

目录

目录	1
1 基于深度嵌入卷积聚类方法的地海杂波无监督分类	1
1.1 引言	1
1.2 卷积自编码网络	3
1.2.1 自编码器	3
1.2.2 卷积自编码器	4
1.3 深度嵌入卷积聚类方法	5
1.3.1 深度嵌入卷积聚类方法结构	5
1.3.2 KL 散度聚类方法	6
1.4 仿真验证	7
1.4.1 参数设置	7
1.4.2 仿真结果分析	7
1.5 小结	8
参考文献	9

1 基于深度嵌入卷积聚类方法的地海杂波无监督分类

1.1 引言

由于有监督学习需要对大量的数据进行打标签，这个过程通常需要花费大量的时间与精力。无监督学习也即聚类算法，由于其不需要数据标签，因此国内外学者对各个领域的聚类算法进行了广泛研究，产生了各种不同类型的聚类算法，例如层次聚类 [1]，基于质心的聚类方法 [2]、基于图的聚类方法 [3]、基于回归模型 [4] 的聚类方法以及基于子空间 [5] 的聚类方法。另一方面，这些聚类方法一般可以分为两类，即生成式和判别式聚类算法。像 K 均值和高斯混合模型 [6] 这样的生成算法使用特征空间的几何特性来明确地表示簇，并且通过输入数据的统计分布来对分类进行建模。与生成式聚类算法不同，判别式方法直接使用分离超平面来识别类别，而不管数据的分布是怎样的。信息论 [7]，支持向量机 [8] 和图谱论 [9] 算法是判别式聚类模型的例子。一般认为，判别式模型与其对应的生成式模型相比往往具有更好的结果，因为他们对数据分布的假设较少，直接将聚类分开，但是他们的训练可能会出现过拟合或陷入局部最优解的问题出现 [10]。我们的 DECC 算法也是一种判别式聚类算法，但它受益于自编码器的辅助重构任务，以缓解我们判别式聚类算法的训练中的这个问题。

虽然，聚类问题已经在各种领域进行了应用研究，但是传统的聚类算法的性能收到数据维数以及数据量的影响，会产生维数灾难的问题。为了解决计算复杂度的问题，以往的研究往往首先将数据投影到一个低维流形中，然后将嵌入的数据聚类到这个新的子空间中 [11]。处理大规模的数据集，也有一些研究只选择数据点的一个子集来加速聚类过程 [12]。

就要在特征空间中数据的局部属性而言，基本的思路通过向目标函数添加先验知识来考虑稀疏性或图形约束 [13]。这类算法可以分为两个阶段：特征学习和聚类。

特征学习指的是一类试图利用一组基本向量或特征来描述一个数据集的学习方法，并且通常这个表示是稀疏的。在实践中，有很多不同的算法可以进行特征学习，包括自编码器、k-均值聚类、高斯混合模型和受限玻尔兹曼机 (RBM) [14]。这些方法都倾向于学习类似的局部滤波器字典 [14]，例如用于自然图像的 Gabor 类边缘滤波器，或者用于 MNIST 数字数据集的书写笔划。虽然 RBM 是可以从数据生成分布中抽样的生成模型 [15]，但是自编码器被训练来优化它们对输入数据的重构。

自编码器 (Autoencoder, AE) 是一种简单的神经网络, 旨在尽可能减少失真的情况下将输入转换成输出。虽然在概念上简单, 但其在机器学习中起着重要的作用。自编码器是 20 世纪 80 年代由 Hinton 和 PDP 小组 [16] 首次提出的, 用输入数据作为标签来解决无监督反向传播的问题。自编码器与 Hebbian 学习规则 [17] 一起为无监督学习提供了一个基本的范例, 并开始解决如何以自组织的方式协调局部事件引起的变化去产生全局学习和智能表现。

最近, 自编码器在深度架构方法 [18, 19, 20, 21] 中再次占据了中心位置, 特别是形式上类似受限玻尔兹曼机的自编码器, 以堆栈形式组织, 然后以无监督的方式自下而上地进行训练, 随后是以有监督学习的方式训练顶层并微调整个架构。考虑到自下而上训练阶段对于最终的任务是不可知的, 因此显然可以用于迁移学习方法。并且实际的实验已经证明这些深层次的体系结构可以在一些具有挑战性的分类和回归问题上得到最优的结果。

