Collaborative Feature Learning from Social Media

Chen Fang[1], Hailin Jin[2], Jianchao Yang[3], Zhe Lin[2]
[1]Department of Computer Science, Dartmouth College. [2]Adobe Research. [3]Snapchat.

Image feature representation plays an essential role in image recognition and related tasks. The current state-of-the-art feature learning paradigm is supervised learning from labeled data [3], which surpasses other well-known hand-crafted feature based methods [4, 5]. However, this paradigm requires large datasets with category labels to train properly, which limits its applicability to new problem domains where labels are hard to obtain.

In this paper, we ask an interesting research question: *Are category-level labels the only way for data driven feature learning?*

There is a surge of social media websites in the last ten years. Most social media websites such as Pinterest have been collecting content data that the users share as well as behavior data of the users. User behavior data are the activities of individual users, such as likes, comments, or view histories and they carry rich information about corresponding content data. For instance, two photos of a similar style on Pinterest tend to be pinned by the same user. If we aggregate the user behavior data across many users, we may recover interesting properties of the content. For instance, the photos liked by a group of users of similar interests tend to have very similar styles.

Approach. We propose a new paradigm for data driven image feature learning which we call *collaborative feature learning.* It is a major departure from the existing paradigms on feature learning such as supervised learning in that we do not rely on category labels at all. The main idea is to learn image features from user behavior data on social media. In particular, we use the user behavior data collected on social media to recover latent representations of individual images and learn a feature transformation from the images to the recovered latent representations.

The proposed approach is a framework that unifies latent factor analysis and deep convolutional neural network for image feature learning from social media. We focus on the simple form of user-item view data in this work to keep our feature learning framework general. Figure 1 provides a high-level overview of our approach. Given a set of content items $\mathcal{I} = \{I_1, \ldots, I_M\}$ and a set of users $\mathcal{U} = \{U_1, \ldots, U_N\}$, the corresponding user-item view data is in the format of a matrix between $\mathcal{I}$ and $\mathcal{U}$, which is denoted as $V \in \mathbb{R}^{M \times N}$. To handle the sparsity and noise in V, and extract compact latent information from it, we use collaborative filtering for implicit feedback data [2] and negative entry sampling [1] to decompose it into the product between the latent factors of content items and users. As the latent factors of content items encode rich information about the similarity between the content items, we then generate pseudo classes for the content items by clustering their corresponding latent factors using K-means. Deep convolutional neural network (DCNN) [3] is then trained based on these pseudo classes in the traditional supervised way. Finally, the trained DCNN can be used to extract features for content data. The details of our approach is covered in the main paper.

Dataset. To validate our new feature learning paradigm, we collect a large-scale image and user behavior dataset from Behance.net, which is a popular social media website for professional photographers, artists, and designers to share their work. We download about 1.9 million artistic images, and for each of the images we obtain the list of users who have viewed it, which results in 326 million view records from 1.9 million users. The density of the view matrix is about 0.0093%.

Experiments. We have done the following experiments to validate our feature learning paradigm.

- Since latent factors from view data should reveal some properties of the content data, we first study the characteristics of our learned latent factors and analyze the information captured. Our empirical study consists of a simple nearest neighbor retrieval experiment on the content items $\mathcal{I}$ under the representation of latent factors, therefore only information from view data (and no visual information) is used. As shown in the main paper, we find strong visual and semantic proximity between query images and their nearest neighbors (NNs), and the observation is consistent across the entire set.

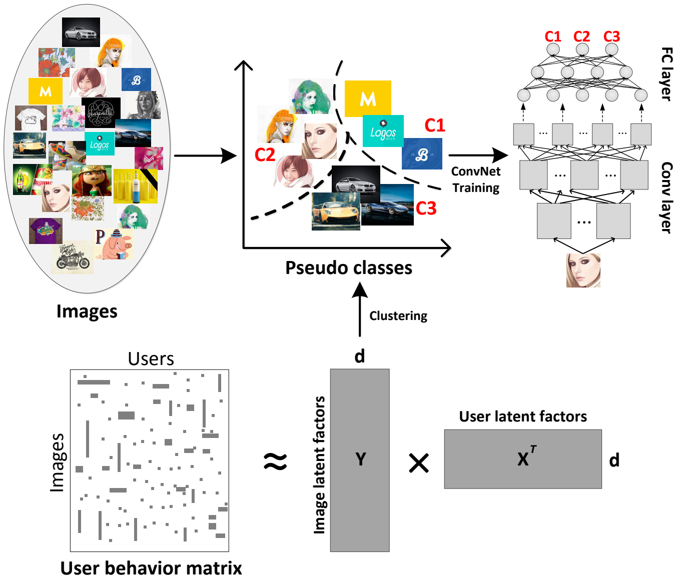

Figure 1: Approach overview.

- Then we investigate the learned neural network based visual feature, and evaluate its performance on the Behance dataset for image similarity/retrieval. We find that our learned feature significantly outperforms the state-of-the-art feature [3] both qualitatively and quantitatively.

- Finally, we apply the learned visual feature as generic image descriptors on standard benchmarks for style recognition and object classification. The experiment results show that our feature generalizes well to other datasets and tasks, and achieves competitive performance compared to other features.

[1] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup '11. In *Proceedings of KDD Cup 2011 competition, San Diego, CA, USA, 2011*, pages 8–18, 2012.

[2] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *International Conference on Data Mining (ICDM)*, 2008.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, 2012.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.

[5] David G Lowe. Object recognition from local scale-invariant features. In *international conference on Computer vision (ICCV)*, 1999.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.