

УДК 004.272 004.032.26

Применение метода экзemplярной нормализации в моделях на основе глубокого обучения для задачи повторной идентификации

© 2020 г. **А. В. Яценко***, **, аспирант; **С. А. Родионов***, канд. физ.-мат. наук;
А. С. Потапов*, доктор техн. наук

*ООО "Сингулярности Лаб", Санкт-Петербург

**Университет ИТМО, Санкт-Петербург

E-mail: yashenkoxciv@gmail.com

Поступила в редакцию 10.03.2020

DOI:10.17586/1023-5086-2020-87-08-52-57

В данной работе исследованы модели для решения задачи повторной идентификации пешеходов на основе глубокого обучения. Рассмотренные модели используют только изображения пешеходов, не требуя дополнительной пространственно-временной информации, которая в разных случаях может различаться. Предлагается модификация исследованных моделей на основе нормализации карт признаков. Представлены результаты оценки качества моделей, демонстрирующие улучшение, в лучшем случае на 10%, относительно результатов моделей без применения предложенной модификации.

Ключевые слова: глубокое обучение, повторная идентификация пешеходов.

Коды OCIS: 150.1135, 100.4996

ВВЕДЕНИЕ

Повторная идентификация пешеходов остаётся сложной задачей, играющей важную роль при автоматизации видеонаблюдения. Хорошее решение данной задачи позволит отслеживать перемещение пешеходов, используя информацию, получаемую с разных камер. Более общее решение может применяться и для объектов других видов.

Ранее наилучшие результаты показывали системы повторной идентификации, использующие вручную заданные признаки изображений и традиционные методы компьютерного зрения. Однако качество таких систем было не вполне достаточным для использования на практике. В последнее время возник интерес к решению этой задачи с помощью моделей глубокого обучения, что связано с общим раз-

витием методов глубокого обучения и появлением новых наборов данных.

Практическое применение такого рода моделей предполагает, что система способна эффективно работать в новых условиях, например, с новым набором камер. И тогда как модели повторной идентификации на основе глубокого обучения существенно превзошли традиционные методы при выполнении обучения и тестирования на изображениях с одного и того же набора камер, при применении на новом наборе камер нейросетевые модели, предобученные на изображениях с других камер, до сих пор могут проигрывать классическим методам, которые не подгоняются к конкретным камерам [1]. В то же время классические методы требуют «ручного» подбора параметров алгоритма для хоть сколько-нибудь улучшения

качества работы в новых условиях, а для моделей глубокого обучения требуется дообучение или адаптация к новым условиям (domain adaption).

В данной работе рассматривается задача повторной идентификации и две модели на основе глубоких свёрточных сетей для её решения. Предлагается модификация на основе нормализации карт признаков, которая позволяет улучшить точность повторной идентификации для одной из моделей на новых наборах данных, отличных от тех, на которых модель обучалась. Проводится анализ результатов и сравнение моделей, обученных на нескольких наборах данных.

ПОСТАНОВКА ЗАДАЧИ

Для решения задачи повторной идентификации используются данные в виде $D = \{X_i, Y_i\}_{i=1..N}$, где X_i — это изображение пешехода, а Y_i — метка или идентификатор пешехода, N — количество пар изображение–метка, представленных для обучения. Для каждого изображения может присутствовать дополнительная информация в виде номера кадра F_i и идентификатора камеры C_i , но в данной работе эта информация не используется. Как правило, данные D разделяются на $D_{s\text{-train}}$ — набор данных, используемый для обучения, и $D_{s\text{-test}}$ — набор, используемый для тестирования модели и обычно содержащий новые идентификаторы, но полученные с того же набора камер, что и $D_{s\text{-train}}$.

