文章编号：1004-4353(2021)04-0350-06

基于迁移学习的少样本朝鲜语古籍文字的识别方法

薛春寒， 金小峰
（延边大学 工学院，吉林 延吉 133002）

摘要：为解决少样本朝鲜语古籍文字识别精度低的问题，提出了一种基于迁移学习的少样本文字识别方法. 首先提出了一种结合传统数据增强和条件深度卷积生成对抗网络的数据增强方法，以此扩充朝鲜语古籍文字图像的训练样本数. 其次，将富样本集预训练得到的模型迁移到少样本数据集的学习任务中，以此实现少样本的朝鲜语古籍文字识别. 实验结果表明，提出的数据增强方法能够满足模型预训练和少样本的学习要求，且 VGG16、ResNet18 和 ResNet50 3 种网络模型在测试集上均获得良好的识别性能，其中 ResNet50 的识别准确率最高(99.72%). 因此，该方法可有效解决小样本的朝鲜语古籍文字识别问题，并可为其他语种的小样本文字识别提供参考.

关键词：文字识别；少样本学习；数据增强；生成对抗网络；迁移学习
中图分类号：TP391.4 **文献标识码**：A

Few-shot optical characters recognition method of Korean historical document based on transfer learning

XUE Chunhan， JIN Xiaofeng
（*College of Engineering*, *Yanbian University*, *Yanji* 133002, *China*）

Abstract: To solve issue of low recognition accuracy in Korean historical document characters recognition with lacking samples, a novel method of transfer learning based few-shot optical character recognition is proposed. First, data augmentation fusing traditional and CDCGAN methods is exploited to expand scale of character-segmented image samples. Second, for few-shot training task implementation with homologous-source tactics, transfers the pre-trained model which is obtained by using abundant dataset to accomplish few-shot optical character recognition of Korean historical document. Experimental results show that proposed data augmentation method meets the requirements of pre-training and few-shot learning. VGG16, ResNet18 and ResNet50 outstand performance on test dataset, and ResNet50 achieves the top accuracy at 99.72%. Therefore, the proposed method is aeffective solution to solveisuue of few-shot Korean historical document optical characters recognition, and the method is valuable for same issue in other languages.

Keywords: optical character recognition; few-shot learning; data augmentation; generative adversarial network; transfer learning

---

收稿日期：2021-04-06
基金项目：吉林省教育厅“十三五”科学技术项目(JJKH20191126KJ)；延边大学外国语言文学世界一流学科建设项目(18YLPY14)
第一作者：薛春寒(1996—)，女，硕士，研究方向为计算机视觉.
通信作者：金小峰(1970—)，男，硕士，教授，研究方向为语音信息处理、计算机视觉、机器人技术.

中国少数民族古籍数字化是“中华古籍数字资源库”的重要组成部分[1]. 与国内其他少数民族古籍数字化研究工作相比，中国朝鮮语古籍的数字化工作相对滞后，因此开展和推进中国朝鮮语古籍的数字化研究是一项迫切的任务. 古籍文字自动识别作为古籍数字化总目标的基础性研究，是建设古籍全文索引库和提供信息检索服务的必要前期环节. 传统的光学字符识别 OCR(optical character recognition)技术由于过度依赖于样本底层视觉特征的选择和提取，使得分类器难以从中提取更高级别的语义特征. 近年来，随着深度学习方法的发展，一些学者提出了基于 LeNet5、AlexNet、VGGNet、ResNet 等[2]的深度网络模型的文字识别方法.

与传统的模式识别方法相比，深度学习方法需要大规模数据集的支撑，而国内少数民族古籍普遍存在存量少、数据集不足的问题，这给利用基于深度学习的古籍文字识别方法研究少数民族古籍带来诸多困难. 基于上述问题，一些学者提出了利用数据增强(data augmentation)和迁移学习(transfer learning)的方法来解决样本不足的问题，并取得了较好的研究结果[3-10]. 为此，本文结合数据增强和迁移学习方法，通过扩充数据集规模和知识迁移，提出了一种少样本朝鮮语古籍文字识别方法，并通过实验验证了本文方法的有效性.

1　朝鮮语古籍数据集的数据增强方法

朝鮮语古籍存在多语种文字混排、文字大小不同、相邻文字黏连等现象. 本文中的朝鮮语古籍文本图像样本均是采用文献[11]的方法获取的(采用文字切分法获取每个文字区域的图像)，所采用的数据增强方法均是指对文本图像进行的. 图 1 为朝鮮语古籍图像的样张.

