Создание универсальных классификаторов текстовых образов на основе сверточных нейросетевых технологий

Кузьмицкий Николай Николаевич
Кафедра "ЭВМ и системы"
Брестский государственный технический университет, Брест, Беларусь
knnbrest@yandex.ru

Аннотация

В данной работе предложен перспективный подход к построению универсальных классификаторов текстовых образов с использованием сверточных нейросетей (CNN). В его основе формирование комитетов CNN, обученных на образах различного масштаба, с последующей селекцией членов. Результаты проведенных экспериментов показали, что созданные классификаторы более эффективны, чем коммерческие системы на рассмотренных базах. При этом была достигнута уникальная точность распознавания тестового MNIST (заглавного NIST): *99.65%* (*98.174%*), что является 3-м (1-м) из лучших опубликованных результатов.

Ключевые слова: *CNN, комитет, селекция, универсальность.*

1. ВВЕДЕНИЕ

Технологиям распознавания текстовых образов (OCR) уделяется повышенное внимание с начала развития области машинного зрения. Причиной тому является множество сфер их потенциального применения (практическая мотивация), а задача распознавания – одна из классических в контексте ИИ (творческая). В результате проведенных исследований:

1) накоплен большой объем информации в форме моделей классификаторов, методов выделения признаков, баз и др. [6].

2) созданы прикладные программные продукты, например, системы обработки банковских документов.

3) проведены эксперименты, показавшие возможность достижения точности распознавания, сравнимой с человеческой на отдельных тестовых множествах [2].

Однако, несмотря на большой прогресс в решении OCR задач, по-прежнему не решенной остается главная – создание универсального классификатора текстовых образов. Помимо объективных причин (их высокой вариативности) не менее важное значение имеет и ориентация разработчиков в первую очередь на специализированные решения в рамках предполагаемых ограничений: шрифтовой, неискаженный, рукопечатный текст и др. При этом во многих приложениях данный подход не позволяет достичь приемлемой точности: начиная от обработки рукописных документов, заканчивая естественными сценами и робототехникой.

Анализ литературы выявил значительный недостаток исследований, связанных с решением главной OCR задачи. Большая часть из рассмотренных работ посвящена созданию методов, способных определять тип текстовых блоков (чаще строк или слов), с целью последующего выбора подходящего классификатора. В частности, авторы [14] применяют подход на основе "мешка слов", SIFT-детекторов и SVM, достигая хороших результатов разделения. Однако, используемые ими типы текстовых образов (машинописный и рукописный) охватывают лишь часть всех возможных, что, как показано в данной работе, может привести к падению точности распознавания на "смешенном" типе. Анализ распределения 822714 образов цифр, проведённый в [12], показал существование устойчивых границ классов, каждый из которых образует несколько крупных центров притяжения (кластеров). Вместе с тем, поиск данных границ затруднен существованием выбросов и аллографов – образов близких сразу нескольким классам. В [10] на примере рукописных баз (MNIST, USPS, DIGITS) и различных методик выделения признаков показана уязвимость классификаторов на основе моделей k-NN, SVM и частично нейросетях, в их способности переносить знания с тренировочного множества на другие. Авторы пришли к выводу, что данная проблема, названная "хрупкостью" (слабостью в ИИ терминологии), присуща всем моделям, основанным на методах машинного обучения.

В представленной работе описан подход к созданию универсальных классификаторов текстовых образов на базе сверточных нейронных сетей (CNN) и их комитетов. При этом основное внимание уделено изучению способности CNN распознавать образы различных типов и эффективному объединению знаний CNN. В ходе проведения экспериментов использовались как известные базы данных, так и новые.

2. ОСНОВНЫЕ МОДЕЛИ И ДАННЫЕ

2.1 Постановка задачи

Под универсальностью классификатора будем понимать его способность сохранять высокую точность распознавания на произвольной выборке входных данных. В зависимости от способа получения нами были выделены три типа текстовых образов: *машинописные, рукописные и синтезированные*, примеры которых приведены на рис. 1. К последнему типу относятся изображения, полученные в результате применения серий пространственных преобразований к образам первых двух. Так, в описываемом исследовании использовалась техника генерации, основанная на "волновых искажениях", образованных суперпозицией функций вида:

$$f_i(x) = (-1)^i \times a_i \times \cos(\pi / l_i \times x)$$

где a_i – амплитуда, l_i – длина i-ой функции. Подобные искажения позволяют имитировать эффекты, возникающие при неблагоприятных окружающих условиях съемки и дефектах аппаратуры (размытии, боковом ракурсе и др.).

