文章编号:1674-2974(2019)02-0097-08 DOI:10.16339/j.cnki.hdxbzkb.2019.02.014

深度神经网络内部迁移的信息几何度量分析

陈力[1,2],费洪晓[2],李海峰[1†],何嘉宝[2],谭风云[2]
(1.中南大学 地球科学与信息物理学院,湖南 长沙 410083;2.中南大学 软件学院,湖南 长沙 410083)

摘 要:使用深度神经网络处理计算机视觉问题时,在新任务数据量较少情况下,往往会采用已在大数据集上训练好的模型权值作为新任务的初始权值进行训练,这种训练方式最终得到的模型泛化能力更好.对此现象,传统解释大多只是基于直觉分析而缺少合理的数学推导.本文将深度神经网络这种网络结构不变下层间的学习转为深度神经网络内部的迁移能力,并将学习过程变化形式化到数学表达式.考虑数据集对训练过程带来的影响,利用信息几何分析方法,确定不同数据集流形之上的度量和联络,实现不同数据集之间的嵌入映射,同时将参数空间的变化也放入流形空间,探究其对学习过程的共同影响,最终实现对这种内部迁移现象的数学解释.经过分析和实验验证可得内部迁移过程其实是一种能使网络可以在更广空间进行最优搜索的变化,有利于模型可以在学习过程中获得相对的更优解.

关键词:深度学习;迁移学习;信息几何
中图分类号:TP183 **文献标志码**:A

Analysis on Information Geometric Measurement of Internal Transfer of Deep Neural Network

CHEN Li[1,2],FEI Hongxiao[2],LI Haifeng[1†],HE Jiabao[2],TAN Fengyun[2]
(1. School of Geosciences and Info-Physics,Central South University,Changsha 410083,China;
2. School of Software Engineering, Central South University, Changsha 410083,China)

Abstract: When deep learning is used to deal with the computer vision tasks, under little number of new task data, the pre-trained model weight based on a very large data is trained as an initial weight to get better generalization ability. At this point, former explanations are based on the intuitive analysis and lack of reasonable mathematical methods. In this paper, deep neural network, which trains on internal layers with fixed structure,changed into internal transfer ability in deep neural network. The changes of learning process are formalized into a mathematical expression. Considering the influence of the data set on the training process, the information geometric analysis method is used to determine the metrics and connections over manifolds of different data sets, which can realize the embedding mapping between different data sets. At the same time, the change of parameter space is also put into a manifold space to explore its common influence on learning process. Finally, a mathematical explanation is provided for the internal transfer phenomenon. Meanwhile, after the analysis and experiments, the process of internal transfer is identi-

* **收稿日期**:2018-03-20
基金项目:国家自然科学基金资助项目(61602525,61603404,41571397,41501442),National Natural Science Foundation of China(61602525,61603404,41571397,41501442)
作者简介:陈力(1992—),男,陕西汉中人,中南大学博士研究生
† 通讯联系人,E-mail:lihaifeng@csu.edu.cn

fied as a change which can make the network search for optimal search in a wider space. Therefore, the model can obtain a relative better solution in learning process.

Key words: deep learning; transfer learning; information geometry

计算机视觉是人工智能非常重要的研究领域,视觉也是人和动物最重要的感觉,至少有 80%以上的外界信息经视觉获得.大数据环境下,图像与视频类数据增长速度达到前所未有的高度.然而,Science 杂志认为现状可描述为:“data-rich but analysis-poor”[1].如何学习到好的“特征”,一直是计算机视觉中的基础性问题[2].传统图像识别方法中,大多通过设计者的先验知识,手工设计特征,如 SIFT[3],HOG[4]等,往往很难真正捕捉到物体的本征特征.近年来深度学习[5]方法的兴起,在图像识别和理解等诸多任务上,获得许多令人印象深刻的成绩.本质上深度学习可以看成是一种“端到端”的特征学习方法[6],借助强大的计算力,通过大量训练样本从低层特征组合成更加抽象的高级特征来揭示事物的属性和特征,这也是其在图像识别应用中成功的重要原因.

虽然借助深度学习技术[7],图像识别问题取得了突破性进展,但深度学习也存在很多局限性.如果新的视觉任务上缺少大量标注的数据,通常需要消耗大量人力物力对数据进行标注和清洗. 而且在许多特定视觉识别任务中,例如糖尿病视网膜病变分析,大量标定的数据往往没有那么容易获得.这使得深度神经网络在小样本学习问题上表现不是非常好,泛化性较差且容易过拟合.对此深度学习提供了一种迁移学习的方法,确定网络结构,在训练过程中,参数不再随机初始化,可以利用已在大库上学习到的收敛模型,将其权值作为新任务网络模型的初始值进行再训练,将这种神经网络的迁移过程叫做微调[8].深度神经网络迁移学习的结果往往比直接在新任务上重新随机初始化参数训练的收敛速度更快,泛化能力更强.Abràmoff 等人[9]利用深度学习和微调等方法,轻易将糖尿病视网膜病变检测能力提高到专家水平.

