$$P(h^{(i)} \mid h^{(i+1)}) = \prod_j^{J^{(i)}} P(h_j^{(i)} \mid h^{(i+1)}) = \text{simgoid}\left(\sum_{k=1}^{J^{(i+1)}} h_k^{(i+1)} W_{jk}^{(i)} + b_j^{(i)}\right) \tag{36}$$

其中,$b^{(i)}$表示第 i 个隐藏层的偏置,$W^{(i)}$表示第 i–1 层和第 i 层之间的权值矩阵,利用逐层训练的方法,可以有效地初始化一个 DBNs 模型.DBMs 是一种层次化的概率无向图模型,每一层单元的激活取决于与之直接相连的上下两层的节点.虽然 DBMs 的计算复杂度高于 DBNs,但是由于 DBMs 每一层单元的激活组合了更加抽象的特征,DBMs 的图像生成能力更加出色.以含有 2 个隐藏层的 DBM 模型为例,其能量函数可以表示如下:

$$E\left(v, h^{(1)}, h^{(2)}\right) = -\sum_{i=1}^{D}\sum_{j=1}^{J^{(1)}} v_i W_{ij}^{(1)} h_j^{(1)} - \sum_{j=1}^{J^{(1)}}\sum_{j'=1}^{J^{(2)}} h_j^{(1)} W_{jj'}^{(2)} h_{j'}^{(2)} - \sum_{j=1}^{J^{(1)}} b_j^{(1)} h_j^{(1)} - \sum_{j'=1}^{J^{(2)}} b_{j'}^{(2)} h_{j'}^{(2)} - \sum_{i=1}^{D} c_i v_i \tag{37}$$

根据能量函数,DBMs 单元的激活概率为

$$P\left(h_j^{(1)} = 1 \mid v, h^{(2)}\right) = \text{simgoid}\left(\sum_i v_i W_{ij}^{(1)} + \sum_{j'} W_{jj'}^{(2)} h_{j'}^{(2)} + b_j^{(1)}\right) \tag{38}$$

$$P\left(h_{j'}^{(2)} = 1 \mid h^{(1)}\right) = \text{sigmoid}\left(\sum_j h_j^{(1)} W_{jj'}^{(2)} + b_{j'}^{(2)}\right) \tag{39}$$

$$P(v_i = 1 \mid h^{(1)}) = \text{sigmoid}\left(\sum_j W_{ij}^{(1)} h_j^{(1)} + c_i\right) \tag{40}$$

DBNs 和 DBMs 模型都可以看作前馈神经的多层神经网络,通常,使用 RBMs 初始化的 DBNs 和 DBMs 是一种无监督模型,无监督初始化的神经网络若想完成监督学习的任务,则必须建立特征与标签之间的映射关系.基于训练后的 DBNs 和 DBMs,综合监督学习的方法,可以完成模式识别任务,常用的监督学习方法有:

(1) 基于 BP 算法的权值微调.

(2) 基于 wake-sleep 算法的认知生成过程.

(3) 基于 Class-RBMs 和分类器的组合.

第 1 种方法是目前最主流的监督学习算法,BP 算法基于梯度下降的思想,其中,有一个相当粗糙的梯度下降法取得了巨大的成功:随机梯度下降(stochastic gradient descent,简称 SGD),在基于监督学习的深度网络(deep neural nets,简称 DNNs)中,SGD 是梯度下降法中最简单的,然而,SGD 算法在训练 DNN 时取得了非常好的效果.至于为什么非常粗糙的算法对神经网络这种复杂的优化问题有效,仍然是一个有待进一步研究的问题.

Wake-sleep 算法是一种基于认知科学的算法:在神经网络中,当训练数据是自上而下生成的时候,那么被用于自上而下(top-down)生成图像的隐藏层单元的状态就可以用于训练自下而上(bottom-up)的认知权值(reco-weights)[56].如果我们已经获得了较好的认知连接(reco-connections),就可以根据前一层的活跃度信息重建下一层的活跃度,从而学习生成权值.给定生成权值(generative weights),算法学习得到认知权值(recognition weights);反之,给定认知权值,算法也可以学习生成权值.在清醒阶段(“wake” phase),认知权值被用于自下而上驱动神经元,相邻层神经元的状态被用于训练生成权值;在睡眠阶段(“sleep” pahse),自上而下地生成连接被用于认知连接的学习,从而生成数据,此时相邻层的神经元状态就可用于学习认知连接.

