Your Click Matters: Enhancing Click-based Image Retrieval performance through Collaborative Filtering

Deepanwita Datta
NTNU
Trondheim, Norway
ddatta.rs.cse13@itbhu.ac.in

Manajit Chakraborty
Università della Svizzera italiana
Lugano, Switzerland
chakrm@usi.ch

Aveek Biswas
UCSD
California, USA
a4biswas@ucsd.edu

Abstract

Image retrieval has been an active research area since the early days of computing. While ensemble, multimodal and hybrid methods coupled with machine learning has seen an upward surge replacing unimodal, heuristic-based methods; a rather new offshoot has been to identify new features associated with images on the web. One such feature is the 'click count' based on the clicks an image or its corresponding text gets in response to a query. Previous state-of-the-art methods have tried to exploit this feature by using its raw count and machine learning. In this paper, we build on this idea and propose a new collaborative filtering based technique to employ the click-log of users from the web to better identify and associate images in response to either a text or an image query. Experiments performed on a large scale publicly available standard dataset having genuine click logs from actual users corroborate the efficacy and significant increase in efficiency of our approach.

1 Introduction

Cross-media retrieval has proven to be an effective solution to search through enormous multi-varied datasets. A fairly common example of such cross-media retrieval is when an image is searched using a text query. Here the textual description of the image content acts as the text query. However, it is not a trivial task to illustrate a non-textual visual content using only text. Hence, a semantic gap is introduced between the user needs and the given (existing) descriptions. Although in literature, a significant amount of work has been devoted to correlating textual and visual information to bridge the semantic gap between the high-level information needs of users and commonly employed low-level features, it continues to be a major challenge. The existing state-of-the-art solutions to this challenge is two-pronged. Few of the existing works (Feng et al., 2014; Zhen and Yeung, 2012) stress on learning mapping functions whereas rest of the works explore the high-level semantic representation of modalities (Karpathy and Fei-Fei, 2015; Reed et al., 2016). Among these, semantic representation based approaches and deep learning based approaches have gained reasonable success. Deep Convolutional Neural Networks (CNNs) are used to learn the latent features and these learned features are utilized as visual and textual semantic representation in these models.

In ACM Multimedia 2015 MSR-Bing Image Retrieval Challenge[1], it was argued that the massive amount of click data from commercial search engines provide a data set that is unique in bridging the semantic and intent gap. Millions of click data *i.e.* clicked image-query pairs, generated from search engines, were collected and released publicly as a new large-scale real-world image click data (*Clickture*) to investigate how to effectively leverage this click-count based data to mitigate the semantic gap. This click data is stored in a large table with multiple rows and three tuples (I, Q, C), indicating that the image I was clicked C times against the search results of a given textual query Q. Wu *et al.* (Wu et al., 2016a) view the entire dataset as a bipartite graph which has two types of vertices, queries and images respectively, and the edge's weight is assigned according to the total number of clicks from all the users. The authors learn a common representation for both image and text query from the perspective of encod-

[1] http://press.liacs.nl/mmgrand/microsoft.pdf

ing the explicit/implicit relevance relationship between the vertices in the click graph. The common representation is obtained as well as any unseen query or image is dealt with by reducing the truncated random walk loss and the distance between the learned representation of vertices and their corresponding deep neural network output.

Thus, the relevance relationship between a text query and its resulting image is obtained purely by measuring the distance between their corresponding learned high-level feature representation. In other words, the cross-modal retrieval on the unseen queries and images is dealt with here in a content-based fashion. As these content-based systems operate solely on feature representations, the definition of similarity in these systems is frequently ad-hoc and not explicitly optimized or generalized for any particular task *i.e.* here cross-modal retrieval. Frequently, the optimization of similarity for ranking affects the quantity of interest. Thus the retrieved items often become coarsely abstracted or potentially irrelevant. To overcome this shortfall, we try to capture relevant similarity information expressed by collaborative filtering.

The motivation behind using collaborative filtering (CF) is that this method produces user-specific recommendations of items based on patterns of ratings or usage without the need for exogenous information about either items or users (Koren and Bell, 2015). Recommendation by collaborative filtering relies on explicit or implicit feedback from the user or indirectly obtained through observing user behavior respectively. In our scenario, for any unseen query, aside from relying on the similarity score of the learned features, we can exploit some previous knowledge *i.e.* implicit feedback of the user. We consider click counts as the implicit feedback from the user. Stemming from this observation, we predict the similarity structure encoded by collaborative filtering data. Finally, we use collaborative filtering for generating a ranked list for cross-modal retrieval. To the best of our knowledge, ours is the first attempt in using collaborative filtering in conjunction with Deep Learning towards cross-modal image retrieval. A rigorous experiment is carried out over the *Clickture* dataset and the experimental results validate our claim. Our method outperforms the current state of the art on the learning and content-based methods.

