深层神经网络中间层可见化建模

高莹莹 [1] 朱维彬 [1]

摘 要 深层神经网络的中间层是隐含的、未知的, 这使得深层网络的学习过程不可追踪, 学习结果无法解释, 在一定程度上制约了深度学习的发展. 本文通过引入先验知识使深层网络的中间层具有明确的含义与显性的影响关系, 即中间层可见化, 从而部分人工干预深层网络的内部结构, 约束网络学习的方向. 基于深层堆叠网络 (Deep stacking network, DSN), 提出两种中间层部分可见的深层神经网络: 输入层部分可见的深层堆叠网络 (Input-layer visible DSN, IVDSN) 和隐含层部分可见的深层堆叠网络 (Hidden-layer visible DSN, HVDSN), 部分可见是为了保留对未知信息的提取能力和一定的容错能力. 以基于文本的言语情感计算为例测试所提网络的有效性, 结果表明先验知识的引入有助于提升深层神经网络的性能; 所提两种网络均可实现中间层的部分可见化, 其中 HVDSN 结构更精简, 性能也更优.

关键词 深层神经网络, 深层堆叠网络, 中间层可见化, 言语情感计算

引用格式 高莹莹, 朱维彬. 深层神经网络中间层可见化建模. 自动化学报, 2015, **41**(9): 1627−1637

DOI 10.16383/j.aas.2015.c150023

Deep Neural Networks with Visible Intermediate Layers

GAO Ying-Ying[1] ZHU Wei-Bin[1]

Abstract The hidden nature of intermediate layers in deep neural networks makes the learning process hard to track and the learned results difficult to explain, which restricts the development of deep networks to some extent. This work focuses on making these intermediate layers visible through prior knowledge, which means giving the intermediate layers definite meanings and explicit interrelationship, in the hope to supervise the learning process of deep networks and guide the learning direction. On the basis of deep stacking network (DSN), we propose two networks in which the intermediate layers are partially visible: the input-layer visible deep stacking network (IVDSN) and the hidden-layer visible deep stacking network (HVDSN). To be partially but not fully visible is to leave room for the unknown and the error. With the application of the text-based detection of speech emotion, the performance of the proposed networks is tested. The results validate that the transparency of intermediate layers is beneficial to improve the performance of deep neural networks. Between the two proposed networks, the HVDSN has a simpler structure and a better performance.

Key words Deep neural network, deep stacking network (DSN), visible intermediate layer, speech emotion detection

Citation Gao Ying-Ying, Zhu Wei-Bin. Deep neural networks with visible intermediate layers. *Acta Automatica Sinica*, 2015, **41**(9): 1627−1637

近年来, 深层神经网络因其在计算机视觉[1−3]、语音识别[4−6] 和自然语言处理[7−8] 等领域获得的成功而被广泛关注. 深层神经网络具有多层非线性映射结构, 通过低层特征到高层特征的逐层抽象得到数据的分布式特征. 然而, 由于网络中间层 (隐含层) 的不可见性, 对于所抽取结果的解释及运用、对网络结构的预判以及训练过程的追踪都造成困难, 这在一定程度上也限制了深度学习在更广阔领域的应用与发展.

对于深层神经网络中间层的研究已有一定的工作基础, 如: 文献 [9] 在传统卷积神经网络的基础上, 通过引入联合目标函数实现对包含输出层和中间层在内每一层的监督; 文献 [10] 将无监督的聚类算法嵌入到深层网络有监督的判别任务中, 实现对输出层和中间任一层的半监督学习; 文献 [11] 提出一种新的深层神经网络 — 深层堆叠网络 (Deep stacking network, DSN), 网络由一系列具有相同或相似结构的单隐层神经网络模块堆叠构成. 区别于传统神经网络的整体训练, DSN 的每个模块均可单独有监督训练, 并通过将前一模块输出累加到下一模块作为部分输入的方式实现对前面结果的利用. 对中间层监督力度的加强, 有助于为目标任务提取到更具区分性的特征, 也在一定程度上缓解了多层神经网络训练过程中易陷入局部最优的难题. 然而, 这距中间层真正意义上的可见化还有一定距离, 中间层的含义仍然是未知的, 我们仍难以干预深层网络内部的学习过程并对学习结果给出解释.

在以往的研究中, 大量存在的可能与目标任务相关的先验知识往往被忽略, 对深层网络的监督仅

收稿日期 2015-01-19 录用日期 2015-05-13
Manuscript received January 19, 2015; accepted May 13, 2015
本文责任编委 王占山
Recommended by Associate Editor WANG Zhan-Shan
1. 北京交通大学信息科学研究所 北京 100044
1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044

仅依赖于有限的基于目标任务的标注数据. 利用先验知识对深层网络的内部结构进行部分人工干预或引导, 可以帮助机器抽取到与目标任务更切合的特征, 尤其在标注数据有限的情况下, 先验知识的引入可以有效地扩充可利用信息. 本文的研究目标在于利用先验知识揭示中间层的具体含义及其内部关系, 使深层网络的中间层变得可见化. 因为仍有未解码信息存在的可能性, 我们将这种可见化定义为"部分可见", 在人工干预的同时保留抽取新信息的能力和一定的容错能力. 先验知识的获取依赖于具体任务, 本文以基于文本的言语情感计算为应用目标, 基于心理学、语音学等相关知识, 对言语情感的生成过程进行设定, 并利用 DSN 作为基础网络, 将其与言语情感生成过程融合, 构建中间层部分可见的深层堆叠网络 (Visible deep stacking network, VDSN). 根据堆叠位置的不同, VDSN 又分为输入层部分可见的深层堆叠网络 (Input-layer visible DSN, IVDSN) 和隐含层部分可见的深层堆叠网络 (Hidden-layer visible DSN, HVDSN). 通过实验分别对这两种网络的性能进行测试, 并将其与未加入中间层监督的深层置信网络 (Deep belief network, DBN)[12] 和未引入先验知识的深层堆叠网络 DSN 对比, 验证所提方法对于深层网络的优化效果.

本文第 1 节介绍中间层部分可见的深层堆叠网络的基本结构与训练算法; 第 2 节给出所提网络在言语情感计算中的应用实例; 第 3 节基于该实例对网络的性能进行测试和验证; 最后给出总结与讨论.

1 网络结构与训练

1.1 基础网络

传统的深层神经网络由输入层 (原始特征)、多个隐含层和输出层串联组成. 网络的训练通常采用反向传播 (Back-propagation, BP) 算法, 通过使输出层与目标值的误差最小反向逐层调整各层间的联系权重, 但该误差函数是一个含多个极小值的非线性空间, 因此在沿梯度下降的方向搜寻误差极小值时常使网络收敛于局部最小; 另一方面由于只对输出层进行了监督, 在利用 BP 算法向其他层反向传播该梯度时会随着网络层数的增加而发生梯度消失[13] 的现象. Hinton 等[12] 将无监督贪心逐层算法引入深层置信网 DBN, 有效地提高了网络的收敛速度, 改善了网络易陷入局部最优的局面. DBN 由一系列受限玻尔兹曼机 (Restricted Boltzmann machine, RBM) 构成, RBM 即包含可视层和隐含层的双层对称网络, 层内无连接, 层间全连接. 通过将可视层映射给隐含层再由隐含层重建可视层, 然后将新的可视层再次映射给隐含层, 反复执行这一过程, 从而实现对网络参数的预训练. 预训练的结果将作为 BP 算法的初始值进行有监督微调. 实验和理论[12, 14] 均证明这一初始值较随机初始值更优, 可以有效缓解网络收敛于局部最小的问题. 但是 DBN 的中间层仍然是 "隐含的"、不可监督也不可解释, 梯度消失的问题在有监督微调过程中也仍可能存在. 本文以 DBN 作为未加入中间层监督和先验知识指引的基础网络.