第 3 种方法是基于 Class-RBMs 以及分类器的监督学习方法.Class-RBMs 是一种基于样本和标签的 RBMs 模型,Class-RBMs 建模输入 x 和标签 y 之间的联合概率分布.其能量函数可以表示如下:

$$E(x, y, h) = -h^T W x - b^T x - c^T h - d^T e_y - h^T U e_y \tag{41}$$

基于能量函数,激活函数可以表示为

$$P(h \mid y, x) = \text{sigmoid}\left(c_j + U_{jy} + \sum_i W_{ji} x_i\right) \tag{42}$$

$$P(x_i = 1 \mid h) = \text{sigmoid}\left(b_i + \sum_j W_{ji} h_j\right) \tag{43}$$

此时,可以求得关于标签 y 和输入 x 的条件概率:

$$P(y \mid x) = \frac{\exp(-F(y, x))}{\sum_{y \in \{1,2,\ldots,C\}} \exp(-F(y, x))} \tag{44}$$

其中,$F(y,x)$为自由能.Class-RBMs 建立了输入数据和标签之间的联合分布,这在一定程度上类似于 BP 算法,不同的是,BP 算法包含了特征逐层抽象的过程.基于 Class-RBMs,在模型堆叠之后直接使用分类器,例如支持向量机(support vector machines,简称 SVMs),也可以获得比较理想的识别效果.

3.2 基于变分自编码和GAN的混合模型

VAEs 模型被广泛地应用于半监督学习和图像生成中,VAEs 是基于贝叶斯原理的有向图模型,分为编码器和解码器两部分,在传统的自编码网络中,从$X\rightarrow Z\rightarrow X'$,$X$表示输入,$Z$是自编码器的隐式表达,$X'$是解码表示.这样的一个过程实现了无监督表征学习.可以学习到隐式表达 Z.VAEs 不同于普通的自编码网络,隐式表达 Z 是概率分布的形式,模型从边缘分布 $P(x)$出发,利用 KL 散度,获得似然函数的变分下界.在 VAEs 中,编码器和解码器可以具有不同的形式,其中最常用的形式为神经网络,编码器和解码器都由神经网络组成,其中假设基于输入 x 的条件概率 $q(z|x)$表示编码器,为了引入变分边界,似然函数可以写为如下形式:

$$\ln(p_\theta(x))=KL(q_\phi(h|x)\|p_\theta(h|x))+L(x,\theta,\phi) \tag{45}$$

其中,L为似然函数中剩余的部分,由于 KL 散度是大于等于 0 的,因此上述的似然函数可以进一步写成如下形式:

$$\ln(p_\theta(x))\geqslant L(x,\theta,\phi)=-KL(q_\phi(h|x)\|p_\theta(h))+E_{q_\phi(h|x)}\left[\ln p_\theta(x|h)\right] \tag{46}$$

其中,$p(h)$是隐层节点的先验概率,一般情况下,假设先验概率为简单的分布形式,例如均值为 0、方差为 1 的标准正态分布,由这个正态分布和概率解码器来生成数据 x,但是使用高斯分布来建模输入数据存在一定的不足,对于图像数据,深度网络在提取特征的过程中其特征是逐步抽象化的,仅使用连续的随机变量来建模图像会导致模型分布过度平滑,为了在抽象特征的基础上实现特征的离散化组合,基于 VAEs 和 RBMs 的混合模型被提了出来,在 VAEs 的基础上,使用 RBMs 作为先验替换传统的标准正态分布,多层卷积网络的基础上,使用 RBMs 建模离散化的高度抽象化的特征,并通过参数化手段,使用 BP 算法训练模型,基于这种方法的图像生成模型可以得到更加清晰、锐利的生成图像.

