Multimodal Multitask Deep Learning for X-Ray Image Retrieval

revision Yu[1], Peng Hu[2], Jie Lin[1], and Pavitra Krishnaswamy[1]

[1] Institute for Infocomm Research, A*STAR, Singapore
[2] College of Computer Science, Sichuan University, Chengdu, Sichuan, China
{yu_revision,pavitrak}@i2r.a-star.edu.sg

Abstract. Content-based image retrieval (CBIR) is of increasing interest for clinical applications spanning differential diagnosis, prognostication, and indexing of electronic radiology databases. However, meaningful CBIR for radiology applications requires capabilities to address the semantic gap and assess similarity based on fine-grained image features. We observe that images in radiology databases are often accompanied by free-text radiologist reports containing rich semantic information. Therefore, we propose a Multimodal Multitask Deep Learning (MMDL) approach for CBIR on radiology images. Our proposed approach employs multimodal database inputs for training, learns semantic feature representations for each modality, and maps these representations into a common subspace. During testing, we use representations from the common subspace to rank similarities between the query and database. To enhance our framework for fine-grained image retrieval, we provide extensions employing deep descriptors and ranking loss optimization. We performed extensive evaluations on the MIMIC Chest X-ray (MIMIC-CXR) dataset with images and reports from 227,835 studies. Our results demonstrate performance gains over a typical unimodal CBIR strategy. Further, we show that the performance gains of our approach are robust even in scenarios where only a subset of database images are paired with free-text radiologist reports. Our work has implications for next-generation medical image indexing and retrieval systems.

Keywords: Content-based image retrieval · Multimodal Representation Learning · Fine-Grained Retrieval · Diagnostic Radiographs.

1 Introduction

The growth in large-scale electronic radiology databases presents rich opportunities for clinical decision support. In particular, multimedia databases such as the Radiology Information System (RIS) and the Picture Archiving and Communication Systems (PACS) store anatomical, pathological and functional information for millions of patients. As such, there is increasing interest in content-based image retrieval (CBIR) capabilities to search these databases and retrieve relevant cases for a range of diagnostic, research and educational purposes [2, 3, 15, 22].

Specific use cases of interest could include differential diagnosis, enhanced assessment of rare conditions, severity assessment, and prognostication based on actualized outcomes in patients with similar images in the database [3]. As manual CBIR is often infeasible, automated CBIR methods are desirable.

Automated CBIR solutions have had substantive successes for forensics, retail, mobile and photo archival applications [25]. However, their uptake in medicine has been limited [2, 3]. While some studies have showed applicability of CBIR for histopathology images [10], there have been far fewer demonstrations of CBIR in the radiology domain. In part, this is because radiology images are often generated by non-light based contrast mechanisms and do not contain color information. Further, meaningful CBIR in radiology requires capabilities to extend beyond gross pixel-based features and assess similarity of semantic content based on domain knowledge or on fine-grained details in the image.

We note that images in radiology databases are often accompanied by textual reports detailing radiologist interpretation and observations. Therefore, we propose a multimodal CBIR framework to leverage the rich semantic information in textual reports alongside visual (and semantic) information in the images. We build on deep learning based multimodal retrieval approaches developed for computer vision applications [12, 13, 30, 31] and adapt these for training with radiology images and text reports. Our framework is amenable to addressing both image queries and multimodal queries, and can be customized to address the fine-grained nature of CBIR in radiology [7, 8, 10, 11, 29]. Our main contributions are as follows:

1. We introduce a customizable deep learning framework that leverages multimodal databases comprising radiology images and free-text reports for content-based image retrieval in radiology.
2. To enhance our framework for fine-grained image retrieval, we present extensions that (a) learn descriptors of abnormal regions in the image and (b) employ triplet loss for metric learning.
3. In extensive experiments on MIMIC-CXR, a real-world radiology multimodal dataset, we demonstrate good performance gains over standard baselines. We further show that these gains remain robust in scenarios where a subset of images lack associated textual reports.

2 Related Work

CBIR in Radiology: Early CBIR efforts in radiology employed handcrafted features [19,20,23]. Other early efforts focused on using regions of interest defined by clinical users to identify database images with visually similar patterns [1]. Recent efforts have explored the more generalizable deep learning based CBIR techniques for retrieval based on modality and/or body part similarity [2, 4, 24, 27], and for retrieval on chest X-Rays [5, 8]. However, such methods are yet to be demonstrated for challenging clinical retrieval use cases with limited user input, and are not set up to fully leverage the multimedia information available in radiology databases.

