CSMATH

阅读报告

姓名：赵磊
班级：博2班
学号：11521062
学院：计算机学院

2016 年 6 月 13 日

深度学习浅谈

1. 深度学习的发展

这是在一篇有关Deep Learning比较新的综述[1]中所写的，简单谈下我的理解。机器学习这个名词已经不再新鲜，之前接触过的支持向量机、梯度下降算法等等，这些都是传统的机器学习算法，机器学习算法的成功主要取决于数据的表达。我们一般猜测，不同的表达会混淆或者隐藏或多或少的可以解释数据不同变化的因素，尽管特定的领域知识可以有助于设计或者选择数据的表达，但通过一般的先验知识来学习表达也是有效的。而且，人工智能AI的发展要求也迫使我们去寻找利用先验知识的更强大的特征学习算法。目前非监督特征学习和深度学习领域的一些近期工作，主要包括概率模型、自动编码机、流形学习和深度网络。

众所周知，机器学习方法的性能很大程度上取决于数据表达（或者特征）的选择。也正是因为这个原因，为了使得机器学习算法有效，我们一般需要在数据的预处理和变换中倾注大部分的心血。这种特征工程的工作非常重要，但它费时费力，属于劳动密集型产业。这种弊端揭露了目前的学习算法的缺点：在提取和组织数据的区分性信息中显得无能为力。特征工程是一种利用人的智慧和先验知识来弥补上述缺点的方法。为了拓展机器学习的适用范围，我们需要降低学习算法对特征工程的依赖性，这样，就可以更快的构建新的应用，更重要的是，在人工智能领域迈出一大步。人工智能最基本的能力就是能理解这个世界（understand the world around us），只有当它能学会如何辨别和解开在观测到的低级感知数据中隐含的解释性因素时才能达到这个目标。

表达学习（亦被江湖称作深度学习或者特征学习）已经在机器学习社区开辟了自己的江山，成为学术界的一个新宠。在一些顶尖会议例如NIPS和ICML中都有了自己的正规军（研究它的workshops）， 2013 年还专门为它搞了一个新的会议，叫ICLR （International Conference on Learning Representations），可见它在学术界得到的宠爱招人红眼。尽管depth（深度）是这个神话的一个主要部分，但其他的先验也不能被忽视，因为有时候，先验知识会为表达的学习献上一臂之力，画上点睛之笔，更容易地学习更好的表达。在表达学习有关的学术活动中最迅速的进展就是它在学术界和工业界都得到了经验性的显著性的成功，第三部分中我们简单的聚焦了几点。

2. 深度学习基础理论

在这一部分中我以深度学习的基础－－－－卷积神经网络CNN[2]为例来简单阐述下深度学习的思想。典型的CNN中，开始几层都是卷积和下采样的交替，然后在最后一些层（靠近输出层的），都是全连接的一维网络。这时候我们已经将所有两维2D的特征maps转化为全连接的一维网络的输入。这样，当你准备好将最终的2D 特征maps输入到1D网络中时，一个非常方便的方法就是把所有输出的特征maps 连接成一个长的输入向量。

(a) 卷积层

在一个卷积层，上一层的特征maps[3]被一个可学习的卷积核进行卷积，然后通过一个激活函数，就可以得到输出特征map。每一个输出map可能是组合卷积多个输入maps的值：

$$x_j^l = f(\sum_{i \epsilon M_j} x_i^{l-1} * k_{ij}^l + b_j^l). \tag{1}$$

这里Mj表示选择的输入maps的集合，有选择一对的或者三个的。每一个输出map会给一个额外的偏置b，但是对于一个特定的输出map，卷积每个输入maps的卷积核是不一样的。也就是说，如果输出特征map j和输出特征map k都是从输入map i中卷积求和得到，那么对应的卷积核是不一样的。

(b) 子采样层对于子采样层来说，有N个输入maps，就有N个输出maps，只是每个输出map都变小了。

$$x_j^\iota = f(\beta_j^\iota down(x_i^{\iota-1}) + b_j^\iota). \tag{2}$$

down(.)表示一个下采样函数。典型的操作一般是对输入图像的不同$n*n$的块的所有像素进行求和。这样输出图像在两个维度上都缩小了n倍。每个输出map都对应一个属于自己的乘性偏置 β 和一个加性偏置b。

(c) 学习特征map的组合

大部分时候，通过卷积多个输入maps，然后再对这些卷积值求和得到一个输出map，这样的效果往往是比较好的。在一些文献中，一般是人工选择哪些输入maps去组合得到一个输出map。但我们这里尝试去让CNN在训练的过程中学习这些组合，也就是让网络自己学习挑选哪些输入maps来计算得到输出map才是最好的。我们用 α ij表示在得到第j个输出map的其中第i个输入map的权值或者贡献。这样，第j个输出map可以表示为：

$$x_j^\iota = f(\sum_{i=1}^{N_{in}} \alpha_{ij}(x_i^{\iota-1} * k_i^\iota) + b_j^\iota). \tag{3}$$

需要满足约束:

$$\sum_i \alpha_{ij} = 1, and\ 0 \le \alpha_{ij} \le 1. \tag{4}$$

这些对变量 α ij的约束可以通过将变量 α ij表示为一组无约束的隐含权值cij 的softmax函数来加强。（因为softmax的因变量是自变量的指数函数，他们的变化率会不同）。

$$\alpha_{ij} = \frac{exp(c_{ij})}{\sum_k exp(c_{kj})}. \tag{5}$$

因为对于一个固定的j来说，每组权值cij都是和其他组的权值独立的，所以为了方便描述，我们把下标j去掉，只考虑一个map的更新，其他map的更新是一样的过程，只是map的索引j不同而已。

