CROSS-MEDIA LEARNING FOR INFORMATION RETRIEVAL

Chen Yun

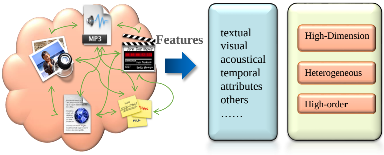
The University of Hong Kong

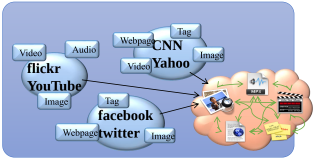
May 31, 2016

Motivation and Background

Properties of Cross-media

- Cross-modality : many kinds of features can be obtained and they have different intrinsic discriminative power to characterize the corresponding semantic
- Cross-collections : the data about a same topic/event may be obtained from multiple sources.





Motivation and Background

Challenge

- Semantic gap between data from different modalities
- Heterogeneity gap between data from multiple sources
- Tremendous amount of cross-media data

Three related cross-learning research topics

- Cross-media retrieval : support similarity search for multi-modal data [1] [2]
- Cross-media ranking : learn ranking function to preserve the orders of relevance for cross-media data [3] [4]
- Cross-media hashing : learn hashing function(s) to faithfully preserve the intra-modality and inter-modality similarities and map the high-dimensional multi-modal data to compact binary codes. [5] [6]

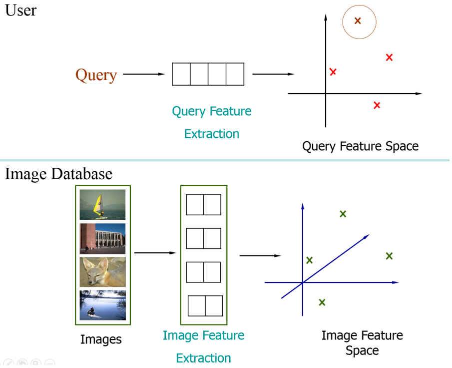
Step 1 : cross-media retrieval for image search engine

Keyword-based image search

- Traditionally, keyword-based image search is performed by leveraging the surrounding texts of images
- Click-through data are natural labeling sources for keyword-based image search

Unique properties of the click-through data

- Noisy with typo and missspelling
- Short query with lots of people name and location name
- Sparse
- Large scale



Task

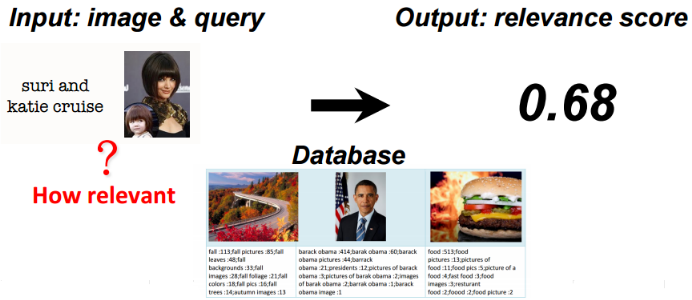


FIGURE 1 – The task of Microsoft Bing Grand Challenge.

Proposed model : click-through-based word embedding (CWE)

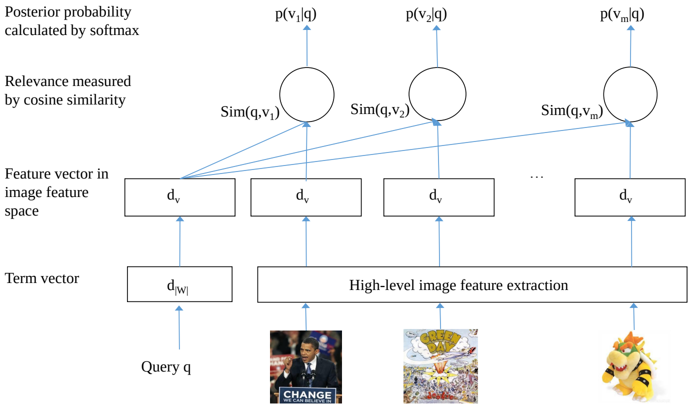


FIGURE 2 – The overall architecture of our proposed model.

