基于深度学习的西夏文献数字化

张光伟

摘　要：西夏文献的数字化是近半个世纪以来西夏文研究学者一直在努力地工作，以往西夏文献的录入几乎都是手工进行，效率非常低。近年来，以深度学习为代表的人工智能技术进行手写文字识别的准确率已经达到或超过人类专家的水平，我们训练的深度神经网络模型在单个手写体西夏字数据集上的识别准确率超过97%。深度学习算法用于西夏字识别需要大量的标注数据，即每个西夏字都需要几十上百的实际例子；从大量的西夏文献中找到西夏字典中的每个字的多个实例需要的工作量巨大，而且有些西夏字的使用频率比较低更难以找到大量实例。我们提出了一种目标导向的混合算法来对西夏字进行标注：使用已有大量实例的西夏字生成虚拟字符表示那些实例数量不足的西夏字。实验证明这些虚拟字符在训练神经网络方面的效能与实际数据近似，让我们能够从大量未标注的西夏字图片中找到那些神经网络尚未识别的西夏字，并进行标注，而且这些虚拟字符可以与使用频率低的西夏字实例一起纳入训练集，以逐步迭代的方式训练识别能力更强的神经网络。在识别单个西夏字的基础上，我们通过在模型中引入循环神经网络单元，实现了西夏文献整列的识别，从而将传统的扫描文献切割为单个字符进行识别工作流程改进为以列为单位进行识别。TDM与列级别的识别模型相结合能够为我们构建高效的西夏文献自动转录系统，进而加速西夏文献的数字化进程。

关键词：深度学习；西夏文；自动识别；训练数据集；虚拟实例；整列识别

西夏文典籍是中国古代西夏王朝的知识载体，是研究中国历史的重要史料。21世纪以来俄藏、英藏、中藏西夏文原始文献陆续刊布，长篇语料的解读成为常态，学者们研究文献的附加工作便是编纂字、词索引，以便学界引用，但这些工作几乎都是手工完成，效率很低，然而制作词汇索引者还不到四分之一，而数量更多的佛经文献才刚刚进入整理与译释阶段。现代各个学科的研究借助计算机能够从根本上提高效率，西夏学的研究也不例外。从20世纪末开始，国内

基金项目：教育部人文社会科学青年基金项目（批准号17YJCZH239），国家社会科学基金西部项目（批准号18XKG003）。
作者简介：张光伟（1982— ），陕西师范大学历史文化学院讲师，主要从事数字人文研究。

外学者在西夏文数字化研究方面不断有成果面世，对西夏文数字化研究产生了深远影响，如李范文教授为西夏文录入设计的四角号码和类似汉字的五笔字型输入法，景永时的“西夏文字处理系统”以及多种西夏文字库等。[①]西夏文研究学者一直在进行的西夏文文献数字化以及建立语料库的努力对西夏学的研究意义重大，但工作量巨大，依靠传统的手工方法在短期内难以完成，例如韩小忙的西夏文世俗文献语料库的构建就花费了十多年时间。因此，研究西夏文自动识别、准确高效的文献转录方法对西夏历史文献数字化、构建文献语料库有非常重要的价值。

当今人工智能技术在很多语言（包括中文、英文等）的手写文字识别任务中已经达到甚至超过人类的水平。本文探讨利用深度学习进行西夏文自动识别相关的方法，包括不同深度神经网络在识别手写西夏文中的性能，训练深度神经网络的西夏文数据集的构建，以及从单个西夏字的识别到整列西夏字识别系统的构建，以期为西夏文的数字化提供一种高效且实际可行的解决方案。

一、基于深度学习的文字识别

近年来深度学习在图像识别、语音识别等领域取得突破性进展，其识别准确率几乎横扫其他传统的机器学习技术。2013年的国际文献分析与识别国际会议ICDAR(International Conference on Document Analysis and Recognition)的手写汉字识别比赛中富士通团队采用改进的卷积神经网络获得了脱机手写汉字识别（识别扫描的整页文稿中的汉字，也称为OCR）的第一名，识别率高达94.77%。[②]深度学习不但在汉字识别领域表现突出，在若干古文献识别领域也有深入研究和应用，例如巴基斯坦的乌尔都语[③]，泰语[④]，德语[⑤]等历史文献的识别。