另一种思路是将 RBMs 和对抗生成网络相结合.GANs 是目前非常有效的生成模型,传统的 GANs 通过对抗的方式最小化模型分布和数据分布之间的 JS 散度,WGANs 在 GANs 的基础上进行了改进,最小化模型分布和数据分布之间的 Wasserstein 距离,但是,WGAN 的训练还存在一定的问题,其训练不稳定且有随时崩溃的风险,且 GANs 对超参数非常敏感,往往需要进行大量的调试和人为干预,才能获得一个比较好的生成模型,为了获得比较稳定且融合 GANs 优势的生成模型,有学者将对抗的思想引入到 RBMs 中,同时最小化数据分布和模型分布之间的 forword KL 散度和模型分布与数据分布之间的 reverse KL 散度,综合自编码器结构,GAN-RBMs 可以结合 VAEs 或自动编码器模型,组成多层的生成模型.

3.3 卷积深度置信网

另一种成功的 DNNs 模型是卷积神经网络(convolutional neural nets,简称 CNNs),不同于预训练的机制,CNNs 从网络拓扑结构上优化 DNNs,利用卷积和池化操作,将局部性信息和不变性信息引入到神经网络中,利用先验信息减少网络参数,进一步降低了计算复杂度.CNNs 在自然图像处理、音频、视频等方面取得了很多研究成果.基于结构的特殊性,CNNs 的训练参数比一般的全连接神经网络的要少得多,为了加速网络的训练,并减缓梯度扩散现象,CNNs 可以使用 ReLU 作为激活单元,并在 GPU 上并行训练.目前在工业界的推广下,除了各种小的修改(Residual Nets、ReLU、BatchNorm、Adam Optimizer、Dropout、GRU、GAN、LSTMs 等)外,神经网络的主要训练方法又回到 30 年前的 BP 算法[57−73].针对图像处理问题,BP 算法将原始的复杂统计问题转化为神经网络的参数调节问题和网络结构的优化问题.这大幅度地降低了 DNNs 研究的门槛,吸引了更多的学者追踪 DNN 的相关研究.同时,GPU 的使用提供了训练 DNNs 的硬件基础.基于 GPU 的深度学习框架,如 CAFFE、TensorFlow 等,为针对 DNNs 的程序设计提供了方便、有力的支持.目前,许多对 DNNs 的研究贡献都集中在神经网络的梯度流上,如:传统的网络采用 sigmoid 函数作为激活函数,然而 sigmoid 函数是一种饱和函数,这会导致梯度扩散问题,为了缓解这个问题,线性整流单元(rectified linear unit,简称 ReLU)以及改进的 Leaky ReLU 被引入到 DNNs 中;为了强调梯度和权值分布的稳定性,ELU 和 SELU 激活函数被引入到 DNNs 中[62];当 DNNs 的深度过大时,尽管使用了非饱和的激活函数,DNNs 的训练还是会面临梯度消失的问题,为此,学者们提出了 highway 网络和 ResNets 模型[65,66].为了稳定参数的均值和方差,BatchNorm 方法被应用到 DNN 的训练中[63].为了缓解过拟合,Dropout 方法和 Weight uncertainty 方法被用于 DNNs[67−70].

基于 RBMs,卷积神经网络可以被用于处理图像识别和图像生成任务,Lee 等学者组合卷积网络和 RBMs,提出了卷积深度置信网(convolutional deep belief nets,简称 CDBNs),通过引入卷积和概率最大池化操作,CDBNs 实现了图像的识别和生成过程.卷积深度置信网的能量函数可以表示如下:

$$E(v,h)=-\sum_{l=1}^{L}\sum_{i,j}c^{l}v_{i,j}^{l}-\sum_{k=1}^{K}\sum_{m,n}h_{m,n}^{k}\left(\sum_{l=1}^{L}(\tilde{W}^{k,l}\times\mathbf{v}^{l})_{m,n}\right)-\sum_{k=1}^{K}\sum_{m,n}b^{k}h_{m,n}^{k} \tag{47}$$

