深度学习在地理国情监测高分遥感影像分类上学习率策略的探讨

杜 娟,周 旭,李广泳,程 涵,陶 舒
(国家基础地理信息中心,北京 100036)

摘 要:学习率是深度学习模型训练中较难设置的超参数之一,设计良好的学习率策略可以显著提高深度学习模型的收敛速度,提升模型的精度水平。本文着眼于遥感影像分类,基于山东省地理国情普查成果数据,利用深度学习 SegNet 网络模型,从学习率是否自适应两个方面研究,选择随机梯度下降方法和 AdaGrad、AdaDelta 方法。其中随机梯度下降方法中学习率策略涉及 fixed、step、multistep 三种,采用语义分割中常用的定量评价指标,验证不同学习率策略下模型精度。研究结果表明,呈下降趋势的学习率策略可以提高模型的精度,应在训练初始阶段设置较大的学习率以加快训练收敛速度,并在训练早期降低学习率;常数型的学习率经验证是一种简单有效的方法,学习率的设置需要依靠经验或者通过反复试验确定;自适应算法具有一定的优势,但有些也需要尝试不同的学习率,并且有时表现不如随机梯度下降。

关键词:深度学习;学习率;高分遥感影像

1 引 言

遥感影像分类作为一种影像信息提取方法,是根据不同谱段的光谱特征、空间结构特征或其他信息,按照某种规则划分,将图像中每个像元点或每块区域划分为不同的类别。地理国情监测以遥感影像信息提取方法为主要技术方法,以高分辨率遥感影像为主要数据源。在地理国情普查的基础上,整合最新的基础地理信息数据及相关部门专题数据,监测全国变化情况,其中遥感影像信息提取可采用计算机自动分类与人工判读解译结合的方式。高分辨率遥感影像中地物的光谱特征更加丰富,同类地物的光谱差异增大,类间的差异减少,导致同物异谱及同谱异物现象更加普遍,而影像中的大量细节和复杂的地物光谱特征使得诸如 k-近邻算法(KNN)、朴素贝叶斯算法(NB)、分类回归树(CART)等基于光谱统计特征的传统分类方法精度低。因此,为了提高遥感影像分类精度,研究者将智能算法应用于高分辨率影像的分类,如神经网络(NN)、支持向量机(SVM)、决策树(DT)等。虽然这些算法能够获得精确度更高的结果,但属于浅层学习算法。由于计算单元有限,浅层网络很难有效地表达复杂函数,所以当样本数量增大、多样性增强时,浅层模型的适用性受到局限。

2006 年 Hinton 等人提出了深度学习的概念,指出多隐层神经网络能够学习到对象更本质的特征,并且其在训练上的复杂度可以通过逐层初始化来有效缓解,同时也掀起了人工神经网络的又一热潮。深度学习架构由多层非线性运算单元组成,通过对原始数据逐层非线性化训练,可以从大量输入数据中学习到更高层次、更加抽象的特征表示,使其在复杂分类上具有很好的效果和效率。然而,超参数的设定困扰着深度模型的训练,因为其不可通过常规方法学习获得,学习率即其中之一。深度学习模型的学习率一般根据函数的特点进行设定调节,作为一种最简单有效的策略,学习率通常为常数型或呈下降趋势的指数型或幂指数型函数。后续有学者提出了一些自适应的算法,如 Duchi 等人 2011 年提出了 Adaptive Gradient (AdaGrad)算法,该算法为所有模型参数单独设计学习率,使每个参数与其梯度历史平方和的平方根成反比,以此保证学习率的下降趋势。该方法首次提出全参数学习率策略,为深度学习模型的学习率自适应调节提供了一个较好的解决思路。Zeiler 于 2012 年在 AdaGrad 算法的基础上将梯度平方和修改为指数加

权的移动平均,提出了 AdaDelta 算法,旨在应用于凸问题时快速收敛。在模型精度水平上,AdaDelta 比 AdaGrad 有了进一步的提升。此后还产生了一些自适应算法,如与 AdaDelta 很像的 RMSProp、Adam 等。

