Capsule Restricted Boltzmann Machine

Yifeng Li
Digital Technologies Research Centre
National Research Council Canada
Ottawa, Ontario K1A 0R6 Canada
yifeng.li@nrc-cnrc.gc.ca

Xiaodan Zhu
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario K7L 3N6 Canada
xiaodan.zhu@queensu.ca

Abstract

We propose a capsule restricted Boltzmann machine by replacing individual hidden variables with encapsulated groups of hidden variables. Our preliminary experiments show that capsule activities can be dynamically determined in context, and these activity spectra exhibit between-class patterns and within class variations.

1 Introduction

The development of deep neural network models is powered by the theory of distributed representation [1] which has achieved mega successes in natural language processing (e.g. word embedding [2]) and computer vision (convolutional nets [3]). However, its limited capacity of supporting complex symbolic manipulations puzzles and reminds people whether revised distributed representational or alternative representational theories exist [4, 5]. Recently, the emergement of capsule nets for classification spurred new discussions and explorations on the next-generation deep learning systems [6, 7]. Meanwhile, generative models have evolved from restricted Boltzmann machine (RBM) based models, such as Helmholtz machines [8] and deep belief nets [9], to variational auto-encoders (VAEs) [10] and generative adversarial networks (GANs) [11] with some successes in image and video generation. Sitting on the shoulders of these accomplishments, we are wondering whether the concept of capsules can be adopted (but unnecessarily identical) to generative models? While there might be many different ways to pursue it, as the first step, we propose in this paper a capsule RBM model that has generative capsules in its hidden layer. In the rest of this paper, we present this model and show some preliminary but interesting results. Certainly, further investigation would still be desirable.

2 Method

Based on theories of exp-RBMs [12, 13, 14] and discriminative capsule nets [6, 7], we propose capsule RBM (or cap-RBM) replacing hidden variables with capsules. We use $\boldsymbol{x}$ to represent the visible vector. The k-th capsule in the hidden layer includes $\boldsymbol{h}_k$ which hosts multiple hidden random variables following any distribution from the exponential family, and z_k a binary random variable indicating whether this capsule is active. An example of such network is displayed in Figure 1.

We assume visible vector $\boldsymbol{x}$ has M units, the hidden layer has K capsules (each contains J units and one switch variable). For notational simplicity, we write $\boldsymbol{h} = \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K\}$. Following the steps of defining exp-RBMs, first of all, the base distributions of cap-RBM can be defined in natural form as

$$p(\boldsymbol{x}) = \prod_{m=1}^{M} e^{\boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{s}_m + \log f(x_m) - A(\boldsymbol{a}_m)}, \quad p(\boldsymbol{h}) = \prod_{k=1}^{K} \prod_{j=1}^{J} e^{\boldsymbol{b}_{k,j}^{\mathrm{T}} \boldsymbol{t}_{k,j} + \log g(h_{k,j}) - B(\boldsymbol{b}_{k,j})},$$

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

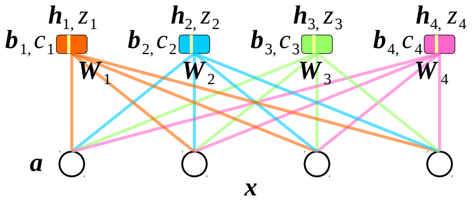


Figure 1: An example of cap-RBM. $\boldsymbol{a}$ contains the bias parameters on the visible units. $\boldsymbol{b}_k$ includes bias parameters on $\boldsymbol{h}_k$. c_k denotes the bias parameter on variable z_k. The capsules ($\boldsymbol{h}_k$ and z_k, $k = 1, \cdots, K$) interact with the visible vector through matrices $\boldsymbol{W}_k$ and $\boldsymbol{\Omega}$ (not shown in figure).