我们认为深度学习在众多历史文献识别中的高准确率，也将在西夏历史文献的识别中发挥重要作用。西夏文已于2016年6月编入Unicode，每个西夏字都有其唯一的编码，共6125个西夏文字收录进Unicode 9.0中的西夏文区块，755个用于现代西夏文研究的部首偏旁添加在西夏文部首区块，西夏文的叠字符号收录于表意文字符号及标点区块[⑥]。虽然现在有数种西夏字的

①景永时：《西夏文数字化的现状与未来》、柳长青：《西夏文计算机数字化现状与展望》，《西夏学》第七辑，上海古籍出版社，2011年，第199—203、204—209页。

② F. Yin, Q. Wang, X. Zhang and C. Liu, ICDAR 2013 Chinese Handwriting Recognition Competition, 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 1464-1470.

③ I. Ahmad, X. Wang, R. Li, and S. Rasheed, Offline Urdu Nastaleeq optical character recognition based on stacked denoising autoencoder. China Communications, vol. 14, no. 1, pp. 146-157, January 2017.

④ T. Kobchaisawat and T. H. Chalidabhongse, Thai text localization in natural scene images using Convolutional Neural Network. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pp. 1-7, 2014.

⑤ J. Richarz, S. Vajda, R. Grzeszick, and G. A. Fink, Semi-supervised learning for character recognition in historical archive documents. Pattern Recognition, vol. 47, no. 3, pp. 1011-1020, Mar. 2014.

⑥ Tangut (Xixia) Script and Unicode. http://unicode.org/~rscook/Xixia/.

输入法，使用 Unicode 或者非 Unicode 的字库，但是西夏字的实际录入仍然比较烦琐，效率较低，且手工录入大量文献的准确率难以保证，甚至出现有些正式出版的刊物、书籍中出现乱码、错码的现象。因此，基于深度学习的高准确度的西夏文献的自动识别将成为一种高效的西夏文献录入方法，从而大大加快西夏文献的数字化进程。

深度神经网络极高的识别准确率需要大量标记数据的支撑，即需要人类专家大量标记的数据作为训练数据集。例如，图像识别领域著名的数据集 ImageNet[①]已经包含超过一千万张经过人类标注的图片；用于手写数字识别的 MNIST[②]数据集包含六万张经过标注的真实手写数字图片；中科院自动化所的手写汉字识别的数据集[③]包含数十万经过标注的真实手写汉字图片。我们构建的西夏字数据集包含直接从扫描版的西夏文献中提取的数量超过十万的手写体西夏字，它们属于 1500 多个不同的西夏字，这个数据集仍在不断地扩充。

文字识别是图像识别的一种，目前在图像识别领域有大量可以借鉴的神经网络模型，如 AlexNet，GoogLeNet，VGGNet，ResNet 等，在构建不同的视觉识别任务模型时可以这些经过验证的模型为基础进行修改，使之能够适应具体识别任务的需要。我们构建的西夏文识别系统在上述多个深度神经网络模型的基础上进行了修改以适用于手写西夏文识别。在我们的标注数据集上训练得到的西夏文识别神经网络模型的识别准确率达到 97% 以上，其中最高的是基于 GoogLeNet 的修改版识别准确率达到了 98% 以上。

除了需要大量的标注数据集以及可选的众多深度神经网络模型之外，深度神经网络的训练对计算设备的处理能力要求很高，所以目前多数的基于深度学习的应用需要在高性能 GPU 计算平台训练。我们目前使用的 GPU 集群由两块 Nvidia Titan V GPU 组成，每个 GPU 拥有 5120 颗 CUDA 核心，640 颗 Tensor 核心以及 12G 显存。在该平台上，不同的神经网络结构使用我们的西夏文数据集进行训练所需时间不同，但大多在四个小时左右达到稳定。

