

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such “autoencoder” networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer “encoder” network

深度学习经典论文分析（四）-Reducing the dimensionality of data with neural networks

CPJS

臭皮匠

每天分享一点点，每天进步一点点

关注

11 人赞同了该文章

收起

章节目录如下：

想要解决的问题

算法的梯度消失

神经网络构成的非线性降维

的是否是一个新问题

复消失的重要性

编码器

采用的新知识

良玻尔兹曼机

定义

目标

训练

toencoder

ep Autoencoder

FRBM的初始化

的关键人物与工作

建人物

jeoffrey Everest Hinton

yavid Rumelhart

Max Welling

imon Osindero

am Roweis

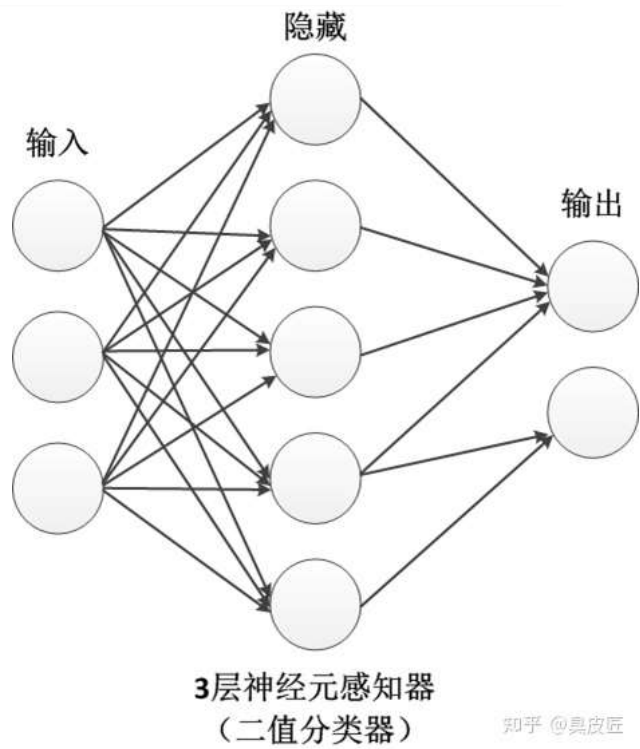
建工作

受限玻尔兹曼机

这个专栏主要是想和大家分享一下深度学习的一些经典论文，具体包含的论文见[目录](#)，在[github](#)中还包括了更丰富的信息，具体包括

1. 深度学习的经典论文-原文（Original paper）
2. 带注释的深度学习经典论文（paper with notes）
3. 论文笔记(notes)
4. 补充的一些其他论文(Supplementary knowledge)
5. 补充的一些知识点(Supplementary paper)

今天的这篇论文是由三巨头之一的Hinton在06年时于Science上发表的。此前，深度学习由于BP算法在1991年被指出存在梯度消失问题时而陷入低谷。当时还没有Adam等优化器，也还没有卷积神经网络，只有多层感知器（MLP，如下图）。到06年这是第一次提出有效解决梯度消失问题的解决方案：**无监督预训练对权值进行初始化+有监督训练微调**，并且重新让深度学习掀起浪潮。虽然现在很少人用这种方法进行初始化了，但本文的重要性仍然不言而喻。



本篇文章目录如下:

- 1 文章想要解决的问题
- 1.1 BP算法的梯度消失
 - 1.2 神经网络构成的非线性降维算法
- 2 研究的是否是一个新问题
- 2.1 梯度消失的重要性
 - 2.2 自编码器
- 3 本文采用的新知识
- 3.1 受限玻尔兹曼机
 - 3.1.1 定义
 - 3.1.2 目标
 - 3.1.3 训练
 - 3.2 Autoencoder
 - 3.3 Deep Autoencoder
 - 3.4 基于RBM的初始化
- 4 相关的关键人物与工作
- 4.1 关键人物
 - 4.1.1 Geoffrey Everest Hinton
 - 4.1.2 David Rumelhart
 - 4.1.3 Max Welling
 - 4.1.4 Simon Osindero
 - 4.1.5 Sam Roweis
 - 4.2 关键工作
 - 4.2.1 受限玻尔兹曼机
 - 4.2.2 PCA
- 5 文章提出的解决方案的关键



6 实验设计

- 6.1 随机二值曲线
- 6.2 MNIST手写数字
- 6.3 Olivetti人脸数据集
- 6.4 文本检索
- 6.5 分类及回归任务

7 采用的数据集及其是否开源

- 7.1 mnist数据集
- 7.2 Olivetti face data set

8 实验结果是否验证了科学假设

9 本文贡献

- 9.1 解决BP算法的缺点
- 9.2 提出DAE进行降维

10 下一步怎么做（本文局限性）

11 重要的相关论文

12 不懂之处

13 专业术语

- 13.1 受限玻尔兹曼机
- 13.2 自编码器
- 13.3 主成分分析
- 13.4 潜在语义分析

14 英语词汇

1 文章想要解决的问题

1.1 BP算法的梯度消失

由于BP算法的局限性（容易出现梯度消失，而且没有优化器），导致初始化的结果对网络的训练结果影响很大：

- 如果随机初始化的值过大，通常会找到较差的局部极小值。个人想法这是因为较大的权值会本身就很接近某些极值或鞍点，会导致BP算法可以“操作”的空间减小了。
- 如果随机初始化的值过小，会导致前几层的梯度过小而难以训练。

