Online multi-core ranking model for image retrieval

Xuezhuan Zhao¹³,a, Lishen Pei²３, Zheng Zhang⁴, Tao Li⁴

¹Zhengzhou University of Aeronautics, Wenyuan Road, Zhengzhou450046, China.
²Henan University of Economics and Law, Jinshui Road, Zhengzhou450046, China.
³Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, Tianjin300300, China.
⁴HeNan Radio Television University, Wenyuan Road, 450046 Zhengzhou, China.

azhaoxuezhuan@zua.edu.cn
lishen198815@126.com

Abstract. With the popularity of social media services, efficient online image retrieval urgently needs to meet the diverse needs of network users. How to use the existing semantic information and social label information to establish a content model to bridge the semantic gap is an urgent problem to be solved. In this article, we propose an efficient online multi-core ranking model (OMKR), which is trained by minimizing the triplet loss of hard negative samples based on multiple query dimensions and complementary feature channels. By optimizing the sorting performance of multi-dimensional queries, the semantic consistency between image sorting and text query input is directly maximized without relying on the intermediate semantic annotation process. A large number of experiments on two social media data sets prove the advantages of our method in terms of retrieval performance.

1. Introduction
Multimedia content search in the network environment is a very challenging task. Due to the diversification of user preferences and the heterogeneity of user behaviour, the popularity of social media services makes this task even more difficult. Online users usually present themselves by transmitting online multimedia to their social circles and providing user-generated content through mobile devices. With the increasing amount of information conveyed by massive online content, how to provide effective and efficient solutions for social media image retrieval has become an increasingly difficult task.

In order to solve this practical problem, a possible paradigm is to construct by combining content information and co-occurrence semantic information through content-based visual analysis[1] and semantic-based analysis[2]. Image retrieval methods in the real world. In content-based analysis, a retrieval model is established based on local and global visual feature representation. Therefore, this article delves into visual content hashing, indexing and similarity learning models. In semantic-based analysis, images and queries (visual or text) are projected into a multi-dimensional semantic space. The similarity between query and database image in semantic space is calculated.

In this work, we propose a new online multi-core ranking (OMKR) method to directly learn image retrieval models without relying on semantic annotation. Our model adopts a learning criterion related to the final retrieval performance based on discriminative learning. It takes a set of training queries
and a set of top-ranked online social media images as input, and outputs a trained model to obtain high ranking performance for new queries. The model combines a variety of visual features to explore the correlation between different query terms, and achieves a good model versatility. OMKR also has an efficient online optimization process, which is based on an online multi-core learning framework. Therefore, it allows learning a large amount of training data.

The article is arranged as follows, and Section 2 briefly reviews the relevant literature. Section 3 introduces the online multi-core ranking. Section 4 provides experimental details and discussion.

2. Related Work
People are committed to modelling different aspects of online multimedia retrieval in order to achieve visual content retrieval. This article has conducted a brief literature review from the following aspects.

For decades, image retrieval has been a core research problem in multimedia research community. Research efforts have been made to bridge the semantic gap between the user queries and the multi-dimensional content representation[1]. For example, Grangier[3] proposed a discriminative kernel based ranking approach for image retrieval by textual queries. Rasiwasia[2] constructed a unified semantic space for cross modal data, which was based on what documents to be retrieved by queries from other modalities. Following this idea, correlation learning from multiple modalities has been comprehensively studied[4]. Zhang[5] proposed an attribute-augmented semantic hierarchy for content based image retrieval. Since then fusing complementary information for visual content modeling has been a widely accepted paradigm to achieve better semantic consistency. Li[6] propose a deep collaborative embedding method for SocialImage tagging, tag-based image retrieval and content-based image retrieval.

As an effective and efficient solution framework for image retrieval and text-image retrieval, online similarity learning[7-9] has been studied extensively. Xia[7] develop a multiple kernel similarity learning for visual search. To address the modality difference between visual and textual modalities, Wu propose an asymmetric similarity learning method for text-image retrieval which aggregates the visual features of different CNN layers. An online low-rank similarity learning method is also proposed in [9] to obtain a low-rank similarity parameter matrix for measuring similarity between image and text. However, existing solution has not taken advantage of the hard negatives for learning the retrieval model, which has been proved to be quite useful in modeling the semantic relation and discovering the true related retrieval results [10].

