Learning Visual Features from Product Title for Image Retrieval

Fangxiang Feng
School of Artificial Intelligence
Beijing University of Posts and Telecommunications
fxfeng@bupt.edu.cn

Tianrui Niu
School of Artificial Intelligence
Beijing University of Posts and Telecommunications
niwtr@icloud.com

Ruifan Li
School of Artificial Intelligence
Beijing University of Posts and Telecommunications
rfli@bupt.edu.cn

Xiaojie Wang
School of Artificial Intelligence
Beijing University of Posts and Telecommunications
xjwang@bupt.edu.cn

Huixing Jiang
Meituan-Dianping Group
Beijing, China
jianghuixing@meituan.com

ABSTRACT

There is a huge market demand for searching for products by images in e-commerce sites. Visual features play the most important role in solving this content-based image retrieval task. Most existing methods leverage pre-trained models on other large-scale datasets with well-annotated labels, e.g. the ImageNet dataset, to extract visual features. However, due to the large difference between the product images and the images in ImageNet, the feature extractor trained on ImageNet is not efficient in extracting the visual features of product images. And retraining the feature extractor on the product images is faced with the dilemma of lacking the annotated labels. In this paper, we utilize the easily accessible text information, that is, the product title, as a supervised signal to learn the features of the product image. Specifically, we use the n-grams extracted from the product title as the label of the product image to construct a dataset for image classification. This dataset is then used to fine-tuned a pre-trained model. Finally, the basic max-pooling activation of convolutions (MAC) feature is extracted from the fine-tuned model. As a result, we achieve the fourth position in the Grand Challenge of AI Meets Beauty in 2020 ACM Multimedia by using only a single ResNet-50 model without any human annotations and pre-processing or post-processing tricks. Our code is available at: https://github.com/FangxiangFeng/AI-Meets-Beauty-2020.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; **Visual content-based indexing and retrieval**; • **Information systems** → *Retrieval models and ranking*.

KEYWORDS

Visual feature learning; Bag of n-grams; Image retrieval; CNN; MAC

ACM Reference Format:
Fangxiang Feng, Tianrui Niu, Ruifan Li, Xiaojie Wang, and Huixing Jiang. 2020. Learning Visual Features from Product Title for Image Retrieval. In *28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3394171.3416296

1 INTRODUCTION

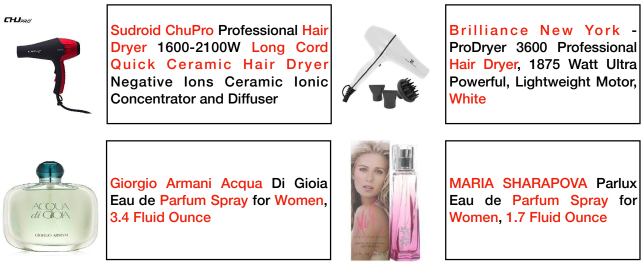
Sudroid ChuPro Professional Hair Dryer 1600-2100W Long Cord Quick Ceramic Hair Dryer Negative Ions Ceramic Ionic Concentrator and Diffuser
Brilliance New York - ProDryer 3600 Professional Hair Dryer, 1875 Watt Ultra Powerful, Lightweight Motor, White
Giorgio Armani Acqua Di Gioia Eau de Parfum Spray for Women, 3.4 Fluid Ounce
MARIA SHARAPOVA Parlux Eau de Parfum Spray for Women, 1.7 Fluid Ounce

Figure 1: Four products with images and titles. The titles provide rich visual information about the product, such as category, brand, color, and capacity.

Over the past decades, online shopping has received increasing attention. With the widespread of mobile devices, using the captured pictures to search for products has become a convenient way to obtain the target products. Currently, most e-commerce platforms provide the function of searching for products with image queries to improve the user's search experience. In academic literature, research on the task involving using an image query to search for images belongs to content-based image retrieval (CBIR).

The key point about CBIR is the feature extraction. In the early days, hand crafted low-level features, such as SIFT [17], PHOW [4], HOG [6] and GIST [18], are widely used in image retrieval tasks. In recent yeas, the visual features learned by deep neural networks for classification purposes are proved to have better performance. However, training such deep models usually requires huge well-annotated labels. Therefore, most existing methods leverage pre-trained models on other large-scale datasets, i.e. the ImageNet dataset, to extract visual features. However, due to the large difference between the product images and the images in ImageNet, the feature extractor trained on ImageNet is not efficient in extracting

MM '20, October 12–16, 2020, Seattle, WA, USA.

