Rapid Clothing Retrieval via Deep Learning of Binary Codes and Hierarchical Search

Kevin Lin
Academia Sinica, Taiwan
kevinlin311.tw@iis.sinica.edu.tw

Huei-Fang Yang
Academia Sinica, Taiwan
hfyang@citi.sinica.edu.tw

Kuan-Hsien Liu
Academia Sinica, Taiwan
liukh@citi.sinica.edu.tw

Jen-Hao Hsiao
Yahoo! Taiwan
jenhaoh@yahoo-inc.com

Chu-Song Chen
Academia Sinica, Taiwan
song@iis.sinica.edu.tw

ABSTRACT

This paper deals with the problem of clothing retrieval in a recommendation system. We develop a hierarchical deep search framework to tackle this problem. We use a pre-trained network model that has learned rich mid-level visual representations in module 1. Then, in module 2, we add a latent layer to the network and have neurons in this layer to learn hashes-like representations while fine-tuning it on the clothing dataset. Finally, module 3 achieves fast clothing retrieval using the learned hash codes and representations via a coarse-to-fine strategy. We use a large clothing dataset where 161,234 clothes images are collected and labeled. Experiments demonstrate the potential of our proposed framework for clothing retrieval in a large corpus.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Deep learning, convolutional neural networks, image retrieval.

1. INTRODUCTION

Online recommendation has become a common way of satisfying our needs on clothes. Image-based clothing retrieval plays an important role for building an online recommendation system and offers a more convenient way for customers to identify their favorite items. For example, when people saw certain clothes they like on a website, the system is expected to recommend other websites containing visually similar clothing items to him/her for price or brand comparisons.

We introduce a deep learning approach for clothes image retrieval in a recommendation system. When a clothes image selected from a large corpus is presented as a query, the system can find similar clothes efficiently in the corpus based on visual appearance. The system architecture is shown in Figure 1. Recent progresses reveal that a deep convolutional network trained on the ImageNet dataset provides rich image representations [12]. Module 1 of our system takes advantage of this representation and uses it as initial weights for further refinement. In module 2, we adapt the image representations by fine-tuning the network on the new domain of a large corpus of clothes images. To facilitate quick search, we enhance the architecture of [12] to learn the binary (hash) codes directly by adding a latent attribute layer with sigmoid outputs in this module. Finally, module 3 is for online search. Fast clothing image retrieval can be obtained through the learned hash codes and representations via a coarse-to-fine strategy.

ICMR'15, June 23–26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749318.

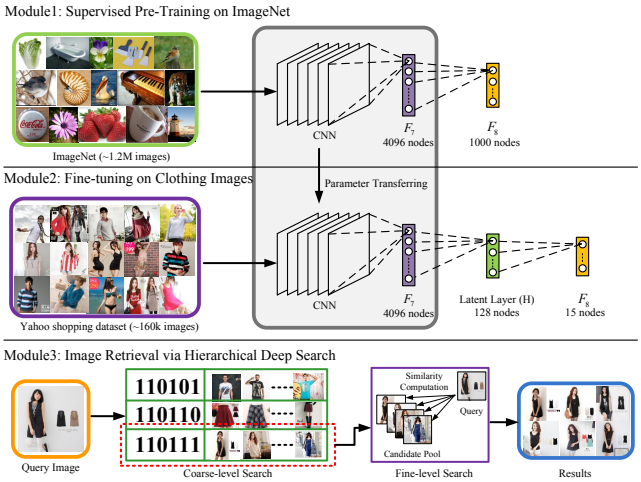


Figure 1: Our framework for rapid clothing retrieval via hierarchical deep search. We use the AlexNet [12] that is pre-trained on ImageNet for learning rich mid-level visual representations (module 1). To facilitate rapid retrieval, we add a latent layer to the network and have neurons in this layer learn hashes-like representations while fine-tuning it on the clothing dataset (module 2). Similar images are retrieved through a hierarchical search that utilizes the hashes-like and F_7 features (module 3).

To properly examine our system in this problem, a large clothing image dataset is needed. We employ a dataset from Yahoo Shopping that contains more than 160,000 clothes images to validate our approach. The dataset has several categories, which can show a great diversity of the images. The remainder of the paper is organized as follows. In Section 2, we review the related work. In Section 3, our proposed hierarchical deep search method is addressed. We conduct extensive experiments and provide discussion in Section 4. Finally, the concluding remarks are given in Section 5.