Online multimedia documents are believed to be correlated to each other on different aspects where such context information is delivered by their meta-data. The context and correlation usually have strong relevance to their semantics. Mcauley [11] showed that image labelling with mere social media meta-data performed equally or even outperformed visual content modelling method. Knowledge discovered from context can be informative for other tasks. For example, friend suggestion can be made by modelling the similarity among users by joint content and context analysis [12]. Heterogeneous user behaviours can be modelled by the social context of online social media and effectively combine the multi-aspect behaviour similarities by multiple kernel learning towards friend recommendation, advertisement and people searching [13]. Our approach captures the potential preference styles from heterogeneous social attributes. Consequently, the user expectation on the retrieval results can be conveniently expressed by weight specification.

3. Online Multiple Kernel Ranking
We propose an Online Multiple Kernel Ranking approach (OMKR) by minimizing the hard-negative-based triplet loss to rank the images according to their semantic consistency with the multi-dimensional textual query input. Compared with the existing approaches, our model directly fits the images into the query space. The better semantic consistency is achieved by combining complementary visual features. We design an online learning procedure which quickly optimizes the ranking model with a large number of training triplets where the negative samples in the triplet are selected from the most similar ones to their positive counterparts, and the model quickly converges in
receiving less than 30 thousand triplets. Consequently, we learn a set of semantic coherent projections which map each image into a low dimensional semantic space where the relevance between the queries and the database images can be directly calculated by inner product.

3.1. Ranking Model

An image database is represented as \( D \in \mathbb{R}^{N \times V} \) where \( N \) denotes the number of images and \( V \) denotes the number of feature dimensions for each image. We denote each image as \( d \in D \) where \( d \in \mathbb{R}^V \). We represent the textual query as an \( M \)-dimensional real value vector \( q = (q_1, \ldots, q_M) \in \mathbb{R}^M \) where there may be multiple non-zero entries for multi-word query input. The score function \( F_w(q, d) \) of an image \( d \) from \( D \) can be written as:

\[
F_w(q, d) = q \cdot f_w(d) = q \cdot (w_1 \cdot d, \ldots, w_M \cdot d)
\]

where \( \gamma(q, d) = [q_1d; \ldots; q_Md] \). For each query \( q \), suppose we have collected the ranking information (\{\it relevant\} or \{\it irrelevant\}) of the images in the database \( D \). We organize the data into a training triplet set \( D_r \) where each triplet is represented as \((q, d^+, d^-) \in D_r \). The ranking function learning is equivalent to minimizing the following primal Ranking SVM (RSVM) objective function[3]:

\[
\min_w \frac{1}{2} \| w \|^2 + \frac{C}{|D_r|} \sum_{(q, d^+, d^-) \in D_r} \xi_{(q, d^+, d^-)}
\]

s.t. \( w \cdot \gamma(q, d^+) - w \cdot \gamma(q, d^-) \geq 1 - \xi_{(q, d^+, d^-)} \),

\[
\xi_{(q, d^+, d^-)} \geq 0
\]

where \( w = [w_1, \ldots, w_M] \) denotes the concatenated discriminative model parameter vector. We can introduce any kernel function \( \kappa: \mathbb{R}^X \times \mathbb{R}^X \rightarrow \mathbb{R} \) for calculating the similarity among images in high dimensional space. Consequently, the discriminative functions \( f_m, m = 1, \ldots, M \) and the score function \( F(q, d) \) can be represented as:

\[
f_m(d) = \sum_{j=1}^M q_m \alpha_j \left( \kappa(d^+_j, d) - \kappa(d^-_j, d) \right)
\]

\[
F(q', d) = \sum_{m=1}^M \sum_{g=1}^G q_m q_g \alpha_m \alpha_g \left( \kappa_g(d^+_j, d) - \kappa_g(d^-_j, d) \right)
\]

When the similarity among images are represented by multiple kernels \( \kappa_g, g = 1, \ldots, G \), according to the represented theorem, the discriminative function and score function are formulated by [14]:

\[
f_m(d) = \sum_{g=1}^G \sum_{j=1}^M q_m q_g \alpha_j \beta_{mg} \left( \kappa_g(d^+_j, d) - \kappa_g(d^-_j, d) \right)
\]

\[
F(q', d) = \sum_{m=1}^M \sum_{g=1}^G \sum_{j=1}^M q_m q_g \alpha_j \beta_{mg} \left( \kappa_g(d^+_j, d) - \kappa_g(d^-_j, d) \right)
\]

