Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики

ОТЧЕТ ПО ПРАКТИКЕ

Учебная практика: технологическая (проектно-технологическая) практика
(наименование практики в соответствии с учебным планом)

Направление подготовки: 01.03.02 Прикладная математика и информатика

Выполнил:

Студент Иванов В.В.
(Ф.И.О.)

Группа ПМ-92

Факультет ПМИ

_____
подпись

«27» декабря 2022 г.

Проверил:

Руководитель от НГТУ Попов А.А.
(Ф.И.О.)

Балл: _____, ECTS_____,

Оценка _____
«отлично», «хорошо», «удовлетворительно», «неуд.»

_____
подпись

«29» декабря 2022 г.

Новосибирск 2022

СОДЕРЖАНИЕ

ВВЕДЕНИЕ В ОПТИЧЕСКОЕ РАСПОЗНАВАНИЕ

Оптическое распознавание символов — это направление распознавания образов, задачей которого является перевод изображений текста в текстовые данные, позволяющий компьютеру быстро получать большое количество информации из внешнего мира.

Создание точных систем оптического распознавания текста на данный момент является непростой задачей и требует дополнительных исследований в связи со специфическими требованиями по разрешению, быстродействию, надежности распознавания и объему памяти, которыми характеризуется каждая конкретная задача. [1]

Большинство алгоритмов, решающих данную задачу, можно поделить на две части: детекцию текста (выделение области) и непосредственно распознавание символов в области детекции. Недостаточно просто распознать отдельные символы на изображении: модели также необходимо уметь собирать их воедино для получения осмысленной синтаксической конструкции.

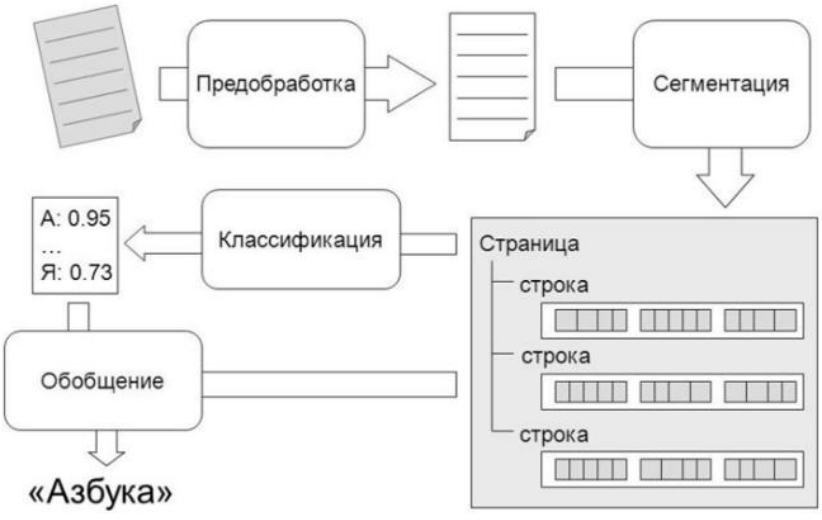

Рис. 1 – Этапы распознавания текста

С задачей распознавания символов связаны такие проблемы, как:

- разнообразие форм начертания символов;

- искажения на изображениях: шумы при печати, засвеченность, размытость и т.д.;
- вариации размеров и положения символов на изображении;
- влияние параметров печати: способ вёрстки, расстояние между строками и т.д.

Актуальность проблемы

На данный момент оптическое распознавание символов широко применяется в следующих задачах:

- автоматизация систем учёта и документооборота (сканирование бланков, анкет, паспортов и т.д.);
- автоматизация распознавания номерных знаков;
- перевод книг и рукописей в цифровой вид;
- оказание помощи слепым и слабовидящим;
- осуществление поиска слов и фраз в больших текстах и анализ больших текстов;
- хранение данных в более компактной форме;
- осуществление электронного перевода слов с использованием камеры телефона.

Классификация текстов

Машинописный текст - наиболее легко читаемый вид текстовых данных на изображении. Существующие модели способны распознавать прямые машинописные символы с очень высокой точностью.

Рукописный текст отличается большим разнообразием форм начертания символов. В отличии от машинописного текста, шрифт изменяется в соответствии с индивидуальным стилем пишущего человека,

из-за чего качество распознавания уменьшается. Тем не менее, существующие модели также способны распознавать и рукописные символы с достаточно высокой точностью.