DSN 打破了传统深层神经网络多个隐含层串联的结构, 将其拆分成一系列含单个隐含层的简单模块, 然后通过堆叠的方式将所有模块组合起来, 即前一模块的输出作为输入传递给下一模块. 每个 DSN 模块由一个线性输入层、一个非线性隐含层和一个线性输出层构成, 且各模块独立训练, 均由目标任务进行监督, 相对于传统神经网络训练复杂度更低, 也避免了因层数过多引起的梯度消失现象; 同时, 各模块间通过堆叠方式进行连接也保留了深层网络逐层抽取更具区分性特征的特性. DSN 实现了深层网络各层的可监督训练, 但是各层均以相同的目标任务为学习对象, 仍然缺乏对于中间层真正含义的发掘, 对于模型结构的设定也缺乏有针对性的指引. 我们将 DSN 作为加入中间层监督但未引入先验知识的网络.

1.2 VDSN

在 DSN 的基础上, 我们加入先验知识的指引, 使其参与到对网络结构的约束中, 从而构建中间层部分可见的深层堆叠网络 VDSN. 与 DSN 不同的是, VDSN 的每一模块都具有不同的子目标, 各子目标按照预先设定的生成顺序依次训练并参与到后续目标及最终目标的学习中. 有别于最终目标的子目标的设定, 使网络的内部结构具有明确的含义和显性的关系, 为网络结构的设计提供依据并指引学习的方向. 至此, 深层网络所描述的特征不再是静态的单一目标的特征, 而是动态衍化的包含一系列子目标的连续过程, 各个环节间可能存在的直接或间接的相互影响通过各模块的训练结果的堆叠进行传递. 先验知识的加入可能只破解了网络结构的部分信息, 尚有未解码信息存在的可能, 各子目标间的相互关系也并非确定的一对一的映射关系. 因此出于完备性考虑, 前面所有模块的结果均会堆叠到后续模块以保证其对后续环节的潜在影响, 原始输入也始终作为输入传递给每个模块.

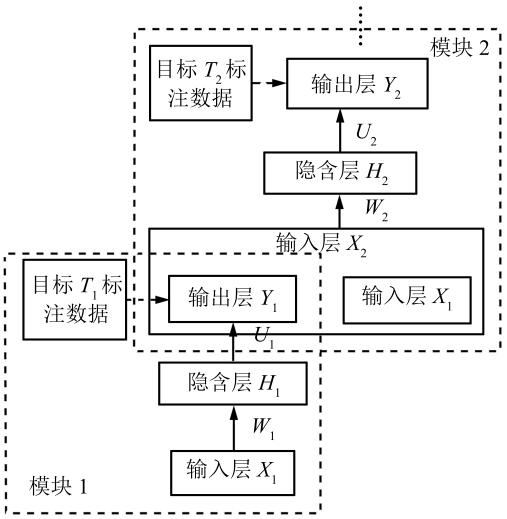
以 X 表示已有的输入特征, Y 表示最终的学习目标, 对于 DSN 以及 DBN 等中间层不可见的深层网络, 网络学习是已知 X 条件下对后验概率 $P(Y|X)$ 进行估计. DBN 通过引入无监督的预训练从而得到观测数据的先验概率 $P(X)$, 相对于单纯的

有监督学习而言, 能够利用大量的无标签数据学习和发现数据中存在的模式, 有助于避免因网络函数表达能力过强而出现过拟合现象. DSN 将 DBN 的逐层无监督训练扩展到逐层有监督训练, 避免了多层神经网络整体的反向误差调整; 同时各模块均由最终目标监督训练, 可以视为在逐层抽取更具区分性的特征. VDSN 延续了 DSN 基本模块的结构以及各模块独立训练的方式, 同时 DBN 中无监督的预训练也可参与到网络参数的初始化; 但是, VDSN 每个模块不再由最终目标监督, 而是依次预测不同的子目标 Y_i, 各子目标与最终目标间构成链式生成关系, 各模块依次估计 $P(Y_1|X)$, $P(Y_2|X,Y_1)$, $P(Y_3|X,Y_1,Y_2)$, $\cdots$, $P(Y_M|X, Y_1, Y_2, \cdots, Y_{M-1})$, M 为子目标数. 由此可见, VDSN 相比于中间层不可见的深层网络虽然也采用逐层训练的模式, 但每层可利用的已知信息在原始输入的基础上得到扩展, 在子目标设定合理的前提下, 先验的引入有助于模型学习到更准确、更具泛化能力的解. 利用先验知识对已知信息进行扩展, 对于数据有限尤其是有标数据难以获得的情况显得尤为重要, 因为仅依赖小规模的数据可能很难依靠网络本身自动发现数据中隐藏的结构模式, 而借鉴相关领域的专业知识对网络结构进行部分人工干预, 可以减少对于训练数据的依赖, 这可被视作从结构设定的角度对网络进行监督或半监督引导.

1.3 IVDSN 与 HVDSN

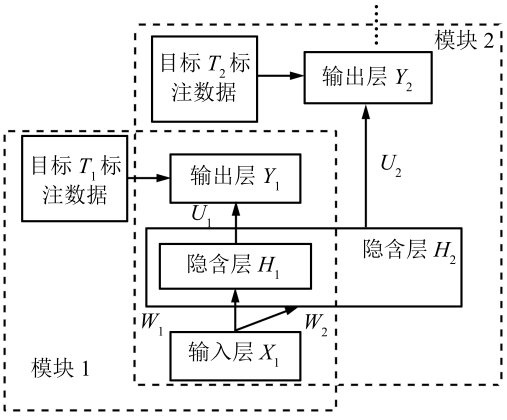
VDSN 的每个模块也由输入层 X、单一隐含层 H 和输出层 Y 构成, 各模块的训练参数有两组: 1) 连接输入层与隐含层的权重矩阵 W; 2) 连接隐含层与输出层的权重矩阵 U. 其中, 隐含层为输入层的非线性映射 $H = \sigma(\boldsymbol{b} + W^{\mathrm{T}}X)$, $\sigma(x) = 1/(1 + \exp(-x))$, 输出层为隐含层的线性组合 $Y = U^{\mathrm{T}}H$. 由此, 各模块可向后面传递的结果有两种可能: 一是将输出层传递给后续模块作为部分输入层, 即两个权重矩阵 W 和 U 均向后传递并对其产生影响; 二是将隐含层传递给后续模块作为部分隐含层, 即仅向后传递连接输入层和隐含层的权重 W. 由于隐含层的值是非线性分布, 而输入层和输出层的值均为线性分布, 因此隐含层没有传递给后续模块的输入层作为部分输入. 第一种堆叠方式与 DSN 的连接模式类似, 但是各模块有不同的训练目标, 每个模块的输出不再是最终目标的估计, 而是预测不同的子目标, 由于训练结果的逐层堆叠发生在后续模块的输入层, 采用这种方式搭建的 VDSN 称为输入层部分可见的深层堆叠网络 (IVDSN). 第二种堆叠方式更直接地体现了中间层部分已知的情况, 由于逐层堆叠体现在隐含层, 将这种网络命名为隐含层部分可见的深层堆叠网络 (HVDSN).