尽管广大学者已经产生了很大的兴趣, 但除了文献 [22, 23, 24] 等, 对自编码器和深层架构的理论认识的文章还很少。除此之外, 使用深度一词可能会造成更多混淆。从计算机科学的角度来看, 深层结构应该对于一些小的 $\alpha > 0$ 有 $n\alpha$ 个多项式大小的层, 其中 n 是输入向量的大小 [25]。但是在 Hinton 等人 [18, 19] 描述的架构中并不是这种情况, 似乎有恒定或最好的对数深度。对于计算机视觉、语音识别和其他典型问题, 有限和对数的深度之间的区别很小, 几乎可以忽略。文献 [26] 对自编码器的理论进行了描述, 提出了一个用自编码器解决线性与非线性问题的通用的数学框架。

后来, 文献 [27] 提出了共同完成特征学习和聚类的算法。深度嵌入聚类算法 [27] 以自学习的方式定义了一个有效的目标函数。定义的聚类损失函数用于同时更新网络和聚类中心的变换参数。但是, 他们忽略了数据属性的保持性, 这可能导致特征空间的破坏。我们通过保留数据生成分布的局部结构并结合卷积层来改进 DEC 算法。深度聚类算法中使用最广泛的神经网络是堆栈自编码器 [27]。SAE 需要逐层预训练, 然后以端对端 (end-to-end) 的方式进行微调。当层次更深时, 预训练过程是十分耗时的。此外, SAE 是建立在全连接的层次上, 对于处理局部信息是无效的。文献 [28] 的工作是第一个在不需要预训练的情况下直接以端到端的方式直接训练 CAE 的试验。

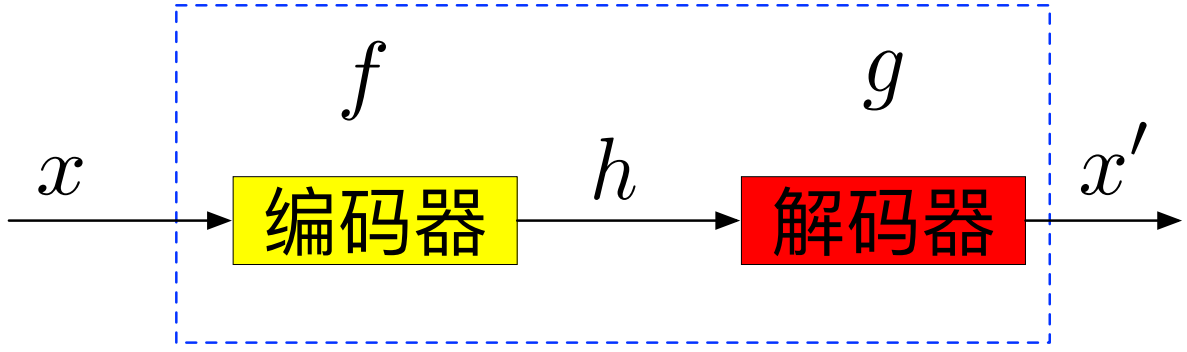


图 1-1 自编码器模型

1.2 卷积自编码网络

1.2.1 自编码器

自编码器是一种利用隐层用来重构输入的神经网络。其模型如图 1-1所示，它由编码器 (Encoder) 和解码器 (Decoder) 两部分组成，本质上是对输入信号的某种变换。编码器将输入信号 $x \in [0, 1]^{n_x}$ 编码到某个隐层的表示 $h \in [0, 1]^{n_h}$ ，然后解码器再将 h 解码为输出 $x' \in [0, 1]^{n_x}$ 。通过最小化 x 与 x' 之间的差距，可以训练得到一个特征表示的映射 h ，其可以用来重构输入。编码和解码公式如下：

$$h_j = \text{sigm}(b_j + \sum_{i=1}^{n_x} w_{ij}x_i) \quad (1-1)$$

$$x'_i = \text{sigm}(a_i + \sum_{j=1}^{n_h} w'_{ij}h_j) \quad (1-2)$$

其中， n_x 和 n_h 分别为输入表示和隐层表示的维数， $\text{sigm}(x) = \frac{1}{1+e^{-x}}$ 是 logistic 损失函数， $\mathbf{W} = [w_{ij}]$ 和 $\mathbf{W}' = [w'_{ij}]$ 为权重矩阵， $\mathbf{a} = [a_i]$ 和 $\mathbf{b} = [b_j]$ 为偏差向量。

