Universidade de Aveiro Departamento de
2019 Eletrónica, Telecomunicações e Informática

Eduardo Miguel Coutinho Gomes de Pinho

Recuperação de Informação Multimodal em Repositórios de Imagem Médica

Multimodal Information Retrieval in Medical Imaging Repositories

Eduardo Miguel Coutinho Gomes de Pinho

Recuperação de Informação Multimodal em Repositórios de Imagem Médica

Multimodal Information Retrieval in Medical Imaging Archives

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Informática, realizada sob a orientação científica de Carlos Manuel Azevedo Costa, Professor do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

Trabalho financiado pela Fundação para a Ciência e Tecnologia (FCT) sob a identificação PD/BD/105806/2014.

o júri / the jury

presidente / president	**Joaquim Manuel Vieira** Professor Catedrático da Universidade de Aveiro
vogais / examiners committee	
	Ana Luísa Nobre Fred Professora Associada da Universidade de Lisboa - Instituto Superior Técnico
	Paulo Martins de Carvalho Professor Associado da Universidade do Minho
	Paulo José Osório Rupino da Cunha Professor Auxiliar com Agregação da Faculdade de Ciências e Tecnologia da Universidade de Coimbra
	Augusto Marques Ferreira da Silva Professor Auxiliar da Universidade de Aveiro
	Carlos Manuel Azevedo Costa Professor Auxiliar da Universidade de Aveiro (orientador)

acknowledgements

A word of thanks to my advisor Carlos Costa, and to Professor José Luís Oliveira, for persistently believing in my abilities and guiding my academic endeavors since day one at IEETA.

A special note of thanks to Luís Bastião Silva for guiding me throughout my years as undergraduate and PhD freshman student, and for giving me the (positive!) realization that talking about Dicoogle encompasses way more than technical matters.

Thank you to everyone else who has been part of the UA.PT Bioinformatics lab at some moment in time (the names of which at this point are too many to enumerate without making the mistake of sorely leaving someone out), for your knowledge, friendship, and laughter.

I thank my family for supporting my decisions in life and for always backing me up in the toughest of circumstances.

Finally, I thank my beloved Andreia for building life-long foundations with me, despite the challenges that came along with it, and the difficult obstacles of her own.

This work has been financed by the *Fundação para a Ciência e Tecnologia* (FCT) under the reference PD/BD/105806/2014.

Palavras-chave

recuperação de informação, imagem médica, recuperação de imagem baseada em conteúdo, *deep learning*, *representation learning*.

Resumo

A proliferação de modalidades de imagem médica digital, em hospitais, clínicas e outros centros de diagnóstico, levou à criação de enormes repositórios de dados, frequentemente não explorados na sua totalidade. Além disso, os últimos anos revelam, claramente, uma tendência para o crescimento da produção de dados. Portanto, torna-se importante estudar novas maneiras de indexar, processar e recuperar imagens médicas, por parte da comunidade alargada de radiologistas, cientistas e engenheiros. A recuperação de imagens baseada em conteúdo, que envolve uma grande variedade de métodos, permite a exploração da informação visual num arquivo de imagem médica, o que traz benefícios para os médicos e investigadores. Contudo, a integração destas soluções nos fluxos de trabalho é ainda rara e a eficácia dos mais recentes sistemas de recuperação de imagem médica pode ser melhorada.

A presente tese propõe soluções e métodos para recuperação de informação multimodal, no contexto de repositórios de imagem médica. As contribuições principais são as seguintes: um motor de pesquisa para estudos de imagem médica com suporte a pesquisas multimodais num arquivo extensível; uma estrutura para a anotação automática de imagens; e uma avaliação e proposta de técnicas de *representation learning* para deteção automática de conceitos em imagens médicas, exibindo maior potencial do que as técnicas de extração de *features* visuais outrora pertinentes em tarefas semelhantes. Estas contribuições procuram reduzir as dificuldades técnicas e científicas para o desenvolvimento e adoção de sistemas modernos de recuperação de imagem médica multimodal, de modo a que estes façam finalmente parte das ferramentas típicas dos profissionais, professores e investigadores da área da saúde.

Ŝlosilvortoj

informretrovo, medicina bildigo, bildoretrovo bazita sur enhavo, profunda lernado, reprezenta lernado.

Resumo

La proliferado de medicina bitbild-akiriloj en malsanulejoj kaj aliaj diagnozejoj verkis grandegajn deponejojn de grandvaloraj datumoj, kiuj ofte ne plene esploriĝas. Krome, la pasintaj jaroj montras kreskon de datumproduktado. Do, la studo de novaj manieroj por indeksi, prilabori, kaj retrovi medicinajn bildojn estas graviĝan subjekton, kiun estas traktonta per la pli larga komunumo de radiologoj, scientistoj, kaj inĝenieroj. Bildoretrovo bazita sur enhavo, kiu agregatas diversajn metodojn, povas esplori la vidan informacion de medicina bildarkivo, kaj povas esti utila por praktikistoj kaj esploristoj. Tamen, la efikeco de la lastaj sistemoj por medicina bildretrovo povas pliboniĝi, kaj siaj integrigo en klinika laborfluo ankoraŭ estas malofta. Ĉi tiu tezo proponas solvojn kaj metodojn por multimodala informretrovo, en la kunteksto de medicinbildiga deponejoj. La ĉefaj kontribuoj estas serĉilo por medicina bildigostudoj, kiu subtenas multimodalajn informpetojn en etendebla arkivo; framo por aŭtomate prinoti medicinajn bildojn por enhavo-malkovro; kaj taksado kaj propono de teknikoj de reprezenta lernado por koncepto-rekono el medicinaj bildoj, kiuj montras pli bonan potencialon ol trajto-eltiraj algoritmoj, kiuj estis antaŭe ofte uzata en similaj taskoj. Ĉiu el ĉi tiuj kontribuoj celas mallarĝi la sciencan kaj teknikan breĉon al la konstruado kaj alpreno de novaj sistemoj por multimodala medicina bildoretrovo, por finfine fariĝi parto de la laborfluoj de medicinaj praktikistoj, instruistoj, kaj esploristoj en sanzorgo.

Keywords information retrieval, medical imaging, content based image retrieval, deep learning, representation learning.

Abstract

The proliferation of digital medical imaging modalities in hospitals and other diagnostic facilities has created huge repositories of valuable data, often not fully explored. Moreover, the past few years show a growing trend of data production. As such, studying new ways to index, process and retrieve medical images becomes an important subject to be addressed by the wider community of radiologists, scientists and engineers. Content-based image retrieval, which encompasses various methods, can exploit the visual information of a medical imaging archive, and is known to be beneficial to practitioners and researchers. However, the integration of the latest systems for medical image retrieval into clinical workflows is still rare, and their effectiveness still show room for improvement.

This thesis proposes solutions and methods for multimodal information retrieval, in the context of medical imaging repositories. The major contributions are a search engine for medical imaging studies supporting multimodal queries in an extensible archive; a framework for automated labeling of medical images for content discovery; and an assessment and proposal of feature learning techniques for concept detection from medical images, exhibiting greater potential than feature extraction algorithms that were pertinently used in similar tasks. These contributions, each in their own dimension, seek to narrow the scientific and technical gap towards the development and adoption of novel multimodal medical image retrieval systems, to ultimately become part of the workflows of medical practitioners, teachers, and researchers in healthcare.

List of contents

List of contents i

List of figures v

List of tables vii

List of acronyms ix

1 Introduction 1

1.1 Motivation 2
1.2 Objectives 2
1.3 Thesis outline 3
1.4 Scientific Publications and Communications 3

2 Computer Systems in Medical Imaging 6

2.1 Medical Imaging Modalities 7
2.2 Picture Archiving and Communication Systems 7
  2.2.1 Definition and typical components 7
  2.2.2 Interoperability 9
  2.2.3 Overall impact 10
2.3 Digital Imaging and Communications in Medicine (DICOM) 11
  2.3.1 DICOM Information Model 11
  2.3.2 Structured Report 13
  2.3.3 Services 14
2.4 Dicoogle 15
  2.4.1 Extensible User Interface 17
  2.4.2 Learning Pack 18
  2.4.3 Dicoogle as a framework in Research 19
2.5 Final Remarks 21

3 Content Discovery and Information Retrieval 22
3.1 Content Discovery . . . 22
3.1.1 Structured and Unstructured Data . . . 22
3.1.2 Data Mining . . . 23
3.1.3 Deep Learning . . . 23
3.1.4 Machine learning for image understanding . . . 25
3.2 Information Retrieval . . . 25
3.2.1 Definition . . . 26
3.2.2 Digital Information Retrieval Systems . . . 27
3.2.3 Semantics and Ontologies . . . 28
3.2.4 Medical Imaging Data and Searching . . . 29
3.3 Content-Based Image Retrieval . . . 30
3.3.1 Definition . . . 30
3.3.2 General Architecture . . . 31
3.3.3 Image Feature Extraction . . . 32
3.3.4 Feature Similarity Measures . . . 34
3.3.5 CBIR in Medical Imaging . . . 34
3.4 Final Remarks . . . 36

4 Multimodal Medical Information Retrieval 38
4.1 Definition . . . 38
4.2 Trending Techniques and Concepts . . . 39
4.2.1 Multi-Query Combination . . . 40
4.2.2 Query Refinement . . . 41
4.2.3 Other Multimedia Content . . . 42
4.3 Multimodal Search Engine Proposal . . . 43
4.3.1 Related Work . . . 43
4.3.2 Architecture . . . 44
4.3.3 Query Formulation and Processing Workflow . . . 45
4.3.4 Multi-Query and Fusion Techniques . . . 46
4.3.5 API and Data Representation . . . 47
4.3.6 Graphical User Interface . . . 48
4.4 Computational Assessment and Results . . . 49
4.5 Final Remarks . . . 52

5 Automated Labeling for Content Discovery 54
5.1 Related Work . . . 55
5.2 Architecture Proposal . . . 57
5.2.1 Dicoogle Integration . . . 57

5.2.2 Multimodal Engine with Content Discovery 58
5.2.3 Classification Database for Object Indexing 59
5.2.4 Organ Recognition Classifiers 60
5.3 Results and Discussion 63
5.3.1 Multimodal Search Engine with Anatomic Database 63
5.3.2 Anatomic Classifiers 64
5.4 Final Remarks 66

6 Representation Learning **67**
6.1 Related Work 69
6.1.1 Auto-encoders 69
6.1.2 Generative Adversarial Networks 71
6.1.3 Representation Learning in the Medical Imaging domain 72
6.2 Comparative Analysis of Unsupervised Learning Methods for Concept Detection 73
6.2.1 Bags of Visual Words 73
6.2.2 Deep unsupervised learning 74
6.2.3 Evaluation Methodology 79
6.2.4 Results and Discussion 82
6.3 Feature Learning for Concept Detection in Medical Images 85
6.3.1 Feature Learning 86
6.3.2 Qualitative Feature Analysis 90
6.3.3 Multi-label Classification 93
6.4 Results and Discussion 94
6.5 Final Remarks 97

7 Conclusion **100**
7.1 Achievements and Limitations 100
7.2 Future Directions 102

References **104**

List of figures

1.1 Diagram containing the chapters of this thesis, where an arrow represents a relation *«contributes to»*. Each component is ultimately applied to medical imaging informatics by enhancing a medical imaging system's search capabilities. 4

2.1 An example of a Picture Archiving and Communication Systems (PACS), representing an archive center, a Radiology Information System, workstations, and modality devices. 8
2.2 A representation of a DICOM object, as a sequence of data elements. 12
2.3 DICOM information hierarchy: Patient-Study-Series-Image. 13
2.4 A DICOM file containing meta-data and an image. A fraction of the DICOM header is detailed to the right. 14
2.5 General architecture of the Dicoogle framework. 16
2.6 Dicoogle Web Core sequence diagram 18
2.7 Use case diagram of a reversibly de-identified medical imaging archive (adapted from Silva *et al.* [4]). 20
2.8 Overview of the architecture of the viewer (adapted from Godinho *et al.* [51]). 20

3.1 A diagram representing a typical Content-based Image Retrieval (CBIR) system. 31
3.2 Diagram depicting the components of the CBIR and their interactions with the Dicoogle platform. 36

4.1 Architecture diagram of the multimodal search engine extension, depicting its key interactions with Dicoogle. 44
4.2 Diagram depicting the multimodal search pipeline 45
4.3 A draft of the multimodal search engine's graphical user interface, depicting a multi-level query, ghost query boxes and a few results from a previous search. . 49
4.4 Scatter plot representing the engine's response time, in seconds, in terms of the number of results obtained from the query. 50
4.5 Chart exhibiting the engine's response time (seconds) over fusion queries with different late fusion algorithms. 51

4.6 Average cumulative response time measured in the third experiment. 52

5.1 The extended architecture of a PACS archive, where Dicoogle is used as the base platform. The multimodal search engine is included. The contributions of this chapter are also highlighted in yellow. 57
5.2 Diagram depicting the main classification database components, their interactions with existing classifiers in the Dicoogle runtime, and an example of selective classification before a given instance. 59
5.3 An example of projecting segmented slices into a list of labels, where the colored lines represent their estimated position. 61
5.4 Diagram depicting the architecture of the Convolutional Neural Network (CNN) proposed for anatomy recognition, to which images (x) are fed to produce the labels (y). 62
5.5 The web-based user interface for the multimodal search engine, exhibiting the first few search results for “liver”. 63

6.1 Diagram depicting the basic auto-encoder structure. 69
6.2 Basic structure of a Generative Adversarial Network (GAN). 71
6.3 Diagram of the sparse denoising auto-encoder. 76
6.4 Diagram of the variational auto-encoder. 77
6.5 Diagram of the bidirectional GAN. 77
6.6 Diagram of the adversarial auto-encoder. 78
6.7 A few samples from the ImageCLEF 2017 concept detection data set, with their respective file ID and trimmed list of concept identifiers. 80
6.8 The 2D projections of the latent codes in the validation set, for each learned feature space. 83
6.9 Architecture of the flipped adversarial auto-encoder with two image levels. . . . 89
6.10 The 2D projections of the features with PCA, for each learned feature space. . 90
6.11 The 2D projections of the features with UMAP, for each learned feature space. 91
6.12 An example of samples produced by the AAE and the Flipped Adversarial Auto-encoder (F-AAE) during the training process. 92
6.13 A sequence of generated samples from the F-AAE by moving the prior code around a circle in the hypersphere, column-first (from left to right, then top to bottom). 92

List of tables

5.1 The specification of the experimented CNN, used as a proof of concept for the 5 organ classifiers. 62
5.2 The results from classifier evaluation using data sets 1 and 2. 65

6.1 A tabular representation of the SimpleNet layers' specifications. 75
6.2 A tabular representation of decoder/generator layers' specifications. 75
6.3 The ten most frequently occurring concepts in the ImageCLEF 2017 training set for concept detection. 80
6.4 The best metrics obtained from logistic regression for each representation learned, where *Mix* is the feature combination of SDAE and AAE. The highest scores are shown in bold. 84
6.5 The best metrics obtained from vector similarity search for each representation learned. The highest scores are shown in bold. 85
6.6 Most frequent concepts in the training set of ImageCLEF 2018 caption. 86
6.7 A tabular representation of the sequential composition of the encoders and discriminators in the networks. 87
6.8 A tabular representation of the sequential composition of the decoders and generators in the networks. 88
6.9 List of submitted runs from the UA.PT Bioinformatics research group to the concept detection challenge. 94
6.10 Results obtained from the submissions to the ImageCLEF 2018 concept detection task with logistic regression. 95
6.11 Results obtained from the submissions to the ImageCLEF 2018 concept detection task with similarity search. 96

List of acronyms

AAE	Adversarial Auto-encoder
ACR	American College of Radiology
AE	Application Entity
ALARA	As Low As Reasonably Achievable
ANSI	American National Standards Institute
API	Application Programming Interface
BiGAN	Bidirectional Generative Adversarial Network
BoC	Bag of Colors
BoVW	Bag of Visual Words
BoW	Bag of Words
CAD	Computer Aided Detection and Diagnosis
CADe	Computer Aided Detection
CADx	Computer Aided Diagnosis
CBIR	Content-based Image Retrieval
CBMIR	Content-based Medical Image Retrieval
CBVIR	Content-based Visual Information Retrieval
CEDD	Color and Edge Directivity Descriptor
CNN	Convolutional Neural Network
CO	Content Object
CT	Computed Tomography
CUI	Concept Unique Identifier
DBMS	Database Management System
DICOM	Digital Imaging and Communications in Medicine
DIM	DICOM Information Model
DIM-SE	DIM service entity
EHR	Electronic Health Record

F-AAE	Flipped Adversarial Auto-encoder
FAST	Features from Accelerated Segment Test
FTRL	Follow-the-regularized-leader
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HIS	Hospital Information System
HL7	Health Level 7
HTML	Hyper-Text Markup Language
ICA	Independent Component Analysis
IE	Information Entity
IHE	Integrating Healthcare Enterprise
IOD	Information Object Definition
IRMA	Information Retrieval in Medical Applications
ISR	Inverted Squared Rank
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
LReLU	Leaky Rectified Linear Unit
MAP	Mean Average Precision
MeSH	Medical Subject Headings
MIME	Multipurpose Internet Mail Extensions
MLP	Multi-layer Perceptron
MPR	Multiplanar Reconstruction
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MRML	Multimedia Retrieval Markup Language
NCI	National Cancer Institute
NEMA	National Electrical Manufacturers Association
NLM	National Library of Medicine
ORB	Oriented FAST and Rotated BRIEF
OWL	Web Ontology Language
P2P	Peer-to-Peer
PACS	Picture Archiving and Communication Systems

PCA	Principal Component Analysis
PDDCA	Public Domain Database for Computational Anatomy
PET	Positron Emission Tomography
PMC	PubMed Central
QBE	Query by Example
RBM	Restricted Boltzmann Machine
RDF	Resource Description Framework
ReLU	Rectified Linear Unit
REST	Representational State Transfer
RIS	Radiology Information System
ROC	Receiver Operating Characteristics
ROI	Region of Interest
RRF	Reciprocal Rank Fusion
SCP	Service Class Provider
SCU	Service Client User
SDAE	Sparse Denoising Auto-encoder
SDK	Software Development Kit
SIFT	Scale Invariant Feature Transform
SOP	Service-Object Pair
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SR	Structured Report
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
UI	User Interface
UID	Unique Identifier
UMAP	Uniform Manifold Approximation and Projection
UMLS	Unified Medical Language System
URI	Universal Resource Indicator
URL	Universal Resource Location
VAE	Variational Auto-encoder
VR	Value Representation
W3C	World Wide Web Consortium
WSI	Whole-slide Imaging
WWW	World Wide Web

XDS	Cross-Enterprise Document Sharing
XDS-i	Cross-Enterprise Document Sharing for Imaging
XML	Extended Markup Language

Chapter 1

Introduction

There came a moment in our history when medical imaging no longer implied film and physical paper-based archives. The digital revolution, like in many other fields, was a remarkable milestone in healthcare systems, leading to improved radiology workflows and new opportunities for data analysis. Digital medical imaging systems in healthcare institutions have become increasingly important over the past few decades, as they are playing a valuable role in medical diagnosis, decision support, and treatment procedures. Research and industry efforts to develop medical imaging equipment, including new acquisition modalities and information systems, are intense and have been grounded by the wide acceptance of medical imaging devices into the digital era.

The number of medical imaging studies is constantly growing, resulting in tremendous large amounts of data produced, which need to be archived and searched, even over distributed networks, making retrieval of these studies an increasingly harder task. Besides supporting local medical image storage and retrieval, these systems provide healthcare practitioners a significant range of new use cases, such as the ability to remotely access multimedia patient information and set up collaborative work environments. The concept of Picture Archiving and Communication Systems (PACS), an umbrella term for digital information systems in medical imaging, is technology-agnostic. On the other hand, Digital Imaging and Communications in Medicine (DICOM) emerged as the common technological standard for computer systems in this context, providing digital information formats for medical images as well as communication protocols for the various components in the system. Multiple PACS solutions with distinct architectures and services are currently in use, spanning from simple models, typically used in small laboratories, to enterprise-wide platforms, mostly used in large hospital networks. The use of digital imaging provides numerous advantages, such as the support for an automatic or semi-automatic interpretation of the available medical data.

1.1 Motivation

Medical imaging repositories are often looked on as “inert bags” of DICOM objects, accessible only through the DICOM “query and retrieve” service. The means by which we currently search for information has been shaped by search engine interfaces, and free searching is currently a common feature expected from any information system. Moreover, typing on a search bar with keywords or phrases of interest, although very common nowadays, is not the only way of obtaining useful information in these systems. Further advancements in the field have granted the ability to search with other kinds of content, such as images, through the process called Content-based Image Retrieval (CBIR). With the automated introspection of medical images provided by a CBIR system, it is feasible to retrieve images by similarity, which can translate to similar clinical cases in the repository. Even more so, the latest trends in machine learning for image recognition exhibit promising methods for interpreting visual data into a feature space that captures information in a descriptive manner, which can also be exploited for content discovery in PACS repositories.

On the other hand, the heterogeneity of information often available in a PACS (medical images, DICOM meta-data and clinical reports to name a few) ought to be combined in order to attain a deeper understanding of what makes studies relevant in a search, in a way that could not be achieved by methods relying on text or visual data exclusively. The application of multimodality in Content-based Medical Image Retrieval (CBMIR) is a challenging issue to be addressed by communities of radiologists, biomedical engineers and computer scientists: new ways to index, process and retrieve medical information must be studied, with a vision that the future of medical practitioners, teachers, and researchers in medical imaging with hold these systems as part of their common workflow.

1.2 Objectives

This thesis aims to encompass research and proposals that contribute to multimodal information retrieval in the context of medical imaging repositories, by exploring new ways to index, process, and retrieve distinct information in these archives.

The following three objectives of this work can be outlined:

1. Perform a study on the latest techniques relevant to the scope of medical image retrieval, including text and image feature extraction, similarity measures, query fusion, and automated medical image analysis.
2. Create and evaluate new techniques for multimodal medical information retrieval, which may be applied to images and respective meta-data, structured reports and other annotated text content.
3. Develop software for the evaluation of Content-based Medical Image Retrieval (CBMIR)

solutions and their subsequent integration to a PACS environment, for its exploitation in clinical and research environments.

1.3 Thesis outline

The outline of the thesis manuscript is as follows:

- **Chapter 2** introduces the reader to the medical imaging field in the modern times, with a special focus on the concept of PACS and the DICOM standard.
- **Chapter 3** starts by making a brief overview of topics in content discovery, including machine learning and deep learning, followed by information retrieval in the digital era, and ends with the definition, concepts and challenges for Content-based Image Retrieval.
- **Chapter 4** describes multimodal information retrieval applied to medical imaging informatics, covering known techniques, existing projects and the development process of a multimodal search engine solution for a PACS archive.
- **Chapter 5** introduces the concepts of computer aided detection and diagnosis, provides a deeper insight into medical image understanding by computer systems, and proposes an architecture for automatic content discovery through classification of medical images in an archive.
- **Chapter 6** presents the concept of representation learning and its application in medical imaging systems, and proposes the use of unsupervised learning methods for automatic concept detection from medical images.
- The document ends with an emphasis on the work achieved and future directions in **Chapter 7**.

With medical imaging informatics as the center point, the thesis envelops this field with the remaining chapters, which cover individual fractions of the state-of-the-art (information retrieval, content discovery, content-based image retrieval, and representation learning), including the scientific output of this PhD, in a way that contributes to systems in digital medical imaging (Figure 1.1).

1.4 Scientific Publications and Communications

Without disregarding the various contributions in the form of open source resources and participation in international campaigns to the communities in medical imaging informatics and machine learning, scientific publication stands nevertheless as a *de facto* measurement of a doctorate's influence. The following journal articles and conference proceedings represent the ones in which a significant (or major) contribution was made, and which is already published at the time of writing.

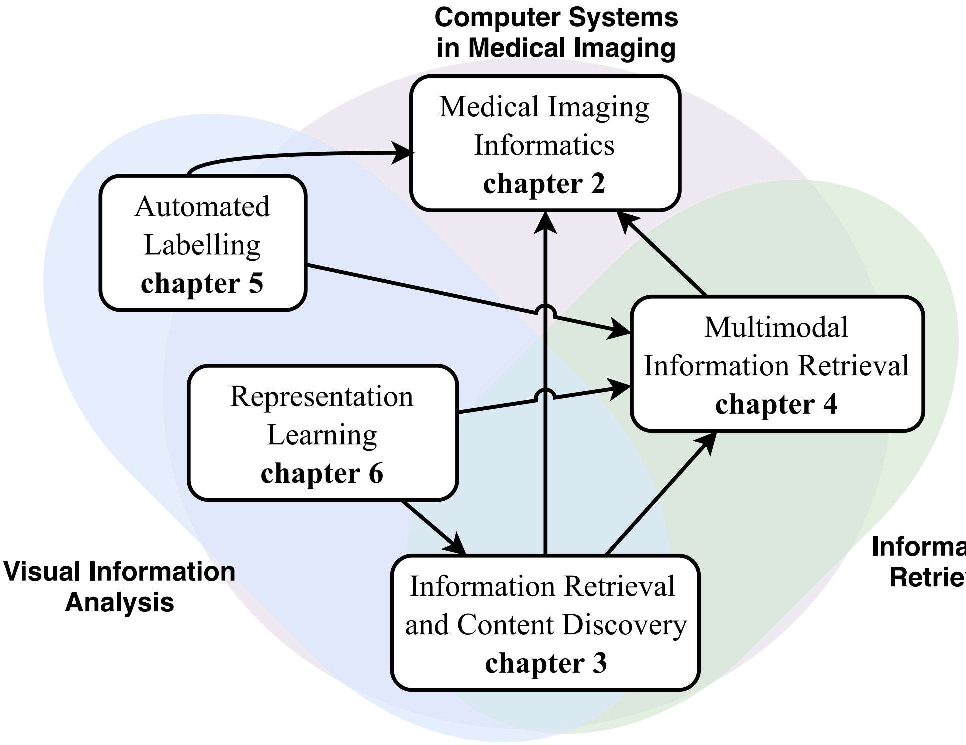

Computer Systems in Medical Imaging
Medical Imaging Informatics chapter 2
Automated Labelling chapter 5
Multimodal Information Retrieval chapter 4
Representation Learning chapter 6
Information Retrieval and Content Discovery chapter 3
Visual Information Analysis
Information Retrieval

Figure 1.1: Diagram containing the chapters of this thesis, where an arrow represents a relation *«contributes to»*. Each component is ultimately applied to medical imaging informatics by enhancing a medical imaging system's search capabilities.

Journal Articles

- Eduardo Pinho, Tiago Marques Godinho, Frederico Valente, Carlos Costa. *"A Multimodal Search Engine for Medical Imaging Studies"*. Springer Journal of Digital Imaging. 2017. [1]
- Eduardo Pinho, Carlos Costa. *"Automated Anatomic Labeling for Content Discovery in Medical Imaging Repositories"*. Springer Journal of Medical Systems. 2018. [2]
- Eduardo Pinho, Carlos Costa. *"Unsupervised Representation Learning for Concept Detection in Medical Images: a Comparative Analysis"*. MDPI Applied Sciences. 2018. [3]
- Jorge Miguel Silva, Eduardo Pinho, Eriksson Monteiro, João Figueira Silva, Carlos Costa, *"Controlled Searching in Reversibly De-identified Medical Imaging Archives"* Elsevier Journal of Biomedical Informatics. 2018. [4]

Conference Proceedings

- Eduardo Pinho, Frederico Valente, Carlos Costa. *"A PACS-oriented multimodal search engine"*. International Conference on Computer Assisted Radiology and Surgery (CARS). 2016. [5]
- Eduardo Pinho, Carlos Costa. *"Extensible Architecture for Multimodal Search Engine in Medical Imaging Archives"*. 12th International Conference on Signal Image Technology & Internet-based Systems (SITIS). 2017. [6]
- Eduardo Pinho, João Figueira Silva, Jorge Miguel Silva, Carlos Costa. *"Towards Representation Learning for Biomedical Concept Detection in Medical Images: UA.PT Bioinformatics in ImageCLEF 2017"*. CEUR Working Notes of CLEF. 2017. [7]
- João Figueira Silva, Jorge Miguel Silva, Eduardo Pinho, Carlos Costa, *"3D-CNN in Drug Resistance Detection and Tuberculosis Classification"*. CEUR Working Notes of CLEF. 2017. [8]
- Eduardo Pinho, Carlos Costa. *"Comparative analysis of unsupervised representation learning methods for concept detection in medical images"*. International Conference on Computer Assisted Radiology and Surgery (CARS). 2018. [9]
- Eduardo Pinho, Carlos Costa. *"Feature Learning with Adversarial Networks for Concept Detection in Medical Images: UA.PT Bioinformatics at ImageCLEF 2018"*. CEUR Working Notes of CLEF. 2018. [10]
- Jorge Miguel Silva, António Guerra, João Figueira Silva, Eduardo Pinho, Carlos Costa. *"Face De-Identification Service for Neuro Imaging Volumes"*. 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018. [11]

Chapter 2

Computer Systems in Medical Imaging

In hospitals and clinics, the requirements of systems for medical information retrieval are ever-increasing. The large amounts of medical studies performed daily during the past decade have reached the point where some use cases are becoming *big data* problems. The Radiology department of the university of Geneva, for instance, has reached 1 TB of total medical image data in 2001 [12]. It is estimated that over 5 billion medical imaging procedures were performed worldwide by 2010 [13], and an estimation in 2012 suggests that the USA has produced over 1 Exabyte of medical imaging data in 2016 [14]. On the other hand, additional studies show that this growth in medical imaging utilization is not transversal to all modalities, demographics, requirements and types of interventions. A population-based cohort study in Taiwan reveals an increase in medical imaging utilization over the years from 1997 to 2008 [15]. In another study, the use of inpatient radiology at an institution shows a decreasing trend after 2009, especially for Computed Tomography (CT). One may also identify a slower growth of medical imaging backed by multiple causes, some of which relate to imposed hospital policies [16]. As medical staff are increasingly aware of associated dangers, they seek to reduce radiation-inducing medical imaging use by encouraging clinical practices for efficiency and harm reduction [17]. Nonetheless, there is a greater tendency towards an increasing rate of acquisition as supply for imaging equipment increases, the level of detail on captured images evolve, and promising modalities are established into clinical practice.