Более высокие показатели могут быть достигнуты только с использованием контекстной и грамматической информации. Например, в ходе распознания искать целые слова в словаре легче, чем пытаться выявить отдельные знаки из текста. Знание грамматики языка может также помочь определить, является ли слово глаголом или существительным. Формы отдельных рукописных символов иногда могут не содержать достаточно информации, чтобы точно (более 98%) распознать весь рукописный текст.

Наиболее сложной и актуальной проблемой оптического распознавания символов является детекция и распознавание машинописного или рукописного текста, для которого характерны различного рода искривления. Текст, имеющий нетипичные формы, то есть отличный от горизонтального, в сообществе исследователей принято называть изогнутым (англ. curved). Частично изогнутый текст называется текстом с множественной ориентацией (англ. multi-oriented).

Общим для большинства алгоритмов распознавания изогнутого текста является процесс выравнивания текста в области детекции по определенному правилу, фактически превращающего искомый текст в прямой машинописный или рукописный.

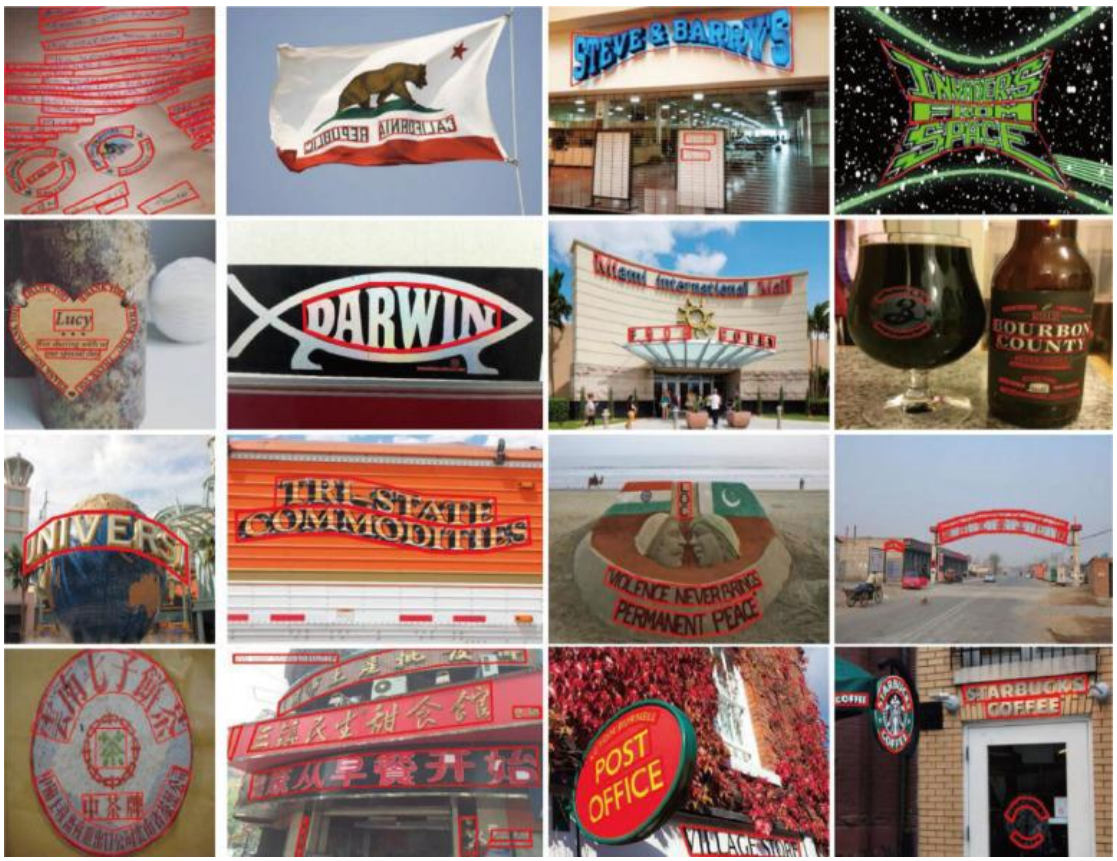

Рис. 2 – Примеры изогнутого текста [2]

НЕЙРОННЫЕ СЕТИ В ОПТИЧЕСКОМ РАСПОЗНАВАНИИ

Многие проблемы алгоритмов распознавания текста на изображении были решены с появлением глубоких нейронных сетей, способных, в частности, справляться с задачей в условиях деформации текста, окклюзий (пересечение текстов) и зашумленности.