对于一组具有 N 个输入向量的给定训练数据集，设每个训练向量为 $x^{(n)}$ 可以被映射到一个隐藏层表示 $h^{(n)}$ ，其重构表示为 $x'^{(n)}$ 。模型参数 $\Theta = \mathbf{W}, \mathbf{a}, \mathbf{b}$ 通过最小化其损失函数进行求取，一般利用均方误差损失函数 (mean squared error, mse)：

$$J(\theta) = \frac{1}{N} \sum_{n=1}^N \|x^{(n)} - x'^{(n)}\|_2^2 \quad (1-3)$$

$$\theta^* = \arg \min_{\theta} J(\theta) \quad (1-4)$$

通常，使用随机梯度下降 (SGD) 或其余基于梯度的优化算法来对公式 1-4求解。

在此基础上，学者对自编码器进行了各种改进，提出了可以学习输入空间中扰动不变的特征的压缩自编码器 [29]、提供有效的变分方法进行训练和生成推理的变分自编码

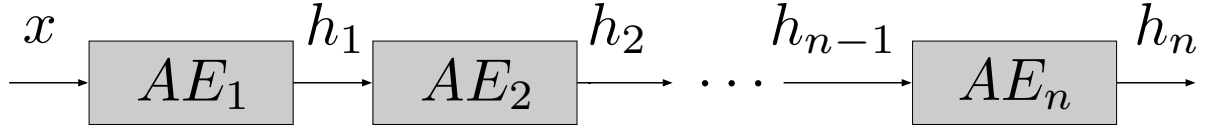


图 1-2 堆栈自编码器模型

器 [30] 以及可以基于连续的数据流在运行中添加或合并隐藏单元的在线增量自编码器 [31]。

1.2.2 卷积自编码器

传统的自编码网络只有一个隐藏层，只能学习出一种特征变化，于是 Bengio[32] 等人在 2007 年的仿照 stacked RBM 构成的 DBN，提出堆栈自编码器（Stacked Auto-Encoder, SAE），为非监督学习在深度网络的应用又添了猛将。堆栈自编码器的基本模型可以参看图 1-2，其通过逐层非监督学习的预训练，也即逐层初始化（Layer-wise Pre-training），来初始化深度网络的参数，替代传统的随机初始化的方法。预训练完毕后，利用训练参数，再进行监督学习训练。

全连接自编码器和深度自编码器都忽略了数据的局部结构信息。这在处理实际大小的输入时不仅是一个问题，而且在参数中引入冗余，迫使每个特征是全局的。然而，在视觉和对象识别领域最成功的模型 [33] 均主要利用了在整个输入空间发生重复的局部特征。基于此，文献 [34] 提出了一种卷积自编码器（Convolutional Auto-Encoders, CAE）。卷积自编码器不同于传统的自编码器，它们的权重在输入向量中的所有位置之间共享，从而可以保持空间局部性。因此重建结果是隐藏局部特征的线性组合。

CAE 架构直观地类似于 AE 的架构，除了权重是共享的。对于输入 x ，第 k 个特征映射的隐藏表示由下式给出

$$h^k = \sigma(x * W^k + b^k) \quad (1-5)$$

A single bias per latent map is used, as we want each filter to specialize on features of the whole input (one bias per pixel would introduce too many degrees of freedom). 其中偏差也是全局共享的， σ 是激活函数（我们在我们所有的实验中使用缩放的双曲正切）， $*$ 表示二维卷积。因为我们希望每个滤波器专注于整个输入空间的特征，另一方面如果每个数据点一个偏差将引入太多的自由度，所以对每一个隐藏层使用相同的偏差。重建过程由下式计算

$$y = \sigma\left(\sum_{k \in H} h^k * \tilde{W}^k + c\right) \quad (1-6)$$

同样的对于每一个输入通道只有一个偏差 c 。 H 表示潜在特征图的组。 \tilde{W} 标识权重的两个维度上的翻转操作。公式 1-5和 1-6中的 2D 卷积操作由上下文决定。一个 $m \times m$ 矩阵和一个 $n \times n$ 矩阵的卷积结果可能是 $(m+n-1) \times (m+n-1)$ (full convolution) 也可能是 $(m-n+1) \times (m-n+1)$ (valid convolution)。其损失函数为最小均方误差 (mean squared error, MSE)

$$E(\theta) = \frac{1}{2n} \sum_{i=1}^n (x_i - y_i)^2 \quad (1-7)$$

就像标准的神经网络一样，反向传播算法被用来计算相对于参数的误差函数的梯度。这可以通过使用以下公式的卷积运算容易地得到：

$$\frac{\partial E(\theta)}{\partial W^k} = x * \delta h^k + \tilde{h}^k * \delta y \quad (1-8)$$