Multimodal Retrieval: There has been longstanding interest in integrating multimodal information to inform medical image retrieval [2,3,15,22]. One study proposed a probabilistic latent semantic analysis method to combine images with short textual descriptors for a modality and body-part similarity retrieval task [4]. However, this approach is limited to short descriptors for the textual modality, focused on single task learning and is not suitable for use cases like differential diagnosis which only allow unimodal query inputs. In contrast, the computer vision literature has advanced deep learning techniques to effectively learn semantic representations from a variety of multimodal data types to enhance CBIR [12, 13, 21, 30, 31]. While these advanced deep learning methods have been demonstrated on natural scene image datasets, they have yet to be translated to domain-specific tasks.

Fine-Grained Image Retrieval: While typical image retrieval approaches focus on macro-level similarities between query and database, fine-grained retrieval seeks to rank images based on similarities in more subtle features. To address this challenge, prior works have employed either attention-based mechanisms or ranking loss terms. Attention-based mechanisms (e.g., Selective Convolutional Descriptor Aggregation (SCDA) [29]) localize the objects of interest by discarding the noisy background and keeping useful deep descriptors. Ranking loss approaches employ triplets of samples to learn relative distances between samples and optimize metric learning [7, 9, 11], and can be beneficial for fine-grained image retrieval [28]. A recent study [8] employed an Attention-based Triplet Hashing (ATH) approach that uses deep triplet loss and a fine-grained attention mechanism for chest X-ray retrieval. However, these approaches have yet to be demonstrated in domain-specific tasks and/or in multimodal retrieval settings.

3 Methods

Problem Formulation: We consider data inputs for the image and text modalities. We denote the set of n_i samples of the i-th modality as $\mathcal{X}_i = \{x_1^i, x_2^i, \ldots, x_{n_i}^i\}$, where x_j^i denotes the j-th data input for the i-th modality. As each data input can be associated with a multiplicity of class labels, the corresponding label matrix for i-th modality is denoted as $\mathcal{Y}_i = [y_1^i, y_2^i, ..., y_{n_i}^i]$. Given a multimodal database $[\mathcal{X}^{\mathrm{D}}, \mathcal{Y}^{\mathrm{D}}]$ and a unimodal image query x^{Q}, the retrieval task is to identify a ranked list of the most similar images from the database.

Multimodal Multitask Deep Learning (MMDL) Framework: Fig. 1 illustrates the proposed MMDL framework. The representation learning task is to learn, for each modality, modality-specific transformation functions. We accomplish this with two feature extraction and encoder networks to project the data inputs from each modality into a common subspace where the intra-class variation is minimized and inter-class variation is maximized.

For the j-th data input, we denote the output of the i-th encoder network as h_j^i. We denote learnable parameters of the i-th modality-specific transformation

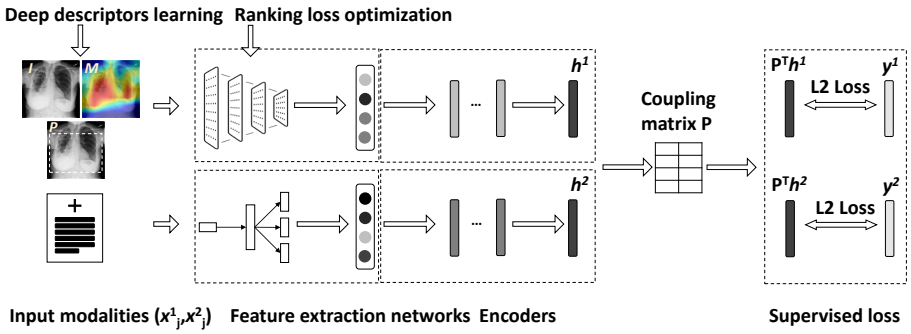

Fig. 1. Proposed Multimodal Multitask Deep Learning (MMDL) Framework. The training framework extracts features from the inputs, reduces dimensionality and projects the multimodal information into a common subspace for semantic representation learning. The encoders and coupling matrix define the common subspace and are learnt during network training. The projected representation is fed into a linear classifier for supervised learning of the multimodal transformation functions. During testing, the query image is transformed into the common subspace so as to rank similar database images for retrieval.

function as Θ_i. Then, the objective function for the i-th encoder network is:

$$\mathcal{L}^i(x_j^i) = \frac{1}{n_i}\sum_{j=1}^{n_i}[||P^T h_j^i - y_j^i||_2], \tag{1}$$

where P is a learnable matrix used to define the common subspace and $\mathcal{L}^i(x_j^i)$ is the supervised loss. The rows of P correspond to the dimensionality of the encoder output while columns of P correspond to the semantic categories. For each modality, the supervised loss $\mathcal{L}$ leverages the labels to enhance the discriminative power of the predefined common space.