所以，本文从**初始化**的角度去解决训练结果不好的问题（相对的，优化器是从BP算法的角度优化）。

1.2 神经网络构成的非线性降维算法

在压缩输入向量的维度方面（降维），常用的是PCA算法（线性降维）。基于1986年由Rumelhart提出的**单层的**自编码器（autoencoder），本文提出一个新的降维算法：由一个非线性的、自适应的**多层的**编码器（encoder）来将高维的数据转化为低维的编码，和一个相似的解码器（decoder）来将低维编码恢复成原始高维数据，整个系统构成一个自编码器（autoencoder，但是称为DAE/deep autoencoder更合适，引自该博客）。该网络用BP算法进行训练，所以也存在上述问题。本文的具体目标就在于优化该算法。

2 研究的是否是一个新问题

不是。无论是梯度消失还是降维算法，都早已提出。梯度消失问题在1991年就被提出，直到06年这篇论文提出前都没有一个好的解决方法。1991年提出梯度消失的问题之后，深度学习就进入了寒冬，可见解决该问题有多重要。而且现今也仍然有人研究其他解决梯度消失/梯度爆炸的办法。

2.1 梯度消失的重要性

由于梯度消失的存在，深层神经网络会变得难以训练，甚至无法收敛。其根本原因在于反向传播训练法则，属于先天不足（引自梯度消失、爆炸原因及其解决方法知识搬运工的博客-CSDN博客梯度消失）。因此，怎么弥补这种缺点就是一个很重要的课题。

2.2 自编码器

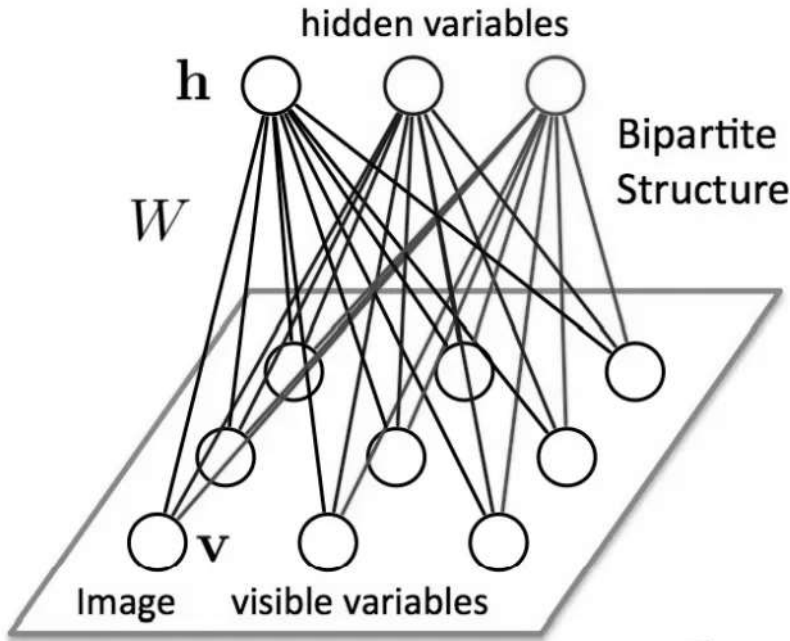
单层的自编码器在86年就已经提出，但是受限于梯度消失而无法使用深层。本文在提出新的初始化方法解决了梯度消失问题后，成功搭建了深层自编码器，并且性能要优于最常用的降维算法PCA。

3 本文采用的新知识

3.1 受限玻尔兹曼机

3.1.1 定义

RBM (Restricted Boltzmann machine) 也是论文作者Hinton提出的，是一种生成式随机神经网络(generative stochastic neural network)。该网络由一些可见单元(visible unit, 对应可见变量，亦即数据样本)和一些隐藏单元(hidden unit, 对应隐藏变量)构成，可见变量和隐藏变量都是二元变量，亦即其状态取{0,1}。整个网络是一个二部图，只有可见单元和隐藏单元之间才会存在边，可见单元之间以及隐藏单元之间都不会有边连接。



知乎 @臭皮匠

其中可见单元构成的向量称为 v ，隐藏单元构成的向量称为 h ，矩阵 W 的 $shape$ 为 $(length\ of\ v, length\ of\ h)$ 。其实实际计算时还有两个 $shape$ 分别与 v 、 h 一样的偏置 b ，但图中没有给出。

更简单一点，可以将 RBM 理解为 输入输出均为2值的单层神经网络，即用明确的 0/1 来表示特征有无。同时这个网络的 W 是“双向”的，即支持从可见层到隐藏层、又从隐藏层回到可见层。这样应该就能理解为什么需要两个偏置了。

RBM网络的目标在于学习两个概率函数



$$P(h_j = 1 | v) = \text{sigmoid}(b_j + W_{j,:} \cdot v)$$

$$P(v_i = 1 | h) = \text{sigmoid}(b_i + W_{:,i}^T h)$$

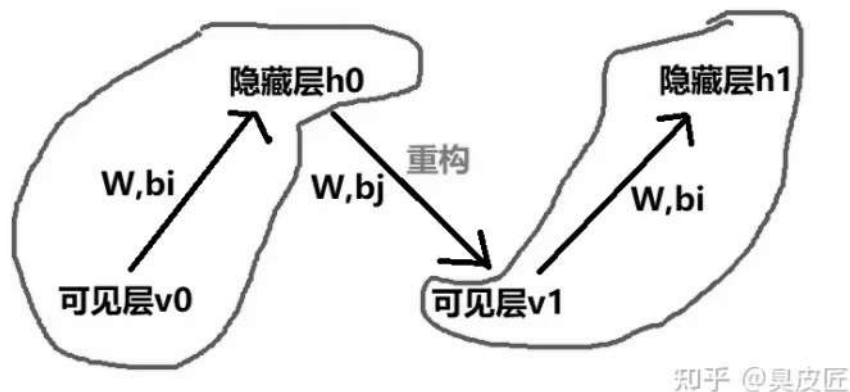
第一个函数决定了给定可见层 v （输入）可以得到什么样的隐藏层 h （并不唯一，因为给出的是概率）。第二个函数则是相反的过程。