δh 和 δy 分别是隐藏层状态和重构结果的变化量。然后使用随机梯度下降来更新权重。

我们的主要思想是，CAE 有利于学习数据的局部结构，避免特征空间的失真。

我们介绍了一种用于分层特征提取的无监督方法，卷积自编码器。虽然 CAE 的完备隐藏表示使得学习比标准自编码器更难，但是我们可以通过添加池化层加强其稀疏性而不需要通过反复试验来设置正则化参数。

1.3 深度嵌入卷积聚类方法

深层嵌入式聚类 (Deep Embedding Clustering, DEC) [27] 算法提供了一种无监督的方式学习特征表示和聚类的思路，通过联合优化神经网络和聚类中心提高了聚类算法的性能和鲁棒性。但是，该算法由于利用了堆栈自编码器，无法利用数据的局部特征信息，因此本章在此基础上提出了深度嵌入卷积聚类方法 (Deep Embedding Convolution Clustering, DECC)，利用卷积自编码器代替传统的堆栈自编码器。

1.3.1 深度嵌入卷积聚类方法结构

首先对于问题进行形式化描述，假设需要将 N 个样本 $X = [x_1, \dots, x_N]$ 聚为 K 个类别，其聚类中心点为 μ_1, \dots, μ_K ，其中每一个样本 $x_i \in \mathbb{R}^{d_x}$ ，由于对于维数较大的样本会产生维数灾难的问题，因此需要引入一个嵌入函数 $\varphi_W : X \rightarrow Z$ ，可以将原始样本映射到嵌入子空间 $Z = [z_1, \dots, z_N]$ ，其中 $z_i \in \mathbb{R}^{d_z}$ 的维数远小于原始样本的维数，也即 $d_z \ll d_x$ 。

本章利用卷积自编码器作为嵌入函数对原始样本进行降维，其基本结构如图 xxx 所示。通过三次卷积操作充分提取输入样本的特征，我们在所有层上均利用 ReLU 激活

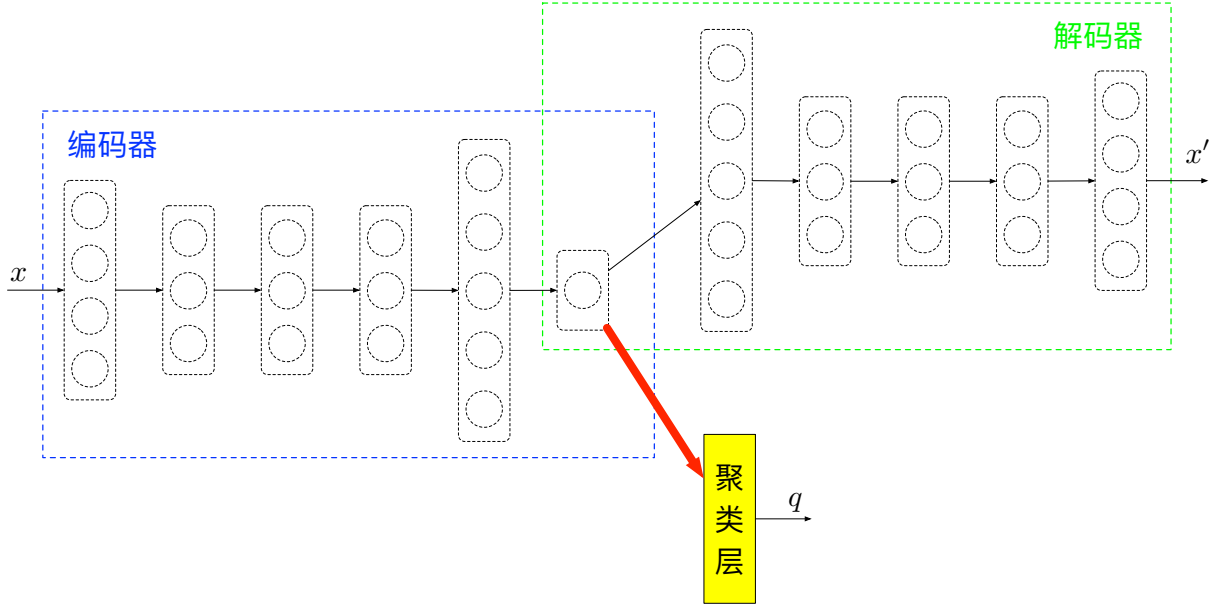


图 1-3 深度嵌入卷积自编码器结构图

函数，并且在编码器后以相反的顺序放置解码器，由于编码器为卷积操作，故解码器为逆卷积 (Deconvolution)。在编码器与解码器中间添加嵌入层，该层的结果被用于进行聚类操作。DECC 的结构如图 1-3 所示。