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3416296

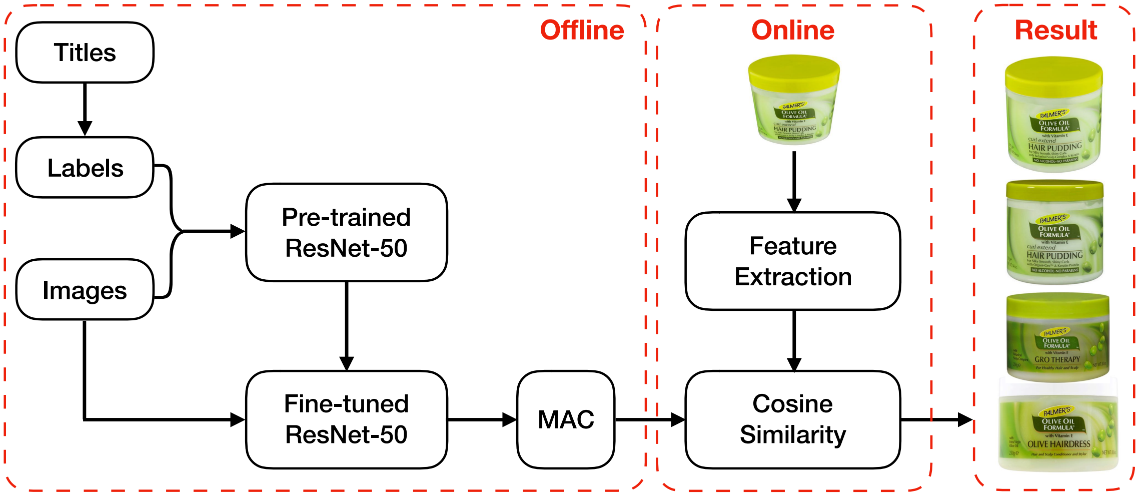

Figure 2: The pipeline of the proposed method.

the visual features of product images. And retraining the feature extractor on the product images is faced with the dilemma of lacking the annotated labels.

In order to solve this problem, in this paper, we utilize the easily accessible text information, that is, the product title, as a supervised signal to learn the features of the product image. As shown in Figure 1, the titles usually provide the concise and prominent description of the products, including rich visual information. They can therefore be used to guide the learning of visual features. Our pipeline for image retrieval is as follows. We first use the n-grams extracted from each product title as the label set of the product images to construct a dataset for image classification. Then, we fine-tuned a pre-trained model on this dataset. Next, the basic max-pooling activation of convolutions (MAC) [21] feature is extracted through the fine-tuned model. Finally, we perform the same feature extraction process on the query image, and retrieve the relevant images by the cosine similarity between features.

We evaluate our method on the Perfect-500K [5] dataset and achieve the fourth position in the Grand Challenge of AI Meets Beauty in 2020 ACM Multimedia by using only a single ResNet-50 [8] model without any human annotations and pre-processing or post-processing tricks.

2 RELATED WORK

CBIR is the procedure of automatically indexing images by the their visual content. And thus the similarity between the visual contents can be calculated for the retrieval of relevant images. Therefore, image features are crucial to the retrieval performance of the CBIR system. In order to extract representative image features, the previous CBIR methods use the hand-crafted low-level features, e.g. SIFT [17], PHOW [4], HOG [6] and GIST [18], which are extracted from the local patches. These local features are usually encoded by the quantization methods, such as Fisher Vector [20] and VLAD [11], to build the global image feature.

Recently, deep CNNs have demonstrated a superior performance over hand-crafted features on a wide range of computer vision tasks, such as image classification, object detection, and semantic segmentation. Inspired by these results, deep CNNs are also applied in the image retrieval domain [2, 7, 14–16, 21, 24, 26]. The primary method is to use the activations of the penultimate fully connected layer from the pre-trained CNNs as the image features. Since then, various pooling methods, e.g. SPoC [3], RMAC [23], RAMAC [16], MS-RMAC [13] and GRMAC [25], are proposed to combine features of the last convolutional layer. As a result of incorporating spatial information, these methods further boost the retrieval performance.