微调的方法也具有很强的技巧性,针对不同的任务需要迁移不同的信息.例如在图像任务中,固定前面几层权值信息, 然后重新训练后几层参数,而在语音识别任务中,会固定后几层参数,需要重新训练前几层的参数. 并且迁移的层数对迁移的效果也会产生不同的影响,Yosinski 等人[10]分析,随着固定迁移层的参数增加,识别效果会先上升然后急速下降,因此迁移学习具有不稳定性和不确定性.深度神经网络的训练过程大多采用类梯度下降算法,虽然类随机梯度下降算法在非凸优化问题上对初始点敏感,但在深度学习中,在高维参数的情况下主要以鞍点的形式存在,即得到的局部最小可以近似等于全局最小[11].可是迁移学习这种表面上只影响了初始化参数而得到的效果又十分突出. 现有的分析大多是基于直觉和特征可视化的直观分析,认为在大库上的信息更加丰富,所以卷积核获取的特征也更加丰富,对于新任务图片的响应,只需要在原有响应上做调整就能很轻松地适用于新任务. 而且由于原有丰富的特征表达能力,这种特征表达在新任务学习过程中难以获取,但又对新任务识别具有非常大的帮助作用,这就使得微调效果变得非常好.但这种解释太过于依赖直觉分析,缺少更好的数学解释,这种层间变化过程很难用传统统计学的理论进行分析.

信息几何是一套研究流形内蕴几何性质变化的理论体系[12],它将概率论、统计学和信息论中许多概念当作概率分布空间的几何结构,使用微分几何的手段进行分析.学习过程中,深度神经网络层之间的变化可以看成是参数概率分布的改变,而这种变化也不断受到数据带来的信息影响,这就为使用信息几何的方法分析深度神经网络学习变化过程提供了可行性.

综上, 本文将深度神经网络在视觉任务下,卷积核变化更新以适应新任务的能力称为其内部迁移能力,并将神经网络在迁移过程中的变化进行形式化,使用严谨的数学模型进行表示.学习过程中,主要变化的是学习到的权值信息,将权值信息的变化放入到流形空间中,同时考虑样本空间,利用信息几何度量的方法对其进行分析,并通过实验重现深度神经网络内部迁移变化的过程. 结果表明在大库上学到的模型可以提供更大的参数探索空间,为深度神经网络迁移变化提供一种可能的数学解释.

1 深度神经网络内部迁移

1.1 深度神经网络符号定义

深度神经网络通常由多个隐含层堆叠而成,输入层所有神经元的连接都与输出层神经元相连.假设有训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,$x_n$ 表示第 n 个样本的输入信息,y_n 表示对应的标签信息. 神经网络由 d+1 层构成,分别记为第 0,1,$\cdots$,d 层.每层的节点数目分别为 $m_i(i=0,1,\cdots,d)$.在本文中进一步假设第 d 层的节点数目为 1,即输出为一维数值.假设每个节点的激活函数表示为 σ, 根据需要可以选择 Sigmoid 函数或者其它 threshold 函数等, 使偏置 $b=x_0$ 对应权值 $w_0=1$,每层线性组合统一为 $\sum_{j=0}^{m_i} w_j x_j$.神经网络最核心的要素是层与层之间的连接权重矩阵,假设第 i 层与第 $i+1$ 层之间的连接权重矩阵 $w_{i(i+1)}\in M_{m_i\times m_{i+1}}(R)(i=0,1,\cdots,d-1)$.假设 x 是一个输入向量,那么输出数值为 $\sigma(W^{\mathrm{T}}_{(d-1)d}\sigma(W^{\mathrm{T}}_{(d-2)(d-1)}\cdots\sigma(W^{\mathrm{T}}_{0\times 1}x)\cdots))$.为了表述方便,定义集合 M 表示与神经网络匹配的连接权重矩阵集合,W 表示集合 M 中的一个元素,具体如下所示:

$$M=M_{m_0\times m_1}(R)\times\cdots\times M_{m_{d-1}\times m_d}(R)$$
$$W=(W_{0\times 1},\cdots,W_{(d-1)d}),W_{i(i+1)}\in M_{m_i\times m_{i+1}}(R) \tag{1}$$

此时神经网络的学习过程简单表述为

$$F_w(x)=\sigma(W^{\mathrm{T}}_{(d-1)d}\sigma(W^{\mathrm{T}}_{(d-2)(d-1)}\cdots\sigma(W^{\mathrm{T}}_{0\times 1}x)\cdots)) \tag{2}$$

1.2 内部迁移学习

在计算机视觉任务中,神经网络在大数据集上学习的信息往往比在小数据集上学习的信息更加完备.通常将在 ImageNet[13]上学习的模型作为新任务的预训练模型,使用训练好的模型的权值信息进行初始化训练,网络结构并不需要发生变化.网络逐层的调整权值以适应新任务的需求,这种内部变化的过程叫做深度神经网络的内部迁移学习.

