

# 基于深度学习的地址分割算法在电力业务辅助录入系统上的应用研究

郭凤婵, 吴毅良\*, 罗序良, 陆庭辉, 刘翠媚

广东电网有限责任公司江门供电局, 广东 江门

收稿日期: 2023年8月3日; 录用日期: 2023年9月1日; 发布日期: 2023年9月8日

## 摘要

现代电力行业中的信息系统通常需要对地址信息进行特定的格式处理以提高数据的统计与分析能力。本文提出一种基于深度学习的地址分割算法, 对地址信息按行政区级进行分割。采用ELU激活层对LSTM网络进行优化以提高网络的整体性能, 另外采用GAN网络对数据集进行了增强, 进一步降低了训练过程中过拟合情况的发生。实验结果表明, 算法对于地址数据的分割平均准确率达到99%, 平均运算时间为0.1秒, 满足辅助录入系统需求。算法利用分割地址信息关联对应的供电局(所), 有效提升了电力业务的办理效率, 具有较好的应用前景。

## 关键词

深度学习, 地址分割, 辅助录入, LSTM网络

# Research on the Application of Address Segmentation Algorithm Based on Deep Learning in Power Service Auxiliary Input System

Fengchan Guo, Yiliang Wu\*, Xuliang Luo, Tinghui Lu, Cuimei Liu

Guangdong Power Grid Co., Ltd., Jiangmen Power Supply Bureau, Jiangmen Guangdong

Received: Aug. 3<sup>rd</sup>, 2023; accepted: Sep. 1<sup>st</sup>, 2023; published: Sep. 8<sup>th</sup>, 2023

\*通讯作者。

文章引用: 郭凤婵, 吴毅良, 罗序良, 陆庭辉, 刘翠媚. 基于深度学习的地址分割算法在电力业务辅助录入系统上的应用研究[J]. 计算机科学与应用, 2023, 13(9): 1655-1664. DOI: 10.12677/csa.2023.139164

## Abstract

Information systems in the modern power industry often require specific formatting of address information to improve data analysis and statistics capabilities. This paper proposes a deep learning-based address segmentation algorithm that separates address information into administrative divisions. The LSTM network is optimized using the ELU activation layer to enhance overall network performance. Additionally, GAN networks are used to augment the dataset, further reducing overfitting during the training process. Experimental results show that the algorithm achieves an average accuracy of 99% in address segmentation with an average computation time of 0.1 seconds, meeting the requirements of auxiliary input systems. The algorithm utilizes the segmentation of address information to associate it with the corresponding power supply bureaus, thereby further enhancing the efficiency of electricity service processing. It has great potential for application in the power industry.

## Keywords

Deep Learning, Address Segmentation, Auxiliary Input, LSTM Network

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

随着电力业务的不断发展，在业务办理终端中需要输入的信息日益增加，为提高业务办理效率，需要采用辅助录入系统来进行常用信息的录入。在常用的电力业务信息中，地址信息是使用最为频繁的信息之一，而地址信息的录入与保存通常需要按省、市、区进行分割以用于进行数据的统计与分析。

目前对于地址分割算法已有一定的研究，常用的地址分割方法主要有以下几种：1) 基于规则的方法[1]：基于预先设定的规则，对地址字符串进行分割。例如，可以设定一些关键词(如省、市、区、街道等)作为分隔符，并通过正则表达式进行模式匹配以实现分割。2) 基于字符匹配的方法[2]：通过对地址字符串中的每个字符进行匹配，来确定分割点。例如，可以根据连续出现的数字或字母判断是否需要分割。3) 基于统计学习的方法[3]：利用统计学习算法(如决策树、朴素贝叶斯、支持向量机等)从已有数据集中学习如何对地址字符串进行分割。4) 基于语言模型的方法[4]：利用语言模型对地址字符串进行建模，以此来分割。例如，可以使用 n-gram 语言模型[5]对地址字符串进行建模，并通过计算分割点左右两侧的概率值来确定最佳分割点。