By introducing the Lagrangian and Karush-Kuhn-Tucker (KKT) condition, we obtain the following dual problem:
minmax αβ\(1-\alpha\) \(\sum_{M} \sum_{G} \beta_{mg} \Theta_{mg}\) α s.t. 0 ≤ α ≤ \(\frac{C}{|D_r|}\), \(\sum_{M} \sum_{G} \beta_{mg} = 1\), ∀ m

where \(a \in \mathbb{R}^{M \times |D_r|}\), \(\beta \in \mathbb{R}^{M \times G}\), and \(\Theta_{mg} \in \mathbb{R}^{[D_r \times |D_r|]}\) is a positive semi-definite matrix where:

\[
\Theta_{mg}(i, j) = q_{mi} q_{mj} \cdot \\
\left( \kappa_g(d^+_i, d^-_j) - \kappa_g(d^+_i, d^-_j) + \kappa_g(d^+_i, d^-_j) \right)
\]

To efficiently learn the ranking model, a large number of training data should be involved. The number of training triplet is approximately \(O(|T| |D|^2)\) where \(|T|\) denotes the number of textual queries for training and \(|D|\) denotes the number of images in the database. Consequently, to optimize the dual problem in Eqn. 5, the prohibitive size of memory is required to load and maintain all the \(\Theta_{mg}\). To efficiently handle big data, we propose an online optimization procedure to optimize the multiple kernel ranking models, which will be introduced later.

3.2. Online Model Learning

The proposed Online Multiple Kernel Ranking (OMKR) algorithm is based on the fusion of two online learning methods: the Perceptron algorithm [15] and the Hedge algorithm [17]. Particularly, for each kernel and each textual query dimension, the Perceptron algorithm is employed to learn a kernel-based classifier with some selected kernel, and the Hedge algorithm is used to update their combination weights.

In this framework, we use \(\theta^t_{mg}\) to denote the combination weight for the \(g\)-th kernel classifier of \(m\)-th query dimension at round \(t\) which is initially set to 1. For each learning round, we update the weight \(\theta^t_{mg}\) by following the boosting style Hedge algorithm where each discriminative function can be treated as a weak learner. The weight update rule can be formulated as:

\[
\theta^{t+1}_{mg} = \theta^t_{mg} \sigma^{z^t_{mg}}
\]

where \(\sigma \in (0,1)\) is a discount weight parameter which is employed to penalize the kernel classifier that performs incorrect prediction at each learning step, and \(z^t_{mg}\) indicates that if the \(g\)-th kernel classifier of the \(m\)-th query dimension makes a mistake on the prediction of the training triplet \((q_j, d^+_i, d^-_j)\), namely, \(q_{mj}(f_{mg}(d^+_i)) - f_{mg}(d^-_j) \leq 0\). When the \(t\)-th training triplet is incorrectly predicted on the \(m\)-th query dimension and the \(g\)-th kernel, the corresponding discriminative sub-model is updated as:

\[
f^{t+1}_{mg}(d) = f^{t}_{mg}(d) + q_{mj} \left( \kappa_g(d^+_i, d^-_j) - \kappa_g(d^+_i, d^-_j) \right)
\]

After receiving a sequence of \(T\) training triplets, denoted by \(D_T = (q_i, d^+_i, d^-_i, t = 1, \ldots, T)\), the number of mistakes \(\Psi\) denoted as:

\[
\Psi = \sum_{t=1}^{T} \left[ w_t \left( \gamma(q_i, d^+_i) - \gamma(q_i, d^-_i) \right) \leq 0 \right] \\
= \sum_{t=1}^{T} \left[ \sum_{m=1}^{M} \sum_{G} (q_{mi} > 0) \beta^t_{mg} z^t_{mg} \geq 0.5 \right]
\]

is bounded as follows:
\[
\begin{align*}
\Psi \leq & \frac{2 \ln(1/\sigma)}{1-\sigma} \min_{1 \leq g \leq G} \sum_{m=1}^{\tau} z_{mg}^g + 2 \frac{MG(\ln M + \ln G)}{1-\sigma} \\
\leq & \frac{2 \ln(1/\sigma)}{1-\sigma} \min_{1 \leq g \leq G} H_{mg} + 2 \frac{MG(\ln M + \ln G)}{1-\sigma}
\end{align*}
\]

