Рекомендации товаров по набору фотографий на основе нейросетевых методов агрегации векторов признаков изображений

К.В. Дёмочкин[1], А.В. Савченко[1]

[1]Национальный исследовательский университет Высшая школа экономики, Большая Печерская 25/12, Нижний Новгород, Россия, 603155

Аннотация. Исследуется задача определения интересов пользователей для рекомендательных систем на основе набора фотографий заказанных или просмотренных ранее товаров. Исследованы нейросетевые методы агрегации векторов признаков изображений, извлеченных с помощью глубоких нейронных сетей. Предложен новый двухэтапный алгоритм, в котором на первом этапе происходит дообучение сверточной нейронной сети, а на втором этапе при помощи последовательного применения методов агрегации neural aggregation network и context gating вычисляется взвешенная сумма векторов признаков всех изображений товаров, ассоциированных с одним пользователем. Экспериментальное исследование для набора данных Amazon Products показало, что F1-мера предложенного подхода оказывается более чем на 20% выше F1-меры традиционного усреднения векторов признаков.

1. Введение

В настоящее время все большее развитие получают визуальные рекомендательные системы (visual recommender system) [1, 2, 3, 4], которые выявляют пользовательские предпочтения (предсказывают категории интересов пользователя) с помощью анализа *набора* фотографий товаров, купленных или просмотренных пользователем ранее. Такие системы могут использоваться либо самостоятельно, либо в качестве составной части существующих рекомендательных систем онлайн-магазинов для быстрой и надежной оценки потенциальных категорий продуктов, которые могут заинтересовать покупателя, основываясь на информации, полученной, например, с приложения на смартфоне.

Категории продуктов, в которых заинтересован один и тот же покупатель, зачастую связаны между собой. Поэтому в данной работе для решения задачи предложено использовать современные методы обучаемой агрегации для извлечения зависимостей между разными фотографиями товаров, принадлежащими одному пользователю. Такие методы разрабатывались для задач распознавания видео данных, например, для верификации и идентификации лиц на видео [5, 6, 7]. Среди них одними из наиболее успешных являются нейро-агрегационный модуль (neural aggregation network) [8] и шлюз контекста (context gating) [9], который применялся в решении, победившем на престижном конкурсе Youtube 8M Large-Scale Video Understanding challenge 2017.

Таким образом, цель настоящей работы состоит в комбинировании известных подходов взвешенной агрегации признаков, использованных ранее в задачах анализа

видеопоследовательностей, для построения визуальных рекомендательных систем. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области распознавания изображений и рекомендательных систем.

2. Предложенный подход

Задача предсказания интересов пользователей по фотографиям заключается в следующем: требуется предсказать наиболее интересные для пользователя категории продуктов на основании набора изображений товаров, который данный пользователь покупал ранее. Каждый продукт принадлежит к одной или более из D категорий. Другими словами, необходимо оценить апостериорные вероятности заказа товаров из D заданных категорий. Для обучения системы для каждого из N пользователей задана коллекция $\{X_n(m)\}$, m=1,2,…, M_n, из M_n изображений продуктов, которые были куплены этим пользователем. Предполагается, что на каждом изображении запечатлён один товар, и каждой фотографии соответствует бинарный вектор $\mathbf{y}$ размерности D, в котором элемент с индексом d равняется 1, если продукт с изображения относится к категории d, и 0 в противном случае.

В настоящей работе предлагается следующий алгоритм, состоящий из двух этапов. На первом этапе применяется традиционный подход с переносом обучения (transfer learning) [10] для извлечения характерных признаков, состоящий в добавлении классификатора к базовой глубокой сверточной нейронной сети, предварительно обученной на большом объеме данных, например ImageNet [11]. В частности, с учетом наложения указанных во введении ограничений на возможности реализации системы на мобильном устройстве, в данной работе применяется нейросетевой модели MobileNet [12]. При этом предлагается разбить все обучающее множество из N наборов изображений на два непересекающихся подмножества размера N_1 и N_2. Первое подмножество будет используется для дообучения (fine-tuning) подходящих признаков. Эта дообученная (fine-tuned) модель затем используется для получения векторов признаков $\mathbf{x}_n(m)$ размерности K для каждого изображения из N_2.