The retrieval performance is also determined by the dataset that the CNN is trained on. Intuitively, the CNN feature extractor trained on the target dataset has effective performance. However, most image retrieval datasets are small or lack label information. One popular method [1, 22] is to first obtain a pre-trained CNN on a large-scale well-annotated dataset, and then retrain the CNN on the target image retrieval dataset. Obviously, this method cannot cope with the situation where the target dataset does not contain annotated labels. In this paper, according to the fact that the product data contains titles, we explore the use of the information in the titles to solve this problem.

3 METHOD

The pipeline of our method for product retrieval task is illustrated in Figure 2. Evidently, the pipeline consists of two stages: the offline stage and the online stage. In the offline stage, we first construct the dataset for image classification by converting product titles into labels. Then, the pre-trained CNN model, i.e. ResNet-50, is fine-tuned on the dataset with the labels and the images. At last, the MAC features of the images to be retrieved are extracted. In the online stage, given an image query, we extract the MAC feature and calculates its similarity score with the retrieved images under the cosine similarity for ranking. We will elaborate the offline stage in the following parts.

The key idea of our method is the construction of the dataset for product image classification. In general, most products on the e-commerce website contain not only the images but also the titles which describe the most compelling information of the product. However, the titles are in the form of sentences, and it is difficult to directly utilize them for visual feature learning. Empirically, in the supervised CNN training, it is more effective to use discrete labels rather than sentences as the semantic supervised signals of images.

To that effect, we adopt a fairly straightforward strategy to convert sentences to discrete labels. In most cases, extracting bag-of-word (BOW) features from textual data is the simplest way to obtain discrete representations. However, the product title dataset is collected from a vertical domain rather than a general domain, and each title usually contains only one sentence. As a result, most titles have a similar pattern, that is, they focus on describing certain attributes, such as category, brand, and capacity. This is beneficial for obtaining the effective semantic words as labels. But, if a single word is used as the basic unit to build the label set, there will be a large overlap between two different product categories. This is not appropriate for fine-grained image retrieval that needs to retrieve exactly the same product. Moreover, the vocabulary generated by BOW features contains too many words, it therefore is not suitable as the label set. Using several consecutive words, that is phrase, as the unit to build the label set is a simple solution to alleviate this problem. This is because even if two phrases contain many common words, they represent two categories that have no correlation at all. Consequently, instead of using the “Bag of Words” features, we use “Bag of n-grams” features as the label of the product image. It is worth noting that n can have multiple values in one label set and each value is greater than 2.

Since an image may be associated with multiple labels in our dataset, we need to fine-tune the pre-trained CNN with the multi-label loss function. As for the specific CNN model, in our experiment, we chose the classical ResNet-50 architecture, which performs well enough, while being simple and reproducible. Formally, given a training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1,...,N}$, where $\mathbf{x}_n \in \mathbb{R}^D$ denote the n^{th} image input and $\mathbf{y}_n \in \{0,1\}^K$ denote the n^{th} multi-label target. We denote the mapping from the input layer to the penultimate layer (pool5) of ResNet-50 by $f(\mathbf{x}_n; \theta)$, where θ is the parameter set of ResNet-50 except for the last layer. The last layer is a fully connected layer with parameter $\mathbf{W}$, which is a $2048 \times K$ matrix since $f(\mathbf{x}_n; \theta)$ produces image features of dimensionality 2048 from pool5 layer of ResNet-50. The softmax cross-entropy loss can be computed as:

$$-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{y_{nk}}{\sum_{k'=1}^{K} y_{nk'}}\log\frac{\exp(\mathbf{w}_k^\top f(\mathbf{x}_n;\theta))}{\sum_{k'=1}^{K}\exp(\mathbf{w}_{k'}^\top f(x_n;\theta))}$$

where $\mathbf{w}_k$ denote the k^{th} column of $\mathbf{W}$.

After the ResNet-50 is fine-tuned, we calculate the max-pooling activation of its last convolutional layer as the retrieval features.