To optimize the transform functions for all m modalities, the system needs to simultaneously minimize the total loss across the m modality-specific networks:

$$\min \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathcal{L}^{\mathrm{i}}. \tag{2}$$

Since the trainable parameters are shared amongst all modality-specific networks, the MMDL learns weights Θ_i and P simultaneously in a joint manner. With this learning strategy, MMDL is able to bridge the heterogeneity gap between various modalities and learn common representations by simultaneously preserving the semantic discrimination and modality invariance. Moreover, the optimized networks trained from multimodal learning could further enhance the performance in the multitask implementation where only one of the modalities is available during querying phase.

During testing, we use the encoder outputs to compute the cosine distance between image query $x^{Q,1}$ and each sample in the multimodal database $[\mathcal{X}^D, \mathcal{Y}^D]$. We then rank database entries by the distance to obtain the retrieval output.

Learning Fine-Grained Image Features: To enhance performance of our multimodal multitask deep learning framework for the fine-grained nature of the content in the images, we present two extensions. First, we employ the Selective Convolutional Descriptor Aggregation (SCDA) method [29] to localize the abnormal regions in the images, and learn the deep descriptors for these abnormal regions. As shown in Fig. 1, we employ SCDA on image I to generate attention masks M and then threshold using these masks to generate the image descriptors or patches P, which then serve as the input for the multimodal multitask retrieval. Second, we employ the triplet loss [11] to optimize the metric learning for better distinctions based on subtle image features. Specifically, the ranking loss optimization is applied to enhance the extracted feature representations for the image modality. Together, these enhancements serve to highlight the fine-grained features in the images for more precise retrieval.

4 Experiment Setup

Datasets: We evaluated our multimodal retrieval framework on the MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0. This dataset contains 377,110 images and paired reports corresponding to 64,579 patients with their annotations categorized into 14 classes [17]. We removed the lateral images and only considered frontal images. From the remaining dataset, we randomly sampled 189,036 image-text pairs for training and 26,750 image-text pairs for validation. From the remainder of the dataset, we selected 4,075 image-text pairs for testing. For the test set selection, we used stratified sampling to represent the class distributions of the native dataset. We tested retrieval performance using each test image as a query against the remainder of the test samples as the database. We generated inputs for multimodal retrieval tasks by first extracting the corresponding features from images and their paired reports. We generated the labels for the images and reports using the CheXpert Natural Language Processing (NLP) labeler with the "U-Zero" setting [16].

Feature Representation Learning: For each image (I), we generate a mask (M)-derived patch (P) using the VGG16 model [26] pre-trained on CheXpert [16] with the SCDA approach. The image patches are further represented by a 1,024-dimensional convolutional neural network (CNN) feature extracted from the final fully connected layer of the DenseNet121 model [14] pre-trained on CheXpert using ranking loss optimization. For reports, we used the free-text from "Findings" and "Impression" sections to extract a 300-dimensional text representation via Doc2Vec model [18] pre-trained on the MIMIC-CXR.

Implementation Details: For each modality, the proposed network includes three fully-connected layers with each layer following a Rectified Linear Unit (ReLU) activation function. The three fully-connected layers have 1024, 1024, 512 hidden units and share weights. We randomly initialized P. We used four

Nvidia RTX 2080 Super GPUs for training the proposed MMDL model with PyTorch. Hyperparameters are in Supplement Table S1.

Baselines: We compared our approach to the conventional unimodal retrieval baseline that applies a fine-tuned DenseNet121 network on the image query (Image D121). For this, we first trained the network on ImageNet [6] and fine-tuned on the CheXpert dataset, then extracted features from average pooling layers, and computed similarity using cosine distance to retrieve relevant database images. For more direct and rigorous comparison with our proposed method, we also implemented a unimodal version of our proposed method using the image query where the network is trained only with image modality (Image MMDL).