其中, $\mathbf{v}\in R^{N_v\times N_v\times L}$, $h\in R^{N_h\times N_h\times K}$, $N_v\times N_v$ 是输入图像的尺寸,L 表示输入图像的通道数,K 表示滤波器的数目,$W^{k,l}\in R^{ws\times ws}$ 为一个 ws 尺寸的滤波器,$\tilde{W}$ 表示矩阵 W 的翻转矩阵.那么,CDBNs 的激活函数可以表示为

$$P(h_{m,n}^{k}=1\mid v)=\text{sigmoid}\left(\sum_{l=1}^{L}(\tilde{W}^{k,l}\times\mathbf{v}^{l})_{m,n}+b^{k}\right) \tag{48}$$

$$P(v_{i,j}^{l}=1\mid h)=\text{sigmoid}\left(\sum_{k=1}^{K}(W^{k,l}\times h^{k})_{i,j}+c^{l}\right) \tag{49}$$

Lee 等人为了实现基于 CDBNs 的图像重构,提出了概率最大池化方法(probabilistic max-pooling),输入图像经过卷积运算,得到的卷积层的输出为 h^k,h^k 可以被划分为 $C\times C$ 大小的图像块,每个图像块α对应一个二值的池化单元 p_α^k,那么,池化层 p^k 的尺寸可以表示为 $N_p=N_h/C$.从直观上看,在概率最大池化中,当池化层单元激活时,有且仅有一个对应的α中的单元激活,当池化层单元灭活时,α中所有单元都不激活.基于最大概率池化理论,CDBNs 的能量函数可以表示如下:

$$\begin{aligned}&E(v,h)=-\sum_{l=1}^{L}\sum_{i,j}c^{l}v_{i,j}^{l}-\sum_{k=1}^{K}\sum_{m,n}h_{m,n}^{k}\left(\sum_{l=1}^{L}(\tilde{W}^{k,l}\times\mathbf{v}^{l})_{m,n}\right)-\sum_{k=1}^{K}\sum_{m,n}b^{k}h_{m,n}^{k}\\&\text{subject to }\sum_{(m,n)\leqslant\alpha}h_{m,n}^{k}\leqslant 1,\quad\forall k,\alpha\end{aligned} \tag{50}$$

基于能量函数,CDBNs 的条件激活概率可以表示为

$$P(h_{m,n}^{k}=1\mid v)=\frac{\exp\left(I(h_{mn}^{k})\right)}{1+\sum_{(m',n')\in\alpha}\exp\left(I(h_{m'n'}^{k})\right)} \tag{51}$$

$$P(p_{\alpha}^{k}=0\mid v)=\frac{1}{1+\sum_{(m',n')\in\alpha}\exp\left(I(h_{m'n'}^{k})\right)} \tag{52}$$

其中, $I(h_{mn}^{k})=\sum_{l=1}^{L}(\tilde{W}^{k,l}\times\mathbf{v}^{l})_{m,n}+b^{k}$, 可见层单元的激活形式与之前的 CDBNs 一致,基于概率最大池化,CDBNs 可以有效地利用网络的层次化结构学习图像逐层抽象的特征,并完成图像的生成工作.

3.4 RBMs与神经网络结合的总结和展望

目前常用的生成模型包括 VAEs 和 GANs 等,常用的判别模型为 CNNs 等,将 RBMs 作为预训练模型应用在 CNNs 中,能够使 CNNs 既可以用于图像识别也可以用于图像生成,且 RBMs 可以为 CNNs 提供更有效的初始化权值,从而促进 CNNs 收敛到更加优秀的局部最优解.但是将 RBMs 作为预训练算法也存在一些问题,首先,RBMs 作为无监督学习算法,并不能保证其特征表达是有利于分类的,随着神经网络层数的增加,使用 RBMs 作为预训练对分类精度带来的提升会越来越不明显,且预训练会非常耗时.如何改变 RBMs 的能量函数和损失函数,从而使 RBMs 得到的特征更有利于多层 CNNs 的分类任务,是 RBMs 未来研究的一个重点问题.其次,作为生成模型,虽然 RBMs 可以有效地与 VAEs 和 GANs 结合,但是作为生成模型本身,RBMs 难以扩展其深度,由于 RBMs 的训练需要采用近似算法,其计算复杂度很高,同样深度下,RBMs 的训练复杂度要远大于 VAEs 和 GANs.如何改进 RBMs 的训练算法和 RBMs 的网络结构,从而扩展 RBMs 的深度,构建更加有效的生成模型也是 RBMs 研究的重点和难点.