在传统的地址信息分割方法中，主要遇到的问题有：1) 地址格式多样性：由于地理位置的复杂性和多样性，不同地区的地址格式也不尽相同。例如，有些地区的地址可能只包含区域、道路或小区名等简单信息，而有些地区则需要包含门牌号或更为详细的信息。2) 语言表达的不确定性：地址信息通常是通过人工输入或语音输入来记录，因此其表达方式存在各种不确定性，例如人们存在不同的口音、语速和语调等。这些因素可能会导致同一地址信息在不同人员输入时存在较大差异。3) 噪声的存在：在实际使用过程中，由于各种原因(如设备故障、信号干扰等)，可能会出现噪声，使得地址信息存在错误或缺损等情况，从而影响地址分割的效果。4) 计算资源的限制：传统方法往往需要针对不同问题设计特定的模型，此类模型的训练需要巨大的计算资源支撑，从而增加了算法设计的难度和时间成本。综合以上因素，电

力业务辅助录入系统中的地址分割问题具有挑战性，需要借助深度学习等先进技术来提高解决效率和准确性。

## 2. 深度学习介绍

深度学习算法是一种特定的机器学习，它的主要思路是模仿人类大脑的神经网络，通过多层次的神经网络模型来解决复杂的任务。它是一种无监督学习算法，可以从大量数据中自动学习特征，并利用这些特征进行数据分类、对象识别、语音识别、自然语言处理等各种任务。深度学习通过神经网络的层层传递，将输入信息从原始数据中抽象出更高级的特征表示。这些特征表示可以在后续的层次中进一步组合和转换，最终得到最终的输出结果。深度学习已经在各个领域取得了显著的成果，如计算机视觉领域中的人脸识别、物体检测、图像分类，自然语言处理领域中的文本生成、情感分析与机器翻译，以及语音识别和推荐系统等。

经过多年的研究与发展，深度学习模型已经产生了很多类型，常见的深度学习模型主要有：卷积神经网络(Convolutional Neural Network, CNN) [6]，主要应用于计算机视觉领域，通过卷积层与池化层来提取图像的特征，并用于图像分类、目标检测等任务。循环神经网络(Recurrent Neural Network, RNN) [7]，主要用于处理序列数据，能够保留先前的状态信息。它的变体包含有长短时记忆网络(Long Short-Term Memory, LSTM) [8]与门控循环单元(Gated Recurrent Unit, GRU) [9]。生成对抗网络(Generative Adversarial Network, GAN) [10]，由生成器和判别器两个模型组成，通过对抗性训练的方式，生成逼真的样本，如图像生成、文本生成等。自编码器(Autoencoder)，通过将输入的数据压缩到低维的编码描述，再解码还原，用于数据降维、特征提取和去噪等任务。注意力机制(Attention Mechanism) [11]，用于处理序列和图像输入，能够将模型的注意力集中在更重要的部分，提升模型性能。残差网络(Residual Network, ResNet) [12]，通过引入跨层连接，有效解决网络训练过程中梯度爆炸与梯度消失问题，具有很深的网络结构。迁移学习(Transfer Learning) [13]，通过预训练好的模型，在新的任务上进行微调，加快训练速度，并在数据较少的情况下获得更好的性能。在自然语言文本处理(Natural Language Processing, NLP) [14]领域，主要应用的是 RNN 和 LSTM 网络。