目前,深度学习已广泛应用于图像分类领域,在高分辨率遥感影像分类方面也已有一些研究成果,如方旭等人将全卷积网络(FCN)算法应用于高分辨率遥感影像的分类中,采用均值漂移(mean-shift)分割算法获得像素之间的空间关系,以改善 FCN 算法在边缘区域的分类效果;冯家文等人在传统 FCN 算法的基础上,在输入端加上经过 Sobel 算子变换的图片信息,从源头增加特征信息的多样性,提升了图像分割的效果;易盟等提出一种端到端的深层结构,通过对航拍图像和 GIS 图像在双通道下多尺度特征信息的融合,利用条件随机场(conditional random field,CRF)改善最终分割结果,实现高分辨航拍图像像素级精确语义分类;汤浩等人引入 FCN,以 ESAR 卫星图像为样本,基于像素点级别构建卷积网络进行训练,得到各像素的初始类别分类概率,结合 CRF 结构得到全局像素类别转移结果,之后进行 RNN 的迭代进一步优化实验结果。但在高分辨率遥感影像分类方面,关于学习率策略的研究相对较少,因此本文以第一次地理国情普查高分遥感影像和地表覆盖栅格数据为基础,通过在 SGD 算法下调整学习率,并纳入可适应算法 AdaGrad 和 AdaDelta 进行对比,提出了一些学习率调整策略,以提升深度学习模型的精度水平。

2 模型及方法

2.1 深度学习模型

遥感影像分类在像素级别上进行分类。2006 年前后提出的卷积神经网络,如 AlexNet 模型、GoogLeNet 模型、VGG 模型,具有强大的特征学习能力,较浅的卷积层学习局部区域特征,而较深的卷积层能够学习更加抽象的特征。这些抽象特征有助于提高识别性能,以对识别图像中的物体进行分类,但无法在图像像素水平上进行分类,因此后续有学者提出了全卷积神经网络模型。结合前期的实验结果,本文模型选择 SegNet 模型。

SegNet 是由剑桥大学设计的全卷积神经网络,其主要动机来源于道路现场理解应用,网络结构如图 1 所示。模型包含一个编码网络和一个对应的解码网络,最后是一个像素级别的分类层。编码网络的前 13 个卷积层与 VGG16 网络相同,编码网络中各层在解码网络中都有层与其对应,模型最后连接一个多级 soft-max 分类器以在像素级别上产生类别概率。

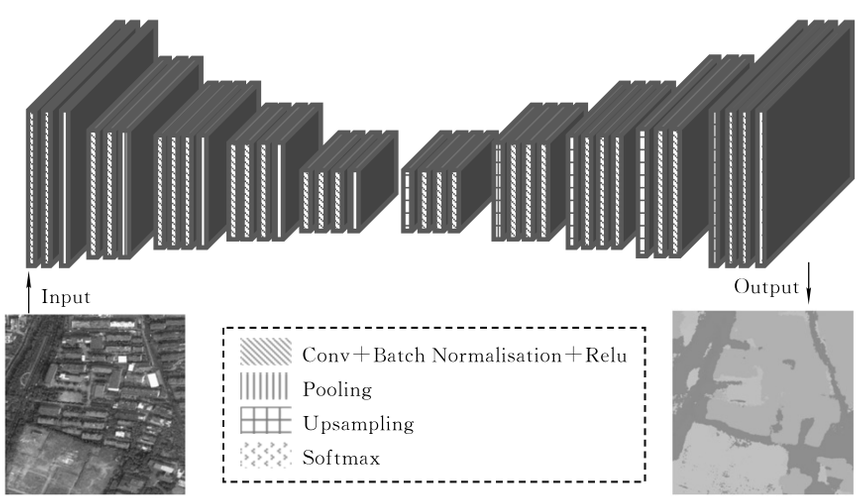

图 1 SegNet 网络模型示意

编码网络与滤波器族卷积得到特征图后,分别进行批归一化(batch normalized,BN),采用 ReLU 激活函数,并为了获得空间小位移的平移不变,选择最大池化方法进行池化。最大池化和下采样导致边界细节损失,因此 SegNet 网络在编码过程中加入对边界信息保存的考虑,引入池化索引方法,考虑内存原

因，仅保存最大池化索引。解码网络则对输入的特征图使用对应编码特征图保存的最大池化索引，进行上采样得到稀疏的特征图，解码技术如图 2 所示。将特征图与可训练的解码滤波器族卷积，从而得到稠密的特征图，继而对特征图进行批归一化，最后将高维的特征图输入 soft-max 层，对每个像素进行分类，得到每个像素属于 k 类的概率。