RBM本质和神经网络一样，学习的也是一种特征提取/映射；而且也可以看成是一种**编码/解码**的过程，从可见层到隐藏层就是编码，而反过来从隐藏层到可见层就是解码（引自受限玻尔兹曼机 (RBM) 原理总结 - 刘建平）。

该网络具体应用可以用到推荐系统上，输入层为推荐物品的编码，隐藏层则是各个物品的特征。由于该网络是支持双向过程（编解码）的，因此可以通过输入用户的历史记录（比如 1-0-0-1-0 表示该用户对第0、3个物品感兴趣），编码得到隐藏层，然后解码重新得到可见层（比如得到 1-1-0-1-0），对比原输入就可以认为该用户对第1个物品也感兴趣。

3.1.3 训练

大致的训练过程和Logistic回归有点不同，使用的不是BP算法，而是**基于对比散度的快速学习算法**（CD）。需要训练的参数为 W 和 b ，这里的 b 代指 v 和 h 的偏置，用下标 i 、 j 区分。 b_i 表示 v_i 的偏置， b_j 表示 h_j 的偏置。



在给定的训练图像 v^0 时，计算隐藏层 h^0 的每个元素 h_j^0 取1的概率：

$$P(h_j^0 = 1 | v^0) = \sigma(\sum_i v_i^0 w_{ij} + b_j)$$

其中 $\sigma(x) = \frac{1}{1+e^{-x}}$ 。

然后在概率函数 $P(h_j^0 = 1 | v^0)$ 中取样得到 $h_j^0 \in \{0, 1\}$ ，即得到隐藏层 h^0 。

在得出 v^0 以及 h^0 后，计算权值 w_{ij} 的第一次得分：

$$F_{data}(w_{ij}) = \langle v_i h_j \rangle_{data} = v_i^0 \cdot h_j^0$$

由于 $v_i^0, h_j^0 \in \{0, 1\}$ ，所以 $F_{data}(w_{ij}) \in \{0, 1\}$ 。

计算得出 h^0 后， h^0 通过重构（Reconstruct）得到 v^1 ： v^1 中的每个元素 v_i^1 都有下式概率为1：

$$P(v_i^1 = 1 | h^0) = \sigma(\sum_j h_j^0 w_{ji} + b_i)$$

然后，再通过 v^1 以类似计算 h^0 的方式计算得出 h^1 。计算权值 w_{ij} 的第二次得分：

$$F_{recon}(w_{ij}) = \langle v_i h_j \rangle_{recon} = v_i^1 \cdot h_j^1$$

同样， $F_{data}(w_{ij}) \in \{0, 1\}$ 。

最终得出权值W的更新公式为：

$$W \leftarrow W + \Delta w_{ij} = W + \epsilon (F_{data}(w_{ij}) - F_{recon}(w_{ij}))$$

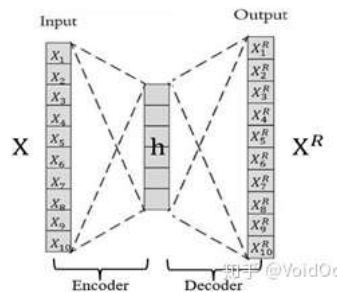
其中 ϵ 为学习率。引自深度学习 --- 受限玻尔兹曼机详解。

偏置b以相似方式更新。可见，在RBM是不需要标签的无监督学习。虽然训练用的不是BP算法，但是也有相当不错的效果。

3.2 Autoencoder

参考深度学习笔记之 —— 自编码器 (AutoEncoder)。

autoencoder是一个单隐含层的、能完成编码、解码工作的神经网络，其结构如下。



该网络包含两个主要部分：**编码器 (Encoder)** 和**解码器 (Decoder)**。该网络的目标是学习一个恒等函数令 $X^R \approx X$ 。由于学习的是恒等函数，所以该网络是无监督的——只需要有输入样本，样本标签与样本一致即可。可能单纯的学习一个恒等函数没什么意思，但是我们可以加上一些额外的限制：比如要先用**更小**的特征进行编码，然后再通过解码将其复原。很明显，这种编码是有损的，但是却仍然能通过这有限的特征数来表达原始数据，这就是**降维**算法要能满足的最低要求。

