УДК 004.932.72’1
DOI 10.25205/1818-7900-2020-18-2-54-61

Автоматическое тегирование изображений одежды

А. Г. Малышев, А. С. Полыгалов, С. А. Алямкин

ООО «Экспасофт»
Новосибирск, Россия

Аннотация

Описывается разработка системы автоматической разметки изображений одежды человекочитаемыми атрибутами (тегами). Подобные системы становятся востребованными в сфере коммерции для пополнения информации об инвентаре и улучшения его организации, а также реализации интерактивного поиска по фотографиям, доступного клиентам. Построенное решение способно выполнять автоматический анализ атрибутов длины, дизайна и цвета для произвольного количества предметов одежды на фотографии. Архитектура решения позволяет менять набор предсказываемых тегов или переходить к решению задач тегирования на других данных.

Ключевые слова

машинное обучение, обработка изображений, компьютерное зрение, тегирование

Для цитирования

Малышев А. Г., Полыгалов А. С., Алямкин С. А. Автоматическое тегирование изображений одежды // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, № 2. С. 54–61. DOI 10.25205/1818-7900-2020-18-2-54-61

Automatic Tagging of Clothing Images

A. G. Malyshev, A. S. Polygalov, S. A. Alyamkin

Expasoft LLC
Novosibirsk, Russian Federation

Abstract

This paper presents a computer vision clothing auto-tagging algorithm. Tagging is highly demanded in e-commerce as a tool to create a rich uniform set of annotations. The annotations improve catalog organization, statistics, and can be used for interactive catalog search by consumer photos. The proposed algorithm predicts length, design, and color attributes for an arbitrary number of clothing items in an image. The modular structure of the proposed system allows reconfiguration for other sets of tags and tagging tasks not related to clothing.

Keywords

machine learning, image processing, computer vision, tagging

For citation

Malyshev A. G., Polygalov A. S., Alyamkin S. A. Automatic Tagging of Clothing Images. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 2, p. 54–61. (in Russ.) DOI 10.25205/1818-7900-2020-18-2-54-61

Введение

Постоянно меняющиеся коллекции одежды, растущее разнообразие стилей и большой ассортимент существенно усложняют как задачу выбора товара для покупателя, так и задачу систематизации инвентаря и рекомендации подходящего предмета для продавца.

Основной способ борьбы с данной проблемой – составление каталогов с объединением товаров в иерархию категорий, в которых проще ориентироваться. Однако разнообразие товаров уже стало настолько высоким, что внутри привычных категорий, таких как «футболка», «водолазка», «толстовка», нужно дополнительное структурирование для эффективного поиска подходящих вариантов.

Дополнительную информацию о товарах могут предоставлять производители, но разные производители часто предлагают несовпадающие наборы атрибутов: для футболок, например, производитель А может описать цвет и посадку, а производитель Б – цвет и рисунок. В таком случае без дополнительной разметки атрибуты «посадка» и «рисунок» не позволяют работать со всей базой, и нужен способ восполнить пробелы. Кроме того, многих атрибутов, способных помочь при поиске, обычно вообще нет, или они недостаточно детальны. Дополнительное преимущество автоматической разметки над ручной – возможность предоставить ее покупателю для автоматического поиска одежды по фотографии.

Для разметки товаров можно нанимать людей, но при большом регулярно обновляющемся инвентаре временные и материальные расходы на такое решение относительно высоки, поэтому в сфере коммерции растет интерес к автоматическим методам разметки.

В данной работе предлагается автоматическая система разметки одежды человекочитаемыми атрибутами (тегами) длины, дизайна и цвета, способная обрабатывать несколько предметов одежды на изображении. Приводятся особенности разработки каждого компонента алгоритма и работы с данными. Архитектура итогового решения обеспечивает обработку произвольного количества предметов одежды на изображении и предусматривает изменения в наборе предсказываемых тегов, а также возможность применения к задачам тегирования на других данных.

