文章编号：1004-4353(2023)02-0183-06

基于 DeepCluster 的朝鲜语古籍文字图像的无监督聚类方法研究

刘晓童， 赵梦玲， 王桂荣， 金小峰
（延边大学 工学院，吉林 延吉 133002）

摘要：为了提高朝鲜语古籍文字图像的标注效率，提出了一种基于 DeepCluster 的朝鲜语古籍文字图像的无监督聚类方法. 首先，基于 DeepCluster 对 AlexNet 卷积网络进行简化；然后，采用 Sobel 滤波器的线性变换消除图像域中的颜色和增加局部图像的对比度；最后，利用数据增强方法强化模型对朝鲜语古籍样本特征的学习能力. 在无标注的朝鲜语古籍文字图像数据集上进行实验显示，该方法的准确率和 NMI 指标比 DCN 方法分别提高了 15.32 个百分点和 0.180. 由此表明，该方法可有效提高文字图像的标注效率，可应用于朝鲜语古籍文字标注数据集的构建中.

关键词：无监督聚类；朝鲜语古籍；DeepCluster；AlexNet 卷积网络；深度学习；图像数据集；文字图像
中图分类号：TP391.1 **文献标识码**：A

Research on unsupervised clustering method of Korean ancient book character images based on DeepCluster

LIU Xiaotong, ZHAO Mengling, WANG Guirong, JIN Xiaofeng
(*College of Engineering, Yanbian University, Yanji* 133002, *China*)

Abstract: An unsupervised clustering method based on DeepCluster for Korean ancient text images was proposed for improve the tagging efficiency of ancient Korean text images. Firstly, the AlexNet convolutional network was simplified based on DeepCluster. Secondly, the linear transformation of Sobel filter was used to eliminate the color in the image domain and increase its local contrast. Finally, data enhancement methods were used to enhance the model for feature learning of Korean ancient text samples. Experiments on unlabeled Korean ancient text images dataset show that the accuracy and NMI metrics of the method improve 15.32 percentage points and 0.180, respectively, compared with the DCN method, thus indicating that the method can effectively improve the efficiency of text image labeling and can be further applied to the construction of Korean ancient character annotation dataset.

Keywords: unsupervised clustering; Korean ancient book; DeepCluster; AlexNet convolutional network; deep learning; image dataset; character image

0 引言

朝鲜语古籍具有多文种混排的特点，尤其以中朝两种文字混排的情况居多. 目前，朝鲜语标注数据

收稿日期：2023-03-20
基金项目：延边大学外国语言文学世界一流学科建设项目(18YLPY14)；国家社会科学基金重大项目(18ZDA306)；延边大学应用基础研究项目(延大科合字(2021)第 2 号)
第一作者：刘晓童(1998—)，女，硕士研究生，研究方向为计算机视觉.
通信作者：金小峰(1970—)，男(朝鲜族)，硕士，教授，研究方向为语音信息处理、计算机视觉.

集的匮乏是影响研究朝鲜语古籍文字识别的关键因素之一. 由于人工标注数据存在效率低和成本高的问题,因此如何利用自动标注方法来构建朝鲜语古籍文字图像数据集,并以此为进一步研究朝鲜语古籍文字的识别方法和实现朝鲜语古籍数字化具有重要的意义. 为此,一些学者对此进行了研究. 例如:苏向东[1]针对蒙古文古籍标注数据集匮乏的情形,提出了一种半自动样本选取方法. 研究显示,该方法可有效提高无标注数据的标注效率,但对未标注数据集只能进行粗分类. 姜丽[2]提出了一种基于 BIRCH 和改进 K 中心点算法的古籍汉字图像聚类方法. 研究显示,该方法可对古籍汉字图像进行有效分类,但作者未利用该方法构建标注数据集. Yang 等[3]针对手写数据集提出了 DCN 方法,研究显示该方法可有效提高图像的聚类质量. 王畅等[4]提出了一种将聚类和跟踪相融合的人脸图像数据集的构建方法,研究显示该方法可提升人脸数据集的生成效率和准确率. Yan 等[5]针对因标注数据集匮乏导致视觉任务研究受限的问题,提出了 Clusterfit 方法. 研究显示,该方法可显著提高预训练模型提取视觉特征的鲁棒性,且模型运用少量的与预训练任务相关的特定信息即可进行聚类,从而使提取的特征更适合于下游任务.