图 1　朝鮮语古籍图像样张

1.1　基于图像变换的增强方法

数据增强方法主要是通过对数据进行多种变换来扩充训练集，以此实现对小数据集的样本扩充. 本文采用随机仿射变换和弹性变换进行数据增强. 随机仿射变换是将空间内的某一个向量随机映射到另一个向量空间，其映射过程相当于对图像进行平移、旋转、放缩、剪切、翻转等操作，或是这些操作的任意组合. 一个集合的仿射变换可表示为 $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$，其中 $\boldsymbol{y}$ 为变换后得到的目标向量，$\boldsymbol{A}$ 为变换矩阵，$\boldsymbol{x}$ 为原始向量，$\boldsymbol{b}$ 为平移向量.

弹性变换增强方法是首先在图像的像素点上的 x、y 方向上生成 $-1\sim1$ 之间的随机数，这些随机数表示像素点在 x、y 方向的移动距离(记为 $\Delta x(x,y)$，$\Delta y(x,y)$)，可表示为：

$$\begin{cases}\Delta x(x,y)=\mathrm{rand}(-1,+1),\\ \Delta y(x,y)=\mathrm{rand}(-1,+1).\end{cases}$$

然后再用标准差为 σ 和均值 μ 为 0 的高斯核对文本图像 I 作卷积运算，以此得到新的增强文本图像 I_σ. 该过程可表示为 $I_\sigma=I\times G_\sigma$，其中 G_σ 是高斯核，$G_\sigma=\dfrac{1}{\sqrt{2\pi}\,\sigma}\mathrm{e}^{-(\Delta x^2+\Delta y^2)/(2\sigma^2)}$.

1.2　基于条件生成对抗网络的增强方法

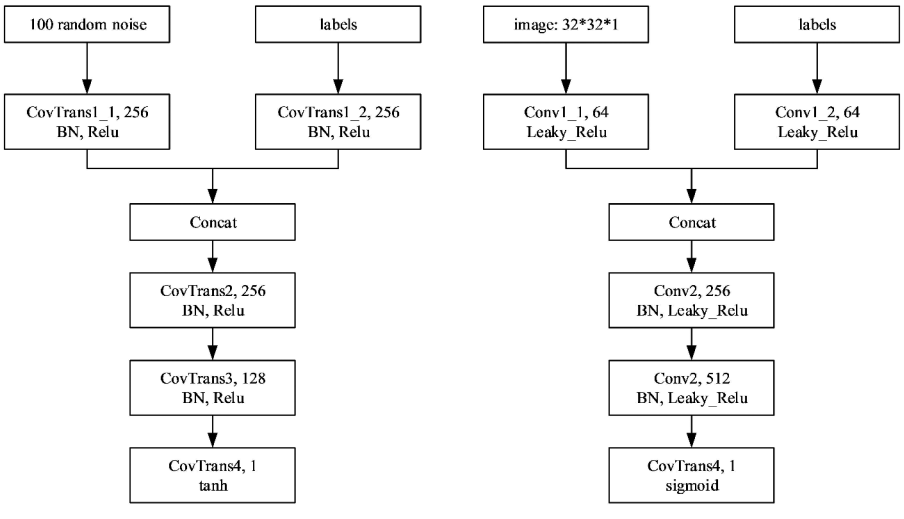
由于在数据样本少和类别样本数严重不平衡时，采用基于图像变换的增强方法生成的新样本会极易陷入饱和状态，因此本文引入 GAN 方法以进一步增强数据. 基础的 GAN 网络模型包含两个核心子网络：生成网络 G(generator) 和判别网络 D(discriminator). G 和 D 的学习实际上是一个博弈过程，G 可以对任意的噪声 z 产生一个伪样本 $G(z)$，而对于任意输入 x，$D(x)$ 能够对其真伪进行判别. GAN 方法的目的就是能够对真实样本 $x\sim P_{\mathrm{data}}(x)$ 实现期望最大化 $\log D(x)$，而对于伪样本 $z\sim P_z(z)$ 实现期望最小化 $\log(1-D(G(z)))$，因此 GAN 的学习过程可表示为：

$$\min_G \max_D V(D,G)=E_{x\sim P_{\text{data}}(x)}[\log D(x)]+E_{z\sim P_z(z)}[\log(1-D(G(z)))]. \tag{1}$$

由于 GAN 网络学习过程所生成的样本是随机和不具有类别的信息,因此本文在式(1)的基础上加上一个类别信息 y 作为条件,以此得到具有类别信息的生成样本.增加类别信息 y 条件后,GAN 网络变为 CGAN 网络,其训练过程可表示为:

$$\min_G \max_D V(D,G)=E_{x\sim P_{\text{data}}(x)}[\log D(x|y)]+E_{z\sim P_z(z)}[\log(1-D(G(z|y)))]. \tag{2}$$

另外,为了提高 GAN 网络的特征提取能力、伪样本的生成质量和收敛速度,本文引入了条件深度卷积生成对抗网络(conditional deep convolution generative adversarial network,CDCGAN).图 2 是本文采用 CDCGAN 生成的网络和判别网络的结构图.