2. RELATED WORK

Liu et al. [14] targeted a system on occasion-oriented clothing recommendation. When a user inputs some occasion, e.g., sports,

wedding or conference, their system can recommend appropriate clothing from the user's own photo album, or pair the user-specified reference clothing (upper or lower body) with the most suitable one from online shops.

Recently, clothing retrieval has drawn more attention due to the huge fashion market. The system proposed by Chen et al. [2] describes clothing by semantic attributes, where a list of nameable clothing attributes are generated. They extract low-level features in a pose-adaptive way and associate with complementary features for learning attribute classifiers. Then, mutual connections between attributes are used to improve the performance from independent classifiers. Their system is evaluated on a clothing attribute dataset including 1,856 images with clothed people.

Liu et al. [15] consider cross-scenario clothing retrieval. They first locate a lot of human parts by a trained human detector. Then, from the part features, they get more reliable one-to-many similarities between the query daily photo and online shopping photos. They collect a daily photo dataset containing 4,321 upper and 4,068 lower body images and an online shopping dataset consisting of 8,293 upper and 8,343 lower body images for performance evaluation. Di et al. [6] consider stylistic visual elements in clothing retrieval. An attribute vocabulary is constructed on a fine-grained clothing dataset by human annotations. The vocabulary is used to train a fine-grained visual recognition system for clothing styles. They then use a Women's Fashion Coat dataset containing 2,092 images in the experiments. Liu et al. [13] introduce the problem of parsing fashion images from color-category tags, such as red-pants and yellow-shirt. They combine the human pose estimation, MRF-based color-category inference, and super pixel-level category classifier to achieve the purpose. Their method is validated on a colorful-fashion dataset containing 2,682 images labeled with pixel-level color-category. Though various researches have been done for clothing matching and retrieval, reliable features for clothes image representation is still demanding. Besides, most approaches are validated on relatively small datasets.

Deep learning has shown its powerfulness in learning good representations from a large corpus. With the introduction of large-scale image datasets and the advances in GPU computing, convolutional neural networks (CNNs) have received great attention recently. Krizhevsky et al. [12] demonstrate that a deep CNN trained on the ImageNet containing more than one million images achieves superior performance on a 1000-class recognition problem. Besides the success on image classification, CNN-based visual representation has also shown improved performance over handcrafted features on digit recognition [3] and pedestrian detection [19].

In addition to visual recognition tasks, deep learning has recently been applied to image retrieval. The work most relevant to ours is the *Deep Search* proposed by Huang et al. [10]. They train a deep CNN that incorporates a tree-structured clothes attributes for feature learning. To deal with cluttered background, they use a human detector to locate the upper body that is fed to the network. The learned features are used for clothing retrieval. Though both our and their work take advantage of deep learning, ours differs from theirs in two aspects: (1) our method does not rely on human detection; and (2) our network learns binary codes, in addition to visual features, to facilitate quick retrieval.

In this paper, we use deep CNN for learning discriminating feature representations capable of identifying different clothes images. Unlike previous approaches using only a small set (such as several thousand) of images for training and evaluation, this work focuses on the circumstance where a large corpus containing more than 160 thousands images for performance comparison. The corpus contains clothes images of heterogeneous types collected from different online shopping stores, including those that are backgroundless or of cluttered backgrounds, with or without human. Nevertheless, we find that favorable performance can still be achieved for such heterogeneous images with the powerful deep learning framework we used, where no pre-segmentation or human detection are performed. Details of our method are described in Section 3.

3. FRAMEWORK

Figure 1 shows our framework for rapid clothing retrieval. It contains (1) supervised pre-training on ImageNet for getting rich image representations, (2) fine-tuning the network on the clothing dataset to learn domain-specific features, and (3) rapid image retrieval through learned binary codes and mid-level representations.

3.1 Supervised Pre-training on ImageNet

We use the AlexNet [12] model from CAFFE [11]. It is pre-trained on more than 1.2 million images in the ImageNet dataset on the 1000-category image classification task. The network contains 8 layers: 5 convolutional layers followed by 3 fully connected layers (F_{6-8}), together with ReLU activations and max-pooling operations. Layers F_6 and F_7 are of 4,096 nodes. Outputs of layer F_7 is fed to a 1,000-way softmax in layer F_8 to produce a probability distribution over the object classes. It has been shown that the 4096-dimensional feature vector extracted from layer F_7 performs favorably against many other handcraft features [12, 18].