Методы анализа изображений

Методы анализа изображений принято делить на классические и нейросетевые, распространившиеся позже. Классические методы основываются на поиске характерных элементов на изображении и вычислении статистик по заранее заданному вручную алгоритму. Основная проблема в разработке и применении этого семейства методов – высокая сложность разработки новых алгоритмов или настройки существующих для решения новой задачи: каждый случай требует сложного алгоритмического описания и индивидуальной настройки параметров. Нейросетевые методы выигрывают у классических возможностью одновременной оптимизации извлечения наиболее эффективных признаков и их обработки. Это обеспечило их лидирующие позиции в задачах анализа изображений, таких как классификация, сегментация и детектирование. В связи с этим в данной работе предпочтение отдано нейросетевым алгоритмам.

Для обработки нескольких объектов на изображении необходимо их обнаруживать и различать, что является задачей детектирования – определения регионов, где есть интересующие предметы. Для экспериментов выбрана библиотека Tensorflow Object Detection API [1] с использованием алгоритма Single-Shot Detector [2] и архитектуры MobileNet v2 [3] в силу простоты использования, наличия предобученных моделей и высокой производительности.

Сама задача тегирования в данной работе решалась как задача классификации. Для экспериментов выбраны классификаторы на основе архитектур Inception v4 [4] и MobileNet v2. Данные архитектуры показали лучшие для своего размера результаты по качеству классификации и позволили проверить зависимость качества от размера сети.

Для задачи распознавания цвета нужно дополнительно определять, какие пикселы принадлежат интересующему предмету – для этого решается задача семантической сегментации. Для экспериментов в силу лучшего на момент обзора качества и высокой производительности выбран алгоритм DeepLab v3+ [5].

Описание решения

В результате проектирования и экспериментов была установлена схема обработки изображений, показанная на рисунке. Изображение сначала проходит процедуру детектирования для устранения лишних деталей и нормализации масштабов, затем выделенная область проходит через алгоритм сегментации, и в итоге по региону и маске происходит предсказание атрибутов длины, дизайна и цвета. Преимущества этого подхода в возможности обработки произвольного количества объектов на изображении и дополнения или замены модулей классификации. Это позволяет дополнять систему тегирования новыми классификаторами по мере их разработки или переходить к решению задачи тегирования на других данных.

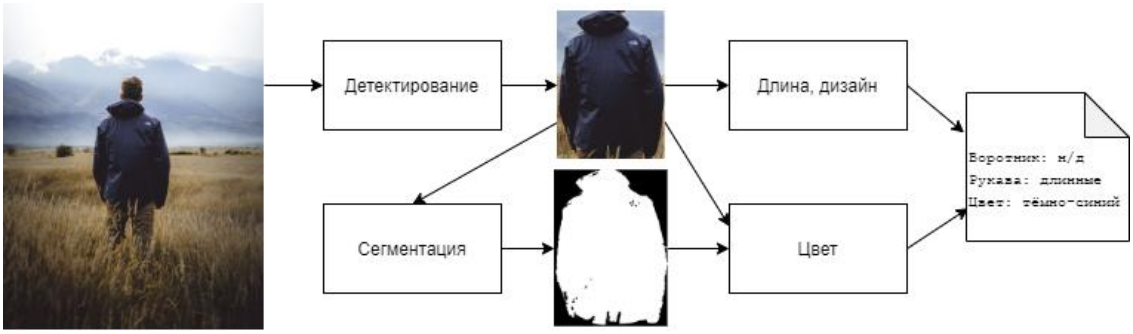

Схема обработки изображений
Image processing sequence

Детектирование

Для обучения и тестирования алгоритма детектирования использовалась подвыборка из коллекций DeepFashion [6] и ModaNet [7]. Выделенный набор данных был вручную провалидирован в целях исключения ошибок разметки. Также были исключены изображения с низкой контрастностью, слишком близким ракурсом, с сильным пересечением предметов и предметами, частично выходящими за кадр. Очистка производилась полуавтоматическим способом: на основе аннотаций изображений строились фильтрующие правила, и корректность отсечения проверялась вручную. Полученные данные балансировались по категории.