Последние модели нейронных сетей выполняют как детекцию, так и распознавание текста, однако существуют также варианты, выполняющие только одну функцию.

Нейронная сеть может быть использована в системе распознавания текста в качестве классификатора. При обучении она получает на вход изображения, анализирует все позиции черных пикселей и выравнивает коэффициенты, минимизируя ошибку. Таким образом, достигается лучший результат распознавания.

Слои свертки и субдискретизации

Практически во всех моделях нейронных сетей для распознавания символов применяется многократная операция свёртки и субдискретизации – основного компонента сверточной нейронной сети.

Процесс свёртки заключается в обходе изображения окном определенного размера, которое позволяет выделить признаки уже на первом слое нейронной сети. Это значительно ускоряет процесс обучения и точность модели.

Свертка происходит в отдельном сверточном слое. В отличии от архитектуры перцептрона, в сверточной нейронной сети при свертке используется лишь ограниченная матрица весов, которую принято называть окном свертки. Веса по-прежнему неизвестны и определяются в ходе обучения.

Окно двигается по всему обрабатываемому слою (в самом начале – непосредственно по входному изображению), формируя после каждого

сдвига сигнал активации для нейрона следующего слоя с аналогичной позицией. То есть для различных нейронов выходного слоя используются одна и та же матрица весов, которую также называют ядром свёртки. Её интерпретируют как графическое кодирование какого-либо признака, например, наличие наклонной линии под определённым углом. Тогда следующий слой, получившийся в результате операции свёртки такой матрицей весов, показывает наличие данного признака в обрабатываемом слое и её координаты, формируя так называемую карту признаков (feature map).

Очевидно, что в сверточной нейронной сети набор весов не один, поскольку в ней происходит кодирование множества различных элементов изображения (например, линии и дуги под разными углами). При этом такие ядра свёртки не закладываются исследователем заранее, а формируются самостоятельно путём обучения сети методом обратного распространения ошибки. Обход каждым набором весов формирует свой собственный экземпляр карты признаков, делая нейронную сеть многоканальной, то есть имеющей множество независимых карт признаков на одном слое. Следует отметить, что при переборе слоя матрицей весов её передвигают обычно не на полный шаг (в соответствии с размером матрицы), а на небольшое расстояние. Так, например, при размерности матрицы весов 5×5 её сдвигают на один или два нейрона (пикселя) вместо пяти, чтобы не перешагнуть признак.

Второй компонент сверточных нейронных сетей – субдискретизирующий слой (слой пулинга). Как правило, этот слой располагается за сверточным. Его задача – уменьшение размерности сформированных карт признаков. Считается, что информация о факте наличия искомого признака важнее точного знания его координат, поэтому из нескольких соседних нейронов карты признаков выбирается максимальный и принимается за один нейрон уплотнённой карты признаков меньшей размерности. То есть, если на предыдущей операции свёртки уже

были выявлены некоторые признаки, то для дальнейшей обработки настолько подробное изображение уже не нужно, и оно уплотняется до менее подробного. За счёт данной операции, помимо ускорения дальнейших вычислений и избавления от лишних деталей (что избавляет от эффекта переобучения), нейронная сеть становится более инвариантной к масштабу входного изображения.

Таким образом, сверточная нейронная сеть, как правило, состоит из нескольких чередующихся сверточных и субдискретизирующих слоев. Это позволяет осуществлять переход от конкретных особенностей изображения к более абстрактным деталям – вплоть до выделения понятий высокого уровня. Нейронная сеть сама вырабатывает необходимую иерархию абстрактных признаков, фильтруя маловажные детали и выделяя существенные. Эти данные объединяются и передаются на обыкновенную полносвязную нейронную сеть, которая тоже может состоять из нескольких слоёв. При этом полносвязные слои уже утрачивают пространственную структуру пикселей и обладают сравнительно небольшой размерностью. [3]