DeepCluster[6]是一种可扩展的无监督学习聚类方法. 由于该方法将无监督聚类与深度神经网络相融合,因此其具有不需要借助已标注数据或特定领域先验知识的优点,并可将学习到的通用特征应用于下游分类任务中. 基于此,本文利用 DeepCluster 聚类方法提出了一种朝鲜语古籍文字图像的聚类方法,并通过实验验证了该方法的有效性.

1 DeepCluster 聚类方法

DeepCluster 网络模型的总体网络架构如图 1 所示. 模型的输出由聚类和分类两个分支构成,且这两个分支共享卷积网络的参数. 聚类分支的任务是将卷积网络提取的特征输入到聚类模型 K-means[7]中进行聚类. 其过程为:首先,利用主成分分析法对卷积网络输出的特征向量进行降维;然后,对降维的特征向量进行线性转换和 L2 归一化;最后,利用 K-means 对特征向量进行聚类(每个聚类分配一个伪标签),以此获得图像的聚类结果. 在对输入样本进行分类的过程中,模型通过误差的反向传播来调整卷积网络的参数.

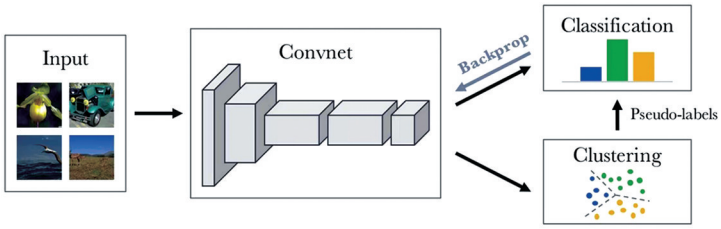

图 1 DeepCluster 网络模型的结构

DeepCluster 网络模型虽然能够在训练过程中实现收敛,但由于其参数采用的是交替迭代聚类和分类的更新方式,因此在学习过程中会得到一些没有意义的平凡解,进而会导致模型在聚类过程中出现空簇和参数平凡化的问题[8]. 为了避免得到平凡解,本文首先对由卷积网络提取的 $n \times d$ 维特征进行 K-means 聚类,以此得到 k 个簇,并将其作为初始的伪标签(形式为 k 维的 one-hot 编码);然后,通过交替使用式(1)和式(2)对特征进行聚类,以此生成伪标签;最后,通过预测生成的伪标签来更新网络参数.

$$F(y_n^*, \boldsymbol{C}) = \min_{\boldsymbol{C} \in \mathbf{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} \left| f_\theta(\boldsymbol{x}_n) - \boldsymbol{C} y_n \right|_2^2, \tag{1}$$

$$\mathscr{L}(\theta^*) = \min_{\theta, \boldsymbol{W}} \frac{1}{N} \sum_{n=1}^{N} l\{gw[f_\theta(\boldsymbol{x}_n), y_n]\}. \tag{2}$$

式(1)中,$f_\theta(\cdot)$ 为卷积网络映射函数,θ 为映射函数的参数,$\boldsymbol{x}_n$ 为训练样本,$y_n \in \{0,1\}^k$ 为 $\boldsymbol{x}_n$ 对应的标签;式(2)中,$l(\cdot)$ 为多项逻辑损失函数,$gw(\cdot,\cdot)$ 为预测伪标签的参数化分类器函数.

2　基于 DeepCluster 的朝鲜语古籍文字图像的聚类方法

2.1　卷积网络结构

基于 DeepCluster 的朝鲜语古籍文字图像聚类模型的结构如图 2 所示. 由于朝鲜语古籍文字图像具有样本稀少的特点(属于小型数据集)，所以本文在模型中选用了 AlexNet 卷积网络.