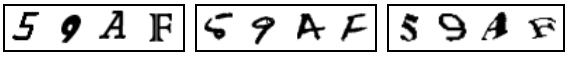

Рисунок 1: Слева направо: машинописные, рукописные и синтезированные текстовые образы.

Представленное исследование было выполнено для модели CNN, которая доказала свою эффективность в решении различных задач машинного зрения. CNN использовались для классификации двух подмножеств текстовых образов: арабских цифр и заглавных букв английского языка, ввиду их большой распространенности в практических приложениях. Оценка универсальности характеристик проводилась по перекрестному распознаванию образов выделенных типов.

2.2 Сверточная нейросетевая модель

В качестве базовой для построения сверточной модели была выбрана архитектура LeNet-5 [5], являющаяся наиболее разработанной в данном классе нейронных сетей. Данная архитектура способна выделять признаки образов в режиме "черного ящика" благодаря следующим особенностям:

1) *локальные рецептивные поля*: нейроны получают входной сигнал от окрестностей нейронов предыдущего слоя, за счет чего сеть обучается двумерной структуре входного образа;

2) *разделяемые веса*: нейроны слоя объединены картами, в которых они обладают общими весами, при этом карты генерируют различные признаки и сокращают количество параметров, настраиваемых в ходе обучения;

3) *пространственные подвыборки:* локальное усреднение откликов карт приводит к синтезу высокоуровневых признаков и повышает инвариантность сети к искажениям.

Отметим, что использованная архитектура, изображение которой приведено на рис. 2, отличалась от классической, т.к. перед выходным в ней отсутствовал RBF слой. Данный факт объясняется целесообразностью его применения при большем числе классов, например, для полного алфавита.

Обучение нейронных сетей проводилось модификацией алгоритма обратного распространения ошибки , основанной на стохастическом диагональном методе Левенберга-Марквардта [5], в online-режиме, с параметром обучения равномерно уменьшающимся от 0.001 до 0.000001 в течение 68 эпох. Перед каждой из них образы тренировочного множества подвергались эластичным ($\sigma = 8$, $\alpha = 36$) и аффинным искажениям (поворот на ±15°, для образов '1', '7', 'I' – ±7°, и масштабирование в пределах ±15%, для каждой размерности). Отметим, важность искажений в повышении обобщающих свойств сети и предотвращении переобучения.

2.3 Базы маркированных образов

Для обучения и тестирования CNN использовались базы, содержащие текстовые образы одного из типов:

1) машинописные: *FONT* – тренировочная часть включает по 1866 шрифтовых образов каждого класса (для цифр и заглавных букв) из [13] и по 508 образов с нормальным и полужирным начертанием из [1] (тестовая часть, содержит по 508 курсивных), *ETL6* – 13830 образов цифр и 38724 заглавных рукопечатных из [3];

2) рукописные: *MNIST* – 60000/10000 образов цифр в тренировочной/тестовой части из [8], *NIST* – 69522/11941 образов заглавных букв (разделены аналогично с [2]), *ETL1* – 14450 цифровых и 37570 заглавных образов из [3];

3) синтезированные: *KNI1* и *KNI2* – по 6000/1000 цифровых и 2500/500 заглавных образов каждого класса в тренировочной/ тестовой части, созданных на базе шрифтовых с помощью методики генерации, основанной на волновых искажениях и поворотах с контролем попарного подобия.

