Классификация символов на основе медиального представления и свёрточных сетей*

Мурзин Д. А., Данилов А. Н., Местецкий Л. М., Рейер И. А., Жариков И. Н., Стрижов В. В.

murzin.da@phystech.edu; andnlv@gmail.com; mestlm@mail.ru; reyer@forecsys.ru; zharikov.i.n@yandex.ru; strijov@phystech.edu

Московский физико-технический институт

В работе рассматривается задача распознавания символов на изображении. Предлагается новый способ построения свёрточной нейронной сети, использующей в качестве входа непрерывное представление цифрового изображения текстовых символов. В качестве тестовых данных используются символы латинского алфавита и цифры в растровом представлении.

Ключевые слова: *классификация символов, непрерывное медиальное представление, свёрточные нейронные сети.*

Введение

Работа посвящена задаче распознавания символов на изображении. Она используется для распознавания текста после сегментации на символы, что 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. Данный формат описания изображения позволяет компьютерам эффективно их обрабатывать. Однако данный формат не является родным для людей, что приводит к сложности придумывания «дискретных» алгоритмов.

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.

Дискретным алгоритмам распознавания символов посвящено большое количество работ. В частности в многих работах рассматривается классификации символов на базе данных рукописных цифр MNIST. Наилучшую точность показали алгоритмы использующие свёрточные нейронные сети [1, 2, 3, 4]. Так в работе [2] предлагается алгоритм классификации, использующий ансамбль из 35 свёрточных сетей, дающий точность 99.77%.

Непрерывным алгоритмам посвящено меньшее число работ. В книге [5] вводятся основные понятия связанные с непрерывным описанием изображения и предлагается алгоритм построения этого описания. В работах [6, 7, 8] рассматриваются подходы к построения классификатора на непрерывных описаниях изображения.

В работе предлагается развить подходы [5, 6]. По дискретному описанию изображения строится непрерывное описание, с помощью алгоритма предложенного в [5]. Это непре-

---

Научный руководитель: Стрижов В. В. Консультанты: Местецкий Л. М, Жариков И. Н.

рывное описание представляет собой граф специального вида. Данное описание преобразуется путём генерации новых признаков для каждой вершины графа. Полученный граф передаётся в классификатор на основе свёрточной нейронной сети, предложенный в [6].

Постановка задачи

В работе решается задача распознавания печатных символов на изображении. Требуется построить классификатор, принимающий описание изображения и возвращающий класс символа, изображённого на изображении. Описание изображения состоит из пары — «дискретного» и «непрерывного» описаний. «Дискретное» описание представляет матрицу пикселей цветов. Непрерывное описание представляет собой граф специального вида. Введём строгие определения.

Определения для «дискретного» описания.

Определение 1. Дискретное изображения высоты h и ширины w — бинарная матрица из h строк и w столбцов. Каждый элемент матрицы описывает цвет одного пикселя изображения: ноль — черный, один — белый.

Определения для «непрерывного» описания.

Определение 2. Жорданова кривая — образ окружности при непрерывном инъективном отображении окружности в плоскость.

Определение 3. Фигура — замкнутая область на плоскости $\mathbb{R}^2$, ограниченная конечным числом непересекающихся жордановых кривых.

Определение 4. Пустой круг фигуры — круг, полностью содержащийся внутри фигуры.

Определение 5. Максимальный пустой круг фигуры — пустой круг, который не содержится ни в каком другом пустом круге этой фигуры.

Определение 6. Скелет фигуры — множество всех центров максимальных пустых кругов фигуры.

Определение 7. Радиальная функция для скелета фигуры — функция, которая каждой точке скелета сопоставляет радиус максимального круга с центром в этой точке.

Определение 8. Медиальное представление фигуры — скелет фигуры с соответствующей медиальной функцией.