(10)

By choosing \( \sigma = \frac{\sqrt{T}}{\sqrt{T} + \sqrt{\ln M + \ln G}} \), we have:

\[
\Psi \leq 2 \left( 1 + \sqrt{\frac{\ln M + \ln G}{T}} \right) \min_{1 \leq m \leq M} H_{mg} + \\
\ln M + \ln G + \sqrt{T(\ln M + \ln G)}
\]

(11)

where \( H_{mg} \) denotes the structured loss on each individual classifier \( f_{mg} \) as:

\[
H_{mg} = \min_{f_{mg}} f_{mg}^2 + 2 \sum_{\tau=1}^{\tau} \sum_{(q_i', d_i', d_j')}.
\]

(12)

As indicated by [16], the proof can be made by essentially combining the proof of the Perceptron [15] and the Hedge algorithm [17]. The details are omitted. It indicates that the error bound of the discriminative function is substantially determined by the error of the best weak learner. The error bound in the above theorem can be further improved from two aspects. First, it can be improved if we further tune the step-size or the margin. Second, it is further improved if we apply hard negative based training scheme, since the error of the best weak learner can be further reduced by minimizing the hard negative learning objective function.

For large scale application, our proposed OMKR model needs to traverse the training data once or very limited times. Consequently, given a textual query \( q' \), we obtain a rank list \( \tau_q \) from the image database which reflects the semantic consistency between the query input and each image.

4. Experimental and Results

In this section, we perform systematical evaluation on two real world social media datasets on social media image retrieval task.

4.1. Datasets

The datasets we used in this paper include: The NUS-WIDE dataset [18] consists of 269,648 images collected from Flickr. The Flickr dataset consists of 3.5 million images collected from Flickr covering wider visual topics than NUS-WIDE.

4.2. Retrieval Performance

We perform extensive experiment to evaluate the retrieval performance of all the compared methods. The MAP measurements on top 500 retrieved results of different methods are shown in Table 1 and 2. We denote the retrieval results by semantic ranking of the original retrieval method, and rank aggregation with both SR and the re-ranking results of the surrounding text as SR+txt, and similarly, SR+lc for location, SR+tm for time, SR+gp for group, SR+cgt for category, SR+id for user ID and SR+all for the weighted aggregation using SR and all the preference ranking lists. By appropriately aggregating the semantic ranking and preference ranking results, our approach achieves much better retrieving performance than other approaches on both datasets.

Firstly, the results indicate that different social attributes carry different implications on the true semantics of the social media images. For example, by aggregating SR and location (SR+lc), the retrieval performance of all the compared approaches is improved over SR. By aggregating SR and user ID (SR+id), our approaches consistently obtain improved results. The upload time information is less relevant to the semantics, since the results by aggregating SR and user ID (SR+id) usually
underperform results on SR. Our method consistently performs the best in all the cases over other approaches.

Table 1: The retrieval performance of the top 500 ranked images in Map (%)

| Datasets | NUS-WIDE        |                |                |                |                |                |                |                |
|----------|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Methods  | SR              | SR+txt         | SR+lc          | SR+tm          | SR+gp          | SR+ctg         | SR+id          | SR+all         |
| PAMIR-PR[19] | 27.6           | 28.3           | 27.9           | 26.9           | 28             | 28.2           | 27.9           | 28.7           |
| MMNN-PR[20]  | 26.3           | 26.9           | 26.3           | 25.5           | 26.9           | 27.3           | 25.9           | 28.9           |
| SCM-PR[2]    | 21.5           | 21.9           | 21.5           | 20.1           | 21.9           | 22.1           | 21.2           | 22.2           |
| OMKR        | 31.7           | 32.2           | 32.0           | 31.8           | 32.1           | 32.5           | 31.9           | 32.6           |

Secondly, the results also imply that by preference ranking and aggregating, the performance of all the semantic based models can be enhanced by incorporating heterogeneous social attributes. For example, the performance of all the approaches on SR+all outperforms SR on both datasets. Generally, by aggregating more social attributes, the retrieval performance of all the methods on SR+all outperforms rank aggregation with single social attribute.