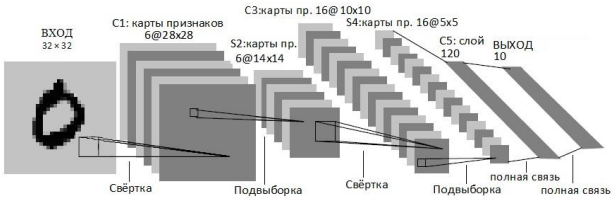

Рисунок 2: Использованная нейросетевая архитектура

Описанные базы по цели их использования можно разделить на две группы: *главные* (FONT, MNIST, NIST, KNI1) – применялись для построения нейронных сетей и анализа универсальности, *контрольные* (ETL, ETL6, KNI2) – для оценки обобщающих свойств итоговых классификаторов. Ввиду различия формата образов баз (размеров, диапазонов яркости), применялась их предобработка по схеме, аналогичной получению MNIST. Заметим также, что в дальнейшем изложении имя базы с префиксами *_dig/_big* и *_test/_train* трактуется как ее цифровая/заглавная и тренировочная/тестовая части соответственно.

3. ПРЕДЛАГАЕМЫЕ РЕШЕНИЯ

3.1 Построение "частных" CNN

Под "частной" будем понимать CNN, построенную на основе образов одного из выделенных типов. Так, для создания трех частных цифровых (заглавных) CNN использовались базы FONT, MNIST (NIST) и KNI1: обучение проводилось на их тренировочных, а проверка на тестовых частях. Результаты тестирования нейросетей, представленные в таблице 1, позволяют отметить следующее: 1) CNN имели более 99% точность распознавания образов своего типа (за исключением CNN_NIST), что подтвердило эффективность выбранной архитектуры сети и методики ее обучения; 2) максимальный уровень ошибок был получен на множестве KNI1, что доказало актуальность рассмотрения синтезированного типа образов; 3) средняя точность сетей 92.58% (91.90%) показала низкую универсальность частных CNN.

3.2 Объединение "частных" CNN в комитеты

Имея высокоточные на образах своего типа CNN, перспективным являлся путь объединения их знаний с помощью методов коллективного распознавания. Хорошие предпосылки для формирования комитетов создавала общая архитектура и методика обучения CNN, обеспечивающие нормировку их откликов, как вероятностей классов, в единый диапазон. Одним из факторов эффективности комитета является схема голосования членов. Из существующих были выбраны наиболее универсальные, для которых победителем является класс: *максимальное (MAX)* – с максимальным откликом членов, *усредняющее (AVER)* – с максимальным средним откликом членов, *большинством (MAJOR)* – с наибольшим числом голосов членов в свою пользу.

CNN	MNIST	FONT_dig	KNI1_dig	Среднее
CNN_MNIST	99.39	95.23	85.06	93.22
CNN_FONT_dig	87.23	99.58	88.06	91.62
CNN_KNI1_dig	86.52	92.97	99.29	92.92
	NIST	FONT_big	KNI1_big	
CNN_NIST	97.69	86.75	81.49	88.64
CNN_FONT_big	89.41	99.15	89.27	92.61
CNN_KNI1_big	89.26	95.148	99.00	94.46

Таблица 1: Точность (в %) "частных" CNN.

Комитет	MNIST	FONT_dig	KNI1_dig	Среднее
MAX_COM_dig	98.23	98,62	98,56	98.47
AVER_COM_dig	97,64	99,11	98,10	98.28
MAJOR_COM_dig	93.82	98.50	98.86	97.06
	NIST	FONT_big	KNI1_big	
MAX_COM_big	96.62	98.27	98.21	97.70
AVER_COM_big	96.60	98.45	97.98	97.67
MAJOR_COM_big	95.09	97.26	95.43	95.92

Таблица 2: Точность (в %) комитетов "частных" CNN.

На основе ранее описанных CNN были сформированы три цифровых (заглавных) комитета с разными схемами учета голосов. Результаты тестирования комитетов, отраженные в таблице 2, показали, что схема MAX является наиболее эффективной. Это объясняется низким уровнем корреляции откликов частных CNN, ввиду отличия типа образов используемых при их обучении. Средняя точность MAX комитета превышала аналогичную одиночных CNN на 5.89% (5.80%), что доказало целесообразность объединения их знаний, однако уровень универсальности комитетов все еще оставался не достаточно высоким.

