Распознавание текста на основе скелетного представления толстых линий и свёрточных сетей*

Мурзин Д. А., Местецкий Л. М., Рейер И. А., Стрижов В. В.
murzin.da@phystech.edu; mestlm@mail.ru; reyer@forecsys.ru;
strijov@phystech.edu
Московский физико-технический институт

В работе рассматривается задача распознавания текста на изображении путём преобразования его в медиальное представление с последующим применением свёрточной нейронной сети для задачи классификации. Данный способ имеет ряд преимуществ по сравнению с классическими дискретными способами распознавания текста. В работе предлагается способ повышения качества распознавания толстых линий за счёт нового способа порождения их описаний. В качестве тестовых данных используются шрифты в растровом представлении.

Ключевые слова: *распознавание текста, непрерывное медиальное представление, свёрточные нейронные сети.*

Введение

Работа посвящена задаче распознавания символов на изображении. Это задача имеет множество применений, от оцифровки старых книг до распознавания рукописного текста.

Существующие методы распознавания текста можно разбить на две группы: «дискретные» и «непрерывные». Дискретные алгоритмы работают с изображением в первоначальном виде, то есть в виде матрицы пикселей. Такой способ обработки изображений близок компьютерам, но не людям, так как мы привыкли различать фигуры и образы, которые являются непрерывными объектами.

С другой стороны, непрерывные алгоритмы построены на использовании таких интуитивных для человека понятий как фигура и форма. Непрерывные алгоритмы устроены примерно следующим образом. Сначала строится непрерывное описание исходного изображения. Это может быть описание границы в виде кривых, либо медиальное представление, то есть набор кривых (скелет) и радиальная функция, которая каждой точке кривой сопоставляет максимальный радиус окружности, лежащей внутри фигуры, с центром в этой точке.

В работе предлагается алгоритм распознавания текста, в котором сначала строится медиальное представление для изображения, с последующим применением свёрточной нейронной сети. Эта сеть состоит из последовательных операций свёртки и уплотнения. В операции свёртки по отдельности рассматривается каждая небольшая часть описания изображения и в ней выделяются характерные паттерны в этой части. Операции уплотнения состоит в уменьшении числа признаков путём замены нескольких частей описания изображения на одну часть, аккумулирующую информацию о найденных паттернах.

Постановка задачи

В работе решается задача распознования рукописных символов на изображении. Рассматриваются два варианта постановки задачи, «дискретный» и «непрерывный», которые

Научный руководитель: Стрижов В. В. Консультант: Местецкий Л. М.

отличаются форматом исходных изображений. Опишем постановку задачи, после чего определим форматы изображений для обоих вариантов.

Пусть задано множество символов $\mathcal{S} = \{s_1, \ldots, s_k\}$ и выборка изображений:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, m\}$$

где $\mathbf{x}_i$ является объектом, описывающим i-ое изображение, а $y_i \in \mathcal{S}$ — символом, на нём изображённым.

Требуется построить алгоритм f, решающий задачу классификации изображений, то есть, принимающий описание изображения в том же формате как в исходной выборке и возвращающий список вероятностей $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_k\}$:

$$f : \mathbf{x} \mapsto (\hat{p_1}, \ldots, \hat{p_k})$$

где $\hat{p}_i$ — предсказание вероятности того что на изображение находится символ s_i, $\forall i\ \hat{p}_i \in [0, 1]$, $\hat{p}_1 + \ldots + \hat{p}_k = 1$. По списку вероятностей можно будет получить предсказание символа на изображении взяв символ с наибольшей вероятностью.

Делаются следующие предположения о выборке:

— Каждое изображение содержит ровно один символ, написанный от руки.
— Каждый символ на изображении полностью содержится в изображении, причём расстояние между символом и границами изображения строго больше нуля
— Каждый символ из множества символов $\mathcal{S}$ встречается достаточно большое число раз в выборке, то есть не существует пар символов $c_1, c_2 \in \mathcal{S}$, таких что символ c_2 встречается много больше раз чем символ c_1. В идеале равномерное распределение на символах (каждый символ встречается равное число раз).

В качестве функции ошибки для оценки качества алгоритма будем использовать перекрёстную энтропию:

$$H(p, \hat{p}) = -\sum_{i=1}^{k} p \log \hat{p}$$

где p — истинный вектор вероятностей (все нули кроме одного элемента), $\hat{p}$ — предсказание вероятностей.