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} e^{c_k z_k - C_k(z_k)}, \tag{1}$$

where $p(\boldsymbol{x})$ and $p(\boldsymbol{h})$ are distributions from the exponential class, and $p(\boldsymbol{z})$ is Bernoulli distributed. $\boldsymbol{a}_m = [a_m^{(1)}, \cdots, a_m^{(R)}]$ and $\boldsymbol{b}_{k,j} = [b_{k,j}^{(1)}, \cdots, b_{k,j}^{(U)}]$ are respectively the natural parameters of x_m and $h_{k,j}$, $\boldsymbol{s}_m$ and $\boldsymbol{t}_{k,j}$ are respectively their sufficient statistics, $A(\boldsymbol{a}_m)$ and $B(\boldsymbol{b}_{k,j})$ are corresponding log-partition functions, and $f(x_m)$ and $g(h_{k,j})$ are base measures. Similar notations apply to $p(\boldsymbol{z})$.

Second, the joint distribution has a similar form as regular RBM: $p(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z}) = \frac{1}{Z} e^{-E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}$, where Z is the partition function, and the energy function can be defined as

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z}) = & -\prod_{m=1}^{M} \big(\boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{s}_m + \log f(x_m)\big) - \prod_{k=1}^{K} \prod_{j=1}^{J} \big(\boldsymbol{b}_{k,j}^{\mathrm{T}} \boldsymbol{t}_{k,j} + \log g(h_{k,j})\big) - \boldsymbol{c}^{\mathrm{T}} \boldsymbol{z} \\ & - \sum_{k=1}^{K} z_k (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{h}_k) - \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{z}, \end{aligned} \tag{2}$$

where the first three terms are bias terms, and the last two terms define interactions between observations and capsules. The interaction term $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{z}$ could be optional in the model.

After that, fortunately we are able to obtain the conditionals in decomposable forms, as below.

$$\begin{cases} p(\boldsymbol{x}|\boldsymbol{h}, \boldsymbol{z}) = \prod_{m=1}^{M} p\big(x_m | \eta(\hat{\boldsymbol{a}}_m)\big), \quad \hat{a}_m^{(r)} = a_m^{(r)} + \delta(r=1) \Big(\sum_{k=1}^{K} z_k (\boldsymbol{W}_k)_{m,:} \boldsymbol{h}_k + \boldsymbol{\Omega}_{m,:} \boldsymbol{z} \Big) & (3) \\ p(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{z}) = \prod_{k=1}^{K} \prod_{j=1}^{J} p\big(h_{k,j} | \eta(\hat{\boldsymbol{b}}_{k,j})\big), \quad \hat{b}_{k,j}^{(u)} = b_{k,j}^{(u)} + \delta(u=1) \Big(z_k (\boldsymbol{W}_k^{\mathrm{T}})_{j,:} \boldsymbol{x} \Big) & (4) \\ p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{h}) = \prod_{k=1}^{K} \mathcal{BE}\big(z_k | \eta(\hat{c}_k)\big), \quad \hat{c}_k = c_k + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{W}_k \boldsymbol{h}_k + (\boldsymbol{\Omega}^{\mathrm{T}})_{k,:} \boldsymbol{x}, & (5) \end{cases}$$

where, $r \in \{1, \cdots, R\}$, $u \in \{1, \cdots, U\}$, without loss of generality we assume the first statistics for x_m and $h_{k,j}$ are respectively x_m and $h_{k,j}$, function $\eta(\cdot)$ maps the natural parameters to the standard forms, and $\delta(\cdot)$ is a Kronecker delta function. See [13] for transformation tables for exponential family distributions in natural and standard forms.

Variable z_k governs whether the k-th capsule can interact with the visible variables, that is, only when $z_k = 1$, the k-th capsule has an impact on $\boldsymbol{x}$. Thus, the model dynamically decides which capsules are active depending on the context. The conditional distribution of $\boldsymbol{x}$ is determined by its interactions with the active capsules and the switch variable. The conditional distribution of $\boldsymbol{h}_k$ depends on its interaction with $\boldsymbol{x}$. In turn, the value of the switch variable $\boldsymbol{z}$ depends on all the other variables. Specifically, this model has the following metrics. (1) An observation $\boldsymbol{x}$ will only activate a portion of capsules, some of which may be ubiquitous while some may be specific. (2) This leads to block-wise structured representation of an observation in the hidden space. (3) Through examining the activity of dozens of capsules, indicated by $\boldsymbol{z}$, it offers a better interpretation compared to investigating the meanings of hundreds or thousands of individual hidden variables in other generative models.