На втором этапе предлагается вычислить итоговый вектор признаков $\mathbf{x}_n$ размерности K, описывающего n-го пользователя, как взвешенную сумму векторов признаков каждой фотографии $\mathbf{x}_n(m)$:

$$\mathbf{x}_n = \sum_{m=1}^{M_n} w(\mathbf{x}_n(m))\mathbf{x}_n(m), \quad (1)$$

где веса w могут зависеть от признаков $\mathbf{x}_n(m)$. Обычно применяются одинаковые веса, что приводит к традиционному усреднению векторов признаков [8]. Однако в данной работе рассматриваются способы обучения весов в (1) на основе нейросетевых методов, в частности нейро-агрегационный модуль с использованием механизма внимания, который был изначально использован для распознавания лиц на видеорядах [8]. Дополнительно применяется context gating [9], модифицирующий норму получаемого вектора признаков (1) для выделения зависимости между категориями (определения часто встречаемых вместе категорий). В результате значения для близких категорий будут увеличены, если они часто встречаются вместе в одном наборе изображений. И, наоборот, для категорий, совместное присутствие которых маловероятна, веса снижаются.

Полная структура модели показана на рисунке 1. В ней агрегированные векторы (1) подаются в полносвязный (fully connected) слой с регуляризацией с помощью метода dropout (обнуления выбранных наугад выходов нейронов предыдущего слоя сети). Так как вектор $\mathbf{y}$ зачастую содержит несколько ненулевых значений из-за того, что товар может принадлежать больше, чем к одной категории, на выходном слое в используется логистическая сигмоида. В результате последний слой выдает оценки апостериорных вероятностей того, что d-я категория является значимой для данного пользователя.

3. Результаты экспериментов

Эксперименты проводились на подмножестве набора данных Amazon Product Data [13] 5-core «Home and Kitchen» (Рисунок 2), в котором находятся только те продукты, с которыми

взаимодействовали как минимум 5 уникальных пользователей и только те пользователи, которые взаимодействовали как минимум с 5 продуктами. Набор данных содержит 547700 записей о N=66519 уникальных пользователей, которые взаимодействовали с 28237 уникальными товарами из D=1000 категорий. Примерами категорий являются «Товары для приготовления», «Контейнеры для хранения», «Кофе» и другие. Для каждого пользователя имеются данные о том, какие товары покупал этот пользователь. Количество товаров на пользователя M_n варьируется от 5 до 40. В среднем у пользователей насчитывается 8 купленных товаров. Каждому пользователю был присвоен вектор **y** размерности D, в котором элемент с индексом d равен 0, если пользователь не взаимодействовал ни с одним товаром из категории d, в противном случае (если пользователь взаимодействовал хотя бы с одним товаром из категории d), соответствующий элемент вектора **y** равен 1.