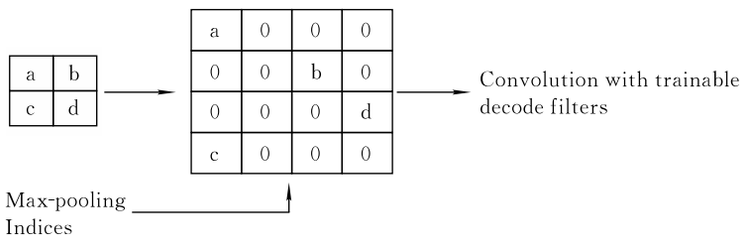

图 2　SegNet 解码示意

2.2　优化算法

1. 随机梯度下降算法

深度学习中应用最多的优化算法是随机梯度下降(stochastic gradient descent, SGD)，其在数据集中随机挑选一部分样本(minibatch)，通过计算它们损失函数梯度的均值，作为梯度下降的依据。但 SGD 学习过程有时较慢，因此后续通过引入动量方法，积累之前梯度指数级衰减的移动平均，旨在加速学习。本文所指的为带动量的随机梯度下降，为简单起见，后续仍以 SGD 称呼，其数学表达式为

$$v_{t+1}=\alpha v_t-\varepsilon_t g_t$$

$$\theta_{t+1}=\theta_t+v_{t+1}$$

$$g_t=\frac{1}{m}\nabla_\theta\sum_{i=1}^{m}L(f(x^i;\theta),y^i)$$

式中，x^i 代表训练集中的样本，y^i 代表其目标值，m 代表小批量数据集内的样本数量，$f(x^i;\theta)$ 表示输入 x^i 时的预测输出，$L(f(x^i;\theta),y^i)$ 为定义在数据集上的损失函数，$\nabla_\theta\sum_{i=1}^{m}L(f(x^i;\theta),y^i)$ 为损失函数的梯度，v_t 表示第 t 步迭代的动量值，α 表示动量参数，ε_t 为第 t 步迭代的学习率，g_t 为第 t 步迭代的梯度，θ 代表参数值。

保证 SGD 收敛的一个充分条件是 $\sum_{k=1}^{\infty}\varepsilon_k=\infty$，且 $\sum_{k=1}^{\infty}\varepsilon_k^2<\infty$。

2. AdaGrad 和 AdaDelta 算法

常数型学习率在很多时候仍然是一种最简单有效的方法，但需要对学习率初值设置有足够丰富的经验。在基本的梯度下降优化中，有一个常见的问题，要优化的参数对于目标函数的依赖是各不相同的。对于某些参数，已经优化到了极小值附近，但是有的参数仍然在梯度很大的地方，这时统一的全局学习率是可能出现问题的。如果学习率太小，则梯度很大的参数会收敛很慢，如果学习率太大，已经优化差不多的参数可能会不稳定。针对这个问题，Duchi 提出了比较有代表性的 AdaGrad 算法，之后陆续有学者提出了 AdaDelta、RMSProp、Adam 等算法，本文选择 AdaGrad 算法和 AdaDelta 算法作为比较。

AdaGrad 算法的基本思想是对每个参数使用不同的学习率，独立地缩放每个参数反比于其所有梯度历史平方值总和的平方根，随着优化过程的进行，对于已经下降较多的参数减缓学习率，反之则保持一个较大的学习率，算法梯度更新公式为

$$\theta_{t+1}=\theta_t-\frac{\varepsilon_0}{\delta+\sqrt{\sum_{s=1}^{t}g_s^2}}*g_t$$

$$g_t=\frac{1}{m}\nabla_\theta\sum_{i=1}^{m}L(f(x^i;\theta),y^i)$$

式中，ε_0 为初始学习率，δ 为小常数。

虽然 AdaGrad 针对不同的变量有不同的学习率，但是初始的学习率仍需要手工设定。如果全局学习率过大，优化同样不稳定；而如果全局学习率较小，随着优化的进行，学习率会越来越小，很可能跳不出局部最优点。AdaDelta 算法在 AdaGrad 的基础上，将历史梯度积累改为当前时间向前一个窗口期内的累积，并受启发于牛顿法和 LeCun 关于牛顿法近似的研究，实现了不用手动设定学习率，具体公式为

