Multi-label Concept Classification in Imaging Entities of Biomedical Literature using CNN and Vision Transformers

Md Mahmudur Rahman, Bikesh Regmi
Computer Science Department
Morgan State University, Baltimore, Maryland
md.rahman, bireg1@morgan.edu

***Abstract*— Biomedical images are frequently used in articles to illustrate medical concepts and highlight regions-of-interests (ROIs) by using annotation markers (pointers) such as different arrows, letters or symbols overlaid on figures. Also, in many cases multiple markers in the same image are often pointing to different concepts relevant to the article. Hence, each image might be assigned with one or more concepts for multi-label classification and object detection based machine-learning tasks. This work reports such a proof-of-concept (POC) experiment by annotating ROIs and classifying (multi-label classification) 200 Chest CT images appeared in biomedical articles with eleven (11) different concept (similar to UMLS) categories such as, ground-glass, bronchi, honeycomb, cyst, nodules, etc. For annotation, we use an online tool (Labelimg) to annotate image ROIs with concepts based on the information content in associated captions. To demonstrate the feasibility of the POC, this study conducts experiments with different Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) using both transfer learning (fine-tuning) and training from scratch. We achieved encouraging results (around 70% micro average precision and recall accuracies) in a test set, whereas the dataset images are in very low resolution, non-uniform lighting conditions and with varied shapes and sizes. Overall, this study demonstrates the effectiveness of deep learning models in multi-label classification in medical images and establishes the feasibility and rationale of the POC. The ultimate goal of this work is to develop a large-scale concept detection framework towards building a visual ontology of images in biomedical articles.**

Keywords—biomedical concepts, classification, multilabel classification, image retrieval, evaluation

I. INTRODUCTION

Due to the ongoing progress in biomedical domain, a wide variety of users, such as patients, researchers, general practitioners, and clinicians often use tools to search for relevant and actionable information from biomedical literature to their individual needs. It creates the need of literature-based informatics for managing the rapidly increasing volume of information in the biomedical domain [1]. For example, according to PubMed Central® (PMC), a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) alone contains more than 8 million articles with an average 3-5 figures as of 2022 [2].

Due to the rapid pace of scientific discovery in medical domain, we are witnessing an exponential growth of biomedical literature for the past few decades. Hence, it is becoming increasingly difficult to search for information in the right place at the right time within large volumes of literature [3, 4]. Until now, little attention is devoted to the use of images in the articles, as the meaning of images cannot be understood by analyzing their content alone. In general, a system searching for images within a collection of biomedical articles commonly represents and retrieves them based on the collateral text, such as captions [5-7]. For example, a basic search in PMC will look at all image captions in the database and retrieve images related to the query topic. Until now, little attention is devoted to the use of images in the articles, as the meaning of images cannot be understood by analyzing their content alone. However, biomedical articles convey information using multiple and distinct modalities, including text and images. Also, the diverse modalities (e.g., X-ray, CT, MRI, US, etc.) constitute an important source of anatomical and functional information for clinical purpose, research, and education. Given that, images are such a crucial source of information within the biomedical domain, using visual features for image classification and search has gained significant popularity during the past three decades [8].

Authors of biomedical articles often use arrows, pointers, and other annotations such as text labels overlaid on figures and illustrations in the biomedical articles to highlight significant areas as ROIs.

979-8-3503-2759-5/23/$31.00 ©2023 IEEE
DOI 10.1109/CSCE60160.2023.00205

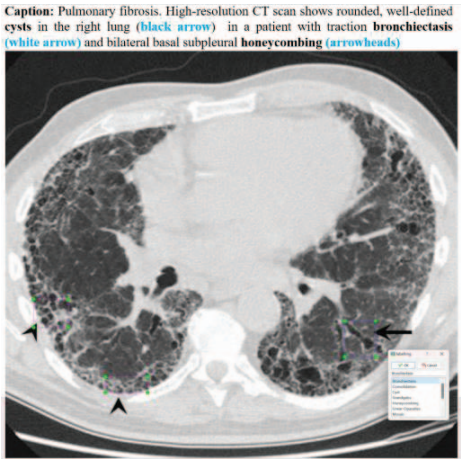

Figure 1: Example Chest CT image with associated caption [23]