These systems should satisfy the needs and improve the performance of medical practitioners, but also be easily integrated into their daily workflow without introducing excessively complex procedures. On the one hand, new solutions should have a visible impact in the field, often in the form of greater accuracy or novel means of medical information retrieval and registering, such as query-by-image searching or tools for writing structured medical reports. On the other hand, the usability of these systems can cross the borderline of whether or not they should be adopted. If the software is found to be highly complex to use, medical practitioners will be reluctant and less inclined to integrate it to their workflow. Therefore, it is important and timely to develop and improve medical imaging computer systems.

This chapter places the reader into the context of computer systems in medical imaging, while describing some of the latest use cases and research challenges of medical imaging informatics. Afterwards, an extended presentation of an open source medical imaging archive software is presented, serving as a technical background for subsequent work in this thesis.

2.1 Medical Imaging Modalities

Medical imaging is the field dedicated to the many challenges of acquiring, analyzing and handling images of organs and tissues. It is coupled with radiology, the specialty of medical image acquisition and interpretation, which emerged in 1895 with the discovery of the X-ray radiography. It was discovered then that electromagnetic radiations could be used to see the internal parts of an opaque object with components of different densities, thus allowing for an observation of the inner workings of a body using non invasive techniques. Subsequent scientific research lead to medical imaging being currently used in other departments, such as cardiology, pathology and ophthalmology, as well as the emergence of new acquisition techniques, such as whole-slide imaging. Both key procedures and devices for acquisition are called *modalities*.

The CT is a very common modality, relying on X-rays to produce several slices of the body along one or multiple axes. The Magnetic Resonance Imaging (MRI) scan combines radio waves with a strong magnetic field in order to acquire these images. Many more modalities are currently in existence, and their usage are strongly dependent on the intended diagnosis. In addition, a combination of multiple different modalities may also be used to obtain more information that could be left unnoticed with a single modality alone [18].

2.2 Picture Archiving and Communication Systems

Before the uprising of digital medical imaging, a radiologist or clinician would need to retrieve and analyze pictures in film from a physical archive kept in the medical facility. Currently, film folders were replaced with digital repositories, usually backed by a Database Management System (DBMS), and images are revised from *workstations* or other kinds of visualization devices such as smartphones. This change of paradigm has greatly contributed to treatment and diagnosis, by providing an improved workflow and better tools to practitioners.

2.2.1 Definition and typical components

PACS is an umbrella-term that defines the composition of software and hardware comprising medical imaging and data acquisition equipment, subsequent storage devices and display subsystems, all of which are integrated by digital networks and end user software.

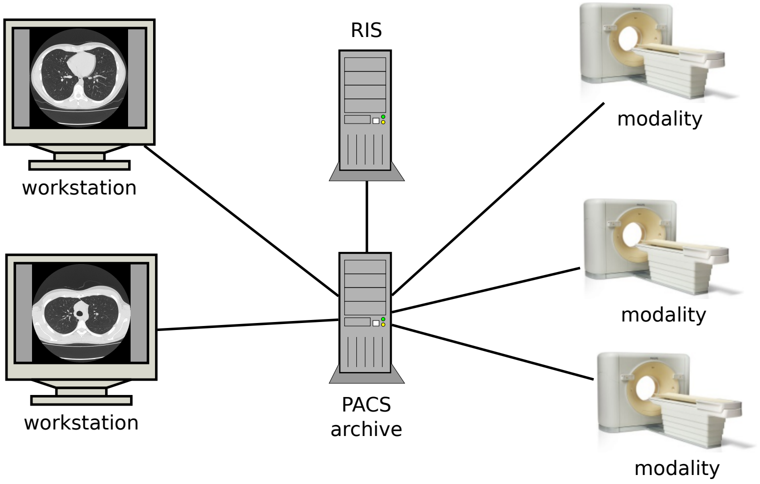


Figure 2.1: An example of a PACS, representing an archive center, a Radiology Information System, workstations, and modality devices.

Figure 2.1 represents a small and basic example of the components comprising a PACS, as well as the communication links between them. The geographic location of each component is not depicted, and may vary upon the facility's structure and whether the interconnection of multiple facilities is considered. This concept does not mandate how large the system should be, what components must be present, nor how they should communicate. In the early days, a PACS for an intensive unit could be as simple as a scanner, a film digitizer, a video monitor, and a base band communication system to link them. On the other hand, a large-scale PACS system in the mid 90's would comprise at least three or four modalities along with multiple workstations inside and outside of the radiology department, and it would be capable of handling over twenty thousand radiological procedures per year [19].

Acting as a central piece to the Picture Archiving and Communication Systems architecture, the archive server implements services for storing and retrieving medical images. By following a common protocol such as Digital Imaging and Communications in Medicine (DICOM), other components may easily communicate with the archive for these purposes. Display workstations are capable of querying the Picture Archiving and Communication Systems archive and exhibiting the images on a display, usually of high resolutions. Modality devices are mainly composed by acquisition scanners used for capturing images of a patient or other artifacts. These systems, with the guidance of a radiology, will transform the captured raw data into images, relying on algorithms which depend on the type of modality. The device may then transfer the digital content to a storage unit or a visualization workstation.

In addition to the components of a PACS, many hospitals and clinics have information systems that contain useful information related with the studies stored, and they usually work in an integrated way. The Radiology Information System (RIS) manages patient-related records for scheduling and tracking examinations. The Hospital Information System (HIS) provides patient registration, pertinent insurance, demographic information, and admission discharge transfer data [20]. All of these information systems, and many others defined by type, usually fall under the umbrella-term of Electronic Health Record (EHR) systems, which embody all medical data for purposes of healthcare, namely setting objectives, planning patient care, documenting their delivery and assessing their outcomes [21].

Distributed PACS

PACS can also be distributed over multiple geographical locations, thus creating a collaborative environment over several institutions. As long as each facility's PACS center is interoperable, a practitioner can seamlessly be provided with medical information and clinical records from institutions in their partnership [22]. Moreover, the effective storage of medical imaging can be outsourced to a cloud [23]. Cloud computing providers offer highly elastic computer system infrastructures that provide on-demand services in a rapidly scalable fashion, all in a pay-as-you-go billing strategy. Cloud storage services are reliable and provide a flexible storage capacity, as well as automatic maintenance and data redundancy. Solutions for *PACS as a Service* are already available, and its impact on healthcare may be quite promising [24]. Nevertheless, a distributed PACS approach also brings several new issues that should be tackled before exploring these systems in practice:

- Medical records are highly confidential, making proper encryption and other security techniques a vital requirement when conceiving these solutions;
- Since it is often unknown *where exactly* the data is contained in the cloud, the specific legislation applied to the data, as well as the level of security provided, are unclear [25];
- Retrieving medical images from the cloud instead of a local repository brings an additional access latency at network-level, which needs to be overcome with cache and pre-fetching mechanisms [26].

2.2.2 Interoperability

A typical PACS is backed by an assortment of technologies for communication between components, whether in the same PACS or a remote one. The DICOM standard (detailed in Section 2.3 currently stands as one of the most important. In addition, the Health Level 7 (HL7)[1] is an American National Standards Institute (ANSI) standard that attempts to provide a comprehensive framework and mechanisms for integrating, sharing, and retrieving electronic health information. Without digital communication standards, integrating a new device to the system would be difficult and may cripple its functionality.

In the case of distributed PACS however, additional concerns are imposed. Given that medical data are usually sensitive, institutions are reluctant to their exchange over a wide area

---

[1] www.hl7.org (last accessed in January 2019)

network. In addition, other legal and ethical issues arise regarding their ownership, integrity and licensure. In order to address this matter, the Integrating Healthcare Enterprise (IHE) initiative [27] has provided guidelines for the integration of information systems in a healthcare environment by establishing content profiles, namely the Cross-Enterprise Document Sharing (XDS), and more specifically, Cross-Enterprise Document Sharing for Imaging (XDS-i) [28]. The latter complements the original profile with concepts found in the medical imaging field, hence contemplating PACS in the XDS workflow. These profiles make feasible and transparent the establishment of new data protection and outsourcing mechanisms in medical workflows, while maintaining the demanded levels of privacy and confidentiality of patient data [29].

2.2.3 Overall impact

The benefits of adopting a PACS-based workflow are significant [30], and the deployment of PACS around the world has brought multiple advantages:

- *Storage was made more durable and more compact.* With the currently available technologies for persistent digital storage, an enterprise hard disk drive can hold several thousands of studies containing hundreds of images. Stored images can be easily backed up and transferred to multiple storage locations.
- *Fast and powerful information systems and technologies* were made for the analysis and retrieval of medical images, leading to useful and flexible image search engines, as well as a foundation for data mining applications. Computer aided diagnosis is a fast-growing field for investigating the automated analysis of provided records for aiding diagnostic procedures. By relying on a PACS, medical records can be easily retrieved in a digital format, making them suitable for data processing.
- *Medical images can be accessed anywhere.* With the adequate provision of web services, a medical practitioner can retrieve and observe clinical records and images at home or while commuting, from a personal computer or even on a mobile device.
- *Enhanced visualization.* With the latest visualization software, images can be easily manipulated for observing any particular detail, ranging from viewport transformations (panning, zooming, and rotating) to color space transformations (adjusting the color window's center and width). In addition, the practitioner can directly perform annotations over images, thus embedding information regarding a diagnosis. Furthermore, a combination of multiple slices of a series can be used to construct a 3D model of the captured body parts.
- *Potential reduction of radiation dose.* Digital acquisition systems can achieve an improved image quality over screen-film radiography without an increased emission of radiation [31], which is known to be harmful and nearly non-decadent. Rather, at some level of radiation, increasing the dose will barely improve the resulting image quality, unlike in the traditional analogue acquisition. Moreover, the information systems in

a PACS can easily monitor the radiation dose in a patient-centric fashion [32]. This behavior of minimizing the amount of radiation dose applied to patients for an accurate diagnosis is attributed to the As Low As Reasonably Achievable (ALARA) principle.
- *Teleradiology*, which is the electronic transmission of medical images from one geographical location to another for the purposes of interpretation and consultation, also became feasible with PACS [33].

2.3 Digital Imaging and Communications in Medicine (DICOM)

Although PACS as a concept is revolutionary in the medical field, its many benefits should not be taken for granted. Much of its success is coupled with related initiatives to achieve interoperability among medical systems. With the combined effort of both industry and academia, the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) started the development of a standard for digital medical imaging, which was named ACR-NEMA. Later on, the name of the standard was changed to Digital Imaging and Communications in Medicine[2] [34] and the DICOM Standards Committee was created for the standard's maintenance.

The DICOM standard facilitates interoperability of medical imaging equipment by specifying [35]:

- A set of protocols for network communications that must be followed by devices claiming conformance to the standard, as well as the syntax and semantics of commands and associated information that can be exchanged using these protocols;
- A file format and medical directory structure to facilitate access to the images and related information such as meta-data and structured reports;
- For media communication, a set of media storage services to be followed by devices claiming conformance to the standard;
- Information that must be supplied with an implementation for which conformance to the standard is claimed.

The latest version of DICOM at the time of writing, 2018e, consists of 21 parts, presented as independent documents. The following sections will cover the essentials of the standard, in compliance to this version.

2.3.1 DICOM Information Model

The DICOM Information Model (DIM) defines the structure and organization of the information related to the communication of medical imaging data [36]. In DICOM, every

---

[2] `http://www.dicomstandard.org`

piece of data, including medical imaging files and network operations, is represented as a DICOM object [37]. This object-oriented approach is made by predefining Information Object Definitions (IODs) and Information Entities (IEs).

IODs behave as templates for specifying attributes that would be contemplated in real-world objects. Serving as an example, IODs are specified for each particular modality, such as CT or MRI. On the other hand, an IE specifies a collection of attributes typically present in a real-word entity (e.g. a patient). IODs are *normalized* when their collection of attributes represents a single IE, whereas *composite* IOD contain properties from multiple real-world entities or their constituent parts. Regardless of which kind, any DICOM object is an instance of one of the IODs formally specified in the standard.

Each attribute in an IOD is encoded as *data elements* (Figure 2.2). An attribute is specified by a tag, a value representation (VR), and a value multiplicity. The tag is in the format `(group.element)` to identify them, where `group` is the group number and `element` is the element number in that group.

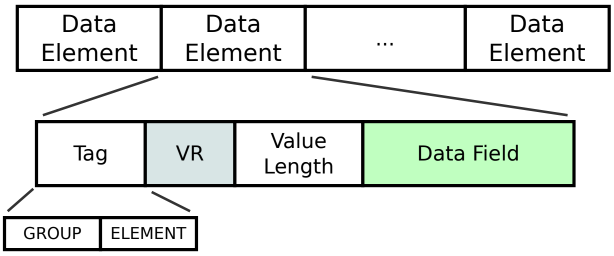


Figure 2.2: A representation of a DICOM object, as a sequence of data elements.

DICOM follows a Patient-Study-Series-Image hierarchy (Figure 2.3). A patient is submitted to one or more studies throughout their life, each identified with a Study Instance UID. These studies can involve multiple series, which in turn contain a sequence of Service-Object Pair (SOP) instances. All parts of the hierarchy are uniquely identified. The DICOM file format is used to contain these images, typically one by one in a file system.

The standard defines an extensive *data dictionary* that maps each tag to a particular piece of information [38]. For instance, the tag `(0x0010,0x0010)` is used to specify the *Patient's Name*. The latest version of the dictionary contains over three thousand definitions. Although many of these data elements are reserved for specific information, all tags with an odd numbered group (e.g. `0x0001`) are private and can be customized to suit a particular need in an institution, by specifying an extended data dictionary.

These DICOM objects are made persistent using the DICOM file format. A typical interpretation of a DICOM file containing an image is presented in Figure 2.4, which separates the header from the actual image content. Although a distinction between the image and the remaining meta-data was made, both are encoded as a group of data elements. The image's pixel data, for instance, is usually kept under the tag `(0x7FE0,0x0010)`. Additional attributes

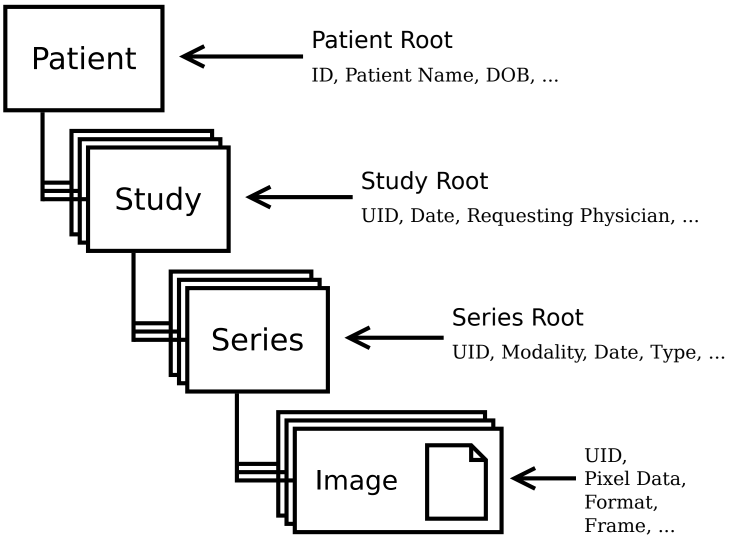

Patient
Patient Root
ID, Patient Name, DOB, ...
Study
Study Root
UID, Date, Requesting Physician, ...
Series
Series Root
UID, Modality, Date, Type, ...
Image
UID,
Pixel Data,
Format,
Frame, ...

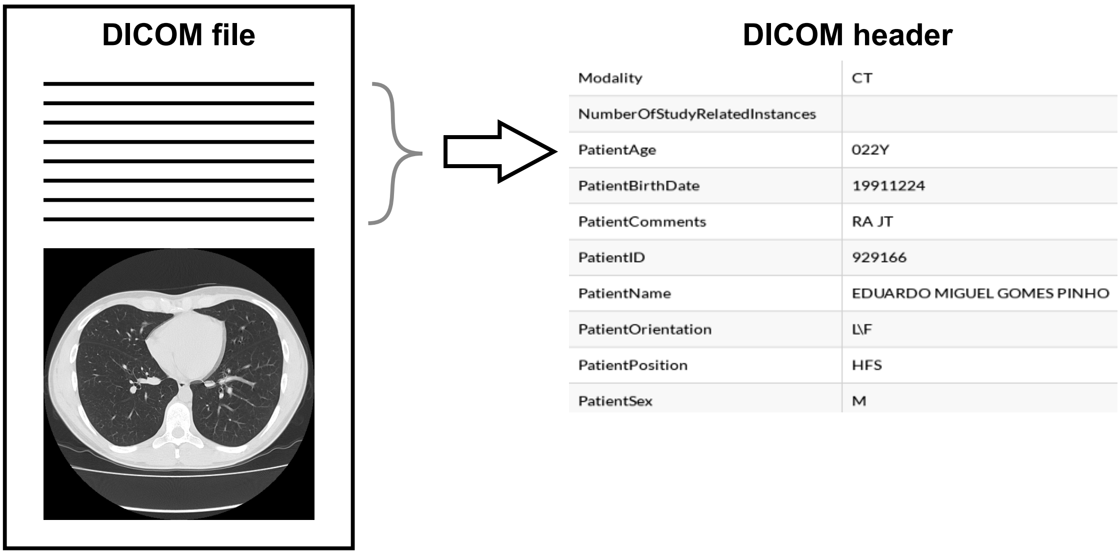
Figure 2.3: DICOM information hierarchy: Patient-Study-Series-Image.

are also contemplated for the interpretation of this pixel data to be possible, such as the image's width and allocated bits per pixel. Furthermore, it is not mandatory for a DICOM file to contain an image: it may hold nothing more than meta-data, or a variable number of media objects, such as video and waveforms.

2.3.2 Structured Report

In spite of the strong focus on images in a PACS archive, there comes a need to combine non-image clinical data with a more complex structure than a collection of simple attributes. For instance, a medical report may wish to refer to multiple media instances and establish relationships between them. Earlier systems have printed images and text into a single image, with the heavy disadvantage of making the information harder to retrieve. Even when producing a purely text-based report, this information needs to be contained in a way that facilitates retrieval, interpretation and transmission across clinical departments.

DICOM defines the Structured Report (SR) entity for these purposes, constituting specific rules for encoding, transmitting and storing imaging diagnostic reports, as well as other kinds of structured documents of low ambiguity [36, 37, 39]. An SR is hierarchically structured, in which information elements, described by a name and a value, compose a data tree. The relationship between any two elements has an explicit type (*contains*, *properties*, . . .). Rather than embedding media content, the report maintains references to multiple kinds of information, which avoids redundancy and favors consistency. Each element is assigned a code as its name, which identifies a concept in the SR. The standard does not impose a specific vocabulary of codes, nor does it apply any constraints over the vocabulary to use in a document. This makes it possible to use a well standardized lexicon such as RadLex [40].



Modality	CT
NumberOfStudyRelatedInstances	
PatientAge	022Y
PatientBirthDate	19911224
PatientComments	RA JT
PatientID	929166
PatientName	EDUARDO MIGUEL GOMES PINHO
PatientOrientation	L\F
PatientPosition	HFS
PatientSex	M

Figure 2.4: A DICOM file containing meta-data and an image. A fraction of the DICOM header is detailed to the right.

Unfortunately, many diagnostic reports, even to this day, are written in free text. The challenge of automated retrieval from reports in free text has been tackled with an assortment of text mining techniques [41]. The proper use of DICOM structured reports brings several benefits, from a better communication with the referring physician to the explicit inclusion of coded semantics [39], which forms useful relationships between concepts that may be used in knowledge-based medical information retrieval (more details in Section 3.2.3). To summarize, the combination of meta-data with DICOM structured reports makes for an invaluable source of information that ought to be processed for the purpose of diagnosis assisted by information systems.

2.3.3 Services

Completing the interoperability layer among medical imaging systems, DICOM also specifies protocols and services to be implemented by medical devices. The DICOM network protocols still abide to the elements encoded under the format tag-length-value, in which some elements may represent network operations. All DICOM communication scenarios contemplate multiple devices called Application Entities (AEs), which may be either Service Class Providers (SCPs) or Service Client Users (SCUs). In both cases, a unique title is used for identification (AETitle).

The base DICOM protocol sits over the TCP/IP network layer, and describes a set of basic services backed by one or more network commands. These commands are particular DICOM objects, always in group `0x0000`, are usually called DIM service entities (DIM-SEs), and define both the request and the response [37]. Query/Retrieve is one of the most relevant services, as it specifies commands for finding and retrieving DICOM Objects. The verification service allows an AE to check the status of another AE. Many other service classes exist, such as Storage, Modality Worklist and Print Management. All services are thoroughly covered in part 4 of the standard [42].

In addition, the standard also defines a set of web services for providing a similar range of communication capabilities over HTTP [43], thus leading to a better integration with web applications. These protocols can either provide the results in an XML-, or JSON-based format. Multiple communication modes are currently presented in the standard. For instance, WADO-RS enables the retrieval of DICOM objects by UID. STOW-RS enables the storage of DICOM objects to the system. QIDO-RS provides the means to search for DICOM studies, series and instances with the use of attribute-based text queries (e.g. `PatientID=11235813`).

More of these services are available and described in part 18 of the standard [43]. Unfortunately, although they may seem to be feature-complete, the standard currently offers poor DICOM object search methods. In fact, these services do not support queries composed of free text (e.g. “John Doe thorax”), nor do they contemplate searches for similarity with other objects (as in Query by Example). Considering the vast amount of information present in DICOM, custom search solutions in a PACS environment have to be developed in order to fully leverage the potential of DICOM data.

2.4 Dicoogle

One of the most noticeable trends in healthcare over the last years is the continuous growth of data volume produced and its heterogeneity. The concept of PACS, alongside the technological endeavors of DICOM-compliant systems, are deeply grounded in medical laboratories, supporting the production and providing healthcare practitioners with the ability to set up collaborative work environments with researchers and academia for study and improve healthcare practice. However, the complexity of those systems and protocols makes difficult and time-consuming to prototype new ideas or develop applied research, even for skilled users with training in those environments.

Dicoogle appears as a reference tool to achieve those objectives through a set of resources aggregated in a form of learning pack. It is an open source PACS archive that, on the one hand, provides a comprehensive view of the PACS and DICOM technologies and, on the other hand, provides the user with tools to easily expand its core functionalities.

Dicoogle[3] [44] is an extensible, platform-independent PACS archive software that replaces the traditional centralized database with a more agile indexing and retrieval mechanism. It was designed with automatic extraction, indexing and storage of all meta-data detected in medical

---

[3] `http://www.dicoogle.com`

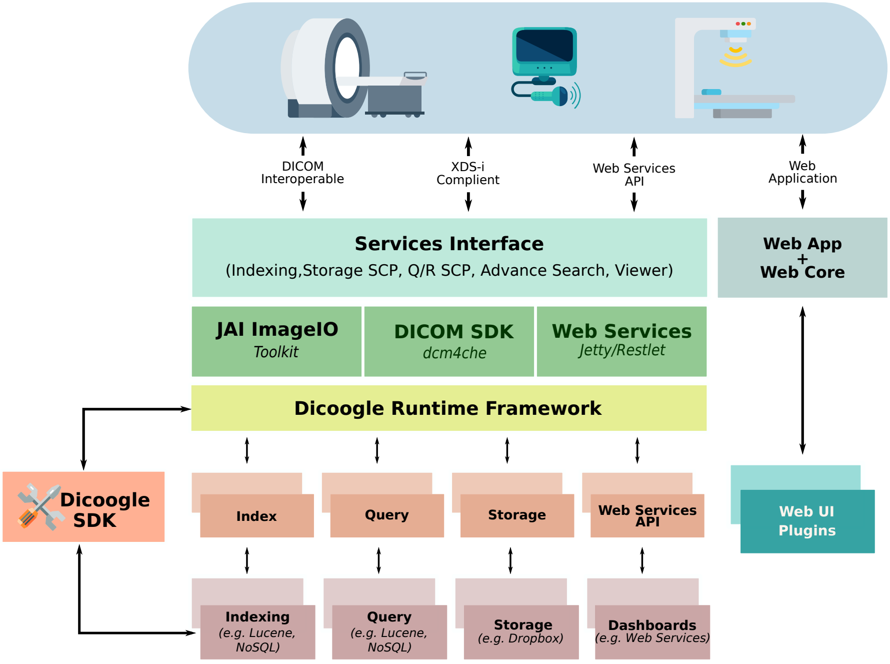


Figure 2.5: General architecture of the Dicoogle framework.

images, including private DICOM attribute tags, without re-engineering or reconfiguration requirements [45], and its software design, which will be described next, favors extensibility without direct manipulation of the core runtime. Furthermore, Dicoogle also supports a collaborative aggregation of multiple PACS in a Peer-to-Peer (P2P) network [46]. P2P groups are self-organized and self-configured, thus taking advantage of a collaborative PACS environment without a centralized management infrastructure.

Figure 2.5 presents the architecture of Dicoogle. The design of the framework is separated in five distinct categories according to its functionality: *Index*, *Query*, *Storage*, *Web Services*, and *Web UI Components*. Each category is tied to a specific Application Programming Interface (API), which is implemented by independently developed modules called plugins. Through these interfaces, the plugins provide operations which are orchestrated by the core Dicoogle platform. The life cycle of each plugin is controlled by Dicoogle's core, as it scans the plugin directory on start-up and identifies the plugins which are loaded according to the set of directives present in the configuration file. Dicoogle core sees the plugins as completely independent from each other, being only accessible via the respective interfaces. If there is the need to share state between plugins, the dependencies are solved inside the `PluginSet` class, which represents a set of plugins and serves as an entry point and management entity to the overall structure of the extension. Here, plugins of several categories are aggregated into one functionally consistent unit, thus simplifying the development and deployment process.

The Dicoogle Software Development Kit (SDK) provides the necessary software components for interconnecting plugins with the core system. Although it is not presented as a center-piece, this SDK simplifies development by establishing the APIs that the plugins must abide to, as well as implementing interfaces with the system for accessing indexed and non-indexed data, dispatching tasks, fetching settings and logging the software's functions.

Dicoogle contains several key features that foster new ways of looking into meta-imaging information for retrospective assessments. Due to the variety of data and their possible interpretation (textual, visual, hierarchical), the system delegates all indexation tasks to *indexer plugins*, thus making them responsible for data indexing in a format that allows quick access to the stored information. These plugins also contain all the required procedures and dependencies for extracting and storing information from DICOM files. Therefore, unlike many other PACS systems, Dicoogle can provide DICOM data indexing and retrieval supported by non-relational databases. A production-ready plugin for using an Apache Lucene index as the database is currently available, and other NoSQL databases were also tested [47].

Dicoogle can also be extended for statistics and task reporting. As an example, some wide-ranging clinical studies require dose metrics that are now increasingly available in DICOM persistent objects [48, 49]. Dicoogle can identify inconsistencies in data and processes by efficiently enabling multiple views over the medical repository, as well as exporting data for further statistical analysis. This tool can be used to audit PACS information data and contribute to the improvement of radiology department practice. At present, the system has been in use in several hospitals and more than 20 million DICOM image meta-data objects have been indexed.