## 3. 数据处理

### 3.1. 数据收集

深度学习数据处理通常包含数据的收集、清洗、预处理等步骤。首先第一步也是最重要一步就是数据集的收集。对地址数据可以通过以下几个途径和网站来实现数据的收集：政府机构网站：许多国家和地区的政府机构网站提供了公共地址数据，例如市政府、统计局、邮政服务等。这些网站通常会提供公共服务，包括地理位置信息、行政区划、邮政编码等。地图和导航服务提供商：知名的地图和导航服务提供商，如 Google Maps、百度地图、高德地图等，它们提供了广泛的地理位置数据，包括地址、经纬度、街道信息等。社交媒体和专业网络：一些社交媒体平台或专业网络，如 LinkedIn、Facebook 等，用户在注册时可能需要提供自己的地址信息。这些平台会对用户提供的地址进行处理和验证，可以作为一种收集地址数据的来源。电商平台：一些大型的电子商务平台，如亚马逊、淘宝等，用户在购物时通常需要提供收货地址。这些平台会保存并管理大量的地址数据，可以作为一个潜在的数据来源。开放数据平台：许多国家和地区的政府机构提供了开放数据平台，其中包含各种公共数据集，包括地址数据。通过这些开放数据平台，可以获取到一些公共数据集，如地理位置数据、行政区划数据等。在以上各种数据平台中，地图服务商能提供相对比较完整的地址数据(具体到门牌号)，且其数据均为公开数据，不涉及到隐私或保密数据，符合本文做深度学习模型训练的数据要求。本文以国内某知名地图作为地址数据收集的主

要来源,通过地图 API 对商户、企业、政府机构、金融机构、景点等多个类别进行在全国范围内进行地址数据采集。本文通过地图 API 共采集各种类别地址数据 100,000 多条。

### 3.2. 数据增强

对于深度学习的数据集,想要达到更理想的训练效果,需要进一步增加训练所用的数据。数据增强(Data Augmentation) [15]是一种通过对原始数据进行变换和扩充来增加训练集样本数量和多样性的技术。常见的数据增强方法有:翻转:对图像数据进行水平或垂直翻转,从而得到新的样本。旋转:对图像进行旋转操作,可以是任意角度的旋转,从而生成具有不同角度的新样本。平移:对图像进行平移操作,将图像在水平或垂直方向上移动一定的距离,生成新的样本。缩放:对图像进行缩放操作,可以放大或缩小图像的尺寸,生成不同尺度的新样本。裁剪:对图像进行裁剪操作,从原始图像中截取出不同区域的子图像,生成新的样本。噪声添加:向图像中添加随机噪声,例如高斯噪声、椒盐噪声等,增加数据的鲁棒性。颜色变换:对图像颜色通道进行转换,如改变亮度、对比度、饱和度等参数,生成新的样本。在进行数据增强时,需要注意保持样本的标签或标注信息的一致性,确保增强后的数据仍然具有相同的语义信息。

以上常见的数据增加方法对于图像数据有较好的作用,而对于纯文本数据,以上的数据增加方法则难以实施,因此本文采用生成对抗网络(Generative Adversarial Network, GAN)来做数据增强。GAN 是一种通过博弈的方式训练生成模型的无监督学习框架。它由生成器(Generator)和判别器(Discriminator)两个网络组成,通过相互竞争的方式来提高生成模型的生成能力。GAN 结构如图 1 所示。