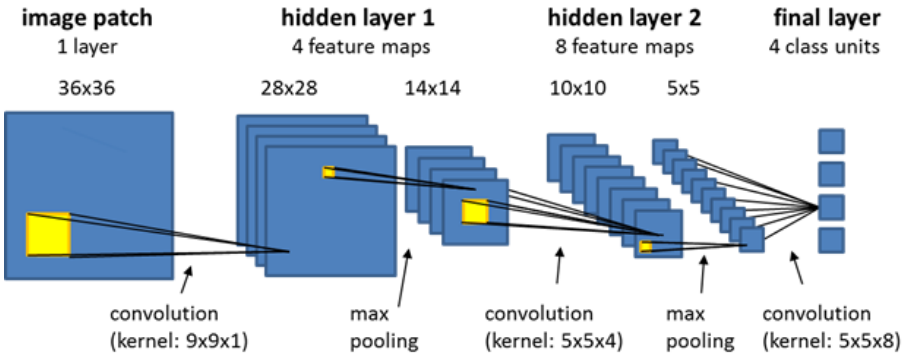

Рис. 3 – Архитектура сверточной нейронной сети

Рекуррентный слой

Рекуррентная нейронная сеть — вторая архитектура для распознавания символов, которая почти всегда используется в комбинации с процессом свертки. Это нейронная сеть с циклами, которые хорошо подходят для

обработки последовательностей. Она применима в таких задачах, где нечто целостное разбито на части, в нашем случае это — текст.

Обучение рекуррентной нейронной сети аналогично обучению обыкновенной, но с небольшим изменением алгоритма обратного распространения ошибки: поскольку одни и те же параметры используются на всех временных этапах в сети, градиент на каждом выходе зависит не только от расчетов текущего шага, но и от предыдущих временных шагов. Эта модификация получила название «алгоритм обратного распространения ошибки сквозь время» (англ. Backpropagation Through Time, BPTT) [4].

Существует множество элементов и разновидностей рекуррентных нейронных сетей. Наиболее интересная в нашем случае — рекуррентная нейронная сеть с краткосрочной памятью (англ. Long short term memory, LSTM).

LSTM-сеть — это нейронная сеть, содержащая LSTM-модули вместо или в дополнение к другим сетевым модулям. LSTM-модуль — это рекуррентный модуль сети, способный запоминать значения как на короткие, так и на длинные промежутки времени. Ключом к данной возможности является то, что LSTM-модуль не использует функцию активации внутри своих рекуррентных компонентов. Таким образом, хранимое значение не размывается во времени, и градиент или штраф не исчезает при использовании BPTT при обучении нейронной сети.

LSTM-блоки содержат три или четыре «вентиля», которые используются для контроля потоков информации на входах и на выходах памяти данных блоков. Эти вентили реализованы в виде логистической функции для вычисления значения в диапазоне [0; 1]. Умножение на это значение используется для частичного допуска или запрещения потока информации внутрь и наружу памяти. Идея заключается в том, что старое значение следует забывать тогда, когда появится новое значение, достойное запоминания. [5]

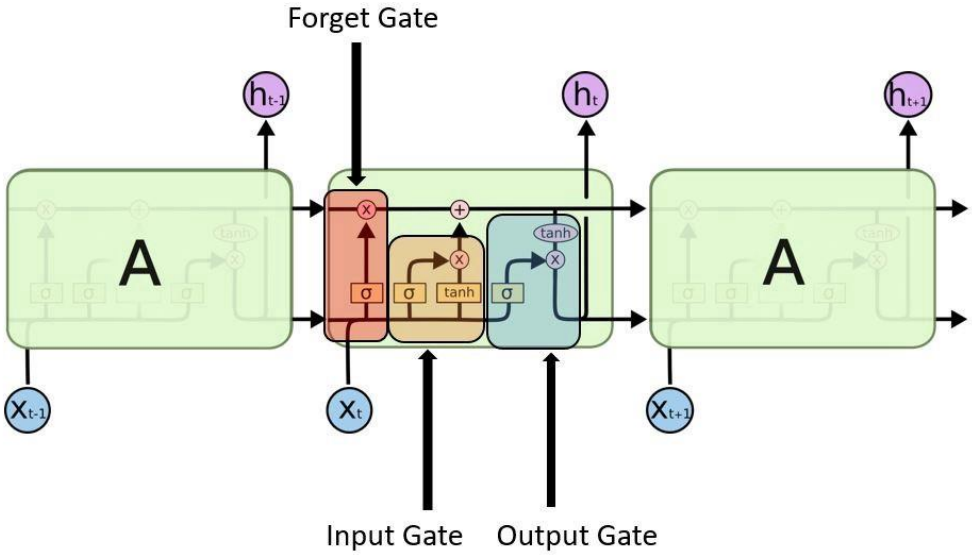

Рис. 4 – Структура LSTM-блока

БИНАРИЗАЦИЯ ИЗОБРАЖЕНИЙ

Несмотря на разнообразие форм текста на изображениях, единым для них остается процесс предварительной обработки исходных данных.