Table 2: The retrieval performance of the top 500 ranked images in Map (%)

| Datasets | FLICKR        |                |                |                |                |                |                |                |
|----------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Methods  | SR            | SR+txt         | SR+lc          | SR+tm          | SR+gp          | SR+ctg         | SR+id          | SR+all         |
| PAMIR-PR[19] | 12.4          | 13.2           | 12.8           | 12.0           | 13.1           | 14.0           | 12.8           | 14.5           |
| MMNN-PR[20]  | 12.1          | 12.9           | 12.6           | 11.8           | 13.2           | 13.6           | 12.3           | 14.0           |
| SCM-PR[2]    | 5.8           | 6.2            | 5.9            | 5.5            | 6.5            | 6.9            | 6.3            | 7.2            |
| OMKR        | 14.3          | 15.1           | 14.9           | 14.4           | 15.8           | 16.2           | 15.1           | 17.8           |

Lastly, the rank aggregation results on Flickr dataset shows that, when processing large scale social media with weak semantic information such as the noisy tags, fusing the semantic relevance delivered in different social attributes will boost the retrieval performance in a more promising manner. Such a claim is made by observing that all of the compared approaches perform at least 13% better on SR+all vs. SR on Flickr dataset, while the largest performance gain on FLICKR is only about 13%, correspondingly.

5. Conclusion

In this paper, we proposed an efficient Online Multiple Kernel Ranking model constructed on multiple query dimensions and complimentary feature channels. By optimizing the ranking performance, the semantic consistency between the image ranking and textual query input is directly maximized without relying on intermediate semantic annotation procedure. Extensive experiments on two social media datasets have demonstrated the advantages of our approach. In future work, we will investigate how to model the online user behaviours in a more comprehensive way to better facilitates the user preference.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61806073, U1904119, 31700858) and the Research Programs of Henan Science and Technology Department (192102210097, 192102210126, 192102210269, 182102210210,182102210447) and the Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (NO. CAAC-ITRB-201607) and the Key scientific research projects of colleges and universities in Henan Province (18A520050, 21A520051).

References

[1] Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-
level semantics. Pattern Recognition, 2007.40(1), 262–282.

[2] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM Multimedia, 2010. pp. 251–260.

[3] Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008. 30(8), 1371–1384.

[4] Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. 2013,( In: SIGMOD)

[5] Zhang, H., Zha, Z.-J., Yang, Y., Yan, S., Gao, Y., Chua, T.-S.: Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In: ACM Multimedia, 2013, pp. 33–42.

[6] Li, Z., Tang, J., Mei, T.: Deep collaborative embedding for social image understanding. IEEE transactions on pattern analysis and machine intelligence, 2018,41(9), 2070–2083.

[7] Xia, H., Hoi, S.C., Jin, R., Zhao, P.: Online multiple kernel similarity learning for visual search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013,36(3), 536–549.

[8] Wu, Y., Wang, S., Song, G., Huang, Q.: Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. IEEE Transactions on Image Processing, 2019,28(9), 4299–4312.

[9] Wu, Y., Wang, S., Huang, Q.: Online fast adaptive low-rank similarity learning for cross-modal retrieval. IEEE Transactions on Multimedia (2019),1-8

[10] Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 2(7), 8 (2017)

[11] McAuley, J., Leskovec, J.: Image labeling on a network: using social-network metadata for image classification. In: ECCV, 2012, pp. 828–841.

[12] Yao, T., Ngo, C.-W., Mei, T.: Context-based friend suggestion in online photo-sharing community. In: ACM Multimedia, 2011, pp. 945–948.

[13] Zhuang, J., Mei, T., Hoi, S.C., Hua, X.-S., Li, S.: Modeling social strength in social media community via kernel-based learning. In: ACM Multimedia, 2011, pp. 113–122.

[14] Bach, F.R., Lanckriet, G.R., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML, 2004, pp.1-6.

[15] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 1958,65(6), 386.

[16] Hoi, S.C., Jin, R., Zhao, P., Yang, T.: Online multiple kernel classification. Machine learning, 2013,90(2), 289–316.

[17] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 1997,55(1), 119–139.

[18] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: CVIR, 2009, pp. 48-56.

[19] Grangier, D., Bengio, S.: A discriminative kernel-based approach to rank images from text queries. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008,30(8), 1371–1384.

[20] Masci J, Bronstein M M , Bronstein A M , et al. Multimodal Similarity-Preserving Hashing[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36(4):824-830.