图 1 给出了 IVDSN 和 HVDSN 两种网络中两个模块及其连接关系示意图. 仅就这两个模块而言, 目标 T_2 相当于最终目标, 而 T_1 相当于其子目标, 模块 1 的训练目标为使输出层 Y_1 和 T_1 的误差最小, 而模块 2 的训练目标为使输出层 Y_2 和 T_2 的误差最小, 各模块独立训练, 分别基于各自目标进行本模块内部由输出层到输入层的反向误差调整. 图 1 (a) 中, 模块 1 的输出 Y_1 传递到模块 2 作为部分输入, 同时为了防止信息丢失和误差累积, 原始输入 X_1 也一并传递作为该模块的部分输入, 因此模块 2 的输入层 $X_2 = [X_1; Y_1]$. 图 1 (b) 中, 各模块输入均为 X_1, 但从模块 2 开始, 隐含层的初始方式发生变化, 即隐含层不再完全隐含, 部分隐含层节点通过模块 1



(a) 输入层部分可见深层堆叠网络 IVDSN

(a) Input-layer visible deep stacking network (IVDSN)



(b) 隐含层部分可见深层堆叠网络 HVDSN

(b) Hidden-layer visible deep stacking network (HVDSN)

图 1　两种中间层部分可见的深层堆叠网络 (VDSN) 示意图

Fig. 1　The two visible deep stacking networks (VDSN)

训练得到的隐含层进行初始化, 同时扩张一部分未知节点 (随机初始化或由 RBM 初始化), 使模块 2 隐含层的初始状态变为 $H_2 = [H_1; H_{\text{unkonwn}}]$.

可见, HVDSN 直接延续了神经网络将隐含层作为中间层的概念, 中间层部分可见、部分未知体现于隐含层的继承与扩张上; 而 IVDSN 将低层模块视作高层模块的中间层, 即各子目标作为实现最终目标的中间过程, 中间层的部分可见化体现于中间过程的部分已知 (各子目标的输出) 和部分未知 (原始输入内未被破解的部分). 网络结构上, IVDSN 的输入层随着训练目标的增加而逐层扩张; 而 HVDSN 的输入层保持不变, 隐含层的规模在逐层扩张. 而从计算角度分析, IVDSN 和 HVDSN 均延续了 DSN 各模块独立训练的训练方式, 无需全局的反向误差调整, 计算量上等同于几个单隐层的浅层神经网络; 其差别仅在于, IVDSN 与 DSN 一样, 输入层维度需要逐层递增, 因此比其他网络具有更多的输入层节点, 而 HVDSN 隐含层维度逐层递增, 因此可能比其他网络具有更多的隐含层节点.

1.4 训练算法

与 DBN 和 DSN 类似, VDSN 的训练也分为无监督的预训练 (参数初始化) 和需要少量标注数据的针对目标任务的微调两个步骤, 训练以模块为单位进行, 各模块的训练目标不同. IVDSN 输入层与隐含层间的权重全部由 RBM 进行初始化, HVDSN 隐含层部分节点由前面模块训练得到的隐含层初始化, 剩余节点由 RBM 初始化.

1) 基于 RBM 的参数预训练

RBM 为包含可视层 $\boldsymbol{v}$ 和隐含层 $\boldsymbol{h}$ 的双层对称网络, 层内无连接, 层间全连接. 模型待训练的参数 $\theta = \{W, \boldsymbol{a}, \boldsymbol{b}\}$, 其中 W 为可见层与隐含层间的连接权重, $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别为可见层和隐含层的偏置向量. 参数 θ 通过最大似然估计的方法学习得到:

$$\hat{\theta} = \arg\max L(\theta) = \arg\max \sum_{n=1}^{N} \log P_\theta\left(\boldsymbol{v}^{(n)}\right) \quad (1)$$

似然概率:

$$P_\theta(\boldsymbol{v}) = \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)\right) \quad (2)$$

能量函数:

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i a_i v_i - \sum_j b_j h_j \quad (3)$$

其中, $\hat{\theta}$ 为最优参数, N 为样本数, $Z(\theta) = \sum_{\boldsymbol{v},\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}; \theta))$ 为归一化因子或称为配分函数 (Partition function).

似然概率的最大值可通过随机梯度上升法求得, 但由于计算该梯度时需要用到归一化因子 $Z(\theta)$, 使其很难直接求解. 基于 RBM 对称结构以及神经元的条件独立性, 可以使用 Gibbs 采样方法得到该梯度的一个近似. 对于一个 K 维随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_K)$, 假设无法求得关于 $\boldsymbol{X}$ 的联合分布, 但知道给定 $\boldsymbol{X}$ 中第 k 个分量 X_k 的条件分布 $P(X_k|X_{k^-})$ 时, 那么可以从 $\boldsymbol{X}$ 的任意状态 (如 $[x_1(0), x_2(0), \cdots, x_K(0)]$) 开始, 利用上述条件分布迭代地对其分量依次采样. 随着采样次数的增加, 随机变量 $[x_1(t), x_2(t), \cdots, x_K(t)]$ 的概率分布将以 t 的几何级数的速度收敛于 $\boldsymbol{X}$ 的联合概率分布 $P(\boldsymbol{X})$. 在 RBM 中进行 t 步 Gibbs 采样时, 首先用一个训练样本初始化可见层的状态 $\boldsymbol{v}_0$, 然后交替进行如下采样: $\boldsymbol{h}_0 \sim P(\boldsymbol{h}|\boldsymbol{v}_0)$, $\boldsymbol{v}_1 \sim P(\boldsymbol{v}|\boldsymbol{h}_0)$, $\boldsymbol{h}_1 \sim P(\boldsymbol{h}|\boldsymbol{v}_1)$, $\cdots$, $\boldsymbol{v}_{t+1} \sim P(\boldsymbol{v}|\boldsymbol{h}_t)$.

当给定可见单元状态时, 各隐单元的激活状态之间是条件独立的, 此时, 隐层第 j 个节点的激活概率为

$$p_\theta(h_j = 1|\boldsymbol{v}) = \sigma\left(b_j + \sum_i v_i W_{ij}\right) \quad (4)$$

由于 RBM 的结构是对称的, 因此当给定隐单元状态时, 各可见单元的激活状态也是条件独立的, 第 i 个可见单元的激活概率为

$$p_\theta(v_i = 1|\boldsymbol{h}) = \sigma\left(a_i + \sum_j h_j W_{ij}\right) \quad (5)$$

以上为可见层为 0-1 分布的情况, 若可见层为实数 (本文中任务输入为实数), 则采用高斯分布, 此时的能量函数变为

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{ij} W_{ij} v_i h_j - \frac{1}{2}\sum_i (v_i - a_i)^2 - \sum_j b_j h_j \quad (6)$$

隐单元的激活概率公式不变, 可见单元的激活概率变为

$$p_\theta(v_i = 1|\boldsymbol{h}) = \mathrm{N}\left(a_i + \sum_j h_j W_{ij}, 1\right) \quad (7)$$

其中, N 表示均值为 $a_i + \sum_j h_j W_{ij}$、方差为 1 的高斯分布.