1.3.2 KL 散度聚类方法

文献 [35] 利用学生 t 分布来衡量嵌入点 z_i 与聚类中心点 μ_k 的相似度：

$$p_{ik} = \frac{(1 + \|z_i - \mu_k\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'} (1 + \|z_i - \mu_{k'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad (1-9)$$

其中， α 为学生 t 分布的自由度， p_{ik} 可以看作样本 i 到类别 k 的概率。

为了定义我们的聚类目标函数，我们使用辅助目标变量 (auxiliary target variable) Q 来迭代地改进模型预测。为此，我们首先使用 Kullback-Leibler (KL) 散度来减小模型预测 P 和目标分布 Q 之间的距离。

$$\mathcal{L}_{kl} = KL(Q||P) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K q_{ik} \log \frac{q_{ik}}{p_{ik}} \quad (1-10)$$

根据文献 [27]，目标分布 Q 的定义如下：

$$q_{ij} = \frac{p_{ij}^2 / f_k}{\sum_k (p_{ik}^2 / f_k)} \quad (1-11)$$

其中 f_k 的定义为

$$f_k = \sum_i q_{ik} \quad (1-12)$$

公式 1-4给出了自编码器重构的损失函数，故最终的损失函数为：

$$\mathcal{L} = \mathcal{L}_{mse} + \beta \mathcal{L}_{kld} \quad (1-13)$$

其中， $\beta > 0$ 为调节两部分损失函数的权重。

1.4 仿真验证

1.4.1 参数设置

本章根据如图设计的卷积自编码器进行特征的降维，该部分进行 30 次迭代，学习方法为 Adam。然后利用随机 20 次 k 均值算法然后从中选取最优结果，作为起始的聚类中心点，此也作为我们的一个对比算法。损失函数的权重系数 $\beta = 0.1$ ，学生 t 的自由度 $\alpha = 1$ 。最终的优化问题可以通过小批量随机梯度下降和反向传播得到有效解决。

1.4.2 仿真结果分析

我们利用第三章中组 C 的训练数据用于本章的实验。其实验结果如表 1-1所示。

表 1-1 辐射源信号数据

方法	无监督聚类精度	时间
CAE + k 均值	66%	12
DECC ($\beta = 0$)	23%	12
DECC	453%	12

此处无监督聚类精度 (unsupervised clustering accuracy, ACC) 的定义为：

$$ACC = \max_m \frac{\sum_{i=1}^n g(l_i, m(c_i))}{n} \quad (1-14)$$

$$g(l_i, m(c_i)) = \begin{cases} 1 & l_i = m(c_i) \\ 0 & l_i \neq m(c_i) \end{cases}$$

其中， l_i 为真实的标签， c_i 为算法的聚类结果， m 是聚类结果与实际标签类别之间的一一对应，其遍历所有可能假设。

1.5 小结

本章结合卷积神经网络与自编码器提出了深度嵌入卷积聚类方法 DECC。其通过在损失函数中将 KL 散度与重构损失两部分结合,使得本章的算法对于两部分可能都进行了考虑。

虽然与有监督学习算法相比,本章的算法精度仍然不够高,但是相比于传统的聚类算法精度有了一定的提高。在一定程度上可以减少打标签的过程。

本章的贡献是:

- 可以以端对端方式训练的卷积自编码器被设计用于从未标记的数据中学习特征。通过在杂波数据中包含空间、时间信息,设计了好于传统堆栈自编码器的卷积自编码器。实验结果证明卷积层、卷积转置层和完全连通层足以构建一个有效的 CAE。
- 在根据聚类损失函数调整网络参数时考虑局部结构特征。实验结果证明,利用局部结构特征有助于稳定训练过程。
- 我们提出了深度嵌入卷积聚类算法来自动聚类地海杂波。DECC 利用 CAE 提取局部特征的优势。
- 利用天波超视距雷达的实际地海杂波数据进行了实验,结果验证了 DECC 算法的有效性。