These annotations are then referenced and correlated with concepts in the caption text or figure citations in the article text. The pointers/markers are good source of information to locate specific visual patterns as ROIs in the image [9]. For example, Fig. 1 shows different arrows/arrowheads pointing to ***honeycomb*** (hexagonal wax cells like structure) and ***bronchial*** (diffuse thickening of the airway walls giving the appearance of thick lines and rings throughout the lungs) and ***cyst*** patterns in a chest CT image (along with caption at the top) as depicted in an article "Idiopathic pulmonary fibrosis" in the Orphanet Journal of Rare Diseases [10]. These annotation markers (e.g., arrows) can assist in extracting relevant ROIs within the image that are likely to be highly relevant to the discussion in the article text. Image regions can then be annotated using biomedical concepts from extracted snippets of text in captions that might be further identified using existing textual ontologies, such as the Unified Medical Language System (UMLS) [11] or RadLex [12]. Such a resource could assist in reducing the semantic gap problem in image classification and retrieval [13].

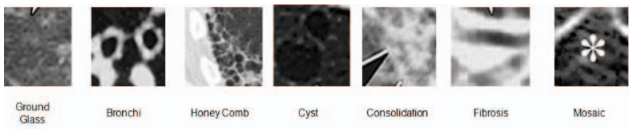

Figure 2: Example visual concept (ROIs) in the thoracic imaging glossary [12]

However, these images are always in low-resolution compared to their clinical counterparts, in varying sizes and lightning conditions and moreover the dataset is highly imbalanced where a few concept categories (patterns) occur more frequently compared to other less frequent ones. Hence, although currently smaller, the dataset still might be considered as a realistic set for evaluating medical image classification and retrieval techniques for images in biomedical articles.

This work presents a proof-of-concept experiment by annotating ROIs and classifying 200 Computed tomography (CT) images of the chest, which appears in biomedical articles obtained through a large benchmark collection [14]. The images are classified into eleven (11) different concept (similar to UMLS) categories such as, ground glass, bronchi, honeycomb, cyst, nodules, etc. The concept detection is considered here as a multi-label classification problem, which involves predicting zero or more concept labels/categories to each image instance. To demonstrate the utility, we trained several CNNs and ViTs based classifiers to automatically assign thoracic imaging concepts to image regions based solely on their appearance. Thus, our methods are capable of automatically mapping the appearance of visual entities within images to a limited set of concepts or terms as shown in Fig. 2, which are in the "imaging observation" (RID5) branch of the RadLex tree [12], included "ground-glass opacity" (RID28531), bronchiole (RID1298), cyst (RID3890), and "honeycomb" (RID35280), among others. We evaluated our methods in the small gold standard set of annotated image regions and descriptions with encouraging results, which are the first steps towards the creation of a visual ontology of biomedical imaging entities.

II. MULTI LABEL CLASSIFCATION

One of the most used capabilities of ML techniques in scientific literature is for classifying content, where each data instance or document is assigned to a class from the set of a priori known classes and the task may be divided into three domains, binary classification, multiclass classification, and multilabel classification. The binary and multiclass classification approaches are well known and widely used in supervised ML, such as text categorization, sentiment and emotion recognition, image classification and object detection, etc. However, many of real life data sets (such as the image in Fig. 1) are too complex to impose the restriction of only one category or label for each data instance. The difference between multi-class and multi-label classification is that the classes are mutually exclusive for the first one, whereas multi-label classification or multi-output classification is a variant of the classification problem where multiple nonexclusive labels may be assigned to each instance.

There are two main methods for tackling a multi-label classification problem [15]. The first kinds are problem transformation methods, which transform the multi-label problem into a set of binary classification problems so that those can be handled using single-class classifiers. On the other hand, algorithm adaptation methods try to address the problem in its full form for directly performing multi-label classification by adapting the algorithms. Most traditional learning algorithms are developed for single-label classification problems [43]. Therefore, a lot of approaches in the literature transform the multi-label problem into multiple single-label problems, so that the existing single-label algorithms can be used.