Перейдём к постановке задачи. Пусть задано множество символов $\mathcal{Y}$ мощности k и выборка изображений, содержащих символы:

$$\mathfrak{D} = \{(I_i, y_i)\}_{i=1}^{m}$$

— I_i — дискретное описание изображения
— $y_i \in \mathcal{Y}$ — класс символа

Требуется построить классификатор f, решающий задачу классификации символов, то есть, принимающий описание изображения в том же формате как в исходной выборке и возвращающий вектор вероятностей $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_k\}$, где $\hat{p}_i \in [0, 1]$, причём $\hat{p}_1 + \ldots + \hat{p}_k = 1$:

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

где $\hat{p}_i$ — предсказание вероятности того что на изображение находится символ s_i.

Классификатор f является композицией трёх алгоритмов:

— $\mu : I \mapsto G$ — алгоритм построения медиального представления. Используются библиотека скелетонизации Никиты Ломова и скрипты для запуска Анны Липкиной [добавить ссылку].
— $g : G \mapsto F$ — алгоритм генерации признаков по медиальному представлению.
— $h : F \mapsto \hat{p}$ — классификатор на основе свёрточных сетей для графов. Используется библиотека DeepChem [добавить ссылку].

Информация о выборке.

В работе предлагается использовать изображения, полученные с помощью генератора символов латинского алфавита и цифр [добавить цитату]. Каждое изображение имеет размер 64×64, а цвета пикселей кодируются числами 0 и 1 (чёрный и белый соответственно). При этом расстояние от символа до границ изображения строго больше нуля.

Функция ошибки.

В качестве функции ошибки для оценки качества классификатора используется кросс-энтропия (cross entropy):

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

где p — истинный вектор вероятностей (все нули кроме одного элемента), $\hat{p}$ — предсказание вероятностей. [Тут не понятно с обозначениями]

Теоретическая часть

Алгоритм скелетонизации μ выдаёт медиальное представление следующего вида: скелет задан в виде плоского графа (вершины — точки плоскости, рёбра — отрезки), радиальная функция задана на каждой вершине этого графа, а значение радиальной функции на рёбрах определяется как взвешенное среднее радиальной функции на концах ребра.

Формально, получаемое медиальное представление является парой:

$$G = \{X, E\}$$

где

— $X = \{\mathbf{x}_u \mid u \in \{1, \ldots, n\}\}$ — описание вершин графа
  — n — число вершин графа
  — $\mathbf{x}_u$ — описание вершины u графа, $\mathbf{x}_u = (x_u, y_u, \rho_u)$
    — x_u, y_u — координаты вершины
    — ρ_u — значение радиальной функции в вершине
— $E = \{(u, v)\}$ — рёбра графа

Дополнительно, каждая вершина имеет степень от одного до трёх.

Алгоритм генерации признаков g преобразует заданное в виде графа медиальное представление, а именно к базовым признакам (x_u, y_u, ρ_u) каждой вершины u добавляется вектор признаков $f = (f_1, \ldots, f_k)$. Формально, получается пара

$$G = \{F, E\}$$

где

— $F = \{\mathbf{f}_u \mid u \in \{1, \ldots, n\}\}$ — признаки вершин графа
  — n — число вершин графа
  — $\mathbf{f}_u$ — признаки вершины u графа, $\mathbf{f}_u = (f_1, \ldots, f_k)$. Первые три признака совпадают с базовыми — (x_u, y_u, ρ_u).
— $E = \{(u, v)\}$ — рёбра графа

Опишем каждый признак, способ его генерации и мотивацию для него.

Степень вершины. Получается непосредственно по графу. Интерес представляет соотношение между количеством вершин степени три и один, а также количество вершин степени четыре. В зависимости от наличия вершин степени три и более символы разбиваются на два класса размера 32 и 31. Вершину степени четыре содержат 8 символов.

Число циклов графа, в котором содержится вершина. Получается предварительным выделением всех циклов графа с помощью поиска в глубину. Предикат наличия циклов в графе разбивает символы на два класса размера 18 и 44.

Средняя площадь циклов, в которых содержится данная вершина. Позволяет разделить символы с большим циклом (0, D, Q) и с маленьким (4, a, p).