3.2 Learning Domain-specific Visual Features

Though the pre-trained CNN has learned rich mid-level image representations, it is non-optimal for a particular domain. Recent studies show that domain adaptations from the AlexNet are effective on transfer-learning tasks [7, 9]. To learn visual descriptors on clothing recognition, we follow this idea to fine-tune the network on the clothing dataset via back-propagation, and thus transfer the learned features to the clothing domain. The learned descriptors will serve as visual signatures for relevant image retrieval.

However, using the fine-tuned F_7 features alone for image retrieval is computationally intensive because it requires an exhaustive search that examines the similarities (based on L_2 norm) between a query and all images in a huge dataset. To facilitate the search critical to real-world retrieval applications, a practical way is to convert the features to binary-valued codes. Such binary strings can be fast compared via Hamming distance or hashing. The binary codes could be generated by random projection (such as locality sensitivity hashing (LSH) [5]) or learning-based techniques (such as [16, 17]). However, they need a second-stage learning or processing via the features constructed for the training data and could result in searching accuracy degradation. In this paper, we enhance the AlexNet structure so that the binary codes can be learned directly from the network and avoid a second-stage learning.

We assume that the final classification results are dependent to a number of h hidden concepts (or hidden attributes) with each attribute *on* or *off*. That is, we associate the input image to a binary-valued outputs (in $\{0,1\}^h$), and the classification depends on these hidden attributes instead of the feature vectors in $\mathbb{R}^{+4096}$ directly. To this end, we add a new layer between F_7 and F_8 as illustrated in the middle row of Figure 1. The role of this fully-connected layer (termed as *latent layer*, H) is to provide an abstraction of the F_7 and serves as a latent proxy that bridges mid-level features and semantics. H is then regulated by its succeeding classification layer F_8 that explicitly encodes semantic information of training samples. That is, images exhibiting similar activation patterns in the latent would have the same label. In our current setting, $h = 128$ and the neurons in H are activated by sigmoid functions so that

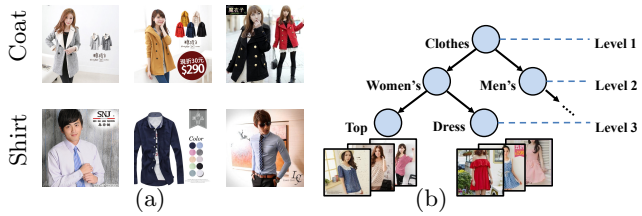


Figure 2: (a) Sample images from the Yahoo shopping dataset. The top row shows the coats images, and the bottom the shirts images from the constructed dataset. (b) Hierarchical labels of Yahoo shopping dataset. The category tree has three levels defined by Yahoo shopping sites.

the activations are approximated to $\{0, 1\}$. Hence, we can obtain binary codes for retrieval when the activations are binarized.

To fine-tune the network, the initial weights of the original layer are set as the pre-trained weights and the latent layer weights are randomly initialized. Stochastic gradient descent (SGD) are performed to refine the weights by maximizing the multinomial logistic regression objective. Through learning, the pre-trained weights evolve to a multi-layer function more suitable for the clothes domain. The latent layer weights can be viewed as evolved from the LSH approach (that uses random projection for binary coding) to a more favorable projection via supervision. With small modifications to a network model, our deep CNN simultaneously learns domain specific image representations and a set of hashing-like (binary coded) functions for rapid image retrieval, while it does not require constructing the hash codes in a separate stage nor dramatically altering the network model with different objective settings.

3.3 Retrieval via Hierarchical Deep Search

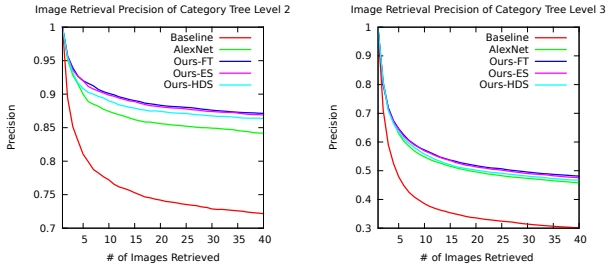
We deploy a coarse-to-fine, hierarchical search strategy for image retrieval. The coarse-level aims to quickly identify a small set of candidates sufficiently similar in terms of the binary-code representation. The rank of each candidate is determined through another search performed based on finer-level image representations.