二、西夏字标注数据集的构建

基于深度学习的西夏文自动识别系统的构建主要包括以下三部分：(1)标注数据集，用于训练和测试深度神经网络的包含单个西夏字的大量图片，这些图片来源于真实的西夏文历史文献的扫描件，而且经过西夏文专家的识别并进行了标注；(2)深度神经网络，在深度学习专家和

① J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

② Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324, November 1998.

③ C. Liu, F. Yin, D. Wang and Q. Wang, CASIA Online and Offline Chinese Handwriting Databases, 2011 International Conference on Document Analysis and Recognition, Beijing, 2011, pp. 37-41.

西夏文专家的共同努力下确定神经网络结构，在训练和测试数据集（两者是将标注数据集划分而得，且没有交集）上进行训练和测试并不断调优，直至网络模型达到符合要求的识别准确率；（3）西夏文 OCR 识别，将训练完成的深度神经网络作为西夏文 OCR 系统的核心进行实际的西夏文献的自动识别。本部分主要介绍西夏字标注数据集的构建情况。

基于深度学习的文字识别一般采用的是监督式学习，即训练和测试的数据是经过专家识别并进行了标注的。深度学习与传统的机器学习相比其优势在于：我们只需要将尽可能多的、准确的标注数据提供给深度神经网络，而无须教给神经网络如何进行分类（即无须提供人工特征用于分类），实际分类能力是深度神经网络从提供给它的训练集中自行学习到的（即“端到端”特性）。基于深度学习的文字识别需要比较大规模的训练与测试数据集，每个字符往往需要几百上千个不同的实例。训练数据集包含了神经网络的学习目标，即数据的标签，我们用训练数据集“教会”神经网络“认识”文字；测试数据集的数据与训练集没有任何交集，在训练神经网络的过程中用来检验其学习效果。训练深度神经网络模型的标注数据集的规模和质量对识别的准确性和适应性是非常重要的。

缺少训练数据集是基于深度学习的西夏文自动识别中最大的困难。系统构建初期，我们从西夏文原始文献的扫描图片中切出大量单个西夏文字，并进行逐一标记：找到的同一个西夏字的所有图片被标记为同样的编号，这个过程为数据标注，如图 1 所示。

图1　西夏文自动识别训练数据集示例（单个字）

训练数据集是实现西夏文自动识别的重要基础设施，因此我们在将其完善之后将会公开，让更多的研究者参与到西夏文自动识别的研究中，从而进一步提高西夏文自动识别的准确率，并共同探讨该方法在西夏学研究中的更为深入的应用。任何训练数据集的获得都是深度学习系统构建过程中最为耗时、花费最大的部分，因此也有很多学者研究如何更高效地进行数据标注，

这也是我们近年来主要的研究内容。

我们使用多标签学习的方法来利用深度神经网络提取西夏字的部首[①]，这种方法与一般的深度神经网络不同：数据集中每个实例都被赋予多个标签。这里每个西夏字的标签是其包含的部首列表，如图 2 所示，这个列表可以视作西夏字部首词典。这个多标签深度神经网络模型是在我们识别西夏字的神经网络模型[②]的基础上修改而来，网络结构的变化主要是最后一层，从输出单个类别变为输出多个类别，即多标签。该方法的学习效果（指标 F1 是准确率和召回率的综合）如图 3 所示。

ID	Handwritten Tangut Character	Radical Vector
5442		[𗍫, 𗍉, 匕, 儿]
5447		[𗍫, 𗍫, 匕, 攵]
5449		[𗍫, 𗍫, 匕, 攵]