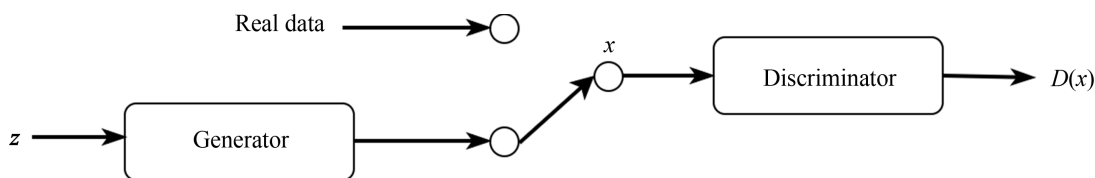


Figure 1. The GAN structure

图 1. GAN 结构

GAN 的基本原理如下:给生成器输入一个随机噪声向量  $z$ ,并将其映射为与真实数据相似的输出  $x$ 。生成器的主要作用是尽可能生成与原数据相近的合成数据。判别器接收真实数据或生成器输出的数据  $x$  作为输入,然后输出  $D(x)$  用于预测输入数据是真实数据还是合成数据。判别器的主要作用是尽可能准确地判断输入数据的真实性。生成器与判别器通过对抗的方式相互竞争、相互学习。在训练过程中,生成器生成合成数据,并将其传递给判别器进行判断。判别器则根据生成器生成的合成数据和真实数据进行判断,并对生成器的输出给出反馈。生成器和判别器交替进行训练,目标是使生成器生成的合成数据逼真度不断提高,同时判别器的判断准确度也不断提高。GAN 的优点在于可以生成高质量的合成数据,能够学习到真实数据分布的潜在特征,并且不需要显式的标签信息。本文采用 GAN 网络生成 200,000 条地址数据,经过数据清理去掉无效数据后行到 170,000 条有效数据。

## 4. 模型设计

### 4.1. LSTM 网络

长短时记忆网络(LSTM)是一种特殊的循环神经网络(RNN),它通过引入门控机制来解决 RNN 中的梯度消失与梯度爆炸的问题。LSTM 通过控制信息的流动,能够在处理长期依赖关系的任务中更好地捕捉和记忆关键信息。LSTM 网络结构如图 2 所示。

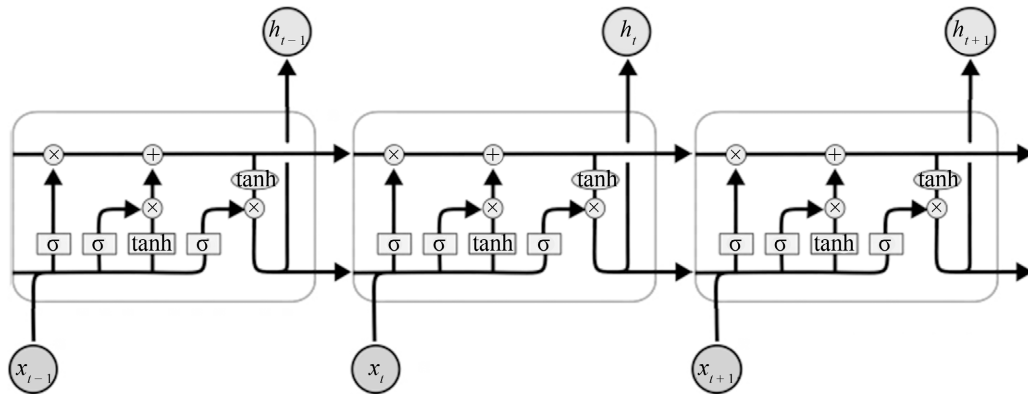


Figure 2. LSTM network structure

图 2. LSTM 网络结构

LSTM 网络由一个或多个 LSTM 单元组成，每个 LSTM 单元内部包含三个不同的门结构：遗忘门、记忆门和输出门。在 LSTM 的单元中，时刻的输入向量和  $t-1$  时刻的状态信息作为输入，通过对内部细胞进行状态单元的补充或移除来更新信息。LSTM 结构的门限由 sigmoid 函数与矩阵点乘组成，遗忘门通过使用 sigmoid 激活函数输出一个 0 到 1 之间的值，来控制对应位置的记忆值的保留程度。值接近 0 表示完全遗忘，值接近 1 表示完全保留。LSTM 网络前向传播可以表示为式(1)~(6)：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

其中， $\tanh$  为双曲正切函数， $\sigma$  为 sigmoid 激活函数； $x_t$  为  $t$  时刻的输入； $f_t$ 、 $i_t$ 、 $o_t$ 、 $C_t$  分别代表  $t$  时刻遗忘门、输入门、输出门信息和细胞状态； $\tilde{C}_t$  表示细胞状态的调制信号； $W_f$ 、 $W_i$ 、 $W_o$ 、 $W_C$  分别代表遗忘门、输入门、输出门信息和细胞状态更新过程中所对应的权值； $b_f$ 、 $b_i$ 、 $b_o$ 、 $b_C$  分别代表遗忘门、输入门、输出门信息和细胞状态更新过程中所对应的偏置。