4 EXPERIMENTS

4.1 Dataset

We evaluate our method on the Perfect-500K [5] dataset released for “AI Meets Beauty” challenge. The dataset contains more than 500,000 images of beauty and personal care items from 14 popular e-commerce sites. Each image corresponds to a product title. The validation set contains 100 image queries and the relevant images for each query. The challenge organizer also provided a private test set for participants to competition. The candidate sets of both the validation set and the private test set are composed of all images of Perfect-500K.

4.2 Evaluation metric

The retrieval performance is evaluated by Mean Average Precision (MAP). Given one image query and the first R top-ranked retrieved images, the average precision is defined as $\frac{1}{M}\sum_{r=1}^{R} p(r) \cdot rel(r)$, where M is the number of relevant image in the retrieved images, $p(r)$ is the precision at r, and $rel(r)$ presents the relevance of a given rank (one if relevant and zero otherwise). Thereafter, the metric MAP is obtained by averaging AP of all the image queries. We report MAP@7 (R=7) in our experiment.

4.3 Implementation Details

As with most work, we use the model trained on ImageNet as our pre-trained model. ImageNet dataset contains two versions: ImageNet-1k (1.28M images with 1,000 classes) and ImageNet-21k (14M images with about 21k classes). In general, the larger the amount of dataset, the more stable and robust the features learned by the model. Therefore, we chose the off-the-shelf model trained on the ImageNet-21k dataset.

In the fine-tuning stage, we need to first convert the product titles into discrete labels. As aforementioned, we extract the “Bag of n-grams” representation of the title. Here, we set the range of n-values to (3, 5). This means that the words in the vocabulary contain at least three consecutive words and at most five consecutive words. We build the vocabulary that only considers the first 3,000 ordered by n-gram frequency across all the titles. We use the CountVectorizer[1] class in scikit-learn [19] to achieve the vectors of titles. More specifically, we set the parameter max_features to 3000, the parameter ngram_range to (3,5), and the default values for other parameters. Since some titles in Perfect-500K dataset do not contain any n-gram in the vocabulary, their corresponding images cannot be used for training. After removing these images, there are 188,407 pairs of image and labels in the Perfect-500K dataset, and the mean number of labels per image is 2.784. We use the BiT toolkit[2] [12] recently released by Google Research to perform this retraining and the pre-training model is the basic BiT-M-R50x1.

4.4 Results

Table 1 presents the results of five deep CNN models with four different pooling methods on the validation set. The first two deep models, i.e. SEResnet-152 [9] and Densenet-201 [10], are trained on the ImageNet-1k dataset. These two models are used by the last year’s challenge champion. The third and fourth models are ResNet-50 trained on the ImageNet-1k and ImageNet-21k dataset, respectively. The last model is also the ResNet-50 pre-trained on the ImageNet-21k dataset but fine-tuned on the Perfect-500K dataset. We can make the following observations from the results.

Firstly, among the first three models which are trained on the ImageNet-1k dataset, the deepest model Densenet-201 shows the

[1] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[2] https://github.com/google-research/big_transfer

Table 1: Performance of different models with different pooling methods on the validation set.

Model	MAC	RMAC	RAMAC	RAMAC
SEResnet-152 ImageNet-1k	0.3616	0.3364	0.3458	0.3486
Densenet-201 ImageNet-1k	0.3868	0.3592	0.4035	0.4102
ResNet-50 ImageNet-1k	0.3253	0.2972	0.3107	0.3133
ResNet-50 ImageNet-21k	0.4158	0.4097	0.4150	0.4152
ResNet-50 Perfect-500k	**0.4980**	0.4484	0.4543	0.4534

Table 2: Performance on the private test set.

Rank	Team Name	Score(MAP@7)
1	USTC_NELSLIP	0.534848
2	GDUT611	0.43729
3	NTU-Beauty	0.405302
4	**Our team, bladefashion**	0.402402
5	toan	0.375517
6	TSST	0.362986
7	WHO_Knows	0.323124

best performance, while the shallowest model ResNet-50 achieves the worst performance. This indicates that the more complex the model, the more effective the features can be extracted. Secondly, the fourth model which is trained on the ImageNet-21k dataset, outperform all models trained on the ImageNet-1k dataset. This shows that the larger the scale of the dataset, the more effective the features can be extracted. Thirdly, fine-tuning the ResNet-50 using the Perfect-500k dataset can further significantly improve the retrieval performance. This demonstrates that the dataset we created can improve the effectiveness of the learned retrieval features. It has to be noticed that these three conclusions are all valid in all pooling methods.