为了更好地分析网络的内部迁移变化,首先将学习过程进行数学抽象.对于训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其中数据集 $X=\{x_1,\cdots,x_n\}=\{x_i\}_{i=1}^n$,每个数据对应相关的数据类别,数据的类别标签是一个用来区分数据的映射, 使用 Y 表示标签集,$Y=\{y_1,\cdots,y_n\}=\{y_i\}_{i=1}^n$,$C_Y$ 表示标签类别数量.则上述映射过程可以表示成

$$\begin{aligned}&\mathrm{label}:X\to Y\\&\mathrm{s.t.},\mathrm{label}(x_i)\in Y\quad\forall i=1,\cdots,n\end{aligned} \tag{3}$$

因此给定一个数据集 X, 可以产生一个与之对应的标签向量

$$\mathrm{label}(X)=(\mathrm{label}(x_i),\cdots,\mathrm{label}(x_n))\in R^{C_Y} \tag{4}$$

当给定一个神经网络, 数据集输入神经网络可以得到输出向量,神经网络的学习过程如下所示,

$$F_x(x)=(F_w(x_1),\cdots,F_w(x_n))\in R^{C_Y} \tag{5}$$

数据集的学习可以表示成在整个假设空间中,选择合适的目标函数,并在给定一个损失函数 loss:$R_+\to R$ 最小的情况下, 寻找神经网络的最优连接权重矩阵, 即如下的优化问题:

$$\begin{aligned}&\min\sum_{x_i\in X}\mathrm{loss}(|\mathrm{label}(x_i)-F_w(x_i)|)+\alpha\Omega(w)\\&\mathrm{s.t.}\quad w\in m\end{aligned} \tag{6}$$

其中 $\Omega(w)$表示正则化项,上式可等价描述为

$$\mathrm{Argmin}_{w\in m}\sum_{x_i\in X}\mathrm{loss}(|\mathrm{label}(x_i)-F_w(x_i)|)+\alpha\Omega(w) \tag{7}$$

由公式(7)可知,学习过程求解最优解属于一个无约束的优化问题. 如果神经网络中选择的激活函数足够光滑(如 Sigmoid 函数),或者写成求偏导数的形式,可以利用既有的优化算法,如随机梯度下降法等实现求解.

此时定义内部迁移学习过程. 给定一个神经网络如下所示:

$$F_w(\cdot)=\sigma(W^{\mathrm{T}}_{(d-1)d}\sigma(W^{\mathrm{T}}_{(d-2)(d-1)}\cdots\sigma(W^{\mathrm{T}}_{0\times 1}x)\cdots)) \tag{8}$$

给定一个损失函数 loss:$R_+\to R$. 给定两个数据集 A,B. 其中 A 数据集较大, 表现在类别丰富度更多,数据量更大.具体可以表示为

$$\begin{aligned}&A=\{a_1,\cdots,a_{n_1}\}=\{a_i\}_{i=1}^{n_1}\\&B=\{b_1,\cdots,b_{n_2}\}=\{b_j\}_{j=1}^{n_2}\end{aligned} \tag{9}$$

根据上文可得关于数据集 A 的学习模型如下所示:

$$\mathrm{Argmin}_{w\in m}\sum_{a_i\in A}\mathrm{loss}(|\mathrm{label}(a_i)-F_w(a_i)|)+\alpha\Omega(w) \tag{P_1}$$

同理可得数据集 B 的学习模型如下所示:

$$\mathrm{Argmin}_{w\in m}\sum_{b_j\in B}\mathrm{loss}(|\mathrm{label}(b_j)-F_w(b_j)|)+\alpha\Omega(w) \tag{P_2}$$

首先求解问题 ,过程如下:

Step 1.确定算法 G;

Step 2.确定停止准则 S;

Step 3.随机选定初始连接权重矩阵 $W_0 \in M$;

Step 4.从 W_0 开始按照停止准则 S 运行算法 G 迭代,得到结果 W_A.

其次求解问题(P_2).为了对比内部迁移学习变化,同时定义在新任务上随机初始化进行学习,可以选择两个不同的过程.

过程一:

Step 1.确定算法 G;

Step 2.确定停止准则 S;

Step 3.随机选定初始连接权重矩阵 $W_0 \in M$;

Step 4. 从 W_0 开始按照停止准则 S 运行算法 G 迭代,得到结果 W_B.

过程二:

Step 1.确定算法 G;

Step 2.确定停止准则 S;

Step 3.初始连接权重矩阵选定问题(P_1)的结果,即是 $W_0:=W_A$;

Step 4. 从 W_A 开始按照停止准则 S 运行算法 G 迭代,得到结果 $W_{B'}$.

过程一属于直接随机初始化权值进行训练,过程二属于内部迁移学习,通常而言 $W_{B'}$模型比 W_B 模型的泛化能力更强,且过程二的收敛速度更快.

2 内部迁移信息几何分析

上文对内部迁移学习进行了形式化的描述,明确了迁移学习的内涵, 构建了迁移学习的模型,但是这样的描述只有形式上的意义,如果需要进一步进行可行、可操作的研究,需要对数据集和连接权重矩阵进行几何化的描述以简化模型并且给予最直观的解释.