图 3 为典型的 AlexNet 卷积网络的结构图. 该网络由 5 个卷积层和 3 个全连接层组成，各卷积层(从左至右)分别有 96、256、384、384 和 256 个滤波器. 由于将批处理规范化作为归一化网络层能够减少对初始化的高度依赖和提高网络的泛化能力，以及能够使参数之间的联系保持不变(参数范围为 0～1)，因此本文在 DeepCluster 中用批处理规范化代替了 AlexNet 卷积网络中的局部响应归一化层. 另外，由于常用的无监督方法通常不能直接将图像域中的不同颜色作为标签，所以本文模型采用基于 Sobel 滤波器的固定线性变换来去除图像域中的不同颜色和增加其局部的对比度[9].

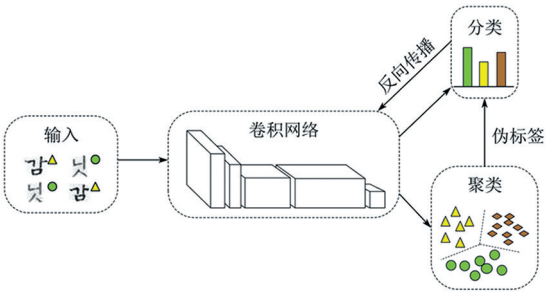

图 2　基于 DeepCluster 的朝鲜语古籍文字图像聚类模型的结构

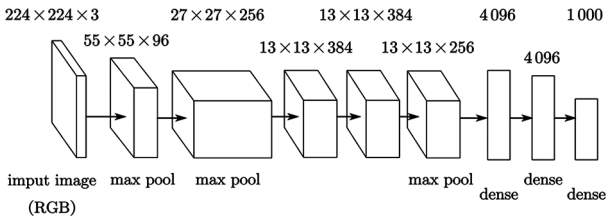

图 3　AlexNet 卷积网络的结构

本文模型对朝鲜语古籍文字图像进行聚类的流程为：

1)生成初始标签. 首先，对 AlexNet 卷积网络架构进行随机权重初始化，并移除最后一个全连接层；然后，利用网络对图像进行特征提取，并前向传递参数，以此获取图像模型的第 2 个全连接层的特征向量. 由 AlexNet 卷积网络的结构可知，此时网络输出的特征向量的维度为 4 096，如图 4 所示. 对 N 张图像重复上述操作过程即可得到一个[N，4 096]的图像特征矩阵.

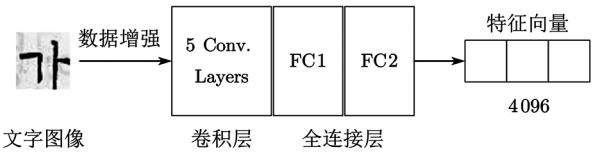

图 4　简化后的 AlexNet 卷积网络对文字图像进行特征提取的示意图

2)生成伪标签. 首先，采用主成分分析法对图像特征进行降维，使特征矩阵由 4 096 维减少至 256 维；然后，对降维之后的特征进行 L2 归一化处理，以此得到 N 幅图像的矩阵[N，256]；最后，利用 K-means 对预处理后的特征进行聚类，以此获得图像及其对应的聚类类别. 由此获得的集群类别结果即为训练模型的伪标签. 生成伪标签的过程如图 5 所示.

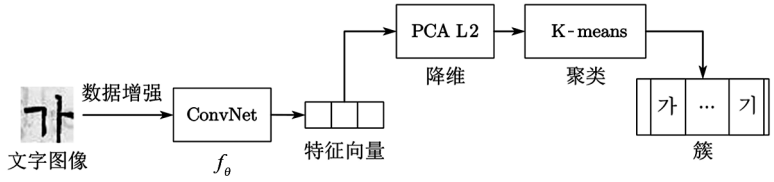

图 5　模型通过特征聚类生成伪标签的示意图

3)判别预测标签和真实集群标签. 首先,创建新批次的图像,以此使每个待聚类的图像有均等被包含于簇内的机会;然后,对待聚类的图像进行随机增强,以此得到图像和其相应的集群;最后,对模型进行训练(批量大小为 256),并运用交叉熵损失对比模型的预测标签和真实集群的标签,以此使得模型能够学习到有用的特征. 模型判别标签的过程如图 6 所示.

