Распознавание текста на основе медиального представления и свёрточных сетей*

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

В работе рассматривается задача распознавания текста на изображении. Предлагается способ построения классификатора на основе обучения свёрточной нейронной сети, использующей в качестве входа медиальное представление цифрового изображения текстовых символов. 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 распознавания рукописных символов на изображении. Рассматриваются два варианта постановки задачи, «дискретный» и «непрерывный». В обоих вариантах первоначальным входом классификатора является дискретное изображение. В «дискретном» варианте классификатор работает непосредственно с этим дискретным изображением, в то время как в «непрерывном» варианте классификатор работает с непре-

---

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

рывным медиальным представлением, являющимся результатом обработки исходного дискретного изображения.

Введём определения, необходимые для постановки задачи.

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.

Определение 2. Дискретное изображения высоты h и ширины w — матрица 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.

Перейдём к постановке задачи. Пусть задано множество символов $\mathcal{S} = \{s_1, \ldots, s_k\}$ и выборка дискретных изображений:

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

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 $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_k\}$:

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

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 вектору вероятностей можно будет получить предсказание символа на изображении взяв символ с наибольшей вероятностью.

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, 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 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.

Постановка задачи (непрерывный случай)

В «непрерывном» варианте входом классификатора является медиальное представление, то есть скелетное представление с заданной на нём радиальной функцией. Введём необходимые определения, в соответствии с [1]:

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.

В работе предлагается использовать алгоритм скелетизации [добавить ссылку], выдающий медиальное представление в следующем формате: 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. Также, дополнительно, каждая вершина имеет степень от одного до трёх.

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

Предлагается по скелету сгенерировать матрицу вещественных чисел, в которой каждый элемент будет описывать точку фигуры символа на соответствующей позиции. Для

этого по скелету строится ограничивающий прямоугольник, расширяется на 10% с каждого края, делится на $n \times n$ прямоугольников равного размера. Рассматривается центр каждого прямоугольника и считается радиус максимального пустого круга с центром в этой точке, содержащегося в фигуре. Для этого перебираются все рёбра скелета, для каждого рассматриваются три точки: концы рёбер и точка на ребре, полученная аналитически из условия равенства нулю производной радиуса окружности. Радиус окружности считается как значение радиальной функции в одной из этих трёх точек минус расстояние от этой точки до центра прямоугольника.

Мы не сразу это поняли, но видимо результат этого алгоритма может быть также получен просто по исходному дискретному изображению (без построения скелета), а именно каждому пикселю изображения сопоставляется расстояние от пикселя до ближайшего белого пикселя (считаем что фон белого цвета, а символ чёрного). Поэтому непонятно даст ли такой такое преобразование улучшение качество по сравнению с CNN над дискретными изображениями и надо ещё подумать над возможными другими признаками.

Литература

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

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

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

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

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

[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.