3.3 Формирование "частных" комитетов

Комитет позволил объединить знания одиночных CNN, однако их перекрестная точность все еще оставалась на низком уровне. С целью его повышения была исследована возможность формирования "частных" комитетов и их объединения в одном. Для создания комитетов применялась процедура "*регулярного масштабирования*": выберем диапазон вариации высоты h и ширины w образа (учитывая, исходный размер в 20×20 использовались [16, 24] для h и [10, 18] – w, с шагом 2 пикселя) и выполним его масштабирование к каждому из допустимых размеров (всего возможны 25 комбинаций с фиксацией h и w, 10 – с фиксацией одного из них). Таким образом, для любого тренировочного/тестового множества можно построить 35 подмножеств аналогичной мощности, обучить 35 CNN и сформировать из них комитет, схема применения которого отражена на рис. 3.

Преимуществами описанной выше процедуры являются: 1) отсутствие необходимости привлекать дополнительные базы данных; 2) обеспечение высокого разнообразия образов; 3) соответствие естественной природе различия образов: шрифтов, наклонов и др. Кроме того, по сравнению с другими методиками, основная цель процедуры – повышение универсальности характеристик классификатора, а не его точности распознавания на отдельном множестве образов. Отметим также существенное значение при изменении h и w выход за пределы 20×20 пикселей (в отличие от [2]). Это позволяет настраивать CNN-фильтры фиксированного размера 5×5 на детектирование уникальных признаков, за сет варьирования степени приближения образа, наподобие "увеличительного стекла". Недостатком процедуры является значительное время обучения: для одной CNN от 3 до 5 часов на ЭВМ стандартной аппаратной конфигурации.

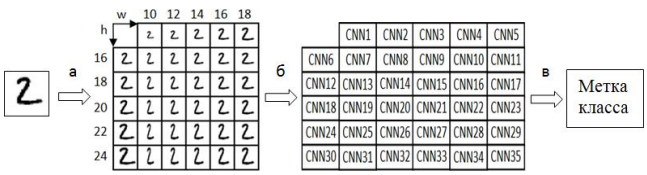

Рисунок 3: Схема применения "частного" комитета CNN: а) регулярное масштабирование; б) распознавание образов; в) объединение решений.

Комитет	MNIST	FONT_dig	KNI1_dig	Среднее
COM_MNIST_dig	99.58	96.85	89.77	95.40
COM_FONT_dig	90.82	99.82	90.44	93.69
COM_KNI1_dig	90.22	96.31	99.86	95.46
COM_COM_dig	98.95	99.27	99.06	99.09
	NIST	FONT_big	KNI1_big	
COM_NIST	97.94	89.28	86.95	91.39
COM_COM_big	96.61	98.43	98.60	97.88

Таблица 3: Точность (в %) "частных" комитетов CNN.

Ввиду указанного недостатка в ходе экспериментов было создано меньше CNN, чем предусматривалось процедурой: построены по 10 сетей для баз MNIST, FONT_dig, KNI1_dig и NIST, из которых были сформированы частные комитеты (*COM*), объединенные в общие (*COM_COM*). Для частных комитетов использовалось AVER голосование (в связи с высокой корреляцией ошибок), для общих – MAX. Результаты тестирования комитетов, представленные в таблице 3, позволили отметить повышение перекрестной точности частных комитетов по сравнению с одиночными CNN на 2.18%, 2.07%, 2.54%, 2.75% для MNIST, FONT_dig, KNI1_dig, NIST соответственно. При этом средняя точность общего цифрового (заглавного) комитета возросла на 0.62% (0.18%). Таким образом, регулярное масштабирование показало свою эффективность даже в своей сокращенной реализации, что доказывает перспективность процедуры.

3.4 Селекция членов комитетов

Основными факторами создания качественного комитета является *точность* и *разнообразие* (степень корреляции ошибок) членов [7]. Результаты экспериментов показали невысокий уровень разнообразия в частных комитетах, построенных хотя и с вариацией, но все же на одном обучающем множестве. Данный факт, наряду с необходимостью уменьшения размеров комитетов с целью их эффективной аппаратной реализации, привел к исследованию вопроса селекции членов, основной задачей которой является выбор подкомитета с точностью не ниже самого комитета.