(a) 生成网络 (b) 判别网络

图 2 基于 CDCGAN 生成的网络和判别网络的结构图

2 基于迁移学习的朝鲜语古籍文字识别方法

领域和任务是迁移学习中的两个重要概念.迁移学习的目的就是利用在源领域学到的知识实现其在目标领域上的学习任务,但迁移学习并不能适合于任何情况,如当源领域和目标领域差异过大时可能会发生负迁移.为此,本文使用已经训练好的预训练模型对全新的目标任务进行训练,即在不改变网络结构的前提下和保持预训练模型的参数基础上,以很小的学习率对模型进行微调,使得模型能够适应新的学习任务.本文提出的基于迁移学习的少样本朝鲜语古籍文字识别方法的流程图见图 3.

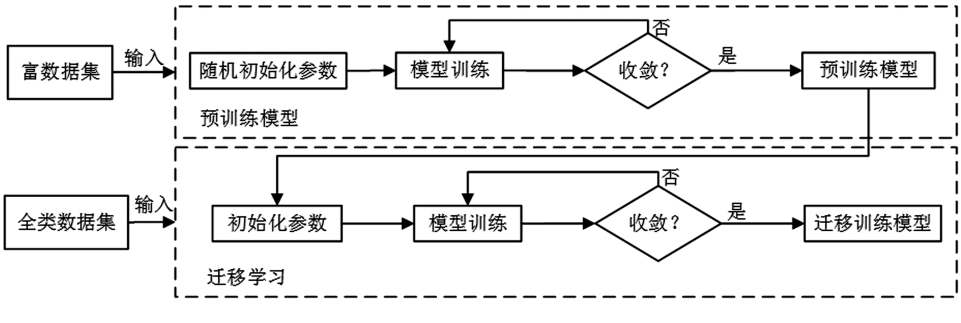


图 3 基于迁移学习的少样本朝鲜语古籍文字识别方法的流程图

在进行迁移学习时,首先按 1.1 和 1.2 中的方法对数据进行增强,然后选取出样本较多的富数据集作为源域预训练模型的数据集,目标域则采用全类数据集(即增强前后的所有数据).源模型到中间模型

的转变是一个典型的基于模型的迁移学习策略，该策略通过将预训练模型和层冻结方法相结合的方法来实现朝鲜语古籍文字识别的训练过程，该过程主要包括如下几个关键阶段：

第 1 阶段：保留预训练模型的权重参数，并将其作为第 1 阶段模型训练过程中的特征提取源.

第 2 阶段：加载预训练模型的参数和权重(除全连接分类层). 定义目标数据集的类别数为 c，并将最后一层全连接层由原来的神经元数改为 c 元的 softmax 分类器，同时冻结预训练模型的卷积层以及池化层的参数和权重，并对新的分类器层进行训练.

第 3 阶段：将在源数据集上已经预训练过的模型的卷积和池化层保留的参数迁移到目标模型上，并与第 2 阶段中训练的新的全连接分类层进行对接，以此得到中间模型.

3　实验结果与分析

实验数据集采用文献[11]中的方法对朝鲜语古籍《同文类解》进行文字切分. 经切分后共得到 1 299 种朝鲜语文字类别，24 390 个文字图像样本(每个样本均做人工类别标注). 由于该数据集中很多的样本数量过于贫乏(样本数小于 30 个)，因此该数据集具有少样本的特点. 本文从数据集中选取了样本数较为充足的 141 个类别作为本文实验数据集，其中将样本数≥100 的 44 类作为富数据集(记为 S1)，将剩余的 97 类作为少样本数据集(样本数为 30～100，记为 S2).

3.1　数据增强实验

采用 1.1 和 1.2 中的数据增强方法对 S1 数据集进行数据扩充，图像变换包括随机仿射结合弹性变换和随机仿射结合噪音扰动两种方法，其中噪音类型包括 Gaussian、Salt&pepper 和 speckle. 由于 CDCGAN 在训练时需要一定数量的样本，因此本文将变换得到的图像样本与原始样本进行合并，并将合并后的样本作为 CDCGAN 训练集. 经实验优化，CDCGAN 的超参数为：epochs＝400，batch_size＝20，learning rate＝0.000 2.