Click-through-based word embedding (CWE)

- Objective function :

$$\arg\max_{\mathbf{W}} \sum_{e_{ij} \in D'} f(c_{ij})(log\sigma(\mathbf{v}_j^{+T}\mathbf{W}^T\mathbf{q}_i) + log\sigma(-\mathbf{v}_k^{-T}\mathbf{W}^T\mathbf{q}_i)) \quad (1)$$

where e_{ij} is a data entry $(q_i, v_j^+, v_k^-, c_{ij})$ in D'

$$f(c) = \begin{cases} (c/c_{max})^{\alpha} & c < c_{max} \\ 1 & otherwise \end{cases} \quad (2)$$

Evaluation : retrieval performance

TABLE 1 – NDCG@25 (%) of different approaches on Dev dataset

Approach	PSI	CCA	CCL	CWE	Random	Upper Bound
	49.91	50.55	50.59	51.12	46.64	67.73

- Normalized Discounted Cumulated Gain at depth d ($NDCG_d$) :

$$NDCG@d = N_r \sum_{i=1}^{d} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{3}$$

where the $rel_i = \{Excellent = 3, Good = 2, Bad = 0\}$ is manually judged relevance for each image with respect to the query. N_r is a normalizer to make the scores for 25 Excellent results 1 : $Nr = \frac{1}{\sum_{i=1}^{d} \frac{7}{log_2(i+1)}}$

Evaluation : retrieval performance

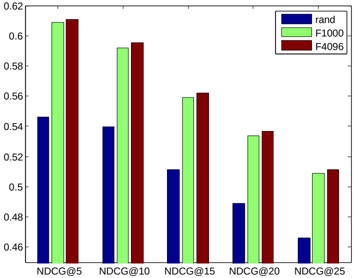


FIGURE 3 – The NDCG value at different depths for CWE using different image features compared with the random method.

Time complexity

- For a query with length l_q, the training complexity is $(l_q + 2) \times d$. It does not scale up with the number of triads in the click log, the image size or the vocabulary size.
- Our model takes only 30 minutes to process 1 million triads until converge on an ordinary PC with 1.4GHz CPU and 8GB RAM, while the state of art CCL takes 32 hours to process the same number of data on a server with 2.4GHz CPU and 128GB RAM.

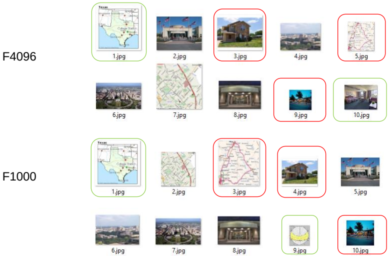
Evaluation : retrieval samples

Top ten images for query uss california

Good
Bad

Top ten images for query college station texas

Good
Bad



Evaluation : quality of word embedding

```
>>> model.most_similar('autumn')
[(u'fall', 0.40435126423835754), (u'fallleaves', 0.37148547172546387), (u'foliag
e', 0.3677600622177124), (u'carissa', 0.3512413203716278), (u'harvest', 0.339123
9047050476), (u'fallvpics', 0.3310385048389435), (u'falldesktop', 0.326423346996
3074), (u'autum', 0.3152065873146057), (u'seasonal', 0.3150034844875335 7), (u'th
anksgivingwallpaper', 0.29540446400642395)]
>>> model.most_similar('china')
[(u'xinjiang', 0.3983587324619293), (u'shandong', 0.31767889857292175), (u'wuxi'
, 0.3097914159297943), (u'westernization', 0.3085956573486328), (u'simbolos', 0.
29337844252586365), (u'acrobatics', 0.2810131907463074), (u'municipality', 0.247
66336381435394), (u'guilin', 0.2474101483821869), (u'lenox', 0.24573944509029388
), (u'macao', 0.2448534071445465)]
```

FIGURE 4 – Most similar words retrieval for given sample words

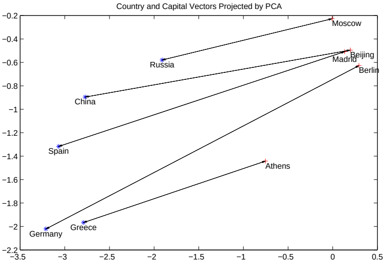


FIGURE 5 – Two-dimensional PCA projection of the 1000-dimensional click-through-based word vectors of countries and their capital cities

Conclusion

- We look at the problem of cross-media retrieval from an image search engine by leveraging the click-through data. There are only limited works about this topic.
- We propose a novel probabilistic model to bridge the semantic gap between images and queries by modeling the conditional probability of an image to be clicked given a query. Negative sampling and an adjusted weighting function has been applied.
- The extensive experiments have demonstrated that our model outperforms state of art in terms of both accuracy and scalability. Thus, our model can be easily applied in larger dataset.

References I

F. Feng, X. Wang, R. Li, and I. Ahmad, "Correspondence autoencoders for cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 1s, p. 26, 2015.

W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, "Effective deep learning-based multi-modal retrieval," *The VLDB Journal*, vol. 25, no. 1, pp. 79–101, 2016.

N. Craswell and M. Szummer, "Random walks on the click graph," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 239–246, ACM, 2007.

V. Jain and M. Varma, "Learning to re-rank : Query-dependent image re-ranking using click data," in *Proceedings of the 20th ACM International Conference on World Wide Web*, pp. 277–286, 2011.

Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization : A procrustean approach to learning binary codes for large-scale image retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2916–2929, 2013.

D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization.," in *AAAI*, pp. 2177–2183, 2014.