编码：编码这个过程即将输入的一个N维向量X（如果是矩阵则可以展开为向量），用公式（6）进行降维成M维。

$$h = \alpha(W_1 \cdot X + b_1) = \sigma(W_1 \cdot X + b_1)$$

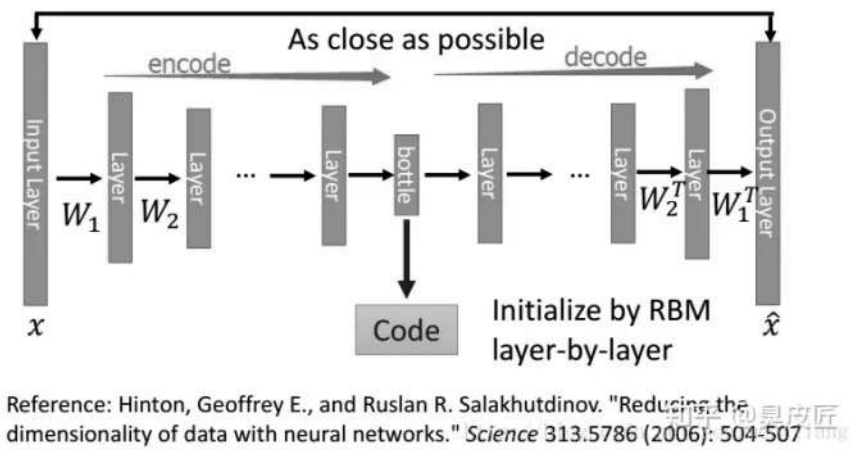
解码：解码与编码相对应，用公式（7）将M维向量升维成N维。

$$X^R = \alpha(W_2 \cdot h + b_2) = \sigma(W_2 \cdot h + b_2)$$

由于是神经网络，所以可以直接使用我们熟悉的BP算法进行训练。

3.3 Deep Autoencoder

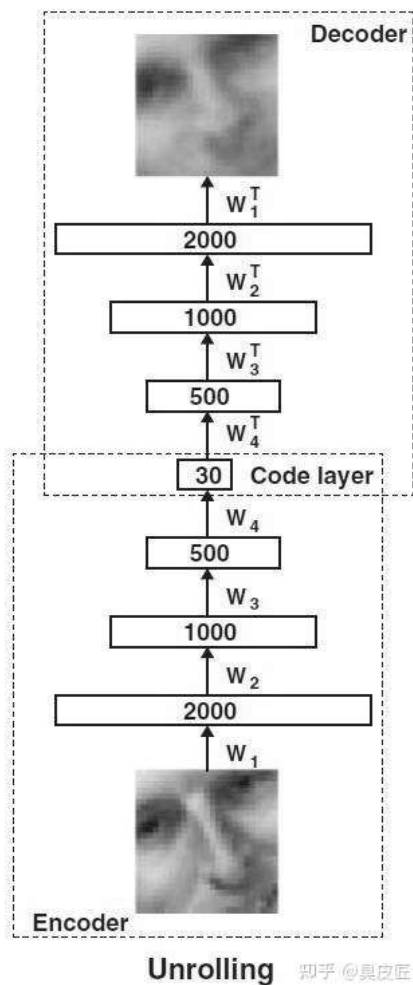
本文提出的DAE (Deep Autoencoder) 是autoencoder的升级版，原理基本相同，只是拥有更深的结构。



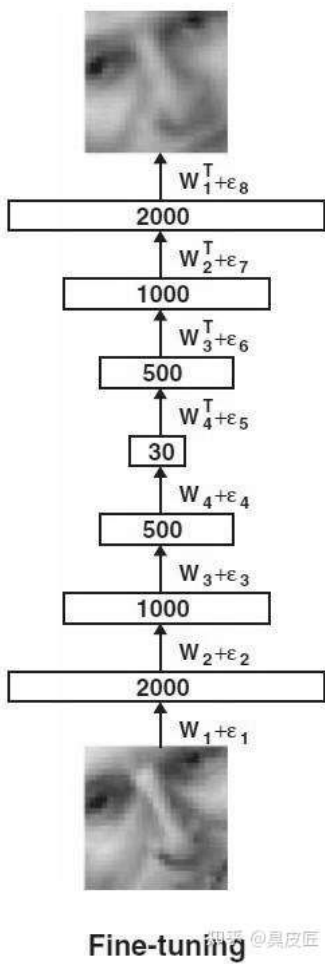
论文中给出的完整的训练该方法的过程如下：

初始化：由于网络更深了，所以梯度消失等问题也会很明显。本文为了解决这个问题，提出了一个逐层的、基于RBM的预训练来进行初始化，这样可以让网络在用BP算法训练前就有一个比较好的起步（good solution）。

展开：在完成初始化后，采用相同的权重构造**encoder**和**decoder**，只是encoder为直接使用，decoder需要进行转置。至于能够直接这样使用的原因，在于RBM初始化带来的特性：RBM本身就兼具编解码能力（通过重构训练解码能力），因此初始化出来的权重也同样兼具编解码能力。



微调：展开后，用BP算法进行微调，直至网络收敛。微调过程中原本只是进行了转置的权重就会不完全相同了，因为展开后两者就互相独立了。

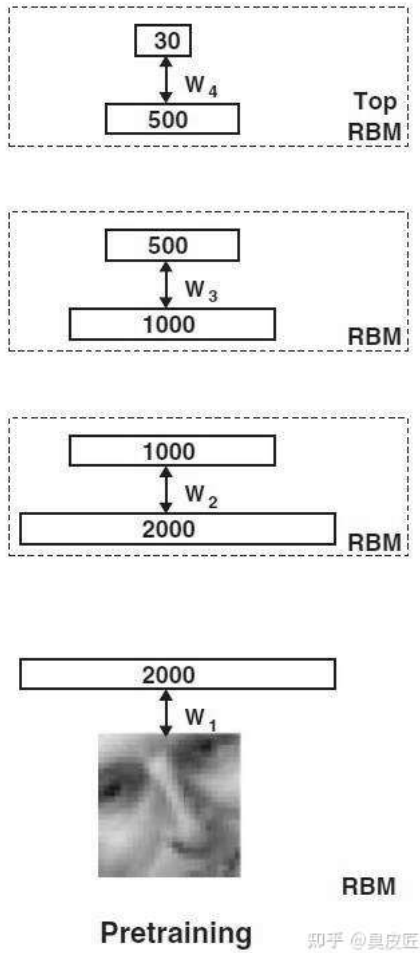


3.4 基于RBM的初始化

以二值图像为例，给出的有效初始化`autoencoder`的方法是：

1. 将`autoencoder`中每层的权重 W_i 视作一个RBM的权重，初始化用较小的随机初始化。
2. 对第一层的RBM进行训练，直至收敛。
3. 用同样的训练样本给第一层的RBM进行预测，得出激活值（隐藏层的值/编码）。将该激活值作为第二层的RBM的训练样本，进行训练。
4. 重复类似第三步的工作，训练所有的权重（RBM）。