$$\theta_{t+1}=\theta_t-\frac{\mathrm{RMS}[\Delta\theta]_{t-1}}{\mathrm{RMS}[g]_t}*g_t$$

$$\mathrm{RMS}[g]_t=\sqrt{E[g^2]_t+\delta}$$

$$E[g^2]_t=\rho*E[g^2]_{t-1}+(1-\rho)*g_t^2$$

$$g_t=\frac{1}{m}\nabla_\theta\sum_{i=1}^{m}L(f(x^i;\theta),y^i)$$

式中，$\mathrm{RMS}[g]_t$ 为第 t 步迭代的梯度均方根(root mean square)，$E[g^2]_t$ 为第 t 步的梯度平方和。

2.3 精度评价方法

本文采用语义分割常用的三种度量方法作为定量评价指标来综合评估模型的精度。

(1)像素精度(pixel accuracy，PA)是最简单的度量，为预测正确的像素占总像素的比例，即 $\mathrm{PA}=\dfrac{\sum_{i=1}^{n_{cl}}n_{ii}}{\sum_{i=1}^{n_{cl}}\sum_{j=1}^{n_{cl}}n_{ij}}$。

(2)均像素精度(mean pixel accuracy，mPA)是 PA 的一种简单提升，是计算每个类内被正确分类像素数的比例，然后求所有类的平均值，即 $\mathrm{mPA}=\dfrac{1}{n_{cl}}\sum_{i=1}^{n_{cl}}\dfrac{n_{ii}}{\sum_{j=1}^{n_{cl}}n_{ij}}$。

(3)平均交并比(mean intersection over union，mIoU)为语义分割的标准度量，其计算真实值(ground truth)和预测值(predicted segmentation)的交集、并集之比后求平均值，即 $\mathrm{mIoU}=\dfrac{1}{n_{cl}}\sum_{i=1}^{n_{cl}}\dfrac{n_{ii}}{\sum_{j=1}^{n_{cl}}(n_{ij}+n_{ji})-n_{ii}}$。其中，$n_{cl}$ 表示类别数量，n_{ij} 表示标签为 i 类预测被分为 j 类的像素数量。

在以上的度量标准中，mIoU 指标简洁且代表性强，所以大多数语义分割论文都使用该指标，本文以 mIoU 为主，PA、mPA 为辅助指标进行评价。

3 基于学习率调整和不同优化算法的实验

3.1 实验环境及数据介绍

实验选择 ubuntu16.04 操作系统，基于 caffe 深度学习框架，采用 CUDA-GPU 加速方案。显卡型号为 NVIDIA TITAN Xp 12G，CUDA 版本为 8.0，cuDNN 版本为 5.1。

本文数据基于第一次地理国情普查成果，训练数据集的质量会影响模型的分类精度，山东普查成果较优，所以本文以山东为研究区域采集数据。数据集由影像数据及对应的标签数据构成，影像数据源主要为数码航空影像、Worldview-2、QuickBird 等，将原始影像数据重采样率设为 1 m，标签数据是将山东省第一次地理国情普查地表覆盖矢量数据栅格化，按照地理国情普查分类体系，选择一级类——耕地、园地、林地、草地、房屋建筑区、道路、构筑物、人工堆掘地、荒漠与裸露地表、水域为研究对象。根据研究区域地物的分布情况，较大尺寸图像有更好的空间全局性，并考虑内存占用，因此数据集尺寸选择 500×500。在训

练集的选取上，已有研究表明，训练集中占比较大的类别其分类精度也较大，但各类别地物在分布上本来就是不平均的。为了尽可能得到平衡样本，减少类别数量差异的影响，本文采用拒绝抽样方法抽取样本组成训练集。训练集样本数量为 10 000，测试集为 4 000，实际抽样时，以山东省全部裁切结果为总体，随机抽取 10 000 张图片，通过标签数据中各类别像素和计算方差，重复此过程 1 000 次后，选择方差最小的值为阈值进行拒绝抽样，继而得到方差小于阈值的训练样本，图 3 为训练数据示例。