Первое, на что следует обратить внимание — это цвет изображения. Для выделения текста не требуется информация о цвете, поэтому от него следует избавиться. В цифровой обработке сообщений это возможно благодаря выполнению процесса бинаризации каждого изображения, входящего в исходную выборку.

Процесс бинаризации заключается в переводе цветного изображения в двухцветное, как правило — чёрно-белое. Соответственно, бинарным называют изображение, в котором каждый пиксел может представлять только один из двух цветов.

Значения каждого пиксела условно кодируются как «0» и «1». Значение «0» условно называют задним планом или фоном, а «1» — передним планом.

Благодаря наличию всего двух цветов размер изображения значительно снижается, сокращаются вычислительные затраты, а точность детекции признаков увеличивается благодаря фильтрации.

Существует также множество методов бинаризации, применяющих дополнительные локальные фильтры к различным типам изображений. [6]

Рис. 5 – Этапы бинаризации

ДЕТЕКЦИЯ ПРИЗНАКОВ

Задача модели детекции — выделение области текста, то есть нахождение набора символов на изображении. Правильность определения области текста напрямую влияет на качество работы распознающих моделей.

TextFuseNet

Основной особенностью TextFuseNet является выделение большего количества признаков и их слияние для более точного определения текстовых областей. TextFuseNet опирается на Mask R-CNN и Mask TextSpotter, рассматривая детекцию текста как задачу сегментации. Выделение признаков происходит на трёх уровнях: символьном, словесном и глобальном.

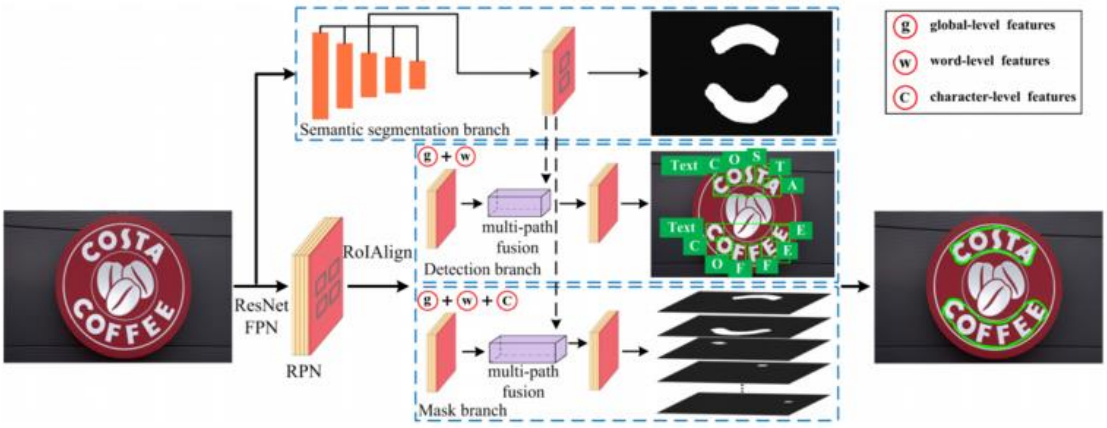

Рис. 6 – Архитектура TextFuseNet

Здесь RPN (Region Proposal Network) используется для генерации предполагаемых текстовых областей, что используется в последующих ветках Detection и Mask.

Сначала в ветви Semantic Segmentation с помощью сегментации определяются признаки на глобальном уровне. Далее в ветви Detection, извлекаются признаки на словесном уровне и объединяются с признаками глобального уровня. Полученное представление используется для регрессии

окружающей рамки и классификации объектов (текста/букв). Наконец, в ветви Mask извлекаются признаки на символьном уровне. Все три уровня признаков (символьный, словесный и глобальный) объединяются, и полученное представление используется при сегментации экземпляров (instance segmentation) для объектов, полученных в ветви Detection. При объединении признаков используется модуль Multi-Path Fusion. [7]