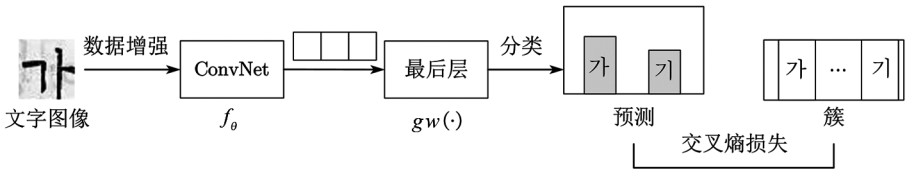

图 6 模型判别标签的示意图

2.2 文字图像的数据增强和模型聚类簇数的判定

为了提高网络的整体学习性能和获得更多的有效图像特征,在将图片输入模型之前,本文利用对输入图像进行随机水平翻转、随机大小变换以及纵横比的裁剪等方式对数据进行了增强.

对图像执行聚类时,首先确定被训练的类别数量. 确定的集群类别数量虽然越多越可对未标注的图像进行更细粒度的分组,但为了便于对聚类结果进行人工判别和标注,本文依据现存的朝鲜语字符类别数量(11 172 个)以及通过设置不同类别数量进行聚类试验,最终将朝鲜语古籍文字图像的聚类簇数设置为 20 000 个.

3 实验结果与分析

3.1 数据集和实验环境

本文所用数据集来源于《同文类解》《阐义昭鉴谚解》和《谚解胎产集》3 本朝鲜语古籍. 对这 3 本古籍进行扫描后共获得文本图像 875 张,其中《同文类解》160 张,《阐义昭鉴谚解》555 张,《谚解胎产集》160 张. 在上述古籍中,《同文类解》收录了对应汉语的朝鲜语对译词和对应满语的朝鲜语对译词,《谚解胎产集》由朝鲜语和汉语两种语言书写,《阐义昭鉴谚解》为手写朝鲜语版本. 对上述古籍文本图像进行文字图像切割后共得到 303 167 张文字图像. 图 7 是切分的部分文字图像的样例.

图 7 部分朝鲜语古籍文字图像的示例

实验硬件环境为 Intel(R) Core(TM) i7-7820X CPU+NVIDIA GeForce RTX 2080(8 GB 显存),软件环境为 Python 3.7.13+Pytorch 1.12.1.

3.2 实验的评价指标

聚类结果评价指标采用准确率和标准化互信息(NMI)[8]. NMI 的计算公式为:

$$NMI(A;B)=\frac{I(A;B)}{\sqrt{H(A)\cdot H(B)}}. \tag{3}$$

其中:A 和 B 为随机变量;$I(A;B)$ 为 A 和 B 的互信息,$I(A;B)=H(A)-H(A\mid B)$;$H(\cdot)$ 为随机变量的信息熵. 由式(3) 可知:若 A 和 B 相互独立,则 $NMI(A;B)=0$;若由 A 可确定 B,或由 B 可确定 A,则 $NMI(A;B)=1$.

3.3 聚类实验结果

实验数据集采用由上述切割得到的无标注的 303 167 张文字图像. 实验中,设置批量大小为 256,初始聚类簇数为20000. 图8为NMI值随迭代轮次的变化趋势. 由图8可以看出:NMI值随迭代轮次的

增加而呈增大趋势. 模型在训练初期时，由于卷积网络未能提取文字图像的有效特征，因此导致聚类效果较差，表现为 NMI 值较小. 当迭代轮次逐渐增加时，模型通过不断更新卷积网络的参数，进而不断提高了模型对不同类别特征的提取能力和聚类效果，表现为 NMI 值逐渐增大. 当迭代轮次达到 500 时（$NMI=0.89$），曲线上升趋势趋于稳定，表明此时模型已收敛.