2.4.1 Extensible User Interface

Dicoogle has been with a modern single-page web user interface, which can be delivered to local and remote users without a previous installation process on the client machine. Moreover, in order to accommodate new use cases from both ends of the software stack, the core Dicoogle project has been augmented with a dynamic pluggable web component architecture. Rather than manually including new menus and actions to the web application, the Dicoogle *Web Core* establishes a backbone for web User Interface (UI) component retrieval and rendering at the browser's run-time. This enables the exposure of new features on the main web application without rebuilding the original source code. This mechanism was developed as a minor contribution to this thesis work, as it makes way for a more seamless integration of novel user interfaces for searching in Dicoogle.

To achieve this in practice, three components were developed:

1. A set of web services in the Dicoogle web server for providing the web-based plugins and their meta-information;

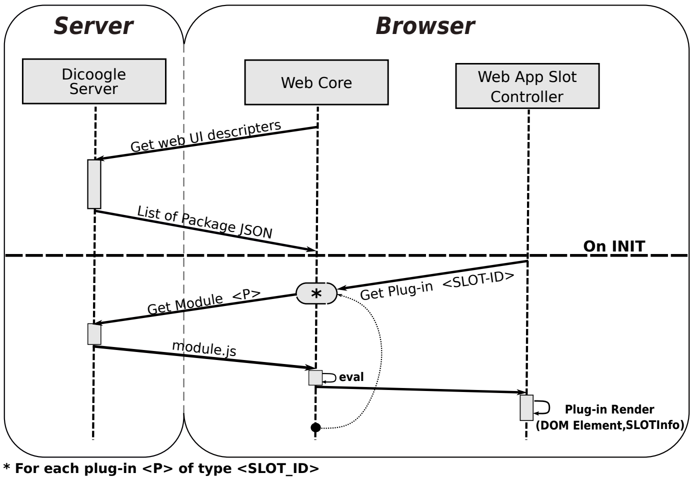


Figure 2.6: Dicoogle Web Core sequence diagram

2. The Web Core component, a JavaScript library that runs in the user's browser for retrieving and exposing the extra user interfaces;
3. A project scaffolding tool to facilitate the creation of web UI plugins.

Each user interface plugin is scoped to a particular view of the application, into divisions called slots. Slots describe the kind of interface that a plugin for the same slot is meant to portray. For example, the `settings` slot is available in the management section. When observing a list of results, a `result-entry` slot is created for each item, so as to perform a particular operation over an item, such as visualize the image in a separate service). In contrast, plugins of the type `result-batch` are scoped to the entire list of results, allowing the exposure of new forms for data visualization, exporting, and manipulation.

The interactions of the Web Core with these slots and the main web application are represented in Figure 2.6. Once the user is logged in, the application fetches the list of web user interface plugins available for that user, using standard asynchronous requests. The actual components, specified as JavaScript modules, are subsequently loaded and rendered to the indicated slot on the page. The Web Core architecture relies on standard HTML5 web components in order to augment slot elements for this functionality.

2.4.2 Learning Pack

Medical imaging informatics is a major subject and teaching offer is becoming increasingly frequent. The notable increase in higher education degrees and courses that combine radiology concepts with medical imaging informatics is a direct consequence of their impact in healthcare. For instance, in the University of Aveiro, it exists a subject (see section 5.3) that is simultaneously attended by students of radiology (technicians) and engineers. The teaching of systems and networks in medical imaging comprises two complementary perspectives:

- Experts in *medical imaging* may understand the main purposes of a PACS archive and know how to use them for research and clinical use, including some understanding of the technical challenges overcome by these systems;
- From a perspective of *computer engineering and software development*, students are given the necessary background of PACS systems and networks, including the DICOM standard, in order to develop end-user solutions.

Therefore, teaching how to use an open source PACS archive in a medical imaging informatics course is an essential counterpart to the development of the actual software. With this in mind, the Dicoogle Learning Pack was created to teach users on both typical and new advances of the field in an academic environment, backed by the Dicoogle archive.

The Learning Pack[4] is available as an open-access static site hosted on GitHub Pages. The site itself is open source, and users of Dicoogle are invited to read and provide their feedback through GitHub's issue tracker. This approach takes advantage of freely available resources for open source projects, while making them accessible to researchers and students alike. The learning pack provides also documentation, configurations and code examples, guiding the user to:

- Set up Dicoogle for the first time on a local machine;
- Index a data set and perform searches over the indexed data;
- Develop plugins for the Dicoogle back-end, in which a background in Java is assumed;
- Develop web user interface plugins for the Dicoogle web application, in which a basic background in JavaScript is assumed;
- As a more advanced topic, build and debug the main Dicoogle application.

2.4.3 Dicoogle as a framework in Research

The following sub-sections describe two recent use cases of Dicoogle employed as a foundation for research in medical imaging informatics.

Controlled Searching in Reversibly De-Identified Medical Imaging Archives

With the technological assistance of the Dicoogle platform, a reversible de-identification mechanism was developed by Silva *et al.* [4]. In this use case, depicted in Figure 2.7, standard medical imaging objects are fully de-identified, including DICOM meta-data and pixel data. At the same time, it provides a reversible de-identification mechanism that retains search capabilities from the original data. The goal was to deploy this solution in a collaborative platforms where data is anonymized when shared with the community, but still searchable for data custodians or authorized entities.

---

[4] `https://bioinformatics-ua.github.io/dicoogle-learning-pack`

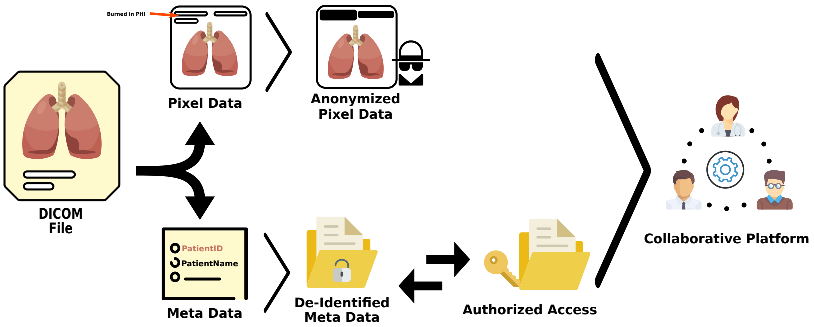


Figure 2.7: Use case diagram of a reversibly de-identified medical imaging archive (adapted from Silva *et al.* [4]).

Whole-slide Imaging and Pathology

The main focus of pathology is based on the detection of morphological anomalies and finding possible relations with functional disorders of tissues, diagnosing a disease.

During the last decades, *digital pathology* has been arising as a new branch of pathology [50]. Digital pathology is the aggregation of hardware and software designed to substitute traditional devices, like microscopes. This field of pathology emerged to fill the need of better conditions of accessibility, cleaning, protection, and storage of the old glass slides, taking advantage of the decreasing cost of digital storage and distribution. The acquisition of images in Whole-slide Imaging (WSI) is performed by whole-slide scanners. The images obtained can have a resolution of several Gigapixels, hence consuming large amounts of space when stored digitally.

The developed system by Godinho *et al.* [51] contributes to a performant and fully DICOM compliant viewer of whole-slide images using Dicoogle as the base PACS archive. Dicoogle was also extended to support the automatic creation of DICOM WSI image pyramids, as well as a tiling engine. The viewer uses standard DICOM web services implemented as Dicoogle plugins, as shown in Figure 2.8.

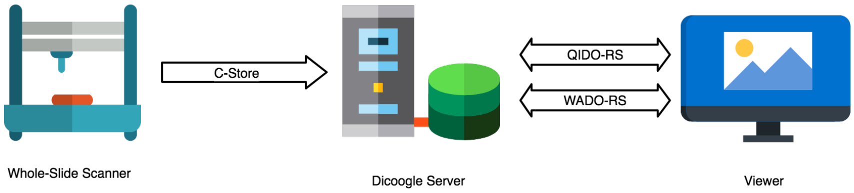


Figure 2.8: Overview of the architecture of the viewer (adapted from Godinho *et al.* [51]).

2.5 Final Remarks

Medical imaging in modern times have heralded a panoply of challenges to be tackled by computer systems, ever since the acceptance of digital imaging, and subsequently PACS as a well known concept in the field. The DICOM standard builds a common ground on how systems in medical imaging represent data and communicate. The increased requirements of medical data storage and retrieval, as well as the emergence of new modalities, have brought researchers in computer science and engineering to investigate these concerns in a fashion that can be accommodated into the PACS mindset, with the archive as a potential cornerstone and a *rendez-vous* point of integration.

Dicoogle, a project that once pioneered in DICOM data mining, fosters high extensibility in its current design, and as such is one of the key components for the realization of this thesis. This proposal will progressively present the use of Dicoogle towards the ultimate goal of multimodal medical information retrieval, making a statement on its current content-based image retrieval capabilities (in Section 3.3.5), and further along the document, a multimodal search engine for medical images (in Section 4.3) and an automated labeling system (in Chapter 5).

Chapter 3

Content Discovery and Information Retrieval

In this chapter, a second foundation is presented under multiple sub-sections: it starts with an overview of several concepts under the content discovery umbrella term, including machine learning and deep learning. The second sub-section explains the concept of information retrieval, followed by content-based retrieval. Each sub-section is bridged to their respective use cases and applications in medical imaging.

3.1 Content Discovery

The term *content discovery* may regretfully possess an overloaded meaning in literature, even under the scope of information systems. For lack of a better term, it is defined in this document as a field encompassing all scientific fields related to the processing and analysis of digital data, including information and knowledge extraction. Although not considered part of the information retrieval domain, searching for information may be perceived as a form of content discovery. The latter encompasses a variety of important concepts in the former, that will be mentioned throughout this document, and as such are specified next.

3.1.1 Structured and Unstructured Data

Digital data can exist in multiple forms. *Unstructured data* refers to data which is not arranged in a way that can be easily accessed by a computer. It is the opposite of *structured data*. A significant amount of research on information retrieval aims to extract information from unstructured documents. Textual data may, or may not, have a structure, depending on the level of semantics involved: a text document will often have a header, footer and a sequence of hierarchically structured sections. However, when attempting to obtain the actual concepts presented in the text, this kind of structure does not match the one that would otherwise be seen in a knowledge base, containing semantic relationships between objects such as *is-a* and *has-a*. Only the latter form can meaningfully inquire the assertions in the text.

3.1.2 Data Mining

Classification is making a decision over documents or fragments on what particular kind (or class) of document they belong to. For instance, an e-mail may be classified as "spam" or "not spam" depending on its content. The process is typically undertaken by a *machine learning* model, which is trained with multiple known samples, and after which a computer may automatically predict the classes of new documents with enough accuracy. When a class family contains more than two classes (for example, when categorizing an e-mail into one of *financial*, *travel*, *social*, *work*, etc.), it is a *multi-class* classification procedure. When predicting multiple binary, non-exclusive classes (such as the presence of *microcalcifications* and/or *calcifications* in a mammogram), we have *multi-label* classification.

Clustering is the task of collecting items into multiple clusters, in a way that similar items will reside in the same cluster [52]. These algorithms are a form of unsupervised classification, and are useful when there are complex semantic concepts for item categorization. This means that the decision of item grouping is driven by the data itself, rather than labels that may be annotated or predicted through classification. This is the case for *k-means clustering*, which assigns data points to one of k sets in a way that minimizes the sum of distances to their respective centroids.

Text mining is the sub-field that regards the extraction of information from various text sources. It is typically tied with **natural language processing**, in which documents built for human understanding are translated so as to be understood by computer systems. Although this subject deviates from the scope of this thesis, it is nevertheless worth noting that a large body of knowledge resides behind documents written in natural language. In the medical domain, the PubMed Central (PMC)[1] is a notable example of a free full-text archive of literature in biomedical and life sciences.

3.1.3 Deep Learning

At the time of writing, deep learning is one of the most prominent and most advertised forms of artificial intelligence in present days [53]. It is often admitted as a subset of *representation learning*, which in turn, is a specific subset of *machine learning*. The concept emerged along with the concept of artificial feedforward neural networks, also called Multi-layer Perceptron (MLP). A feedforward neural network defines a model as a function f composed of multiple operations in the form of a directed acyclic graph. A layer-based perspective on feedforward networks is usually made, in which a model is composed of a sequence of layers $f_1, f_2, ...f_k$, with trainable parameters (also called weights) and an appended

---

[1] `https://www.ncbi.nlm.nih.gov/pmc/`

non-linearity, and coupled together in a chain: $f = f_k \circ f_{k-1} ... \circ f_1$. Each layer learns a new representation of the original data, thus contributing to a hierarchical interpretation, with each level creating more complex abstractions. The last layer of the network is often called the *head* of the model, and establishes its output domain. For example, a layer with fully connected (dense) networks followed by a softmax activation may serve as the output of a multi-class classifier. Neural networks are trained to minimize a *loss function*, which defines how far the model is from the expected outputs. In practice, an optimization algorithm based on gradient descent is usually employed, where gradients are backpropagated through the network, leading to the update of all weights across layers. Depth enables these models to capture complex patterns with less patterns than with shallow networks. Alternatives to layers of fully connected neurons exist. For instance, it is very common to use convolutional layers in image recognition, as they create a spatial interpretation of the data. These layers comprise multiple trainable kernels (or filters), and usually result in a significantly smaller number of parameters than a dense layer.

Additional attention was granted to deep learning when a deep Convolutional Neural Network (CNN) significantly outperformed existing methods in the ImageNet classification benchmark [54]. As speculated in their work, the training of a relatively large number of hidden layers (5 convolutional layers, and 2 densely connected layers) was made possible with a varied assortment of techniques (Rectified Linear Unit (ReLU) activations, max-pooling, drop-out [55], etc.), along with the computational power of Graphics Processing Units (GPUs). In recent times, state-of-the-art neural network architectures can exhibit a much larger number of layers (some of which experiment with more than 100 layers), although other lines of research focus on reducing the capacity of these models while achieving competitive performance [56]. One of the unique benefits of deep learning resides in its capability of working well with original data directly, even without a feature extraction step. Nevertheless, some shortcomings of deep learning approaches do exist:

- Deep neural networks usually require significantly more computational resources and time for training and inference.
- As these models have a significantly larger number of trained parameters than other methods, they are more prone to overfitting the training data.
- Although neural networks reach their full potential with depth, this also requires large amounts of data to provide meaningful outcomes [57].
- Training and designing a deep neural networks involves a form of expertize that is quite distinct from other machine learning algorithms, in the form of hyper-parameter tuning (as in, adjusting the learning rates and related properties of the model), normalization layers, auxiliary losses for regularization, and many more details which can more or less influence the outcomes of an experiment.
- Even in the face of a large data set, neural networks may be victim of spurious biases

and correlations between factors that should not be dependent. As deep learning is still a hot topic in research, multiple studies emerge every year with the goal of attaining a better reasoning on their performance, and on how they understand the problems that they were designed to solve [57].

- Research papers in deep learning are also a pertinent source of poor claims of improvements, either from a lack of time and computational resources to better assess the given technique, or from the inability to identify the causes of the empirical gains observed [58]. For instance, although a paper could be promoting a specific algorithm by showing that its use results in better metrics at the given task(s), the actual improvement could be, for the most part, the consequence of better hyper-parameter tuning.

Despite these drawbacks, deep learning has become fairly ubiquitous in data science, as a consequence of its significant prospects of problem solving with high-dimensional data, easy access to literature (most authors in the field publish pre-prints to their work) and open software frameworks for training and deploying models (such as TensorFlow [59]). Given the significant importance of *representation learning* in this thesis, it will be covered with greater detail in Chapter 6.

3.1.4 Machine learning for image understanding

Machine learning has also been widely used in CBMIR for the last few decades [60]. The use of statistical methods of automated decision making and modeling made its way into Computer Aided Detection and Diagnosis (CAD) solutions, expanding the possibilities of research on this subject. For instance, by taking existing clinical information from human-diagnosed medical cases, a CAD system may apply supervised learning techniques in order to predict the diagnosis of future cases. That is, by assuming that the effective diagnosis can be modeled by a function f, where $y = f(x)$ and x is a vector of reported data regarding the case, algorithms such as Support Vector Machines (SVMs) can learn previous data points representing clinical occurrences (pairs in the form $(x, f(x))$) and obtain a sufficiently approximated model for f. As a concrete example, the detection of microcalcifications in mammograms is one of the most studied fields where the use of machine learning has shown very bright results [61]. A more extensive overview of the state-of-the-art in CAD is provided in Chapter 5.

3.2 Information Retrieval

Although in the 90's society has preferred to obtain information from other people rather than proper information retrieval systems, searching for information on the web has become a standard in the last two decades, with the overall increase of search engine quality.

Information retrieval did not begin with the web, however. The field began with scientific publications and library records, but enlarged its domain of appliance to other professional activities, such as healthcare. Nevertheless, the World Wide Web (WWW) is currently the largest "super"-source of information, unleashing publication at the scale of tens of millions of content creators. The current amount of data on the web is too large to be indexed by a single machine in its entirety. However, modern search engines have managed to attain high-quality results within a fraction of a second per query, while handling hundreds of millions of searches a day over billions of web pages.

The previous chapter, depicting the state of medical imaging informatics in recent years, also reveals increasing utilization, complexity, and demand for the various components of a PACS. The act of searching through a medical repository is part of several workflows in practitioners, and benefit from being improved from multiple dimensions:

- As in many other use cases for computer systems, retrieving data *quickly and efficiently* is an important matter that can be addressed in many ways, depending on the expected requirements.
- When shifting healthcare organizations towards digital imaging, the adoption of a patient-centric model becomes more pertinent: systems that *distribute and federalize* medical data in a geo-distributed PACS would benefit from receiving information about a patient across institutional boundaries, while providing additional fault tolerance through redundancy [22].
- And also, as will be the focus of this thesis, the attribution of relevance in the data upon a search.

3.2.1 Definition

The definition of information retrieval may become excessively broad. For instance, something as simple as checking a person's phone number in the yellow pages is a form of information retrieval. Manning, Raghavan, and Schütze [62] defined it as thus: "information retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." Throughout this thesis, additional care is taken in the nature of the data, namely in what form the documents exist and what they can contain. While it is true that text is a rather usual form of keeping data, as articles, books, web pages and health records are mostly of this form, documents in a certain domain may hold image, audio or other multimedia content, in which other techniques must be applied so as to further scrutinize them.

In medical imaging, documents to retrieve may assume the form of digital imaging studies in their various forms, articles in medical imaging literature, or even online pages containing professional medical information. While these sources of data are not disregarded, the contributions of this thesis have an increased focus on the nature of medical imaging repositories and the methods through which their information retrieval capabilities can be improved.

3.2.2 Digital Information Retrieval Systems

Information retrieval is tightly related to a large assortment of techniques, technologies, challenges and trends. Perhaps the most important concept that the success of digital information retrieval can be attributed to is the **database**, which is a device (or group of devices) dedicated to the structured storage and retrieval of digital data.

Indexing is the process of generating and handling data structures for a quicker look up of relevant information. For instance, rather than traversing the entire data set for people named John, an index of names will speed up the process by narrowing down the range of candidate entries to a smaller list, potentially only containing people named John. Look-up tables may be hierarchical, thus subdividing the process into a tree traversal by portions of text.

Querying

Query formulation is the process in which a human-readable question is translated to a computer-readable inquiry to a source of information. In a very similar fashion to computer programming, these questions need to be laid out in an unambiguous format expected by the system. This process is typically performed by the end user of the information system, but may be guided by the system itself using automated processes or improved user interfaces. Searching and querying may appear interchangeably in information retrieval, meaning that the retrieval of a particular result usually involves a search procedure. The intermediate representation between the human language and the computer is a query language. Structured Query Language (SQL), for example, stands as the foundational language for querying a relational database, which in turn has influenced other kinds of DBMSs. SPARQL Protocol and RDF Query Language (SPARQL), as an example, is standardized by the World Wide Web Consortium (W3C) and is prominently used in semantic databases [63].

Boolean queries represent a combination of conditions imposing whether each document in the archive is part of the expected output (as in *yes* or *no*). The human readable query *"Which computed tomographies of male patients were performed by Dr. Jake Quarter*[2] *since the 8th of January 2015?"* could translate to the following list of filters applied to all series (in DIM-specific terms) of a PACS archive:

- the *patient's sex* is *male*;
- the *acquisition modality* is *CT*;
- the *performing physician* is *Dr. Jake Quarter*;
- the series' *acquisition date* is *2015-01-08 or later.*

Each filter would be translated to a language that depends on the underlying system, and the filters are applied simultaneously in this example, making an *intersection* (boolean

---

[2] fictional name

operation *AND*) of the results that would be obtained from each individual query. A boolean query language may also admit other boolean set operations, such as a union (*OR*).

For questions which do not have a clear *yes* or *no* response from a document, we have **score-based queries**, which rely on algorithms determining which documents in the system seem to be the most relevant for the given query. The algorithm attributes a *score* to each document based on certain metrics, with the aim of highlighting resemblances of a document with the given query, while negatively emphasizing dissimilarities with the same query. For instance, an open query with the words *"lung"* and *"cancer"* will likely yield documents containing both words, followed by (potentially less relevant) documents containing only one of them.

Query refinement may also be present in a search. When it is a manual refinement, the user will rewrite the query and try again, with hope that the new query will bring the user more pleasing results. In an automatic refinement, the search engine offers suggestions or directly modifies the query with potentially more useful terms. Moreover, queries are not limited to a textual language, and may directly accept multimedia content as a part, or a whole, of a query. This capability is covered by the domain of content-based retrieval and is further described in Section 3.3.

3.2.3 Semantics and Ontologies

Effectively, the semantics of a search will lie in the user's point of view [64]. That is, if the user is looking for images of skin cancer, a Query by Example where the query image contains (and points to) cancer cells, low-level features may suffice to achieve the desired results without relating such features to the concept of skin cancer in the system itself. However, some terms in an image's annotation may be enhanced into a semantic concept with the use of ontologies [65]. An ontology is a description of a group of concepts and the relationships between them, with the goal of formally defining a shared understanding of a domain. By associating content to semantic concepts, a traversal through the network of relationships improves the information retrieval capability by employing this reasoning mechanism over these associations [66]. The application of semantics in CBIR can bring ambiguity in a generic context, given that the same image (or other media) can bear different annotations or similarities depending on the external knowledge considered. In medical imaging however, the contents of an image are unambiguously linked to semantic concepts depending on the context [64].

Previous studies dating as far as the late 1990's have approached the challenge of representing a medical image interpretation process as a network of knowledge-based computations [67]. More recent work can be found in [64]. Semantic retrieval is necessarily backed by ontologies with a wide vocabulary of medical terms forming a knowledge base of the medical domain. The National Library of Medicine (NLM) conceived the Unified Medical Language System (UMLS) [68], a project composed by multiple resources medical knowledge bases. One of them is the Medical Subject Headings (MeSH), a thesaurus for subject indexing and searching of journal articles, currently used in PubMed[3] and many other institutions. The National Cancer Institute (NCI) has also developed a thesaurus[4] containing a broad terminology of more than 20 thousand concepts and 70 thousand terms [69]. Under the scope of radiology, the American College of Radiology (ACR) established an index of medical terms for radiological diagnoses [70]. The RadLex project, emerged later on, has outperformed the ACR index with a larger lexicon for medical information retrieval and organization [40]. As of version 3.12, the RadLex[5] ontology contains over 74 thousand medical terms. The use of field-specific ontologies such as these makes semantic retrieval feasible.

Medical information retrieval may also take advantage of the semantic web paradigm [71], which regards modern technologies for exploiting and delivering knowledge for machine and human consumption. By resorting to information and knowledge representation structures, such as the Resource Description Framework (RDF) [72] and the Web Ontology Language (OWL) [73], interconnected networks of knowledge can be derived from heterogeneous sources of medical data, including public medical articles and structured reports.

3.2.4 Medical Imaging Data and Searching

Searching for information in the medical domain is heavily influenced by the group in which the user is tied to. In [74], a great distinction is made between *laypeople*, who rely on readily available online services to identify problems and respective solutions in their daily lives, and medical experts such as doctors and radiologists. Another notable discrimination in a user's profile is made for teachers and speakers, who may collect specific images and loops with the intent to explain and discuss specific cases [75]. Last but not least, users of information systems to conduct *research* may perform searches of a wide degree of diversity, depending on their intended focus.

From a technical perspective, the DICOM standard already covers multiple communication modes for searching in an archive. For instance, WADO-rs enables the retrieval of DICOM objects by Unique Identifier (UID). STOW-rs enables the storage of DICOM objects to the system. QIDO-rs provides the means to search for DICOM studies, series and instances with the use of attribute-based text queries (e.g. `PatientID=11235813`). More of these services are available and described in part 18 of the standard [43]. The standardization of network and web APIs enable software developers in medical imaging informatics to integrate and iterate on user-facing search mechanisms independently from the archive's querying functionalities.

As presented, a DICOM-compliant archive provides QIDO-rs or query/retrieve network services for searching medical imaging data. Unfortunately, although they may seem to

---

[3] `http://www.ncbi.nlm.nih.gov/pubmed`

[4] `https://ncit.nci.nih.gov/ncitbrowser` (last accessed in January 2019)

[5] `https://www.rsna.org/en/practice-tools/data-tools-and-standards/radlex-radiology-lexicon` (last accessed January 2019)

be feature-complete, the standard currently offers poor DICOM object search methods. In particular:

- These services do not support queries composed of free text (e.g. "John Doe thorax");
- Queries other than text-based are not contemplated. In Query by Example, the system should be able to admit a medium such as an image, or a region thereof, as the query itself.

Considering the vast amount of information present in DICOM, custom search solutions in a PACS environment have to be developed in order to fully leverage the potential of DICOM data. One of the initial goals of Dicoogle was to make searching DICOM data more flexible, with the introduction of extensive meta-data indexing mechanisms [45]. By stepping away from relational databases, the user can perform queries over any of the indexed fields, including attributes admitted by the standard as private to an institution, without reengineering the software.

3.3 Content-Based Image Retrieval

Humans are gifted with a fine perception of the world and our surroundings. In addition, although it does not often come to our minds, the human brain is also great at recognizing patterns and filtering noise from our own perceptions. As a person looks at an object (e.g. a laptop), they can identify the various constituent parts (keyboard, display, touch-pad, . . .), position it in space ("on a desk approximately 2 meters away") and tell it apart from coexistent artifacts that are less relevant to that object (the table, an external monitor, a VGA cable, a mouse, . . .). When having a conversation, an ordinary person would not take the sound coming from a stereo on a music channel as part of it. These innate capabilities are key for a human being to retrieve information. On the other hand, "teaching" the same skill to a computer is a challenge with multiple levels of complexity.

3.3.1 Definition

CBIR, or Content-based Visual Information Retrieval (CBVIR), can be defined as the concept of extracting and using information stored in some kind of visual medium. The acronym CBIR may occasionally be used to refer other kinds of media, such as audio, video, and others. Regardless of the involved kind of content, CBIR systems leverage capabilities usually seen in textual information retrieval, such as searching, to multimedia content. The most frequent feature in these systems is Query by Example (QBE), which performs a search over items similar to the one provided in the query. In medical imaging, the research of CBIR is commonly narrowed down to image retrieval, given that the largest amount of medical information is in the form of digital images. CBMIR is also a common expression in the field.

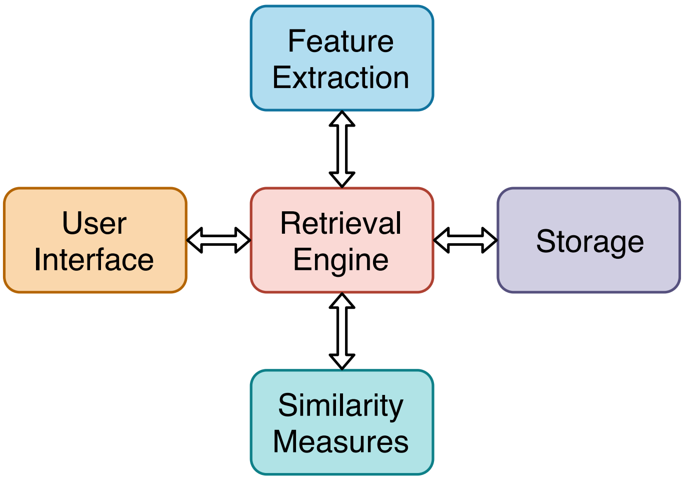


Figure 3.1: A diagram representing a typical CBIR system.

The first occurrences of image storage and retrieval in computers were in the early 1980's [76], nearly alongside QBE techniques [77]. The concept of CBIR arrived a few years after, with IBM's QBIC project [78]. Nowadays, many attempts were made to let computers learn, understand and annotate visual content, exhibiting notable improvements during the early years [79].