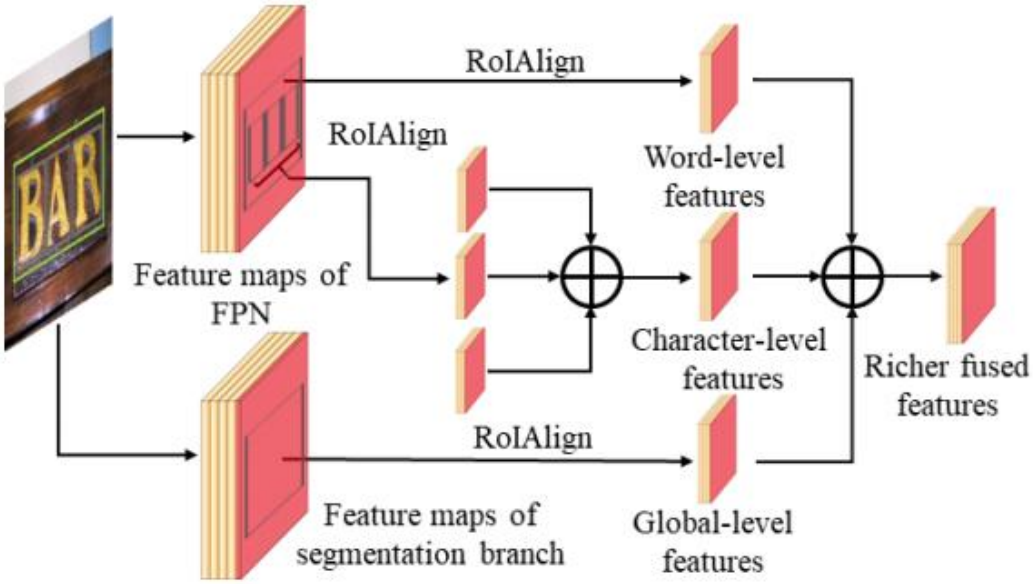

Рис. 7 – Структура модуля Multi-Path Fusion

РАСПОЗНАВАНИЕ ПРИЗНАКОВ

При распознавании происходит последовательная обработка выделенных при детекции признаков с последующим предсказанием символов в тексте.

Когда выбрана область изображения, содержащая текст, система распознавания переходит к классификации всех символов на изображении с целью определения языка, на котором написан текст. В случае с машинописным текстом также зачастую определяется шрифт.

Это возможно благодаря методу сопоставления с образцом (англ. pattern matching). В нашем случае метод сопоставляет символы на исходном изображении с набором эталонных символов, отражающих алфавит языка и, в случае с машинописным текстом, шрифт.

На данном этапе система распознавания уже обладает информацией о контексте, поскольку выделяет символы в конкретное множество. Это значительно упрощает работу с последующими изображениями и позволяет повысить показатель точности.

Контекст также может быть создан вручную. Например, если мы заранее зададим язык и шрифт машинописного текста или вовсе обучим модель исключительно на изображениях текста с одинаковым языком и шрифтом, то система перейдет непосредственно к распознаванию символов. Если же выборка содержит изображения с текстами на разных языках и выполненных разным шрифтом, понадобится сопоставление с образом.

END-TO-END МОДЕЛИ

Сквозные, или end-to-end модели являются наиболее актуальным типом архитектур нейронных сетей, объединяющих функции детекции области и распознавания символов.

CRNN

Данная модель состояит из CNN блока, который выделяет признаки, и BiLSTM блока, который обрабатывает полученные признаки. Выход из последнего блока декодируется в текст.

Основная идея работы заключается в обработке признаков разного уровня, то есть локальных, которые представляют собой признаки отдельных частей символов, и глобальных, признаков целого изображения. Входное изображение с текстом сперва разбивается на компоненты, которым соответствуют определенные части изображения. Далее между компонентами модель извлекает важную информацию, используя локальные и глобальные признаки, а затем формирует многогранное представление о тексте на изображении. В итоге получается последовательность признаков, в которой закодирован текст.

На входе нейронной сети расположен блок STN (Spatial Transformer Network), который выполняет выпрямление текста. Изображение с входным размером (32, 100) уменьшается до размера (32, 64) и подаётся в CNN сеть с полносвязным слоем в конце. На выходе получаем координаты ключевых точек, определяющих нижнюю и верхнюю границы текста. После этого выполняются преобразования исходного изображения, в результате которых текст выпрямляется.

Главном преимуществом использования STN по сравнению с другими методами выпрямления текста является отсутствие необходимости в разметке, так как сеть обучается благодаря градиентам, распространяющимся от сети распознавания.