Coarse-level Search. For an image, we extract the output of latent layer as its signature and obtain the binary code H^j ($j = 1 \cdots h$) by a threshold 0.5. Let $\Pi = \{I_1, I_2, \ldots, I_n\}$ denote a set of n training images and $\Pi_H = \{H_1, H_2, \ldots, H_n\}$ be the set of their corresponding binary codes. Given a query image q, the training samples with the Hamming distances to H_q less than a threshold form the pool of m image candidates, $P = \{I_1^c, I_2^c, \ldots, I_m^c\}$, which are further compared at the fine level.

Fine-level Search. We use the features of layer F_7 to identify the top k ranked images from the candidate pool P. Let V_q and V_i^P denote the feature vectors of the query image q and of the image I_i^c from the pool, respectively. We use the Euclidean distance between two feature vectors as a similarity measure. The similarity s_q^i between the query image q and the ith candidate image is defined as $s_q^i = \|V_q - V_i^P\|$. Each candidate I_i^c is ranked in ascending order by the similarity; hence, top k ranked images are identified.

4. EXPERIMENTS

We use a dataset consisting of plentiful clothing product images with a complete hierarchical category tree. It contains 161,234 clothes images collected by crawling the images from the Yahoo shopping sites. Figure 2(a) shows some samples images from our dataset. The images in each category demonstrate high levels of variation in human poses and appearances and are of noisy backgrounds or backgroundless. Some photos contain customer logos and advertisements, which are challenging to image classification and retrieval. The clothes-specific category tree is defined according to the product database from the Yahoo shopping sites. Each node in the tree represents different type of clothing, such as Men's, Women's, and so on. Figure 2(b) shows some examples of the image and with its hierarchical labels. The root of the tree represents the category Clothes. The second level consists of nodes Women's and Men's, while the third level contains a total of 15 leaf nodes representing 15 categories of clothing, such as Top, Dress, Coat, and so on. Table 1 exhibits the details of the 15 clothing categories and the number of images for each category.



(a) Category labels at level 2 (b) Category labels at level 3

Figure 3: Image retrieval precision of the Yahoo shopping sites. (a) and (b) compare the retrieval performance on different features. The fine-tuned networks perform favorably against the AlexNet and handcrafted features in spite of k and semantic labels. Ours-HDS that utilizes both latent and F_7 descriptors achieves comparable performance to others that consider F_7 features alone.

Evaluation Metrics. Our framework is designed to recommend clothes items similar to the one(s) customers like on a website by searching the entire dataset. Because of this inherent characteristic in the online recommendation system, we do not divide the data into training and test sets. We randomly select 1,000 query images from the dataset for the system to retrieve visually similar ones also from the dataset. We use a ranking based criterion [15] for evaluation. Given a query image q and a similarity measure, a ranking of n images in the dataset can be assigned. We evaluate a ranking of top k images with respect to a query image q by a precision:

$$Precision@k = \frac{\sum_{i=1}^{k} Rel(i)}{k}, \quad (1)$$

where $Rel(i)$ denotes the ground truth relevance between a query q and the i-th ranked image. Here, we consider only the category label in measuring the relevance so $Rel(i) \in \{0, 1\}$ with 1 for the query and the ith image with the same label and 0 otherwise.

Performance Comparison on Different Features. The visual features play an important role in retrieving relevant images. To investigate how features affect the performance, we compare the features learned in our model with others learned by different networks. We consider the following scenarios: (1) Baseline: DPM [8] is used for human detection and the one with the highest confidence score is selected as the representative. We extract handcrafted features including HOG [4], LBP [1], and color features. Concatenating them leads to a 22,459-dimensional feature vector. (2) AlexNet: F_7 features from the pre-trained network on ImageNet. (3) Ours-FT: F_7 features from the fine-tuned AlexNet on the clothing data with only the classification layer being replaced. (4) Ours-ES: F_7 from our network, fine-tuned on the clothing data with a new latent layer, and (5) Ours-HDS: F_7 and latent features from our network. Note that we conduct an exhaustive search based on L_2-norm when either basline or F_7 features alone are used in retrieval; a hierarchical search is performed when F_7 and latent features are considered.