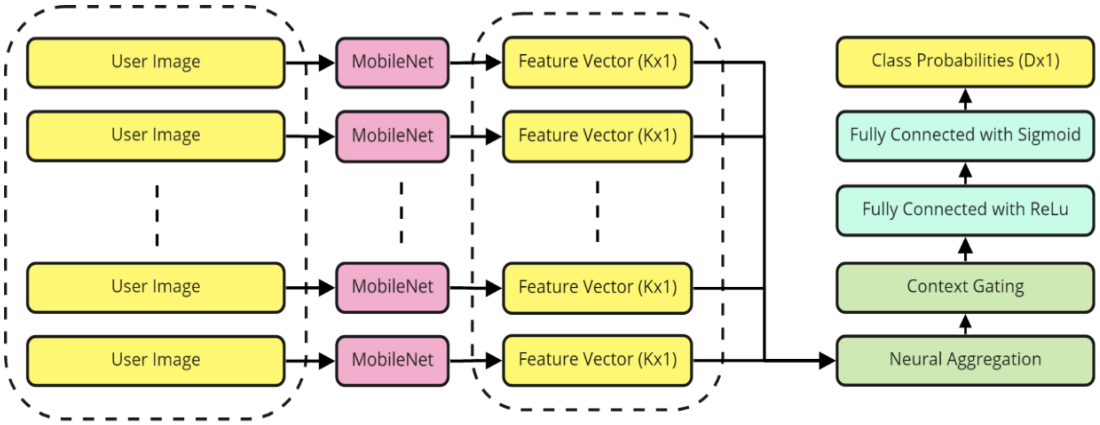

Рисунок 1. Предлагаемая нейросетевая архитектура для рекомендаций товаров по набору изображений.

Рисунок 2. Примеры изображений из набора данных “Home and Kitchen”.

Все нейросетевые модели реализовывались с использованием библиотеки Keras. На этапе предобработки изображения были приведены к размеру 224x224 пикселя, а RGB значения пикселей нормализованы в промежутке [-1;1] для соответствия формата входных данных MobileNet v1. Сначала на 70% изображений (N_1=0.7N=46563), выбранных наугад, была дообучена базовая модель MobileNet. В связи с тем, что каждый товар принадлежит всего нескольким категориям, целевые векторы получаются разреженными. Для уменьшения несбалансированности классов в качестве целевой функции использовалась взвешенная перекрестная энтропия. Были протестированы следующие значения веса положительного класса: {10, 36, 72, 140}. Наиболее точное решение получалось для веса положительного класса, равного 36. Экспериментально выявлено, что добавление скрытого слоя с 2048 нейронами и случайным занулением (dropout) с 50% вероятностью выходов предыдущего слоя приводит к наиболее высокой точности распознавания одиночных изображений. В процессе обучения сначала были зафиксированы веса 22 первых слоев, после чего модель дообучалась с помощью оптимизатора ADAM со скоростью обучения (learning rate) 0.001, beta_1 = 0.9, beta_2

= 0.999 в течение 10 эпох. После этого веса всех слоев дообучались еще 20 эпох со скоростью обучения (learning rate) = 0.0001.

Исследовались три способа агрегации векторов признаков: традиционное усреднение (Average), нейро-агрегационный модуль (Neural Aggregation), и предложенная последовательная комбинация двух известных методов (Neural Aggregation + Context Gating). В первом случае вычислялось среднее значение векторов признаков. Во втором были соединены два блока внимания (attention block) согласно статье [8]. В предложенном подходе к двум последовательно соединенным блокам внимания был добавлен context gating слой [9], который принимает вектор признаков (1) и динамически взвешивает его элементы, используя обучаемые коэффициенты.

После того, как векторы признаков всех изображений пользователя агрегируются в один вектор $\mathbf{x}_n$ (1), он подается на полносвязный слой с 2048 нейронами, после чего производится предсказание значимости категорий с помощью полносвязного слоя с сигмоидальной функцией активации. Для обучения весов взвешенной суммы использовалось 70% от второй выборки размера N_2, а на других 30% пользователей тестировались алгоритмы после обучения. В качестве целевой функции также использовалась взвешенная перекрестная энтропия с весом положительного класса, равным 36. Модель обучалась с помощью оптимизатора ADAM с learning rate = 0.001, beta_1 = 0.9, beta_2 = 0.999. Зависимость F1-меры от количества рекомендаций k продемонстрирована на Рисунке 3.