通过上述过程，LSTM 可以灵活地处理时间序列数据，实现长期依赖关系的建模和记忆。每个 LSTM 单元自身就负责控制信息的保留和遗忘，通过堆叠多个 LSTM 单元，可以构建更复杂的 LSTM 网络，以应对更复杂的任务需求。

## 4.2. 模型优化

LSTM 网络具备较强的表达能力，但也容易在训练样本较少或数据噪声较大的情况下出现梯度消失的问题。本文采用引入 ELU 激活层的方法来解决 LSTM 网络容易梯度消失的问题，以提高预测的准确性。ELU 结合了两种激活函数，即 sigmoid 和 ReLU。激活函数被应用于加权组合的源，或者结果被用作后续阶段的输入。当感知器的突触是一个数学函数时，该组件的输出应始终在 0 和 1 之间波动。在深度学习模型或多层认知系统中，ReLU 是一种非线性活动变量。输入变量  $x$  可以用来描述这个算法。它表明 ReLU 的输出是 0 和传入信号中的最高值。它在左侧具有软饱和度，可以提高对噪声的鲁棒性，在右侧没有饱

和度, 通过在正区间中取输入  $x$  本身来缓解这一问题。梯度分散使其能够缓解梯度消失的问题。与在 Sigmoid 函数上训练模型相比, 使用 ReLU 训练得到的网络会迅速收敛或者所需时间大大减少。使用 ReLU 设计的方法中, 拟合问题明显可见。这是由于快速发散导致的。使用 ReLU 进行模型开发时, 模型的效率明显提高。与 ReLU 相反, 在 ELU 中具有负符号, 允许以比批量归一化更低的计算成本将平均单元激活向零移动。ELU 具有额外的变量, 它必须是一个正数, 这与其他训练算法不同。除负值外, ELU 和 ReLU 非常相似。ELU 超参数控制负净输入的 ELU 饱和值。ELU 公式如式(7)~(8):

$$f(x) = \begin{cases} x, (x > 0) \\ \alpha(e^x - 1), (x \leq 0) \end{cases} \quad (7)$$

$$f'(x) = \begin{cases} 1, (x > 0) \\ f(x) + \alpha, (x \leq 0) \end{cases} \quad (8)$$

与其他激活函数单元相比, ELU 都有较好的改进学习特性, 平均激活接近零可以让学习更快。当输入参数较小时, ELU 减少了前向传播和传输到下一层的信息的变化。本文通过引入 ELU 层代替 sigmoid, 使模型既具有鲁棒性, 又具有低复杂度, 简化了该操作并缓解了梯度消失的问题。改进后的输出门如式(9), 改进后的 LSTM 单元结构如图 3 所示。

$$o_t = \text{elu}(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

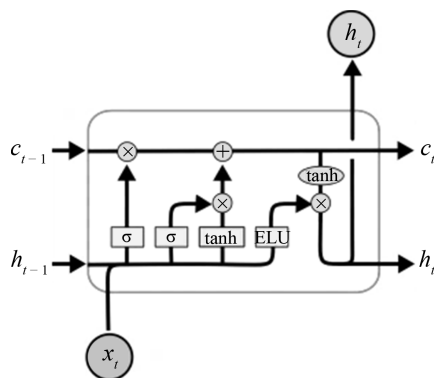


Figure 3. Improved LSTM cell structure  
图 3. 改进的 LSTM 单元结构

本文使用 Pytorch 框架搭建算法开发环境, 通过对原始的 LSTM 与改进的 LSTM 模型分别进行训练, 得到 2 个不同算法的模型。使用原始收集数据与增强后的数据集作为训练样本。训练配置参数如表 1 所示。