This work focuses on the first approach of multi-label classification (e.g., detection task is divided into a series of multiple binary classification problems) methods using both deep CNNs and ViTs. Since, it is not feasible to use the stastandard LabelBinarizer class for multi-class classification, the scikit-learn library's MultiLabelBinarizer class is used, which transforms the labels of each image to a vector with a total eleven (11) categories where one-hot encoding transforms categorical labels from a single integer to a vector. The concept labels of images in the dataset are extracted from a manually annotated csv file and added in a list.

III. CNN AND VISUAL TRANSFORMER (VIT)

Over the past decade, Deep Neural Networks (DNNs) and more specifically CNNs have shown state of the art performances in different medical image analysis tasks, such as disease classification, tumor segmentation, and lesion detection [16]. A convolutional layer in CNN is characterized by sparse local connectivity and weight sharing and are often followed by a non-linear activation, pooling, and fully connected (or dense) layers.

A. CNN Architectures

Three well known and popular CNN architectures, such as Xception [17], DenseNet-121 [18], and ResNet-50 [19] are experimented in this study by training the dataset both from scratch and using transfer learning (TL) with fine tuning approach. Xception and ResNet networks use skip connections and multiple convolutional and max-pooling blocks in each layer. The training of the CNNs is performed by minimizing a loss function using gradient descent-based methods and backpropagation of the error with following configuration:

- Number of nodes in the output layer matches the number of labels.
- Keep activation function of the classification (output) layer in our models to sigmoid, which enables to perform multi-label classification with Keras.
- Treat each output label as an independent Bernoulli distribution and to penalize each output node independently, the ***binary cross-entropy*** loss function is used rather than the commonly used ***categorical cross-entropy***.

B. Vision Transformer (ViT)

Transformers are a type of DL architecture, based primarily upon the self-attention module, that were originally proposed for language translation task in NLP. Recent works have shown that transformers can fully replace the standard convolutions in DL networks by operating on a sequence of image patches, giving rise to ViTs [20]. These ViT models continue the long-lasting trend of removing hand-crafted visual features and inductive biases from models to leverage the availability of larger datasets coupled with increased computational capacity. Capitalizing on these advances in Computer Vision, the medical imaging field has also witnessed growing interest for Transformers in segmentation, detection, classification, reconstruction, synthesis, registration, clinical report generation, and other tasks [21]. Being inspired by the success of ViTs in computer vision and its application in medical imaging fields in recent years [21,22], this work also experimented with different ViT models trained from both scratch and fine-tuned for classification using transfer learning. The set-up of the ViT encoder consists of a patch encoder that receives an input of 224 x 224 images and produces a dense projection of the patches and a positional embedding representative of each patch's locality. 14 x 14 x 3 overlapping patches are obtained from the image (e.g., 256 patches per image) and are flattened into a 588-D array. The Patch Encoder layer linearly transforms a patch by projecting it into a vector of size projection_dim = 64-D in the experiment.

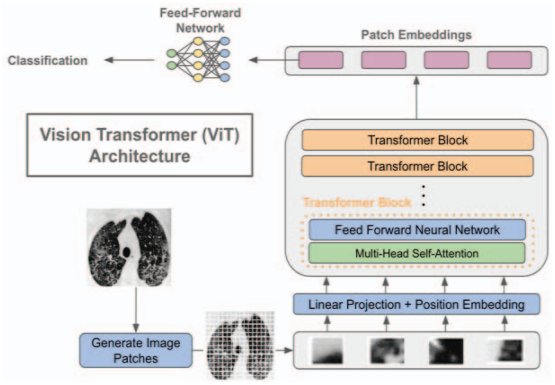

Figure 3: Process flow diagram for ViT based classification

As shown in Fig. 3, the patch projections and positional embeddings are fed into the encoder stack which is a layer of 4 encoders consisting of a multi-head attention layer that provides the attended representation of the features, a skip connection and an intermediate dense layer that projects the visual feature representation into the specified dimension size. The Transformer blocks produce a [*batch_size* =8, *num_patches* =256, *projection_dim* =64] tensor, which is processed via the classifier head with sigmoid (like CNN classification) to produce the final class probabilities output.