Необходимо получить отображение изображения пешехода X_i в идентификатор этого пешехода — Y_i . Для этого могут применяться различные методы и модели [2–4]. Один из базовых подходов заключается в обучении глубокой классификационной сети, которая позволяет оценить условную вероятность идентификатора при заданном изображении $p_w(y | x)$ и известных параметрах сети w , оптимизируя функцию перекрёстной энтропии: $L(w) = -E [y \ln(p_w(y | x))]$. Классификационная сеть представляет собой композицию функций $s(f(x | w_f) | w_c) \rightarrow y$ аппроксимирующих $p_w(y | x)$. При этом обучается латентное представление, $z = f(x | w_f)$, где w_f — параметры модели, порождающей латентное представление, а w_c — параметры классификационного слоя. Выходы классификационной модели

будут строго соответствовать классам в обучающей выборке и не смогут назначать новые идентификаторы, например, для пешеходов из $D_{s\text{-test}}$.

В этой связи, хотя сеть и обучается классифицировать изображения обучающей выборки, при тестировании вместо выхода её классификационного слоя используется латентное представление. При этом классификация производится с помощью поиска ближайшего соседа в пространстве латентного кода. Таким образом, задача заключается не только в том, чтобы получить отображение $X_i \rightarrow Y_i$, но и в получении такого латентного кода, который можно эффективно применять для классификации новых данных. Получается, что задачу можно сформулировать и в виде кластеризации образов обучающей выборки по латентному представлению z . В таком случае, латентный код может оптимизироваться непосредственно, например, с применением триплет-функции ошибки [5]: $L(z_a, z_p, z_n)_w = \max(0, (z_a - z_p)^2 - (z_a - z_n)^2 + m)$.

На практике требуется применять модели повторной идентификации к новым данным D_t , которые получены с новых камер. Далее к D_t применяется тот же алгоритм, что и к $D_{s\text{-test}}$. Однако D_t не содержит идентификаторов пешеходов. Для сопоставления изображений пешеходов применяется алгоритм ближайшего соседа, использующий латентное представление, полученное как выход предпоследнего слоя каждой модели.

Классические методы повторной идентификации «непредвзяты» относительно конкретного набора данных кроме набора изображений, на которых оценивалось качество работы метода, и поэтому могут работать сравнительно хорошо на новых данных. Однако может потребоваться изменение параметров алгоритма для наилучшей работы, что можно считать «переобучением» в случае с классическими подходами. Модели глубокого обучения часто превосходят классические методы на обучающем наборе данных, но точность повторной идентификации на новых данных сильно снижается. Для повышения точности повторной идентификации при работе с D_t применяют методы адаптации: 1) применение к новым условиям без какой-либо модификации модели (cross-dataset evaluation); [5–7] 2) дополнение новых данных метками и дообучение модели

(pseudo labeling); [8–10] 3) «выравнивание» пространства признаков на основе новых данных (latent space alignment); [8] 4) использование моделей переноса стилей для дополнения данных (data augmentation) [11].

Далее рассматриваются две передовые модели применительно к задаче повторной идентификации. Предлагается модификация, позволяющая сопоставить эффективность и способность к адаптации к новым условиям для каждой модели.

МОДИФИКАЦИЯ МОДЕЛЕЙ

Рассматриваемые модели содержат глубокую свёрточную сеть ResNet-50 как модель для выделения основного признакового представления. В качестве модификации предлагается добавить экземплярную нормализацию [12] после третьего блока ResNet-50, заменяя, таким образом, слой пакетной нормализации.

Экземплярная нормализация применяется в основном в моделях переноса стиля [13, 14]. Работа слоёв экземплярной нормализации схожа с работой пакетной нормализации и заключается в применении следующих модификаций к картам признаков, подающимся на вход: $y_{tijk} = (x_{tijk} - \mu_{ti}) / (\sigma_{ti}^2 + \varepsilon)^{1/2}$, где μ_{ti} — среднее каждой карты признаков каждого изображения, а σ_{ti}^2 — среднее квадратическое отклонение, y_{tijk} — результат экземплярной нормализации, x_{tijk} — входные карты признаков. Индексы $tijk$ соответствуют номеру примера в пакете данных t , глубине, высоте и ширине входной карты признаков i, j, k соответственно.

Так как для визуальной идентификации пешехода используется информация о его стиле, в статье [10] предлагается использовать экземплярную нормализацию в классификационных моделях для повторной идентификации пешеходов. Предполагается, что такой подход позволяет лучше выделить информацию о стиле пешехода. Нужно отметить, что в применении к задаче идентификации пешеходов слово «стиль» означает отличительные визуальные черты отдельного пешехода, которые включают тип, фасон, цвет одежды и некоторые другие.