Table 2 presents the overview of the performance on the private test set. We submit the fine-tuned ResNet-50 with the simple MAC feature since it has the best performance on the validation set. As a result, our method is ranked at the fourth place with MAP@7 of 0.402402. It is worth to mention that our score is only less than 0.003 lower than the third place, and we just use a single ResNet-50 model with the simplest pooling method, i.e. MAC.

4.5 Visualization

Figure 3 presents the retrieval results of two typical kinds of queries. The first four queries are pictures of real objects taken by the user, and the last four queries are advertisement pictures. As it shows, our method can retrieve products that match these queries from two typical application scenarios.

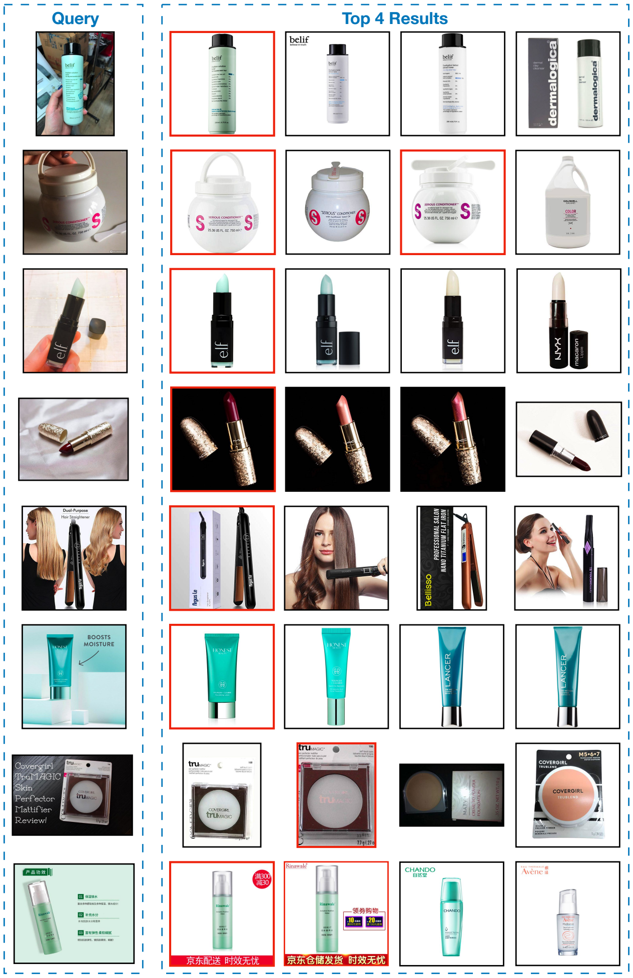

Figure 3: Visualization of the retrieval results. Relevant matches are shown with red bounding box.

5 CONCLUSION

In this work, we propose a simple yet effective method to learn the visual representation of product. Our key idea is to build the image classification dataset based on product image and title information. Empirically, the pre-trained deep CNN that fine-tuned on this dataset can improve the effectiveness of the features for product image retrieval task. In practice, in order to construct a better label set, we highly recommend integrating high-frequency n-grams with domain dictionary. According to our experience, for products whose title is Chinese, the proposed method can achieve a high-quality label set without performing the word segmentation.

6 ACKNOWLEDGMENTS

We thank the reviewers for their comments and suggestions. This paper is partially supported by NSFC (No. 61906018), MoE-CMCC “Artificial Intelligence” Project (No. MCM20190701) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Ahmad AlZu'bi, Abbes Amira, and Naeem Ramzan. 2017. Content-based image retrieval with compact deep convolutional features. *Neurocomputing* 249 (2017), 95–105.

[2] Artem Babenko and Victor S. Lempitsky. 2015. Aggregating Local Deep Features for Image Retrieval.. In *ICCV*. IEEE Computer Society, 1269–1277.

[3] Artem Babenko and Victor S. Lempitsky. 2015. Aggregating Local Deep Features for Image Retrieval.. In *ICCV*. IEEE Computer Society, 1269–1277.