4 总结与展望

本文综述了 RBMs 和神经网络在理论研究和应用中的进展.在过去十年中,深度学习逐渐成为人工智能研究的主流方向,许多学者致力于该领域,并将概率图模型应用到深度学习中.目前已有大量研究结果证明了

RBMs 模型的有效性.然而,仍存在一些值得进一步研究的问题:RBMs 模型的算法理论问题需要进一步研究,如缓解 RBMs 中过拟合的方法、加快 RBMs 模型的训练以及提高 RBMs 模型建模实值数据的能力.Carlson 等学者发现,RBMs 的目标函数由 Shatten-∞范数限定,并提出了在赋范空间中更新参数的 SSD 算法.目前常用的缓解过拟合问题的方法有:权值衰减、Dropout 方法、DropConnect 方法和 Weight-uncertainty 方法等.如何获得图像处理中有效的抽象化特征也是 RBMs 研究的重点.已知 RBMs 的特征表达可以结合 CRFs 应用到图像分割和标注中.相反地,CRFs 中的图像分割和标记结果是否也可用于 RBMs 的特征提取中,以提高特征表达的能力?这也是我们今后的研究中关注的问题.目前除了向量神经网络(capsule nets)的训练方式不同外,神经网络的训练是基于 BP 算法的,其特征表示和特征学习仍然是一种黑箱的形式.这个问题也为基于梯度的 RBMs 算法带来了相同的困扰.如何在 RBMs 模型中引入新的训练方式也是接下来我们研究的重点.

References:

[1] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning. MIT Press, 2009.

[2] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786):504–507.

[3] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18:1527–1554.

[4] Hinton GE. Products of experts. In: Proc. of the Int'l Conf. on Artificial Neural Networks. 1999,1:1–6.

[5] Hinton GE. A practical guide to training restricted Boltzmann machines. In: Neural Networks: Tricks of the Trade. Berlin, Heidelberg: Springer-Verlag, 2012. 599–619.

[6] Ravanbakhsh S. Learning in Markov random fields using tempered transitions. In: Advances in Neural Information Processing Systems. 2009. 1598–1606.

[7] Welling M, Rosen-Zvi M, Hinton G. Exponential family harmoniums with an application to information retrieval. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2004. 1481–1488.

[8] Ravanbakhsh S, Poczos B, Schneider J, *et al*. Stochastic neural networks with monotonic activation functions. arXiv: 1601.00034v4, 2016. 573–577.

[9] Osindero S, Hinton G. Modeling image patches with a directed hierarchy of Markov random fields. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2007. 1121–1128.

[10] Larochelle H, Bengio Y, Louradour J, *et al*. Exploring strategies for training deep neural networks. Journal of Machine Learning Research, 2009,1(10):1–40.

[11] Salakhutdinov R, Hinton GE. Deep Boltzmann machines. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2009. 448–455.

[12] Salakhutdinov R, Hinton GE. An efficient learning procedure for deep Boltzmann machines. Neural Computation, 2012,24(8): 1967–2006.

[13] Hinton GE, Salakhutdinov R. A better way to pretrain deep Boltzmann machines. In: Advances in Neural Information Processing Systems. 2012,3:2447–2455.

[14] Goodfellow I, Mirza M, Courville A, Bengio Y. Multi-prediction deep Boltzmann machines. In: Advances in Neural Information Processing Systems. 2013. 548–556.

[15] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002,14(8):1711–1800.

[16] Jordan MI, Ghahramani Z, Jaakkola TS, *et al*. An introduction to variational methods for graphical models. Machine Learning, 1999,37(2):183–233.

[17] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984,6(6):721–741.

[18] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: Proc. of the Int'l Conf. on Machine Learning. Helsinki, 2008. 1064–1071.

[19] Tieleman T, Hinton GE. Using fast weights to improve persistent contrastive divergence. In: Proc. of the Annual Int'l Conf. on Machine Learning. Montreal, 2009. 1033–1040.

[20] Desjardins G, Courville A, Bengio Y, *et al*. Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines. Technical Reprot, 1345, University of Montreal, 2009.

[21] Desjardins G, Courville A, Bengio Y, *et al*. Parallel tempering for training of restricted Boltzmann machines. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. 2010. 145–152.