В коллекции DeepFashion присутствуют аннотации обрамляющих прямоугольников (bounding boxes), ключевых точек (landmarks), категории одежды и прочих атрибутов, не используемых в этой работе.

При детальном изучении выяснилось, что обрамляющие прямоугольники построены по ключевым точкам, что позволяет размножать данные поворотом и перспективными искажениями и генерировать по ним новые корректные прямоугольники.

Много полезной информации удалось извлечь из ключевых точек. Ключевые точки описывают положение частей одежды в кадре: например, концов рукавов, плеч, подола. Каждая точка описана ее координатами на изображении и флагом видимости. Если точка не видна, но ее положение можно угадать по окружающим частям одежды, она указывается как неви-

димая, но с заполненными координатами. Если точку указать невозможно – она не применима к данному виду одежды, как положение молнии на футболке, или не попала в кадр, ее координата не заполняется. Используя положения, видимость и заполненность точек на многих изображениях удалось пополнить имеющуюся информацию и проверить качество съемки. По длине прилегания области, содержащей объект, к краю изображения установлено, на каких изображениях съемка произведена слишком близко, и объект не полностью попадает в кадр. По пересечению областей, содержащих предметы, установлено, на каких фотографиях два человека стоят друг за другом. По видимости и заполненности ключевых точек воротника получена информация, спереди снят объект или сзади. Если на человеке размечено несколько предметов одежды, они расположены один над другим, и информацию о ракурсе можно распространить на них. Также убраны слишком мелкие по площади предметы. Наконец, отсеяны темные и низкоконтрастные фотографии, где слишком низкая средняя интенсивность и много шума или малый разброс значений.

Коллекция ModaNet содержит разметку областей, где находятся предметы, в виде попиксельных масок, что позволяет использовать эти данные для обучения как детектированию, так и сегментации. К этой коллекции также применимо размножение с помощью пространственных искажений. Для того чтобы эту коллекцию объединить с DeepFashion, произведена переразметка категорий одежды и убраны неактуальные категории, такие как шарфы, сумки и обувь. Как и в случае с DeepFashion, отсеяны слишком крупные и мелкие предметы.

Коллекции объединены и сбалансированы по категориям и положению объектов в кадре, чтобы избежать переобучения под конкретный стиль фотографии. Итоговый результат сохранен как последовательность изображений в бинарном формате TF Records, требуемом со стороны TF OD API. Балансировка происходила посредством равновероятного семплирования из каждой категории с размножением изображений из маленьких подмножеств. В частности, данный подход к организации данных гарантирует разнообразие изображений на каждом шаге обучения, что ускоряет сходимость алгоритма и улучшает качество итоговой модели.

Обучение производилось с применением метода transfer learning [8] с модели, обученной для детектирования на коллекции MS COCO [9], с поиском сложных примеров с высокой уверенностью ошибочного предсказания (> 0,99) и размножением данных в процессе обучения случайной обрезкой и отражениями по горизонтали.

Классификация атрибутов длины и дизайна

На примере атрибутов длины и дизайна изучены особенности обработки визуально близких атрибутов с упорядоченностью и без нее. Для обучения применялась коллекция FashionAI [1], состоящая из студийных фотографий, на каждой из которых позирует один человек в различных окружениях. На каждой фотографии размечен ровно один атрибут длины или дизайна элемента одежды. Разметка выполнена в формате «да / нет / наверное» для всех атрибутов; для случаев, когда атрибут не применим или соответствующая ему часть одежды не видна, предусмотрено значение атрибута «Invisible».