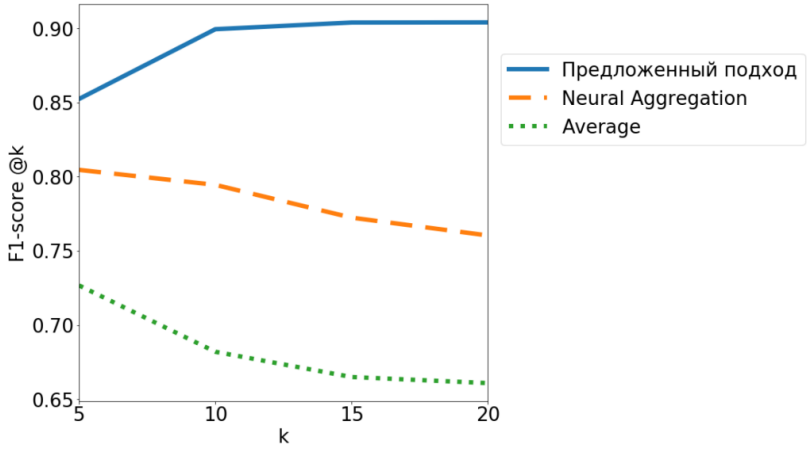

Рисунок 3. Зависимость F1-меры от количества рекомендаций *k*.

В таблице 1 представлены значения точности и полноты precision@k и recall@k [14]. Лучшие результаты выделены жирным шрифтом.

Таблица 1. Precision@k и Recall@k для различных подходов к агрегации.

k	Метод	Precision@k	Recall@k
5	Average	0.704867	0.749925
	Neural Aggregation	0.772574	0.839458
	Предложенный подход	**0.792203**	**0.922438**
10	Average	0.797340	0.595867
	Neural Aggregation	0.901716	0.710123
	Предложенный подход	**0.91846**	**0.881151**
15	Average	0.815469	0.561431
	Neural Aggregation	0.932418	0.710123
	Предложенный подход	**0.942565**	**0.868210**
20	Average	0.820141	0.553453
	Neural Aggregation	0.943513	0.636783
	Предложенный подход	**0.947498**	**0.864384**

Стоит отметить, что самый высокий показатель F1-меры был достигнут в предложенной комбинации Neural Aggregation [8] с Context Gating [9]. Так, наша F1-мера оказалась на 12-35% выше по сравнению с традиционным усреднением векторов. Добавление шлюза контекста (Context Gating) к нейро-агрегационному модулю позволило улучшить качество предсказаний на 5-14%.

4. Заключение

В настоящей статье исследовалось применение различные подходов обучаемой агрегации, использованных ранее в анализе видеоданных, для предсказания предпочтений пользователей на основании изображений товаров, купленных этими пользователями. Экспериментально показано, что с помощью нейро-агрегационного модуля [8], к которому был добавлен шлуюз контекста (Context Gating) [9] достигаются результаты на 34% лучше, чем простое усреднение векторов (Рисунок 3, Таблица 1).

Основным направлением для дальнейшего исследования является создание на основе предложенного подхода полноценной мобильной рекомендательной системы для рекомендации значимых категорий пользователям, исходя из изображений на их мобильном устройстве. Также планируется провести сравнительный анализ эффективности предложенного метода и традиционных рекомендательных систем, таких как колаборативная фильтрация (collaborative filtering) и факторизационные машины (factorization machines) [15]. Наконец, необходимо провести эксперименты на других открытых наборах данных, например на наборе данных AmazonFashion, в котором представлена информация о покупках пользователями различных предметов одежды.

5. Литература

[1] Hidasi, B. Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations / B. Hidasi, M. Quadrana, A. Karatzoglou, D. Tikk // Proceedings of the 10th ACM Conference on Recommender Systems, 2016. – P 241-248.

[2] Shankar, D. Deep learning based large scale visual recommendation and search for e-commerce / D. Shankar, S. Narumanchi, H.A. Ananya, P. Kompalli, K. Chaudhury // arXiv preprint, 2017. – ArXiv: 1703.02344.