图 9 是模型训练稳定时部分聚类结果中的簇. 由图 9 可以看出，图像尽管受到了多种干扰（如尺寸不同、切分不准确以及噪声等），但模型的聚类结果仍是准确的.

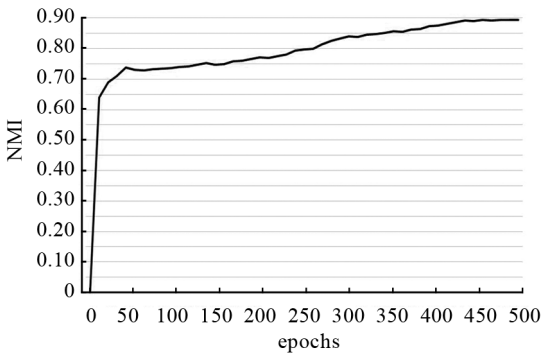

图 8　NMI 值随迭代轮次的变化趋势

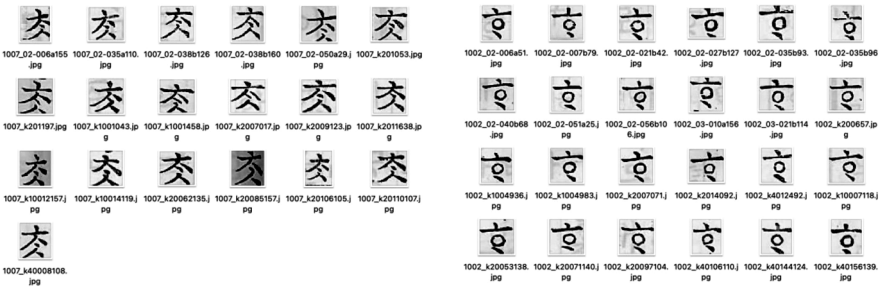

图 9　聚类结果的部分示例

为了验证本文方法的优越性，将本文方法与 DCN 方法进行了对比实验. 实验中，均使用上述切分的数据集（303 167 张无标注的朝鲜语古籍文字图像）. 实验结果见表 1. 由表 1 可以看出，本文方法的准确率和 NMI 值比 DCN 方法分别提高了 15.32 个百分点和 0.180. 该结果表明，本文方法对文字图像的聚类性能显著优于 DCN 方法对文字图像的聚类性能.

表 1　本文方法和 DCN 方法的聚类结果

聚类方法	评价指标	
	准确率/%	NMI
DCN	68.24	0.715
本文方法	83.56(↑15.32)	0.895(↑0.180)

4　结论

研究表明，本文提出的基于 DeepCluster 的朝鲜语古籍文字图像聚类方法的准确率和 NMI 值比 DCN 方法分别提高了 15.32 个百分点和 0.180，因此该方法可为构建朝鲜语古籍数据集提供参考. 在今后的工作中，我们将探讨适用于小样本朝鲜语古籍数据集的聚类方法，以进一步提高构建小样本标注数据集的有效性.

参考文献：

[1]　苏向东. 基于深度学习和知识策略的蒙古文古籍识别研究[D]. 呼和浩特：内蒙古大学，2016.

[2]　姜丽. 基于 BIRCH 和改进 k 中心点算法的古籍汉字图像聚类研究[D]. 保定：河北大学，2012.

[3]　YANG B, FU X, SIDIROPOULOS N D, et al. Towards K-means-friendly spaces: simultaneous deep learning and clustering[C]//International Conference on Machine Learning. Sydney: PMLR, 2017:3861-3870.

[4]　王畅，金璟璇，金小峰. 聚类与跟踪相结合的人脸数据集生成方法研究[J]. 延边大学学报（自然科学版），2019，45(3)：221-227.

