ИНТЕГРИРОВАННАЯ СИСТЕМА ОБЕСПЕЧЕНИЯ СЕГМЕНТАЦИИ И РАСПОЗНАВАНИЯ МАШИНОЧИТАЕМЫХ БЛАНКОВ

И.К. Квасникова, Н.Г. Авдеев
А.В.Лепустин
Томский политехнический университет
foxkik@tpu.ru

Введение

Контроль знаний, умений и навыков является важным звеном учебного процесса. В России одним из самых массовых тестирований является Единый государственный экзамен (ЕГЭ) – форма государственной итоговой аттестации (ГИА) по образовательным программам среднего общего образования. Для ознакомления учеников 11 классов с заданиями ЕГЭ, а также его процедурой проведения проводятся диагностические работы.

Существующий формат проведения данного мероприятия ставит перед организаторами задачу анализа информации, представленной в виде изображений – произвести так называемое off-line распознавание уже написанного на бумаге текста.

Задача обработки и распознавания изображений относится к разряду трудно формализуемых задач и является одной из наиболее важных на сегодняшний день.

Целью данной работы является разработка интегрированной системы распознавания, позволяющей автоматизировать этапы проведения диагностических работ в региональном центре обработки информации Томской области, связанные с обработкой рукописных текстов.

Существующий бизнес-процесс

Общее количество участников и их широкая территориальная распространенность, недостаточная оснащенность школ техническими средствами накладывают свои ограничения на возможные методы организации и проведения государственной итоговой аттестации.

В настоящее время при проведении ЕГЭ используются контрольные измерительные материалы (КИМ), представляющие собой комплексы заданий стандартизированной формы, а также специальные бланки для оформления ответов на задания, которые затем обрабатываются в региональных центрах обработки информации (РЦОИ). Таким образом, проведение диагностических работ напрямую связано с печатью, сканированием и обработкой бланков: регистрационных, а также бланков ответов №1 и №2 (для заданий с кратким и развернутым ответом соответственно).

В настоящее время исполнение этих задач разделено между автоматизированной информационной системой «Репетитор» и программой распознавания бланков ABBYY FormReader, а обмен данными осуществляется через экспортирование и импортирование файлов в формате TIFF и CSV. На рис. 1 этот процесс проиллюстрирован для одного бланка.

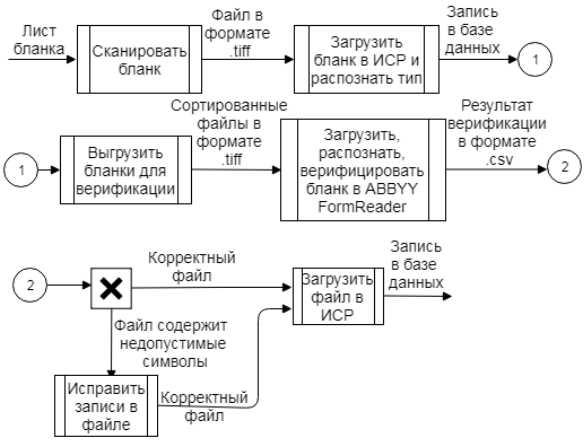

Рис. 1. Стадии обработки бланков

Особый интерес вызывает верификация и следующие за ней процессы. Верификация проводится путем сравнения на экране монитора символов, внесенных в машиночитаемые формы, в бланки ЕГЭ или в протокол проверки, с теми же символами, полученными в результате их распознавания.

Сотрудник-верификатор проверяет правильность распознавания символов и, в случае необходимости, вносит исправления. После завершения верификации результат распознавания сохраняется в формате CSV и снова подгружается в ИСР. При этом оператор следит за корректностью файлов .csv и при необходимости вручную вносит исправления.

Этапы, связанные с дополнительным экспортом изображений, их отдельной обработкой и импортом результата в ИСР, сопряжены с дополнительными трудозатратами и являются местом, где особенно требуется вмешательство оператора. К тому же, при печати и обработке бланков ИСР руководствуется жестко заданными в программном коде правилами, что делает систему неадаптивной к изменениям в процедуре.

В связи с этим задача устранения существующих узких мест бизнес-процесса и внедрение собственных шаблонов и алгоритмов обработки является особенно актуальной.

Некоторые процессы при печати и обработке бланков можно адаптировать к возможным изменениям путем внедрения шаблонов. Дополнительное вмешательство оператора, связанное с переносом данных из одной системы в другую, предполагается решить путем создания модуля распознавания в самой ИСР.