3.3.2 General Architecture

A content-based image retrieval system usually comprises the following main components (as depicted in Figure 3.1):

1. **Storage**: Backed by a DBMS, the storage provides access to all multimedia files to process, as well as the means of saving and retrieving features associated to that content. In a PACS, the file data source may also be decoupled from the feature set database.
2. **User Interface**: It is the main component facing the end users, allowing them to perform CBIR requests and collect/observe results. For purposes of integration with third party software platforms, this component may be replaced, or augmented, with an API, so as to expose the functionality to other computer programs.
3. **Feature extraction**: This component retrieves a descriptor of the object in the form of a vector of features. The obtained descriptor aims to contain the most essential properties of the original image, in a way that can be compared with other feature sets.
4. **Similarity Measures**: It contains algorithms establishing distance (and sometimes semi-distance) measures between feature sets, hence estimating the semantic distance between their associated objects or concepts.
5. **Retrieval engine**: Triggered by the interface (2), the engine relies on the remaining components (1,3,4) to drive queries into a list of results.

The existing number of techniques applied to these solutions are overwhelming at first, as many of them were inherited from generic information retrieval techniques and span over multiple disciplines, including computer vision, artificial intelligence and human-computer interaction [12, 80, 81]. The following topics make a presentation of some of the techniques, trends and aspects often found in medical and visual information retrieval, which is later on continued to the more focused scope of multimodal medical information retrieval (in Chapter 4).

3.3.3 Image Feature Extraction

When performing pure visual retrieval, in any field, a particular interpretation of visual data is required. The matrix of values constituting the outcome of cameras, imaging modalities, or other sensors, exist in a form of near-maximal information retention. However, this form is usually unsuitable for content-based retrieval. The naive pixel-based approach towards image comparison consists in making a correlation from one segment to another in a different image by observing the similarity in these pixels. Direct image comparison in raw pixel space represents multiple problems:

- It is *very slow*, most notably in larger images. The image resolutions involved in medical imaging (some of which to the scale of Gigapixels) would make this procedure too computationally intensive, and impractical. Images are also hard to index for similarity searching, due to their large dimensionality, and thus failing to take advantage of faster image look-ups.
- Unless more complex metrics are employed, is *excessively sensitive* to less relevant visual nuances, such as translation, rotation, and lighting differences.
- Pixel data also contains multiple forms of *noise*, such as those derived from deficiencies in image capture (as in white, seemingly random grain captured by modalities), or from parts of an image which do not relate to the main object of study (for example, a computed tomography scan of the lungs will also include portions of the body around the lungs).

Overall, the *curse of dimensionality* strongly applies when attempting to obtain a measure of similarity between high dimensional items, such as images. In a data domain of very high dimensionality, only a few factors are relevant.

In order to cope with large image data sets containing a significant amount of irrelevant or redundant information, CBIR systems include a feature extraction stage as a mapping $X \rightarrow Z$, from the digital image's domain to a new space of *visual features*. In this context, a feature is a piece of information describing a property of an image. These features may be either global or local. The former kind of features is the result of processing the entire image content. Local features are obtained when defining a Region of Interest (ROI), which is meant to highlight peculiar properties found in an image, so as to encounter similar cases in the system. Hence, feature extraction provides a smaller representation of visual information.

Visual feature spaces can be assessed on their levels of transparency and complexity. By using digital image processing techniques, a set of *low-level features* can be extracted from the image in its digital form, related to the presence of shapes, textures and colors. Histograms are often employed to quantify such characteristics into feature vectors. As a notable example, Color and Edge Directivity Descriptor (CEDD) combines color and texture information factors of an image into one compact 54-byte descriptor [82].

Feature extractor functions may also comprise multiple levels of abstraction, or they can be defined as an extensive sequence of non-linear operations. For the purpose of object detection, it is not uncommon to rely on visual key-point extraction algorithms, specifically designed to recognize local parts of visual data with scale and orientation invariance. Scale Invariant Feature Transform (SIFT) is an early example of key-point extractor and descriptor [83], which detects local changes in the image using differences of Gaussian filters, and encodes them into a representation with resilience to shape distortion and illumination [83]. This enabled the algorithm to match an object in an image with the same object in another image, even if rotated or slightly distorted. SIFT established a milestone in image recognition and detection, and other key-point extraction algorithms followed over the years. To name a few examples, Speeded Up Robust Features (SURF) emerged as a faster solution to key-point extraction [84], and Oriented FAST and Rotated BRIEF (ORB) combines existing algorithms to produce an open, yet competitive alternative to SIFT [85].

The number of key-points captured by these algorithms may vary, depending on the image and the choice of parameters of the algorithm, and as such does not directly provide a suitable global descriptor for CBIR. Fortunately, the bags of words method from text-based retrieval can be adapted to arbitrary key-points [86]. From a previously acquired image data set, their respective key-point descriptors are collected to serve as template key-points. A *visual vocabulary* (also called a codebook) of a size k (of which varying ranges and magnitudes have been tried in literature) is then obtained by performing a clustering algorithm, such as k-means clustering, on all template key-point descriptors and retrieving the centroids of each cluster, yielding a fixed size list of key-point descriptors $\mathcal{V} = \{V_i\}$. Once a visual vocabulary is available, an image's Bag of Words (BoW) is constructed with a quantization process: by determining the closest visual vocabulary point and incrementing the corresponding position in the BoW for each image key-point descriptor. In other words, for an image's BoW $B = \{o_i\}$, for each image key-point descriptor d_j, o_i is incremented when the smallest Euclidean distance from d_j to all other visual vocabulary points in V is the distance to V_i. We can picture the bag of visual words as a histogram of visual descriptor occurrences, which can be used as a global image descriptor [87].

The same quantization process can be used for colors, as seen in Bags of Colors (BoCs)

[88], where dominant color values in an image are used instead of visual key-points.

In recent years, *deep learning* methods have been observed in visual feature extraction as well. Deep neural networks have shown superior image recognition capabilities in contrast to former methods [54], which contributed to this increasing use and prominence of deep learning.

Semantic features, on the other hand, represent a particular concept with a direct meaning to the end users (e.g. pulmonary bullae with a certain size in some positions). The extraction of these features is often more challenging, requiring complex techniques, and sometimes human intervention for a reasonable accuracy. These features in medical images provide a greater level of usefulness, as they cross the bridge over the semantic gap towards a preliminary diagnosis (e.g. detection of bullous emphysema from pulmonary bullae [89]).

3.3.4 Feature Similarity Measures

The choice of algorithm for determining feature similarity is also quite relevant in CBIR. In fact, some types of features may require specific feature similarity measures, posing no scientific meaning otherwise.

The definition of *similarity* usually employed by these systems is that of a function resolving the feature distance between two objects [90]. In this case, a greater distance translates to a greater dissimilarity between them. The L_2-norm, or the Euclidean distance, is one of the fastest measurements, where each feature vector of length N is pictured in an N-dimensional Euclidean space. The cosine similarity, which can be computed using an inner product between vectors, is commonly used in embeddings already normalized to unit norm. The earth mover's distance, or Wasserstein distance, is another metric that relies on variably sized histograms [91]. Many other algorithms for the purpose of measuring the similarity between features can be found in the literature, such as the signature quadratic form distance [92] and the Bhattacharyya distance [93]. The increased complexity of the feature similarity metric, although more expensive to compute in practice, can be beneficial if shown to provide superior results in evaluation methods.

3.3.5 CBIR in Medical Imaging

Content-based information retrieval, especially to the extent of image retrieval, holds great potential in medical applications. Not only can a CBMIR system determine the level of similarity with existent images, its combination with available meta-data can also provide support for CAD. Unfortunately, these solutions were not yet fully leveraged to the radiology practice, in which CBMIR still had little impact. The continued efforts in image processing, medical informatics and information retrieval are slowly creating conditions for the integration of content-based information retrieval in radiology workflows [80, 94].

Given the importance of assessing new CAD systems [95], some public initiatives emerged, providing data sets establishing a ground truth and/or specific tasks to solve. ImageCLEF is a series of benchmarking challenges, based on the CLEF campaign, which drives researchers to create solutions to a multitude of automated image understanding tasks [96]. Each participant of a challenge is provided a data set and a task to perform based entirely on a computer system's decision making techniques. The participants are then encouraged to implement one or more solutions to the task, evaluate the solution and submit the outcomes and conclusions. Evaluation measures exist for the purpose of assessing the performance of each submission. The Mean Average Precision (MAP) is a typical evaluation measure in these challenges, but many others may be requested, depending on the task involved.

Medical image classification and annotation challenges emerged in 2004 [97], serving as a catalyst for the development of high performance retrieval systems. Since the beginning of the ImageCLEFmed campaign, the data sets were enlarged to over 300 thousand images [96]. Many of the novel medical image retrieval techniques were presented as ImageCLEF submission reports. More recently, the ImageCLEF medical tasks have been focused on automatic concept detection, diagnosis, and question answering from medical images [98].

Furthermore, one of the most important trends in CBMIR is the combination of medical imaging with text content. This particular subject, although still under the scope of CBIR, will be covered in Chapter 4 due to its heavy importance in this thesis.

CBIR in Dicoogle

When thinking of a practical mindset towards these methods, CBIR offers practitioners an additional tool for a quick identification of similar medical studies. The integration of CBIR into clinical use comes with multiple challenges to be tackled individually, one of which is how a PACS archive can provide this form of searching. Dicoogle has previously been extended to support CBMIR using a profile-based approach [99]. This solution supports selective and automatic extraction and comparison of visual features depending on a chosen CBIR profile (Figure 3.2). When indexing medical images, Dicoogle will run the installed feature extraction methods and store feature sets associated to the new images in a Lucene database. An additional Graphical User Interface (GUI) was provided as a graphical extension to Dicoogle, thus enabling users to easily query the database by visual similarity.

The profile-based approach stems from the practical need to automatically (or manually) specify feature sets that are more suitable for the intended search. As an example, the positions and prevalence of micro-calcifications in a mammogram may provide a more useful similarity metric among a repository of mammography studies than the texture of the breast [99]. A general feature extraction profile is available for the purpose of providing query-by-example on any kind of medical image. For particular cases, a modality-specific profile can be selected, which will choose the appropriate feature extraction and similarity metrics for that image modality, as such, this was implemented as a Dicoogle CBIR plugin, which augments the PACS archive with a feature extraction indexer, a web service and an internal query API

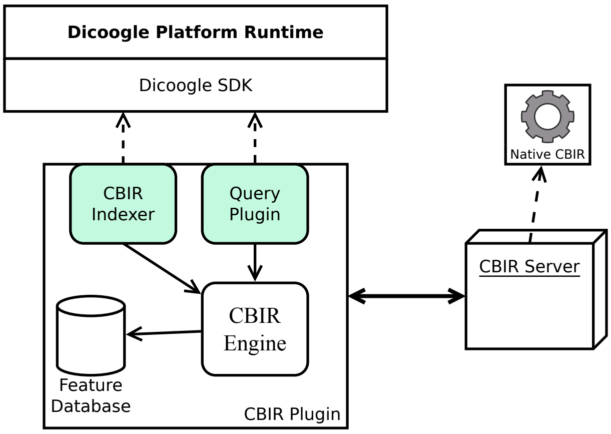


Figure 3.2: Diagram depicting the components of the CBIR and their interactions with the Dicoogle platform.

implementation for QBE.

Image processing systems in an academic context often show little concern for integration in other systems. Technological limitations may arise from integrating CBIR with other technologies: software bindings for a target programming language to native computer vision or machine learning libraries may not exist or have a multitude of issues constraining their use. Furthermore, image processing applications developed for a native target, usually made in C or C++, may contain bugs that are hard to detect and could compromise the entire process in case of a crash.

We have experienced technology fatigue while working with the Java binding of OpenCV [100], which were used for the development of the original CBIR system. This was overcome with the creation of a server component which extracts features using the native interface of OpenCV. These feature extractors were then exposed over the network, becoming accessible in a technology-independent fashion, including the CBIR extension presented. This allowed more existing feature extraction procedures to be ported to the system, as not all OpenCV functions were available in the Java bindings at the time. Later on, the concept was expanded with feature specification, key-point descriptor extraction and classification components. This set of specifications and components were placed under the umbrella term *MIRbiome* (word play for Multimedia Information Retrieval biome).

3.4 Final Remarks

The topics described in this chapter are a necessary foundation to the contributions presented in the following chapters. The ideas to take away from these sections may be summarized in the following form:

Traditional information retrieval seeks to answer queries by collecting the most relevant documents from a set of data. In a PACS archive, this involves identifying images or clinical studies that are more relevant to a particular case, thus rewarding the user with an automatic second opinion with potentially useful insights. Although the DICOM standard provides a common way to search for medical images in a compliant archive, their querying capabilities are too restricted when considering recent searching paradigms (namely free text searching and QBE).

With Content-based Image Retrieval, the same user would be able to search for similar studies without formulating a text based query. However, the archive will usually contain information in heterogeneous forms: in the case of most medical imaging modalities, this is keyword-based meta-data containing textual information, and pixel data. Taking a PACS archive's search engine into a multimodal mindset is, as revealed next, a crucial stage forward in the CBIR framework and its sound application in healthcare.

Chapter 4

Multimodal Medical Information Retrieval

The use of CBIR alone in medical imaging can only go so far: when a medical doctor performs a search on a system during their work, the original intent is to obtain information about a clinical case, rather than just an image [101]. The combination of medical image feature sets with non-visual data, such as DICOM meta-data and structured reports, can provide complementary information and increase the accuracy of the overall decision making process [102]. At the moment, it is proven to be highly beneficial for medical decision support systems [80].

For these reasons, visual feature extraction and comparison is already contemplated by many CBMIR systems (including Dicoogle's CBIR extension), but only stands as one of the pillars of a multimedia retrieval system. Therefore, studying the concept of multimodality in medical imaging informatics, as well as new and better ways to employ it, is a great step towards this goal. This chapter provides an overview of techniques for multimodal information retrieval, including their application in medical imaging systems, and proposes a new multimodal search engine for PACS archives.

4.1 Definition

The definition of multimodality in the domain of medical imaging informatics can become confusing, due to its notable similarity with the concept of medical imaging acquisition modality. The latter, described in Section 2.1, refers to the technologies, formats and hardware used for capturing images of organs and tissues. Hence, *multimodality* in medical imaging has been mentioned in some literature as the acquisition, combination and registration of images from multiple acquisition modalities [103, 18], in which they may be better exploited for more relevant information. For instance, Positron Emission Tomography (PET) scans can be combined with CT scans, with a procedure named *image registration*, to achieve the PET-CT multi-modality, thus easing diagnosis. An overview and discussion of these techniques can be found in [18].

On the other hand, multimodality in the context of information retrieval refers to the theories, algorithms, systems and challenges of indexing and searching for multiple modes (kinds) of data, which may include content from meta-data, free text, images or other multimedia sources [104, 105, 106]. This thesis grabs on to this definition, applied to the available information in medical imaging repositories, very often narrowed down to medical images and textual annotations. In [107], this definition is regarded as the enhancement of retrieval using non-image data.

In order to prevent ambiguities throughout the document, **multimodality** shall refer to multiple modes / kinds of data, thus considering a superset of using medical images of any acquisition modalities, as well as meta-data, medical reports and other kinds of annotations. The other definition, as in [18], is rephrased to refer to multiple acquisition modalities, or use thereof, when such a distinction is required.

4.2 Trending Techniques and Concepts

Quite often, the architecture for multimodal information retrieval in the scope of medical imaging informatics is tightly coupled with the goals of CBIR: techniques for query refinement, expansion and combination are usually presented as part of the image retrieval engine [108, 109]. On the other hand, some approaches to CBIR are exclusively dependent on the existence of multiple modalities. This section complements the previous topic of CBIR, exposing some of the currently existing approaches and solutions to multimodal content-based information retrieval in medical imaging informatics. A great part of these techniques were first exhibited as part of benchmarking initiatives. ImageCLEF contains several medical image retrieval tasks where both visual and textual features are involved [97]. The VISual Concept Extraction Challenge in RAdioLogy (VISCERAL[1]) contains a *Retrieval* challenge specifically designed for the evaluation of multimodal information retrieval techniques in medical imaging. A review of the latest evaluation tasks and results from the VISCERAL 2015 Retrieval benchmark can be found in [110]. The Retrieval 2 benchmark, created shortly after, is a continuously running benchmark with the same challenges and goals.

There are search engines for multimedia content retrieval that do not perform CBVIR over images, thus relying solely on annotated content. GoldMiner[2] [111] is one of such search engines, designed for retrieving medical images obtained from peer-reviewed journals. It boast a better performance than generic search engines for the task of radiology image search, and creates semantic relations between search terms and concepts assigned to images generated from their captions. BioText [112] is similar in concept, providing access to scientific literature

---

[1] `http://www.visceral.eu`

[2] `http://goldminer.arrs.org`

on biology. They can only function as expected with the annotated content, and do not attempt to measure similarities between images based on their actual content. A multimodal approach, by definition, would also require the interpretation of additional media content for retrieval.

4.2.1 Multi-Query Combination

The topic of query fusion is one of the foundations for multimodal information retrieval, as it involves, at some step of the process, fusing (merging) content originated from multiple modalities. Where this step takes place in the retrieval procedure can significantly affect the design of the database, as well as the outcomes of each search. In ImageCLEF, more than a hundred submissions from 2003 to 2009 attempted to combine text-based with image-based retrieval. Typically, query fusion may be classified as *early* fusion or *late* fusion, although hybrid versions exist [113].

In **early fusion**, each feature vector associated to a unimodal query is merged together into a unique feature space, regardless of the kind of query involved (both textual and image features reside in the same vector or data model). The resulting model is only then used to query the system, thus relying on a single decision rule over all sources of information. The major drawback involved is having to face the *curse of dimensionality*, where the broad dimension of the heterogeneous feature space is usually very sparse, which leads to a generalization problem. Therefore, the challenges inherited from using early fusion lies in making a data fusion model with an appropriate feature weighting mechanism. Integrating both textual and visual contents in a document-term matrix may suffice, and potentially provide good performance with simple visual terms, as in [114]. More complex approaches exist, such as those based on probabilistic Latent Semantic Analysis (pLSA) [115]. This model was used for visual vocabulary pruning in [116], but an extended version of pLSA was developed for the actual fusion of visual and textual terms [117].

Late fusion involves applying an algorithm over multiple ranked lists of results, with the aim of obtaining a single stream of greater performance. This is a particular case of the *rank fusion* concept, in which individual ranking preferences of several judging entities are given a "concensus" [118]. They are, in general, the most preferred and utilized in benchmarks for the past few years, and some diversity of algorithms in this scope have emerged. In [79], late fusion is defined as a task for obtaining combination rules across multiple decision streams, using a certain amount of data with ground truth as a validation set. However, this technique is preferably named *fusion learning*. Although such definition represents a fair suggestion to rely on machine learning techniques, the concept itself does not demand such an approach. Late fusion algorithms may be as simple as a boolean operator applying a restriction to the results effectively contained on the merged result list (`AND`, `OR`, `LEFT`, `RIGHT`) and/or a reordering algorithm for that list, for each unimodal query.

Result ordering algorithms may be based on the score of each result entry (representing the system's individual appreciation of relevance for that result), or be based on the rank of said result document on the full list [119–121]. Some algorithms however, may rely on concepts from both rank-based and score-based fusion techniques, such as the Inverted Squared Rank (ISR) [122].

In score-based fusion, most techniques are based on a linear combination of results. CombMAX and CombMIN discard all score values but the highest or the lowest, respectively. These strategies attempted, on the one hand, to minimize the probability that relevant documents would become poorly ranked (in CombMAX), and on the other hand, preventing poor documents from being highly ranked (in CombMIN). This approach is inherently flawed, since only one or the other can be applied [120]. However, they may still perform well in specific cases, and other algorithms may attempt to combine the two [123]. CombSUM makes a summation of all similarity values, thus giving each query the same weight. CombMNZ applies a greater weight to results retrieved by more modalities, by multiplying the score summation with the number of sub-queries that contemplate each document.

For score-based techniques to be useful, a normalization of all scores should take place, in order to balance the importance of documents obtained over the existing range of modalities [124], which may not be commensurable and have distinct rank-similarity curves [119]. This procedure is not required in rank-based fusion. For instance, the Reciprocal Rank Fusion (RRF) algorithm, in spite of being simple, can yield equal or better results when compared to a few other methods, without requiring similarity scores or a normalization process [125].

Some medical imaging search systems have (fully or partially) adopted multi-query techniques. In fact, a survey performed in 2012 suggests that a perfect medical imaging search system would let a user combine visual and textual information [126]. For instance, the IRMA project's medical image retrieval engine supports intersections and conjunctions of results with those previously performed by the user, by keeping a tree-structured history of all queries [108]. A multimodal medical search engine to retrieve information from medical articles was presented in [122], supporting a multi-query combination of text and a set of images. The multimodal search engine for histopathology case retrieval by Jimenez-del-Toro *et al.* [127] also employs late fusion of text-based and image-based queries.

4.2.2 Query Refinement

Query refinement, often found in generic information retrieval systems, is a kind of search technique for providing the means of tweaking a query in order to achieve higher quality results. In a purely manual refinement, the user will rewrite the query and try again, often using the information retrieved with the previous search, with hope that the new query will bring the user more pleasing results. In an automatic refinement, the search engine may offer suggestions or directly modify the query during the process.

Query expansion is a particular kind of refinement often mentioned as a type of fusion [113], in which a query is modified with new search terms or, for the case of multimodal queries, additional media content. In a rule-based approach, some query terms may be translated to their synonyms. In addition, a query performed in one modality may be used to expand another one of a different modality, thus performing what is currently named as *inter-media feedback*. The most common form of inter-media feedback is textual query expansion by inclusion of annotated content from an image-based retrieval. Image-based query expansion, although less common, is also possible by adding features correlated to a text query's most relevant terms [128].

Relevance feedback is an information retrieval process in which a user performs a sequence of queries towards a desired set of results by "tuning" the query in each step. A system with automatic (or implicit) relevance feedback performs log analysis over a user's behavior [129]. Manual (or explicit) relevance feedback relies on the user's immediate feedback on the relevant information. For instance, the user may be able to select results deemed relevant to the search, or those that are irrelevant. In medical image retrieval, relevance feedback is known to be a powerful technique that can significantly improve performance. In the IRMA project, relevance feedback is achieved per result entry with a slider bar, thus providing input for a continuous numerical value [108].

Relevance feedback has also been applied in multimodal medical information retrieval. In the context of the ImageCLEF 2012 ad-hoc image retrieval task [130], Markonis, Schaer, and Müller [131] constructed a mechanism simulating multiple iterations of perfect relevance feedback (which means that the user detects all relevant images on each step, on top of a retrieval engine combining late fusion of text and visual retrieval. The results show that having a sufficiently large number of result entries k for the user to check for relevance will significantly improve retrieval metrics after within the first four iterations (MAP 0.349 with $k = 100$, against ≈ 0.165 with no relevance feedback). The improvements that came with the inclusion of visual retrieval are significantly lower: a plain text-based approach with the same parameters would yield a MAP of 0.335.

4.2.3 Other Multimedia Content

Although audio and video content may be present in medical applications, images constitute a much greater part of all studies in an institution, thus why the study of visual information retrieval is more relevant in the field, and has been addressed more often during the last decades. Nevertheless, research on systems performing multimodal information retrieval on other sorts of media has been done. AALIM [132] is one such system, which also performs data mining on image, video and audio patient data for cardiac decision support.

4.3 Multimodal Search Engine Proposal

As part of one of the contributions of this doctorate, Dicoogle was extended with a new engine in order to provide multimodal search capabilities. This extension aims to:

1. Create an interoperability layer among different sources and information modalities in the Dicoogle environment, namely text-based and image-based query providers, as well as potentially other information modalities in the future.
2. Integrate state-of-the-art query fusion techniques and leverage the potential of Dicoogle's CBIR support and its profile-based approach to be put both in the clinical practice and image retrieval benchmarking scenarios.
3. Exploit a flexible and usable search user interface relying on well regarded paradigms in the field, such as *query-by-example*, *relevance feedback* and *query expansion*.

4.3.1 Related Work

Quite often, the architecture for multimodal information retrieval in the scope of medical imaging informatics is tightly coupled with the goals of CBIR: techniques for query refinement, expansion and combination are usually presented as part of the image retrieval engine, which also contemplate text-based retrieval in some cases. Multimodal information retrieval has had its impact in a multitude of fields, and several tools and techniques for CBMIR have emerged over the last two decades [80], [133]. In [117], the authors cover the state-of-the-arton multimodal medical information retrieval in three perspectives, one of which is the latest research done in CBMIR.

The NovaMedSearch engine [134] exhibits the similar goals of supporting multimodal queries with a simple and intuitive user interface, for medical case-based retrieval. Our work, in contrast, is not tightly coupled to specific sources of data and shows a greater concern of integrating the engine to a PACS. The Khresmoi project also stands out. It is a large EU-funded project with the goal of conceiving a multi-lingual and multimodal search and access system for biomedical information [135]. The main user interface is based on the ezDL project, but an alternate interface was developed, called Shambala [136]. Markonis *et al.* [137] have covered the use of Khresmoi for the retrieval of medical images in a PACS archive and the biomedical literature. The search engines developed under this project are backed by ParaDISE [138], a CBIR system featuring an architecture with scalability and extensibility in mind. Rahman *et al.* [139] also present an interesting multimodal framework with an embedded hierarchical image classifier and a fixed pipeline of fusion strategies for medical image retrieval. It was our intention in this new architecture to be as flexible as possible in the kinds of queries that can be created, by supporting query trees of arbitrary depth, configurable transformation and fusion strategies, and the possibility of including classifiers as a dedicated source of data, which do not have to rely on the system's extracted features.

4.3.2 Architecture

Figure 4.1 presents a top-level view of the proposed multimodality search plugin and its interactions with the Dicoogle runtime framework. It was developed as a plugin that does not contain feature extraction, similarity measures, or direct means of querying a database. Such tasks are delegated to existing query providers by categorization of their modality.

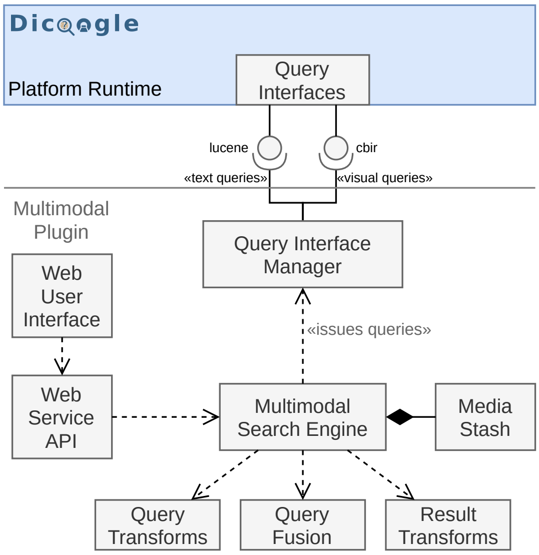


Figure 4.1: Architecture diagram of the multimodal search engine extension, depicting its key interactions with Dicoogle.

The multimodal retrieval engine requires access to two modality interfaces of the Dicoogle core platform:

- *Textual meta-data*: Typical text queries are fundamental to support the DICOM query plugin. This interface is based on Lucene and it was decided to keep this option at the multimodal level, by relying on Lucene's query format as the text query interface, which is sufficiently flexible for most text-based searching tasks.
- *Image content* (CBIR): image queries provide either a Universal Resource Indicator (URI) of an already indexed image, or an object containing the image proper. These queries can optionally be followed by the name of a CBIR profile, thus focusing on a particular group of features and measures.

These categories provide a thin layer of portability among plugins that accept the same kind of query object and follow the same API.

The two entry points for multimodal queries are the RESTful API (Section 4.3.5) and the user interface (Section 4.3.6), which are both web-based. At the back-end, the *multimodal search engine* processes the queries and makes calls to the Dicoogle query providers. The *query interface manager* contains query adapters for this purpose, which will be provided according to the kind of query requested by the engine.

4.3.3 Query Formulation and Processing Workflow

The major difference between a simple text query and a multimodal one is that it may contain information infeasible or too expensive to be fully represented in a textual format. If the user wants to perform a query-by-example over a local file, this object needs to be uploaded before or alongside the remaining description of the query. In order to achieve this, the multimodality platform contemplates a *media object stash*. It is a container for temporary multimedia content, which is tagged with a unique identifier (`uid`). With this approach, multimedia files not already indexed by Dicoogle are transferred before the effective query descriptor is sent (Section 4.3.5). The object will later on be retrieved by `uid`, and have its feature set extracted and processed by the CBIR module.