数据增强扩充后的样例如表 1 所示. 从表 1 可看出：传统的数据增强方法通过控制参数取值虽然能够有效地扩充新样本，但过度的形变和噪声扰动会使新样本严重失真；而 CDCGAN 产生的新样本能够丰富训练数据集，所以该方法可以避免模型训练过程出现的模式坍塌现象.

表 1　数据增强后的样本样例

类别	原始样本样例	扩充样本样例		
		随机仿射＋弹性变换	随机仿射＋噪音扰动	CDCGAN
라				
버				
허				
밤				

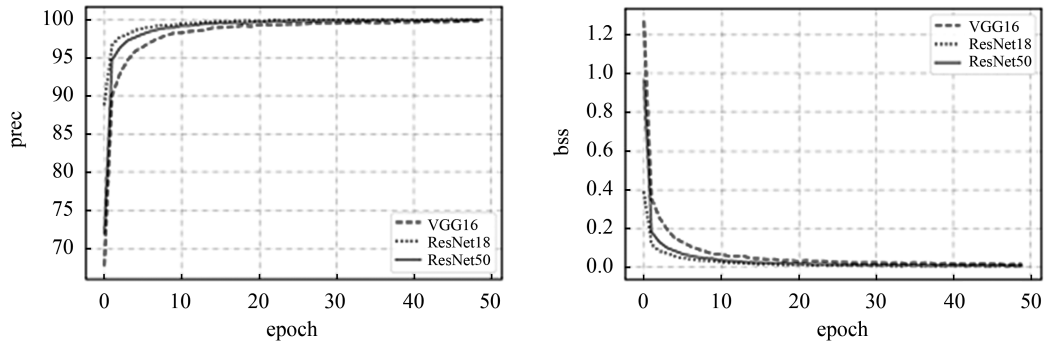
为了提高模型性能，对 S1 和 S2 分别进行数据增强，并将增强后的数据集分别记为 N1 和 N2. N1 和 N2 的样本分布以及数据集的划分情况见表 2. 由表 2 可以看出，合并后的数据集(记为 N)的样本数量(94 400)满足深度网络模型的训练要求.

表 2　数据集 N1 和 N2 的样本划分和样本分布

数据集	样本类别数	每类总样本数	总样本数	数据集划分(每类)		
				训练样本	验证集样本	测试集样本
N1	44	1 020	44 000	800	200	20
N2	97	520	50 440	400	100	20
N	141	—	94 400	—	—	—

3.2 预训练模型的对比实验

在数据集 N1 上分别利用 VGG16、ResNet18 和 ResNet50 卷积神经网络模型对数据进行预训练，并对这 3 种模型使用不同的优化方法和超参数进行调优. 经训练和调优后，本文最终取 batch_size＝16，learning_rate＝0.000 1. 为了减轻模型的过拟合现象，本文使用 L2 正则化和 dropout 方法对模型进行约束，其中神经元失活率取 0.5. 为了提高训练效率，本文使用 Adam 优化方法对模型进行优化. 图 4 是 3 种卷积神经网络模型总体的查准率和损失函数随迭代次数的变化情况. 本文以交叉熵作为刻画预测值和真实值之间差距的损失函数.



(a) 查准率随迭代次数的变化　　(b) 损失函数随迭代次数的变化

图 4　查准率和损失函数随迭代次数的变化情况

由图 4(a)可以看出：当 3 种模型的迭代次数小于 10 次时，其查准率均呈上升趋势；当迭代次数超过 10 次后，3 种模型的查准率逐渐趋于稳定(说明训练模型已经收敛). 由图 4(b)可以看出，整体损失值随迭代次数的增加而不断下降，其中 ResNet18 和 ResNet50 的下降趋势相同，并优于 VGG16.

表 3 是 3 种模型在数据集 N1 上的实验结果. 实验中查准率(presicion，P)的计算公式为：

$$P=\frac{TP}{TP+FP}, \tag{3}$$

其中 TP 是正确识别的样本数，FP 是错误识别的样本数. 由表 3 可以看出，3 种模型在验证集和测试集上的分类精度均较高，且较为接近. 这表明这 3 种模型在 N1 训练集中能够学习到有效的特征，扩充后的数据集规模能够满足 3 种网络模型的训练和分类要求，因此可将训练的结果作为迁移学习的预训练模型应用于少样本朝鲜语古籍文字的识别当中.

