Multimodal Medical Image Retrieval based on Latent Topic Modeling

Mandikal Vikram
Department of Information Technology,
National Institute of Technology
Karnataka, Surathkal
15it217.vikram@nitk.edu.in

Aditya Anantharaman
Department of Information Technology,
National Institute of Technology
Karnataka, Surathkal
15it201.aditya.a@nitk.edu.in

Suhas BS
Department of Information Technology,
National Institute of Technology
Karnataka, Surathkal
15it110.suhas@nitk.edu.in

Sowmya Kamath S.
Department of Information Technology,
National Institute of Technology
Karnataka, Surathkal
sowmyakamath@nitk.edu.in

Abstract

Modern medical practices are increasingly dependent on Medical Imaging for clinical analysis and diagnoses of patient illnesses. A significant challenge when dealing with the extensively available medical data is that it often consists of heterogeneous modalities. In this work, we seek to leverage both the visual and textual modalities to improve the image retrieval. We propose a Latent Dirichlet Allocation (LDA) based technique for encoding the visual features of the medical images. We then explore early fusion and late fusion techniques to combine the textual and visual features. The proposed late fusion technique obtains a higher mAP than the state-of-the-art on the ImageCLEF 2009 dataset.

1 Introduction

The proliferation of medical image data from medical institutions, documented in digital forms is a valuable asset for diagnostic medical informatics. The primary objective of medical image retrieval task is to retrieve images which are the most relevant from a given clinical perspective. The varying subjectivity and shallow context sensitivity of the image annotations are significant hurdles faced by text retrieval techniques. Hence, more recent approaches have focused on combining visual and textual techniques for a multi-modal approach to image retrieval. One of the significant challenges in medical image retrieval is that the low-level visual and textual features do not directly correspond to the high-level medical concepts, in other words, there exists a semantic gap between the two. Another challenge is the way in which the visual and textual features are integrated, which needs to be specifically addressed to account for the diverse information contained in medical images. In this work, we show that the visual features of medical images can be effectively encoded by the latent topics derived using the LDA[1] from the generated SIFT features[2]. We combine these visual features with the textual words to further improve our performance.

2 Methodology

Figure 1 illustrates the proposed approach. We adopt a Visual Bag-of-Words(VBoW) model inspired by [3] to represent the visual features. The Scale-invariant Feature Transform [2] is used to detect

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

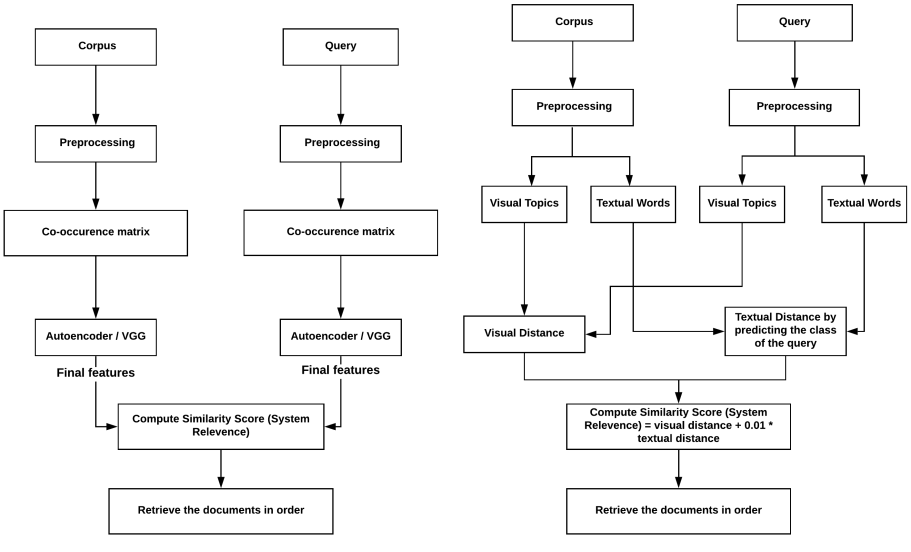
the salient points and obtain the feature descriptors of a given image. The feature descriptors of all the images in the corpus are clustered into some k clusters using the K-means algorithm. The hence obtained k centroids are referred to as the "visual words". All the SIFT descriptors of an image are represented by their respective nearest visual words, i.e. the cluster center of their respective cluster. The images are modeled as a histogram of visual words to represent the corpus as a bag of visual words. In contrast to the classic VBoW model, we adopt an approach where the images are represented as a histogram of latent topics. Latent topics are derived from the visual words using the Latent Dirichlet Allocation (LDA)[1] and are dubbed as the "visual topics". We observe that words mapped to a given visual topic are semantically closer. Hence, it would be more efficient to represent a document using these few visual topics than using the visual words. We find that the proposed LDA based approach exceptionally models the visual features in medical images and is a lot more effective than the PLSA based latent topic modeling used in [3]. We attribute this to the Dirichlet prior on the per-document distribution which prevents the well-known overfitting in PLSA to a large extent.

