Классификация символов на основе медиального представления и свёрточных сетей*

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

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

Московский физико-технический институт

В работе рассматривается задача распознавания символов на изображении. Предлагается новый способ построения свёрточной нейронной сети, использующей в качестве входа медиальное 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. 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 распознавания рукописных символов на изображении. Требуется построить классификатор, принимающий описание изображения и возвращающий класс символа, изображённого на изображении. Описание изображения состоит из пары —

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

«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. $\mathcal{C}$ — множество цветов, которые может принимать один пиксель изображения.

В работе всегда предполагается $\mathcal{C} = \{0, 1\}$, где ноль соответствует белому цвету, а 1 чёрному.

Определение 2. Дискретное изображения высоты h и ширины w — матрица из h строк и w столбцов: $I = [c_{ij}] \in \mathcal{C}^{h \times w}$. Каждый элемент матрицы описывает цвет одного пикселя изображения.

Определения для «непрерывного» описания.

Определение 3. Жорданова кривая — образ окружности при непрерывном инъективном отображении окружности в плоскость.

Определение 4. Фигура — замкнутая область на плоскости $\mathbb{R}^2$, ограниченная конечным числом непересекающихся жордановых кривых.

Определение 5. Пустой круг фигуры — круг, полностью содержащийся внутри фигуры.

Определение 6. Максимальный пустой круг фигуры — пустой круг, который не содержится ни в каком другом пустом круге этой фигуры.

Определение 7. Скелет фигуры — множество всех центров максимальных пустых кругов фигуры.

Определение 8. Радиальная функция для скелета фигуры — функция, которая каждой точке скелета сопоставляет радиус максимального круга с центром в этой точке.

Определение 9. Медиальное представление фигуры — скелет фигуры с соответствующей медиальной функцией.

Перейдём к постановке задачи. Пусть задано множество символов $\mathcal{Y}$ и выборка изображений:

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

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

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

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

где $\hat{p}_i$ — предсказание вероятности того что на изображение находится символ s_i, $\hat{p}_i \in [0, 1]$, $\hat{p}_1 + \ldots + \hat{p}_k = 1$.

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.
— $h : F \mapsto y$ — классификатор на основе свёрточных сетей для графов. Используется библиотека DeepChem [добавить ссылку].

Формальное описание алгоритмов

Алгоритм скелетонизации h 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, $\mathbf{x}_u = (x_u, y_u, radial_u)$
    — x_u, y_u — координаты вершины
    — $radial_u$ — значение радиальной функции в вершине
— $E = \{(u, v)\}$ — рёбра графа

Дополнительно, каждая вершина имеет степень от одного до трёх.

Алгоритм генерации признаков f преобразует заданное в виде графа медиальное представление, а именно описание $(x_u, y_u, radial_u)$ каждой вершины u заменяется на 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, $\mathbf{f}_u = (f_1, \ldots, f_k)$
— $E = \{(u, v)\}$ — рёбра графа

Информация о выборке.

В работе предлагается использовать изображения, полученные с помощью генератора символов латинского алфавита и цифр [добавить цитату]. Каждое изображение имеет размер 32×32, а цвета пикселей кодируются числами от 0 до 255 (оттенки серого, 0 — белый, 255 — чёрный).

Делаются следующие предположения о выборке:

— Каждое изображение содержит ровно один печатный символ, полученный с помощью генератора
— Каждый символ на изображении полностью содержится в изображении (то есть расстояние между символом и границами изображения строго больше нуля).

~~images/cnn_graph.png~~

Функция ошибки.

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

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

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

Базовый вычислительный эксперимент

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

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

— 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.

Обучение сети будет осуществляться методом обратного распространения ошибки.

Вычислительный эксперимент

...

Теоретическая часть

Генерация признаков.

...

Классификатор на основе CNN для графов.

Свёрточная нейронная сеть для графов использует три базовые операции: свёртки, активации и пулинга.

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:

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