Аннотация

Задача идентификации авторов для рукописных текстов является актуальной задачей в области компьютерного зрения, заключающейся в определении количества авторов набора рукописных документов и их кластеризации по писателям. Данная работа посвящана решению данной задачи путем использования различных архитектур сверточных нейронных сетей, а также различных методов кластеризации. В результате работы было протестировано множество моделей и методов, и предложены улучшения полученных результатов.

Содержание

1 Введение

1.1 Актуальность

Идентификация авторов рукописных текстов является актуальной задачей в области компьютерного зрения. Среди сфер использования данной технологии можно выделить анализ исторических документов, обработку рукописных текстов в судебной практике, кластеризацию огромного количества рукописных текстов в образовательной сфере, а также разметку датасета по писателям в автоматическом режиме, что может помочь улучшить качество работы генеративных нейронных сетей, обученных на рукописных текстах.

1.2 Постановка задачи

Работы по данной теме выделяют онлайн и оффлайн методы рапознавания авторов. Онлайн метод подразумевает обработку рукописного текста, который представлен в виде временных фрагментов штрихов, из которых извлекается уникальная информация о писателе. В свою очередь, оффлайн метод проводит анализ изображения уже написанного рукописного текста, из которого извлекаются признаки, по которым выявляется автор.

Задачу идентификации авторов можно решать в постановке как задачи классификации, так и задачи кластеризации. В случае задачи классификации каждый автор представляется из себя отдельных класс, который модель предсказывает, имея на вход рукописный текст. В случае задачи кластеризации, не зная заранее множество авторов и их количество, рукописные фрагменты разбиваются на кластера, каждый из которых предположительно написал один человек. Стоит отметить, что если задача решена в постановке кластеризации, то она решена в постановке классификации, так как в случае успешной кластеризации, можно сопоставить полученные кластера уже известным классам.

Данная дипломная работа будет решать задачу идентификации авторов в формулировке оффлайн кластеризации. Имея на входе документы с рукописным текстом, нужно определить количество писателей и кластеризовать тексты по авторам. Документы могут из себя представлять как полноценные тексты на бумаге, так и отдельно написанные от руки слова или предложения. Обученной модели на стадии inference могут подаваться тексты писателей, которых она не видела во время обучения.

2 Обзор существующих методов

Исследования в области идентификации авторов рукописных текстов проводились в течении многих лет, и улучшали постепенно результаты, предлагая различные методы и идеи извлечения и обработки признаков рукописного текста.

Для выявления признаков из полученного изображения современные научные работы в основном делают выбор на сверточных нейронных сетях. Используются различные архитектуры, включая ResNet-18 (источник), ResNet-50 (источник), VGG (источник). Данные модели показали хорошие результаты в классификационной постановке задачи, где их применяли в качестве энкодеров.

Научные работы предлагают различные варианты обучения энкодера. Например, некоторые исследования обучают энкодер в паре с полносвязной нейронной сетью, используя функцию потерь CrossEntropy (источник). Также есть исследования в области применения сиамской архитектуры обучения энкодера на данной задаче (источник).

Существует также несколько способов извлечения фрагментов из рукописного текста для дальнейшего извлечения признаков. Один из самых простых способов заключается в нарезания рукописного текста на слова или просто на фрагменты определенной ширины (источник). В некоторых работах из рукописного текста извлекаются самые информативные элементы почерка, которые обнаруживаются различными алгоритмами обнаружения углов (corner-detectors), например, HARRIS и FAST. После прохождения через сверточную нейронную сеть, полученные эмбеддинги потом агрегируются различными способами. Например находится среднее арифметическое векторов или используется алгоритм агрегации VLAD.

В целях значительного увеличения датасета и, в последствии, улучшения качества обучения, существует идея синтетической генерации датасета рукописных текстов, используя шрифты, похожие на рукописный текст, и применяя аугментацию (источник).

