

手写化学方程式配平程序

NJU CS2017 徐国栋

一、 目录

- 简介和开发历程
- 相关实现的介绍

二、 简介和开发历程

手写化学方程式配平程序的功能是对从电子画板手写输入的完整化学方程式进行实时识别和最终的配平。

以 Qt 作为开发框架，利用 Qt 构建图形界面、进行像素处理。配平和识别部分以 C++ 作为唯一编程语言。程序不依赖任何外存文件，模型和字典利用代码直接在内存中初始化。利用 Qt 的跨平台特性编译出桌面应用程序和安卓应用程序。在 VS2017-release 测试中，1080×720 分辨率的画板界面，实时最高内存占用为 38MB。

我从 2018 年 10 月开始构思并逐步完成相关代码，10 月制作出包含三百多个化学方程式的文本文件，11 月完成了基于元素守恒的配平程序，12 月加入了得失电子守恒的配平判据，2019 年 1 月完成了基于笔画走向的手写识别程序，2 月利用开源 C++ 项目 MojoCNN 完成了基于深度学习模型的手写识别程序，3 月初制作了一个包含三千多个英文单词和化学式的字典来优化深度学习模型的识别结果，程序具备可用性。

三、 相关实现的介绍

1、 配平

自底向上构建元素、物质、方程类，对以字符串形式输入的化学方程式以一定规则进行解析，这里没有用正则表达式，而是以括号为最小单元进行循环解析直到解析完成，解析过程中大小写敏感。

配平的数学本质是解线性方程组，利用元素守恒和得失电子守恒列出该方程组并作高斯消元，有唯一解的方程组被认为是合法的。

元素守恒判据直接从方程类获得，解析完字符串的方程类存储了该方程的所有原始信息。

得失电子守恒判据从化合价表中获得，当且仅当元素守恒判据不足以列出合法线性方程组时，寻找得失电子守恒判据。通过查表的方式获得化合价信息，预存了三百多个物质的化合价信息。得失电子守恒判据遵循同种元素化合价升降“只靠拢不交叉”的规则。

2、 CNN

采用 MojoCNN 作为深度学习的 C++ 代码，在一个包含三十八万张 28×28 像素的英文字母与数字的训练集上达到 85% 的测试准确率（该模型在 MNIST 上的测试准确率可以超过 99.6%）。

训练模型的配置为：

5×5 卷积层步长 1 深度 8，3×3 池化层步长 3，

1×1 卷积层步长 1 深度 16，

5×5 卷积层步长 1 深度 8, 2×2 池化层步长 2,
softmax = 62。

模型固化的方法是把模型数据输出为 c++ 代码的形式, 例如把所有权重输出为 `vector<float>`, 程序启动时直接从内存初始化模型。该程序使用的模型占用 300KB 左右的内存空间。

3、字典

为了区分诸如 '1' 'l' 'I' 等字符, 使用字典优化深度学习模型的识别结果, 即从目标像素矩阵向前传播后得到的 62 个字符权重中, 选出较高的多个字符, 结合前缀从字典中择优选择。

常规来讲, 前缀树是一种时间复杂度 $O(1)$ 的高效字典结构, 但是在字符种类较多的情况下, 空间复杂度不可承受。

本程序借鉴前缀树的思想, 对一个单词, 前三个字符进行哈希结点存储, 为此开辟了 $64 \times 64 \times 64$ 的数组空间, 后面更多的字符采用字符串链表集中存放。在近四千个单词的情况下, 占用内存 10MB。在时间复杂度上, 所有单词的平均查找次数小于 10, 在化学式查找上, 平均查找次数小于 3。

4、画板

画板的功能独立于识别算法, 画板顺序提供所有输入点的坐标信息, 字符识别的时机是抬笔/鼠标左键松开。画板对加号和等号的判别是基于笔画边框的。

目前画板提供的纠错机制是：

双击画板撤销上一次输入、单击[复位]归零画板状态。