Softmax函数的导数表示为：

$$\frac{\partial \alpha_k}{\partial c_i} = \delta_{ki}\alpha_i - \alpha_i\alpha_k. \tag{6}$$

这里的δ是Kronecker delta。对于误差对于第ι层变量 α i的导数为：

$$\frac{\partial E}{\partial \alpha_i} = \frac{\partial E}{\partial u^\iota}\frac{\partial u^\iota}{\partial \alpha_i} = \sum_{u,v}(\delta^\iota \circ (x_i^{\iota-1} * k_i^\iota))_{uv}. \tag{7}$$

最后就可以通过链式规则去求得代价函数关于权值ci的偏导数了：

$$\frac{\partial E}{\partial c_i} = \sum_k \frac{\partial E}{\partial \alpha_k}\frac{\partial \alpha_k}{\partial c_i} = \alpha_i(\frac{\partial E}{\partial \alpha_i} - \sum_k \frac{\partial E}{\partial \alpha_k}\alpha_k). \tag{8}$$

3. 深度学习的应用

(a) 语音识别与信号处理

语音也是神经网络诞生时期最早的一个应用之一，例如卷积（或者时延）神经网络（Bengio在 1993 年的工作），当然，HMM在语音识别成功之后，神经网络也相对沉寂了不少。到现在，神经网络的复活，深度学习在语音识别领域可谓大展拳脚，重展雄风，在一些学术派和工业派人士（Dahlet al., 2010; Deng et al., 2010; Seide et al., 2011a; Mohamedet al., 2012; Dahl et al., 2012; Hinton et al., 2012）的努力下取得了突破性的成果，使得这些算法得到更大范围的应用，并且实现了产品化。例如，微软在2012年发布了它们的语音识别MAVIS (Microsoft Audio Video Indexing Service)系统的一个新版本，这个版本是基于深度学习的（Seide et al., 2011a）。对比现有的一直保持领先位置的高斯混合模型的声学建模

方法，他们在四个主要的基准测试集中把错误率降低了 30% 左右（例如在RT03S数据库中从27.4%的错误率降到 18.5%）。在 2012 年，Dahl 等人再次书写神话，他在一个小的大词汇量语音识别基准测试集中（Bing移动商业搜索数据库，语音长40小时）的错误率降到 16% 与 23% 之间。

表达学习算法还被应用的音乐方面上，在四个基准测试集中，比当前领先的polyphonic transcription (Boulanger-Lewandowskiet al., 2012)在错误率上取得了5%到30%之间的提升。深度学习还赢得了MIREX (Music Information Retrieval)音乐信息检索竞赛的冠军。

(b) 目标识别

在 2006 年，深度学习的开始，主要聚焦在MNIST手写体图像分类问题上（Hinton et al.,2006; Bengioet al., 2007），它冲击了SVMs在这个数据集的霸主地位（1.4%的错误率）。最新的记录现被深度网络占据着：Ciresanet al.(2012)声称他在这个任务的无约束版本（例如，使用卷积架构）的错误率是 0.27% ，而Rifaiet al.(2011c)在MNIST的knowledge-free版本中保持着 0.81% 的错误率 。

在最近几年，深度学习将其目光从数字识别移到自然图像的目标识别，而最新的突破是在ImageNet数据库中把领先的 26.1% 的错误率拉低到 15.3% (Krizhevskyet al., 2012)。

(c) 自然语言处理

除了语音识别，深度学习在自然语言处理中也有很多应用。symbolic 数据的分布式表达由Hinton在 1986 年引入，在 2003 年由Bengio等人在统计语言模型中得到第一次的发展，称为神经网络语言模型(neural net language models)(Bengio,2008)。它们都是基于学习一个关于每个单词的分布式表达，叫做word embedding。Collobertet al.(2011)在此基础上增加了一个卷积架构开发了一个SENNA系统，它在语言建模、部分语音标记、节点识别、语义角色标记和句法分解中共享表达。SENNA接近或者超于目前的在这些任务中的那些领先方法，它比传统的预测器要简单和快速。学习word embeddings可以以某种方式与学习图像表达结合，这样就可以将对文本和图像的理解联系起来。这个方法被成功运用到谷歌的图像搜索上，利用大量的数据来建立同一空间中图像与问题之间的映射(Weston et al.,2010)。在 2012 年，Srivastava等将其拓展到更深的多模表达。

神经网络语言模型也被通过隐层增加循环来改进(Mikolovet al., 2011)，与当下领先的平滑n-gram语言模型相比，不仅在复杂度上降低，还降低了语音识别的错误率（因为语言模型是语音识别系统的一个重要组成部分），这个模型还被应用到统计机器翻译上面(Schwenk et al., 2012; Leet al., 2013)，改进了复杂度和BLEU分数。递归自动编码机(Recursive auto-encoders)（产生循环网络）在全句释义检测(full sentence paraphrase detection)上也达到了现有的领先水平，是以前技术的两倍分数(Socheret al., 2011a) 。表达学习还用到了单词歧义消除(word sense disambiguation)上(Bordeset al., 2012)，取得了准确率从 67.8% 到 70.2%的提升。最后，它还被成功运用到情感分析(Glorotet al., 2011b; Socher et al., 2011b) 上，并超越现有技术。

参考文献

[1] Representation Learning: A Review and New Perspectives, Bengio Y., Courville A.,& Vincent P., 2012.

[2] Notes on Convolutional Neural Networks , Jake Bouvrie , November 22, 2006.

[3] Visualizing Higher-Layer Features of a Deep Network, Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent , June 9th, 2009.