2 Related Works

A fundamental image retrieval technique is to search for images by textual queries. The conventional image search engines leverage the benefits of associated or surrounding text which are generally collected from the data like image captions, tags, comments *etc.* To train such systems by labeled text-image pairs human intervention is necessary. However, such human labeling is expensive, time-consuming and quite cumbersome. These labeled data is unreliable as quite often they suffer from noise. Expressing an image entirely through a concise set of keywords, keyphrases or free text is a non-trivial task even for humans let alone a system. The problem is compounded when the user has limited to no knowledge of how the search system or IR work. To alleviate these problems of cross-view learning, the use of click-through data have gained momentum (Pan et al., 2014). Cross-View learning creates a common latent subspace where the data from different modalities like text, image *etc.*, can be compared with each other easily.

On the other hand, the click-through data is available in huge amount and is relatively easy to access. Also, this click-through data helps in better understanding a query. In the work by Pan *et al.* (Pan et al., 2014), the distance between mappings of query and image in the latent subspace is reduced and the inherent structure is preserved back to each original space. Once the mapping is done, and the latent representations are acquired, the next step is to compute the distance among these representations. Hence, choosing an appropriate similarity function becomes crucial as it is the key to make the cross-modal similarity tractable. He *et al.* (He et al., 2016) propose a deep and bidirectional representation learning model to address the issue of imagetext cross-modal retrieval. The authors adopt two convolutional deep neural networks to extract semantic representation from both raw image and text data and calculate cosine distance among those. Subsequently, a bidirectional network is learned from the matched and unmatched image-text pairs for training to capture the property of the cross-modal retrieval. This learning framework uses maximum likelihood criterion and optimizes the network through backpropagation and stochastic gradient descent.

Similarly, in the paper by Wang *et al.* (Wang et al.,

2015), the authors propose a supervised framework based on a deep neural network which captures the intra-modal and inter-modal relationships efficiently. The proposed model requires only a little prior knowledge to exploring high-level semantic correlation and also it can tackle the situation if any modality is missing. While most of the recent works focus on learning semantic representation, the work by Wu *et al.* (Wu et al., 2016b) concentrates on distance metric learning which is essential to improve similarity search for content-based retrieval. Usually, single modal distance metric learning methods suffer from some critical issues such as choosing the dominant feature from diverse feature representation, learning a distance metric on the combined high-dimensional feature space which is very time-consuming *etc.*. To overcome these issues, the authors proposed a multi-modal distance metric learning scheme called online multi-modal distance metric learning (OMDML), which can learn an optimized distance metric on each individual feature space and learn to find an optimal combination of diverse types of features.

Our work in this paper adopts the approach of using convolutional neural networks as suggested by He *et al.* (He et al., 2016). The reason for choosing this over other learning methods is that while training stage might take longer than some other methods, CNN usually supersedes others when it comes to the accuracy of learning. It should be noted that we have modified the settings of CNN to fit our problem and adapted it to our needs.

3 Methodology

Our model consists of two phases, training and testing, as is the case with any learning based technique. A click graph is used as labeled training data where the number of click count is treated as a label between a text query-image pair *i.e.* if any click is present between any text query and image, the image must be relevant to the text query. This assumption stems from the fact that a user usually clicks on an image against a text query only if she finds it relevant and useful. Here, the number of click counts reinforces how strongly relevant the image is against the query or vice versa. Thus a labeled query-image pair is learned through the click count. In the testing phase, relevant images from the test set are retrieved against any given test query and ranked. Hence, we perform cross-modal ranking over the new images and queries that are not involved in the training click graph.

3.1 Obtaining feature vector representations of query and documents

Multimodal objects from different feature spaces are present in the click graph. So, the first step of our model consists of projecting a feature vector representation of multimodal data into a common dimensional space. Here, image and text are the two different sources of information where the dimension of an image depends on the pixel intensity, and the number of pixels present in the image and vocabulary size of bag-of-words denotes the dimension of the text. Let us say if M is the dimension of image feature and N is the dimension of text feature then our objective is to come up with a common latent vector space of dimension D for both the image and the text.