最终，得到一组经过预训练的权重，这样就视作初始化完成。下图是一个有4层神经网络的`encoder`的预训练过程。只不过样本不是二值的，而是实数。



如果输入数据是实数，而非二值，论文中采用了如下几种措施：

- 每一个可见层都用Logistic函数（\sigma函数）进行激活，让其取值在[0, 1]间
- 训练时，高层的可见层的样本为上一层隐藏层的激活概率
- 训练时，除了最后一层（顶层，比如 W_4 ）每个隐藏层仍然是随机的二值
- 训练时，最后一层的隐藏层从方差为1、均值由该层的可见层的决定的高斯分布（正态分布）中取样。

4 相关的关键人物与工作

4.1 关键人物

本篇论文引用的文献并不多，引用的最多的是论文作者Hinton本人。但是他最后致谢了几位和他一起讨论的大佬：David Rumelhart（D. Rumelhart）、Max Welling（M. Welling）、Simon Osindero（S. Osindero）和Sam Roweis（S. Roweis）。

下图的搜索结果均在Semantic Scholar搜索得到。

4.1.1 Geoffrey Everest Hinton

大家都很熟悉的名字了，不愧是三巨头。

4.1.2 David Rumelhart

与Hinton于1986年一起提出BP算法，这也是他的主要贡献。



D. Rumelhart

Publications 147
h-index 55
Citations 63,466
Highly Influential Citations 2,522

[Follow Author...](#)

[Claim Author Page](#)

Author pages are created from data sourced from our academic publisher partnerships and public sources.

PUBLICATIONS Publications 388 Influence

Search Publications

[Q](#) Co-Author ▾ Has PDF More Filters Sort by Most Influ... ▾

Learning internal representations by error propagation

D. Rumelhart, Geoffrey E. Hinton, R.J. Williams · Mathematics, Computer Science · 3 January 1986

This chapter contains sections titled: The Problem, The Generalized Delta Rule, Simulation Results, Some Further Generalizations, Conclusion

17,765
 777
 PDF
 View via Publisher
 Save
 Alert
 Cite
 Research Feed

Learning representations by back-propagating errors

D. Rumelhart, Geoffrey E. Hinton, R.J. Williams · Computer Science · Nature · 1 October 1986

TLOD We describe a new learning procedure, back-propagation, for networks of neurone-like units. Expand

16,465
 921
 PDF
 View on Springer
 Save
 Alert
 Cite
 Research Feed

Parallel Distributed Processing: Explorations in the Microstructure of Cognition

James L. McClelland, D. Rumelhart · 1985

2,451
 109
 Save
 Alert
 Cite
 Research Feed

Forward Models: Supervised Learning with a Distal Teacher

Michael I. Jordan, D. Rumelhart · Psychology, Computer Science, Cogn. Sci. · 1 July 1992


TLOD Internal models of the environment have an important role to play in adaptive systems. Expand

1,487
 91
 PDF
 View via Publisher
 Save
 Alert
 Cite
 Research Feed

4.1.3 Max Welling

高通技术副总裁、阿姆斯特丹大学机器学习首席教授。曾和他的一名学生提出GCN（图卷积神经网络，下图引用第二多）。下图引用最多的论文讲述的是一个将贝叶斯和神经网络结合而成的变分自编码器。

他本人有一个主要工作是研究概率模型，所以也可以发现下图引用第一、三多的论文都用到了贝叶斯。





M. Welling

Publications 273



h-index 65


Citations 33,576

Highly Influential Citations 6,273

Author pages are created from data sourced from our academic publisher partnerships and public sources.

 Publications
  Influence





Co-Author
Has PDF
More Filters

Sort by Most Influe...

Auto-Encoding Variational Bayes

Diederik P. Kingma · M. Welling · Mathematics, Computer Science · ICLR · 20 December 2013



TLOU We introduce a stochastic variational inference and learning algorithm that scales to large datasets and, under some mild differentiability conditions, even works in the intractable case. Expand

 9,663
  2195
 PDF
 View PDF on arXiv
 Save
 Alert
 Cite
 Research Feed

Semi-Supervised Classification with Graph Convolutional Networks

Thomas Kipf · M. Welling · Computer Science, Mathematics · ICLR · 9 September 2016



TLOU We present a scalable approach for semi-supervised learning on graph-structured data that is based on an efficient variant of convolutional neural networks which operate directly on graphs. Expand

 5,055
  1489
 PDF
 View PDF on arXiv
 Save
 Alert
 Cite
 Research Feed


Bayesian Learning via Stochastic Gradient Langevin Dynamics

M. Welling · Y. Teh · Mathematics, Computer Science · ICML · 28 June 2011

TLOU In this paper we propose a new framework for learning from large scale datasets based on iterative learning from small mini-batches. Expand

 1,114
  220
 PDF
 View Paper
 Save
 Alert
 Cite
 Research Feed

Recommended Authors



Yoshua Bengio

817 Publications, 245,182 Citations

4.1.4 Simon Osindero

和Hinton一起提出了深度置信网络（下图第一篇）；重要贡献还有提出了Conditional GAN的（下图第二篇）。

Simon Osindero

Publications241

h-index21

Citations17,884

Highly Influential Citations2,030

Follow Author...