Для этого использовались по 10 CNN, обученных ранее на базах MNIST и NIST с вариацией размеров: (h = 16, 18, 20, 22, 24, *w/h*), (h = 20, w = 10, 12, 14,), (h = 18, w = 18*)*, (w = 22, *w/h)* (где *w/h* = пропорционально), имеющие точности 99.14% – 99.44% и 97.39% – 97.79% на тестовых частях баз. Селекция CNN проводилась с помощью алгоритма *EPIC* [7], в основе которого лежит тезис: члены с бóльшим (меньшим) числом верных (ложных) предсказаний в меньшинстве наиболее (наименее) полезны (вредны) при формировании комитета. Выходом алгоритма является перечень членов в порядке убывания их вкладов в общую эффективность, который трактовался двумя способами: 1) путем пошагового добавления членов строилось десять подкомитетов и выбирался самый точный; 2) подкомитет инициализировался первым членом, добавление остальных проводилось пока мог быть найден член, повышающий общую точность. Алгоритмом были сформированы по два подкомитета (из 5 и 6 членов) с точностью распознавания: тестового MNIST – 99.59% и 99.6%, NIST – 98.02% и 98.08%, превышающей уровень всего комитета, при этом количество членов, по сравнению с ним, было значительно уменьшено.

В базовой реализации EPIC корректность предсказания членами оценивается бинарными величинами, поэтому возникло предположение об их замене вероятностными откликами CNN. Так, модифицированный EPIC позволил

сформировать *комитеты с 99.65% точного распознавания тестового MNIST и 98.174% заглавного NIST!* Данные результаты являются 3-м и 1-м из лучших для данных точек отсчета, при этом по сравнению с ближайшими из [8] и [2], использовались гораздо менее громоздкие архитектуры CNN.

Отметим также, что комитеты в [2] содержали по 35 членов, в отличие созданных: MNIST – 4 члена (h = 16, 24, w / h), (h = 20, w = 10, 18), NIST – 6 (h = 16, 20, 24 w / h), (h = 20, w = 14) (w = 22, *w/h),* (h = 18, w = 18). Выбор данных сетей выглядел неоднозначным с позиции точности (10, 1, 9, 6-ая для MNIST) однако с точки зрения разнообразия он не являлся случайным, т.к. данные CNN были созданы при значительно отличных вариациях масштабов символов.

3.5 Формирование комитетов "общих" CNN

Кратким итогом описанных выше исследований является: "частные" CNN обладают низкой универсальностью, комитет является хорошим средством объединения их знаний, его размер ограничен практической эффективностью, поэтому необходима селекция членов. "Частные" CNN – наиболее слабое звено исследования, поэтому в его заключительной части был выполнен переход к "общим" CNN. Их обучение проводилось на смешенных выборках исходных баз. В частности, для построения цифрового (заглавного) множества MKF (NKF) использовалось по 20000/5000 (20540/5200) образов тренировочных/тестовых частей баз MNIST (NIST), FONT и KNI1 (количество образов обеспечивало равное представительство классов и их типов).

Результаты тестирования обученных CNN, приведенные в таблице 4, показали, что средняя точность "общих" CNN не только значительно превышала точность "частных", но и их комитетов: MAX_COM_dig на 0.65%, MAX_COM_big – 0.73% и даже COM_COM_dig на 0.03%. Причина данного факта заключается в том, что хотя базы MKF и NKF содержали меньше образов, однако различие их типов заставило CNN выделять более универсальные признаки, т.е. вместо объединения знаний выполнялась их интеграция. Эксперименты по схеме: регулярное масштабирование + селекция членов позволили построить комитеты, CNN в которых обучались на образах размеров: (h = 20, 16, 24, w/h), (h = 20, w = 10, 18, 14), (w = 20, w/h). Их средняя точность, отраженная в таблице 4, была максимальной полученной. Для сравнения в таблице 4, 5 также приведены лучшие результаты систем [9, 11] (при разных значениях OCR/ICR параметров).