在采样步数 t 足够大时, 可以得到服从 RBM 所定义的分布的样本, 但是该方法训练效率仍然不高. 2002 年, Hinton[15] 提出了 RBM 的一个快速学习

算法, 即对比散度 (Contrastive divergence) CD-n 算法, 仅需 n (通常 $n=1$) 步, Gibbs 样本便可得到足够好的近似. 在算法初始阶段, 可见单元的状态被设置成一个训练样本, 然后利用式 (4) 计算所有隐层单元的激活概率并抽取其二值状态. 在所有隐层单元的状态确定之后, 根据式 (5) 或式 (7) 确定第 i 个可见单元 v_i 取值为 1 的概率, 进而产生可见层的一个 "重构 (Reconstruction)". 这样, 在使用随机梯度上升法最大化对数似然概率时, 各参数的更新规则为

$$\Delta W_{ij}=\varepsilon\left(\langle v_i h_j\rangle_{\text{data}}-\langle v_i h_j\rangle_{\text{recon}}\right) \tag{8}$$

$$\Delta a_i=\varepsilon\left(\langle v_i\rangle_{\text{data}}-\langle v_i\rangle_{\text{recon}}\right) \tag{9}$$

$$\Delta b_j=\varepsilon\left(\langle h_j\rangle_{\text{data}}-\langle h_j\rangle_{\text{recon}}\right) \tag{10}$$

其中, ε 为学习速率, $\langle\cdot\rangle_{\text{recon}}$ 表示一步重构后模型定义的分布期望.

2) 基于批量梯度下降的参数微调

BP 算法是常用来对神经网络参数进行微调的有监督学习算法, 因为预测误差由输出层向输入层反向逐层传播而得名. 基于最小均方误差准则, 优化方法是梯度下降法. 梯度下降法有两种迭代求解思路: 随机梯度下降 (Stochastic gradient descent) 和批量梯度下降 (Batch gradient descent). 随机梯度下降通过计算每条样本的损失函数并对参数求偏导得到对应的梯度, 以此来迭代更新参数, 即最小化每条样本的损失函数; 批量梯度下降则是最小化所有训练样本的损失函数, 每次迭代要用到所有训练集的数据. 对比而言, 随机梯度下降法的迭代速率要高于批量梯度下降法, 但不是每次迭代得到的损失函数都朝着全局最优的方向, 对于有多个极小值的非凸问题则可能收敛到局部最优; 而批量梯度下降法最终求解的是全局最优解. DSN 的训练[16] 基于批量梯度下降法, 同时采用矩阵计算的形式, 便于实现算法的并行计算.

网络训练的目标是更新 W 和 U, 使输出矩阵 Y 与标注数据 T 的平方误差 E 最小:

$$E=\|Y-T\|^2=\text{tr}\left[(Y-T)(Y-T)^{\text{T}}\right] \tag{11}$$

式中, tr 表示求矩阵的迹. E 关于 U 的偏导数即梯度为

$$\frac{\partial E}{\partial U}=2H\left(U^{\text{T}}H-T\right)^{\text{T}} \tag{12}$$

令该梯度为 0, 因该函数是一凸优化问题, 可以得到 U 的闭合形式的解:

$$U=\left(HH^{\text{T}}\right)^{-1}HT^{\text{T}} \tag{13}$$

U 的值与 W 有关, 因为 H 由 W 确定, 因此, 计算 E 关于 W 的梯度需要考虑 W 与 U 的关联关系, 将式 (13) 代入 E 关于 W 的梯度计算公式, 可以得到:

$$\frac{\partial E}{\partial W}=2X\Big[H^{\text{T}}\circ(1-H)^{\text{T}}\circ \left[H^{+}\left(HT^{\text{T}}\right)\left(TH^{+}\right)-T^{\text{T}}\left(TH^{+}\right)\right]\Big] \tag{14}$$

其中, $H^{+}=H^{\text{T}}(HH^{\text{T}})^{-1}$, 符号 "$\circ$" 代表元素相乘的内积运算.

上式中, 每条样本在每次迭代中所起的作用一样. 为了加快算法的收敛速率, 文献 [16] 引入一个权重矩阵 Λ 以加强对预测误差大的样本的关注. Λ 为一对角阵, 对角线上的元素 $\lambda_{ii}=(N/E\|\boldsymbol{y}_i-\boldsymbol{t}_i\|^2+1)/2$, 其中, i 为样本索引, N 为样本数量, 样本预测误差越大, 其对应的 λ_{ii} 值越大, 由此使得预测误差大的样本在全部样本的预测误差中具有较大的权重, 即在参数更新中起更主要的作用. 如果该样本训练误差降低了, 在下次迭代中该样本所占权重也会变小. 通过这种方式使参数向误差最有效降低的方向更新, 提升了算法的收敛速度, 同时一定程度上降低了网络学习陷入局部最优的可能. 我们在此基础上增加了关于样本数量不均衡问题的调整, 即在计算对角阵元素 λ_{ii} 时不仅考虑该样本预测误差的影响, 同时考虑该类型样本数目的影响, 样本数较少的数据在全局误差中赋予一个比大类别样本更大的权重, 从而均衡不同类别样本在参数更新中所起的作用. 调整后的 λ_{ii} 变为 $\lambda_{ii\_ad}=(N/E\|\boldsymbol{y}_i-\boldsymbol{t}_i\|^2/(N_k+1)+1)/2$, N_k 为该样本所属类型的子集大小, k 指示样本类别. 引入 Λ 后, 平方误差 E 变为 $E=\text{tr}[(Y-T)\Lambda(Y-T)^{\text{T}}]$, U 的最优解变为

$$U=\left(H\Lambda H^{\text{T}}\right)^{-1}H\Lambda T^{\text{T}} \tag{15}$$

E 关于 W 的梯度公式变为

$$\frac{\partial E}{\partial W}=2X\Big[H^{\text{T}}\circ(1-H)^{\text{T}}\circ \left[H^{+}\left(H\Lambda T^{\text{T}}\right)\left(TH^{+}\right)-\Lambda T^{\text{T}}\left(TH^{+}\right)\right]\Big] \tag{16}$$

其中, $H^{+}=\Lambda H^{\text{T}}(H\Lambda H^{\text{T}})^{-1}$.

2 基于 VDSN 的言语情感计算

言语情感计算是赋予计算机感知、理解和表达情感的能力的重要手段, 也是人机交互过程需要解决的关键问题之一. 本文以从文本计算言语情感为例, 说明可见化深层网络的建模过程以及中间层部分可见化的作用.

2.1 方案描述

言语情感通常与其他情感一样被描述为高兴、悲伤、愤怒、惊讶等若干离散的状态[17−18], 言语情感检测也通常被当作一个多分类任务[19−21], 对于可能影响言语情感产生及发展的各种因素及中间过程则经常被忽略. 近年来, 一种新的情感理论 — 认知评估理论[22], 逐渐被情感计算的研究者们接受并采纳. 在该理论中, 情感被定义为持续一段时间的连续过程, 而非某些离散状态. 情感由对过去经验和当前情境的评估触发, 涉及认知、动机、感受、生理、表情等一系列变化, 评估结果决定其他部分的反应, 反应结果又会反馈回去影响评估的结果. 由此, 情感由评估结果和反应模式共同表示, 不再拘泥于有限数目的离散状态, 而是形成情感的分布式表示, 内部构造及其发展衍化的过程也可得到体现. 基于认知评估理论, 我们提出了专门针对言语情感产生机制的描述模型[23]. 模型从认知、心理、生理和发音四个视角刻画情感, 各视角内容分别表示为一个由若干维度支撑的超平面; 各超平面构成体现言语情感内部复杂结构的多层立体空间. 同时, 基于朗读学[24]和播音学[25] 中创作主体从接触文稿刺激到将其转换为携带恰当感情色彩的有声语言的创作过程, 我们将基于文本生成言语情感的过程概括为认知评价 → 心理感受 → 生理反应 → 发音调整四步, 分别对应着言语情感生成过程中在认知、心理、生理和行为上发生的变化.