假设数据集先验位于某个分布族之中,即假设有一个分布族 F,这个族构成的流形记为 M_F,坐标卡系记为 f,其上的 Fisher 度量记为

$$\mathrm{d}s_F^2 = \sum g_{ij,F}\mathrm{d}f^i \times \mathrm{d}f^j \tag{10}$$

根据 Amari 信息几何理论[14],可以确定流形 M_F 上的度量 $\mathrm{d}s_F^2$ 与相容的联络 D_F,和联络形式 ω_F 以及对应的 Christoffel 系数 $\Gamma_{ij,F}^k$.

假设数据集 A 的概率分布来源于分布族 A,则其对应的某些分布构成流形 M_A, 显然 M_A 是 M_F 的子流形,假设嵌入方式为

$$\phi: M_A \to M_F \tag{11}$$

根据嵌入方式的拉回, 可得到流形 M_A 的坐标卡系为 $a:=\varphi^*(f)$,其上的 Fisher 度量记为

$$\mathrm{d}s_A^2 = \sum g_{ij,A}\mathrm{d}a^i \times \mathrm{d}a^j = \varphi^*(ds_F^2) \tag{12}$$

根据 Amari 信息几何理论[14],可以确定流形 M_A 上的度量 $\mathrm{d}s_A^2$ 与相容的联络 D_A,和联络形式 ω_A 以及对应的 Christoffel 系数 $\Gamma_{ij,A}^k$.显然 M_A 上的联络、联络形式和 Christoffel 系数还可以通过嵌入映射 ϕ 的拉回得到,即

$$D_A = \phi^*(D_F), \omega_A = \phi^*(\omega_F), \Gamma_{ij,A}^k = \phi^*(\Gamma_{ij,F}^k) \tag{13}$$

前文已知数据集 B 在规模上远远小于数据集 A 的规模,且同作为计算机视觉任务,假设数据集被采样的分布在结构上相似. 这个基本的假定在数学上可用子流形来表示,即假设数据集 B 的概率分布来源于分布族 B,某些可能分布构成流形 M_B,那么上面的基本假设可以表示为一个嵌入映射

$$\varphi: M_B \to M_A \tag{14}$$

根据嵌入方式的拉回, 可得到流形 M_B 的坐标卡系为 $b:=\varphi^*(a)$,其上的 Fisher 度量记为

$$\mathrm{d}s_B^2 = \sum g_{ij,B}\mathrm{d}b^i \times \mathrm{d}b^j = \varphi^*(ds_A^2) \tag{15}$$

根据 Amari 信息几何理论[14],可以确定流形 M_B 上的度量 $\mathrm{d}s_B^2$ 与相容的联络 D_B,和联络形式 ω_B 以及对应的 Christoffel 系数 $\Gamma_{ij,B}^k$.显然 M_B 上的联络、联络形式和 Christoffel 系数还可以通过嵌入映射 φ 的拉回得到,即

$$\begin{aligned} D_B &= \varphi^*(D_A) \\ \omega_B &= \varphi^*(\omega_A) \\ \Gamma_{ij,B}^k &= \varphi^*(\Gamma_{ij,A}^k) \end{aligned} \tag{16}$$

由流形 M_A 到 M_F 的嵌入和从 M_B 到 M_A 的嵌入

$$\varphi: M_B \to M_A, \phi: M_A \to M_F \tag{17}$$

可以产生一个从 M_B 到 M_F 的直接嵌入

$$\phi \cdot \varphi: M_B \to M_F \tag{18}$$

根据嵌入方式的拉回, 可得到流形 M_B 的坐标卡系为 $b:=(\varphi \cdot \phi)^*(f)$,其上的 Fisher 度量记为

$$\mathrm{d}s_B^2 = \sum g_{ij,B}\mathrm{d}b^i \times \mathrm{d}b^j = (\varphi \cdot \phi)^*(\mathrm{d}s_F^2) \tag{19}$$

显然 M_B 上的联络、联络形式和 Christoffel 系数

还可以通过嵌入映射 $\varphi\cdot\phi$ 的拉回得到,即

$$D_B=(\varphi\cdot\phi)^*(D_F)$$
$$\omega_B=(\varphi\cdot\phi)^{**}(\omega_F)$$
$$\Gamma^k_{ij,B}=(\varphi\cdot\phi)^*(\Gamma^k_{ij,F}) \tag{20}$$

固定神经网络的拓扑结构,那么所谓学习,即确定连接权重矩阵 W.一般而言,神经网络的节点个数是大规模的,具体确定一个连接权重矩阵是不现实的,需要对连接权重矩阵空间进行简化.假设神经网络连接权重矩阵也落在某种分布族 G 之中,这个分布族构成的流形结构为 M_G,坐标卡系为 θ,其上的 Fisher 度量记为

$$\mathrm{d}s^2_G=\sum g_{ij,G}\mathrm{d}\theta^i\times\mathrm{d}\theta^j \tag{21}$$

根据 Amari 信息几何理论[14],可以确定流形上的度量 $\mathrm{d}s^2_G$ 与相容的联络 D_G,和联络形式 ω_G 以及对应的 Christoffel 系数 $\Gamma^k_{ij,G}$.

学习过程不仅与数据集相关,还与神经网络的联接分布相关,因此考察学习的过程,即考察流形 $M_A\times M_G$,$M_B\times M_G$ 上的曲线过程,不同的学习算法和机制对应于曲线的不同性质.