Table 1. The training configuration parameters  
表 1. 训练配置参数

| 训练配置        | 具体参数 |
|-------------|------|
| Batch size  | 20   |
| Hidden size | 1500 |
| Num steps   | 40   |

## Continued

|                   |      |
|-------------------|------|
| Init scale        | 0.05 |
| Max grad_norm     | 10   |
| Epoch start_decay | 16   |
| Max epoch         | 60   |
| Lr decay          | 0.65 |
| Dropout           | 1.1  |

在训练数据集前需要分别搭建 2 个网络模型为改进前后的 LSTM 网络。同时，为减少不同量纲和数量级产生的影响，需将数据集进行 0~1 标准化处理。本文选用以下几种模型评价指标来对辨识效果进行评估，包括 ACC (精确度)、MAPE (平均相对百分比误差)、RMSE (均方根误差)及  $R^2$  (拟合优度)等。其中，MAPE 和 RMSE 的值越小，说明模型建立的效果越好； $R^2$  衡量的是回归方程整体的拟合度， $R^2$  的值越接近 1，说明模型的拟合效果越好。根据仿真结果可以得到改进前后的 LSTM 网络评价指标见表 2。

**Table 2.** Comparison of evaluation indicators before and after LSTM improved

**表 2.** 改进 LSTM 前后评价指标的对比

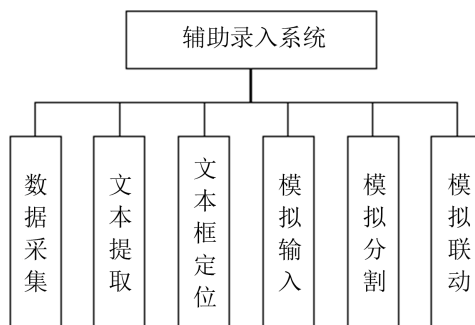
| 检测模型     |     | ACC    | MAPE   | RMSE   | $R^2$  |
|----------|-----|--------|--------|--------|--------|
| LSTM     | 训练集 | 0.9873 | 0.0126 | 0.2658 | 1.2690 |
|          | 测试集 | 0.9869 | 0.0118 | 0.2569 | 1.1256 |
|          | 验证集 | 0.9865 | 0.0124 | 0.2954 | 0.6952 |
| 改进的 LSTM | 训练集 | 0.9955 | 0.0035 | 0.2116 | 1.1243 |
|          | 测试集 | 0.9964 | 0.0034 | 0.2220 | 1.0226 |
|          | 验证集 | 0.9897 | 0.0052 | 0.2305 | 0.8932 |

从表 2 可看出，采用 ELU 改进的 LSTM 网络总体性能比原 LSTM 网络有明显的优化效果。

## 5. 系统设计与应用

### 5.1. 系统架构

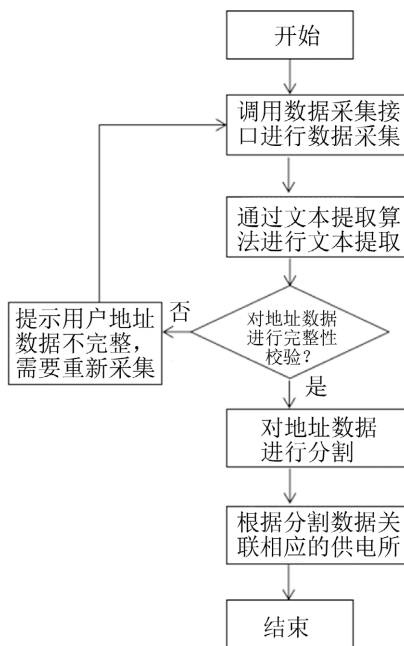
在电力业务终端中，采用辅助录入系统可提升信息录入的准确性与效率。辅助录入系统可以通过格式识别和校验、数据提取和转换、快捷输入、自动保存和恢复、数据智能验证等功能，辅助用户进行信息录入。地址数据是最为常用的数据之一，地址数据在录入后需要按一定格式来保存，通常会按不同等级的行政区对地址数据进行分割，以方便进行数据的统计与分析。为提高录入效率，辅助录入系统需要结合多项技术来完成开发，主要的功能模块包含有：数据采集、文本提取、文本定位、模拟输入、数据分割、数据联动六大模块。其中，数据联动模块可用分割后的地址数据与对应的供电所进行关联，数据录入完成后自动调取供电所信息，进一步提升了业务办理速度。辅助录入系统架构如图 4 所示。