The model parameters are denoted by $\boldsymbol{\theta} = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}, \boldsymbol{\Omega}\}$, where we simply write $\boldsymbol{b} = \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_K\}$ and $\boldsymbol{W} = \{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K\}$. Similar to exp-RBM, the gradients w.r.t. these model parameters can be computed in the form:

$$\Delta_\theta = \frac{1}{N}\sum_{n=1}^{N}\Big(\mathrm{E}_{p(\boldsymbol{h},\boldsymbol{z}|\boldsymbol{x}_n)}\big[\frac{\partial E(\boldsymbol{x}_n, \boldsymbol{h}, \boldsymbol{z})}{\partial \boldsymbol{\theta}}\big] - \mathrm{E}_{p(\boldsymbol{x},\boldsymbol{h},\boldsymbol{z})}\big[\frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial \boldsymbol{\theta}}\big]\Big). \tag{6}$$

The derivatives of the energy function w.r.t. the model parameters can be computed as below:

$$\frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial a_m^{(r)}} = -s_m^{(r)}, \quad \frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial b_{k,j}^{(u)}} = -t_{k,j}^{(u)}, \quad \frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial \boldsymbol{c}} = -\boldsymbol{z},$$

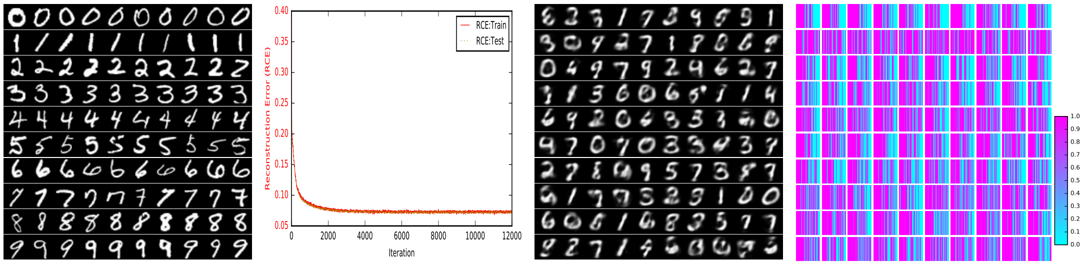
$$\frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial \boldsymbol{W}_k} = -z_k \boldsymbol{x}\boldsymbol{h}_k^{\mathrm{T}}, \quad \frac{\partial E(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{z})}{\partial \boldsymbol{\Omega}} = -\boldsymbol{x}\boldsymbol{z}^{\mathrm{T}}. \tag{7}$$

In Equation (6), sampling from the joint distribution can be realized by Gibbs sampling with the conditional distributions; the data-dependent term can be estimated by mean-field variational approximation with the conditional distributions over $\boldsymbol{h}$ and $\boldsymbol{z}$ while fixing $\boldsymbol{x}_n$.

A range of modifications could be done for this model. For instance, a generative convolutional mechanism could be integrated in the visible layer, which may be particularly useful for complex inputs, e.g. large images or signals. More hidden capsule layers could be naturally added to form capsule deep generative models (cap-DGMs). These models could be either undirected or directed.