定理 1　当神经网络结构不变,且数据集数据量大小,以及类丰富度小于数据集 A 时,神经网络在数据集 B 上获得的连接权重矩阵即流形 $M_B\times M_G$ 为神经网络在数据 上流形 $M_A\times M_G$ 的子流形.

证　对于流形 $M_A\times M_G$,其上的 Fisher 度量为

$$\mathrm{d}s^2_{A+G}=(\mathrm{d}a,\mathrm{d}\theta)\begin{pmatrix}g_{ij,A}&0\\0&g_{ij,G}\end{pmatrix}\begin{pmatrix}\mathrm{d}a^{\mathrm{T}}\\\mathrm{d}\theta^{\mathrm{T}}\end{pmatrix} \tag{22}$$

根据 Amari 信息几何理论[14],可以确定流形 $M_A\times M_G$ 上的度量 $\mathrm{d}s^2_{A+G}$ 与相容的联络 D_{A+G},和联络形式 ω_{A+G} 以及对应的 Christoffel 系数 $\Gamma^k_{ij,A+G}$. 它们之间显然有如下的简单关系:

$$D_{A+G}=\begin{pmatrix}D_A&0\\0&D_G\end{pmatrix};\omega_{A+G}=\begin{pmatrix}\omega_A&0\\0&\omega_G\end{pmatrix} \tag{23}$$

对于流形 $M_B\times M_G$,其上的 Fisher 度量为

$$\mathrm{d}s^2_{B+G}=(\mathrm{d}b,\mathrm{d}\theta)\begin{pmatrix}g_{ij,B}&0\\0&g_{ij,G}\end{pmatrix}\begin{pmatrix}\mathrm{d}b^{\mathrm{T}}\\\mathrm{d}\theta^{\mathrm{T}}\end{pmatrix} \tag{24}$$

根据 Amari 信息几何理论[14],可以确定流形 $M_B\times M_G$ 上的度量 $\mathrm{d}s^2_{B+G}$ 与相容的联络 D_{B+G},和联络形式 ω_{B+G} 以及对应的 Christoffel 系数 $\Gamma^k_{ij,B+G}$. 它们之间显然有如下的简单关系

$$D_{B+G}=\begin{pmatrix}D_B&0\\0&D_G\end{pmatrix};\omega_{B+G}=\begin{pmatrix}\omega_B&0\\0&\omega_G\end{pmatrix} \tag{25}$$

流形 $M_A\times M_G$ 和流形 $M_B\times M_G$ 之间有自然的嵌入关系

$$(\varphi,id):M_B\times M_G\rightarrow M_A\times M_G \tag{26}$$

因此流形 $M_B\times M_G$ 是流形 $M_A\times M_G$ 的子流形.

证毕.

整个学习过程如图 1 左所示,在问题(P_1)中通过随机初始化的模型 W_0 在数据库 上进行训练得到最终模型权值 W_A.在问题(P_2)中,过程一是使用随机初始化模型 W_0 重新进行训练得到最后模型 W_B,过程二是使用得到的模型权值 W_A 作为初始值,在数据库 B 上进行训练得到模型 $W_{B'}$. 通常情况下,模型 $W_{B'}$比模型 W_B 具有更好的泛化能力.

根据定理 1,以及内部迁移信息几何分析可知,问题(P_2)的学习过程一,可当作在 $M_B\times M_G$ 流形上进行探索,它属于 $M_A\times M_G$ 的子流形.因此模型随机初始化永远也逃不出 $M_B\times M_G$ 的流形空间, 模型参数在优化过程中的探索空间有限.而通过 W_A 进行初始化的模型可以在整个 $M_A\times M_G$ 流形上进行探索,则能很容易达到模型较为理想的参数 $W_{B'}$.整个过程可简化为如图 1 右所示.