Разметка была переведена в более строгий формат: оставлены только объекты, где однозначно определяется атрибут (ровно один ответ «да» либо ровно один ответ «наверное» при отсутствии «да»). Результат проверен вручную. Выяснилось, что значение «Invisible» для атрибутов длины использовалось очень неаккуратно: как правило, оно не означало, что длину определить нельзя, и фотографии с ним ничем не отличались от фотографий с прочими значениями атрибута длины. Для лучшей интерпретируемости результатов такие образцы были убраны из коллекции.

[1] tianchi.allyun.com. Fashionai global challenge. URL: https://tianchi.aliyun.com/competition/rankingList.htm?spm=5176.11409106.5678.4.6510d751R1Fce0&raceId=231649.

Обучение производилось с предобученных на ImageNet [10] моделей с настройкой параметров обучения по процедуре, описанной в [11]. Для определения оптимальной стратегии обработки данных проведена серия экспериментов. Сравнивались модели, предсказывающие атрибуты длины и дизайна вместе и по отдельности. За счет присутствия в кадре только одного человека удалось проверить влияние обрезки изображения вокруг интересующего предмета одежды по обрамляющему прямоугольнику. Результаты показывают, что выгодно производить предсказания длины и дизайна в отдельных нейросетях и что обрезка имеет положительное влияние на качество предсказаний – предположительно, из-за отсечения лишних деталей на фоне. Также установлено, что влияние качественных изменений (обрезка) превосходит выгоду от перехода к многократно более тяжеловесной архитектуре. Дополнительно обнаружено, что при предсказании визуально близких упорядоченных атрибутов (длины) в большинстве случаев ошибка приводит к предсказанию соседнего класса, что позволяет для задачи поиска по каталогу объединить соседние классы и получить выдачу высокой точности.

Сегментация

Сегментация в данной работе является вспомогательным компонентом для задачи определения цвета. Получаемые маски используются для сбора пикселов, принадлежащих интересующему предмету.

Задача семантической сегментации обычно решается как задача попиксельной классификации для определения, какие пикселы принадлежат фону, а какие – предмету. У данного подхода есть недостаток, заключающийся в том, что на самом деле нас интересует не просто точность классификации, а точность соответствия предсказанной маски эталонной: как хорошо воспроизводится форма предмета, есть ли в предсказанной маске пропуски, выходит ли она за границы эталона. Численно это выражается мерой Жаккара – отношением площади пересечения к площади объединения, обладающей инвариантностью к масштабу и хорошо согласующейся с восприятием качества локализации. Для непосредственной оптимизации меры Жаккара предложена функция потерь [имени] Ловаса [12]. Для экспериментов данная функция потерь была добавлена в библиотеку для DeepLab v3+. Итоговая функция потерь – взвешенная сумма многоклассовой кросс-энтропии и функции Ловаса. Для обучения и тестирования использована коллекция ModaNet с той же предобработкой, что для задачи детектирования.

Анализ предсказаний базового решения с «наивной» предобработкой данных (обрезка ровно по обрамляющему прямоугольнику с приведением к размеру входа нейросети) показал несколько характерных особенностей в ошибках предсказания: часто происходит предсказание семантически близкого класса (например, куртка вместо футболки); контрастные элементы могут предсказываться как фон; большинство ошибок в определении края объекта случается около границ изображения.

Первым потенциальным улучшением стала проверка того, как влияет сохранение соотношения сторон на качество предсказаний. В базовом решении эта информация терялась из-за обрезки ровно по прямоугольнику. Для устранения этого прямоугольники увеличиваются до квадрата – таким образом, предметы любой формы могут быть поданы на вход нейросети без искажений. Из-за увеличения обрамляющего прямоугольника возникает два дополнительных вопроса. 1. Где расположить предмет внутри нового прямоугольника? Рассмотрены два варианта: по центру и в случайном месте. 2. Что делать, если часть прямоугольника не помещается на изображении: обрезать не поместившиеся части, закрасить черным, заполнить значениями на краю изображения или заполнить отражением? Лучше всего сработало центрирование и заполнение отражением.

