Enhancing the efficiency of text-image based tasks through multimodal solutions

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

by

Deepanwita Datta

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY)

VARANASI- 221 005

Roll No: 13071502

December 2017

Chapter 7

Conclusions

It has been established in the literature that unimodal solution to image retrieval do not produce optimum performance and hence most of the latest research has pivoted to multimodal solutions. This thesis studied and presented several such issues and challenges related to images and tried to solve them under different circumstances. Basically, we tried to address three such facets–(1) Image Retrieval (2) Image Classification and (3) Image Annotation. While the first part of the thesis described the work done towards solving the issue of multimodal retrieval using query reformulation, the second half focused on ancillary but very integral issues associated with image retrieval and used various learning techniques, statistical measures and artificial intelligence techniques to address them. Query expansion, a commonly employed query reformulation technique, has been widely studied in the field of text retrieval. We studied the effect of text query expansion for the

case of image retrieval. Our study revealed that judicious query expansion improves the image retrieval efficiency as well. For textual query expansion we selected some highly important keyphrases from relevant documents through relevance feedback and appended them with the original or user-provided query. To find the top most informative or important terms we adopted few well established keyphrase extraction techniques; *tf-idf* based (TQEM), KEA (KQEM) and Topic decomposition based (TMQEM), aside from proposing a novel one which is based on Mutual Information (MIQEM). MIQEM is a graph model, enriched semantically by WordNet and the semantic association between words are captured by Mutual information where as TMQEM is a PageRank based approach which basically works on decomposed topics generated by a Topic decomposition model which is Latent Dirichlet Allocation. Our comparative study revealed that TMQEM had the best performance overall. Although our earlier proposed method, MIQEM, was the next best in performance, its novelty lie in that it is entirely unsupervised and effectively captures the semantic association between words and was the first step towards formulating TMQEM.

For the image retrieval model, the textual and visual features are combined using optimal combination parameters learned by employing Fisher-LDA and it is also established that the combined query improves the image retrieval efficiency than any single modal system. We believe that this is the first work that studies query expansion using keyphrase extraction in light of multimodal retrieval and also puts forward a new algorithm for the same. A detail set of experiments is carried out on the ImageCLEF 2010-11 dataset to validate our claims.

Image annotation is an indispensable aspect of image retrieval, especially for image indexing. Generally conventional systems and tools segregate the image into discrete objects and annotate them separately. But sometimes, specially for complex scenes, such systems fail to capture the essence of the entire scene. Thus the information which is conveyed by the entire image as a whole, are missing in those cases. To combat this, we proposed a probabilistic model with two different measures namely co-occurrence count and mutual information which supersedes the performance of the baselines comprehensively. Our model is an entire graph-based approach where a *concept graph* is formed including all concepts in the database. An insightful community detection technique, *Infomap*, is used to cluster the graph and identify the similar or overlapping concepts. Finally, a greedy walk is performed on the clusters of concepts extracted from the concept graph. To this end, best possible annotations for an image are ranked according to their informativeness and the top most annotations are selected to be the finally predicted annotations. Through a set of experiments, we tested and compared our proposed system against few state-of-the-art system and finally we concluded that our system is the best performing among them.

We hypothesized that multimodal solution performs better in any task than unimodal one. Since image classification is also a closely related topic to image retrieval, we tested our hypothesis on it as well. For the classification task, first relevant features are identified and combined in such a way that the best result will be obtained.

Generally, the weight assignment of each feature is done according to their importance or discriminatory power for classification. We contended that an intelligent optimization method such as Hill Climbing can perform better where other recent existing work focuses on average combination and weighted average combination of kernels to achieve the same. Since gradient descent approaches like Hill Climbing suffers from its inherent shortfall *i.e.* get stuck into local maxima, we proposed a modified version of the same named Extended Hill Climbing to reinforce the same. It is established in the state-of-the-art that SVM works well in multiple kernel learning framework, we choose SVM as the core building block of our classification task. Thorough experiments on a publicly available standard dataset validates our claim and proves that our proposed approach significantly outperforms the other existing combination methods in image classification task.