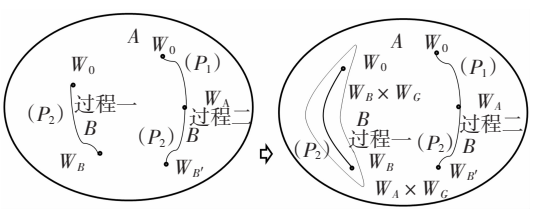


图 1　参数空间下权值迁移变化

Fig.1　Weight transfer change in parameter space

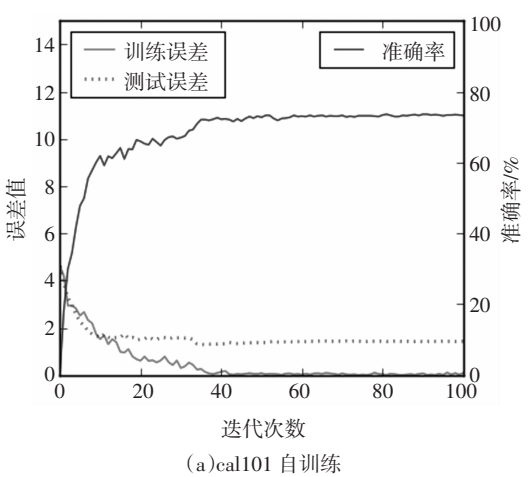
3　实验与分析

为了更好地体现不同数据集之间模型迁移的效果,选用 3 种计算机视觉数据集.具体实验设置数据集 为 ImageNet[13],数据集大小约为 138 G,其中包含 1 281 167 张图,1 000 类, 每类大约 1 300 张.数据集 B 采用 Cal101[15],数据集约 131 M,其中包含 102 类,每类大约 50 张图.数据集 C 使用 Cal256[16],数据大小约 1.2 G,其中包含 257 类,每类约 110 张图片. 其中数据集 A 数据量大小和类别数都大于数据集 B,C.根据前文分析,数据集 B,C 可以实现到数据集 A 的嵌入映射.同理数据集 B 也能变化为数据集 C 的嵌入映射.例如,ImageNet 包含大多数 Cal101 中的类别,而未包含的类别可以通过形态变化[17]的方法实现嵌入映射.同理可实现 Cal101 嵌入映射到 Cal256 数据集. 深度神经网络结构使用 AlexNet[18],

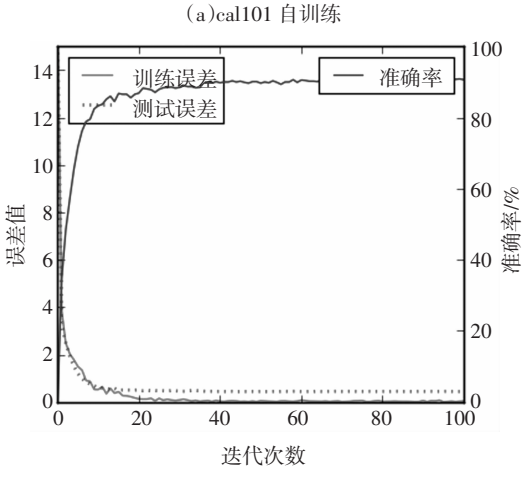
GoogleNet[19].

分析不同数据集上的学习问题,首先解决问题(P_1),随机初始化神经网络进行训练,通过不断的迭代,最终收敛得到模型 W_A.其次对比过程一和过程二网络训练的差别.

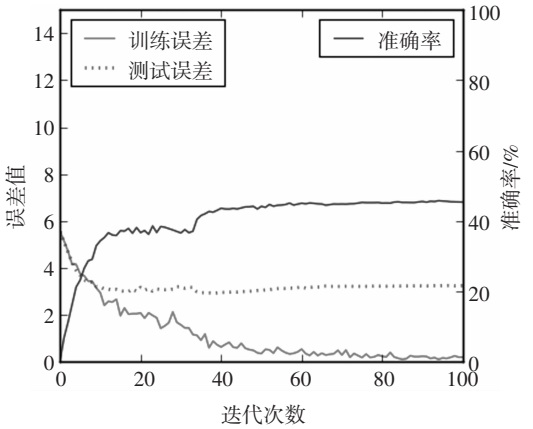
过程一通过在数据集上直接初始化网络权值进行训练.过程二通过在 ImageNet 上训练得到的模型 W_A 的权值进行初始化,然后进行微调.为了更好地分析大库对小库的影响,再次通过在数据集 Cal256 上训练得到的模型作为新任务 Cal101 的初始化值进行训练.图 2 表示 Alexnet 网络在不同数据集和情况下的表现,图 3 所示 GoogleNet 网络在几种情况下的表现.分析实验结果可知,在 AlexNet 网络中,对于 Cal101 数据集而言,随机初始化训练的网络迭代约 60 轮左右时收敛,且在测试集上的准确率为 73.275 9%. 而通过模型对网络进行微调,可以发现网络在迭代 40 次左右时已经开始收敛,且在测试集上的准确率为 90.625%,远超过重新训练权值的结果,具有非常好的泛化能力. 即使是通过 Cal256 数据集得到的网络模型进行微调,最终也能得到比直接初始化网络得到的结果更好.