Минимальный угол между рёбрами, исходящими из вершины. Полезен для вершин степени два, так как позволяет отделить символы похожие на прямые (i, l) от символов с изгибами (w, z)

Расстояние до ближайшей вершины степени один.

Сумма углов поворота между рёбрами на пути к ближайшей вершине степени один. Позволяет отделить немного изогнутые символы (j, f, 9) от символов с прямыми линиями (k, w, 4)

Длина максимальной прямой линии, в которой содержится текущая вершина. Прямая линия — путь в графе, такой что угол между каждой парой соседних рёбер отличается не более чем на 10° от 180°. Позволяет отделить символы с длинными линиями (p, q, F, M).

Угол между максимальной прямой линией и горизонталью.

Алгоритм классификации h — это свёрточная нейронная сеть для графов, которая использует три базовые операции: свёртки, активации и пулинга.

`images/cnn-graph.png`

[не вставляется картинка]

Операция свёртки производится независимо для каждой вершины u графа. Обозначим $d = \text{degree}(v)$ — степень вершины v. Рассмотрим вершину v, смежную с ней. Обозначим $l = \text{distance}(u, v)$ — евклидово расстояние между вершинами u и v. Вершине v соответствуют признаки $f_1, \ldots, f_{k_1}$. Операция свёртки состоит из двух этапов:

— для каждой смежной с u вершины v строится промежуточные вектора признаков, на основе степени вершины u и расстояния l
— промежуточные вектора всех смежных с u вершин складываются, в результате получается новый вектор признаков вершины u

Для построения промежуточных векторов используются матрицы $W_{d,l} \in \mathbb{R}^{k_1 \times k_2}$ (k_1 — число признаков до операции свёртки, k_2 — после) и вектор $b_{d,l} \in \mathbb{R}^{k_2}$. Расстоянию l сопоставляется матрица W и вектор b путём дискретизации вещественного значения расстояния l, а именно рассматриваются положительные вещественные числа $0 = b_1 < b_2 < \ldots$, разбивающие положительную числовую прямую на классы $\mathbb{R}_+ = [b_1, b_2) \bigcup [b_2, b_3) \bigcup \ldots$.

Тогда в результате операции свёртки получаются новые признаки для вершины u:

$$h_{\text{conv}}(u) = \sum_{(u,v)\in E} \left(W_{d,l} f_v + b_{d,l}\right)$$

Операция активации также производится независимо для каждой вершины u графа. Рассматриваются вектор признаков вершины u, вектора признаков всех смежных с u вершин. К данным векторам применяется операция максимума и получается новый вектор признаков для вершины u:

$$h_{\text{relu}}(u) = \max(f_u, \max_{(u,v)\in E} f_v)$$

Операция пулинга применяется к группе вершин $(u_1, \ldots, u_k)$ графа. Данные вершины заменяются одной, с вектором признаков, равным сумме векторов признаков исходных вершин:

$$h_{\text{pool}}(v_1, \ldots, v_k) = \sum_{i=1}^{k} (f_{v_i})$$

Базовый вычислительный эксперимент

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

[Структура сети будет уточняться по ходу эксперимента]

$$\text{INPUT} \to [[\text{CONV} \to \text{RELU}] \times 2 \to \text{POOL}] \times 2 \to \text{FC}$$

— 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 $f(x) = \ln(1 + \exp(x))$
— POOL — слой пулинга. Каждая группа пикселей 2×2 уплотняется в один пиксель, путём взятия максимума.
— FC — полносвязный слой.

Вычислительный эксперимент

...

Литература

[1] Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. Regularization of neural networks using dropconnect. 2013. https://cs.nyu.edu/ wanli/dropc/.

[2] Dan Ciresan, Ueli Meier, and Jurgen Schmidhuber. Multi-column deep neural networks for image classification. 2012.

[3] Ikuro Sato, Hiroki Nishimura, and Kensuke Yokoi. Apac: Augmented pattern classification with neural networks. 2015.

[4] Jia-Ren Chang and Yong-Sheng Chen. Batch-normalized maxout network in network. 2015.

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

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

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

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

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

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

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

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

[13] 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.