Следующим проверено предположение об улучшении предсказаний за счет увеличения обрамляющего прямоугольника для обеспечения дополнительного контекста около границ. Увеличение происходило на 10, 25 и 50 %, лучше всего сработало увеличение на 25 %.

Далее проверено влияние добавления функции потерь Ловаса к кросс-энтропии. Лучше всего сработало сложение кросс-энтропии и функции Ловаса с весом 0,5.

Также проверено использование поиска сложных пикселов для акцентирования обучения на областях, где предсказание происходит наименее надежно. При использовании этой процедуры градиент подсчитывался по 1, 10, 50 % изображения с наивысшим значением функции потерь. Ни один из вариантов не дал улучшения относительно обучения без применения поиска, поэтому в финальном варианте поиск не используется.

Некоторые ошибки, такие как небольшие пробелы внутри масок или ошибки классификации, достаточно просто описываются и стабильно воспроизводятся, чтобы можно было их исправить алгоритмически. Небольшие пробелы в масках и ложные срабатывания устраняются путем переразметки аномальной области в класс окружающего предмета или фона. Ошибки классификации устраняются на основе знания ожидаемого класса объекта, предсказываемого детектором, характерных ошибок (футболки путаются с куртками, но не со штанами) и характерного распределения площадей для каждого ожидаемого класса: перекрашивая области, наиболее выбивающиеся из распределения, в целевой класс, можно понять, какая из областей на изображении была классифицирована неправильно.

Предсказание цвета

Задача предсказания цвета в данной работе решается как задача классификации. Основная сложность состоит в том, что работа происходит с малой коллекцией изображений с большим количеством оттенков: изображения распределены приблизительно по 1 000 оттенков, которые объединены в 100 промежуточных и 20 основных цветов. Прямое решение задачи классификации при этом дает неудовлетворительный результат: много ошибок, причем при ошибках часто предсказывается совершенно непохожий цвет. Для исправления этой проблемы решено проверить методы обучения распознавания сходства, распространенные в задаче распознавания лиц. Задача распознавания лиц обладает схожими проблемами: небольшое количество данных, высокое количество классов (уникальных лиц), высокие требования к качеству предсказаний. Обучение распознаванию сходства заставляет нейросеть, во-первых, располагать схожие объекты ближе в пространстве признаков, что должно решить проблему с предсказанием непохожих цветов в случае ошибки, а также улучшить надежность предсказаний в целом. Для тестирования выбраны показывающие лучшие результаты методы Triplet loss [13] и ArcFace [14]. Принцип работы обоих методов – группировать похожие объекты и обеспечивать настраиваемую минимальную дистанцию до непохожих, что повышает устойчивость к вариациям во входных данных и помогает в задаче классификации.

В базовом решении обнаружено, что цвет фона может влиять на предсказание. Для улучшения качества предсказаний протестировано применение масок сегментации и аугментации фона. Маски либо добавлялись к изображению четвертым каналом, либо использовались для закраски всех пикселов, не относящихся к изображению, в черный цвет. Лучше сработал первый вариант. В качестве аугментации применялась замена фона на случайное изображение, геометрические искажения и искажения цвета. Положительный эффект дали все аугментации, кроме искажения цвета.

Поскольку разметка цвета иерархическая (есть три уровня подробности разбиения), протестированы подходы к обучению модели предсказывать цвета по более мелкому разбиению и затем предсказывать по более грубому, а также к обучению модели предсказывать цвета сразу на нескольких уровнях подробности. Лучше сработал последний вариант, когда модель одновременно учится упорядочивать раскраски по схожести по самому мелкому разбиению и предсказывать цвета по мелкому, среднему и крупному разбиениям.

В дополнение к Triplet loss протестирован декодирующий модуль для предсказания цвета как абсолютной величины для обеспечения дополнительной информации во время обучения. Положительного влияния на качество предсказаний этим методом добиться не удалось.