IV. EXPERIMENTS AND RESULTS

To validate the assumptions of the proof-of-concept, we experimented with a manually annotated ground truth dataset of 200 lung CT images with eleven different concept labels (Fig. 4), which is a subset of images under a much larger ImageCLEFmed benchmark [14].

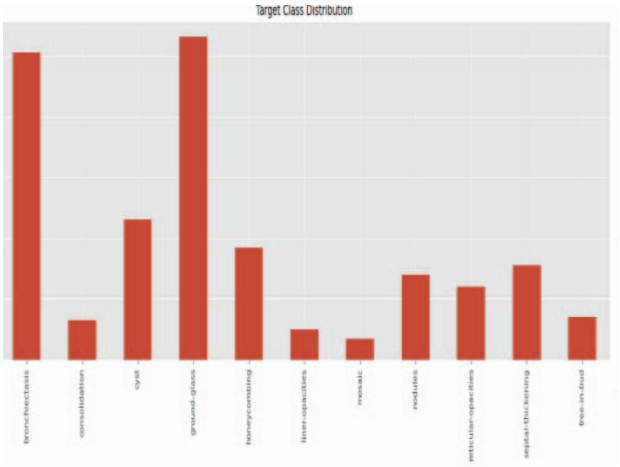

Figure 4: Distribution (frequency) of eleven concept categories.

image	labels
31922.jpg	nodules bronchiectasis septal-thickening
31924.jpg	bronchiectasis
31925.jpg	bronchiectasis mosaic
35859.jpg	bronchiectasis ground-glass
35860.jpg	bronchiectasis ground-glass nodules
35861.jpg	bronchiectasis ground-glass
35862.jpg	ground-glass bronchiectasis septal-thickening consolidation
35863.jpg	bronchiectasis septal-thickening nodules ground-glass
38765.jpg	bronchiectasis septal-thickening consolidation

Figure 5: Example multilabel annotation in the csv file.

The annotation is saved in a csv file where each image is associated with one or more labels (Fig. 5). Also, an online annotation tool, ***LabelImg*** [23] is used to annotate the ROIs (with coordinates information) and annotations (Fig. 1) are saved as XML files in PASCAL VOC format and YOLO text file format for future exploration of object (ROI) detection using techniques, such as different versions of R-CNN and YOLO algorithms.

There exists around one-fourth images (out of 200) in this dataset (Fig. 5) , which contain only a single label (category), such as 18 images with a label "cyst", 16 images with label "bronchiectasis" and 14 images with "ground-glass" label. Since, it's a small dataset currently, almost half of the multi-labels occurred only once in the dataset (Fig. 5), hence makes the training and model generation difficult.

A. Pre-processing with Image Augmentation

All dataset images are resized to 224 x 224 (except 299 x 299 for Xception model) pixels and scaled the raw pixel intensities to the range [0, 1] and stored as NumPy arrays. After that, labels are binarized for multi-class classification by utilizing the scikit-learn library's MultiLabelBinarizer class, which actually transforms the concept labels into a vector that encodes which concepts are present in the image. The high imbalance in the label frequency results in a huge bias towards the multilabel classification problem. Hence, data augmentation (scaling, rotation, flipping, etc.) is also applied while training as we have only a handful of images per concept class. The images are randomly rotated (25 degrees), horizontally and vertically shifted by a factor of 0.2, sheared by 0.2, and randomly horizontally flipped.

The goal of applying data augmentation is to increase the generalizability of the model. Applying a (small) amount of these transformations to an input image will change its appearance slightly, but it does not change the class label – thereby making data augmentation a very natural, easy method to apply to deep learning for computer vision tasks. The dataset is divided into random training (80%) and testing (20%) subsets where different accuracies are measured in the testing sets to compare different models and feasibility of the classification.