言语情感的产生过程是连续的、动态的, 各步骤之间存在直接或间接的相互影响, 前面步骤的结果会影响后续步骤的反应, 后续步骤的反应又会反馈回去进一步影响前面成分的变化 (出于简化计算考虑, 这里暂不考虑反馈). 图 2 给出了在不考虑反馈的情况下, 从输入文本生成言语情感各环节内容的过程及其内部关系. 文本分析得到的特征作为原始输入特征, 言语情感产生过程的每一步对应深层堆叠网络的一个模块. 文本特征会依次传送给每个模块, 同时, 每个模块的生成结果也会对其后续模块产生影响. 当以发音方式作为最终目标, 其他三个步骤则为中间过程; 类似地, 每个环节都可称作其后续环节的中间过程. 当直接从文本特征预测各部分内容而不考虑其中间环节的影响时, 即为中间层 “隐含” 的情况.

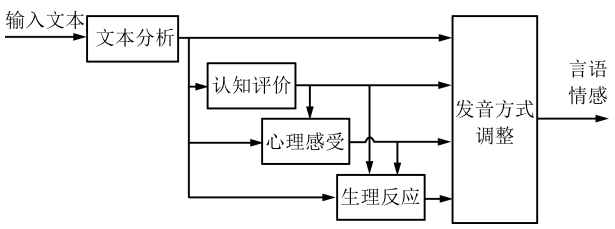
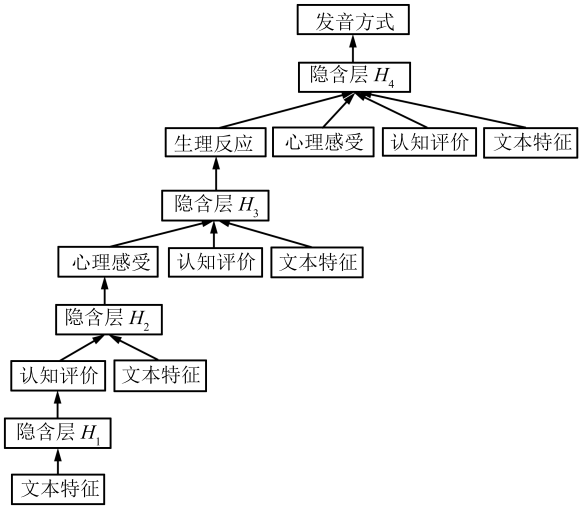


图 2 基于文本的言语情感产生过程及其内部关系示意图

Fig. 2 The framework of the text-based producing process for the components of speech emotion and the interactions among them

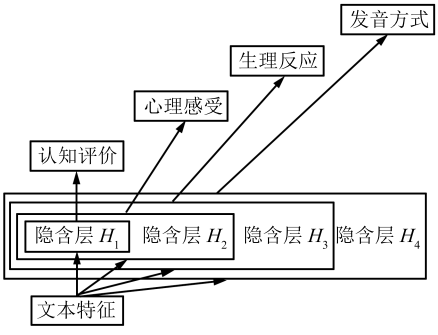
2.2 网络结构

基于图 2 所示的言语情感产生过程, 采用两种 VDSN 搭建了文本–情感计算模型. 当采用 IVDSN 时, 前面模块的输出层依次堆叠到后面模块的输入层, 如图 3 (a) 所示. 每个模块为含单一隐含层的神经网络, 由下至上依次对言语情感的四种成分进行预测. 采用 HVDSN 时, 前面模块的隐含层逐层堆叠作为后面模块部分隐含层的初始值, 如图 3 (b) 所示, 同样采用含单一隐含层的神经网络作为基本模块. 可以看出, IVDSN 形成明显的多层级联型网络, 而 HVDSN 则构成嵌套型的网络, 整体结构仍为单



(a) 基于 IVDSN 的言语情感计算

(a) The computing network based on IVDSN



(b) 基于 HVDSN 的言语情感计算

(b) The computing network based on HVDSN

图 3 基于 VDSN 的言语情感计算网络结构

Fig. 3 The computing networks of speech emotion based on VDSN

隐层神经网络. 不同于传统神经网络由隐含层数目决定网络深度的方式, 堆叠网络作为深层网络时网络深度由模块数决定, 其中, IVDSN 和 HVDSN 的模块数与经由的中间环节的个数有关, 本例中二者最终都构成了深度为 4 的深层网络.

两种网络对于未知信息和已知信息的获取及处理方式不同. IVDSN 中, 已知信息与未知信息均位于输入层, 已知信息来源于子目标的输出, 未知信息仍隐含于文本特征中; 而 HVDSN 的已知信息与未知信息均位于隐含层, 二者均是对文本特征的非线性表示, 差别仅在于已知信息由其他子目标进行了有监督微调, 而未知节点仅做了无监督预训练 (这一部分节点单独做无监督预训练, 不包含已知节点). 两种网络的训练方式上文已提到, 即各模块按生成顺序依次单独训练, 分别进行本模块内部的无监督预训练和有监督微调, 每个模块有各自不同的训练目标, 输出层每个节点对应各自子目标的一个维度, 其中, 认知评价由 5 个维度表示; 生理反应包含 2 个维度; 发音描述包含 7 个维度; 心理感受采用分层表示的方法描述对于感受类别不同粒度的划分, 每个类别作为心理感受的一个维度, 由上至下依次有 4 维、7 维、19 维和 43 维 (本文实验采用最细致的划分, 即 43 维).

2.3 特征提取与降维

文本特征作为深层神经网络的原始输入, 其维度决定输入层的节点数, 进而影响网络的规模和学习复杂度. 向量空间模型 (Vector space model, VSM) 是最常用的文本特征表示方式, 将文档映射到几千维的词典空间, 采用词典词在当前文档中出现的词频或其他变体形式表示文档, 是一种稀疏的文本表示方法. 直接以此为输入特征会增加网络空间和计算时间的开销, 当标注数据有限时还易造成过拟合.

本文使用狄里克雷分配 (Latent Dirichlet allocation, LDA) 模型[26] 来对文本的向量空间表示进行降维. LDA 模型是文本语义分析模型的一种, 也被称作主题模型, 通过在文档与词语之间增加一个语义空间来抽取文档的语义信息, 该空间的每个维度称作一个主题, 每个主题由词语的概率分布表示, 每篇文档则表示成这些主题的概率分布, 文档–主题分布被用作文本特征的降维表示.

主题数 (文本特征的维度) 由模糊度确定[26]. 模糊度是常用来评价模型对数据刻画能力的指标, 它与单个词语的平均似然度成反比, 值越低则表示模型对数据的刻画能力越好. 模糊度的计算公式为

$$perplexity = \exp\left(-\frac{\sum_{d}^{D}\log\left(p\left(w_d\right)\right)}{\sum_{d}^{D}N_d}\right) \tag{17}$$

其中, D 为数据集的文档数, N_d 为每篇文档包含的词语数, $\log p(w_d)$ 表示文档中每个词语的对数似然概率.

3 网络性能测试

3.1 数据集与评价指标

测试数据来自中央人民广播电台《新闻与报纸摘要》节目的播音稿, 以篇章为单位, 总共收录 29 109 篇新闻播音稿, 时间跨度从 2006 年到 2013 年, 内容涵盖教育、医疗、科学、经济、文化、艺术、娱乐、军事以及政治等各领域, 文档平均长度为 227 字/篇. 综合考虑篇幅、内容、情感类型和强度等因素, 从中筛选出 150 篇有代表性的语篇进行人工标注. 标注人员由同一课题组的三名师生担任, 三人独立标注, 标注结果通过高斯加权整合成最终的标注信息. 标注内容基于言语情感的多视角描述模型, 其中认知评价和发音描述的每个维度具有正负极性, 每个极性分为弱、中、强不同的强度, 加上中性构成七级刻度; 心理感受和生理反应的每个维度分为无、弱、中、强四级刻度. 由于输入文本特征为 [0, 1] 区间内的实数, 为方便输出值在输入层的堆叠, 需对标注数据进行基于刻度范围的归一化处理, 其中, 认知和发音每个维度的标注范围为 [−3, 3], 因此标注值均除以 6; 心理和生理的标注范围为 [0, 3], 标注值除以 3.