**Figure 4.** The assisted entry system architecture  
**图 4.** 辅助录入系统架构

## 5.2. 业务流程设计

辅助录入系统业务流程设计主要原则主要包括有：**提升用户体验**：在数据录入上提供多种录入方式，如手动录入、拍照录入、语音录入等；**数据校验**：在用户输入了基础数据后，系统进行数据校验与清理，保障所录入数据的合格与完整性；**系统集成与接口**：在系统中提供对不同录入方式的接口兼容，方便日后扩展更多的录入方式；**错误处理和异常情况**：对于硬件或软件产生的错误，录入系统自动进行基本故障处理，自动恢复系统功能，对于无法恢复的故障给用户明确提示；**性能和扩展性**：在业务流程中，要简化用户操作，从用户交互上提升整体的录入效率，同时对软件算法进行优化以提升后台的处理速度。**数据联动**：充分体现辅助录入系统的智能化，自动根据地址信息调取所属供电所信息，提高系统业务输效率。辅助录入系统算法流程如图 5 所示。



**Figure 5.** Process of the assist entry system  
**图 5.** 辅助录入系统流程

## 5.3. 性能分析

本文辅助录入系统基于安卓 11 版本进行开发，通过在 Android Studio 上的模拟器进行仿真环境运行



测试, 测试数据采用国内某知名地图服务商提供的地址数据, 分别进行不同省市自治区的数据测试, 以评价算法的运算性能。本次测试采用四段分割法, 分别分割省级、市级、区(县)级、详细地址信息, 每级地址采用逗号进行分隔, 系统根据分割后的地址数据自动关联对应的供电局(所)信息, 并检测算法的运算时间, 测试例子如图 6, 测试结果如表 3 所示。

```
E:\Project\YinZhang\venv\Scripts\java.exe
广东省,佛山市,顺德区, 校林路与衡山路交叉口北100米, 佛山供电局大良供电所, t=0.09
广东省, 茂名市, 茂南区, 官山四路2号, 广东电网公司茂名供电局, t=0.09
广西壮族自治区, 北海市, 海城区, 北部湾东路27号, 广西电网北海供电局, t=0.10
云南省, 普洱市, 西盟佤族自治县, 勐卡路485号, 西盟供电局, t=0.11
重庆市, 九龙坡区, 红狮西路5号, 重庆杨家坪供电局, t=0.09
内蒙古自治区, 呼和浩特市, 新城区, 爱民街27号, 呼和浩特供电局新城分局, t=0.12
新疆维吾尔自治区, 乌鲁木齐市, 天山区, 延安路139号, 乌鲁木齐市西郊供电局, t=0.12
```

Figure 6. Test example

图 6. 测试例子

Table 3. Analysis of address data test results

表 3. 地址数据测试结果分析

| 行政区类别         | 数据量 | 平均准确率(%) | 平均运算时间(s) |
|---------------|-----|----------|-----------|
| 省 - 市 - 区     | 280 | 100      | 0.09      |
| 省 - 市 - 县     | 300 | 100      | 0.09      |
| 省 - 自治州 - 县   | 180 | 98.9     | 0.10      |
| 直辖市 - 区       | 120 | 99.1     | 0.12      |
| 自治区 - 市 - 县   | 160 | 98.7     | 0.11      |
| 自治区 - 自治州 - 区 | 90  | 98.9     | 0.12      |