Перейдём к описанию форматов изображений для обоих вариантов постановок.

Постановка задачи (дискретный случай)

Сначала введём определение множества цветов:

Определение 1. $\mathcal{C}$ — множество цветов, которые может принимать один пиксель изображения. В работе всегда предполагается $\mathcal{C} = \{0, 1\}$, где ноль соответствует белому цвету, а 1 чёрному. Другими возможными вариантами могут быть $\mathcal{C} = \{0, 1, \ldots, 255\}$ — оттенки серого и $\mathcal{C} = \{0, 1, \ldots, 255\}^3$ — цветовое пространство RGB.

В данной постановке описание изображения $\mathbf{x}$ представляет собой матрицу из h строк и w столбцов: $\mathbf{x}_i = [c_{ij}] \in \mathcal{C}^{h \times w}$. Каждый элемент матрицы описывает цвет одного пикселя изображения. Ответ $y_i \in \mathcal{S}$ — символ, находящийся на изображении $\mathbf{x}_i$.

В работе предлагается использовать базу данных рукописных изображений MNIST [1]. В ней каждое изображение имеет размер 28×28, а цвета пикселей кодируются числами от 0 до 255 (оттенки серого, 0 — белый, 255 — чёрный).

Постановка задачи (непрерывный случай)

В данной постановке описанием изображения является скелетное представление с заданной на нём радиальной функцией. Введём необходимые определения, в соответствии с [2]:

Определение 2. Фигура — связная область на плоскости $\mathbb{R}^2$, такая что её граница представляет собой дизъюнктное объединение конечного числа отрезков.

Определение 3. Пустой круг фигуры — круг, полностью содержащийся внутри фигуры.

Определение 4. Максимальный пустой круг фигуры — пустой круг, который не содержится ни в каком другом пустом круге этой фигуры.

Определение 5. Скелет фигуры — связный граф на плоскости, такой что каждая точка каждого ребра графа является центром максимального пустого круга.

Определение 6. Радиальная функция для скелетного представления — функция, которая каждой точке скелетного представления сопоставляет радиус максимального круга с центром в этой точке.

Определение 7. Медиальное представление фигуры — скелет фигуры с соответствующей медиальной функцией.

В работе предлагается использовать выборку, в которой медиальное представление имеет следующий вид: скелет задан в виде графа, радиальная функция задана на каждой вершине этого графа, а значение радиальной функции на рёбрах определяется как взвешенное среднее радиальной функции на концах ребра.

Также, дополнительно, каждая вершина имеет степень от одного до трёх.

Базовый алгоритм

В качестве базового алгоритма используется свёрточная нейронная сеть для задачи в дискретной постановке. Предлагается использовать следующую структуру сети:

$$INPUT -> [[CONV -> RELU] * 2 -> POOL] * 3 -> FC$$

- INPUT — входной слой, имеет размеры $28 \times 28 \times 1$
- CONV — слой свёртки. Фильтры имеют размер 3×3. Также используется увеличение пространственных размеров на 2 в каждой размерности предыдущего слоя путём дополнения одинарной линией из нулей с каждой стороны.
- RELU — слой активации. Используется функция $f(x) = \max(0, x)$
- POOL — слой пулинга. Каждая группа пикселей 2×2 уплотняется в один пиксель, путём взятия максимума.
- FC — полносвязный слой.

Обучение сети будет осуществляться методом обратного распространения ошибки.

Литература

[1] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits, 1998. http://yann.lecun.com/exdb/mnist/.

[2] Леонид Моисеевич Местецкий. *Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры.* Физматлит, 2009.

[3] Солдатова Ольга Петровна and Гаршин Александр Александрович. Применение сверточной нейронной сети для распознавания рукописных цифр. 2010.

[4] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. 2003.

[5] Dan Claudiu Cires, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. Convolutional neural network committees for handwritten character classification. 2011.

[6] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. Emnist: an extension of mnist to handwritten letters, 2017. https://www.nist.gov/itl/iad/image-group/emnist-dataset.

[7] Aleksey Morozov. Low data drug discovery with one-shot learning. 2017.

[8] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. 2016.

[9] Визильтер Ю.В., Горбацевич В.С., and Желтов С.Ю. Структурно-функциональный анализ и синтез глубоких конволюционных нейронных сетей. 2018.