Также, в исследованиях распознования лиц применяется техника обучения Metric Learning, которая помогает в той области получить более репрезентативные эмбеддинги. Так, используя функцию потерь ArcFace удалось достичь значительного улучшения результата в задачи классификации фотографий лиц людей (источник). Не исключено, что данный метод хорошо себя может показать и на рукописных текстах.

3 Разработанные решения

Исходя из вышеописанных работ, можно составить общую архитектуру решения поставленной задачи. Рукописные тексты сначала проходят через стадию предобработки, во время которой улучшается качество самого рукописного текста, а также происходит его разбивка на фрагменты, либо путем нарезания на слова/части одинаковой ширины, либо путем применения алгоритма нахождения углов для получения максимально репрезентативных элементов почерка. Далее, эти фрагменты поступают в энкодер, который представляет из себя сверточную нейронную сеть, в результате чего получаются эмбеддинги. Далее, эти эмбеддинги при необходимости агрегируются в глобальный эмбеддинг фрагмента текста, если ранее был применен corner-detector. Наконец, применяется алгоритм уменьшения размерности эмбеддингов для улучшения качества кластеризации и применяется сам алгоритм кластеризации.

3.1 Препроцессинг и агрегирование

На вход энкодеру не подается целое изображение документа рукописного текста, так как в нем может содержаться лишняя информация, и энкодеру может быть сложно извлечь репрезентативные признаки из него.

Одним из самых простых решений этой проблемы является нарезание текста на фрагменты одинаковой ширины. Если учитывать, что высота одной строки текста одинакова, то при обучении сети не придется менять размер фрагментов или применять паддинг, чтобы их объединить в батч, тем самым сохраняя все информацию, содержащуюся во фрагменте, и не допуская смещения модели во время ее обучения.

3.1.1 Детекторы углов

Однако даже в уже нарезанном фрагменте может содержаться лишняя информация, так как прямые линии, содержащиеся в почерке, и пустые элементы на бумаге не содержат много информации, по которой можно различить автора. В связи с этим можно энкодеру подавать только фрагменты, содержащие самую важную информацию, например, углы и персечения. Найти подобные участки могут помочь так называемые детекторы углов. Существует множество алгоритмов в данной области. Самыми классическими являются Harris и FAST (источник).

3.1.2 Агрегирование средним и VLAD

После того, как энкодер обработал куски рукописного текста, на выходе мы имеем множество локальных эмбеддингов. Существует несколько способов получения глобального вектора, содержащего репрезентативные признаки текста. Одним из них является взятие арифметического среднего от всех локальных эмбеддингов. ... (Описать VLAD)

3.2 Обучение энкодера

Мы поняли как подать сверточной нейронной сети изображение, чтобы на выходе получить вектор. Но теперь нужно обучить энкодер выдавать именно репрезентативные и кластеризуемые эмбеддинги. Для этого существует несколько способов, которые будут описаны далее.

3.2.1 Auto-encoder

Во время выполнения данной работы мы хотели добиться минимального использования меток авторов во время обучения модели. Использование архитектуры автоэнкодера является одним из самых простых способов получения данного результата. Автоэнкодер представляет из себя комбинацию двух сверточных нейронных сетей, одна из которых называется энкодером, а вторая декодером. Данная архитектура обучается на задаче восстановления изображения, которое подается в начале модели. Предполагается, что после обучения энкодер научится ”сжимать” подаваемое на вход изображение в промежуточное состояние, представленное в виде вектора определенной размерности, таком образом, чтобы имеющийся декодер умел уже восстанавливать исходное изображение. Соответственно, промежуточное состояние содержит достаточно информации для восстановления изображения, и, в силу гладкости нейронной сети как обычной математической функции, можно построить гипотезу, что эмбеддинги одинаковых почерков должны находится близко друг к другу.