Claim Author Page

Author pages are created from data sourced from our academic publisher partnerships and public sources.

Publications284Influence

Search Publications

Q

Co-Author

Has PDF

More Filters

Sort by Most Influ...

A Fast Learning Algorithm for Deep Belief Nets

Geoffrey E. Hinton, Simon Osindero, Y. Teh · Mathematics, Computer Science · Neural Computation · 1 July 2006

TLDR We derive a fast, greedy algorithm that can learn deep, directed belief networks one layer at a time, provided the top two layers form an undirected associative memory. Expand

11,229

1073

PDF

View on arXiv Preprint

Save

Alert

Cite

Research Feed

Conditional Generative Adversarial Nets

M. Mirza, Simon Osindero · Computer Science, Mathematics · arXiv · 8 November 2014

TLDR We introduce the conditional version of generative adversarial nets, which can be constructed by simply feeding the data, y, we wish to condition on to both the generator and discriminator. Expand

6,370

618

PDF

View PDF on arXiv

Save

Alert

Cite

Research Feed

Cross-Dimensional Weighting for Aggregated Deep Convolutional Features

Yannic Kalantidis, Claudiu Mitea, Simon Osindero · Computer Science · ECCV Workshops · 13 December 2015

TLDR We propose a simple and straightforward way of creating powerful image representations via cross-dimensional weighting and aggregation of deep convolutional neural network layer outputs. Expand

267

42

PDF

View PDF on arXiv

Save

Alert

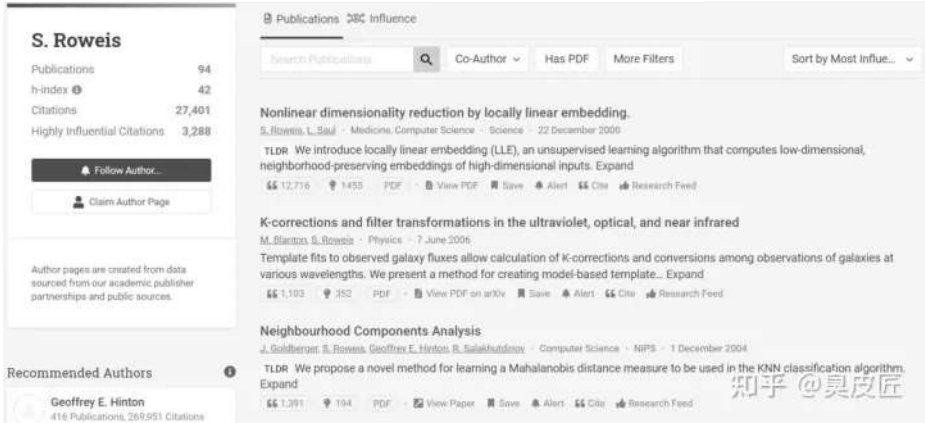
Cite

Research Feed

Recommended Authors

K. Kavukcuoglu

4.1.5 Sam Roweis



4.2 关键工作

4.2.1 受限玻尔兹曼机

RBM于1986年第一次提出，后因Hinton于2000年代发明了新的训练方法（CD算法）后而真正出名并广泛应用。

更具体内容的上文有，此处不再赘述。

4.2.2 PCA

主成分分析（Principal Component Analysis，PCA）是一种对高维度特征数据预处理方法。主要思想是从原始的空间中顺序地找一组**相互正交**的坐标轴，新的坐标轴的选择与数据本身是密切相关的。该算法是线性降维算法，相对来说更简单同时效果也不算差，因此是主流的降维算法之一。由于原理比较硬核，故此处不展开细讲，感兴趣的同学自行了解。

5 文章提出的解决方案的关键

5.1 DAE

深层自编码器具有非线性、自适应等特点，这些特点让其拥有了比线性降维算法PCA更优秀的性能。

5.2 用RBM进行预训练

限于当时的发展，相对于随机初始化，这种初始化方法可以为BP算法提供一个很好的起点，相当于直接把原来随机的起点换成了在最优解附近的起点，这无疑让BP算法可以轻易到达最优解。

6 实验设计

6.1 随机二值曲线

数据集：第一个实验用人造二值曲线数据集。曲线是通过先生成三个随机的二维点（6个数）后再绘制的。

BP算法微调：损失函数用像素级对数损失函数

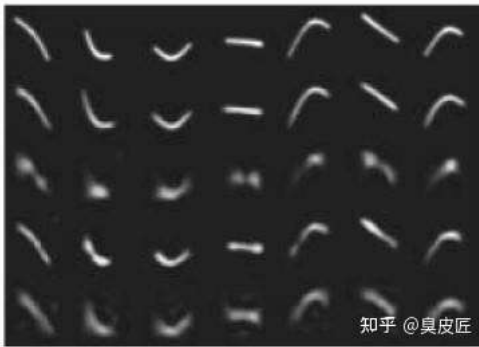
$$-\sum_i p_i \log \hat{p}_i - \sum_i (1 - p_i) \log(1 - \hat{p}_i)$$