图2　以西夏字部首为多标签学习的数据集示例

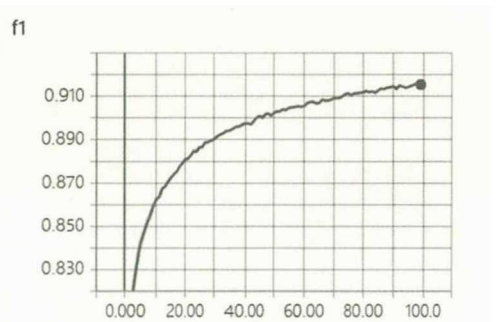

图3　多标签学习效果（F1指标）

多标签学习能够使用训练好的神经网络提取任何输入神经网络的西夏字的部首，这些监测到的部首列表可以作为从大量未标注数据集中标注西夏字的一种方法。这为我们构建西夏文训练数据集提供了一个可行的方案，但此方法对于那些使用频率低的西夏字无能为力，因此我们开发了目标导向的混合方法（Target-Directed Mixup，TDM）[③]来为这些字符构造虚拟的西夏字符，并用于训练神经网络来识别这些低频字或尚未标注的西夏字。对于未标注数据集中有大量支持数据的那些字符，我们可以使用 TDM 来从中找到真实的实例并加入训练集中，从而扩充西夏文识别模型的能力；那些罕见字就可以直接使用 TDM 的虚拟字符来代替，并训练神经网络来识别它们。如图 4 所示，经过三次混合得到的虚拟实例训练出

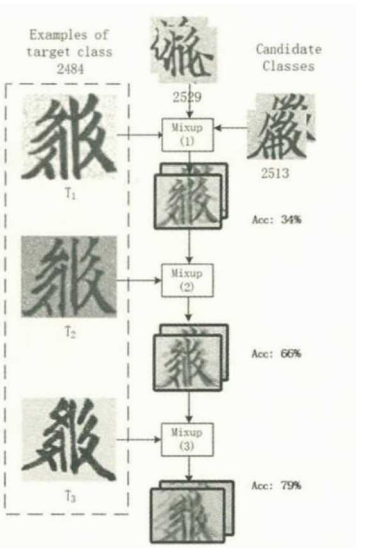

图4　目标导向的混合方法

① G. Zhang and Y. Zhao, Learning Radicals From Tangut Characters, 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, 2018, pp. 373-378.

② G. Zhang and X. Han, Deep Learning Based Tangut Character Recognition, 2017 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, 2017, pp. 437-441.

③G. Zhang and Y. Zhao, Target-Directed MixUp for Labeling Tangut Characters, 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 2019, pp. 202-207.

的神经网络模型对相应的真实西夏字的识别准确率已经能够达到79%，能够有效地在未标注数据集中进行相关西夏字的查找并标注。

三、西夏文识别系统

（一）单个西夏字的识别

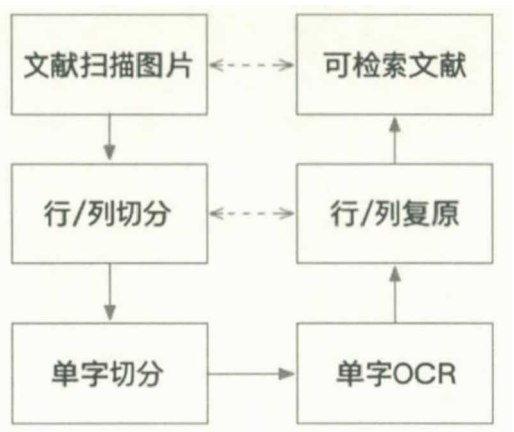

图5　单个西夏字识别流程

西夏文识别系统可以分为单个西夏字的识别和整列西夏字的识别，前者可以用于西夏文学习或研究中单个西夏字的识读，后者可以用于西夏文献的自动转录。传统的文献识别的过程一般需要将包含文字的文献图片进行切分，将单个字切出来，识别系统将每个字识别之后，按照原始文献的顺序将识别出的文字进行排列，其流程如图5所示。