Each image in the ImageCLEF 2009 dataset is represented by its IRMA (Image Retrieval in Medical Application) code [4]. We extract the textual words from the last two axes (A and B) which describe the region of the body and the biological system to which the part in the image belongs. We assume that the query has both the textual and visual features, in case it lacks the textual description, then only the visual features will be used for the retrieval.

Fig. 1(a) depicts the early fusion approach. In this approach, the co-occurrence matrix of the visual topics and the associated textual words are generated for all the images. As this co-occurrence matrix is most often sparse, it cannot be used as a feature on its own. We try to extract the relevant features from this using an autoencoder and a VGG-16 [5] based model. In the autoencoder model, we use the output of the bottleneck layer of the trained autoencoder as the compact feature representation of the inputs. In the VGG-16 based model, we use the first 4 layers of the VGG-16 model to obtain a compressed representation of the co-occurrence matrix. The image similarity computed is the Euclidean distance between the obtained compressed image features. Let x_i denote the i^{th} document in the corpus and let q denote the query, then their distance (or dis-similarity) in the early fusion approach $\mathcal{D}_e(x_i, q)$ is given by

$$\mathcal{D}_e(x_i, q) = \|\mathcal{F}_e(x_i) - \mathcal{F}_e(q)\|_2 \tag{1}$$

where $\mathcal{F}_e(d)$ denotes the compressed early fusion feature of the document d.


Corpus
Query
Preprocessing
Preprocessing
Co-occurence matrix
Co-occurence matrix
Autoencoder / VGG
Autoencoder / VGG
Final features
Final features
Compute Similarity Score (System Relevence)
Retrieve the documents in order
Corpus
Query
Preprocessing
Preprocessing
Visual Topics
Textual Words
Visual Topics
Textual Words
Visual Distance
Textual Distance by predicting the class of the query
Compute Similarity Score (System Relevence) = visual distance + 0.01 * textual distance
Retrieve the documents in order

(a) Early Fusion Process

(b) Late Fusion Process

Figure 1: Proposed Visual Topic Modeling based Approach for Multi-modal Medical Image Retrieval

Fig. 1(b) depicts the late fusion approach which is an ensemble model that builds on both the visual and textual features. Here, we leverage on the class information provided in the ImageCLEF 2009 dataset[6]. We train a Support Vector Machine (SVM) with RBF kernel to predict the class of the image when presented with just its textual features. For a query q and document x_i, the visual distance $\mathcal{D}_v(x_i, q)$ is defined as

$$\mathcal{D}_v(x_i, q) = \|\mathcal{F}_v(x_i) - \mathcal{F}_v(q)\|_2 \tag{2}$$

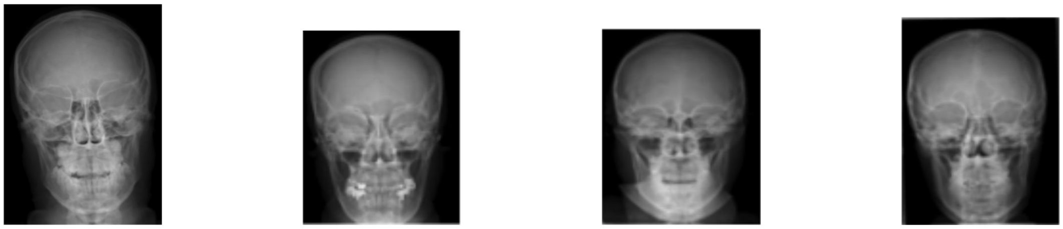
where $\mathcal{F}_v(d)$ denotes visual feature of the document d i.e. the latent topics of document d. This is supplemented with the textual distance $\mathcal{D}_t(x_i, q)$ given by

$$\mathcal{D}_t(x_i, q) = (\mathcal{C}(x_i) - \mathcal{C}(q))^2 \tag{3}$$

where $\mathcal{C}(d)$ denotes the class predicted for the document d by the text-feature based SVM. We define the textual loss as in equation 3 since labels which are closer to each other represent semantically similar classes in the ImageCLEF 2009 datatset. Thus, the total distance $\mathcal{D}_l$ for a document x_i and query q in late fusion approach $\mathcal{D}_l(x_i, q)$ is given by

$$\mathcal{D}_l(x_i, q) = \mathcal{D}_v(x_i, q) + \lambda\mathcal{D}_t(x_i, q) \tag{4}$$