自编码器模型：编码器各层的shape为：(28 \times 28)=784 - 400 - 200 - 100 - 50 - 25 - 6。个人认为最值得学习的是最后的特征数为6，这对应了生成该曲线的6个数，并不是拍脑门一想然后用6个的；解码器是对称的；最后一层6维的特征层是用线性激活，其余均是Logistic激活（\sigma函数）。

结果：



- 最终该DAE正确学会将输入的784维用6个特征来表示，并且明显优于PCA算法
- 同时，还实验了在没有预训练的情况下，即使用BP训练非常久，DAE也总是重构数据的平均值（没看懂这句，原话为always reconstructs the average of the training data）
- 只有单隐藏层的Autoencoder虽然可以在没有预训练的情况下学习，但是加入预训练后能大大减少训练总耗时
- 用相同的参数量分别搭建深层、浅层的自编码器，深层在测试集有着更低的错误率；但随着参数量的增加，这种优势会逐渐消失



第一行为随机曲线样本；第二行为特征数为6的autoencoder；第三行为使用6个主成分的logistic PCA；第四、五行分别为使用18个主成分的logistic PCA和标准PCA；均方误差分别为1.44, 7.64, 2.45, 5.90。

6.2 MNIST手写数字

BP算法微调：损失函数用像素级对数损失函数

$$-\sum_i p_i \log \hat{p}_i - \sum_i (1 - p_i) \log(1 - \hat{p}_i)$$

自编码器模型：编码器各层的shape为：(28 \times 28)=784 - 1000 - 500 - 250 - 30。最后一层30维的特征层是用线性激活，其余均是Logistic激活（\sigma函数）。

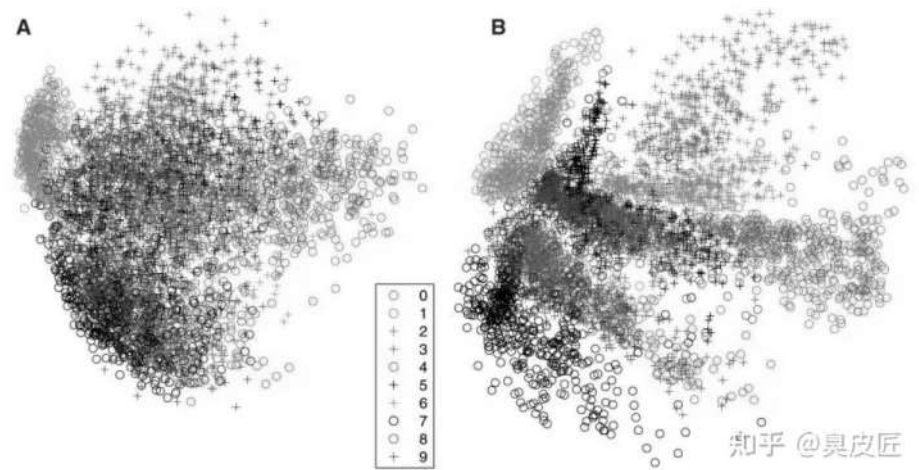
训练集和测试集：60,000个训练数据以及10,000个测试数据。

结果：做了两个测试，第一个是手写数字压缩到30维特征；第二个是压缩到2维特征。

1. 第一行为第一行为随机手写数字样本；第二行为降维至30个特征的autoencoder；第三、四行分别为30个主成分的logistic PCA和标准PCA；均方误差分别为3.00, 8.01, 13.87。



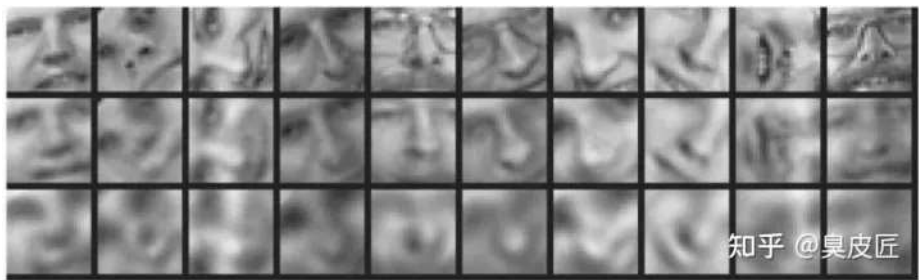
1. 下图一共有10类（0~9），每类有500个数字（对应500个点），在降维到2维后进行可视化。A为2个主成分的PCA算法的可视化结果；B为特征数为2的DAE的可视化结果。



6.3 Olivetti人脸数据集

模型：基本同上，模型改为 (25\times25)=625 - 2000 - 1000 - 500 - 30。

结果：第一行为随机人脸样本；第二行为30个特征的`autoencoder`；第三行为30个主成分的PCA。均方误差分别为126、135。



6.4 文本检索

数据集：804,414个新闻故事，并且用2000个最常见的词干组成概率向量。即用 一个2000维的向量表示一个故事，每个元素是对应词干的频率（a vector of document-specific probabilities of the 2000 commonest word stems）。