参考文献

- [1] Heller KA, Ghahramani Z. Bayesian hierarchical clustering[C]. In Proceedings of the 22nd international conference on Machine learning. ACM2005:297–304.
- [2] Lloyd S. Least squares quantization in PCM[J]. IEEE transactions on information theory. 1982, 28(2):129–137.
- [3] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transactions on pattern analysis and machine intelligence. 2000, 22(8):888–905.
- [4] Wang H, Nie F, Huang H. Multi-view clustering and feature learning via structured sparsity[C]. In International Conference on Machine Learning. 2013:352–360.
- [5] Agrawal R et al. Automatic subspace clustering of high dimensional data for data mining applications[M]vol. 27. ACM. 1998.
- [6] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood[J]. IEEE transactions on pattern analysis and machine intelligence. 2000, 22(7):719–725.
- [7] Li H, Zhang K, Jiang T. Minimum entropy clustering and applications to gene expression analysis[C]. In Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE. IEEE2004:142–151.
- [8] Xu L et al. Maximum margin clustering[C]. In Advances in neural information processing systems. 2005:1537–1544.
- [9] Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm[C]. In Advances in neural information processing systems. 2002:849–856.
- [10] Raina R et al. Classification with hybrid generative/discriminative models[C]. In Advances in neural information processing systems. 2004:545–552.
- [11] Roth V, Lange T. Feature selection in clustering problems[C]. In Advances in neural information processing systems. 2004:473–480.

-
- [12] Shinnou H, Sasaki M. Spectral Clustering for a Large Data Set by Reducing the Similarity Matrix Size.[C]. In LREC. 2008.
- [13] Tian F et al. Learning Deep Representations for Graph Clustering.[C]. In AAAI. 2014:1293–1299.
- [14] Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning[C]. In Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011:215–223.
- [15] Hinton G. A practical guide to training restricted Boltzmann machines[J]. Momentum. 2010, 9(1):926.
- [16] Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation[R]. Tech. rep.California Univ San Diego La Jolla Inst for Cognitive Science1985.
- [17] Hebb DO. The organization of behavior: A neuropsychological theory[M]. Psychology Press. 2005.
- [18] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks[J]. science. 2006, 313(5786):504–507.
- [19] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets[J]. Neural computation. 2006, 18(7):1527–1554.
- [20] Bengio Y, LeCun Y et al. Scaling learning algorithms towards AI[J]. Large-scale kernel machines. 2007, 34(5):1–41.
- [21] Erhan D et al. Why does unsupervised pre-training help deep learning?[J]. Journal of Machine Learning Research. 2010, 11(Feb):625–660.
- [22] Montufar G, Ay N. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines[J]. Neural Computation. 2011, 23(5):1306–1319.
- [23] Sutskever I, Hinton GE. Deep, narrow sigmoid belief networks are universal approximators[J]. Neural computation. 2008, 20(11):2629–2636.
- [24] Baldi P, Hornik K. Neural networks and principal component analysis: Learning from examples without local minima[J]. Neural networks. 1989, 2(1):53–58.

-
- [25] Clote P, Kranakis E. Boolean functions and computation models[M]. Springer Science & Business Media. 2013.
- [26] Baldi P. Autoencoders, unsupervised learning, and deep architectures[C]. In Proceedings of ICML Workshop on Unsupervised and Transfer Learning. 2012:37–49.
- [27] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]. In International Conference on Machine Learning. 2016:478–487.
- [28] Dundar A, Jin J, Culurciello E. Convolutional clustering for unsupervised learning[J]. arXiv preprint arXiv:151106241. 2015.
- [29] Rifai S et al. Contractive auto-encoders: Explicit invariance during feature extraction[C]. In Proceedings of the 28th International Conference on International Conference on Machine Learning. Omnipress2011:833–840.
- [30] Kingma DP, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:13126114. 2013.
- [31] Zhou G, Sohn K, Lee H. Online incremental feature learning with denoising autoencoders[C]. In Artificial Intelligence and Statistics. 2012:1453–1461.
- [32] Bengio Y et al. Greedy layer-wise training of deep networks[C]. In Advances in neural information processing systems. 2007:153–160.
- [33] Lowe DG. Object recognition from local scale-invariant features[C]. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Ieeevol. 21999:1150–1157.
- [34] Masci J et al. Stacked convolutional auto-encoders for hierarchical feature extraction[J]. Artificial Neural Networks and Machine Learning–ICANN 2011. 2011, :52–59.
- [35] Maaten Lvd, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research. 2008, 9(Nov):2579–2605.