3.2.2 Сиамская нейронная сеть

Другая идея обучения энкодера исходит из того факта, что мы хотим получить именно кластеризуемые эмбеддинги. Это значит, что эмбеддинги текстов одного автора должны находиться на максимально близком расстоянии, а эмбеддинги различных авторов – на далеком, чтобы потом алгоритм кластеризации смог отделить ”облака” векторов. Существует

множество методов обучения нейронных сетей, которые непосредственно закладывают вышеописанное свойство в процесс обучения. Одним из таким методов является архитектура сиамской нейронной сети (SNN). Она представляет из себя пару идентичных нейронных сетей, веса которых непосредственно связаны. Во время обучения подаются пары изображений как от одного автора, так и от разных авторов. Сеть обучается таким образом, чтобы евклидово расстояние между векторами текстов от одного автора было минимально, а между векторами от разных авторов – как минимум равнялось какому-то гиперпараметру α.

Если говорить формально, определим функцию потерь для данной сети следующим образом. Пусть x_i, x_j – изображения почерка, которые подаются SNN на вход, c_k – множество изображений от автора с номером k, α – числовой гиперпараметр. Функцией отсечения назовем:

$$\text{ReLU}(y, \alpha) = \begin{cases} y, & 0 \leq y < \alpha \\ \alpha, & y \geq \alpha \end{cases}$$

Она будет использоваться в формуле функции потерь. Мы не хотим наказывать нейронную сеть во время обучения за то, что эмбеддинги находятся слишком далеко. Поэтому если расстояние между ними будет больше α, то мы будем отсекать его по заранее заданному гиперпараметру.

Определим целевую функцию:

$$Target = \begin{cases} 0, & x_i, x_j \in c_k \\ \alpha, & x_i \in c_k \text{ and } x_j \in c_q \end{cases}$$

Если изображения принадлежат одному автору (находятся в одном множестве c_k), то расстояние между ними должно быть равно нулю. Иначе, если они от разных авторов, то расстояние должно быть как минимум α.

Наконец, определим функцию потерь:

$$\text{Loss}(Target, emb_1, emb_2) = ||\text{ReLU}(||emb_1 - emb_2||, \alpha) - Target||$$

Как мы видим, перед тем как сравнивать расстояние между эмбеддингами и значение target, мы производим отсечку, чтобы не наказывать сеть за слишком далекие друг от друга эмбеддинги.

3.2.3 Обучение на задаче классификации

Альтернативной идеей обучения энкодера является обучение его на задаче классификации, с последующим отсечением классификатора. Стандартной архитектурой при обучении на задаче классификации является связка сверточной и полносвязной нейронных сетей.

Можно простроить гипотезу, что эмбеддинги, которые выдает энкодер во время обучения, обладают геометрическими свойствами, которые позволяют классификатору понять к какому автору рукописный текст действительно относится.

Функция потерь в данной ситуации играет огромную роль, так как именно от нее зависит каким образом полученные эмбеддинги будут расположены в пространстве. Обычно при обучении классификатора применяют стандартную функцию SoftMax:

$$L_{\text{SM}} = -\log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{N} e^{W_j^T x_i + b_j}},$$

где $x_i \in \mathbb{R}^d$ – эмбеддинг i-го изображения от автора с номером y_i, d – размерность пространства эмбеддингов, W и b задают линейное преобразование. Однако, данная функция потерь никак не способствует близости эмбеддингов от одного автора и дальности эмбеддингов из разных классов, и при большом количестве авторов пространство векторов будет плохо кластеризуемо [источник].

Поэтому в области распознования лиц используют другую функцию потерь для задачи классификации [источник].

$$L_{\text{ArcFace}} = -\log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{N} e^{s\cos(\theta_j)}}$$

При обучении с функцией потерь ArcFace эмбеддинги распространяются по гиперсфере радиуса s, и меняется именно угловое расстояние между ними. При этом обеспечивается геодезическое расстояние m между эмбеддингами разных классов. Это свойство может дать хорошо кластеризуемые вектора.

3.3 Обучение на синтетическом датасете

3.4 Уменьшение размерности

3.5 Кластеризация

Существует множество алгоритмов кластеризации, некоторые из которых принимают на вход уже известное количество кластеров, которое в нашем случае является неизвестным.