Once all required media content is stored, the multimodal search takes place following the pipeline described next (Figure 4.2):

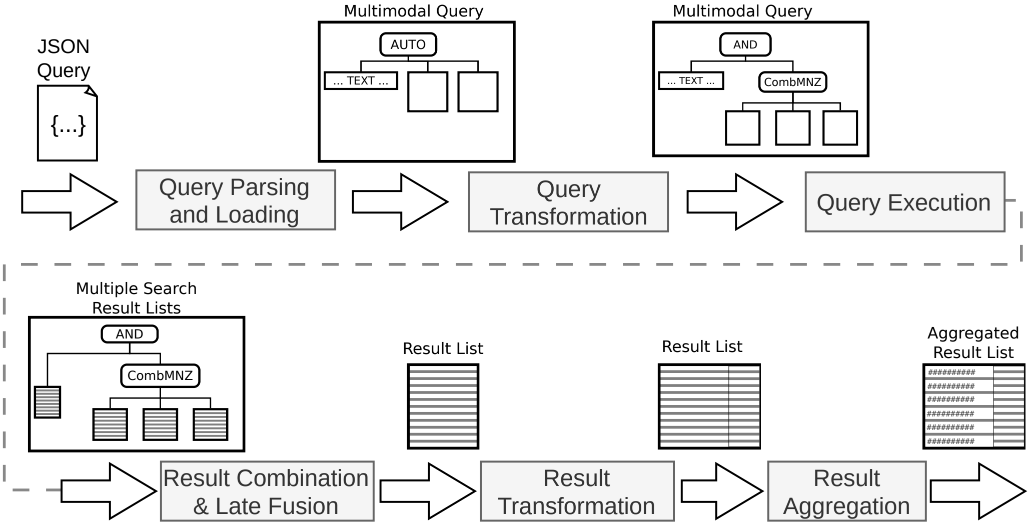


Figure 4.2: Diagram depicting the multimodal search pipeline

1. The user will formulate and send a query using the available web-based user interface. External systems may also construct and perform queries using the REST API.
2. The engine will pre-process the query by traversing it through a fixed series of query transformation functions. This step is where query expansion and score normalization

techniques are applied.

3. The multimodal query will be split into unimodal queries that will be invoked on the Dicoogle core runtime, by adapting it to one or more query providers with the same interface. The operation will yield multiple result streams that must be properly merged.
4. The result lists are combined into a single list using late query fusion techniques, which are detailed in Section 4.3.4.
5. Before returning the final outcome, the results may undertake a series of transformations. This may include augmenting the results from the CBIR engine with additional useful fields not previously contemplated, such as the attributes contained in the archive's DICOM meta-data source.

This pipeline is established at Dicoogle deployment time, and can be extended with more query / result transformers and fusion strategies. Moreover, the platform can be configured to abort the pipeline process if an error occurs in one of the steps, or simply ignore the failing part.

4.3.4 Multi-Query and Fusion Techniques

This system features multimodal queries based on a combination of multiple text and image Content Objects (COs) forming a tree structure of queries. A "leaf" is a unimodal query, which can be handled by the Dicoogle core runtime through an adaptation procedure in the *query interface manager* (Figure 4.1). Any other tree node represents a query fusion process. For identification purposes, each node contains a *content object key* (CO key) property, which functions as an index for that query fusion's child nodes. In order to uniquely identify a node in the multimodal query, a concatenation of CO keys is made, comprising the path from the root to the intended node (e.g. "`1.2`" is the third query of a query fusion, which is the second query of the top-level query fusion).

A few known late fusion algorithms are also contemplated. Boolean combinations intersect (*and*) or union (*or*) a list of results without concerning about the ordering of results. CombSUM (eq. 4.1), CombMNZ (eq. 4.2), CombMAX (eq. 4.3) and CombMIN (eq. 4.4) were also added to the initial foster of query fusion strategies, as described in [113]. d is the document of a result, N_j is the number of sub-queries performed in the fusion, $S_j(d)$ is the score of d in the sub-query j and $F(d)$ the number of occurrences in the sub-queries:

$$S(d) = \frac{1}{\sum_{j=1}^{N_j} {}^{1}/_{S_j(d)}}, \tag{4.1}$$

$$S(d) = \frac{1}{\sum_{j=1}^{N_j} \left({}^{1}/_{S_j(d)} \right) \times F(d)}, \tag{4.2}$$

$$S(d) = \arg \min_{j=1:N_j} S_j(d), \tag{4.3}$$

$$S(d) = \arg \max_{j=1:N_j} S_j(d). \tag{4.4}$$

Each result list may yield score values that are inadequate for comparison among different queries, since they may follow disparate score distributions and ranges [140]. Therefore, each result list needs to be normalized before a score-based fusion between other lists takes place. Rather than having a single implementation, the platform allows a client to select one out of multiple score normalization strategies. The algorithms currently included are min-max (proposed in [124]), min-sum and min-var (the last two proposed in [140]). This "freedom of choice" was deemed relevant due to the fact that some score normalization algorithms offer more robustness against outliers, thus increasing performance when fused by strategies that are particularly sensitive to them [140].

The search results of CBIR queries in Dicoogle have a distance-based score. That is, the value 0 represents the highest score possible, whereas higher values relate to greater dissimilarity or irrelevance among objects. This irregularity is addressed by automatically converting distance-based scores to a non-negative "higher is more relevant" range before each normalization.

4.3.5 API and Data Representation

A multimodal query representation format was specified as part of this proposal. It was designed to be simple and easy to use by web-based applications, have a low memory footprint, and support some degree of extensibility. Other means of describing multimedia queries are already available but were unsuitable for the given requirements. The Multimedia Retrieval Markup Language (MRML) [141] is a standard defining an XML-based communication protocol for performing queries to compliant multimedia retrieval systems. Although an implementation exists and the format could be extended to support multimodal queries, the protocol is unsuitable for the web, the official website[3] is no longer available at the time of writing, and the standard has not had any significant impact during the last years. Therefore, making an additional effort to make the system MRML compliant did not seem to be worthwhile. For this work, a data schema in JavaScript Object Notation (JSON) was employed for the complete description of multimodal queries. JSON has the advantage of producing files with less overhead and facilitating query construction and parsing. This is especially useful in web applications, the runtime environment of which have built-in JSON support. Other systems can also easily read and write queries with the aid of JSON libraries.

As expressed, the proposed system composes a set of web services providing four main

---

[3] `http://www.mrml.net`, currently unavailable

resource endpoints (relative to Dicoogle's base Universal Resource Location (URL) for web services):

- `/multimodal/search` is the endpoint for performing queries. A query JSON object of the query is uploaded with a POST operation, which will follow with a response containing the outcome of the search.
- `/multimodal/stash` provides the means to store media objects for use in future queries. A store operation will accept either a media content (e.g., of Multipurpose Internet Mail Extensions (MIME) type `image/png`, `application/dicom`, . . .) or a multi-part data form containing the same item (MIME type `multipart/form-data`). This content type was added in order to support file uploads purely based on an Internet browser's implementation of Hyper-Text Markup Language (HTML) version 5.
- `/multimodal/ui` is used to retrieve the user interface and will be consumed by an Internet browser.
- `/multimodal/fusion` simply returns a list of query fusion strategies made available, as a JSON array of (value,label) pairs.

4.3.6 Graphical User Interface

The system also provides a new graphical user interface (Figure 4.3) to exploit and use query fusion techniques made available with the plugin's search engine. As a key concept of interaction, each unimodal query in the multimodal query tree is represented in a box. *Ghost boxes*, which are empty and will not take part in the search, are shown to allow the user to introduce more queries in the tree.

The drag-and-drop interaction paradigm was significantly exploited for this interface. An image from a previous result can be dragged and dropped over a query object box in order to become part of the query. If the box was "ghosted", a new child query is contemplated, and another ghost box is placed next to it. Text queries can still be performed by typing on a text input box. If the fore-mentioned query box already contained an image, the text input will instead provide the unimodal query's meta-options. In a multi-query fusion, the user can choose one of the available fusion strategies from a drop-down list, or leave the "Automatic" option selected. Once issued, the results are shown as a grid of images, all of which can be dragged and dropped for a manual form of relevance feedback. Not only a grid layout of results is more fitting for drag-and-drop operations, but it is also known that some radiologists prefer this layout to a list of items [142].

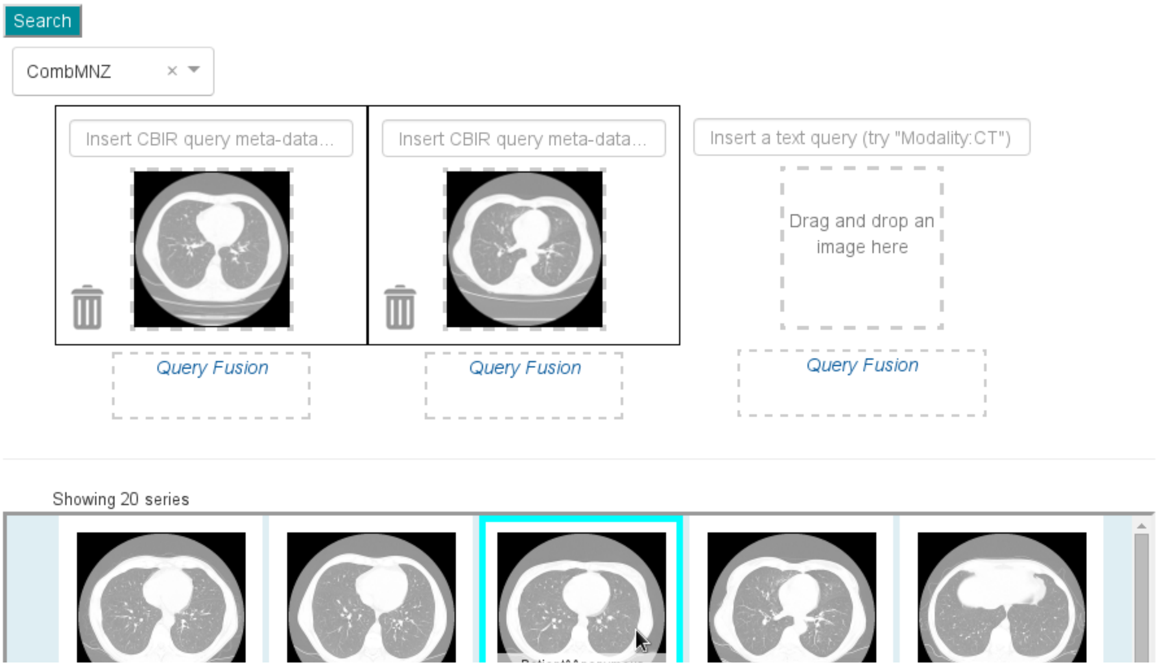

Search
CombMNZ
Insert CBIR query meta-data...
Insert CBIR query meta-data...
Insert a text query (try "Modality:CT")
Drag and drop an image here
Query Fusion
Query Fusion
Query Fusion
Showing 20 series

Figure 4.3: A draft of the multimodal search engine's graphical user interface, depicting a multi-level query, ghost query boxes and a few results from a previous search.

4.4 Computational Assessment and Results

The methods described in this document allow us to capture the immense heterogeneity of medical image archives by supporting content discovery services on top of multiple data formats. More specifically, they enable the combination of multiple providers into a single search interface where users may search for interesting artifacts using a unified query language. However, featuring such functionality alone is not enough. It is of major importance that our services are offered in a performant manner according to the requirements of the medical imaging environment, which are known to be demanding. The following section presents a series of trials devised to ensure the proposed architecture's computational performance in real world medical institutions.

In this picture, scalability raises a concept of major importance. In computer science, scalability refers to a system's ability to maintain its performance indicators with increasing levels of load. The performance indicators of an interactive system, such as ours, reflect the throughput of successful requests it can handle, in our case, search requests. On the other hand, the load factors reflect the number of concurrent requests, which is an indicator of how many users are using the system at the same time, as well as their complexity. In a scalable system, the rate of degradation of the performance indicators with the increasing amount of load would be as close to zero as possible. In such a utopia, the system would be capable of handling an infinity of users simultaneously.

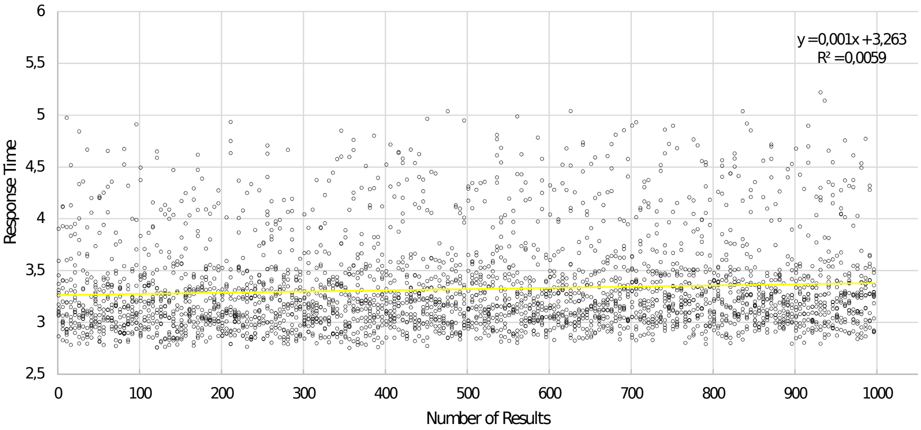


Figure 4.4: Scatter plot representing the engine's response time, in seconds, in terms of the number of results obtained from the query.

Taking this formal definition into consideration, we devised three experiments in order to understand the degree of scalability of the proposed system. Firstly, we wanted to capture how the system responds to the complexity of the search operations. As a result, the first experiment relates the number of returned results with the time necessary to handle the search task. On the same note, the second experiment relates the latency of the search task when it is used the fusion operator. The last experiment was designed to analyze the system response in simultaneous tasks processing. The experiments were conducted using an Intel® Core^{TM} i7-3770 CPU @ 3.40GHz × 8 with 12 GiB to run the Dicoogle instance. The data set used for these tests was retrieved from the clinical case archive of the Belarus Tuberculosis Portal[4] [143]. Four hundred one clinical cases were indexed, consisting of 62,198 medical images.

The first experience involved using a fixed query composed of two images and a meta-data search for the keyword "CHEST." The two images were examples of pulmonary CTs. The number of results requested in the search procedure (as configurable by the web service) varied between 1 and 1000. This query was designed merely for experimental purposes: although it makes little sense to search for 1000 related artifacts, these tests are bound to a hypothesis where this solution will scale by the number of results. In total, 3000 search operations were collected. The resulting distribution of the results can be analyzed in the scatter-plot in Figure 4.4. All time measurements are in seconds unless otherwise specified.

Empirically, it is perceptible that the increasing number of returned results has little effect on the search service time. Nevertheless, we computed a linear regression using the least squares method. The results (≈ 0.001 s) confirm a very slight variation rate, meaning that the number of returned artifacts has very little impact on the system's search response time. More concretely, it is expected an aggravation of 1 ms in the search time per retrieved result.

---

[4] `https://www.tuberculosis.by`

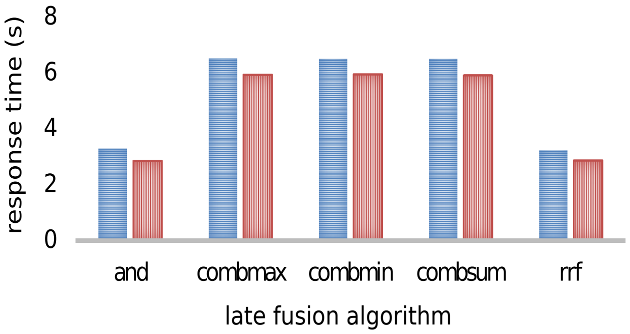


Figure 4.5: Chart exhibiting the engine's response time (seconds) over fusion queries with different late fusion algorithms.

In practice, this raises no concerns on its own regarding the system's scalability. What may appear to be large response times in general are a consequence of the naive visual feature index implemented for CBIR in Dicoogle, which is inefficient for a full data set traversal at the time of writing [99].

The second experiment, the results of which are presented in Figure 4.5, introduced a performance comparison of the fusion operators described in this document. The goal was to discover if any of the proposed operators were impractical in a real world environment. We tested four fusion algorithms, which were best applicable to our queries. The performed queries were a combination between a meta-data query and an image query, the latter of which were filtered by an additional image query to refine the results. As expected, neither of the fusion operators proved to be impractical to use; however, we noticed that the CombSUM, CombMAX, and CombMIN operators took considerably more time than the others. This is an expected behavior, as these strategies require the scores of all considered search results to be normalized, as explained in Section 4.3.4. Both the intersection operator (*and*) and RRF do not require score normalization.

The last experiment is actually more interesting, as it evaluates the system performance in a multi-user environment. Initially, we asked an experienced user to perform six search operations over the data set, involving late fusion operator and without any restriction of complexity. We recorded not only the queries inserted, but also the idle time spent by the user between each query. This capture was assumed as reasonably demonstrative of a regular usage pattern of the system. Afterwards, a query simulator was developed according to the user's searching patterns recorded. The program performs the searches as a regular user, with a variable delay between queries modeled by a normal distribution (avg=10 s, std=3 s). The experiment consisted in running multiple instances of this program simultaneously and independently from each other, thus mimicking a regular usage of the system with a variable, but controlled, number of users. We tested up to nine users simultaneously, as we think that it is a fairly high number of concurrent users in a PACS of a central hospital.

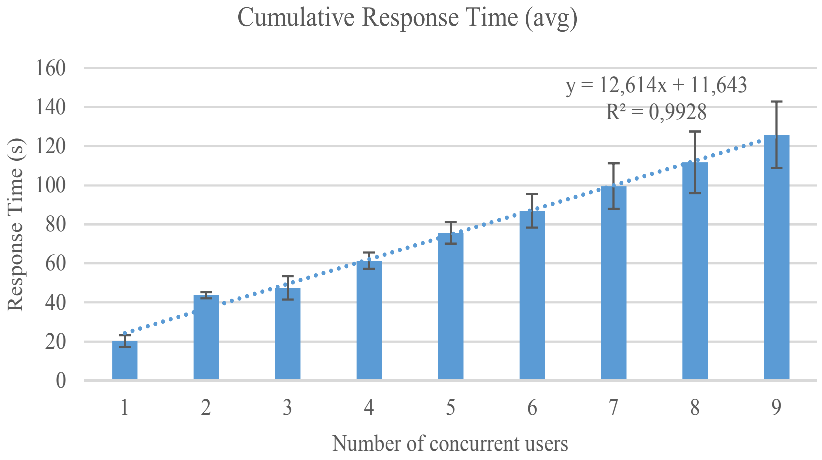


Figure 4.6: Average cumulative response time measured in the third experiment.

Figure 4.6 shows the average cumulative response time of all searches for each user in the multiple test cases. As it is perceivable, the escalation of the cumulative response time with the increasing number of concurrent requests is best fit by a linear regression. As opposed to an exponential fitting function, a linear regression ensures that the proposed methods are easily scalable to a multitude of users, provided that adequate hardware is used. The regression also shows that a penalty of 12.5 s per concurrent request is to be expected.

4.5 Final Remarks

In this chapter, an architecture for multimodal information retrieval was proposed, with the main objective of being usable in real world PACS scenarios, including real time search operations over medical imaging repositories and DICOM compliant archives. Its backbone was designed to be extensible, supporting new algorithms without major changes to the software, thus providing a multimodal layer of abstraction over the large domain of existing retrieval algorithms, such as feature extractors and model representations. At this level, such techniques will not be incurred a cost in retrieval quality when integrated with the platform. Rather, it allows researchers to discover improvements by combining multiple sources. A proof of concept was built as an extension to Dicoogle, although the decoupling of the architecture from this system can be done by relying on other query provider manager implementations. The performance and scalability of this architecture were evaluated, and the results demonstrate that the proposed solution can be used in real world environments.

Certain improvements regarding the kind-based query provider interfaces can and should be considered for an integration to clinical workflows to be considered.

- Relevance feedback, as a means to disambiguate the user's perception of relevance, would become an indispensable process [144]. Being able to mark a query image as relevant or irrelevant, and even to outline regions of interest in an image, can be translated into intuitive user interactions for boosting the results of a search in the clinical routine [131].

- How one would build a single query from a series of instances, when the system only acknowledges instances as independently indexed items, requires an additional layer of complexity. A 3D CT volume, as an example, may be seen as a list of instance-level queries to be feature-fused, using a feature fusion technique (for instance, Rocchio's algorithm [145]). On the other hand, this strategy does not scale to the depth of a volume, and can include slices without relevant content into the query. In order to overcome these difficulties, additional instance filtering processes may be adopted, on both queries and results, based on an automated detection of visual cues and anatomic regions. A form of search result enhancement is presented in Chapter 5.
- Furthermore, these interfaces could be expanded to also return suggestions of query expansions and auto-completion when such an expansion is possible, by relying on indexed DICOM tags, RadLeX terms, or CBIR profile names, to name a few. A complete integration of these sources would involve leveraging adequate query specification constructs into their interfaces, for use in the multimodal search engine, and are posed as technically solvable challenges.

Chapter 5

Automated Labeling for Content Discovery

A DICOM object contains some useful attributes, such as patient demographics and radiation exposure. On the other hand, there are certain search terms that the practitioner or researcher would find useful, but may not yield any search results. For instance, information about the presence of certain organs or anatomical regions in a particular image or slice can be limited, vague, or non-existent in some archives.

The concern of usefulness of DICOM meta-data stems from the high degree of freedom found in the content of some DICOM fields. As an example, it is not uncommon for the fields *Study Description* and *Series Description* to contain relevant content regarding the purpose of the imaging study and the anatomic region being considered, although this is critically dependent on the quality of the modalities, as well as the radiologist's effort for coherence and completeness. The DICOM attribute *Body Part Examined* may exhibit different levels of granularity for images of the same body parts, while still conforming to the standard. This particular field is defined at series-level, and is not multi-valued, meaning that only the main part intended for examination is recorded. As such, this attribute would not convey any information about anatomic regions in a whole body scan. Although certain guidelines may be suggested when defining these fields, they are often not enforced, and even so, could hardly be applicable in the scope of more than one institution. In the scope of a single hospital, these tags still may not be sufficiently accurate for automatic anatomic categorization [146]. The DICOM tag *Anatomic Region Sequence*, which aims for a better indication of anatomic regions in a study, may be used for keeping precise anatomic information, albeit only applicable to a small set of modalities. While the DICOM standard may be capable of specifying anatomic regions for most use cases, obtaining that information still requires an analysis that is out of the scope of imaging modality machines.

At the opposite spectrum of existing information in a PACS archive, there is Content-based Image Retrieval (CBIR), previously covered in Chapter 3, which enables a purely visual approach to search, and does not require manually annotated content. However, the latest research on CBMIR suggests that the use of visual information alone is less likely to exhibit as much performance as the use of textual data [80, 144], since the latter conveys a much shorter semantic gap. That is, a keyword such as *"pancreas"* can be unambiguously bound to a semantic concept, whereas there is no direct link from low-level visual features to such a concept. Manual annotation or proper image recognition mechanisms are required to establish this link. Nevertheless, textual features can be combined with visual features in order to improve the performance of retrieval.

It is with a multimodal perspective of a medical imaging repository that an automated content discovery mechanism is considered beneficial: the automatic detection of complex patterns in visual content introduces new concepts which were previously unavailable for medical image retrieval. In light of these ideas, this chapter is dedicated to the subject of automated object recognition in medical images, followed by the design of an architecture for information enrichment in a standard PACS archive through automatic extraction and indexing of visual information. The presented solution, albeit sufficiently generic to be applied in other open source archives, was implemented as an extension to Dicoogle, and aims to be used in production and research environments. The solution includes a new classification database system, along with a specification for classifier integration, thus enabling flexible queries over the indexed data. The final outcome is a system with substantially enhanced medical image search capabilities than a typical PACS repository, by combining and orchestrating automated content discovery with multimodal querying.

5.1 Related Work

Classification of medical imaging is an important concept already applied in Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx) systems (both often seen under the more generic acronym CAD), and a multitude of automatic and semi-automatic techniques can be found in the literature [147]. When CAD systems are available, integration with the local PACS archive is mutually useful: the archive can provide medical images and meta-data to the CAD components for analysis, whereas the latter parts can return the respective outcomes for indexing in the PACS databases, making the information accessible from a single entry point. This integration is not trivial, but great efforts have been made from a technological point of view [148].

A traditional CAD system would segment and classify body structures by creating algorithms that reflect a proper understanding of the human anatomy. For example, in [149] the authors aim for coarse automatic classification of CT scans by making a specific algorithm based on known statistics of a CT volume's density features. As another example, the work in [150] includes a series of image processing steps, such as thresholding, region growing and contour detection, for the 3D delimitation of multiple parts of a torso, including the separation of skin, muscle and fat. Fully automated subdivision of the human cerebral cortex in Magnetic Resonance (MR) scans is also known to be valid and practical [151].

The use of deep neural networks has aroused significant interest in the scientific community, as they have demonstrated state-of-the-art performance in image recognition. Unlike traditional techniques, which rely on hand-crafted and human-refined feature extraction algorithms, these networks learn visual feature extractors based on the given input domain, leading to a better detection of relevant visual nuances. The growing application of deep learning in medical imaging is no exception in this regard, to the point of becoming a prevalent, and often the primary, method of choice for image recognition tasks [152].

Roth *et al.* [153] have tackled the very similar challenge of recognizing anatomical regions using CNNs, in which specific image transformation methods were added to augment the training set. The target regions were coarse-grained (pelvis, neck, lungs, . . .), which allowed the authors to build a data set using only manually corrected DICOM meta-data for the labels on each data point. Since modalities do not usually identify fine-grained organs (such as *spleen* or *thyroid gland*), a reliable data set for these cases must be built through other means. In [154], the authors have explored the use of CNNs for medical image classification by modality, using an ensemble of classifiers and fine-tuned models. In addition, CNNs have also allowed for an almost perfect classification of frontal and lateral radiographs [155]. A fully automatic segmentation is also possible with CNNs models. In the biomedical domain, the U-net was proposed [156]. In [157], CNNs were used for segmentation of fat areas in CT scans. A more extensive survey of deep learning in medical image analysis can be found in [152].

Previous projects have identified benefits in content-based enrichment of information retrieval systems in medical imaging. In the Information Retrieval in Medical Applications (IRMA) project, images are classified into a tree of semantic categories described as a sequence of codes [158]. This hierarchical categorization concept was shown to provide a high amount of content understanding and fulfill real-world medical image retrieval requirements. In the Khresmoi project, 3D volumes are mapped to an atlas of the human anatomy, and landmarks in the volume are automatically localized for each volume, with the goal of providing useful information to radiologists before a diagnosis or observation is made [137]. The multimedia system EIR aims to be a reliable CAD framework in gastrointestinal endoscopy, in which the image analysis methods are automatic [159]. Coarse classification of X-Ray images has also been proposed for improving the retrieval and computational performance of an interactive radiography image retrieval system [160]. These works show that there is a pertinent interest in automated content discovery in medical imaging, as it can accommodate use cases from a wide variety of medical departments. Our proposal achieves these requirements through a direct integration with a PACS archive, so as to be applicable to multiple imaging modalities, as well as to a wider range of end users.

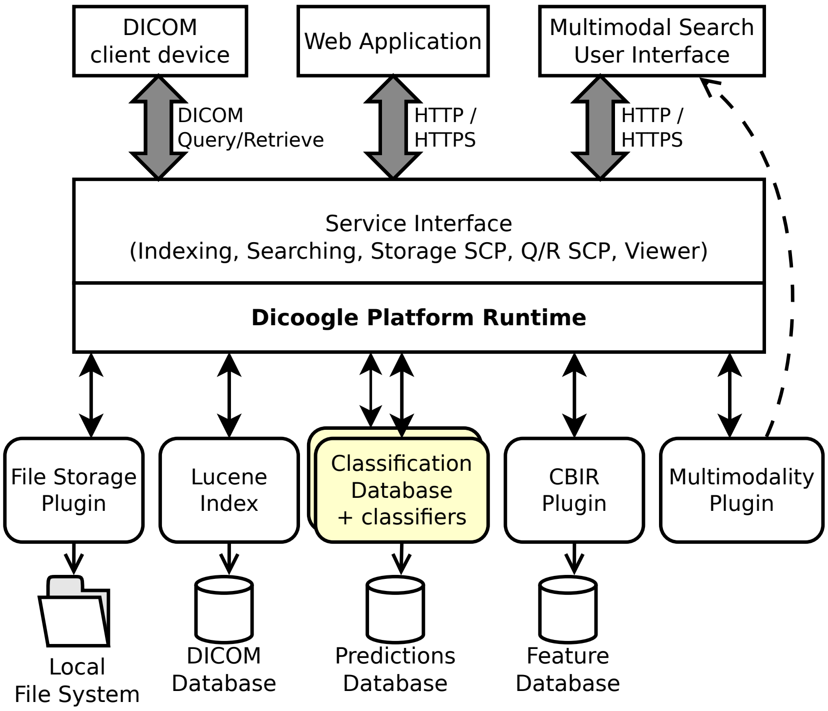


Figure 5.1: The extended architecture of a PACS archive, where Dicoogle is used as the base platform. The multimodal search engine is included. The contributions of this chapter are also highlighted in yellow.