Результаты, представленные в работе [15], демонстрируют, что добавление экземплярной нормализации позволяет улучшить точность отождествления пешеходов. Однако экс-

перименты, проведённые в этой работе, основаны на одной архитектуре классификатора и нескольких наборах данных. Эксперименты, проведённые в данной статье, основаны на подходе из [15], но этот подход применяется к нескольким моделям и наборам данных.

МОДЕЛЬ ЭКЗЕМПЛЯРНОЙ ПАМЯТИ ДЛЯ АДАПТАЦИИ К НОВЫМ ДАННЫМ

В работе [10] предлагается модель, которая для адаптивного обучения использует экземплярную память (ЭП). В основе модели лежит глубокая свёрточная сеть ResNet-50 [16], которая состоит из блоков вида свёрточный слой \rightarrow пакетная нормализация [17] \rightarrow нелинейная активация. Распространённой практикой является предобучение такого рода моделей на наборе данных ImageNet [18]. Далее модель специализируется для конкретной задачи.

Основа модели в виде ResNet-50 позволяет построить представление размерности 4096 для входного изображения. Далее это представление передаётся в полносвязный слой, который выдаёт представление размерности 4096 — \mathbf{z} , которое используется для обучения классификационного слоя и помещается в экземплярную память.

Обучение параметров классификационного слоя w_c производится с помощью целевой функции перекрёстной энтропии. Экземплярная память представляет хранилище для \mathbf{z} , полученных по изображениям из данных для адаптации. Каждому изображению изначально устанавливается уникальная метка. При обучении на данных для адаптации оптимизируется персональная инвариантность, то есть каждое представление классифицируется как уникальное. Представление, инвариантное камере, обучается с помощью дополнения данных, для чего используется перенос представления между камерами, где применяется модель из работы [14]. Для формирования общего представления для отдельных пешеходов производится поиск ближайшего соседа в пространстве \mathbf{z} для каждого представления из целевого набора данных (данных для адаптации), далее минимизируется расстояние между найденным представлением и тем, что использовалось при поиске.

МОДЕЛЬ «МЕШКА ТРЮКОВ» ДЛЯ ПОВТОРНОЙ ИДЕНТИФИКАЦИИ

В работе [7] описывается модель, также основанная на ResNet-50, но обучающаяся по двум комплементарным функциям ошибки: классификации и триплет-функции [5].

Архитектура модели устроена таким образом, что выход сети ResNet-50 используется в триплет-функции ошибки. Далее следует пакетная нормализация, выход которой передается классификационному слою.

Данная модель не предполагает какой-либо процедуры адаптации к новым данным и может оцениваться только по выходу после пакетной нормализации.

Рассматриваемые модели схожи в том, что используют ошибку классификации при обучении и базовую сеть ResNet-50 для извлечения признаков. Однако результат первой можно отнести к методам адаптации 2-го, 3-го и 4-го класса. Работа второй из рассмотренных моделей относится к 1-му классу адаптации.

ЭКСПЕРИМЕНТЫ

В работе [15] указывается, что экземплярная нормализация позволяет улучшить точность повторной идентификации для новых наборов

данных. Этот подход был протестирован на базовой модели классификатора. Задача повторной идентификации подразумевает сравнение признаков представлений двух независимых снимков пешеходов и, как указывалось ранее, для этого используется алгоритм ближайшего соседа, работающий с латентными представлениями. Пешеходы в свою очередь отличаются стилевыми особенностями и положением, позой. Выравнивание контраста, которое выполняет экземплярная нормализация оказывается, полезной и позволяет значительно улучшить работу модели «мешка трюков» (МТ) на новых данных.

Для обучения моделей использовались наборы данных MARKET (МКТ) [19], MSMT17 (MSMT) [11], DUKE (DK) [20], CUHK03 (СК) [19], SYRI (SY) [20], SYSU (SU) [21], VIPER (VI) [22], WARD (WA) [2]. Все они содержат изображения пешеходов. Отличие заключается в количестве камер, ракурсе съёмки и самих пешеходах.