因为情感的多维空间表示, 采用能够刻画空间距离的均方根误差 (Root-mean-square-error, RMSE) 作为网络性能的评价指标, RMSE 越小, 网络性能越好. 计算公式为

$$RMSE = \sqrt{\frac{\sum_{i}^{N}\left\|\boldsymbol{y}_i - \boldsymbol{t}_i\right\|^2}{N}} \tag{18}$$

其中, i 为样本索引, N 为测试样本数, $\boldsymbol{y}_i$ 为输出向量, $\boldsymbol{t}_i$ 为标注数据.

3.2 参数设置

3.2.1 文本特征维度

第 2.3 节提到, 文本特征维度即主题数根据模糊度确定. 以 25 为步长, 在 [25, 200] 范围内, 我们测试了上述数据集的模糊度随特征维度的变化, 如图 4 所示. 从图 4 可以看出, 模糊度在特征维度等于 125 时达到最小, 说明此时文本模型对数据集的

刻画能力最佳. 因此, 在之后的实验中我们将文本特征维度设为 125.

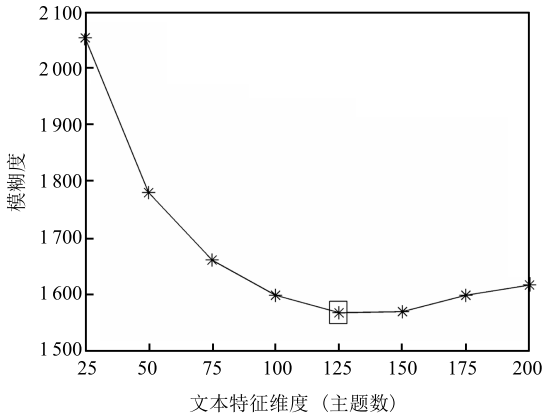


图 4 不同文本特征维度下文本模型模糊度分布曲线

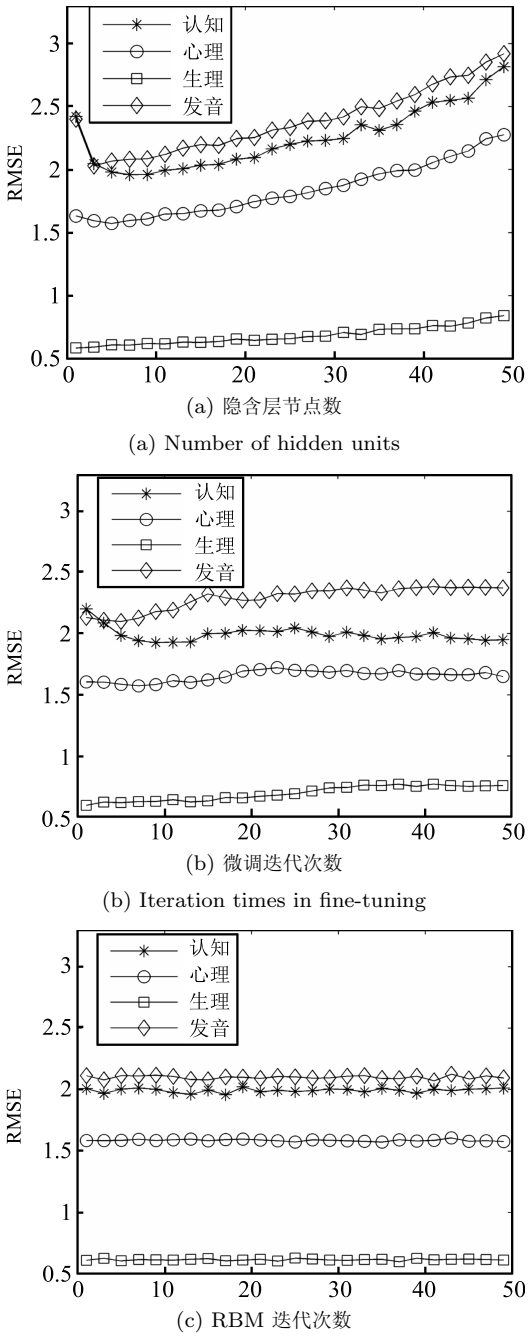
Fig. 4 The distributed curves of perplexities of the textual modules with different dimensionalities

3.2.2 网络训练参数

以从文本特征直接生成言语情感各部分内容而不考虑其他环节影响的预测结果为评估指标 (此时两种可见网络没有差别), 分别测试了隐含层节点数、微调阶段的迭代次数以及采用 RBM 进行预训练的迭代次数三个参数的影响.

因为标注样本有限, 采用五折交叉验证 (5-fold cross validation) 的方法进行网络的性能验证, 即: 将全部标注数据随机分成五等份, 每次 (One fold) 取一份作为测试集, 其余部分作为训练集 (预训练时加入无标注数据). 此外, 由于 RBM 初始化过程中的随机性会造成结果的振荡, 因此五折交叉验证重复运行 10 次, 取 10 次结果的均值作为最终的结果. 以下实验均采用同样的实验方式.

图 5 (a) 给出隐含层节点数变化对于四种情感成分预测结果的影响, 此时微调和 RBM 的迭代次数均设为 5. 可以看出, 随着隐含层节点数的增加, 认知、心理和发音三种成分的 RMSE 均先下降之后上升, 可见隐含层节点数并非越多越好, 尤其在当前训练数据有限的情况下, 隐含层节点过多很容易造成网络的过拟合. 由于 HVDSN 的隐含层节点数是逐层扩张的, 在避免过拟合的同时也要兼顾每个模块的隐含层节点不会过少而造成欠拟合, 因此将最底层认知模块隐含层节点数设为 4, 依次递增 2, 即心理、生理和发音模块的隐含层节点数分别设为 6、8、10. 为了使结果具有可比性, 其他网络的隐含层节点数设置与 HVDSN 相同. 图 5 (b) 给出微调阶段的迭代次数对预测结果的影响, 此时隐含层节点数按前面实验得到的数据设置, RBM 迭代次数设为 5. 可以看出, 迭代次数较少时各成分的 RMSE 变化不明显或有所下降, 随着迭代次数进一步增加, RMSE 均有所上升, 因此将微调阶段迭代次数设为 5. 图 5 (c) 为 RBM 的迭代次数对于训练结果的影响. 可以看出, 四种成分的预测结果变化均不显著, 因此也将其设置为 5, 进一步增加迭代次数会降低训练速度且不会使结果有明显改善.



(a) 隐含层节点数

(a) Number of hidden units

(b) 微调迭代次数

(b) Iteration times in fine-tuning

(c) RBM 迭代次数

(c) Iteration times of RBM

图 5 不同网络训练参数下四种言语情感成分的预测结果 (RMSE)

Fig. 5 The RMSEs of the four affective components with different network training parameters

3.3 中间层可见化效果验证

网络的性能由两个实验进行测试和验证, 实验 1 对从文本特征生成言语情感各部分内容的所有可能路径进行测试和对比, 从而验证中间环节对于后续任务的影响, 即中间层已知对预测任务的作用; 实验 2 将 VDSN 与其他中间层不可见的深层网络对比, 从而进一步验证中间层部分可见化对于网络的优化作用.