3 Experimental Results



(a) TEXT: skull query

(b) First image retrieved

(c) Second image retrieved

(d) Third image retrieved

Figure 2: A sample query and top 3 images retrieved

As mentioned before, we used the ImageCLEF 2009 dataset, a standard and open dataset used widely by several state-of-the-art works, for validating the proposed medical image retrieval task. We extensively evaluated the retrieval performance using several standard metrics such as mAP, gmAP, precision@k and NDCG (Normalized Discounted Cumulative Gain). The metric mAP is calculated from the area under curve (AUC) of the precision-recall curve (using the 11 point scale), averaged over 100 random queries. GmAP also is calculated similar to mAP, it represents the geometric mean taken over 100 random queries, whereas the mAP represents the arithmetic mean. Precision@k is a metric that is used to judge the relevance of the top few documents returned by the model as highly similar to the query image. We performed experiments to evaluate the $top - k$ retrieval performance at k= 5, 10 and 20, using the precision@5 (p@5), precision@10 (p@10) and precision@20 (p@20) metrics. NDCG is important as it assigns a higher score for a ranking list with relevant documents at the top and the non-relevant documents at the bottom, thus can be used to judge the quality and performance of similar image retrieval of the proposed method. The observations have been summarized in Tables 1 and 2 as well as in Fig. 3 and Fig. 4. A sample query and the corresponding top 3 documents retrieved are shown in Figure 2. The value of λ in equation 4 was experimentally determined to be 0.01.

We compare our work with the state-of-the-art model proposed by Cao et al [3] who also used the ImageCLEF 2009 dataset. Our proposed approach outperformed Cao et al [3] on both fronts - just visual model and multimodal feature model. Cao et al [3] reported that their visual approach obtained an mAP of **0.0101** which less when compared to the mAP of **0.283** achieved by our model. This clearly indicates that our LDA based vocabulary modeling was much more effective in representing the visual features than PLSA which is used by Cao et al [3]. Our proposed Late fusion multimodal approach obtained an mAP of **0.326** which outperformed Cao et al's [3] multimodal approach, with its mAP of **0.2909**.

Table 1: mAP, gmAP and NDCG on the ImageCLEF2009 using our various approaches

Approach	mAP	gmAP	NDCG
Visual features only	0.283	**0.226**	0.702
Early fusion using the Co-occurrence matrix	0.155	0.138	0.587
Early fusion using the Autoencoder features	0.16	0.143	0.596
Early fusion using the VGG features	0.162	0.145	0.61
Late fusion	**0.326**	0.159	**0.722**

Table 2: precision@5 and precision@10, precision@20 using our various approaches

Model	p@5	p@10	p@20
Visual features only	**0.567**	**0.5**	**0.465**
Early fusion using the Co-occurrence matrix	0.264	0.169	0.122
Early fusion using the Autoencoder features	0.268	0.165	0.123
Early fusion using the VGG features	0.27	0.187	0.134
Late fusion	0.526	0.466	0.445

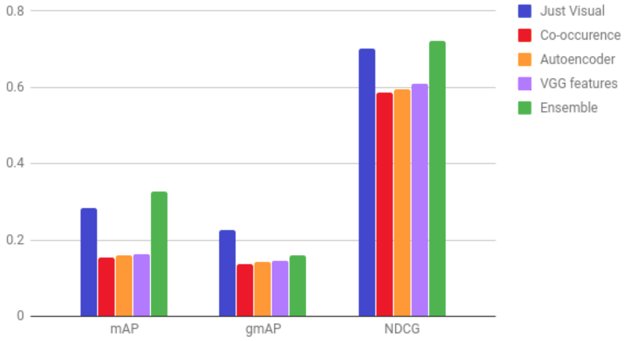


Figure 3: Comparison of mAP, gmAP and NDCG

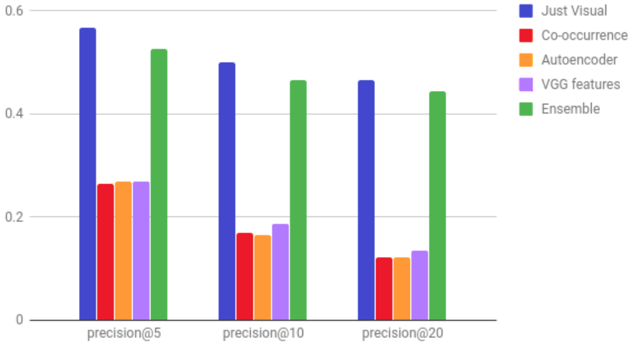


Figure 4: Comparison of precision@5 and precision@10, precision@20

4 Concluding Remarks

We conclude that the LDA based visual topics model the medical images exceptionally well. We explore multimodal medical image retrieval approaches that incorporate both visual and textual features and improve the retrieval performance by leveraging the textual features. We observe that the semantic relationships between textual and visual words play a significant role in effective retrieval, hence, we intend to explore this avenue further.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[3] Yu Cao, Shawn Steffey, Jianbiao He, Degui Xiao, Cui Tao, Ping Chen, and Henning Müller. Medical image retrieval: a multimodal approach. *Cancer informatics*, 13:CIN–S14053, 2014.

[4] Thomas Martin Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B Wein. The irma code for unique classification of medical images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 5033, pages 440–452. International Society for Optics and Photonics, 2003.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] William Hersh, Henning Müller, and Jayashree Kalpathy-Cramer. The imageclefmed medical image retrieval task test collection. *Journal of Digital Imaging*, 22(6):648, 2009.