В табл. 1 представлены результаты, полученные при обучении моделей с дополнительным слоем экземплярной нормализации. Все модели протестированы на наборе данных MARKET.

Как видно из табл. 1 при дополнении модели экземплярной памяти слоем экземплярной

Таблица 1. Точность сопоставления на данных MARKET

Данные	МТ	МТ с ЭН	ЭП	ЭП с ЭН
DK	54,6	64,8	75,7	75,2
MSMT	–	–	75,3	–
DK + СК	69,8	76,9	76,1	76,2
MSMT + DK + СК + SY	78,4	78,3	78,4	–
MSMT + DK + СК + SY + SU + VI + WA	76,9	78,3	72,6	–

Таблица 2. Точность сопоставления на данных MSMT17

Данные	МТ	МТ с ЭН	ЭП	ЭП с ЭН
МКТ	15,0	23,9	–	–
DK	21,2	30,7	29,4	–
МКТ + DK	22,4	–	–	–
DK + СК	21,7	37,8	29,1	–
DK + СК + VI + WA	26,5	–	–	–
МКТ + DK + СК + SU	28,6	36,9	23,5	–
МКТ + DK + СК + SU + SY + VI + WA	–	–	22,6	–

нормализации улучшения получить не удаётся. Более того, с увеличением набора данных для обучения результаты модели (ЭП) только ухудшаются, а добавление экземплярной нормализации только незначительно улучшает точность. Из табл. 2 видно, что модель МТ с увеличением объёма данных для обучения улучшает точность и на новых данных, а добавление экземплярной нормализации и во все позволяет улучшить результат примерно на 10%. Тенденция, присущая модели экземплярной памяти, при обучении с большим набором данных, сохраняется и в этом случае, а различие в точности составляет 13,4%. Важно отметить, что процедура и параметры обучения не менялись в зависимости от используемого набора данных.

Для того, чтобы объяснить этот результат, потребовалось визуально изучить наборы данных MARKET и MSMT17. В результате, можно заключить, что MSMT17 содержит изображения с большей изменчивостью, чем изображения в MARKET. Кроме того, конфигурация камер в MSMT17 сильно отличается. Некоторые пешеходы не представлены на некоторых камерах. Поэтому метод экземплярной памяти, размечающий целевой

набор данных, склонен к ошибке второго рода при выборе позитивных и отрицательных примеров.

ЗАКЛЮЧЕНИЕ

В работе была рассмотрена задача повторной идентификации пешеходов. Рассмотрены две базовые модели для решения этой задачи. Для улучшения работы рассмотренных сетей предложена модификация, которая заключается в добавлении слоя экземплярной нормализации.

Добавление соответствующего слоя позволило улучшить точность повторной идентификации. Представлены таблицы для сравнения эффективности базовой и модифицированной модели. Кроме того, показано, что добавление данных для обучения позволяет значительно улучшить результат при тестировании на MARKET. Однако такой подход не срабатывает при тестировании на MSMT17 и улучшает точность незначительно, а в некоторых случаях и ухудшает.

Предложенный метод, а именно добавление экземплярной нормализации после нескольких блоков глубокой сети, можно эффективно применять для обучения модели «мешка трюков».