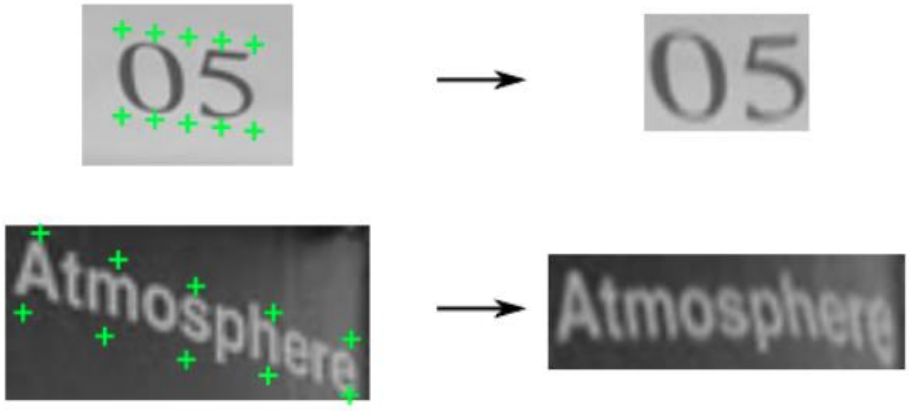

Рис. 8 – Примеры выпрямления текста с помощью STN

Полученное изображение подаётся на вход блоку Patch Embedding, суть которого заключается в разбиении изображения на патчи (малые части одного размера) и представлении каждого патча в виде вектора. К выходу Patch Embedding добавляется Position Embedding для передачи информации о позиции каждого патча на изображении.

Ключевая часть архитектуры заключается в разделении полученных компонент, описывающих отдельные области изображения, на два вида признаков: локальные и глобальные. Локальные признаки представляют собой паттерны наподобие штрихов, извлеченные из определенной области изображения. В них закодированы морфологические признаки и корреляция между компонентами определенной части изображения. Такие признаки помогают идентифицировать каждый отдельный символ, что позволяет справиться, например, с различением цифры “0” и буквы “O” на изображении.

Глобальные признаки определяют зависимости между всеми компонентами. Так как каждая часть изображения может относиться либо к тексту, либо не к тексту, то вычисление глобальных признаков может установить долгосрочную зависимость между компонентами текста, снижая влияние не текстовых компонент и придавая большее значение текстовым.

Для извлечения глобальных и локальных признаков авторами архитектуры были разработаны слои global mixing и local mixing. [8][9]

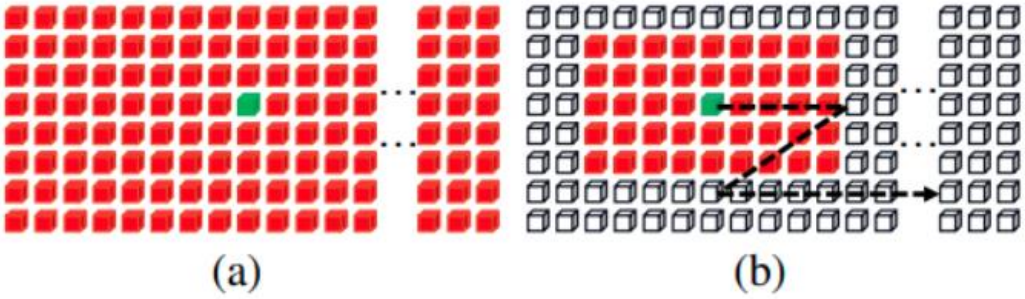

Рис. 9 – Процесс извлечения глобальных и локальных признаков

MANGO

Разделение модели на этапы детекции и распознавания вызывает трудности в обучении, так как результат распознавания сильно зависит от результата детекции, но обучать два этих слоя одновременно и взаимно — сложная задача.

Разработчики MANGO отказались от подобной архитектуры и делегировали обе задачи одному единственному слою. Именно поэтому MANGO невозможно протестировать на качество детекции — в этой модели этап детекции неразделим с этапом распознавания.

Особенность этой модели в том, что данные, спустя условный этап детекции, имеют такой вид, что код на этапе распознавания представляет собой легковесный инструмент. Это возможно благодаря тому, что этап детекции уже включает в себя элементы распознавания.