[22] Cho KH, Raiko T, Ilin A. Parallel tempering is efficient for learning restricted Boltzmann machines. In: Proc. of the Int'l Joint Conf. on Neural Networks. 2010. 605–616.

[23] Salakhutdinov R. Learning in Markov random fields using tempered transitions. In: Advances in Neural Information Processing Systems. 2009. 1598–1606.

[24] Welling M, Hinton GE. A new learning algorithm for mean field Boltzmann machines. In: Proc. of the Int'l Conf. on Artificial Neural Networks. Springer-Verlag, 2002. 351–357.

[25] Montavon G, Müller K, Cuturi M. Wasserstein training of restricted Boltzmann machines. In: Advances in Neural Information Processing Systems. 2017.

[26] Fisher C, Smith A, Walsh J. Boltzmann encoded adversarial machines. arXiv: 1804.08682, 2018.

[27] Krizhevsky A. Learning multiple layers of features from tiny images [MS. Thesis]. Department of Computer Science, University of Toronto, 2009.

[28] Cho KH, Ilin A, Raiko T. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines. In: Proc. of the Int'l Conf. on Artificial Neural Networks. Berlin, Heidelberg: Springer-Verlag, 2011. 10–17.

[29] Ranzato M, Krizhevsky A, Hinton GE. Factored 3-way restricted Boltzmann machines for modeling natural images. Journal of Machine Learning Research, 2010,9:621–628.

[30] Ranzato M, Hinton GE. Modeling pixel means and covariances using factorized third-order Boltzmann machines. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2010. 2551–2558.

[31] Courville A, Bergstra J, Bengio Y. A Spike and Slab restricted Boltzmann machine. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, 2011. 233–241.

[32] Courville AC, Bergstra J, Bengio Y. Unsupervised models of images by Spike and-Slab RBMs. In: Proc. of the Int'l Conf. on Machine Learning. Washington, 2011. 1145–1152.

[33] Goodfellow IJ, Courville A, Bengio Y. Spike-and-Slab sparse coding for unsupervised feature discovery. arXiv Preprint arXiv: 1201.3382, 2012.

[34] Huang. H, Toyoizumi. T. Advanced mean-field theory of the restricted Boltzmann machine. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2015,91(5).

[35] Goodfellow IJ, Courville A, Bengio Y. Large-scale feature learning with Spike-and-Slab sparse coding. In: Proc. of the Int'l Conf. on Machine Learning. Edinburgh, 2012.

[36] Courville A, Desjardins G, Bergstra J, *et al*. The Spike-and-Slab RBM and extensions to discrete and sparse data distributions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014,36(9):1874–1887.

[37] Kuleshov V, Ermon S. Neural variational inference and learning in undirected graphical models. In: Advances in Neural Information Processing Systems. 2017.

[38] Nair V, Hinton G. Rectified linear units improve restricted Boltzmann machines. In: Proc. of the Int'l Conf. on Machine Learning. 2010. 807–814.

[39] Yang E, Ravikumar P, Allen G, *et al*. Graphical models via generalized linear models. In: Advances in Neural Information Processing Systems. 2012.

[40] Tran T, Phung DQ, Venkatesh S. Mixed-variate restricted Boltzmann machines. In: Proc. of the Asian Conf. on Machine Learning. 2011. 213–229.

[41] Nguyen TD, Tran T, Phung D, *et al*. Latent patient profile modelling and applications with mixed-variate restricted Boltzmann machine. In: Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining. 2013. 123–135.

[42] Tran T, Phung DQ, Venkatesh S. Cumulative restricted Boltzmann machines for ordinal matrix data analysis. In: Proc. of the Asian Conf. on Machine Learning. 2012. 411–426.

[43] Tran T, Phung DQ, Venkatesh S. Thurstonian Boltzmann machines: Learning from multiple inequalities. In: Proc. of the Int'l Conf. on Machine Learning. 2013. 46–54.

[44] Feng F, Li R, Wang X. Deep correspondence restricted Boltzmann machine for cross-modal retrieval. Neurocomputing, 2015,154: 50–60.

[45] Zhao F, Huang Y, Wang L, *et al*. Learning relevance restricted Boltzmann machine for unstructured group activity and event understanding. Int'l Journal of Computer Vision, 2016,119(3):329–345.