5.2 Architecture Proposal

5.2.1 Dicoogle Integration

As highlighted throughout Chapter 2, the Dicoogle open source archive was selected as a base platform for developing this thesis work. In the previous chapter, a multimodal engine was proposed as a Dicoogle extension and, in this section, we will describe the necessary Dicoogle components, adaptions and development performed to support the proposed anatomic labeling architecture for automatic content discovery in medical imaging repositories. The solution was also instantiated as an extension to Dicoogle, since its mindset of providing a robust PACS backbone with a plugin-based architecture allows for rapid exploration and deployment of innovative solutions.

The complete Dicoogle archive architecture with automatic anatomic labeling for content discovery is depicted in Figure 5.1. When augmented with CBIR capabilities, the full architecture becomes a solution for automatic medical image classification and retrieval using combinations of text and image queries.

5.2.2 Multimodal Engine with Content Discovery

In Dicoogle, information retrieval is powered by indexing and query plugins. In technical terms, each of the two are mapped to a common Java interface, as specified in the Dicoogle SDK. This pair of concepts have enabled Dicoogle to support a variety of use cases, which cover information exploration and retrieval mechanisms, as well as data representations and search paradigms (such as query-by-example). Formally, they can be defined as follows:

- *Indexing plugins* are the components responsible for organizing stored data for the purpose of fast information retrieval. Its interface defines a function $\{I_i\} \rightarrow \mathcal{R}$, receiving a sequence of persistently stored DICOM instances I_i and producing an index report $\mathcal{R}$, providing some output regarding how many items were indexed successfully, and whether any errors emerged in the process.
- *Query plugins* provide this information by defining a function $Q \rightarrow L = \{R_i\}$, where Q is a query object and L is a sequence of search results $R_i = U_i \times D_i$, each uniquely identified by a URI U while containing arbitrary data in the dictionary D. In a common provider of DICOM meta-data, the query Q contains a string which complies to a specific syntax, namely the standard Apache Lucene query language. In CBIR, an identifier of an DICOM instance in the archive or an actual image is provided instead, yielding visually similar instances with the search outcome.

As part of the proposed architecture, a programmatic API was created and documented to support the development of classifiers in Dicoogle, so that they can be included in deployment time for use by other plugins with sufficient awareness of classification sources [1]. This API relies on a subset of the query interface, and defines how class families, classes and predictions should be named and identified in the system. More specifically, classifiers implement a function

$$Q \rightarrow L = \{P_i\}$$

, where Q is either an identifier of a DICOM instance in the archive or an actual image, and L is a list of unambiguously identified classification predictions P_i. Predictions are uniquely specified by relying on the search result's URI (as in, the U_i in each P_i). All instances of U follow the format `class://{classifier}/{criterion}#{class}`, where `{classifier}` is the unique identifier of the classifier, `{criterion}` represents the class family (e.g. *liver*), and `{class}` is the predicted class (e.g. *true* for binary class families). Classifiers are not tied to a particular algorithm, library or machine learning framework.

Usually, these organ detection procedures are not applicable to all medical images. For example, a particular classifier may only work with axial transverse CT scans of the thorax, whereas another might only expect cranial caudal mammograms. With this requirement in

---

[1] Available on GitHub: `https://github.com/Enet4/dicoogle-classification-api`

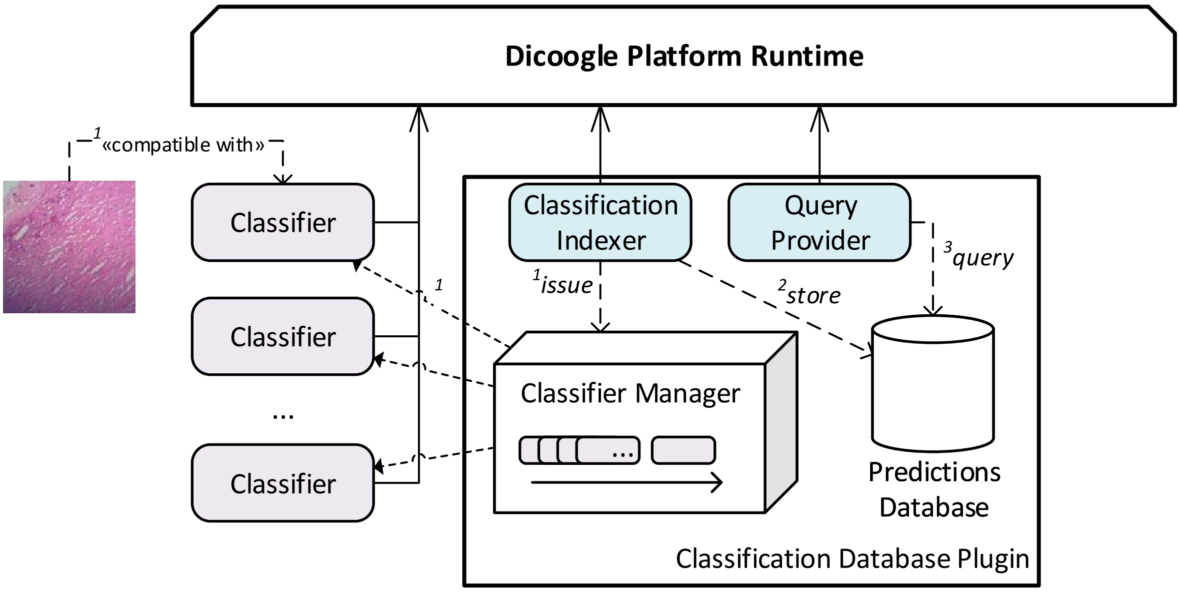


Figure 5.2: Diagram depicting the main classification database components, their interactions with existing classifiers in the Dicoogle runtime, and an example of selective classification before a given instance.

mind, a rule set is specified for each predictor, imposing conditions to images that can be fed to the classifier for certain criteria. These constraints may be either based on DICOM meta-data, such as *Modality* or *ImageType*, or inferred from the outcomes of other classifications. In order to ensure that a particular classifier is invoked only after another, a *depends-on* relation between the classifiers is established by configuring the database accordingly. Therefore, each criterion can have one or more classification dependencies, which are resolved by the database, as further detailed in the next sub-section. The ultimate value in this logic is that the architecture can be extended with a wide array of classifiers, and they will be automatically selected based on the incoming images. As such, the presented architecture can be applied to other modalities and image types.

5.2.3 Classification Database for Object Indexing

A classification database was also developed to support the proposed system, as depicted in Figure 5.2. The solution is integrated into Dicoogle, thus leveraging the available indexing and query capabilities. This extension comprises the necessary components for the architecture to fulfill the three essential tasks: (1) issue requests for classification, triggered by the platform to all installed indexers, including the *Classification Indexer*, on an indexing task; (2) store and index the obtained predictions in a persistent database; and (3) query for results, as made available by the extension's *Query Provider*.

The *Classifier Manager* maintains a record of classifiers and respective criterion rule sets in the system, so as to properly invoke the right classifiers for the given instances. As also often employed in Dicoogle for DICOM meta-data storage and retrieval [45], the *Prediction Database* was built as an independent Apache Lucene[2] instance. The use of Lucene is deemed acceptable for its query parsing, indexing and storage capabilities. In particular, Lucene enabled the database to accept multiple forms of querying, as described next:

- When indicating a binary classification criterion such as "`pancreas`", the system will output all files with a *positive* prediction of pancreas recognition. Content which has been annotated as not having pancreas can also be retrieved by indicating the negative class: "`pancreas:false`".
- The query may include the classifier's identifier for a more specific indication of the class family: "`mammo/microcalcifications`".
- By specifying the URI to an instance in the archive, namely using "`uri:"file:/CT/20163112/1.dcm"`", the user will obtain all automatically annotated content from the given file.

5.2.4 Organ Recognition Classifiers

As a proof of concept for a PACS-integrated image classifier, we have conceived a few binary classifiers with the purpose of identifying whether a specific organ is present in a CT slice.

For this goal, we have taken two data sets containing CT axial scans with professionally segmented organs. Our first data set is composed of 40 CT volumes, comprising whole body and chest+abdomen scans, with and without contrast enhancement, and providing 3D segmentations of the aorta, the liver, and the spleen, making a total of approximately 25,000 CT image slices. In order to provide a reproducible experiment, the second data set was obtained from the Public Domain Database for Computational Anatomy (PDDCA) [161], version 1.4, which is freely downloadable [3]. This database gathers 48 CT head and neck volumes, each containing a fair assortment of 3D segmentations and landmarks. For the experiments presented next, we took these volumes, making approximately 7,300 CT slices, and their respective segmentations of the brain stem and the left parotid gland.

Data processing and augmentation

The volumes used for training and validation were concatenated and their slices were treated as individual data points. The 3D masks of each segmentation were projected and concatenated over the transversal axis (which is the direction of the patient's height), so as to obtain the respective labels. Figure 5.3 shows this procedure for a single volume. The

---

[2] `https://lucene.apache.org`

[3] `http://www.imagenglab.com/newsite/pddca/` (accessed in January 2019)

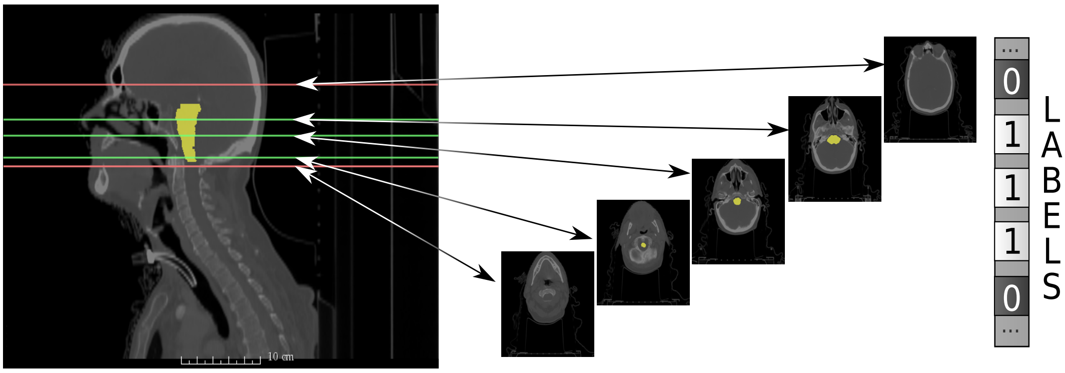


Figure 5.3: An example of projecting segmented slices into a list of labels, where the colored lines represent their estimated position.

projection yields a positive outcome in a slice if the mask exhibits a positive value in at least one pixel of that slice.

The image input was linearly transformed to have approximately a mean of zero and unit variance. Moreover, the image input was scaled down to a resolution of 128x128.

Each data set was split into 5 partitions (folds) with near-same distribution of CT volumes. In the first data set, the same number of contrast enhanced and non-contrast enhanced volumes were placed in each fold. In order to reduce the chances of a biased partitioning, the models were evaluated using 5-fold cross-validation, rather than holding out a single partition for all validation purposes, as often employed in deep learning.

Using each fold separately, an artificially enlarged data set was generated, in order to increase the robustness of the neural network model, and so reduce the chance of overfitting. Augmentation was made by performing slight random label-preserving transformations to the images. Each slice was rotated in a range of -5 to 5 degrees around the center, and then translated in a range of -8 to 8 pixels, on both axes separately, with the concrete amounts sampled randomly from a uniform distribution. Four rotations were combined with four translations, resulting in sixteen times the number of original slices, plus the original images. In the CNN's training phase, the augmented slices were included along with the original images, while leaving out the augmented volumes related to the test fold.

Convolutional neural network model and evaluation

For the proposed classification task, we designed and trained a Convolutional Neural Network (Figure 5.4). The evaluated neural network model follows a fairly similar classification approach to AlexNet [54], adapted to reflect the distinct problem at hand. More specifically, the AlexNet model was simplified in terms of the number of layers and kernels. Moreover, the local response normalization followed by each convolution was replaced with batch normalization [162]. Table 5.1 describes each layer of the network in greater detail, where

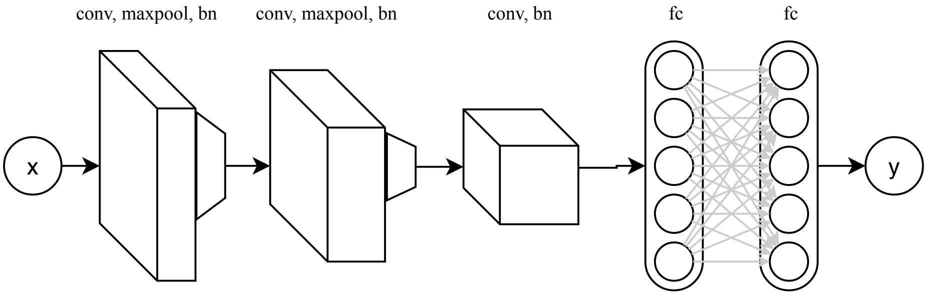


Figure 5.4: Diagram depicting the architecture of the CNN proposed for anatomy recognition, to which images (x) are fed to produce the labels (y).

Table 5.1: The specification of the experimented CNN, used as a proof of concept for the 5 organ classifiers.

Layer type	Parameters	Output shape
conv2D	48 filters, 7x7, stride 2, BatchNorm	64x64x48
max pool	3x3, stride 2	32x32x48
ReLU		
conv2D	64 filters, 5x5, stride 2, BatchNorm	16x16x64
max pool	3x3, stride 2	8x8x64
ReLU		
conv2D	96 filter, 3x3, stride 1, BatchNorm	8x8x96
ReLU		
FC	128, BatchNorm	128
ReLU		
dropout	50%	
FC	128, BatchNorm	128
ReLU		
dropout	50%	
FC	1	1

conv2D is a two-dimensional convolutional layer, BatchNorm stands for batch normalization with a moving average decay of 0.999, ReLU is a rectified linear unit activation, and FC is a fully connected neural network layer. The weights of the network were initialized with variance scaling (as in [163]), and L_2-norm regularization of the weights was included with a weight decay of 0.0005. The final neuron approximates a logit of the binary prediction (as in, tending to $+\infty$ on a positive prediction and $-\infty$ on a negative prediction).

The model was trained with the Adam optimization algorithm [164] with an initial learning rate of 0.001, which was decayed to 0.0001 at half of the training steps. The mini-batch size was 200 for the first data set and 100 for the second data set. The training phase, without augmented data, was composed of 12,000 training iterations (steps) for the first data set and 4,000 steps for the second.

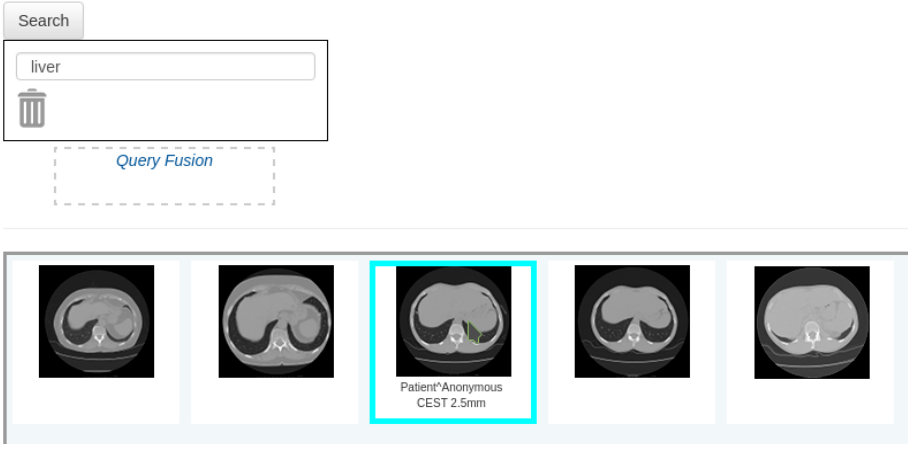


Figure 5.5: The web-based user interface for the multimodal search engine, exhibiting the first few search results for "liver".

5.3 Results and Discussion

5.3.1 Multimodal Search Engine with Anatomic Database

The specification for classifiers in a medical imaging archive and the classification database[4] were implemented and successfully integrated into Dicoogle. The developed solution is available as open source software, with the goals of facilitating deployment and establishing a common programmatic environment when integrating classifiers with this extensible archive software. Apart from the actual classifiers, the classification database is independent from the remaining components of the PACS archive.

The inclusion of the database as a Dicoogle query provider makes it available for searching when the system contains a search engine with multimodal search capabilities, such as the one designed in Chapter 4. Such a tool provides a usable combination of multiple search sources in the archive, including text and image based providers. The aforementioned solution's user interface is presented in Figure 5.5. Although the capability of searching by visual similarity is not required, it provides additional value alongside the classification database.

The benefits of this proposal are made more apparent with the following examples:

- Users can combine keyword-based DICOM queries with queries for annotated regions, which were previously unspecified in the DICOM objects. While searching for a particular organ could previously bring incomplete results (as they would largely depend on DICOM meta-data), the new system taps into automatically annotated data to produce more relevant entries.
- The results of a visual query-by-example search can be restricted to a known anatomical region, either explicitly indicated by the user or automatically extracted from the given

---

[4] Available on GitHub: `https://github.com/Enet4/dicoogle-class-db`

image, thus improving the accuracy and speed of visual retrieval. For example, a query-by-example using a CT scan of the liver can limit similarity searching to other CT scans presenting the same organ, thus scoping to potentially as much as roughly 18% of the images in the data set, an estimation based on the ratio between CT scans presenting the liver and the total number of CT scans, from the experimented data sets. In a flat visual index, the same query can be processed roughly five times faster as a consequence of this approach. Moreover, as a possible extension to this concept, each category attained from these predictions can have its own visual feature index, resulting in even more performance improvements.

- When the user is focused on one particular image, the system can present a list of the entire content automatically discovered in it. This information can aid a medical doctor as a second opinion when writing reports, thus reducing human error in the process.

The experimented classifiers were fit for primary axial CT scans on the transversal axis. A PACS archive can detect whether the images are applicable for classification by considering the attributes *Modality* (`0008,0060`), *ImageType* (`0008,0008`) and *ImageOrientationPatient* (`0020,0037`), which are required to contain a valid value for a DICOM compliant CT image [36]. This also means that, by training and installing additional classifiers, anatomic labeling of a volume in multiple planes is possible in the presence of Multiplanar Reconstruction (MPR), resulting in individually classified instances across more volume planes (such as sagittal or coronal axes). Since the second classifier, due to the distinct nature of the data set, is unaware of anatomical regions lower than the shoulders, an additional criterion for classification may be either retrieved from *BodyPartExamined* (`0018,0015`) if such a field is suitably filled, or by taking the outcome from another classifier providing CT image categorization, such as the one conceived in [153], thus keeping coherence on the intended domain of classification.

Concerning the performance of the proposed architecture, the final solution is also extremely rapid. The classifier can predict the labels of a whole body CT series stored in a hard drive as a volume with GZip compression, in approximately 7.45 seconds, on a system with an NVIDIA Tesla K80 graphics card. These measurements assume batched processing, where each slice is processed in parallel on a single GPU. Considering that the benchmarked series tended to be 860 slices large on average, this automated annotation procedure can classify as fast as 115 slices per second. As classification is typically performed only once in the indexing phase, the system would still tolerate slightly larger indexing times. The system's speed is, therefore, deemed more than adequate for the use case of automated CT volume annotation.

5.3.2 Anatomic Classifiers

The metrics retrieved from the CNN's evaluation process are presented in Table 5.2, where all metrics were averaged across the 5 folds of cross-validation, and the precision and recall

Table 5.2: The results from classifier evaluation using data sets 1 and 2.

object	accuracy	F_1 score	precision	recall
aorta	97.19%	96.01%	96.18%	95.86%
liver	97.48%	93.14%	92.22%	94.21%
spleen	95.79%	77.77%	88.45%	70.02%
brainstem	98.36%	94.60%	96.46%	92.87%
L parotid	97.07%	89.48%	92.18%	86.97%

were measured in terms of the positive outcome (where the organ is present).

Despite the simple architecture of the neural network, the trained models have achieved a best F_1 score of 96.01% (for aorta detection) and best accuracy of 98.36% (for brain stem detection). Spleen and left parotid detection were presented as the hardest tasks, which is also expected due to the distinct characteristics of each organ, with the size being a major one. As an example in the first data set, the spleen, occurring in approximately 9% of all slices on a whole body volume, was significantly harder to detect than the aorta or the liver, which would appear at the same volume in 45% and 18% of the slices, respectively.

When used in retrieval alongside other providers, a higher precision means that less noise will be passed into the results of a query, whereas a higher recall implies an increased coverage of the region. In the specific use case of domain filtering for CBIR (restricting visual similarity search to an anatomical region), the recall of the classifier may also be considered to prevent useful samples from being stripped out in the process. A higher specificity, on the other hand, will lead to less irrelevant entries, which can make visual query-by-example faster by reducing the number of similarity tests. For these reasons, we also consider the F_1 score of these classifiers as a fair metric for evaluating the performance of these models. Considering the obtained metrics, the probability of capturing at least one true positive in a volume, for any of the 5 organs, is very high. When used with a medical image viewer for clinical use (e.g. screening), a practitioner can easily observe adjacent slices starting from at least one slice containing the region to examine. These factors make us conclude that the classifiers are still reasonably appropriate for practical use, notwithstanding the possible difficulty of recognizing smaller anatomic regions or organs with severe abnormalities.

The results obtained are hard to compare with the state-of-the-art, since other solutions rely on different data sets and outputs. Organ recognition challenges with associated data sets often aim for a complete segmentation of organs or lesions, rather than a plain slice-wise identification. In order to favor reproducible research on slice-level classification for organ detection in the future, a public data set was hereby specified and implemented.

In Roth *et al.* [153], the classified organs are physically exclusive in each slice, which means that the presence of one of the five considered regions would automatically imply the non-presence of the other four. The subtle differences in the data set significantly influence a comparison with our results. An overall error rate of 5.9% was achieved in their work, in contrast to our average error rate of 2.77% over the 5 classifiers. The area under the Receiver

Operating Characteristics (ROC) curve for liver detection alone in this work is 0.999. Also regarding liver detection, the work in [149] achieved 0.5% of error rate in identifying a position near the center of the liver. Unlike our work, the goal of determining a center of the organ is performed with respect to the volume, where the non-presence of the organ is not applicable.

5.4 Final Remarks

In this chapter, an automated labeling architecture was proposed, described and validated, specifically for the improvement of multimodal content discovery in real-world medical imaging repositories. The solution was designed and developed to enhance a PACS archive's search capabilities, by adding enriched information that would only be available with a CAD system integration. This approach, in fact, resembles some of the capabilities seen in CAD, as the latter may feature automatic anatomic recognition. Still, the proposed system does not aim to invalidate or replace this concept, and integrating a complete CAD system is still deemed useful with the proposed architecture, and can even complement the archive with additional classifiers.

The particular challenge of organ detection in CT scans is still considered feasible and worthwhile with the use of deep neural networks, in spite of the need for large and varied data sets. We also believe that the performance, both in terms of speed and prediction quality, will increase with the emergence of more powerful hardware and higher quality prediction models.

The classification API and database for Dicoogle were made freely available in public code repositories. Some of the concepts presented, such as the prediction identifiers, can be translated to other systems. An unambiguous definition of these predictions also allows a system to bind classes and class families to concepts in an ontology, for subsequent semantic search [165], or to improve an otherwise visual retrieval with semantic term dissimilarity metrics [166].

Chapter 6

Representation Learning

In an era of a steadily increasing use of digital medical imaging, automatic image recognition poses an interesting prospect for novel solutions supporting clinicians and researchers. Taking this a step further, there is increasing interest in letting a model learn a meaningful representation for the various use cases in the field, as is the case for anatomic region detection. Representation learning field is growing fast in recent years [167], and many of the breakthroughs in this field are occurring in deep learning methods, which have also been strongly considered in healthcare [168].

Representation learning, or *feature learning*, can be defined as the process of learning a transformation from a data domain into a representation that makes other machine learning tasks easier to approach [167]. Given a sufficiently descriptive set of data samples from a distribution of data X, the goal is to learn a function $f : X \rightarrow Z$ that maps individual samples to a *latent space*. These outputs of f, also called latent codes, may then be used in various other tasks as descriptive and meaningful representations of the original data. In contrast, the concept of feature extraction previously defined in Chapter 3 refers to the more general concept of mapping data to another feature space, usually stated when this mapping is obtained with handcrafted algorithms rather than learned from the original data. Multiple properties of a representation can be outlined when evaluating the usefulness of learned representations. As an example, learned representations can have a lower dimensionality than the original data to exploit the existence of an intrinsic lower dimension manifold in a high dimensional data domain, which is the case for images. The task of transforming high dimensional data into a new representation of lower dimension is specifically referred to as *dimensionality reduction* [169], which can be seen as a sub-set of all representation learning methods. In contrast, a representation is over-complete when its dimensionality is greater than the original data's.

The prominence of representation learning in this thesis is justified by the following aspects of the medical imaging field.

- *Access to medical imaging data*: while multiple initiatives for the provision of medical imaging data sets exist, the process of annotating these sets with useful information is exhaustive and requires medical expertise, as it often nails down to a medical diagnosis. In the face of few to no annotations, unsupervised learning stands as a possible means of feature extraction for a measurement of relevance, leading to more powerful information retrieval and decision support solutions in digital medical imaging. Although unsupervised representation learning is very unlikely to achieve the performance of a supervised learning approach, the latter requires an exhaustive process from experts to obtain annotated content. Unsupervised learning, which avoids this issue, can also provide a few other benefits, including transferability to other problems or domains, and can often be bridged to supervised and semi-supervised techniques.

- *Automatic representation design:* throughout the literature of CAD, there is a history of feature extraction methods, specifically designed for the task at hand. While feature sets tailored by field experts are not to be disregarded, these features may either be hard to extract by a computer, which is often the case when analyzing medical images, or may overlook other important factors that would have led to a more accurate prediction. In information retrieval, the most important visual nuances may differ significantly with each query, suggesting that keeping a large amount of various factors is more important than in a specific detection task. Representation learning brings the opportunity to create suitable feature sets with enough data to learn from.

Feature learning is often mentioned in unsupervised learning and semi-supervised learning methods. Unlike a classification or regression problem, feature learning does not require a known target property for the model to predict. However, the presence of additional properties of the data, even if only available to one portion of the data set, may be included in the learning process, thus producing more informative features.

This chapter introduces representation learning and its applications in medical imaging. In addition, an assessment of unsupervised mid-level representation learning approaches for images in the biomedical literature is made. We establish a hypothesis that a sufficiently powerful representation of images would enable a medical imaging archive to automatically detect biomedical concepts with some level of certainty and efficiency, thus improving the system's information retrieval capabilities over non-annotated data. Representations are built using an ensemble of images from biomedical literature. The learned representations were validated with a brief qualitative feature analysis, and by training simple classifiers for the purpose of biomedical concept detection. The outcomes show that feature learning techniques based on deep neural networks can outperform techniques that were previously common-place in image recognition, and that models with adversarial networks, albeit harder to train, can improve the quality of feature learning.

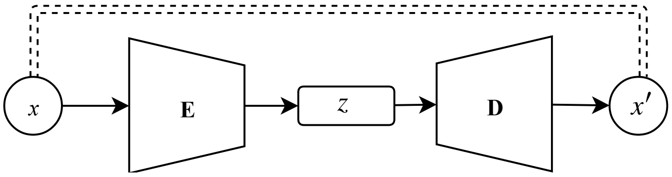


Figure 6.1: Diagram depicting the basic auto-encoder structure.

6.1 Related Work

Representation learning can be achieved using a wide range of methods. One of the earliest forms of representation learning is Principal Component Analysis (PCA), a linear transformation $f(\mathrm{x}) = \mathrm{h} = \mathrm{W}^T\mathrm{x} + b$ that maximizes the variance of orthogonal directions in h using an orthogonal basis matrix W. In other words, the transformation makes an attempt at creating de-correlated factors while capturing high variability among data points in the set. Given its linear nature, Principal Component Analysis (PCA) performs poorly when attempting to represent complex patterns in the data. Independent Component Analysis (ICA) represents a wider family of algorithms for finding a linear transformation with minimal statistical component dependence [170].

Sparse coding is the concept of separating inputs into a linear combination of disentangled bases [171]. That is, a set $\vec{b_0}, \vec{b_1}, ..., \vec{b_k} \in \mathbb{R}^n$ is learned so that each input x can be described as a vector $\vec{s} \in \mathbb{R}^k$ so that $x = \sum_i^k \vec{b_i} s_i$. Untied to a specific algorithm, a wide family of sparse coding techniques exist. The Bag of Visual Words (BoVW) algorithm described in Chapter 3, for instance, is a form of sparse coding where the bases are modeled through the learning of the visual word vocabulary. In image recognition, algorithms based on bags of visual words have been prevalent, as they have shown superior results over other low-level visual feature extraction techniques [172, 173]. More recently however, deep learning methods have been favored, very often surpassing traditional methods in image recognition tasks.