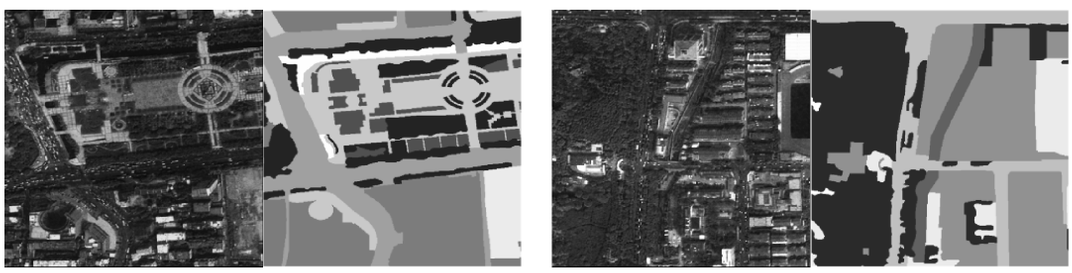

图 3　原始影像与标签示例

3.2　实验结果与分析

为基于实验数据得到学习率调整策略，本部分从使用 SGD 算法手动调节学习率和使用自适应学习率算法出发，通过设置基础学习率、学习率调整策略、学习率调整步长等参数进行对比分析。训练时批量大小统一设为 4，总迭代次数为 100 000 次，即运行 40 代(epoch)，采用像素精度、均像素精度、平均交并补进行对比分析。

1. 调整学习率

1)常数型

常数型学习率，即在整个训练过程中学习率不变，学习率分别设置为 0.001、0.01、0.1，结果如表 1 所示。学习率为 0.001 时，PA、mPA、mIoU 的值都是最高的；随着学习率增加，精度减少；而当学习率为 0.1 时，迭代 8 000 次后，损失值溢出，说明学习率过大。

表 1　不同常数型学习率水平下精度对比

Base_lr	Lr_policy	Iteration	PA	mPA	mIoU
0.001	**Fixed**	**100 000**	**0.733 4**	**0.514 9**	**0.40**
0.01			0.682 4	0.427 8	0.317 9
0.1			—	—	—

2)step

一个呈下降趋势的学习率策略可以显著提高模型的收敛速度，本文选择 step 方法，即经过一定步长后，学习率按照规则进行改变，规则为

$$lr = base\_lr * gamma\hat{}(floor(iter/stepsize))$$

式中，lr 代表当前学习率，base_lr 代表基础学习率，gamma 代表缩放因子，iter 代表当前迭代次数，stepsize 为步长。实验中 gamma 设定为 0.1，stepsize 设定为 20 000，基础学习率分别设置为 0.001、0.01、0.1，结果如表 2 所示。当基础学习率为 0.01 时，精度水平最高；当基础学习率为 0.1 时，损失值溢出，学习率过大；而基础学习率为 0.001 时的精度不敌 0.01，说明学习率过小。

表 2　step 学习策略下不同学习率水平的精度对比

Base_lr	Lr_policy	Iteration	PA	mPA	mIoU
0.001	Step	100 000	0.704 4	0.453 6	0.338 4
0.01			**0.730 3**	**0.498 9**	**0.395 1**
0.1			—	—	—

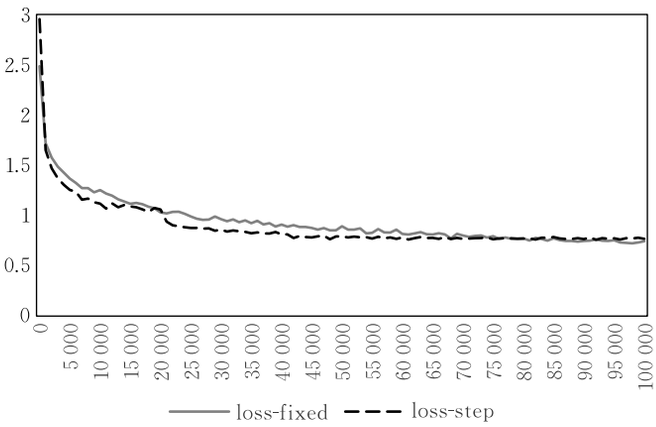

图 4 fixed 和 step(Base_lr=0.01)学习策略下损失曲线对比

3)multistep

将以上两种方法下精度水平最高的训练进行对比,损失曲线如图 4 所示。可以看出,step 方法下损失曲线开始下降较快,训练后期较为平缓,因为学习率很小,最终精度稍低于常数型学习率。为防止训练后期学习率衰减过多而变得太小,可以通过设定步长大小进行控制,但 step 方法是均匀等间隔变化,损失值并不是均匀减小的,因此加入 multistep 方法,以期提高模型精度。

