASSP, 2013.