We obtain a common representation through Convolutional Neural Network (CNN). Some pre-trained model such as inception-v3 model[2] can be used to learn the proper representation of input data. The learned representations account for the variations associated with the features. Any established CNN model consists of layers like convolutional filtering, local contrast normalization, max-pooling and finally fully connected neural network layers. For our model, we eliminate the last fully connected layer of the CNN (the output layer) and retrieve the image vectors from the penultimate layer. This is done since we are not interested in the classification of the images but instead in the generated embeddings of the images. The raw image features (embeddings) are directly fed into the model to get a latent representation. However, the text queries are represented in vector space model (bag-of-words). So, all words present in the vocabulary are inserted into a vector lookup table. Finally, a D dimensional representation is learned for each word from the lookup table. Test query usually consists of multiple words. So, the entire query can be represented by summing up all their corresponding word vectors. The obtained latent representations are used for learning purpose as depicted in the next subsection.

3.2 Learning from labeled click data

In Recommendation System, Collaborative Filtering (CF) models capture the interaction between

[2]https://www.kaggle.com/google-brain/inception-v3

users and item based on the rating. A rating indicates the preference of an individual user towards a particular item. High values of rating indicate a stronger preference of the user towards the particular item. The rating values are by nature either implicit or explicit feedback provided by the user or collected from user behavior or history. Perceiving some resemblance with the inherent nature of collaborative filtering with the characteristics of our dataset, we hypothesize that learning from clicked data through collaborating filtering may increase the retrieval performance. In this scenario, the text query and the images corresponding to the query play the role of *user* and *item* respectively. We treat click-count of each query-image pair as an implicit rating and train our model from these labeled query-image pairs and the rating matrix.

3.3 Prediction of click count for unseen query

Model-based collaborative filter predicts users' rating of unrated items. CF engines are more versatile, in the sense that they can be applied to any domain, and with some care could also provide cross-domain recommendations. Also, CF works best when the user space is large, which is the case for image searching where thousands of users are looking for images over the web every second. Taking a cue from this fact, we choose a model-based collaborative approach to predict the click count for an unseen query. The collaborative filter tries to predict ratings or click counts by characterizing both the query and image. Let us consider that the learned latent vector for any query q is V_q and the learned latent vector for any image i is V_i such that for a given query q, R measure the extent of relevance the query has with images that are highly relevant. The interaction R between query q and image i can be captured by the following Equation 1:

$$R = V_q^T V_i \tag{1}$$

where, the dot product between two vectors $x, y \in R^f$ is defined as in Equation 2.

$$x^T y = \sum_{k=1}^{f} x_k y_k \tag{2}$$

Thus, the predicted click count becomes R which is calculated using the Equation 1.

3.4 Calculating similarity score

By incorporating predicted click count between the unseen query and each image in the dataset, we calculate the similarity between each pair of the images. Let us consider, the predicted click count for the two images i_u and i_v against the nth query q_n as $R_{n,u}$ and $R_{n,v}$ respectively. Then the similarity measure between any two images, $S_{u,v}$, can be calculated by using the following Equation 3:

$$S_{u,v} = \frac{\sum_{n \in I}(R_{n,u} - \bar{R_n})(R_{n,v} - \bar{R_n})}{\sqrt{\sum_{n \in I}(R_{n,u} - \bar{R_n})^2}\sqrt{\sum_{n \in I}(R_{n,v} - \bar{R_n})^2}} \tag{3}$$

where $\bar{R_n}$ is the average click-count for those images and I denotes the entire image set. It can be observed from the above equation that the similarity measure depends on how much the click-count for a pair of images deviates from the average rating for those images. So, the similarity measure is purely dependent on the predicted click-count. As stated earlier, we calculate all the similarities between each pair of the images present in the dataset and based on the similarity score we rank the images against the each query. Thus a final ranked list is prepared and we select top-most images as the most relevant retrieved ones.