BP算法微调：损失函数用多分类交叉熵损失函数

$$-\sum_i p_i \log \hat{p}_i$$

自编码器模型：编码器各层的shape为：2000 - 500 - 250 - 125 - 10。

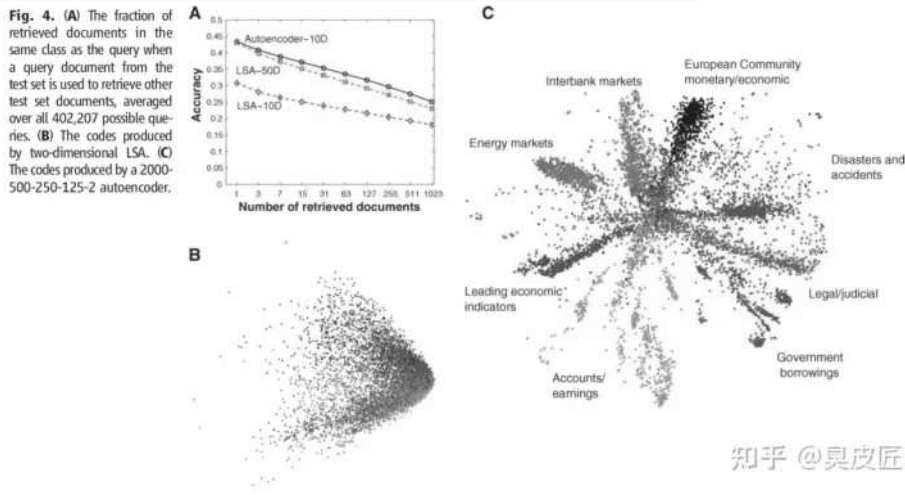
训练集和测试集：20,000个训练数据以及10,000个测试数据。激活函数同上。

结果：最终得到的为10维特征的向量，用向量间的余弦衡量两者的相似性。对比的算法是LSA（latent semantic analysis，一个很出名的基于PCA的文本检索算法）和 local linear embedding（非线性降维算法）。

下图A，纵轴是准确率，横轴是检索文档的数量。可以看出同样10维的情况下DAE的准确率明显比LSA高，甚至比50维的LSA还优秀。

下图B，是2维LSA的可视化。

下图C，是2维DAE的可视化。对比B，明显能将不同类型的文档分离开来。



6.5 分类及回归任务

在MNIST数据集上，随机初始化的神经网络的最优错误率为1.6%；SVM（support vector machines，支持向量机）则为1.4%；结构为 784 - 500 - 500 - 200 - 10的网络在采用本文提出的预训练方法后错误率下降到1.2%。

预训练很好地增加了网络的泛化能力，因为它确保了权值中的大部分信息都是只来自图像的重建，而标签中的非常有限的信息仅用于进行微调。

7 采用的数据集及其是否开源

本文主要采用了两个数据集，一个是总所周知的mnist手写数字数据集，另一个则是Olivetti人脸数据集，均开源。

7.1 mnist数据集

数据集网址[在这](#)。

7.2 Olivetti face data set

数据集网址[在这](#)。

8 实验结果是否验证了科学假设

是。本文从定性和定量两方面进行分析，结果均显示本文的DAE要优于PCA及其衍生模型（LSA）。

不过个人认为，虽然PCA是使用最广泛的降维算法，和他进行比较合情合理；但是应该也需要和其他的非线性降维算法比较，这样会更有说服力。

9 本文贡献

9.1 解决BP算法的缺点

虽然本文的主要工作都是在降维算法方面，但是提出的初始化方法并不局限于编码器，还适用于分类、回归任务。因此是第一次提出能有效解决BP算法缺陷的初始化方案，让深度学习在06年又一次掀起热潮。

9.2 提出DAE进行降维

本文提出的Deep Autoencoder降维算法性能对比PCA相当优秀。



预训练虽好，但是对比直接使用BP需要花费更多的时间，方法也算不上简单。所以需要设计其他更快捷且有效的初始化方法。但是如果用我们现在的角度（2020年）来看，BP算法的局限性已经可以通过Adam等优化器很好地解决了。既然初始化和优化器两方面都基本充分发展了，那剩下的可以做的就是用新的算法来调整权重等参数。比如可以结合强化学习的思想，通过试错的方式就行调整。

11 重要的相关论文

1. G. E. Hinton, Neural Comput. 14, 1711 (2002). 引用了两次，讲的是有关RBM的研究。

12 不懂之处

实验部分的文档检索有一些疑惑：

- a vector of document-specific probabilities of the 2000 commonest word stems不知道和我理解的是不是一致
- 文档检索（Document retrieval）的目标、任务我不是特别清楚

13 专业术语

13.1 受限玻尔兹曼机

Restricted Boltzmann machine, RBM。

13.2 自编码器

Autoencoder / Auto Encoder。

13.3 主成分分析

Principal Component Analysis, PCA。

13.4 潜在语义分析

Latent Semantic Analysis, LSA。本质上是分析词语的潜在语义，以此做到检索文本时，可以通过分析潜在语义来判断目标词与当前词是不是近义词而选择要不要检索。

14 英语词汇

restricted, 受限的；

component, 成分；组成部分；组件、元件；

unit variance, 方差为1、单位方差

retrieval, 检索

reconstruction, 重构、重建

下一篇：深度学习经典论文分析（六）-Deep Residual Learning for Image Recognition

感谢@Kingsley Alien对笔记后期的精心校对

编辑于 2021-11-10 01:50

深度学习（Deep Learning） 经典 文献笔记



写下你的评论...



还没有评论，发表第一个评论吧

文章被以下专栏收录



深度学习-论文阅读
介绍一些深度学习领域的经典论文

推荐阅读



精读深度学习论文(25)
Siamese Network
清欢守护者 发表于Bob学步



精读深度学习论文(4)
Inception V3
清欢守护者



精读深度学习论文(31)
SiameseFC
清欢守护者 发表于Bob学步



2021...
(Con
忆臻