[5]　YAN X, MISRA I, GUPTA A, et al. Clusterfit: Improving generalization of visual representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: ICCV, 2020:6509-6518.

[6] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering for unsupervised learning of visual features[C]//Proceedings of the European Conference on Computer Vision. Munich: ECCV, 2018:132-149.
[7] COATES A, NG A Y. Learning feature representations with K-means[J]. Neural Networks: Tricks of the Trade, 2012:561-580.
[8] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: An astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus: ICCV, 2014:806-813.
[9] BOJANOWSKI P, JOULIN A. Unsupervised learning by predicting noise[C]//International Conference on Machine Learning. Sydney: PMLR, 2017:517-526.
[10] WENG Y, ZHANG N, YANG X. Improved density peak clustering based on information entropy for ancient character images[J]. IEEE Access, 2019,7:81691-81700.
[11] 陈扬,王金亮,夏炜,等.基于特征自动提取的足迹图像聚类方法[J].计算机科学,2021,48(S1):255-259.
[12] ZHAO H, CHU H, ZHANG Y, et al. Improvement of ancient shui character recognition model based on convolutional neural network[J]. IEEE Access, 2020,8:33080-33087.
[13] WANG X, GUPTA A. Unsupervised learning of visual representations using videos[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago: ICCV, 2015:2794-2802.
[14] 魏银华.基于Python的古汉语文本聚类应用研究[D].大连:大连理工大学,2018.
[15] 李丁园,李晓杰.基于多尺度残差卷积自编码器的图像聚类方法[J].吉林大学学报(信息科学版),2022,40(4):684-687.

(上接第148页)

6 结束语

本文研究了带形状参数的双三次Bezier三角曲面的光滑拼接,该曲面不仅保留了传统Bezier曲面的一些优良性质(插值性、边界性质、凸包性、几何不变性等),而且在不改变控制顶点的条件下可通过形状参数调节曲面形状.为提高Bezier曲面在复杂曲面造型中的构图能力,本文给出了带形状参数的双三次Bezier三角曲面的u向与u向、v向与v向、u向与v向间的G^1拼接定理及其算法.实例计算显示,曲面的形状参数和法矢模比例因子可分别调整曲面的整体形状和局部形状,并且其几何意义明显.即:在给定范围内,形状参数越大,曲面在整体上越靠近控制网格;法矢模比例因子越大,相邻公共边界线的两排控制顶点越靠近边界线处的控制顶点.本文研究结果有效解决了Bezier曲面难以进行局部调节的缺点.在今后的工作中,我们将对高次曲线曲面的光滑拼接进行研究.

参考文献:

[1] 施法中.计算机辅助几何设计与非均匀有理B样条[M].北京:高等教育出版社,2001:306-454.
[2] 黄吴戟,王志国.B样条曲面光滑拼接方法[J].机械制造与自动化,2020,49(3):71-74.
[3] 吴丽娟,李博,ABEYSINGHE ARACHCHIGE S S,等.B样条曲面拼接算法的设计与实现[J].沈阳师范大学学报(自然科学版),2019,37(6):549-553.
[4] 吴丽娟,张心慈,任海清,等.CNSBS曲面拼接方法的设计与实现[J].沈阳师范大学学报(自然科学版),2021,39(2):178-181.
[5] 王崇.基于近似光滑的样条曲面拼接方法研究[D].长春:吉林大学,2022.
[6] 杨军,黄倩颖.一类广义Bezier曲线及其性质研究[J].南昌航空大学学报(自然科学版),2020,34(4):19-24.
[7] 师晶.一种基于几何约束的插值曲线的参数连续性[J].沈阳大学学报(自然科学版),2019,31(1):78-83.
[8] 喻德生,徐迎博,曾接贤.一类双参数类四次三角Bézier曲线及其扩展[J].计算机工程与应用,2013,49(18):180-186.
[9] 胡钢,吉晓民,白晓波.广义带多参Bézier-like曲面及其拼接条件[J].计算机集成制造系统,2016,22(2):501-515.