4 Experimental Setup

In this section we list the possible requirements for the experiment. To run Convolutional Neural Network for learning the latent representations over the large set of images, we have used cloud computing services of Google Cloud TPU [3] through 10 different instances. The learning is done by a pre-trained model through *ImageNet*[4] *i.e.* inception-v3 model[5]. The last fully connected layer of the Convoluted Neural Network, *i.e.* the penultimate layer of the CNN, is extracted using TensorFlow[6]. The dimension of all the learnt image vectors are kept to 2048. The other libraries which aid this process are NumPy[7], SciPy[8], scikit-learn[9], pickle[10] *etc.*

[3] https://cloud.google.com/tpu/
[4] www.image-net.org/
[5] https://cloud.google.com/tpu/docs/inception-v3-advanced
[6] https://www.tensorflow.org/
[7] https://www.numpy.org/
[8] https://www.scipy.org/
[9] https://scikit-learn.org/stable/
[10] https://docs.python.org/3/library/pickle.html

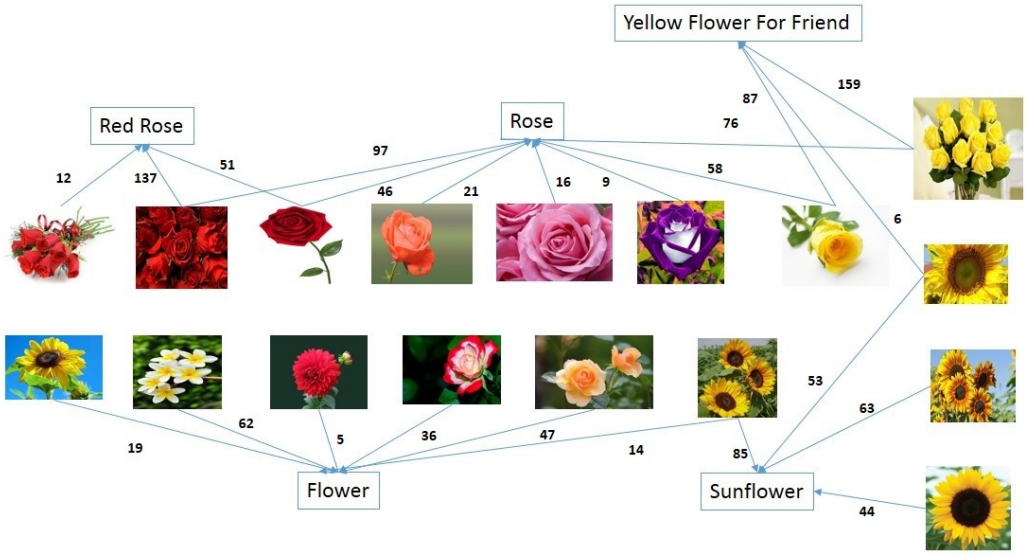

Figure 1: An example of the subgraph of the click graph

Dataset Our experiments are performed over an established real world dataset, Clickture 2014 (Microsoft, 2014), released by Microsoft as part of an Image Retrieval Challenge in 2015. Commercial image search engines like Google, Bing record clicks against queries to capture the user behaviour. Insightful usage of the recorded click-logs may lead to better cross-modal retrieval. The dataset comprises of two parts: (a) the training dataset and (b) the testing dataset or *Dev* dataset. The training dataset, consists of 1 million images and 11.7 million unique queries, is a sample of user click log which is a large table consisting of text queries, its associated images and number of of clicks for each query-image pair. An example of a subgraph of the Clickture dataset is depicted in Figure 1.

The click count between an image and a query is calculated from different users at different times. There are at least 23.1 million query-image pairs which have click count equal or more than 1. The *Dev* Dataset, which has $79,926$ query-image pairs generated from $1,000$ queries, is composed to have consistent query distribution, judgment guidelines and quality of a test dataset. For performance evaluation, manually annotated relevance measurement which is purely qualitative (labeled as Excellent, Good and Bad) is provided with the dataset.

5 Results and Analysis

In this work, each image is ranked by its respective Discounted Cumulated Gain (DCG) measure against the test queries. To calculate DCG, we sort the images against each query based on the final similarity score, obtained from our process. DCG for each query is computed as depicted in the following Equation 4. This metric was the official metric for the MSR- Bing Image Retrieval Challenge 2014 and 2015[11].

$$DCG_{25} = Z_{25} \sum_{i=1}^{25} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{4}$$

where $rel_i = \{Excelent = 3; Good = 2; Bad = 0\}$ is the manually judged relevance for each image with respect to the query, and $Z_{25} = 0.01757$ is a normalizer to make the score for 25 Excellent results. Here, we report the final evaluation metric as the average of for all queries present in the test set. We choose the comparative methods (baselines) against our proposed system from the base paper (Wu et al., 2016a). The comparative methods are as follows:

1. Bag-of-Words similarity based ranking method (BoWDNN-R)

[11] http://press.liacs.nl/mmgrand/microsoft.pdf

Model	DCG
BoWDNN-R	50.89
CCL	50.59
PAMIR	50.17
PSI	49.91
CMRNN	50.71
MRW-NN	51.04
Our Proposed Model (CNN+CF)	$\mathbf{79.88}^{p}$

Table 1: A comparison of various image retrieval methods

2. Click-through-based Cross-view Learning (CCL)

3. Passive-Aggressive Model for Image Retrieval (PAMIR)

4. Polynomial Semantic Indexing (PSI)

5. Cross-Model Ranking Neural Network (CMRNN) and

6. Multimodal Random Walk Neural Network (MRW-NN) respectively.

From Table 1, we can conclude that there is marked improvement in terms of retrieval performance when compared to other state-of-the-art techniques. The gain in terms of DCG is also statistically significant (indicated with a superscript p). Hence, we can safely conclude that considering click-counts as ratings and formulating the problem of image retrieval as item recommendation yields significantly better performance. The possible reason why our system performs better than others could be attributed to the fact that we did not rely on either latent semantic representation based learning or collaborative filtering individually. Instead, we proposed a new model that incorporates the learned feature representations from CNN as user and items, which possibly negated the shortfall of both techniques.

6 Conclusion and Future Work

Image retrieval has been one of the focal points of information retrieval systems since the early days of computing. Recent techniques have focused on various learning techniques to minimize the semantic gap between the query intent of users and the actual information retrieved by IRs. The same applies to image retrieval as well. While hybrid and multi-modal systems have shown superior performance when compared to unimodal retrieval systems, the problem of capturing the user information need ideally continues to be a challenge. Click counts offer a new dimension to aid in better understanding user's information need concerning images and when used judiciously can significantly improve the corresponding IR's performance. In this paper, we have applied the knowledge embedded within the clicks by using a collaborative filtering technique as an implicit feedback mechanism to enhance the latent representation based similarity computation. Our proposed technique performs superlatively against the state-of-the-art baseline over a real-world dataset.

As part of our future work, we would like to address the irregularities in the retrieval mechanism in the absence of modalities and would like to explore and suggest techniques to handle common problems associated with collaborative filtering methods. We would also like to study our method's effectiveness and scalability for real-time data. This work is a part of a larger project where we aim to integrate the retrieval model with image classification and automatic image annotation techniques proposed by us in our earlier works.

References

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal Retrieval with Correspondence Autoencoder. In *Proceedings of the 22Nd ACM International Conference on Multimedia*. ACM, New York, NY, USA, MM '14, pages 7–16. https://doi.org/10.1145/2647868.2654902.

Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan. 2016. Cross-Modal Retrieval via Deep and Bidirectional Representation Learning. *IEEE Transactions on Multimedia* 18(7):1363–1377. https://doi.org/10.1109/TMM.2016.2558463.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-

Semantic Alignments for Generating Image Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yehuda Koren and Robert Bell. 2015. *Advances in Collaborative Filtering*, Springer US, Boston, MA, pages 77–118. https://doi.org/10.1007/978-1-4899-7637-6_3.

Microsoft. 2014. Clickture project. https://www.microsoft.com/en-us/research/project/clickture/. Accessed: 2019-04-22.

Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based Cross-view Learning for Image Search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '14, pages 717–726. https://doi.org/10.1145/2600428.2609568.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

C. Wang, H. Yang, and C. Meinel. 2015. Deep Semantic Mapping for Cross-Modal Retrieval. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. pages 234–241. https://doi.org/10.1109/ICTAI.2015.45.

Fei Wu, Xinyan Lu, Jun Song, Shuicheng Yan, Zhongfei Mark Zhang, Yong Rui, and Yueting Zhuang. 2016a. Learning of Multimodal Representations With Random Walks on the Click Graph. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 25(2):630642. https://doi.org/10.1109/tip.2015.2507401.

P. Wu, S. C. H. Hoi, P. Zhao, C. Miao, and Z. Y. Liu. 2016b. Online Multi-Modal Distance Metric Learning with Application to Image Retrieval. *IEEE Transactions on Knowledge and Data Engineering* 28(2):454–467. https://doi.org/10.1109/TKDE.2015.2477296.

Yi Zhen and Dit-Yan Yeung. 2012. A Probabilistic Model for Multimodal Hash Function Learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '12, pages 940–948. https://doi.org/10.1145/2339530.2339678.