6.1.1 Auto-encoders

Auto-encoders are models which learn to reconstruct the original inputs when subjected to an information bottle-neck (Figure 6.1). The basic form of an auto-encoder is a pair of parametric functions: the encoder $E(x) \rightarrow z$ and the decoder $D(z) \rightarrow x'$, where the goal is to minimize the difference between x and $x' = D(E(x))$. When using neural networks for the encoder and the decoder, the auto-encoder can be trained like other feed-forward neural networks through gradient descent, with the main objective of minimizing $d(x, D(E(x)))$. For the dissimilarity function d, the mean squared error or binary cross-entropy are commonly employed.

Unlike linear models, auto-encoders can be composed of an arbitrary number of layers with non-linearities in-between, both in the encoder and the decoder, as in an artificial neural network. The first examples of deep auto-encoders were trained by iteratively stacking additional layers [174]. However, multiple improvements to the convergence and stability of deep neural network training have been proposed in literature since then, leading to recent work on auto-encoders proposing end-to-end training directly.

A meaningful encoding in a traditional auto-encoder requires the middle layer z to have a lower dimensionality, so as to prevent the components of the model from learning the identity function without learning anything about the given data distribution. On the other hand, the auto-encoder has been extended or adapted to constrain this bottleneck in other ways, subsequently enabling these models to create useful representations of higher dimensionality. A few examples of regularized auto-encoders follow.

- Sparse auto-encoders are regularized in such a way as to yield a sparse representation, that is, with a high number of features set to zero. Sparsity of the latent features can achieved with various mechanisms. A ReLU activation at the end of the encoding process will turn negative values to zeros, thus contributing to sparse representations. As another example, an absolute value penalization can be applied to the vector z, by adding the extra minimization goal of keeping the sum of this vector small in magnitude. Sparsity-imposing models benefit from the useful properties also exhibited in sparse coding methods.
- In denoising auto-encoders, the original data x is deliberately corrupted in some way to produce noisy samples $\tilde{x}$. Its goal is then to learn the pair of functions (E, D) so that $D(E(\tilde{x}))$ is closest to the original input x. The aim of making E a function of $\tilde{x}$ is to force the process to be more stable and robust to noise by requiring samples to lie on the same manifold as those without noise, thus leading to higher quality representations [175].
- The Variational Auto-encoder (VAE) employs a stochastic sampling and distribution divergence technique to create variational inference [176]. In other words, these models also enable the generation of samples which resemble the original training data, even without data inputs, which would not be easily approachable in other kinds of auto-encoders. The encoder learns to produce a distribution (usually a Gaussian distribution $\mathcal{N}(\mu, \sigma I)$), and the respective latent codes are obtained with a *reparameterization* trick: given a vector ϵ sampled from the unit Gaussian distribution $\mathcal{N}(0, I)$, the latent code becomes $z = \epsilon \times \sigma + \mu$. The VAE's loss function combines the reconstruction loss with the Kulback-Leibler divergence between the distribution of codes sampled from E and $\mathcal{N}(0, I)$.

Some forms of regularization can be applied simultaneously, as is the case for the Sparse Denoising Auto-encoder (SDAE). Section 6.2.2 shows an example of a Sparse Denoising Auto-encoder (SDAE).

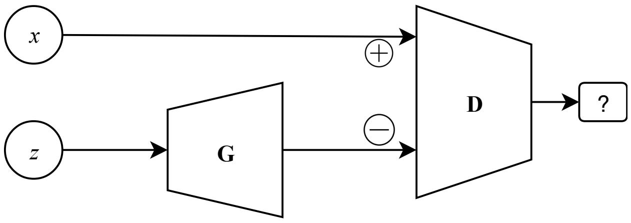


Figure 6.2: Basic structure of a GAN.

6.1.2 Generative Adversarial Networks

Research on representation learning is even more intense on the ground-breaking concept of GAN [177]. Initially conceived in 2014, the GAN establishes a min-max game between two mutually learning models (as depicted in Figure 6.2):

- a *generator*, the goal of which is to produce realistic data samples;
- a *discriminator* network, which learns to discriminate generated samples from real ones.

The original min-max game of a basic GAN, can be described in mathematical terms as two loss functions,

$$L_D = -\mathbb{E}_{\mathrm{x}\sim p_{\mathrm{x}}}[\log D(\mathrm{x})] - \mathbb{E}_{\mathrm{z}\sim p(\mathrm{z})}[\log(1 - D(G(\mathrm{z})))], \tag{6.1a}$$

$$L_G = \mathbb{E}_{\mathrm{z}\sim p(\mathrm{z})}[\log(1 - D(G(\mathrm{z})))], \tag{6.1b}$$

where x is a data sample and z is a stochastic prior code which is sampled from an arbitrary distribution. Very often, z is sampled from a random standard distribution, such as a uniform distribution ($U(-1,1)$) or a normal distribution ($\mathcal{N}(0,I)$). However, a *non-saturating* version of the generator is preferred in practice,

$$L_G = -\mathbb{E}_{\mathrm{z}\sim p(\mathrm{z})}[\log D(G(\mathrm{z}))], \tag{6.2}$$

where the loss function is adjusted to maximize the probability of samples being perceived as real, rather than minimizing the probability of samples being perceived as fake.

As the two components improve, this system will ideally reach a state where it can no longer improve: the generator will ultimately produce visually-appealing samples of as much resemblance as the original data, and the discriminator will reach an average of 50% in discrimination accuracy, as the fake samples become too difficult to tell apart from real data. In other words, the system will ideally reach a *Nash* equilibrium [178, 179]. The impressive quality of the samples generated by GANs in literature have led the scientific community into devising multiple variants and applications to this approach.

While GANs are generally seen as devised primarily for image synthesis, the interest of GANs for representation learning has also not been disregarded. The basic GAN architecture does not provide a means to encode samples to their respective prior code, which may seem as a barrier towards its use in feature learning. However, multiple approaches were devised to overcome this aspect, thus being able to use GANs in feature learning scenarios. A few of these approaches are described next.

- *Learned approximate inference*, also known as *latent regression* in other works, was suggested alongside the GAN, consisting in training an auxiliary encoder network to predict the code z based on samples x' generated by a trained model [177]. However, this method has been shown to perform very poorly, mainly as a consequence of the encoder never observing real data [180].
- It is possible to exploit the discriminator's learned features, by retrieving this network's activations at some of its deep layers when fed with samples. These activations may then be used as feature for other tasks, such as classification [181].
- The Bidirectional Generative Adversarial Network (BiGAN), depicted in Figure 6.5, addresses this concern by including an encoder component, which learns the inverse process of the generator [180]. Rather than only observing data samples, the BiGAN discriminator's loss function depends on the code-sample pair. A stochastic version of this model has been proposed around the same time [182].

The loss functions used in GANs have been applied in the bottleneck vector of an auto-encoder instead of the samples, resulting in the Adversarial Auto-encoder (AAE) [183]. This model benefits from the latent code inference capabilities of an auto-encoder, as well as of the variational inference of a VAE if well regularized. The adversarial loss function is similar to the original [177], where the discriminator learns to distinguish the distribution $q(z \sim E(x))$ from a prior distribution $p(z)$:

$$L_D = -\mathbb{E}_{\mathrm{z} \sim p_{\mathrm{z}}}[\log D(\mathrm{z})] - \mathbb{E}_{\mathrm{x} \sim p(\mathrm{x})}[\log(1 - D(E(\mathrm{x})))], \tag{6.3a}$$

$$L_E = -\mathbb{E}_{\mathrm{x} \sim p(\mathrm{x})}[\log D(G(\mathrm{x}))]. \tag{6.3b}$$

Section 6.2.2 and Section 6.3.1 show concrete examples of Adversarial Auto-encoders evaluated in this work.

6.1.3 Representation Learning in the Medical Imaging domain

Representation learning has been notably used in medical image retrieval, although even in this decade, handcrafted visual feature extraction algorithms are frequently considered in this context [96, 184]. Nonetheless, although the interest in deep learning is relatively recent, a wide variety of neural networks have been studied for medical image analysis [152, 185], as they often exhibit greater potential for the task [186]. The use of unsupervised learning techniques is also well regarded as a means of exploiting as much of the available medical imaging data as possible [187]. In particular, Restricted Boltzmann Machines (RBMs) have been used for multimodal medical information retrieval [117]. Representation learning methods based on auto-encoders have been considered in medical data and are still prevalent to this day [188]. Unsupervised representation learning has also been used as an intermediate step towards the diagnosis of Alzheimer's disease and mild cognitive impairment, using stacked auto-encoders [189].

GANs were also experimented with in medical imaging for a variety of use cases, exhibiting capabilities ranging from quality improvement in low dose CT scans [190], translation between modalities [191], and many more [192]. However, the direct use of a GAN's generated sample in clinical use is still discouraged, since the generated images can produce misleading image characteristics [193].

6.2 Comparative Analysis of Unsupervised Learning Methods for Concept Detection

In this section, an assortment of representation learning techniques are presented, including their evaluation in a biomedical concept detection problem. We have considered a set of unsupervised representation learning techniques, both traditional (as in, employing classic computer vision algorithms) and based on deep learning, for the scope of images in the biomedical domain. These representations were subsequently used for the task of biomedical concept detection. Namely:

- we have experimented with creating image descriptors using BoWs, for two different visual key-point extraction algorithms.
- With the use of modern deep learning approaches, we have designed and trained various deep neural network architectures: a Sparse Denoising Auto-encoder (SDAE), a Variational Auto-encoder (VAE), a Bidirectional Generative Adversarial Network (BiGAN), and an Adversarial Auto-encoder (AAE).

6.2.1 Bags of Visual Words

For each data set, images were converted to grayscale without resizing and visual key-point descriptors were subsequently extracted. We employed two key-point extraction algorithms separately: Scale Invariant Feature Transform (SIFT) [83], and Oriented FAST and Rotated BRIEF (ORB) [85]. While both algorithms obtain scale and rotation invariant descriptors, ORB key-points are significantly faster to compute. The key-points were extracted and their respective descriptors computed using the OpenCV 3 library [100]. Each image would yield a variable number of descriptors of fixed size (128-dimensional for SIFT, 32-dimensional for ORB).

In the event that the algorithm did not retrieve any key-points for an image, the algorithm's parameters were adjusted to loosen edge detection criteria. In SIFT, the contrast threshold (`contrastThreshold`) was changed from 0.04 to 0.01, the edge threshold (`edgeThreshold`) was increased from 10 to 20, and the sigma of the Gaussian on the first octave (`sigma`) was modified from 1.6 to 1.4. With ORB, the FAST key-point detection threshold (`fastThreshold`) was changed from 20 to 10. Nonetheless, such cases of images yielding no key-points were residual in the data set (less than 0.05% of all images), and occurred mostly in images with little to no potential for visual understanding. As a last resort, images without any meaningful features (e.g. images of solid color or gradients) were given an empty bag of words (all zeros).

All procedures described henceforth are the same for both ORB and SIFT key-point descriptors. From the training set, 3,000 files were randomly chosen and their respective key-point descriptors collected to serve as template key-points. A visual vocabulary (codebook) of size $k = 512$ was then obtained through k-means clustering. The remaining steps for building the BoVWs are described in Section 3.3.3.

6.2.2 Deep unsupervised learning

Modern representation learning techniques often rely on deep learning methods. We have considered a set of deep convolutional neural network architectures for inferring a late feature space over biomedical images. These models are composed of parts with very similar numbers of layers and parameters, in order to obtain a fairer comparison in the evaluation phase. This also means that the models will have very similar resource requirements (time, processing power, and memory) in the feature extraction phase.

Training samples were obtained through the following process: images were resized so that its shorter dimension (width or height) was exactly s_g pixels. Afterwards, the sample was augmented by feeding the networks random crops of size $s \times s$ (out of 9 possible kinds of crops: 4 corners, 4 edges and center). Validation images were resized to fit the $s \times s$ dimensions. For all cases, the images' pixel RGB values were normalized to fit in the range [-1, 1]. Unless otherwise mentioned, the networks assumed a rescale size to $s_g = 96$ and a crop size $s = 64$.

Models with an encoding or discrimination process for visual data were based on the same convolutional neural network architecture, described in Table 6.1. These models were influenced by the work on deep convolutional generative adversarial networks [181]. Each encoder layer is composed of a 2D convolution, followed by a normalization algorithm and a non-linearity. The specific normalization and activation employed depend on the particular model, and are concretely indicated in their respective sub-sections. At the top of the network,

Table 6.1: A tabular representation of the SimpleNet layers' specifications.

Layer	Kernels	Size/Stride	Details
conv1	64	5x5 /2	normalization + non-linearity
conv2	128	5x5 /2	normalization + non-linearity
conv3	256	5x5 /2	normalization + non-linearity
conv4	512	5x5 /2	normalization + non-linearity
conv5	512	5x5 /2	normalization + non-linearity
fc	nb		linear

Table 6.2: A tabular representation of decoder/generator layers' specifications.

Layer	Kernels	Size/Stride	Details
fc	4,096		reshaped to 2x2x1024, linear
dconv5	512	5x5 /2	normalization + ReLU
dconv4	256	5x5 /2	normalization + ReLU
dconv3	128	5x5 /2	normalization + ReLU
dconv2	64	5x5 /2	normalization + ReLU
dconv1	3	5x5 /2	linear

global average pooling is performed, followed by a fully connected layer, yielding the code tensor z. The Details column in both tables may include the normalization and activation layers that follow a convolution layer, where ReLU stands for the standard, non-leaking rectified linear unit $max(0, x)$.

The decoding blocks replicate the encoding process in inverse order (Table 6.2). It starts with a fully connected network from the latent (or prior) code vector into a series of convolutional layers. Convolutions in these blocks are transposed (also called *fractionally-strided convolution* in literature, and *deconvolution* in a few other papers). The weights of the network were randomly initialized with a Gaussian distribution.

Sparse Denoising Autoencoder

The first tested deep neural network model is a common auto-encoder with denoising and sparsity constraints (Figure 6.3). In the training phase, a Gaussian noise of standard deviation 0.05 was applied over the input, yielding a noisy sample $\tilde{x}$. As a denoising auto-encoder, its goal is to learn the pair of functions (E, D) so that $x' = D(E(\tilde{x}))$ is closest to the original input x. The aim of making E a function of $\tilde{x}$ is to force the process to be more stable and robust, thus leading to higher quality representations [175].

Sparsity was achieved with two mechanisms. First, a ReLU activation was used after the last fully connected layer of the encoder, turning negative outputs from the previous layer into zeros. Second, an absolute value penalization was applied to z, thus adding the extra minimization goal of keeping the code sum small in magnitude. The final decoder loss function

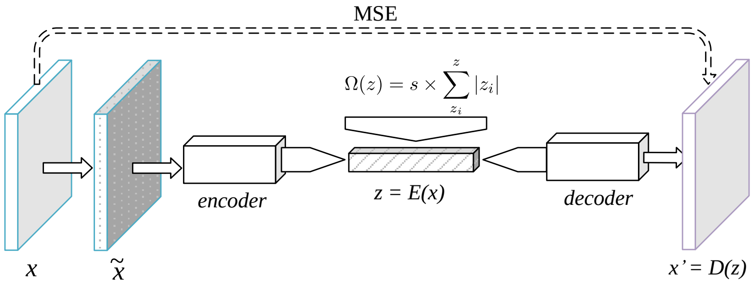


Figure 6.3: Diagram of the sparse denoising auto-encoder.

was therefore

$$\mathcal{L}(E, D) = \frac{1}{r} \sum_{i=0}^{\mathrm{r}} \left(x_i - x_i'\right)^2 + \Omega(z), \tag{6.4}$$

where

$$\Omega(z) = s \times \sum_{i}^{z} z_i$$

is the sparsity penalty function, r is the number of pixels in the input images (64×64), and x represents the original input without synthesized noise. s is the sparsity coefficient, which we left defined as $s = 0.0001$. This network used batch normalization [162] and standard ReLU activations.

Variational Autoencoder

The encoder of the variational auto-encoder (Figure 6.4) learns a stochastic Gaussian distribution which can be sampled from, by minimizing the Kulback-Leibler divergence with a unitary Gaussian distribution [176]. The architecture of this network resembles the SDAE, with a relevant exception at the encoding process: in order for the encoder to produce a sample distribution, the final dense layer was replaced with two layers of equal dimensions μ and σ. The encoded sample is obtained with the reparameterization trick: given $\epsilon \sim \mathcal{N}(0, I)$, the latent code becomes $z = \epsilon \times \sigma + \mu$. The VAE's loss function was

$$\mathcal{L}(E, D) = \frac{1}{r} \sum_{i=0}^{\mathrm{r}} \left(x_i - x_i'\right)^2 + \mathcal{KL}[\mathcal{N}(\mu, \sigma) || \mathcal{N}(0, I)],$$

where $E(x) := (\mu, \sigma)$ and $\mathcal{KL}$ is the Kulback-Leibler divergence from the learned distribution.

As in the SDAE, convolutional layers were followed by batch normalization [162] and ReLU activations.

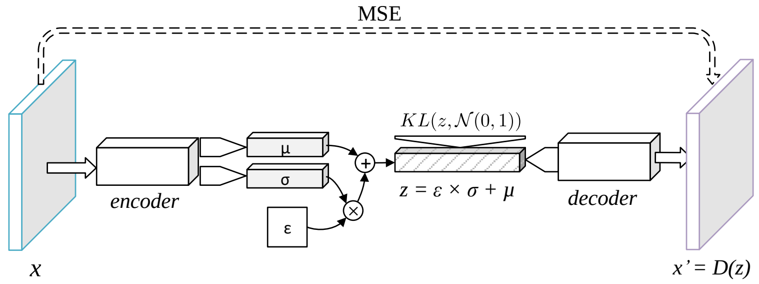


Figure 6.4: Diagram of the variational auto-encoder.

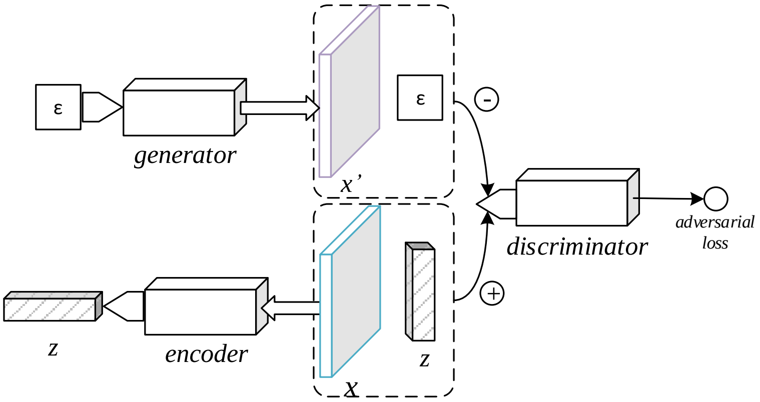


Figure 6.5: Diagram of the bidirectional GAN.

Bidirectional GAN

In order to exploit the GAN's potential for representation purposes, we have trained a BiGAN, depicted in Figure 6.5, thus including an encoding component that learns the inverse process of the generator [180]. The training process can be formally defined into eq. 6.5. In plain words, the encoder's goal is to learn to fool the discriminator into thinking that the sample-code pair (x, $E(x)$) has a fake sample:

$$V(G,E,D) = \min_{G,E} \max_{D} \mathbb{E}_{\mathrm{x}\sim p_{\mathrm{x}}}[\log D(\mathrm{x}, E(\mathrm{x}))] + \mathbb{E}_{\mathrm{z}\sim p_{\mathrm{z}}}[\log (1 - D(G(\mathrm{z}), \mathrm{z}))]. \tag{6.5}$$

The *Encoder* component relies on the same convolutional neural network design as the rest, with an exception suggested in [180]: the original data was fed to the encoder with a size s of 112x112, cropped from images with the shortest dimension resized to 128 pixels (as in, $s_g = 128$). Images were still down-sampled to 64x64 in order to be fed to the discriminator. The discriminator, in turn, was conceived as follows.

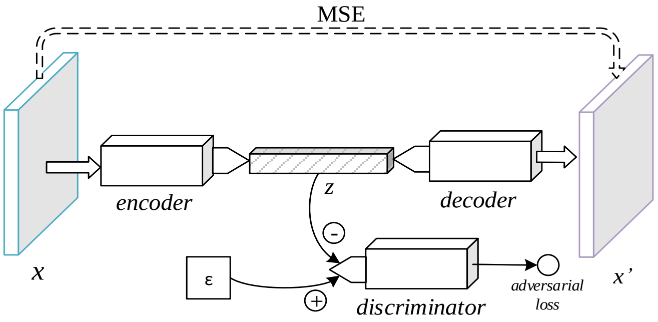


Figure 6.6: Diagram of the adversarial auto-encoder.

1. Like in the encoder, the image was processed by the convolutional neural network described in Table 6.7 with $nb = 128$.
2. The prior code z was fed to two fully connected layers with the output shape Bx64 (where B is the batch size).
3. The two outcomes (1) and (2) were concatenated to form a tensor of shape Bx192, followed by 2 fully connected networks of shape Bx512.
4. Finally, a fully connected layer with a single neuron (Bx1) produces the output $D(x, z)$.

As in [180], all constituent parts of the GAN were optimized simultaneously in each iteration. The encoder and the discriminator of this model used layer normalization [194] and *leaky ReLU* with a leaking factor of 0.2 on all layers except the discriminator's last convolutional layer indicated in step (1) and the final output stated in step (4).

Adversarial Auto-Encoder

The Adversarial Auto-encoder, as described in Section 6.1.2, is an auto-encoder in which a discriminator is added to the bottleneck vector [183]. While reducing the L_2-norm distance between a sample and its decoded form (as in other auto-encodes), the AAE includes an adversarial loss for distinguishing the encoder's output from a stochastic prior code, thus serving as a regularizer to the encoding process.

Our AAE (Figure 6.6) used a simple code discriminator composed of 2 fully connected layers of 128 units with a *leaky ReLU* activation for the first two layers, followed by a single neuron without a non-linearity. During training, the discriminator is fed a prior z sampled from a random normal distribution $\mathcal{N}(0, I)$ as the *real* code, and the output of the encoder $E(x)$ as the *fake* code. The model uses layer normalization [194] on all except the last layers of each component, and leaky ReLU with a leaking factor of 0.2. Like in [180], all three components' parameters were updated simultaneously in each iteration.

Network Training Details

The networks were trained through stochastic gradient descent, using the Adam optimizer [164]. The α_1 hyperparameter was set to 0.5 for the BiGAN and the AAE, and 0.9 for the remaining networks.

Each neural network model was trained over 206,000 steps, which is approximately 100 epochs, with a mini-batch size of 64. The base learning rate was 0.0005. The learning rate was multiplied by 0.2 halfway through the training process (50 epochs), to facilitate convergence.

All neural network training and latent code extraction was conducted using TensorFlow, and TensorBoard was used during the development for monitoring and visualization [59]. Depending on the particular model, training took on average 120 hours (a maximum of 215 hours, for the adversarial auto-encoder) to complete on one of the GPUs of an NVIDIA Tesla K80 graphics card in an Ubuntu server machine.

6.2.3 Evaluation Methodology

The previously described methods for representation learning were aimed towards addressing the domain of biomedical images. A proper validation of these features was made with the use of the data sets from the ImageCLEF 2017 caption challenge [195], specifically the portion corresponding to the concept detection task. In 2017, the focus of the challenge was on automated medical image understanding for a quicker decision making in clinical diagnoses. As the first sub-task of the caption prediction challenge, the goal of the concept detection task is to conceive a computer model for identifying the individual concepts from medical images, from which full captions could be composed. In this context, a concept is a single point of an ontology, referring to a particular meaning.

This task was accompanied with a collection of images, originally seen in open access biomedical journals from the PubMed Central, and encompassing a wide variety of modalities (radiology, biopsy, pathology, and general diagrams, among others). Images were formatted in JPEG, with resolutions varying from 400 to 800 pixels in either dimension. Each image is also coupled with the respective caption (which was not used in this work), along with the extracted list of biomedical concepts. These concepts were written as Concept Unique Identifiers (CUIs) from the UMLS[1], which specifies a large vocabulary for biomedical information systems. A few examples from the collection are shown in Figure 6.7, and a small insight of the concepts present in these images may be obtained from Table 6.3. The full data set was divided into three parts: the *training set* (164,614 images), the *validation set* (10,000 images) and the *testing set* (10,000 images). The testing set's ground truth was hidden during the challenge, but was later on provided to the participants.

Each of the set of features, learned from the approaches described in the previous section, were used to train simple classifiers for concept detection. In both cases, the same training

---

[1] `https://www.nlm.nih.gov/research/umls`

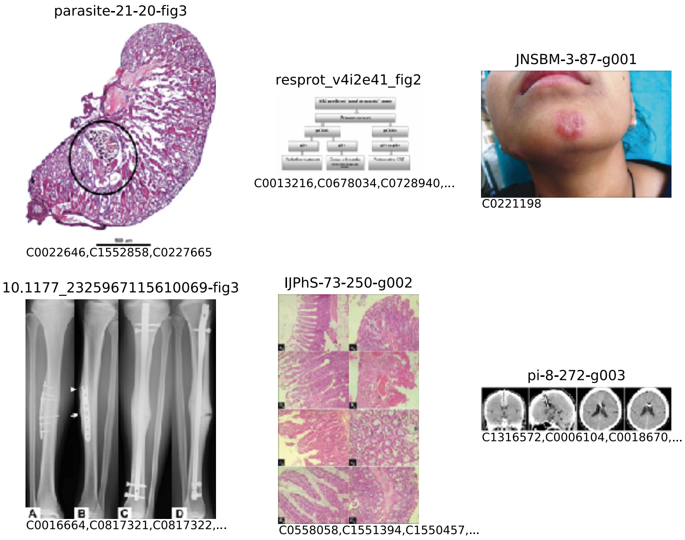

parasite-21-20-fig3
C0022646,C1552858,C0227665
resprot_v4i2e41_fig2
C0013216,C0678034,C0728940,...
JNSBM-3-87-g001
C0221198
10.1177_2325967115610069-fig3
A
B
C
D
C0016664,C0817321,C0817322,...
IJPhS-73-250-g002
C0558058,C1551394,C1550457,...
pi-8-272-g003
C1316572,C0006104,C0018670,...

Figure 6.7: A few samples from the ImageCLEF 2017 concept detection data set, with their respective file ID and trimmed list of concept identifiers.

Table 6.3: The ten most frequently occurring concepts in the ImageCLEF 2017 training set for concept detection.

CUI	Occurrences in training set	Textual description
C1696103	17,998	image-dosage form
C0040405	16,217	X-ray computed tomography
C0221198	14,219	lesion
C1306645	10,926	plain X-ray
C0577559	9,769	mass (lump, localized mass)
C0027651	9,570	tumor
C0441633	9,289	diagnostic scanning
C0817096	5,602	thorax
C1317574	5,039	note
C0087111	4,983	therapy

and validation folds from the original data set were considered, after being mapped to their respective feature spaces. In addition, data points in the validation set with an empty list of concepts were discarded.

These simple models were used to predict the concept list of each image by sole observation of their respective feature set. Therefore, the assessment of our representation learning methods is made based on the representation's descriptiveness without learning additional complex feature spaces, and so, on the effectiveness of capturing high-level features from latent codes alone.

Logistic regression

Aiming for low complexity and classification speed, we performed logistic regression with mini-batch gradient descent for concept detection, treating the UMLS terms as labels. More specifically, linear classifiers were trained over the features, one for each of the 750 most frequently occurring concepts in the training set. All models were trained using FTRL-Proximal optimization [196] with a base learning rate of 0.05, an L_1-norm regularization factor of 0.001, and a batch size of 128. Since the biomedical concepts are very sparse and imbalanced, the F_1 score was considered as the main evaluation metric, which was calculated with respect to multiple fixed operating point thresholds (namely, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, and 0.2), for each sample, and averaged across the 750 labels. The threshold which resulted in the highest mean F_1 score on the validation set is recorded, and the respective precision, recall, and area under the ROC curve were also included. Subsequently, the same model and threshold were used for predicting the concepts in the testing set, the F_1 score of which was retrieved with the official evaluation tool from the ImageCLEF challenge[2].

Since it is also possible to combine multiple representations with simple vector concatenation, we have experimented training these classifiers using a mixture of features from the SDAE and AAE latent codes. This process is often called *early fusion*, and is contrasted with *late fusion*, which involves merging the results of separate models. Each model undertook a few dozens of training epochs until the best F_1 score among the thresholds would no longer improve. In practice, training and evaluation of the linear classifiers was done with TensorFlow.

k-nearest neighbors

A relevant focus of interest in representation learning is its potential in information retrieval. While concept detection is not a retrieval problem, and the use of retrieval techniques is a naive approach to classification, it is fast and scales better in the face of multiple classes. Furthermore, it enables a rough assessment of whether the representation would fare well in

---

[2] Available on the official website: `https://www.imageclef.org/2018/caption`

retrieval tasks where similarity metric were not previously learned, which is the case for the Euclidean distance between features.