[4] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. 2007. Image Classification using Random Forests and Ferns. In *ICCV*. IEEE, 1–8.

[5] Wen-Huang Cheng, Jia Jia, Si Liu, Jianlong Fu, Jiaying Liu, Shintami Chusnul Hidayati, Johnny Tseng, and Jau Huang. 2020. Perfect Corp. Challenge 2020: Half Million Beauty Product Image Recognition. https://challenge2020.perfectcorp.com/.

[6] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. 886–893.

[7] Yuanqiang Fang, Wengang Zhou, Yijuan Lu, Jinhui Tang, Qi Tian, and Houqiang Li. 2018. Cascaded Feature Augmentation with Diffusion for Image Retrieval.. In *ACM Multimedia*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 1644–1652.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385 cite arxiv:1512.03385Comment: Tech report.

[9] Jie Hu, Li Shen, , and Gang Sun. 2018. Squeeze-and-excitation networks.. In *CVPR*. IEEE Computer Society, 7132–7141.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks.. In *CVPR*. IEEE Computer Society, 4700–4708.

[11] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. 2012. Aggregating Local Image Descriptors into Compact Codes. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 9 (2012), 1704–1716.

[12] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2019. Conditional Generative Adversarial Nets. http://arxiv.org/abs/1912.11370 cite arxiv:1912.11370.

[13] Yang Li, Yulong Xu, Jiabao Wang, Zhuang Miao, and Yafei Zhang. 2017. MS-RMAC: Multiscale Regional Maximum Activation of Convolutions for Image Retrieval. 24, 5 (2017), 609–613.

[14] Jian Han Lim, Nurul Japar, Chun Chet Ng, and Chee Seng Chan. 2018. Unprecedented Usage of Pre-trained CNNs on Beauty Product.. In *ACM Multimedia*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 2068–2072.

[15] Zehang Lin, Haoran Xie, Peipei Kang, Zhenguo Yang, Wenyin Liu, and Qing Li. 2019. Cross-domain Beauty Item Retrieval via Unsupervised Embedding Learning.. In *ACM Multimedia*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2543–2547.

[16] Zehang Lin, Zhenguo Yang, Feitao Huang, and Junhong Chen. 2018. Regional Maximum Activations of Convolutions with Attention for Cross-domain Beauty and Personal Care Product Retrieval.. In *ACM Multimedia*. ACM, 2073–2077.

[17] David G. Lowe. 1999. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2 (ICCV)*. IEEE Computer Society, USA, 1150.

[18] A. Oliva and A. Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int'l Journal of Computer Vision* 42(3) (2001), 145–175.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[20] Florent Perronnin and Christopher R. Dance. 2007. Fisher Kernels on Visual Vocabularies for Image Categorization.. In *CVPR*. IEEE Computer Society. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2007.html#PerronninD07

[21] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features off-the-shelf: an Astounding Baseline for Recognition. (2014). http://arxiv.org/abs/1403.6382 cite arxiv:1403.6382Comment: version 3 revisions: 1)Added results using feature processing and data augmentation 2)Referring to most recent efforts of using CNN for different visual recognition tasks 3) updated text/caption.

[22] Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung. 2019. Deep Aggregation of Regional Convolutional Activations for Content Based Image Retrieval.. In *MMSP*. IEEE, 1–6.

[23] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations.. In *ICLR (Poster)*.

[24] Qi Wang, Jingxiang Lai, Kai Xu, Wenyin Liu, and Liang Lei. 2018. Beauty Product Image Retrieval Based on Multi-Feature Fusion and Feature Aggregation.. In *ACM Multimedia*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). ACM, 2063–2067.

[25] Jun Yu, Guochen Xie, Mengyan Li, Haonian Xie, and Lingyun Yu. 2019. Beauty Product Retrieval Based on Regional Maximum Activation of Convolutions with Generalized Attention.. In *ACM Multimedia*. ACM.

[26] Yi Zhang, Linzi Qu, Lihuo He, Wen Lu, and Xinbo Gao. 2019. Beauty Aware Network: An Unsupervised Method for Makeup Product Retrieval.. In *ACM Multimedia*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 2558–2562.