ЛИТЕРАТУРА

1. *Poongothai E., Suruliandi A.* Survey on colour, texture and shape feature analysis for person re-identification technique // *Advances in Vision Computing: An International Journal (AVC)*. 2016. V. 3. № 3. doi: 10.5121/avc.2016.3303
2. *Jaderberg M., Simonyan K., Zisserman A.* Spatial transformer networks // *Advances in Neural Information Processing Systems*. 2015. V. 28. P. 2017–2025.
3. *Potapov A., Shcherbakov O., Zhdanov I.* HyperNets and their application to learning spatial transformations // *27th International Conference on Artificial Neural Networks*. Rhodes, Greece. October 4–7, 2018. P. 476–486.
4. *Sabour S., Frosst N., Hinton G.E.* Dynamic routing between capsules // *Advances in Neural Information Processing Systems*. 2017. V. 30. P. 3856–3866.
5. *Hermans A., Beyer L.* In defense of the triplet loss for person re-identification [электронный ресурс] 2017. — Режим доступа: <https://arxiv.org/abs/1703.07737>, свободный.
6. *Zhang F.* View confusion feature learning for person re-identification // *The IEEE International Conference on Computer Vision (ICCV)*. Seoul, Korea. October 27 – November 2, 2019. P. 6639–6648.
7. *Luo H., Gu Y.* Bag of tricks and a strong baseline for deep person re-identification [электронный ресурс] 2019. — Режим доступа: <https://arxiv.org/abs/1903.07071>, свободный.
8. *Potapov A., Rodionov S., Latapie H.* Metric embedding autoencoders for unsupervised cross-dataset transfer learning [электронный ресурс] 2018. — Режим доступа: <https://arxiv.org/abs/1807.10591>, свободный.
9. *Rodionov S., Potapov A., Latapie H.* Improving deep models of person re-identification for cross-dataset usage // *Artificial Intelligence Applications and Innovations*. 2018. P. 75–84. doi: 10.1007/978-3-319-92007-8_7
10. *Zhong Z., Zheng L.* Invariance matters: exemplar memory for domain adaptive person re-identification // *Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA. June 16–20. 2019. P. 598–607.

11. *Wei L., Zhang S.* Person transfer GAN to bridge domain gap for person re-identification // Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA. June 18–22. 2018. С. 79–88.
12. *Ulyanov D., Vedaldi A.* Instance normalization: The missing ingredient for fast stylization [электронный ресурс] 2016. — Режим доступа: <https://arxiv.org/abs/1607.08022>, свободный.
13. *Zheng Z., Yang X.* Joint discriminative and generative learning for person re-identification // Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. June 16–20. 2019. P. 2138–2147.
14. *Zhong Z., Zheng L.* Camera style adaptation for person re-identification // Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA. June 18–22. 2018. P. 5157–5166.
15. *Jia J., Ruan Q.* Frustratingly easy person re-identification: generalizing person re-ID in practice // 30th British Machine Vision Conference. Cardiff, UK. September 9–12, 2019. P. 117–131.
16. *He K., Zhang X.* Deep residual learning for image recognition // Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, Nevada, USA. June 26 –July 1. 2016. P. 770–778.
17. *Ioffe S., Szegedy C.* Batch normalization: accelerating deep network training by reducing internal covariate shift // International Conference on Machine Learning. Lille, France. July 7–9, 2015. P. 448–456.
18. *Deng J., Dong W.* ImageNet: A large-scale hierarchical image database // IEEE Conference on Computer Vision and Pattern Recognition. Miami Beach, FL, USA. June 20–21, 2009. P. 248–255.
19. *Li W., Zhao R.* DeepReID: Deep filter pairing neural network for person re-identification // IEEE Conference on Computer Vision and Pattern Recognition. Columbus. Ohio, USA. June 24–27. 2014. P. 152–159.
20. *Bak S., Carr P.* Domain adaptation through synthesis for unsupervised person re-identification // Computer Vision — ECCV. 2018. P. 193–209.
21. *Zheng L., Shen L.* Scalable person re-identification: A benchmark // IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. 2015. P. 1116–1124.
22. *Ristani E., Solera F., Zou F.* Performance measures and a data set for multi-target, multi-camera tracking [электронный ресурс] 2016. — Режим доступа: <https://users.cs.duke.edu/~tomasi/papers/ristani/ristaniBmvt16.pdf>, свободный.
23. *Wu A., Zheng W.-S., Yu H.-X., Shaogang Gong, Jianhuang Lai.* RGB-infrared cross-modality person re-identification // IEEE International Conference on Computer Vision. Venice, Italy. October 22–29, 2017. P. 5390–5399.
24. *Gray D., Brennan S.* Evaluating appearance models for recognition, reacquisition, and tracking // IEEE International Workshop on Performance Evaluation for Tracking and Surveillance. Rio de Janeiro, Brazil. October 14, 2007. V. 3. № 5. P. 1–7.