multistep 与 step 很相似,不同之处为 step 是根据 stepsize 等间隔变化,而 multistep 为非等间隔变化,需要人为设定 stepvalue,即训练达到设定的 stepvalue 后根据 step 方法中的规则更新学习率。训练过程中 stepvalue 可根据需要设置多个,为避免学习率过小,本文仅设置一个 stepvalue。基础学习率为 0.01,gamma 为 0.1,stepvalue 分别设定为 10 000、30 000、50 000、70 000、90 000,以观察学习率在训练中哪个阶段衰减有利于精度提升,结果如表 3 所示。结果显示基础学习率为 0.01,经过第 stepvalue 次迭代后衰减为 0.001, stepvalue 为 10 000 时精度水平最高,而随着 stepvalue 的增加,精度水平降低。

表 3 step 学习策略下不同学习率水平的精度对比

Base_lr	Lr_policy	Iteration	stepvalue	PA	mPA	mIoU
0.01	Multistep	100 000	**10 000**	**0.732 1**	**0.526 1**	**0.415 2**
			30 000	0.730 8	0.517 1	0.403 4
			50 000	0.725 8	0.499 9	0.393 2
			70 000	0.728 9	0.493 8	0.390 4
			90 000	0.714 2	0.484 6	0.380 9

2.自适应优化算法

加入 AdaGrad、AdaDelta 自适应优化算法进行对比,AdaGrad 的基础学习率分别设定为 0.001、0.01、0.1,AdaDelta 算法不用指定全局学习率,所以学习率设定为 1,结果如表 4 所示。AdaGrad 算法中学习率为 0.01 和 0.1 时,损失值溢出,说明学习率设置过大;当学习率为 0.001 时,像素精度达到 73.55%,mIoU 约为 39.51%,而 AdaDelta 的像素精度达到 74.84%,mIoU 为 40.74%,高于 AdaGrad 算法。

表 4 AdaGrad、AdaDelta 算法下不同学习率的精度对比

Method	Base_lr	Iteration	PA	mPA	mIoU
AdaGrad	0.001	100 000	0.735 5	0.500 1	0.395 1
	0.01		—	—	—
	0.1		—	—	—
AdaDelta	1.0	10 000	0.748 4	0.514 8	0.407 4

通过以上实验结果,可以得到以下结论:

(1)对比 SGD 算法下的三种学习率策略,表明在训练初始阶段,应设置较大的学习率以保证训练的损失值快速下降,加快训练收敛,并在训练较早时期降低学习率,这样的策略对于提升模型的精度水平是有利的,但应防止训练后期学习率过小。

(2)常数型的学习率在本文实验中经验证是一种简单有效的方法,不过学习率的设置需要依靠经验或者通过反复试验确定,对于使用已有开源模型的情形,学习率的设定可以借鉴已有的研究。

(3)自适应算法中,AdaGrad 的 mIoU 与 SGD 下的 step 策略处于同一量级,但 PA 和 mPA 则优于 step 策略;AdaDelta 算法的精度高于 SGD 下的常数型和 step 策略,但略低于 multistep 方法下取得的最高精度。上述结果说明自适应优化算法有一定的优势,但对于 AdaGrad 来说,也需要尝试不同的全局学

习率，对 AdaDelta 来说，不用手动设定学习率是其优点，但其表现有时比不过学习率经过设计的 SGD。

3. 基于地理国情普查数据的分析

比较以上实验中表现最优和最差的两个模型的各类别精度，即学习率为 0.01 时的 multistep 方法和学习率为 0.01 时的 step 方法，如表 5 所示。结果表明，除林地、房屋建筑区和水域外，根据 multistep 得到的各类别精度都有一定程度的提升，最终平均像素精度提升大约 2.7%。将两种实验下得到的权重模型应用于预测训练、测试集以外的数据，以图 5 两例进行说明。图 5 中，(a)1、(a)2 表示原始影像，(b)1、(b)2 表示标签数据(标签与影像不对应的部分，可能缘于后期经过了时点核准)，(c)1、(c)2 表示利用 multistep(Base_lr=0.01)策略下权重模型得到的预测示例，(d)1、(d)2 表示利用 step(Base_lr=0.01)策略下权重模型得到的预测示例。通过将图 5(c)、图 5(d)与图 5(b)对比，发现图 5(c)1 的道路比图 5(d)1 更清晰准确，因为 multistep 方法中道路类别的精度为 58.49%，相比于 step 方法下的 55.09%，精度提升约 3.4%；，而图 5(c)2 的道路也比图 5(d)2 更清晰，此外图 5(c)2 中预测出人工堆掘地，而图 5(d)2 则没有，因为就人工堆掘地类别，multistep 比 step 方法的精度高 15.59%。