Table 1: 15 clothing categories and the number of images for each category in the Yahoo shopping dataset.

Top	Camis	Dress	Formal Dress	Pant Suit	Coat	Skirt	T-shirt	Polo	Shirt	Vest	Jacket	Trendy Man	Suit	Sweater	**Total**
61052	6608	27826	3719	5523	12099	420	22273	3187	6361	2119	7395	686	1000	964	161234

Figure 4: Top 5 retrieved images by different features. The red bounding boxes show the representative objects by the DPM. The blue check marks indicate the query and retrieved images share the same label; the black crosses indicate otherwise.

Besides, to better understand the relationship between the learned representations and semantics, we vary the number of class labels while fine-tuning the network and evaluate retrieval performance on the learned features. We use the 2 labels from level 2 and 15 from level 3 for investigation.

Figure 3 shows the precision at k of the Yahoo shopping sites on handcrafted features and different CNN features learned with variable semantic labels. Generally, the CNN features either from AlexNet or fine-tuned models attain improved performance over the handcrafted ones regardless of how k and the number of semantic labels differ. All the fine-tuned models perform favorably against the AlexNet. This is because the fine-tuned ones are well-adapted to the new domain and learn domain-specific image representations. Among all the fine-tuned models, Ours-FT achieves the best performance. With a latent layer added to the network, Ours-ES achieves comparable performance to Ours-FT. Ours-HDS gives similar performance to Our-FT and Ours-ES (less than 1% drop in accuracy), but taking the least search time (more details in the computational time analysis).

Exemplar Retrieval Results. The top 5 images retrieved by different features are shown in Figure 4. Clearly, the baseline method retrieve images of great diversity. On the other hand, visually similar images can be retrieved by the deep features. The fine-tuned models retrieve more images with the same label as the query than AlexNet and the baseline. Notably, Ours-HDS's results are close (or identical) to Ours-ES's. This indicates the learned binary codes are informative and of discriminative power.

Computational Time. For an image with a size of 500×500 pixels, extracting CNN features takes about 0.06 seconds on a machine with Geforce GTX 780 GPU and 3 GB memory. The search is carried out on the CPU mode with a MATLAB implementation. Performing an exhaustive search using the handcrafted features for a query takes about 30 seconds; CNN features takes 5-6 seconds. In contrast, our hierarchical search takes 0.6 seconds. In sum, benefiting from the binary codes, our proposed hierarchical deep search is 10x and 50x faster than an exhaustive search using the F_7 and handcrafted features, respectively.

5. CONCLUSIONS

We have presented a deep CNN framework for rapid clothing retrieval in a recommendation system. To facilitate search, we add a latent layer to the network for learning binary codes that can be used to quickly identify a pool of image candidates for later refinement. Experimental results on a large clothing dataset demonstrate that our hierarchical, coarse-to-fine search attains a 10x and 50x speed-up in retrieval compared to an exhaustive search using CNN and handcrafted baseline features, respectively. In the future, we plan to embrace attributes, apart from category labels, to provide the collected clothes images with more comprehensive description and to test our method on the images with attribute annotations.

6. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. PAMI*, 28(12):2037–2041, 2006.

[2] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *Proc. ECCV*, 2012.

[3] D. C. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Proc. CVPR*, 2012.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. SCG*, 2004.

[6] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Proc. CVPR Workshops*, 2013.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proc. ICML*, 2014.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[10] J. Huang, W. Xia, and S. Yan. Deep search with attribute-aware deep network. In *Proc. ACM MM*, 2014.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[13] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Trans. Multimedia*, 16(1):253–265, 2014.

[14] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *Proc. ACM MM*, 2012.

[15] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proc. CVPR*, 2012.

[16] W. Liu, C. Mu, S. Kumar, and S.-F. Chang. Discrete graph hashing. In *Proc. NIPS*, 2014.

[17] M. Norouzi, D. J. Fleet, and R. Salakhutdinov. Hamming distance metric learning. In *Proc. NIPS*, 2012.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. CVPR*, 2014.

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. ICLR*, 2014.