To further compare with other state-of-the-art (SOTA) unimodal retrieval approaches, we also implemented the ATH method [8]. Specifically, we adapted the ATH pipeline for multi-label inputs and retrained the model on the entire MIMIC-CXR dataset (14 classes) to evaluate performance (ATH). Further, we integrated the ATH method within our MMDL framework and repeated the previous evaluation to obtain a more rigorous comparison of performance (ATH-MMDL).

Experiments: In a first experiment, we focused on training with a database comprising complete image and text pairs, and evaluating with unimodal image queries. For this, we performed retrieval using our proposed method and the above baselines. In a second experiment, we simulated the scenario where not all radiology images in the database would have paired textual reports. We considered scenarios where only 25%, 50%, 75%, or 90% of the images had associated reports, and compared the performance against the completely paired image-text experiment. For these experiments, we first trained the networks with the image-text pairs and then utilized the remaining images to further train the network.

Evaluation Metrics: In each experiment, we evaluated the mean Average Precision (mAP) score for all testing samples and the Average Precision (AP) for each individual class for the Top 10 retrieved images. Retrieval is considered accurate if any of the positive labels in the retrieved image overlap with the query image. We also computed the mAP for the Top 1, 5, 20, and k (where k is number of test queries) retrieved images and the Top 1 and Top 5 Accuracy (Acc-1 and Acc-5 with sigmoid probability of the top 1 and top 5 predicted labels).

5 Results

Fig. 2 shows results for the two aforementioned multimodal retrieval experiments on the MIMIC-CXR dataset. The first experiment focuses on retrieval of database images for a given image query, and evaluates our multimodal multitask framework against the standard baseline unimodal image retrieval approach (Panel A). The second experiment focuses on comparing the effectiveness of retrieval models trained in scenarios where not all database images have accompanying text reports (Panel B).

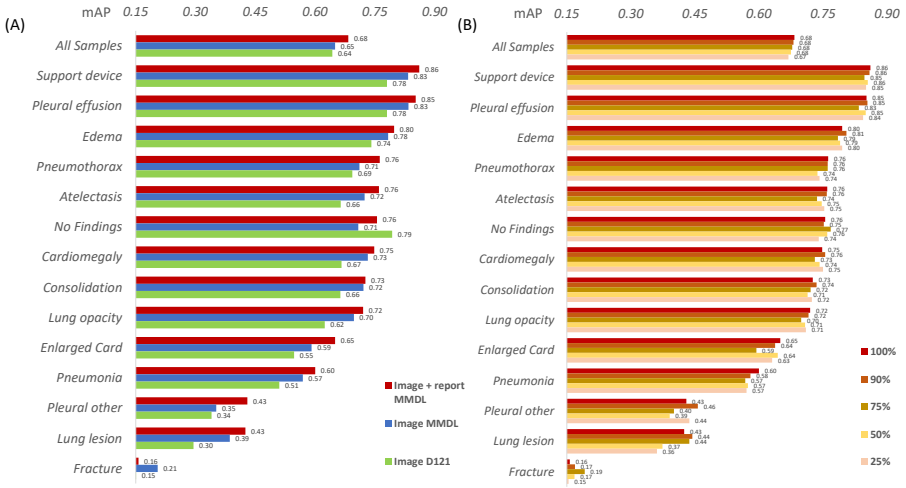

Fig. 2. Performance comparison for proposed MMDL method and baselines: A) All samples mAP and individual class AP for unimodal and multimodal retrieval tasks. (B) All samples mAP and individual class AP for multimodal retrieval in scenarios with incomplete data coverage, as indicated by percentage of images paired with textual reports.

Performance of Multimodal vs. Unimodal Retrieval: Fig. 2 (A) reports mAP of our proposed multimodal method and two unimodal image retrieval baselines (with DensetNet121 and with our MMDL framework). For the unimodal retrieval task, our MMDL framework (Image-MMDL) performs comparably to the standard approach of learning features with DenseNet121 (Image-D121). We note that inclusion of semantic features from the textual reports with our proposed multimodal approach provides a 4% boost in mAP over the conventional DenseNet121-based unimodal retrieval baseline. Supplement Table S2 details additional performance metrics for varying numbers of the top retrieved images. Our results suggest that the weight sharing strategy in our multimodal multitask learning approach is able to enhance semantic consistency amongst the modalities.

Comparison with SOTA Unimodal Baselines: We benchmarked our results against the recent unimodal ATH retrieval baseline and found that the ATH baseline achieves all samples mAP of 0.515. When integrated within our MMDL framework, the ATH-MMDL baseline achieves all samples mAP of 0.541. In both cases,the ATH results are lower than our DenseNet121-based unimodal retrieval implementations (0.643 for Image-D121, and 0.650 for Image-MMDL). Supplement Table S2 also details additional performance metrics for the ATH baseline implementations using varying numbers of top retrieved images.