3.3.1 中间环节的作用

实验 1 以从文本特征直接生成言语情感各成分而不考虑其他环节影响的情况作为基准 (表 1 中斜体行), 分别与考虑了中间环节的各种情况作对比. 考虑中间环节的情况分为包含全部中间环节和部分中间环节的多种情况, 表 1 列出了不考虑反馈情况下各成分所有可能的预测方式和结果, 分别采用 IVDSN 和 HVDSN 两种网络.

表 1　IVDSN 和 HVDSN 采用不同路径预测言语情感各成分结果 (RMSE)

Table 1　The predicted results (RMSE) of each emotional component through different paths by IVDSN and HVDSN

	IVDSN	HVDSN
文本 → 认知	**2.07**	**2.07**
文本 → 心理	1.59	1.58
文本 → 认知 → 心理	**1.55**	**1.55***
文本 → 生理	0.63	0.64
文本 → 认知 → 生理	0.62	0.64
文本 → 心理 → 生理	0.62	0.62
文本 → 认知 → 心理 → 生理	**0.62**	**0.65**
文本 → 发音	2.16	2.16
文本 → 认知 → 发音	2.11	2.11
文本 → 心理 → 发音	2.13	2.11
文本 → 生理 → 发音	2.11*	2.11**
文本 → 认知 → 心理 → 发音	2.09**	2.08*
文本 → 认知 → 生理 → 发音	2.11***	2.09*
文本 → 心理 → 生理 → 发音	2.12**	2.09*
文本 → 认知 → 心理 → 生理 → 发音	**2.11****	**2.09**

*: p-value < 0.05;　**: p-value < 0.01;　***: p-value < 0.001

从表 1 可以看出, 在采用这两种网络的情况下, 中间环节的加入均可一定程度降低预测误差. 对各模块考虑中间环节的情况与未考虑中间环节的情况作差异显著性分析, 可以得出: 当采用 IVDSN 时, “文本 → 生理 → 发音”、“文本 → 认知 → 心理 → 发音”、“文本 → 认知 → 生理 → 发音”、“文本 → 心理 → 生理 → 发音” 以及 “文本 → 认知 → 心理 → 生理 → 发音” 的结果均优于 “文本 → 发音” 的结果, 且差异显著, 说明增加中间环节有助于发音方式的预测; 当采用 HVDSN 时, “文本 → 认知 → 心理” 的预测效果要显著优于 “文本 → 心理”, 说明认知的加入有助于心理感受的预测, “文本 → 生理 → 发音”、“文本 → 认知 → 心理 → 发音”、“文本 → 认知 → 生理 → 发音”、“文本 → 心理 → 生理 → 发音” 的预测效果也显著优于直接从文本预测发音. 心理和认知对于生理的预测结果提升作用不明显, 原因可能是由于该成分预测误差本来就小, 因此可提升空间不大. 这些结果反映出中间环节对于后续任务的影响, 考虑中间环节的影响可以提升网络的预测效果. 最终确定的言语情感四种成分的预测方式如表 1 中粗体行所示, 按生成顺序将结果逐步累积参与到后续目标的预测, 与图 3 所示网络拓扑结构一致.

3.3.2 与其他深层神经网络对比

实验 2 中, DBN 作为未对中间层进行监督且未引入先验知识的神经网络, DSN 作为对中间层进行监督但未引入先验知识的网络, IVDSN 和 HVDSN 则为对中间层进行监督且引入先验知识的网络, 即中间层部分可见的深层网络. 通过 DSN 与 DBN 的对比, 可验证选用 DSN 作为基础网络, 即增加对中间层监督的必要; IVDSN、HVDSN 分别与 DSN 比较则可以验证引入先验知识将中间层部分可见化的效果; 另外, IVDSN 与 HVDSN 也可以进行比较以测试两种堆叠方式的优劣. 除此之外, 实验 2 还考虑了网络深度对其性能的潜在影响. 对于 DBN 来说, 网络深度由隐含层数目决定, DSN、IVDSN 和 HVDSN 的网络深度则由模块数决定.

表 2 列出了网络深度从 1 到 4 的不同网络由文本特征预测发音方式的结果. 其中, IVDSN 和 HVDSN 网络深度为 1 表示从文本直接预测发音的情况; 网络深度为 2 表示从文本经认知或心理或生理某个中间环节预测发音的情况, 表 2 中结果为三种情况的平均值; 网络深度为 3 表示从文本经认知、心理、生理中某两个中间环节预测发音的情况, 共有三种组合方式 (如表 1 所示), 表 2 中结果为三种情况的平均值; 网络深度为 4 则表示从文本经认知、心理和生理三个中间环节预测发音的情况.

DSN 相对 DBN 预测效果有明显提升 (p-value $= 0.0002$), 验证了对中间层进行监督对于深层神经网络性能的优化作用. IVDSN 与 DSN 相比, 性能又有进一步提升 (p-value $= 0.02$), 说明在对中间层进行监督的基础上, 不同子目标的设定可以进一步提升网络性能; 同理, HVDSN 的性能也显著优于 DSN (p-value $= 0.03$). 在 HVDSN 与 IVDSN 的对比中, HVDSN 的性能更优 (p-value $= 0.04$), 一方面可能由于 IVDSN 输入层维度多于 HVDSN 的输入层, 因此网络规模上 HVDSN 更精简, 这在数据有

限情况下会对训练结果产生一定影响; 另一方面可能与已知信息和未知信息的相对作用有关, HVDSN 中, 已知信息与未知信息的节点数相差不大, 且已知节点占主导, 而 IVDSN 中, 两部分特征维度相差较大且未知信息占主导, 因此 IVDSN 中已知信息的作用可能被削弱, 这也从侧面反映出信息可见化的重要; 此外, 数值分布情况也可能对结果造成影响, HVDSN 中, 已知信息与未知信息属于同样的数值分布 (均为隐含层节点), 而 IVDSN 中, 子目标输出值与文本特征的分布不同, 可能会对两部分信息的融合产生影响.

表 2 不同深层神经网络在不同深度条件下发音方式预测结果 (RMSE)

Table 2 The predicted results (RMSE) of utterance manner by different deep networks and in different depths

Depth	DBN	DSN	IVDSN	HVDSN
1	2.28	2.17	2.16	2.16
2	2.27	2.13	2.12	2.11
3	2.27	2.15	2.11	2.09
4	2.27	2.15	2.11	2.09

纵观网络深度对各深层网络的影响, DBN 和 DSN 随网络深度的增加对网络性能的提升效果不如 IVDSN 和 HVDSN 明显, 这意味着在数据规模一定的情况下, 单纯增加网络深度并不总是能提升网络的性能, 在此基础上增加一定的先验知识的指引, 可以进一步优化网络的结构.

4 讨论

子目标的引入可以视作在最终目标的基础上引入了联合目标, 网络训练过程中既要满足最终目标的误差最小, 同时也要满足各子目标的误差最小. 联合目标的引入相当于在经验风险上增加了一个正则项, 即在最小化经验误差的同时约束网络的结构. 从贝叶斯估计的角度来看, 该约束对应于模型的先验分布. 如果先验合适, 则得到的学习结果更倾向于真解, 这也解释了引入子目标之后测试误差更小的原因. 联合目标的训练可以先于最终目标而分步完成, 也可与最终目标同步训练. 文献 [9−10] 在对中间层进行监督或半监督学习时都采用联合目标与最终目标同步训练的做法. 同步训练的方式沿用深层神经网络全局运用反向误差传递的训练模式, 代价函数仍是一个含多个极小值的高度非凸空间, 因此可能使网络最终收敛于局部最小; 同时, 在反向传播的过程中还可能发生梯度消失的情况. 我们采用分步训练的方式, 每个模块为一浅层神经网络, 代价函数由高度非凸变成凸函数或近似凸函数, 降低了网络训练复杂度, 并提升了网络收敛于全局最优的可能.