B. Train the model

All the models (CNNs and ViTs) are built by initializing the Adam optimizer and compiled using binary cross-entropy rather than categorical cross-entropy to treat each output label as an independent Bernoulli distribution where the labels are not disjoint. After training is complete, the models and label binarizes are saved to disk and loaded later during prediction in the test set. For training of the models from scratch, a learning rate = 0.001 and for pre-trained models a learning rate = 0.0001 is used and all the models are trained with 100 epochs with batch size = 8.

C. Result Analysis

For evaluating the performances of different models, measuring simple accuracy is not sufficient when working with a class-imbalanced data set, like this one, where there is a significant disparity between the class labels. Hence, aggregate metrics like ***macro, micro, weighted*** and ***sampled avg*** are calculated as those give us a high-level view of how the models are performing.

For example, Fig. 6 shows the macro, micro, weighted and sampled avg precision, recall and F1-scores for the test samples based on using Xception model and training the model from scratch.

	precision	recall	f1-score	support
bronchiectasis	0.71	0.65	0.68	23
consolidation	0.00	0.00	0.00	2
cyst	0.50	0.44	0.47	9
ground-glass	0.76	0.65	0.70	20
honeycombing	0.80	0.50	0.62	8
liner-opacities	0.00	0.00	0.00	1
mosaic	0.00	0.00	0.00	3
nodules	0.33	0.33	0.33	3
reticular-opacities	0.50	0.20	0.29	5
septal-thickening	0.20	0.11	0.14	9
tree-in-bud	1.00	0.67	0.80	3
micro avg	0.64	0.48	0.55	86
macro avg	0.44	0.32	0.37	86
weighted avg	0.59	0.48	0.52	86
samples avg	0.62	0.47	0.51	86

Figure 6: Classification accuracy (test set) report for the Xception model

The low avg. accuracies (in the range of 45-65%) are because the dataset size is currently small and there is not simply enough

representation of different concept labels in these low-resolution and highly varied images. As can be observed in Fig 7, the classifier even obtained zero (0) precision, recall and F1-scores for three concept labels (e.g., consolidation, linear opacities, and mosaic patterns).

TABLE I. ACCURACY IN TEST SET FOR DIFFERENT MODLE CONFIGURATIONS

Method	Micro_Avg Precision	Micro_Avg Recall	Micro_Avg F1-Score	Weighted_ Avg_F1-Score
DenseNet-121 (Scratch)	0.70	0.44	0.54	0.52
DenseNet-121 (Pre-trained)	0.53	0.48	0.50	0.34
Xception (Scratch)	**0.71**	0.52	**0.60**	**0.57**
Xception (Pre-trained)	0.56	0.31	0.40	0.37
ResNet-50 (Scratch)	0.59	0.37	0.46	0.33
ResNet-50 (Pre-trained)	0.59	0.47	0.52	0.41
ViT (Scratch)	0.65	0.37	0.47	0.40
ViT_B_16 (Pre-trained)	0.30	**0.70**	0.42	0.52
ViT_l_32 (Pre-trained)	0.35	0.66	0.46	0.53

Table I shows the aggregate metrics, such as micro avg precision, recall, and F-scores and also weighted avg F-scores for different classifiers based on using CNN and ViT models and training both from scratch and fine tuning with TL. For pre-trained ViTs, both the ViT-Small model (***ViT-B/16***) and ViT-Large model (***ViT-L/32***) from original paper [20] are used. It is observed from Table 1 that Xception model (scratch) performed better compared to other models in terms of micro avg precision, and micro and weighted avg F1-scores. In addition, it seems the pre-trained ViTs achieved good avg recalls, however their precisions are very low (30-35%) compared to other models. As mentioned in the original paper [20], the quality of the model is affected not only by architecture choices, but also by parameters such as the learning rate schedule, optimizer, weight decay, etc. In practice, it's recommended to fine-tune a ViT model that was pre-trained using a large, high-resolution dataset. Overall, the accuracy (precision, recall and F1-scores) in the range of 60-70% are satisfactory considering all other facts related to the problem domain, types of images and current small dataset size.