4. ЗАКЛЮЧЕНИЕ

Исследование показало низкую перекрестную точность CNN, одной из лучших моделей классификаторов, в распознавании текстовых образов с типом отличным от ее тренировочных.

Классификатор	MNIST	FONT_dig	KNI1_dig	Среднее
CNN_ MKF	98.82	99.01	99.55	99.12
COM_ MKF	99.15	99.56	99.86	99.52
SmartZone	97.65	95.07	89.04	93.92
Nicomsoft OCR	97.50	97.44	85.84	93.59
	NIST	**FONT_big**	**KNI1_big**	
CNN_ NKF	97.17	98.10	98.07	97.78
COM_ NKF	97.68	98.80	98.81	98.43
SmartZone	92.07	93.85	86.46	90.79
Nicomsoft OCR	85.55	97.76	85.26	89.52

Таблица 4: Точность (в %) "общих" CNN, их комитетов и сторонних классификаторов на главных базах.

Классификатор	ETL1	ETL6	KNI2_dig	Среднее
COM_ MKF	99.05	99.78	99.25	99.36
SmartZone	96.90	99.43	93.65	96.66
Nicomsoft OCR	95.53	98.06	94.10	95.89
	ETL1	**ETL6**	**KNI2_big**	
COM_ NKF	98.19	99.51	99.18	98.96
SmartZone	92.55	97.42	87.99	92.65
Nicomsoft OCR	92.69	95.97	89.43	92.69

Таблица 5: Точность (в %) комитетов "общих" CNN и сторонних классификаторов на контрольных базах.

Подход, основанный на объединении знаний "частных" CNN и их интеграции с помощью единой выборки типов в "общих" доказал свою эффективность в повышении универсальности. Методика создания комитетов CNN, обученных на разных масштабах образов, и их селекция позволили достичь уникальной точности распознавания тестового MNIST и заглавных букв NIST. Перспективными направлениями продолжения исследования является регуляризация множеств образов, развитие архитектуры CNN и их комитетов.

5. ССЫЛКИ

[1] T.E. Campos, B.R. Babu *Character Recognition in Natural Images*. VISAPP (2), 2009, pp. 273-280.

[2] D.C. Ciresan, U. Meier *Multi-column deep neural networks for image classification*. CVPR, 2012, pp. 3642-3649.

[3] ETL character databases: http://projects.itri.aist.go.jp/etlcdb/.

[4] P.J. Grother *Nist special database 19 – handprinted forms and characters database*. NIST, Tech. Rep., 1995.

[5] Y. LeCun, L. Bottou *Gradient-Based Learning Applied to Document Recognition*. Proceedings of the IEEE, 86(11), 1998, pp. 2278-2324.

[6] C.-L. Liu, H. Fujisawa *Classification and Learning Methods for Character Recognition: Advances and Remaining Problems*. Machine Learning in Document Analysis and Recognition, 2008, pp. 139-161.

[7] Z. Lu, X. Wu *Ensemble pruning via individual contribution ordering*. KDD, 2010, pp. 871-880.

[8] MNIST database: http://yann.lecun.com/exdb/mnist/.

[9] Nicomsoft OCR: http://www.nicomsoft.com/products/ocr/.

[10] A.K. Seewald *On the Brittleness of Handwritten Digit Recognition Models.* ISRN Machine Vision, vol. 2012, Article ID 834127, 10 pages.

[11] SmartZone: http://www.accusoft.com/smartzone.htm.

[12] S. Uchida, R. Ishida *Character Image Patterns as Big Data*. ICFHR, 2012, pp. 479-484.

[13] J.J. Weinman, E. Learned-Miller *Scene text recognition using similarity and a lexicon with sparse belief propagation* IEEE Trans. on PAMI, 31 (10), 2009, pp. 1733–1746.

[14] K. Zagoris, I. Pratikakis *Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm*. ICFHR, 2012, pp. 103-108.

Об авторах

Кузьмицкий Николай Николаевич – аспирант, кафедры "ЭВМ и системы" БрГТУ. Его адрес: knnbrest@yandex.ru