Ablation Studies: Next, we performed ablation studies to systematically evaluate the effect of each component (multimodal multitask learning, deep descriptors learning, and ranking loss optimization) in our framework. The results,

in supplementary Tables S3 and S4, show that the injection of semantic features via multimodal multitask deep learning (Table S4) provides 3.0% increment in mAP over the comparable unimodal image-only baseline (Table S3); the deep descriptor learning provides 0.4% increment in mAP; while the ranking loss optimization offers 1.1% increment in mAP.

Performance with Incomplete Coverage of Reports: Fig. 2 (B) reports the mAP with MMDL retrieval models trained with incomplete multimodal data. In practical scenarios, only 75-90% of the images might have paired reports. As a worst case, perhaps only 25-50% of the images may have accompanying reports. Intuitively, greater coverage of textual reports should offer higher mAP. Yet, when only 75-90% of the images have associated textual reports, the mAP across classes is maintained within 98.1-99.7% of the ideal 100% coverage scenario. This shows that performance gains from our approach are robust to incompleteness in multimodal data. Even in the worst case setting when 25-50% of the images have associated textual reports, our MMDL approach still offers a 1.9-2.5% boost in mAP over the conventional DenseNet121 unimodal retrieval baseline.

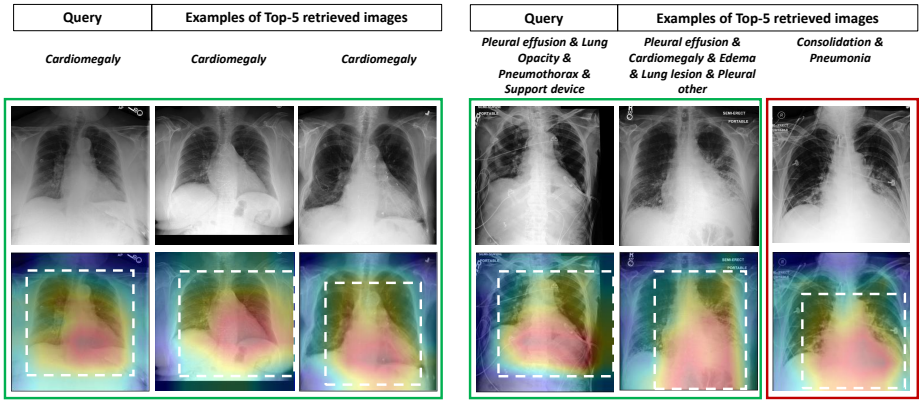

Fig. 3. Exemplar query images with two of the top 5 retrieved cases. In each case, original images are overlaid with attention masks and image patches (white dotted box) generated by the deep descriptor learning step. The correctly retrieved results are marked with green box while incorrectly retrieved results are marked with red box.

Visualization of Retrieved Images: For further characterization, we present exemplar results on 2 queries from the testing set in Fig. 3. For each query, the test image and associated ground truth labels are shown, alongside two of the top 5 retrieved cases and their labels. We observe that the returned images have good overlap in labels with respect to the query images when correctly retrieved. Further, there is good overlap in the visual features between query and returned images despite inherent inter-subject variations. For the second query, although the pleural effusion is inconspicuous, the correctly retrieved example also corresponds to a mild pleural effusion. Furthermore, even for the wrongly retrieved

example, there is a sensible semantic match - specifically the query corresponds to a lung opacity but the returned image corresponds to consolidation and pneumonia which are sub-types of lung opacities. Finally, we note that the full text report for this case mentioned an effusion, although the CheXpert labeler did not infer pleural effusion as it was based on the impression section only [16]. These illustrations highlight the potential of the proposed method to identify semantically similar X-Ray images, especially when accurate retrieval relies on fine-grained features.

6 Conclusion

We proposed a deep learning framework that learns common semantic representations for CBIR based on a combination of radiology images and textual reports. Comprehensive experimental results on a public Chest X-Ray dataset demonstrate the effectiveness of our framework to boost retrieval performance. Future work to improve report-based feature extraction could offer additional performance improvements. Our work establishes a new baseline for content-based medical image retrieval and has implications for practical clinical use cases in differential diagnosis, cohort selection and prognostication.