我们在标注的数据集上通过对一些经典的深度卷积神经网络（Deep Convolutional Neural Network，DCNN）如AlexNet，GoogLeNet，DenseNet，MobileNet，以及SqueezeNet进行修改用于识别西夏字。由于文字识别中的输入图片一般使用单通道，因此我们对这几个神经网络进行了修改，在保持原有网络基本结构不变的前提下，将输入的通道数量改为单通道。其中，AlexNet是最早提出的深度神经网络，也是深度神经网络在性能上获得突破的网络，目前仍然用于很多视觉识别任务；在AlexNet之后，大量涌现出更深的网络如GoogLeNet，VGGNet，ResNet等，这些网络中我们都进行了相应的尝试，其中GoogLeNet，DenseNet能够达到比AlexNet更好的效果，VGGNet和ResNet虽然更深但并没有表现性能方面的提升。由于我们的目标不但包括在较高性能的平台上使用，如PC、服务器等，AlexNet、GoogLeNet等训练好的模型相对比较大，可以在这些平台运行，但计算能力较弱的移动设备是将来的进行西夏文献识别应用的重要平台，这些模型对于移动平台而言就有些大而难以顺利运行，因此我们也对MobileNet和SqueezeNet等轻量模型的变体进行了测试，它们能够达到与AlexNet，GoogLeNet和DenseNet近似的准确率，但它们训练的模型更小，适合在移动设备运行。

为了更快更好地训练深度神经网络，包含西夏字的图片在被深度神经网络处理之前，首先需要经过图片的预处理，我们使用两个预处理步骤：（1）统一尺寸，将所有图片统一调整为224x224像素；（2）二值化，将所有图片处理为白底黑字。

（二）整列西夏字的识别

传统的单字符识别需要对扫描文献进行字符级别的切分，把每个字符从中分割出来并进行

标注，识别之后的每个字符根据原始文献的位置进行文本结构的还原。这种方式在实际使用中由于字符的检测与切分效果在很大程度上取决于文献的整洁程度，以及字符与字符之间是否有连接的部分，如果有，则切割效果往往比较差，从而导致识别效果不好。西夏文献与中文历史文献中的文字的书写顺序一样，是垂直的，而且常见的西夏文献的列之间的分割比较明显，整列切分是比较容易。因此，以列为单位的西夏文识别将能够极大地提高我们构建西夏文献的自动识别系统的效率，其工作流程如图6所示。

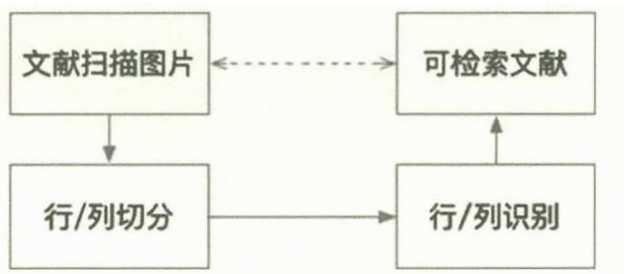

图6　整列西夏文识别流程

循环神经网络（Recurrent Neural Network，RNN）是主要处理序列数据的一种深度神经网络，我们前面使用的是深度卷积神经网络（DCNN），其输入输出的大小都是固定的，因此无法处理变长的序列。卷积循环神经网络（Convolutional Recurrent Neural Network，CRNN）[①]是将DCNN和RNN进行组合，其优势在于：（1）可以直接从序列标签来学习，而不需要序列中单个字符级别的标注（无须给出单个字符的具体位置）；（2）具有和DCNN一样的能够直接从图片数据中学习表征的能力，既不需要手工的特征也不需要切分和定位；（3）拥有RNN产生标签序列的能力；（4）序列长度没有限制，只需要统一输入图片的宽度；（5）参数数量比标准的DCNN模型要少得多，空间需求也少得多。