3 Experiments

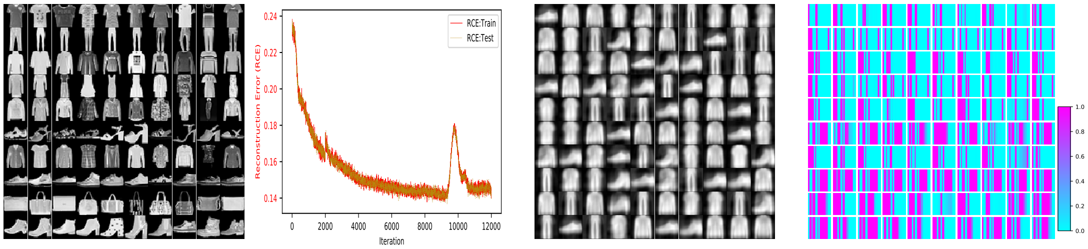
We have preliminarily investigated the performance of cap-RBM on MNIST data (`http://yann.lecun.com/exdb/mnist`, see Figure 2a for some examples from the data). We let both base distributions of $\boldsymbol{x}$ and $\boldsymbol{h}_k$ be Bernoulli, and set K=40, J=16, initial learning rate to 0.02 (gradually decreased), batch size to 100, and number of epochs to 20. We tracked the reconstruction error of training and test samples, which reduced quickly along model learning (see Figure 2b). Figure 2c shows 100 images generated by Gibbs sampling with the learned model. We obtained the capsule activities (values of $\boldsymbol{z}$) of the actual images as in Figure 2a using mean-field approximation, and displayed them as spectra in Figure 2d. Interestingly, class-wise patterns (such as patterns in classes 0 and 1) and within-class variations can be observed from the spectra. Furthermore, we find that digits sharing similar parts tend to have partially similar patterns (such as digits 1, 4, 7, and 9 all have vertical strokes). By contrast, a Bernoulli-Bernoulli RBM with 640 hidden variables was run, but its hidden states couldn't clearly exhibit such patterns (results not shown due to page limit).



(a) Actual Images. (b) Reconstruction Errors. (c) Generated Images. (d) Capsule Activity.

Figure 2: Performance of cap-RBM on MNIST.

We also explored cap-RBM on the Fashion-MNIST data [15] (see Figure 3a for some examples from its test set). We let both base distributions of $\boldsymbol{x}$ and $\boldsymbol{h}_k$ be Gaussian, and set K=20, J=16, and the initial learning rate to 0.005. Figure 3b shows the reconstruction errors of training samples and test samples as the learning proceeded. Figure 3c gives some images generated from the model after training. Figure 3d displays the capsule activities corresponding to actual test images in Figure 3(a). Again, one can see that the capsule activity spectra exhibit class-specific patterns (e.g. T-shirt/top (row 1) versus Trouser (row 2)). Close classes tend to share similar spectral patterns (e.g. Sandal (row 6) and Sneaker (row 8)). Furthermore, within each class, variations can be observed among capsule activity spectra of samples.



(a) Actual Images. (b) Reconstruction Errors. (c) Generated Images. (d) Capsule Activity.

Figure 3: Performance of cap-RBM on FASHION-MNIST.

4 Conclusion

In this study, we explored using capsules in generative models. Our preliminary results, as a proof of concept, corroborate that our cap-RBM model is able to dynamically activate capsules depending on the context, and the activity of capsules offers a new way to interpret representations of observations in hidden space. Our next task is to test the model on more complicated data in comparison with benchmarks. Our method will be extended to deep generative models. The reasonability of our model from the cognitive science perspective is currently under discussion.

Acknowledgments

We thank the reviewers for valuable feedback. This research is supported by the National Research Council Canada.

References

[1] G.E. Hinton, J.L. McClelland, and D.E. Rumelhart. Distributed representations. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, pages 77–109. MIT Press, Cambridge, MA, 1986.

[2] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 2:1137–1155, 2003.

[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, December 1989.

[4] J.A. Fodor and Z.W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[5] G. Hinton. Aetherial symbols. In *AAAI Spring Symposium on Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*, 2015.

[6] S. Sabour, N. Frosst, and G.E. Hinton. Dynamic routing between capsules. In *Neural Information Processing Systems*, pages 3856–3866, 2017.

[7] G. Hinton, S. Sabour, and N. Frosst. Matrix capsules with EM routing. In *International Conferences on Learning Representations*, 2018.

[8] P. Dayan, G.E. Hinton, R. Neal, and R.S. Zemel. The Helmholtz machine. *Neural Computation*, 7:1022–1037, 1995.

[9] G.E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[10] D.P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[11] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[12] M. Welling, M. Rosen-Zvi, and G Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, pages 1481–1488, 2005.

[13] Y. Li and X. Zhu. Exponential family restricted Boltzmann machines and annealed importance sampling. In *International Joint Conference on Neural Networks (IJCNN)*, pages 39–48, July 2018.

[14] Y. Li and X. Zhu. Exploring Helmholtz machine and deep belief net in the exponential family perspective. In *ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, July 2018.

[15] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017.