[3] Andreeva, E. Extraction of Visual Features for Recommendation of Products via Deep Learning / E. Andreeva, D.I. Ignatov, A. Grachev, A. Savchenko // Proceedings of International Conference on Analysis on Images, Social Networks and Texts – AIST, 2018.

[4] Zhai, A. Visual discovery at pinterest / A. Zhai, D. Kislyuk, Y. Jing, M. Feng, E. Tzeng, J. Donahue, T. Darrell // Proceedings of the 26th International Conference on World Wide Web Companion – International World Wide Web Conferences Steering Committee, 2017. – P 515-524.

[5] Соколова, А.Д. Упорядочивание данных в системах видеонаблюдения на основе технологий глубокого обучения / А.Д. Соколова, А.В. Савченко // Сборник трудов ИТНТ-2018. – Самара: Новая техника, 2018. – С. 946-952.

[6] Никитин, М.Ю. Нейросетевая модель распознавания человека по лицу в видеопоследовательности с оценкой полезности кадров / М.Ю. Никитин, В.С. Конушин, А.С. Конушин // Компьютерная оптика. – 2017. – Т. 41, № 5. – С. 732-742. DOI: 10.18287/2412-6179-2017-41-5-732-742.

[7] Sokolova, A.D. Cluster analysis of facial data in video surveillance systems using deep learning / A.D. Sokolova, A.V. Savchenko // Computational Aspects and Applications in Large-Scale Networks - NET 2017. – New York LLC: Springer. – 2018. – Vol. 7. – P 113-120.

[8] Yang, J. Neural Aggregation Network for Video Face Recognition / J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, G. Hua // CVRP. – 2017. – Vol. 4(6). – P 7.

[9] Miech, A. Learnable pooling with Context Gating for video classification / A. Miech, I. Laptev, J. Sivic // arXiv preprint, 2017. – ArXiv: 1706.06905.

[10] Pan, S.J. A survey on transfer learning / S.J. Pan, Y. Qiang // IEEE Transactions on Knowledge and Data Engineering. – 2010. – Vol. 22(10). – P 1345-1359.

[11] Krizhevsky, A. ImageNet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G.E. Hinton // Advances in neural information processing systems. – 2010. – P. 1097-1105.
[12] Howard, A.G. MobileNets: Efficient convolutional neural networks for mobile vision applications / A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, H. Adam // arXiv preprint, 2017. – ArXiv: 1704.04861.
[13] McAuley, J. Image-based recommendations on styles and substitutes / J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015. – P 43-52.
[14] Herlocker, J.L. Evaluating collaborative filtering recommender systems / J.L Herlocker, J.A. Konstan L.G. Terveen, J.T. Riedl // ACM Transactions on Information Systems – TOIS. – 2004. – Vol. 22(1). – P 5-53.
[15] Zhou, Y. Large-scale parallel collaborative filtering for the netflix prize / Y. Zhou, D. Wilkinson, R. Schreiber, R. Pan // International Conference on Algorithmic Applications in Management, 2008. – P 337-348.

Благодарности
Статья подготовлена в результате проведения исследования (№ 17-05-0007) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2017-2018 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

Visual Product Recommendation using Neural Aggregation Network and Context Gating

K.V. Demochkin[1], A.V. Savchenko[1]

[1]National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis, Bolshaya Pecherskaya str. 25/12, Nizhny Novgorod, Russia, 603155

Abstract. In this paper we focus on the problem of user prediction in visual product recommender systems based on the given set of photos of products purchased by the user previously. We studied neural aggregation methods for image features extracted by the deep neural networks. We propose the novel two-stage algorithm. At first, the image features are learned by fine-tuning the convolutional neural network. At the second stage, we sequentially combine the known learnable pooling techniques (neural aggregation network and context gating) in order to compute a single descriptor for particular user as a weighted average of image features. It is experimentally shown for the Amazon product dataset that F1-measure for our approach is more than 20% higher when compared to conventional averaging of the feature vector.