Также в дополнение к Triplet loss протестировано использование декодирующих модулей, реализующих Arcface. Наилучший результат получен при замене всех модулей классификации на Arcface и сохранении Triplet loss в качестве ограничения на векторы признаков.

Заключение

Разработана и реализована расширяемая архитектура системы тегирования. Полученная система позволяет обрабатывать изображения одежды и обеспечивать поиск по каталогу по фотографии.

Рассмотрены особенности реализации каждого этапа обработки изображения. Исследованы особенности подготовки данных и работы с визуально близкими упорядоченными и неупорядоченными атрибутами, влияние локализации на качество классификации, приемы улучшения качества классификации в случае большого количества классов.

Полученная архитектура, как и приемы предобработки изображений, настройки параметров обучения и анализа предсказаний, универсальна и может применяться в будущих системах, решающих смежные задачи на других данных.

Список литературы / References

1. **Huang J., Rathod V., Sun C. et al.** Speed / accuracy trade-offs for modern convolutional object detectors. In: arXiv:1611.10012, 2016.
2. **Wei Liu, Dragomir Anguelov, Dumitru Erhan et al.** Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, 2016, p. 21–37.
3. **Sandler M., Howard A., Zhu M. et al.** Mobilenetv2: Inverted residuals and linear bottlenecks. In: arXiv:1801.04381, 2018.
4. **Szegedy C., Ioffe S., Vanhoucke V., Alemi A.** Inception-v4, inceptionresnet and the impact of residual connections on learning. In: arXiv:1602.07261, 2016.
5. **Chen L.-C., Zhu Y., Papandreou G. et al.** Encoder-decoder with atrous separable convolution for semantic image segmentation. In: arXiv:1802.02611, 2018.
6. **Z. Liu, P. Luo, S. Qiu et al.** Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 1096–1104.
7. **Zheng S., Yang F., Kiapour M. H., Piramuthu R**. Modanet: A large-scale street fashion dataset with polygon annotations. In: arXiv:1807.01394, 2018.
8. **Tan C., Sun F., Kong T. et al.** A survey on deep transfer learning. In: arXiv:1808.01974, 2018.
9. **Tsung-Yi Lin, Michael Maire, Serge Belongie et al.** Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014. Eds. David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars. Cham, Springer International Publishing, 2014, p. 740–755.
10. **Russakovsky O., Deng J., Su H. et al.** Imagenet large scale visual recognition challenge. In: arXiv:1409.0575, 2014.
11. **Page D.** On hyperparameter tuning and how to avoid it. URL: https://myrtle.ai/ how-to-train-your-resnet-5-hyperparameters/.
12. **Berman M., Triki A. R., Blaschko M. B.** The lov´asz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: arXiv:1705.08790, 2017.

13. **Schroff F., Kalenichenko D., Philbin J.** Facenet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, Jun.
14. **Deng J., Guo J., Xue N., Zafeiriou S.** Arcface: Additive angular margin loss for deep face recognition. In: arXiv:1801.07698, 2018.

Материал поступил в редколлегию
Received
25.06.2020

Сведения об авторах

Малышев Александр Григорьевич, разработчик, ООО «Экспасофт» (Новосибирск, Россия)
a.malyshev@expasoft.tech

Полыгалов Александр Сергеевич, ведущий разработчик, ООО «Экспасофт» (Новосибирск, Россия)
a.polygalov@expasoft.tech

Алямкин Сергей Анатольевич, технический директор, ООО «Экспасофт» (Новосибирск, Россия)
s.alyamkin@expasoft.com

Information about the Authors

Alexander G. Malyshev, developer, Expasoft LLC (Novosibirsk, Russian Federation)
a.malyshev@expasoft.tech

Alexander S. Polygalov, lead developer, Expasoft LLC (Novosibirsk, Russian Federation)
a.polygalov@expasoft.tech

Sergey A. Alyamkin, CTO, Expasoft LLC (Novosibirsk, Russian Federation)
s.alyamkin@expasoft.com