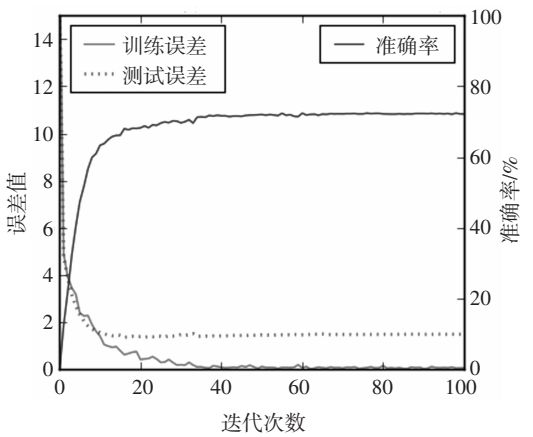
(a)cal101 自训练



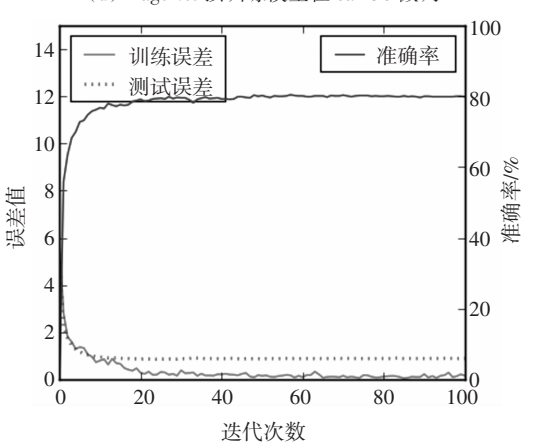
(b)ImageNet 预训练模型在 cal101 微调



(c)cal256 自训练



(d)ImageNet 预训练模型在 cal256 微调

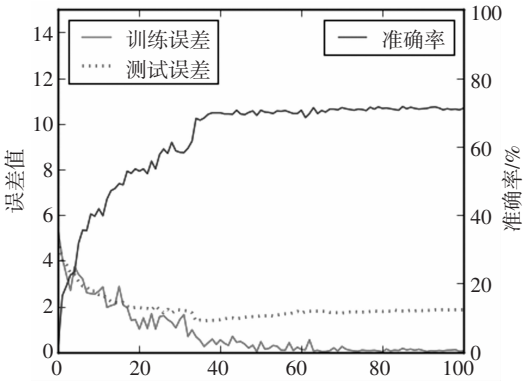


(e)cal256 预训练模型在 cal101 微调

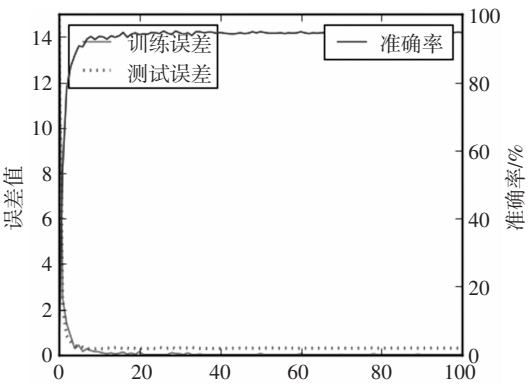
图 2 AlexNet 网络在不同数据集下训练变化

Fig.2 The AlexNet network training changes under different data sets

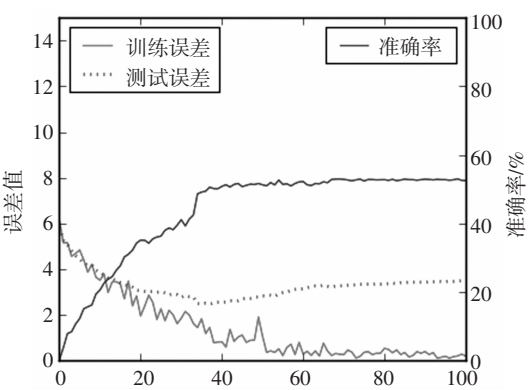
同理对比图 3,GoogleNet 的最终正确率虽跟 AlexNet 在各个数据集上不同,但都表现出同样的性质. 通过在大数据集上学习得到的模型对小数据集进行微调,模型可以得到比直接在小数据集上学习更好的表现.具体在各个任务上,各网络最终的准确率如表 1 所示.



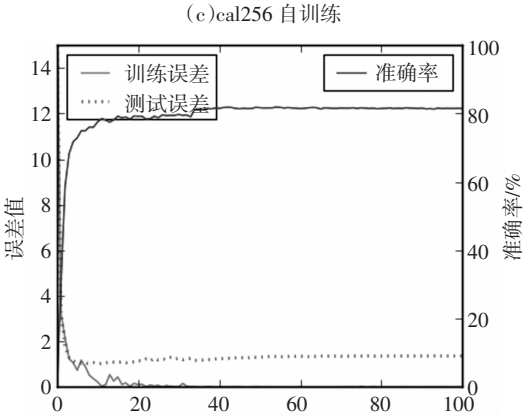
(a)cal101 自训练



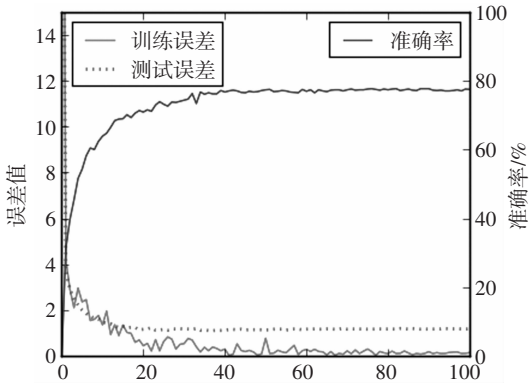
(b)ImageNet 预训练模型在 cal101 微调



(c)cal256 自训练



(d)ImageNet 预训练模型在 cal256 微调



(e)cal256 预训练模型在 cal101 微调

图 3　GoogleNet 网络在不同数据集下训练变化

Fig.3　The GoogleNet network training changes under different data sets

表 1　不同条件下网络准确率

Tab.1　Network accuracy rate under different conditions

数据集与模型	准确率/%
Cal101–AlexNet–自训练	73.275 9
Cal101–AlexNet–ImageNet 微调	90.625
Cal101–AlexNet–Cal256 微调	79.849 1
Cal101–GoogleNet–自训练	71.120 7
Cal101–GoogleNet–ImageNet 微调	94.612 1
Cal101–GoogleNet–Cal256 微调	77.478 5
Cal256–AlexNet–自训练	45.377 6
Cal256–AlexNet–ImageNet 微调	72.135 4
Cal256–GoogleNet–自训练	52.506 5
Cal256–GoogleNet–ImageNet 微调	81.412 8