表 3　3 种模型在数据集 N1 上的实验结果

模型	训练集损失	训练集准确率/%	验证集损失	验证集准确率/%	测试集准确率/%
VGG16	0.013	99.66	0.140	98.72	99.21
ResNet18	0.006	99.88	0.062	99.07	99.77
ResNet50	0.004	99.86	0.051	99.10	99.48

3.3 迁移学习的对比实验

由于数据集 N2 中的每类样本的数量都小于 N1 的规模，无法直接用于模型的学习，因此本文借助迁移学习中的预训练模型(基于 N1 得到的预训练模型)学习 N2. 为了证明同源迁移的优势，本文将基于 ImageNet[12] 的预训练模型用于 N2 的训练. 在迁移学习的过程中，冻结除全连接层以外的所有网络参数，且迁移学习的训练只用于更新全连接层的参数. 迁移学习的实验结果见表 4.

表 4　基于 N1 和 ImageNet 的预训练模型在 N2 数据集上的训练结果

预训练模型	模型	训练集损失	训练集准确率/%	验证集损失	验证集准确率/%	测试集准确率/%
ImageNet-based	VGG16	0.005	99.14	0.399	90.58	98.35
	ResNet18	2.423	45.73	2.372	44.18	36.08
	ResNet50	2.302	45.24	2.291	44.92	43.04
N1-based	VGG16	0.003	99.10	0.101	97.69	99.80
	ResNet18	0.073	98.20	0.060	97.31	99.48
	ResNet50	0.090	97.31	0.068	98.10	99.18

由表 4 可以看出，基于 N1 的预训练模型的实验结果均明显优于基于 ImageNet 的预训练模型. ImageNet 预训练模型的实验效果低的原因是 ImageNet 预训练模型的源域和目标域的图像结构差异较大，使得其在训练 N2 时只能从 ImageNet 预训练模型中获取图像的通用知识.

为了验证全类数据集的分类识别效果，本文采用在数据集 N1 上学习获得的预训练模型对测试集 N 进行分类实验，结果显示 VGG16、ResNet18、ResNet50 3 种网络模型的识别精度分别为 99.54%、99.48%、99.72%. 由此可以看出，3 种网络模型采用本文提出的迁移学习方法进行文字识别时均可达到良好的分类效果，其中 ResNet50 的识别精度最高(99.72%). 这表明，采用这 3 种模型并利用 N1 数据集训练得到的预训练模型对少样本 N2 数据集进行迁移学习是有效的，因此这 3 种模型可用于识别少样本的朝鲜古籍文字图像.

4　结论

研究表明，本文提出的基于迁移学习的少样本朝鲜语古籍文字识别方法可有效解决少样本的朝鲜语古籍文字识别问题，且可为其他语种少样本的文字识别研究提供参考. 本文在研究中只对类样本数≥30 的数据集进行了数据增强和识别，在今后的研究中我们将对样本数少于 30 的数据集进行研究.

参考文献：

[1] 杨凡. 大数据框架下古籍数字化发展趋势研究[J]. 图书馆学刊，2017，39(9)：74-77.

[2] 张亚倩. 卷积神经网络研究综述[J]. 信息通信，2018，191(11)：27-29.

[3] 柴伟佳，王连明. 卷积神经网络的多字体汉字识别[J]. 中国图像图形学报，2018，23(3)：410-417.

[4] 陈善雄，王小龙，韩旭，等. 一种基于深度学习的古彝文识别方法[J]. 浙江大学学报(理学版)，2019，46(3)：4-12.

[5] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Cambridge：MIT Press，2014：2672-2680.

[6] 俞彬. 基于生成对抗网络的图像类别不平衡问题数据扩充方法[D]. 广州：华南理工大学，2018.

[7] PAN S J，YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge & Data Engineering，2010，22(10)：1345-1359.

[8] ZEILER M D，FERGUS R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision. Berlin：Springer，2014：818-833.

[9] TANG Y，WU B，PENG L，et al. Semi-supervised transfer learning for convolutional neural network based Chinese character recognition[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Piscataway：IEEE，2018：441-447.

[10] JANGID M，SRIVASTAVA S. Handwritten devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods[J]. Journal of Imaging，2018，4(2)：41.

[11] 刘星辰，金小峰. 朝汉混排古籍的文字切分方法[J]. 计算机工程与应用，2020，56(11)：135-141.

[12] JIA D，WEI D，SOCHER R，et al. ImageNet：A large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition. Miami，FL，USA：IEEE，2009：248-255.