Acknowledgements: Research efforts were supported by funding and infrastructure for deep learning and medical imaging research from the Institute for Infocomm Research, Science and Engineering Research Council, A*STAR, Singapore. We thank Victor Getty, Vijay Chandrasekhar and Ivan Ho Mien from the Institute for Infocomm Research, A*STAR for their valuable inputs. We also acknowledge insightful discussions with Jayashree Kalpathy-Cramer at the Massachusetts General Hospital, Boston, USA.

References

1. `https://contextflow.com/`
2. Ahmad, J., Sajjad, M., Mehmood, I., Baik, S.W.: Sinc: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. PloS one **12**(8), e0181707 (2017)
3. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: current status and future directions. Journal of digital imaging **24**(2), 208–222 (2011)
4. Cao, Y., Steffey, S., He, J., Xiao, D., Tao, C., Chen, P., Müller, H.: Medical image retrieval: a multimodal approach. Cancer informatics **13**, CIN–S14053 (2014)
5. Chen, Z., Cai, R., Lu, J., Feng, J., Zhou, J.: Order-sensitive deep hashing for multimorbidity medical image retrieval. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 620–628. Springer (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

7. Dodds, E., Nguyen, H., Herdade, S., Culpepper, J., Kae, A., Garrigues, P.: Learning embeddings for product visual search with triplet loss and online sampling. arXiv preprint arXiv:1810.04652 (2018)
8. Fang, J., Fu, H., Liu, J.: Deep triplet hashing network for case-based medical image retrieval. Medical Image Analysis **69**, 101981 (2021)
9. Gómez, R.: Understanding ranking loss, contrastive loss, margin loss, triplet loss, hinge loss and all those confusing names. Raúl Gómez blog (2019)
10. Hegde, N., Hipp, J.D., Liu, Y., Emmert-Buck, M., Reif, E., Smilkov, D., Terry, M., Cai, C.J., Amin, M.B., Mermel, C.H., et al.: Similar image search for histopathology: Smily. NPJ digital medicine **2**(1), 1–9 (2019)
11. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
12. Hu, P., Peng, D., Wang, X., Xiang, Y.: Multimodal adversarial network for cross-modal retrieval. Knowledge-Based Systems **180**, 38–50 (2019)
13. Hu, P., Zhen, L., Peng, D., Liu, P.: Scalable deep multimodal learning for cross-modal retrieval. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 635–644 (2019)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
15. Hwang, K.H., Lee, H., Choi, D.: Medical image retrieval: past and present. Healthcare informatics research **18**(1), 3 (2012)
16. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 590–597 (2019)
17. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
18. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014)
19. Liu, Y., Rothfus, W.E., Kanade, T.: Content-based 3d neuroradiologic image retrieval: Preliminary results. In: Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database. pp. 91–100. IEEE (1998)
20. Ma, L., Liu, X., Gao, Y., Zhao, Y., Zhao, X., Zhou, C.: A new method of content based medical image retrieval and its applications to ct imaging sign retrieval. Journal of biomedical informatics **66**, 148–158 (2017)
21. Mourão, A., Martins, F., Magalhaes, J.: Multimodal medical information retrieval with unsupervised rank fusion. Computerized Medical Imaging and Graphics **39**, 35–45 (2015)
22. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. International journal of medical informatics **73**(1), 1–23 (2004)
23. Pilevar, A.H.: Cbmir: Content-based image retrieval algorithm for medical image databases. Journal of medical signals and sensors **1**(1), 12 (2011)
24. Qayyum, A., Anwar, S.M., Awais, M., Majid, M.: Medical image retrieval using deep convolutional neural network. Neurocomputing **266**, 8–20 (2017)
25. Schaer, R., Otálora, S., Jimenez-del Toro, O., Atzori, M., Müller, H.: Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. Journal of pathology informatics **10** (2019)

26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
27. Sklan, J.E., Plassard, A.J., Fabbri, D., Landman, B.A.: Toward content-based image retrieval with deep convolutional neural networks. In: Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging. vol. 9417, p. 94172C. International Society for Optics and Photonics (2015)
28. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1386–1393 (2014)
29. Wei, X.S., Luo, J.H., Wu, J., Zhou, Z.H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing **26**(6), 2868–2881 (2017)
30. Zhen, L., Hu, P., Peng, X., Goh, R.S.M., Zhou, J.T.: Deep multimodal transfer learning for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems (2020)
31. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10394–10403 (2019)