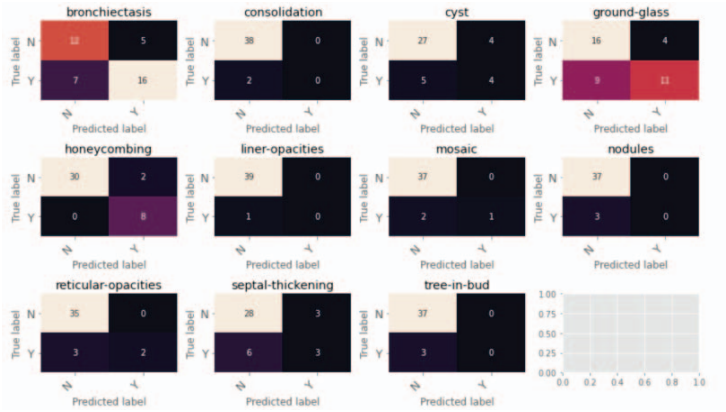

Figure 7: Multilabel Confusion Matrix (test set) for the Xception model

The class-wise multilabel confusion matrix is also generated (Fig. 7) using ***sklearn*** library to evaluate the accuracy of the classification, and output confusion matrices for each concept class. The output of the confusion matrices in Fig. 8 also confirmed the reason of low accuracies for certain class labels (Fig. 7), such as linear opacities, mosaic etc.

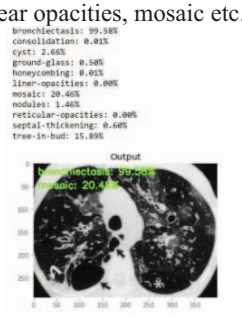

Figure 8: Output result of the test image, 31927.jpg

Fig. 8 shows the classification probabilities of different class labels and top two labels with associated probabilities are overlaid in a sample test image in the original ImageCLEFmed dataset with associated caption "CT scan at the level of the upper lobes in a 26-year-old woman demonstrates mild to moderate signs of **bronchiectasis** and peri bronchial wall thickening. **Mosaic** perfusion, bullae (straight arrows), emphysema (*), and an area of **consolidation** (curved arrow) are also seen" [14]. From the output we can figure out that it correctly predicted "bronchiectasis" and "mosaic" patterns and confused probably "consolidation" with "tree-in-bud" pattern.

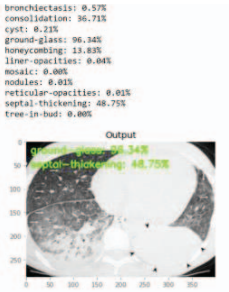

Figure 9: Output result of the test image, 62167.jpg

Fig. 9 shows the classification probabilities of another test image (62167.jpg) in the ImageCLEFmed dataset [14] with associated caption as "Acute systemic lupus erythematosus pneumonitis. CT scan reveals extensive ***ground-glass*** attenuation throughout both lungs (arrows), interlobular ***septal thickening,*** bilateral lower lobe ***consolidations*** (complete on the left side [arrowheads]), and minimal pleural effusion" [14]. The output shows this time it correctly predicted the "***ground-glass", "septal-thickening"*** and ***"consolidation"*** class labels with higher probabilities.

V. CONCLUSIONS

This work presents a proof-of-concept study to demonstrate the effectiveness of images appeared in biomedical articles as a valuable resource for ML and Information Retrieval tasks, such as concept-based classification and image search. It shows the potential to improve the retrieval of biomedical literature by targeting the visual content in articles, a rich source of information not typically exploited by conventional bibliographic or full-text databases. It is expected that this work can be extended further to generate more data (training ground truth) which would offer building blocks for the development of advanced information retrieval systems aided by a visual ontology. The main limitation of this study is that the models (networks) are unable to predict on data they were never trained on using Keras networks for multi-label classification. In future, we plan to work on DL based object (ROI) detection based on the annotations gathered using LabelImg tool [23]. Overall, the impact of this work is substantial due to many applications such as digital libraries and image search engines for teaching and training purposes require effective and efficient techniques to categorize and access images.

ACKNOWLEDGMENT

The work is supported by an NSF grant (#2131207), entitled, "CISE-MSI: DP: IIS:III: Deep Learning Based Automated Concept and Caption Generation of Medical Images Towards Developing an Effective Decision Support System (DSS)".