[46] Larochelle H, Mandel M, Pascanu R, *et al*. Learning algorithms for the classification restricted Boltzmann machine. Journal of Machine Learning Research, 2012,13(1):643–669.

[47] Lee T, Yoon S. Boosted categorical restricted Boltzmann machine for computational prediction of splice junctions. In: Proc. of the Int'l Conf. on Machine Learning. 2015.

[48] Chen CLP, Zhang CY, Chen L, *et al*. Fuzzy restricted Boltzmann machine for the enhancement of deep learning. IEEE Trans. on Fuzzy Systems, 2015,23(6):2163–2173.

[49] Johnson MJ, Duvenaud D, Wiltschko AB, *et al*. Composing graphical models with neural networks for structured representations and fast inference. arXiv: 1603.06277, 2016.

[50] Lee H, Grosse R, Ranganath R, Ng AY. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2009. 609–616.

[51] Lin M, Chen Q, Yan S. Network in network. arXiv: 1312.4400.

[52] Norouzi M, Ranjbar M, Mori G. Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 2735–2742.

[53] Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in Neural Information Processing Systems. 2009. 1096–1104.

[54] Lee H, Grosse R, Ranganath R, Ng AY. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM, 2011,54(10):95–103.

[55] Chen L, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Computer Science, 2014,(4):357–361.

[56] Hinton GE. To recognize shapes, first learn to generate images. Progress in Brain Research, 2007,165(6):535–547.

[57] Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines. In: Proc. of the Int'l Conf. DBLP, 2008.

[58] Carlson D, Cevher V, Carin L. Stochastic spectral descent for restricted Boltzmann machines. In: Proc. of the Int'l Conf. on Artificial Intelligence and Statistics. San Diego, 2015.

[59] Telgarsky M. Representation benefits of deep feedforward networks. Computer Science, 2015,15(8):1204–1211.

[60] Chui CK, Li X, Mhaskar HN. Neural networks for localized approximation. Mathematics of Computation, 1994,63(208):607–623.

[61] Eldan R, Shamir O. The power of depth for feedforward neural networks. In: Proc. of the Annual Conf. on Learning Theory. 2016. 907–940.

[62] Shaham U, Cheng X, Dror O, Jaffe A, *et al*. A deep learning approach to unsupervised ensemble learning. arXiv Preprint arXiv: 1602.02285, 2016.

[63] Djork-Arné C, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). Computer Science, 2015.

[64] Klambauer G, Unterthiner T, Mayr A, *et al.* Self-normalizing neural networks. In: Proc. of the NIPS. 2017.

[65] Srivastava RK, Greff K, Schmidhuber J. Highway networks. Computer Science, 2015.

[66] He KM, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. [doi: 10.1109/CVPR.2016.90]

[67] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the 32nd Int'l Conf. on Machine Learning. 2015. 448–456.

[68] Srivastava N, Hinton GE, Krizhevsky A, *et al*. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014,15:1929–1958.

[69] Wan L, Zeiler M, S. Zhang, *et al*. Regularization of neural networks using dropconnect. In: Proc. of the Int'l Conf. on Machine Learning. 2013. 1058–1066.

[70] Zhang N, Ding SF, Zhang J, Xue Y. Research on point-wise gated deep networks. Applied Soft Computing, 2017,52:1210–1221.

[71] Zhang J, Ding SF, Zhang N, Xue Y. Weight uncertainty in Boltzmann machine. Cognitive Computation, 2016,8(6):1064–1073.

[72] Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. Computer Science, 2015,5(1):36.

[73] Chung J, Gulcehre C, Cho KH, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling. Eprint Arxiv, 2014.

张健(1990－),男,山东泰安人,博士生,主要研究领域为深度学习,玻尔兹曼机.

丁世飞(1963－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为人工智能,模式识别,机器学习,数据挖掘.

张楠(1991－),男,博士生,CCF 学生会员,主要研究领域为机器学习,玻尔兹曼机.

杜鹏(1994－),男,硕士生,主要研究领域为深度学习,数据挖掘.

杜威(1994－),男,硕士生,主要研究领域为深度学习,强化学习.

于文家(1994－),男,硕士生,主要研究领域为深度学习,生成对抗网络.