子目标的设定涉及到先验知识的融合与运用, 本文中, 我们大量引入了相关领域的研究成果, 汇总出可能与目标任务有关的多方面影响因素, 并进一步推断出其可能的相互关系和组织结构, 最后通过实验验证该结构的合理性. 除此之外, 子目标还可有另一种解读, 如在自然语言处理中常涉及的韵律层级结构, 低层小尺度单元的训练任务可视作高层大尺度单元的子目标, 从而由细到粗汇总出整个句子或篇章的训练结果 (句子级情感分析常用的做法); 或者反之, 高层大尺度单元也可先于低层小尺度单元进行处理, 从而为小尺度单元提供上下文参考. 在图像处理领域, 这种由细到粗或由粗到细的层级结构仍然存在, 在人脸识别的研究中已经发现, 某些中间层学习到了组成人脸的结构特征 (边缘或器官), 验证了人脑视觉系统由具体到抽象逐层组合低层特征的分级处理模式. 但是, 基于深层网络自动发现文本或图像中结构特征的研究都离不开大规模数据的支持, 组建的网络规模也很庞大, 我们希望通过人为设定某些中间层的子目标, 从而使网络能快速高效地学习到期待的特征, 减少数据和网络规模的开销.

5 总结与展望

针对深层神经网络中间层不可追踪、难以解释的问题, 本文提出两种中间层部分可见的深层神经网络, 利用先验知识有针对性地指引深度学习的中间过程. 网络基于深层堆叠网络 (DSN) 构建, 根据堆叠的位置不同, 提出输入层部分可见 (IVDSN) 和隐含层部分可见 (HVDSN) 两种堆叠网络. 以言语情感计算为例, 明确言语情感产生过程中的相关环节及其相互影响, 并将其与深层堆叠网络融合而使深层网络的中间层可见化, 即具有明确的意义和显性的影响关系. 实验表明, 中间层的可见化对于深层网络的性能有优化作用; IVDSN 和 HVDSN 两种网络均可实现中间层部分可见化的建模, 且均具有良好的可扩展性和简便易操作的训练过程, 其中 HVDSN 结构更精简, 性能也更优.

本文中给出的是中间层可见的深层堆叠网络在级联型生成关系中的应用实例, 除此之外, 还可应用于层级结构等包含关系. 子目标的获取依赖于领域知识, 领域知识的引入在训练资源有限的情况下很有必要. 目前, 两种网络的基本模块均为单隐层结构. 在今后的研究中, 还可根据需求增加每个基本模块的隐含层个数, 即每个模块也是一个多层神经网络, 从而提升对每个子目标的学习效果. 对于 IVDSN 网络, 增加隐含层个数不影响网络模块间的堆叠方式, 而 HVDSN 模块间可考虑全部继承或部分继承前面模块的隐含层, 并通过增加每个隐含层

节点数或继续增加隐含层的个数来水平扩张或垂直扩张隐含层.

References

1 Yoo H J. Deep convolution neural networks in computer vision: a review. *IEIE Transactions on Smart Processing and Computing*, 2015, **4**(1): 35−43

2 Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, 2014. 1717−1724

3 Zhang C, Zhang Z Y. Improving multiview face detection with multi-task deep convolutional neural networks. In: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). Steamboat Springs, CO: IEEE, 2014. 1036−1041

4 Sainath T N, Kingsbury B, Saon G, Soltaua H, Mohamed A, Dahlb G, Ramabhadran R. Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 2015, **64**: 39−48

5 Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: an overview. In: Proceedings of the 2013 International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 8599−8603

6 Bengio S, Heigold G. Word embeddings for speech recognition. In: Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech. Singapore: ISCA, 2014. 1053−1057

7 Le Q V, Mikolov T. Distributed representations of sentences and documents. In: Eprint Arxiv, 2014. 1188−1196

8 Kiros R, Zemel R S, Salakhutdinov R. A multiplicative model for learning distributed text-based attribute representations. In: Eprint Arxiv, 2014. 2348−2356

9 Lee C Y, Xie S N, Gallagher P, Zhang Z, Tu Z W. Deeply-supervised nets. In: Eprint Arvix, 2014. 562−570

10 Weston J, Ratle F, Mobahi H, Collobert R. Deep learning via semi-supervised embedding. *Neural Networks: Tricks of the Trade*. Berlin Heidelberg: Springer, 2012. 639−655

11 Deng L, Yu D, Platt J. Scalable stacking and learning for building deep architectures. In: Proceedings of the 2012 International Conference on Acoustics, Speech, and Signal Processing. Kyoto: IEEE, 2012. 2133−2136

12 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

13 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR: W & CP, 2010. 249−256

14 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

15 Hinton G E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002, **14**(8): 1711−1800

16 Yu D, Deng L. Accelerated parallelizable neural network learning algorithm for speech recognition. In: Proceedings of the 2011 Annual Conference of the International Speech Communication Association. Florence, Italy: ISCA, 2011. 2281−2284

17 Ekman P. An argument for basic emotions. *Cognition and Emotion*, 1992, **6**(3−4): 169−200

18 Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech. *Speech Communication*, 2003, **40**(1−2): 5−32

19 Calvo R A, Mac K S. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 2013, **29**(3): 527−543

20 Trilla T, Alias F. Sentence-based sentiment analysis for expressive text-to-speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(2): 223−233

21 Bellegarda J R. A data-driven affective analysis framework toward naturally expressive speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(5): 1113−1122

22 Moors A, Ellsworth P C, Scherer K R, Frijda N H. Appraisal theories of emotion: state of the art and future development. *Emotion Review*, 2013, **5**(2): 119−124

23 Gao Ying-Ying, Zhu Wei-Bin. A study of a transcription system for speech emotion. *Chinese Journal of Phonetics*, 2013, **4**: 71−81
(高莹莹, 朱维彬. 言语情感描述体系的试验性研究. 中国语音学报, 2013, **4**: 71−81)

24 Zhang Song. *Recitation Science*. Beijing: Communication University of China Press, 2007.
(张颂. 朗读学. 北京: 中国传媒大学出版社, 2007.)

25 Zhang Song. *China Broadcasting Science*. Beijing: Communication University of China Press, 2003.
(张颂. 中国播音学. 北京: 中国传媒大学出版社, 2003.)

26 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 2003, **3**: 993−1022

高莹莹 北京交通大学信息科学研究所博士研究生. 主要研究方向为情感语音合成与机器学习.
E-mail: 10112060@bjtu.edu.cn
(**GAO Ying-Ying** Ph. D. candidate at the Institute of Information Science, Beijing Jiaotong University. Her research interest covers expressive speech synthesis and machine learning.)

朱维彬 北京交通大学信息科学研究所副教授. 主要研究方向为语音识别, 语音合成与机器学习. 本文通信作者.
E-mail: wbzhu@bjtu.edu.cn
(**ZHU Wei-Bin** Associate professor at the Institute of Information Science, Beijing Jiaotong University. His research interest covers speech recognition, speech synthesis, and machine learning. Corresponding author of this paper.)