Распознавание бланков

Для решения задачи распознавания было решено применить нейронные сети. Обучающая вы-

борка была сформирована из файлов бланков репетиционных экзаменов. Исходными данными являлись xls-файл с ответами и tiff-файлы бланков.

Все изображения были сжаты до размера 28x28 пикселей (именно такой размер изображения будет использоваться во время проведения диагностики) и отцентрированы следующим образом: для каждого изображения был вычислен центр масс, а затем центр масс был совмещен с центром изображения.

Для программы репетиционного экзамена будут использоваться 3 нейронных сети для распознавания следующих множеств символов:

- цифры, минус, запятая;
- латинские буквы;
- русские буквы.

Использование 3 нейронных сетей необходимо для того, чтобы избежать неоднозначности в распознавании символов (цифра 0 и буква О, цифра 3 и буква з и т. д.).

Для реализации нейронных сетей были рассмотрена SharpLearning.Neuro – библиотека для машинного обучения с открытым исходным кодом, предоставляющая инструментарий для обучения с учителем для задач классификации и регрессии, а также оптимизации и проверки обученных моделей [1]. В качестве функции активации использовалась функция Softmax – нормализованная экспоненциальная функция, которая применяется в машинном обучении для задач классификации, когда количество возможных классов больше двух [2].

Изначально для распознавания был использован однослойный перцептрон. Из исходной выборки была выделена стратифицированная тестовая выборка, где количество элементов в каждом классе для тестирования составляет ~ 10% от количества элементов в соответствующем классе исходной выборки. Была исследована зависимость ошибки распознавания от количества нейронов на скрытом слое. Для сети, используемой для распознавания символов латинского алфавита, результат представлен на рис. 2.

Для нейронной сети, используемой для распознавания цифр, минусов и запятых было выбрано 650 нейронов на скрытом слое, для сети, использующейся для распознавания латинских символов – 750 нейронов, т.к. дальнейшее увеличение нейронов не приводит к существенному уменьшению точности распознавания, но приводит к увеличению времени работы нейронной сети.

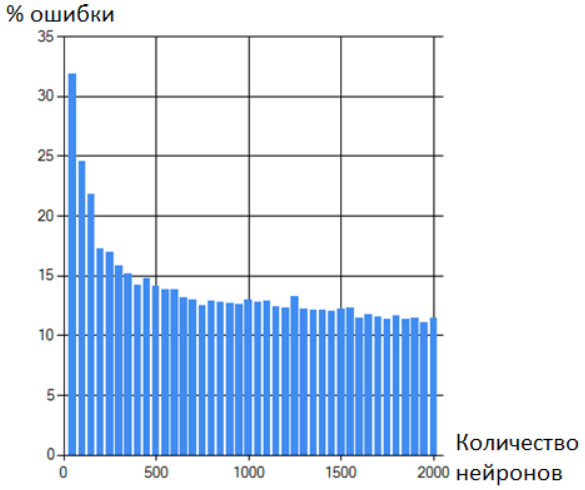

Рис. 2. Зависимость ошибки распознавания от количества нейронов на скрытом слое для латинских букв

Так же была исследована зависимость ошибки распознавания от количества итераций обучения, и было замечено, что уменьшение ошибки не происходит после 20-ой итерации, следовательно, при обучении целесообразно использовать именно это количество итераций.

Заключение

В ходе выполнения работы был изучен существующий в копании бизнес-процесс и архитектура информационной системы, решающая данную проблему на текущий момент. В результате было описано решение задачи шаблонизации бланков, а также была протестирована архитектура перцептрона с одним скрытым слоем.

На сегодняшний день самые передовые системы распознавания образов построены на базе сверточных нейронных сетей, поэтому при дальнейшей работе планируется протестировать СНС с активационными функциями типа ReLU и tanh.

Список использованных источников

1. SharpLearning: [Электронный ресурс] / GitHub – URL: github.com/mdabros/SharpLearning (дата обращения: 20.07.2018).
2. Softmax [Электронный ресурс] / Википедия, свободная энциклопедия – URL: https://ru.wikipedia.org/wiki/Softmax (дата обращения: 20.07.2018).
3. Калиновский И.А. Метод нейросетевого детектирования лиц в видеопотоке сверхвысокого разрешения. Дисс. на соиск. уч. степ. кандидата технических наук, Томск, 2016. - 191 с.