На вход инструменту распознавания подается так называемая позиционно-ориентированная маска внимания. Она представляет собой конкатенацию двух других масок: маски областей текста и многослойной маски точек внимания символов. Каждый слой маски символов сопоставлен с соответствующим слоем маски областей текста. Данные, представленные в таком виде (особенно важно сопоставление между двумя масками), сильно облегчают распознавание (выполняя часть работы по распознаванию заранее). Поэтому на данном этапе можно оставить лишь легковесный

инструмент и не писать для распознавания отдельный слой. Первоначальная обработка изображения происходит с помощью ResNet50 остаточной свёрточной нейронной сети. [10]

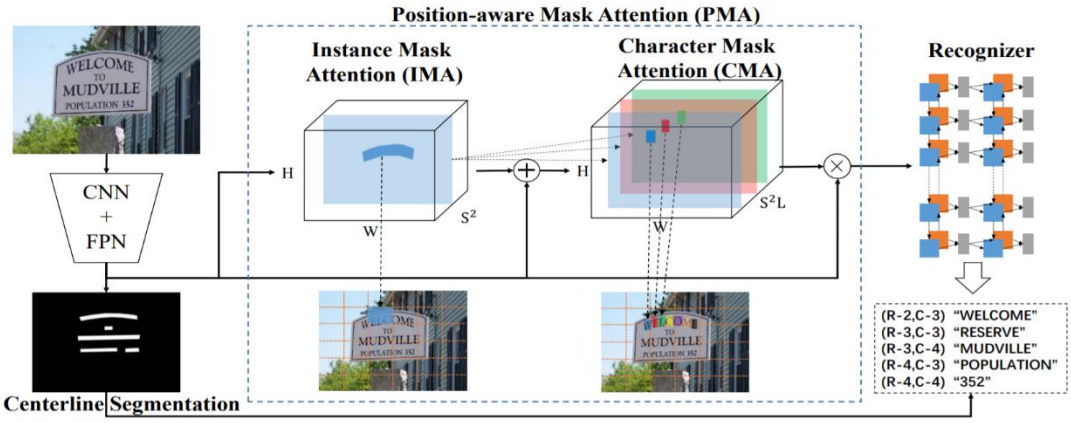

Рис. 10 – Архитектура MANGO

ЗАКЛЮЧЕНИЕ

В рамках учебной практики было приобретено понимание основных архитектур нейронных сетей, используемых в областях компьютерного зрения и обработки естественного языка. Получен опыт использования методов обработки изображений, включая метод бинаризации, имеющий множество специальных вариаций, используемых в широком спектре задач, связанных с распознаванием образов.

Данная работа является заготовкой для выпускной квалификационной работы, предполагающей более глубокое изучение архитектур нейронных сетей и методов оптического распознавания символов, а также проведение собственных исследований для сравнительного анализа преимуществ и недостатков с возможными предложениями по модификации существующих архитектур.

СПИСОК ИСТОЧНИКОВ

1. Илларионов А.А., Обзор методов распознавания текста [Электронный ресурс]. URL: https://alley-science.ru/domains_data/files/15May2019/OBZOR%20METODOV%20RASPOZNAVANIYa%20TEKSTA.pdf (дата обращения 27.12.22)
2. Curved scene text detection via transverse and longitudinal sequence connection [Электронный ресурс]. URL: https://www.sciencedirect.com/science/article/pii/S0031320319300664 (дата обращения 27.12.22)
3. Going deeper with convolutions [Электронный ресурс]. URL: https://arxiv.org/pdf/1409.4842.pdf (дата обращения 27.12.22)
4. Backpropagation Through Time and Vanishing Gradients [Электронный ресурс]. URL: https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-3/ (дата обращения 27.12.22)
5. Understanding LSTM Networks [Электронный ресурс]. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (дата обращения 27.12.22)
6. Исрафилов Х.С., Исследование методов бинаризации изображений [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/issledovanie-metodov-binarizatsii-izobrazheniy (дата обращения 27.12.22)
7. TextFuseNet: Scene Text Detection with Richer Fused Features [Электронный ресурс]. URL: https://github.com/ying09/TextFuseNet (дата обращения 27.12.22)
8. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition [Электронный ресурс]. URL: https://arxiv.org/abs/1507.05717 (дата обращения 27.12.22)
9. SVTR: Scene Text Recognition with a Single Visual Model [Электронный ресурс]. URL: https://arxiv.org/pdf/2205.00159.pdf (дата обращения 27.12.22)

10. A Closer Look at the Robustness of Vision-and-Language Pre-trained Models [Электронный ресурс]. URL: https://www.researchgate.net/publication/347398428_A_Closer_Look_at_the_Robustness_of_Vision-and-Language_Pre-trained_Models (дата обращения 27.12.22)