表 5　multistep(Base_lr=0.01)和 step(Base_lr=0.01)两种策略下各类精度对比

class	各类别精度		差值
	multistep	step	
耕地	0.909 6	0.907 0	0.002 6
园地	0.459 3	0.408 0	0.051 3
林地	0.635 3	0.734 4	−0.099 1
草地	0.504 4	0.376 8	0.127 6
房屋建筑区	0.836 1	0.851 0	−0.014 9
道路	0.584 9	0.550 9	0.034 0
构筑物	0.463 8	0.442 4	0.021 5
人工堆掘地	0.193 3	0.037 4	0.155 9
荒漠	0.001 0	0.000 0	0.001 0
水域	0.673 4	0.680 8	−0.007 4
总体	0.526 1	0.498 9	0.027 2

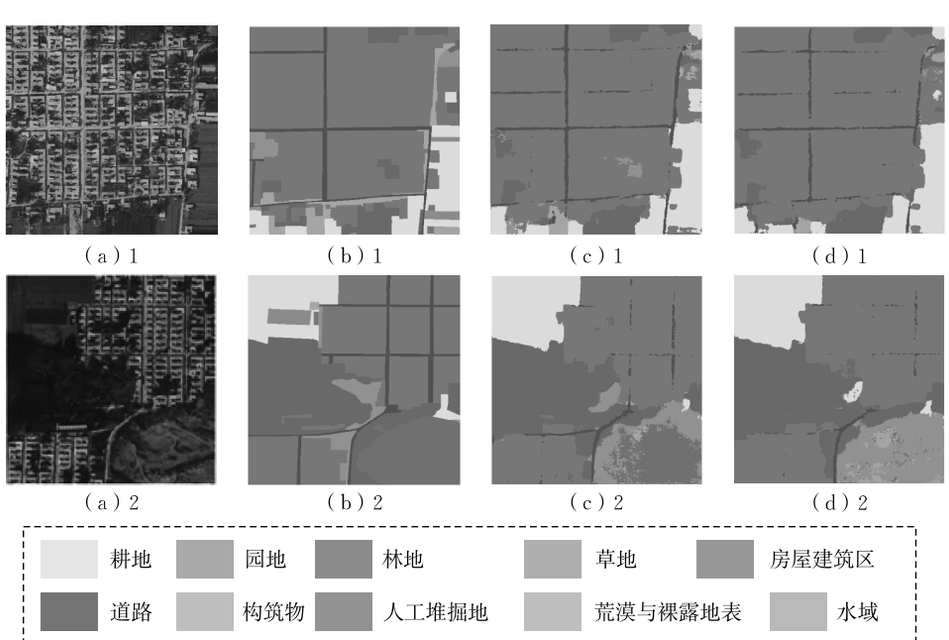

图 5　multistep(Base_lr=0.01)和 step(Base_lr=0.01)两种方法预测示例

4 结束语

本文基于地理国情普查成果,将 SegNet 模型应用于高分辨率遥感影像分割,以山东省为研究区域,从学习率是否自适应调整出发,比较随机梯度下降方法中 fixed、step、multistep 等学习率调整策略和 AdaGrad、AdaDelta 方法下的精度评价指标,提出了一些适用的学习率调整策略建议。目前,深度学习已广泛应用于图像处理领域,遥感影像分类领域也受到业内人士的关注,训练网络模型,学习率是较难设置的超参数之一。本文的研究是在高分遥感影像分类中探索细化的学习策略,其结果应能为遥感影像分类中训练深度学习模型提供支持。

参考文献:(略)

第一作者简介:杜娟,女,1987 年生,工程师,主要从事于空间统计方面工作。E-mail:mhwgo_jane@163.com