通过对比随机初始化训练的网络与微调学习方法的网络,还可以发现在学习过程中使用微调的网络,训练误差和测试误差下降速度非常快,并且在网络开始收敛时,误差的波动都较小.这些实验结果都说明了,通过在大数据集上学习的模型作为新任务的初始化权值信息,可以使得模型的参数探索的空间更大,使之能找到相对较优的结果.

4　结　论

深度神经网络的内部迁移过程本质上只是参数的初始化方法不同.根据 Im 分析[20],即使得到的局部最小值不同,其表现的泛化能力并没有太大的

差异,这与 Dauphin 等[21]分析的结果相符.然而通过大数据集训练的网络进行参数初始化,往往可以得到一个更好的结果.本文通过形式化整个学习过程,将这种参数信息的变化放入流形空间中. 在考虑参数变化的同时,也融合了样本的分布信息.通过信息几何理论对学习过程中流形的变化进行讨论. 分析可得大数据集下训练的网络作为权值更新,与小数据集上重新训练相比,隐含的包含了原样本的数据空间,使得其具备更大的探索空间且更容易找到一个更好的模型参数. 通过多组实验对比重现这一过程,该分析也为深度神经网络内部迁移过程提供了一种可能的数学解释. 并且在深度迁移学习问题中有关迁移变化过程中定量的分析还缺少合理的手段,该方法可进一步探究深度神经网络在学习过程中获取的有用信息量,以及相对应丢失的冗余信息量,探索网络学习过程中的不变性分析,试图打开学习过程的黑盒,实现不同应用场景下又快又准的迁移.

参考文献

[1] CLERY D,VOSS D. All for one and one for all [J]. Science, 2005,308(5723):809—809.

[2] DENèVE S,MACHENS C K. Efficient codes and balanced networks[J]. Nature Neuroscience,2016,19(3):375.

[3] NG P C,HENIKOFF S. SIFT: predicting amino acid changes that affect protein function[J]. Nucleic Acids Research,2003,31(13): 3812—3814.

[4] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2005:886—893.

[5] HINTON G E,OSINDERO S,TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation,2014,18(7):1527—1554.

[6] LEVINE S,FINN C,DARREL T,*et al*. End-to-end training of deep visuomotor policies [J]. Journal of Machine Learning Research,2016,17(1):1334—1373.

[7] LECUN Y,BENGIO Y,HINTON G. Deep learning [J]. Nature, 2015,521(7553):436—444.

[8] OUYANG W,WANG X,ZHANG C,*et al*. Factors in fine tuning deep model for object detection with long-tail distribution[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2016:864—873.

[9] ABRàMOFF M D,LOU Y,ERGINAY A,*et al*. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning [J]. Investigative Ophthalmology & Visual Science,2016,57(13):5200.

[10] YOSINSKI J,CLUNE J,BENGIO Y,*et al*. How transferable are features in deep neural networks? [C]//Advances in Neural Information Processing Systems 2014. 2014: 3320-3328.

[11] DAUPHIN Y,PASCANU R,GULCEHRE C,*et al*. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization[J]. Mathematics,2014,111(61):2475—2485.

[12] AMARI S I. Information geometry of statistical inference -an overview [C]//Information Theory Workshop,2002. Proceedings of the 2002 IEEE. IEEE,2002: 86-89.

[13] DENG J,DONG W,SOCHER R,*et al*. ImageNet: A large-scale hierarchical image database [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2009: 248—255.

[14] AMARI S,NAGAOKA H. Methods of information geometry [M]. American Mathematical Society,2000:13-206.

[15] LI F F,FERGUS R,PERONA P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories [J]. Computer Vision and Image Understanding,2007,106(1):59-70.

[16] GRIFFIN G,HOLUB A,PERONA P. Caltech-256 object category dataset [EB/OL]. http://www.vision.caltech.edu/Image_Datasets/Caltech101,April 5,2006.

[17] SCHMITZER B,SCHNORR C. Globally optimal joint image segmentation and shape matching based on Wasserstein modes[J]. Journal of Mathematical Imaging & Vision,2015,52(3):436—458.

[18] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks [C]// International Conference on Neural Information Processing Systems. Curran Associates Inc,2012:1097—1105.

[19] SZEGEDY C,LIU W,JIA Y,*et al*. Going deeper with convolutions [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2015:1—9.

[20] IM D J,TAO M,BRANSON K. An empirical analysis of the optimization of deep network loss surfaces [J]. ArXiv Preprint ArXiv:1612.04010,2016.

[21] DAUPHIN Y N,PASCANU R,GULCEHRE C,*et al*. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization [C]// International Conference on Neural Information Processing Systems. MIT Press,2014:2933-2941.