REFERENCES

[1] D. Demner-Fushman, S. K. Antani, M. S. Simpson and G. R. Thoma, "Annotation and retrieval of clinically relevant images," International Journal of Medical Informatics, vol. 78(12), e59–e67, 2009

[2] Z. Lu. "Pubmed and beyond: a survey of web tools for searching biomedical literature", , vol. baq036, 2011

[3] M. S. Simpson, D. You, M. M. Rahman, Z. Xue, D. Demner-Fushman, S. K. Antani and G. R. Thoma G., "Literature-based biomedical image classification and retrieval, " Comput Med Imaging Graph. Vol. 39, pp. 3-13, 2015

[4] M. M. Rahman, D. You, M. S.Simpson, S. K. Antani, D. Demner-Fushman and G. R. Thoma, "Interactive cross and multimodal biomedical image retrieval based on automatic region-of-interest (ROI) identification and classification," International Journal of Multimedia Information Retrieval. 3. 131-146. 10.1007/s13735-014-0057-9., 2014

[5] E. K. Charles and Cheng, "GoldMiner: A Radiology Image Search Engine, The Practice of Radiology", vol. 188 (6), pp. 1475-1478, 2007.

[6] M. A. Hearst, A. Divoli, et al., "Biotext search engine: beyond abstract search. Bioinformatics, " vol. 23(16), pp. 2196–2197, 2007

[7] S. Xu, J. McCusker and M. Krauthammer, "Yale Image Finder (YIF): a new search engine for retrieving biomedical images," Bioinformatics. vol. 24 (17):1968-70, 2008

[8] H. Muller¨, N. Michoux, D. Bandon, A. Geissbuhler, "A review of contentbased image retrieval systems in medical applications clinical benefits and future directions," Int J Med Inform, vol. 73, pp. 1–23, 2014

[9] You D, Antani SK, Demner-Fushman D, Rahman MM, Govindarju V, Thoma GR. (2010) Biomedical Article Retrieval Using Multimodal Features and Image Annotations In Region-based CBIR, Document Recognition and Retrieval XVII. Edited by Likforman-Sulem, Laurence; Agam, Gady. Proceedings of the SPIE. San Jose, CA. 7534:75340V-75340V-12

[10] E. B. Meltzer and P.W. Noble, "Idiopathic pulmonary fibrosis," Orphanet J Rare Dis, vol. 3(8), https://doi.org/10.1186/1750-1172-3-8, 2008

[11] D. Lindberg, B. Humphreys and A. McCray, "The unified medical language system," Methods Inf Med, vol. 32(4), pp. 281–291, 1993

[12] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," Radiographics, vol. 26(6), pp. 1595–7, 2006

[13] R. Datta, D. Joshi, J. Li and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," ACM Comput Surv., 40(2), pp. 1–60, 2008.

[14] H. Müller, A. Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. K. Antani and E. Ivan, "Overview of the ImageCLEF2012 Medical Image Retrieval and Classification Tasks," In: The Working Notes for the CLEF 2012 Labs and Workshop, Rome, Italy, 17–20 Sept 2012

[15] J. Bogatinovski, L. Todorovski, S. Džeroski and D. Kocev, "Comprehensive comparative study of multi-label classification methods," Expert Systems with Applications, vol. 203,2022, 117215, ISSN 0957-4174

[16] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," Med Image Anal 42:60–88, doi: 10.1016/ j.media.2017.07.005, 2017

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. of the IEEE conference on computer vision and pattern recognition, pp. 1251—1258, 2017

[18] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Densely Connected Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[19] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[20] D. Alexey, B. Lucas, K. Alexander, et al., " An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929

[21] K. Han, Y. Wang, H. Chen, et al., "A Survey on Visual Transformer," *ArXiv, abs/2012.12556*, 2020

[22] S. Fahad, K. Salman, W. Z. Syed, et al., "Transformers in Medical Imaging: A Survey," arXiv:2201.09873, 2022

[23] LabelImg for Labeling Object Detection Data: https://blog.roboflow.com/labelimg/