CRNN模型的训练也需要大量的标注数据，我们在已有单字数据集的基础上构建了列级别的训练数据集。我们训练的整列西夏文识别的CRNN模型对实际西夏文文献中文本列的识别已经能够达到：（1）识别其中包含西夏字符的数目，（2）整列字符的识别准确率达到90%以上，如图7所示。西夏文整列识别神经网络模型训练完成之后，我们根据图6设计实现了西夏文识别系统。当整页手写体西夏文献图片输入该系统之后，系统在后台首先对文献图片进行列级别的切分，接着切分得到的西夏文本列逐个被输入上述训练好的整列西夏文献识别神经网络模型，经过处理之后，该模型输出其推测结果，即该文本列所实际包含的西夏字，并呈现给用户。本系统以Web的形式提供服务，用户将需要识别的西夏文献图片上传到本系统，识别完成之后，用户将看到如图7所示的界面，鼠标移到其中的文本列上时识别的西夏字就会弹出显示，用户可以直接将整页识别结果下载。由于识别准确率难以达到100%，因此本系统在识别结果预览之外还提供了手动更正的功能，即用户可以在本系统对于少数识别错误的内容进行修改，之后再下载保存。由此可见，整列识别的西夏文OCR系统已经具备了实用性。

① B. Shi, X. Bai, and C. Yao, An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, Nov. 2017.

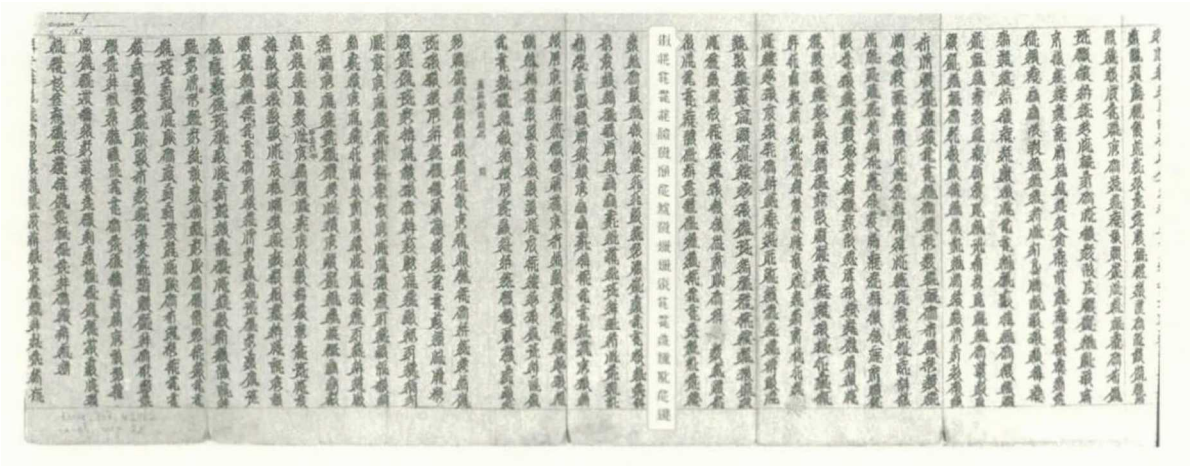

图7　西夏文识别系统界面

四、总结与展望

本文研究使用深度学习方法进行西夏文自动识别系统的设计、构建和实现，我们通过构建西夏文自动识别标注数据集，设计并训练深度神经网络模型，实现了对单个西夏字在训练集上的识别准确率达到97%以上，整列西夏字识别神经网络对实际文献中包含的文本列的识别准确率达到90%以上，因此该系统已经具备相当程度的实用性。在构建本系统的过程中，训练数据集的构建是最为耗时及花费最为昂贵的部分，我们也将近年来在西夏文训练数据集构建的工作进行了介绍。我们希望在学界的共同努力下，在西夏文字自动、准确、高效的识别基础上，实现西夏文献的自动转录，从而加快西夏文献数据库的构建，提高西夏文献研究和使用的水平。

（责任编辑：张笑峰）