测试数据选取比例根据国内不同省、市、自治区进行选择, 以满足数据的多样性。从测试结果可知, 算法的平均准确率达到 99% 以上, 平均运算时间为 0.1 秒, 算法性能取得较好的效果, 满足电力业务辅助录入系统的需求。

## 6. 结束语

本文主要针对电力业务辅助录入系统中对地址数据特定格式化的需求, 提出一种基于深度学习的地址分割算法, 对地址信息按行政区级进行分割。算法采用 ELU 激活层对 LSTM 网络进行优化以提高网络的整体性能, 另外采用 GAN 网络对数据集进行了增强, 有效解决了训练过程中发生过拟合的问题。通过实验数据测试可知, 算法对于地址数据的分割平均准确率达到 99%, 平均运算时间为 0.1 秒, 且利用分割地址信息关联对应的供电局(所), 进一步提升了电力业务的办理效率。算法应用于电力业务辅助录入系统中可有效提高录入效率, 具有较好的应用前景。

## 基金项目

本文由“南网高层次人才特殊支持计划”项目资助。

## 参考文献

- [1] Sarawagi, S. and Cohen, W. (2004) Semi-Markov Conditional Random Fields for Information Extraction. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Banff, 7-11 July 2004, 344-351.
- [2] Navarro, G. (2000) A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, **33**, 31-88. <https://doi.org/10.1145/375360.375365>
- [3] Kipf, T.N. and Welling, M. (2016) Semi-Supervised Classification with Graph Convolutional Networks. arXiv preprint arXiv:1609.02907. <https://doi.org/10.48550/arXiv.1609.02907>
- [4] Lafferty, J., McCallum, A. and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Berkshires, Massachusetts, 28 June-1 July 2001, 282-289.
- [5] 何晓明, 洪亲, 蔡坚勇, 等. 基于 n-gram 中英文字符串分割算法实现[J]. 电脑知识与技术, 2012, 8(23): 5530-5533. <https://doi.org/10.3969/j.issn.1009-3044.2012.23.009>
- [6] Yin, W., Kann, K., Yu, M., et al. (2017) Comparative Study of CNN and RNN for Natural Language Processing. arXiv preprint arXiv: 1702.01923. <https://doi.org/10.48550/arXiv.1702.01923>
- [7] Tarwani, K.M. and Edem, S. (2017) Survey on Recurrent Neural Network in Natural Language Processing. *International Journal of Engineering Trends and Technology (IJETT)*, **48**, 301-304. <https://doi.org/10.14445/22315381/IJETT-V48P253>
- [8] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [9] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. arXiv preprint arXiv: 1406.1078. <https://doi.org/10.48550/arXiv.1406.1078>
- [10] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. and Bengio, Y. (2014) Generative Adversarial Nets. *Advances in neural information processing systems (NIPS)*, Montreal, 8-13 December 2014, 2672-2680.
- [11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. and Polosukhin, I. (2017) Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, 4-9 December 2017, 5998-6008.
- [12] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 27-30 June 2016, 770-778.
- [13] Pan, S.J. and Yang, Q. (2009) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. <https://doi.org/10.1109/TKDE.2009.191>
- [14] Chowdhury, G.G. (2003) Natural Language Processing. *Annual Review of Information Science and Technology (ARIST)*, **37**, 51-89. <https://doi.org/10.1002/aris.1440370103>
- [15] Simard, P.Y., Steinkraus, D. and Platt, J.C. (2003) Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. *7th International Conference on Document Analysis and Recognition (ICDAR 2003)*, Edinburgh, 6 August 2003, 958-962. <https://doi.org/10.1109/ICDAR.2003.1227801>