A modified form of the k-nearest neighbors algorithm was used as a second means of evaluation. Each data point in the validation set had its concepts predicted by retrieving the k closest points from the training feature set (henceforth called neighbors) in Euclidean space and accumulating all concepts of those neighbors into a boolean sum of labels. This tweak makes the algorithm more sensitive to very sparse classification labels, such as those found in the biomedical concept detection task. All natural numbers from one to five were tested for the possible k number of neighbors to consider. Analogous to the logistic regression above, the k which resulted in the highest F_1 score on the validation set was regarded as the optimal parameter, and predictions over the testing set were evaluated using the optimal k.

The actual search for the nearest neighbors was performed using the Faiss library, which contributed to a rapid retrieval [197]. Feature fusion was not considered in the results, as they did not seem to bring any improvement over singular representations.

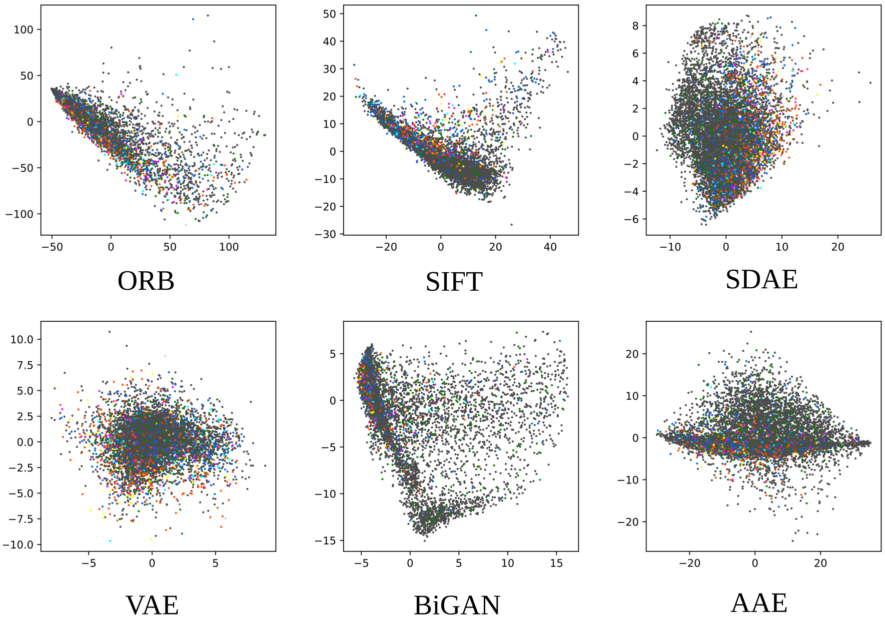
6.2.4 Results and Discussion

Qualitative Results

Each representation learning approach described in this work resulted in a 512-dimensional feature space. Figure 6.8 shows the result of mapping the validation feature set of each representation learned into a two-dimensional space, using PCA. The three primary colors were used (red, green, and blue) to label the points with the three most commonly occurring UMLS terms in the training set, namely `C1696103` (*image-dosage form*), `C0040405` (*X-ray CT*), and `C0221198` (*lesion*), each painted in an additive fashion.

While extreme outliers were removed from the figures, it can be noted that the ORB, SIFT and BiGAN representations had more outliers than the other three representations. A good representation would enable samples to be linearly separable based on their list of concepts. Even though the concept detection task is too hard for a clear cut separation, one can still identify regions in the manifold in which points of one of the frequent labels are mostly gathered. The existence of concentrations of random points in certain parts of the manifold, as further observed from the classification results, is noticeable mostly in poorer quality representations.

The latent space regularization in representations based on deep learning is also apparent in these plots: both the AAE (with the approximate Jensen-Shannon divergence from the adversarial loss) and the VAE (with the Kullback-Leibler divergence) manifest great similarity with a normal distribution.



ORB SIFT SDAE

VAE BiGAN AAE

Figure 6.8: The 2D projections of the latent codes in the validation set, for each learned feature space.

Linear Classifiers

Table 6.4 shows the best resulting metrics obtained with logistic regression on the validation set, followed by the final score on the testing set. *Mix* is the identifier given for the feature combination of SDAE and AAE. We observed that, for all classifiers, the threshold of 0.075 would yield the best F_1 score. This metric, when obtained with the validation set, assumes the existence of only the 750 most frequent concepts in the training set. Nonetheless, these metrics are deemed acceptable for a quantitative comparison among the trained representations, and have indeed established the same score ordering as the metrics in the testing set. The adversarial auto-encoder obtained the best mean F_1 score in concept detection, only superseded by a combination of the same features with those from the sparse denoising auto-encoder.

These metrics, although seemingly low, are within the expected range of scores in the domain of concept detection in biomedical images, since the classified labels are very scarce. As an example, only 10.9% of the training set is positive for the most frequent term. For the second and third most frequent terms, the numbers are 9.8% and 8.6% respectively. The mean number of positive labels of each of the 750 most frequent concepts is 876.7, with a minimum of 203 positive labels for the 750th most frequent concept in the training set. We find that most concepts in the set do not have enough images with a positive label for a valuable classifier.

The scores obtained here can be compared with the results from the ImageCLEF 2017

Table 6.4: The best metrics obtained from logistic regression for each representation learned, where *Mix* is the feature combination of SDAE and AAE. The highest scores are shown in bold.

Type	AUC	Val. F_1 score	**Test F_1 score**
ORB	0.699	0.138	0.0967
SIFT	0.753	0.133	0.0952
SDAE	0.781	0.151	0.1029
VAE	0.760	0.140	0.0924
BiGAN	0.781	0.141	0.0989
AAE	0.787	*0.159*	**0.1080**
Mix	0.789	*0.161*	**0.1105**

challenge. The best F_1 scores on the testing set, without the use of external resources that could severely bias the results, were 0.1583 and 0.1436 [195]. In the former, the participants performed multi-label classification directly over the images using a neural network model that was pre-trained with the ImageNet and COCO data sets [198]. The use of additional information outside of the given data sets is known to significantly improve the results. In the list of submissions where no external resources were used, these techniques were only outperformed by the submissions from the IPL team [199], which have extracted global image descriptors using dense SIFT bags of words and quad-tree bags of colors. While recognizing the fact that the work also relied on building global features in an unsupervised manner, the representations obtained in this work are significantly more compact in size, and thus more computationally efficient in practice.

The combined representation of concatenating the feature spaces of the SDAE and AAE have resulted in even better classifiers. Although the results of the combined representation are shown here, this improvement is not to be overstated, given that it relies on a wider feature vector and on training two representations that were meant to perform individually. Another relevant observation is that the representations based on BoWs were generally less effective for the task than deep representation learning methods. Although SIFT BoWs have resulted in a slightly better area under the ROC curve, the chosen operating points led to ORB slightly outperforming SIFT.

k-Nearest Neighbors

The results of classifying the validation set with similarity search are presented in Table 6.5. The presence of lower F_1 scores than those with linear classifiers is to be expected: the linear classifier can be interpreted as a model which learns a custom distance metric for each label, whereas k-NN relies on a fixed Euclidean distance metric. With k-nearest neighbors, the best mean F_1 score of 0.0751 was obtained with the SDAE. The AAE follows with a mean F_1 score of 0.0691. The form of passive fitting over the validation set, from the choice of k, is much less

Table 6.5: The best metrics obtained from vector similarity search for each representation learned. The highest scores are shown in bold.

Kind	k	Val. F_1 score	**Test F_1 score**
ORB	4	0.043	0.0418
SIFT	3	0.060	0.0567
SDAE	2	*0.080*	**0.0751**
VAE	4	0.036	0.0345
BiGAN	3	0.047	0.0473
AAE	2	0.072	0.0691

greedy than the training process of the logistic regression, which included a choice of operating threshold and halting condition based on the outcome from the validation set. Therefore, it is expected that the final F_1 scores on the testing set do not deviate as much from the scores obtained on the validation set.

6.3 Feature Learning for Concept Detection in Medical Images

The previous section was extremely important, since it allowed us to perform a practical evaluation of state-of-the-art unsupervised learning techniques. The experience and lessons learned were fundamental for experimenting new representation learning methods, through which other machine learning tasks can be employed more efficiently, including biomedical concept detection and CBIR. This section presents our proposal that was evaluated on the 2018 edition of the ImageCLEF caption challenge, obtaining the highest scores of the year.

Table 6.6 presents a list of the most frequent, where duplicated or very similar concepts were removed. In contrast with the previous data set (Table 6.3), this data set included over a hundred thousand different concepts, and exhibited a significant number of noisy labels.

This year's task was accompanied with two data sets containing various images from biomedical literature: the *training set*, comprising 223,859 images, included the list of concepts from the UMLS dictionary associated to each image. The *testing set*, composed of 9,938 images, had its annotations hidden from the participants.

We have addressed the concept detection task in two phases. First, mid-level representations of the images were chosen and built;

- as a classical approach, bags of visual words were used as image descriptors, obtained from the clustering of visual key-points;
- two kinds of deep neural networks for unsupervised feature learning were designed and trained for the purpose of visual feature extraction. Both networks were a hybrid form of GAN with an auto-encoding mechanism.

Afterwards, the concept detection problem were treated as a multi-label classification

Table 6.6: Most frequent concepts in the training set of ImageCLEF 2018 caption.

term	count	description
C1706368	77,003	And
C1696103	20,164	image-dosage form
C1837463	19,491	narrow face (physical finding)
C1546708	19,253	marrow
C0771936	19,079	yarrow flower
C1704653	17,527	cell device component
C0043194	14,079	Wiskott-Aldrich syndrome
C0015726	12,991	scared
C0040405	12,530	X-ray computed tomography
C0423899	12,424	intelligence above average
C1550655	12,350	patient
C1261259	12,217	Wright stain
C0523207	11,853	Hematoxylin and Eosin
C2087366	11,539	left
C1306645	10,390	plain X-ray

problem: the three representations were validated by training classifiers of low complexity over the new representations.

6.3.1 Feature Learning

Bags of Visual Words

After converting the images to grayscale, without resizing, visual key-point descriptors were extracted using ORB [85]. The implementation in OpenCV [100] was used for ORB key-point extraction and descriptor computation. Each image would yield a variable number of descriptors of size 64. The remaining steps are equivalent to the steps described in Section 6.2.1.

Adversarial Auto-encoding networks for feature learning

The following sections describe two GAN-based deep neural network models for the unsupervised extraction of visual features from biomedical images. Both architectures will result in an *encoder* of samples into a latent code z, which is subsequently used as a global descriptor.

The networks presented next abide to similar specifications: encoders and discriminators were built according to Table 6.7. The *code discriminator* is an exception to these two forms, and is instead specified inline with the description of the Adversarial Auto-encoder. Both architectures are composed of a sequence of blocks of convolutional layers, where each are followed by a normalization procedure and leaky ReLU activations with a leakiness factor of 0.2. All convolutional layers relied on a kernel of size 3x3. The encoding network ends

Table 6.7: A tabular representation of the sequential composition of the encoders and discriminators in the networks.

layer	kernel size/stride	output shape
conv($nb = 64$, LReLU)	3x3 /2	32x32x64
conv($nb = 128$, LReLU)	3x3 /1	32x32x128
conv($nb = 128$, LReLU)	3x3 /2	16x16x128
conv($nb = 256$, LReLU)	3x3 /1	16x16x256
conv($nb = 256$, LReLU)	3x3 /2	8x8x256
conv($nb = 512$, LReLU)	3x3 /1	8x8x512
conv($nb = 512$, LReLU)	3x3 /2	4x4x256
conv($nb = 512$, LReLU)	3x3 /1	4x4x512
flatten	N/A	8192
fc(nb=h)	N/A	h

with a fully connected layer, where h is equal to the size of z for the encoder, and 1 for the discriminator.

The decoding network, which is used for decoders and generators, replicate the encoding process in inverse order (Table 6.8). It starts with a mapping of the prior (or latent) code features to a feature space of dimensions 4x4x512 using a fully connected network. Each 2-layer block is composed by a convolutional layer, followed by a transposed convolution of stride 2 (also called *fractionally-strided convolution*). Each block doubles the height and width of the output, as a consequence of the strided convolution, until the intended image output dimensions are reached. The last convolution maps these activations to the RGB pixel value domain.

Seeing the success of the Adversarial Auto-encoder in Section 6.2.2, the same kind of model was again included in this experiment for feature learning. Instead of convolutional layers, the code discriminator is composed of three fully connected networks of 1,024 neurons each, with layer normalization [194] and Leaky Rectified Linear Unit (LReLU) after each one. The fourth layer, as in the remaining discriminators, ends with a single output neuron. Additionally, the bottleneck vector is regularized with an additional code discriminator network.

When the mappings of the AAE are inverted, this results in the Flipped Adversarial Auto-encoder (F-AAE) [200]. In this architecture, a basic GAN is augmented with an encoding network $E(x') = z'$, which learns to reconstruct the original prior code leading to the given generated sample (Figure 6.9). As a means of stabilizing the the GAN training process, the architecture was adjusted to handle two levels of samples at different levels of the network. The generator produces two images x' and x'_{small}, the latter four times smaller in side (16x16), whereas the discriminator receives both images for discriminating the sample as a whole. The

Table 6.8: A tabular representation of the sequential composition of the decoders and generators in the networks.

layer	kernel size/stride	output shape
fc($nb = 8192$, LRelU)	N/A	8192
reshape(4x4)	N/A	4x4x512
conv($nb = 512$, LRelU)	3x3 /1	4x4x512
dconv($nb = 512$, LRelU)	3x3 /2	8x8x512
conv($nb = 256$, LReLU)	3x3 /1	8x8x256
dconv($nb = 256$, LReLU)	3x3 /2	16x16x256
conv($nb = 128$, LReLU)	3x3 /1	16x16x128
dconv($nb = 128$, LReLU)	3x3 /2	32x32x128
conv($nb = 64$, LReLU)	3x3 /1	32x32x64
dconv($nb = 64$, LReLU)	3x3 /2	64x64x64
conv($nb = 32$, LReLU)	3x3 /1	64x64x64
conv($nb = 3$, tanh)	1x1 /1	64x64x3

adversarial training formula is equivalent to the original GAN's:

$$V(G, D) = \min_G \max_D \mathbb{E}_{\mathrm{x} \sim p_{\mathrm{x}}}[\log D(\mathrm{x})] + \mathbb{E}_{\mathrm{z} \sim p(\mathrm{z})}[\log 1 - D(G(\mathrm{z}))]. \tag{6.6}$$

This two-level GAN is influenced by the progressive GAN [201]. After two convolutional blocks of the generator, a 1x1/1 convolution of 3 kernels is used to convert the feature maps into a smaller RGB image. The network progresses as usual to the other two blocks in order to produce the 64x64 image. At the two-level discriminator, the smaller image is mapped to a 64-channel feature map with another 1x1 convolution, and concatenated with the features after the second convolutional block. Unlike the progressive GAN, the two levels are generated and updated simultaneously.

This architecture employs the concept of latent code regression described at the end of Section 6.1.2, interleaved with the GAN training process. The encoder never gets to see real data, and the generated samples, while potentially making great approximations of the intended distribution, will usually not achieve a perfect result. The application of the F-AAE in this task was not envisioned as a potentially competitive solution to feature learning, but rather as a means to obtain quantitative results in contrast to the AAE.

Image Preprocessing and Augmentation

Training samples were preprocessed in the following fashion: images were resized to the square resolution of 96 pixels. Afterwards, each incoming sample was cropped to a random square of size 64x64, yielding the final *real* samples. Images in the testing set, on the other hand, were only resized to fit the 64x64 dimensions. For all cases, the images' pixel RGB

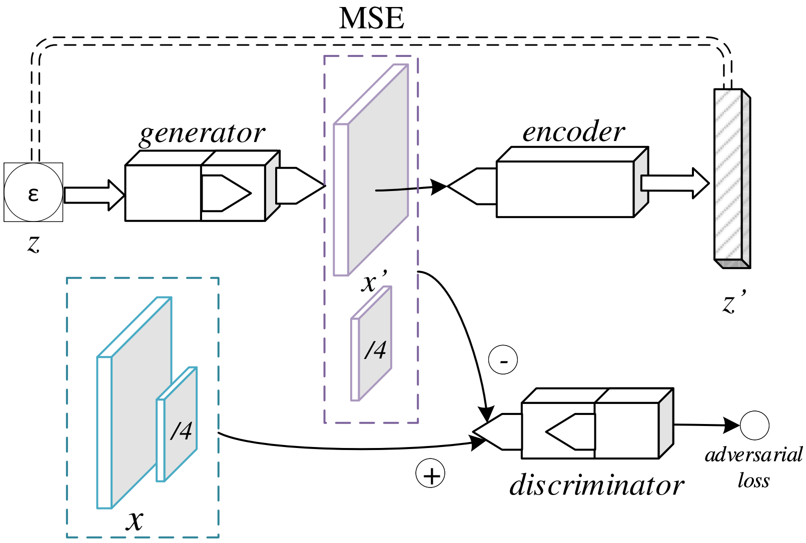


Figure 6.9: Architecture of the flipped adversarial auto-encoder with two image levels.

values were normalized with the formula $\mathrm{n}(v) = {}^{v}/_{127.5} - 1$, thus sitting in the range [-1, 1].

Network Training Details

The GANs were trained with sequential 3-step iterations of stochastic gradient descent: (1) reconstruction, (2) generator training, and (3) discriminator training. The prior codes were sampled from a 1,024-dimensional surface of a hypersphere. The parameters were initialized with a random Gaussian distribution with a standard deviation of 0.002. Training took place through stochastic gradient descent, using the Adam optimizer [164] with a learning rate of 10^{-4} and the hyperparameters $\beta_1 = 0.5$ and $\beta_2 = 0.99$ for all three optimization steps. The AAE was trained for 140,000 iterations, the F-AAE for 210,000 iterations, with a mini-batch size of 32. This is approximately 20 epochs over the training data for the AAE and 30 epochs for the F-AAE.

Depending on its purpose, specific regularizers and stabilization methods were also added to the network.

- In the (convolutional) discriminator, batch normalization [162] was employed after every convolutional layer of the encoding blocks. As a technique devised in [201] to prevent mode collapse, the across-minibatch standard deviation of the last convolutional layer activations was injected into the same activations as an additional feature map, yielding a new tensor of output 4x4x513 (and a flattened layer of 8,208 features). Moreover, we included an auxiliary loss component $0.001 \times \mathbb{E}[D(x)^2]$ to prevent the output from

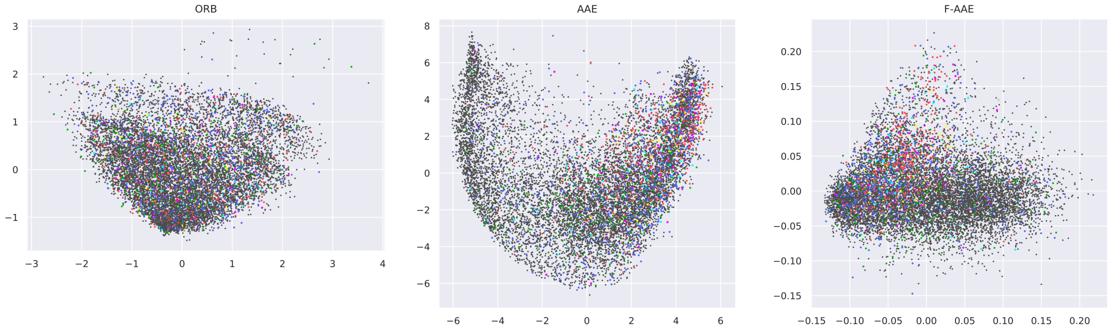


Figure 6.10: The 2D projections of the features with PCA, for each learned feature space.

drifting too far away from zero [201].

- In the final activations of the encoder, a regularization loss was employed so that the codes would approach a unit norm, which is an invariant in the hypersphere distribution: $\epsilon_{sphere} \times |\|z\|_2 - 1|$, where ϵ_{sphere} was set to the constant 0.001.
- In the generators of samples, pixelwise feature vector normalization [201] was added immediately after every convolutional layer in the decoding blocks (before the leaky ReLU): $b_{x,y} = a_{x,y} / \sqrt{\frac{1}{N}\sum_{j=0}^{N-1}(a^j_{x,y})^2 + \epsilon}$, where $\epsilon = 10^{-8}$.
- The introduction of some noise in GANs is known to stabilize the training process and contribute to an easier convergence [202]. We added drop-out layers [55] with a 50% drop rate at the second to last layer of the discriminator, and after the fully connected layer in the generator (25% drop rate).

TensorFlow [59] with GPU support was used to train both neural networks, as well as to retrieve the final features of each image in the two data sets.

6.3.2 Qualitative Feature Analysis

In a similar fashion to prior art, an attempt at visualizing the features of each model follows. Each representation learning approach described in this work resulted in a 1,024-dimensional feature space. A stratified portion of 5% of the training set was taken, and their respective PCA 2D embeddings were built. The visualizations are presented in Figure 6.10. The three primary colors were used to label the points with, respectively, the concepts `C1696103` (*image-dosage form*), `C0040405` (*X-ray CT*), and `C0221198` (*lesion*), each painted additively.

The projection obtained using PCA in this case is not very conclusive, as there is no clear feature disentanglement that would contribute to an aggregation of similar samples (as in, with more labels in common). As an attempt to obtain more powerful visualizations (colloquially speaking), we trained another dimensionality reduction algorithm, Uniform Manifold Approximation and Projection (UMAP) [203]. It is an graph-based non-linear

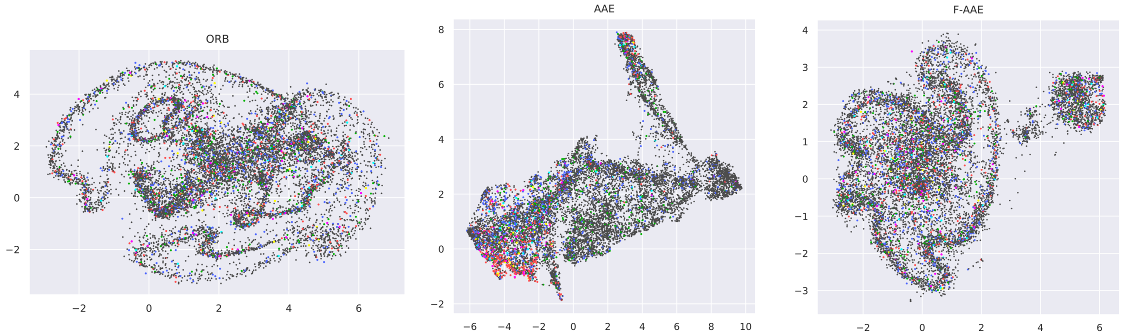


Figure 6.11: The 2D projections of the features with UMAP, for each learned feature space.

manifold learning technique which attempts to preserve the structure among neighboring data samples in a lower dimension, and can perform effectively even under a very high data dimensionality. An open implementation of UMAP is also available[3]. For this particular task, the algorithm was configured with 15 as the number of neighbors (`n_neighbors`) and 0.01 as the minimum distance (`min_dist`). The samples were then mapped to this two-dimensional embedding, as presented in Figure 6.11. The second set of plots, while still quite noisy, reveal a better mode discrimination in the features from the AAE, as it managed to aggregate a significant portion of samples with the label `C1696103` (in red) in the same region of the manifold. This effect is not as noticeable in the other feature spaces learned.

The original concept of GAN shows ground-breaking results in the field of data synthesis: as the generator learns to create realistic samples (in this case, images in the biomedical domain), retrieving new content from a trained generator can be done by feeding it with prior codes. The F-AAE benefits from this perk, as it is an extension to the original GAN. The AAE, on the other hand, presents blurrier images as a consequence of mean squared error being used as the reconstruction loss. This can be seen in Figure 6.12, where a few samples were retrieved nearly at the end of training.

Other patterns of the learned representation can be obtained with an observation of samples in a close manifold. Figure 6.13 shows one generated sample of the trained F-AAE which has been moved in its own latent space around one of the circles of the hypersphere. This exploration of the feature space does not ensure that the images stay in the same modality or retain semantic concepts, but unlike an interpolation in pixel space, intermediate positions still exhibit sharp visual features, and can be considered as “real-looking” as the starting image, within the capabilities of this GAN.

---

[3] https://github.com/lmcinnes/umap

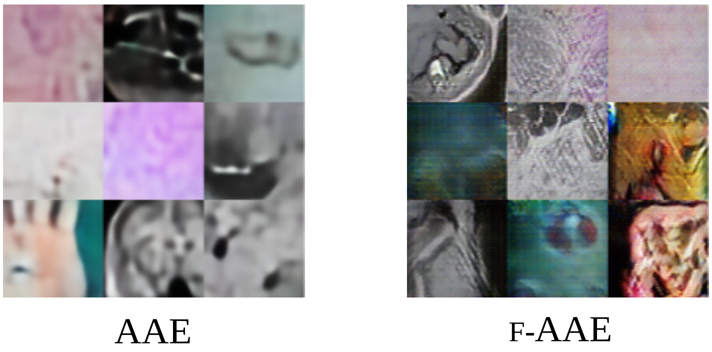


Figure 6.12: An example of samples produced by the AAE and the F-AAE during the training process.

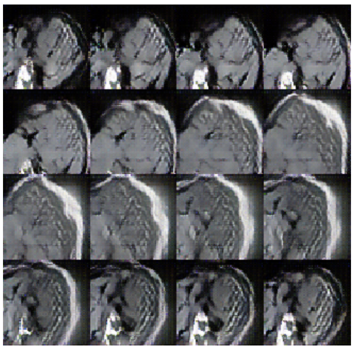


Figure 6.13: A sequence of generated samples from the F-AAE by moving the prior code around a circle in the hypersphere, column-first (from left to right, then top to bottom).

6.3.3 Multi-label Classification

For each of the three representations learned, a multi-label classifier was applied to the resulting features, where the concepts of the image associated to the feature vector were treated as labels. With the purpose of evaluating the descriptiveness of these features, algorithms of low complexity were chosen. We experimented with logistic regression and a variant of k-nearest neighbors.

Logistic regression

After mapping all images in the training and testing sets into the latent feature space, the training set was split to create a fine-tuning (or validation) set containing approximately 10% of the data points. Afterwards, linear classifiers were built for multi-label classification of biomedical concepts. As the number of possibly existing concepts in the images is very high, only the 500 terms most frequently occurring in the training set were considered. The label vectors were built based on a direct mapping from the UMLS term identifier to an index in the vector. The reverse mapping was kept for producing the textual list of concepts.

The linear classifiers were trained for each of the three representations, using Follow-the-regularized-leader (FTRL)-Proximal optimization [196] with a base learning rate of 0.05, L_1- and L_2-norm of 0.01, and a batch size of 64. After each epoch, the classifiers were evaluated based on their precision, recall, and mean F_1 score averaged against the samples in the separate fine-tuning set, with respect to multiple fixed operating point thresholds: 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, and 0.2. Classifiers were trained until the best score among these thresholds would no longer improve, and the model with the maximizing threshold was then used to predict the concepts in the testing set. These procedures were implemented in TensorFlow.

After the classifiers were experimented for all three representations, an attempt to recognize more concepts was made with the AAE representation: a set of new linear classifiers were trained, as above, but for the following 1,000 most frequent concepts (after the 500), with even lower thresholds: 0.01, 0.0124, 0.025, 0.0375, 0.05, 0.075, and 0.1. The same procedure was performed again, for the 1,000 most frequent concepts following the other 1,500. The three prediction files were merged with a concatenation of the concept lists to form this final submission.

Similarity search

In this classification method, the images' features were placed in a dense feature index, and the k most similar points were used to determine which concepts are present in each sample. While the use of similarity searching techniques is a naive approach to classification, it enables a rough assessment of whether the representation would fare well in retrieval tasks

Table 6.9: List of submitted runs from the UA.PT Bioinformatics research group to the concept detection challenge.

Rank	Run file name	Kind	Classifier	**Test** F_1
1	aae-500-o0-2018-04-30_1217	AAE	linear(500)	0.1102
2	aae-2500-merge-2018-04-30_1812	AAE	linear(2500)	0.1082
3	lin-orb-500-o0-2018-04-30_1142	ORB	linear(500)	0.0978
9	faae-500-o0-2018-04-27_1744	F-AAE	linear(500)	0.0825
11	knn-ip-aae-train-2018-04-27_1259	AAE	k-NN(cosine)	0.0570
12	knn-aae-all-2018-04-26_1233	AAE	k-NN(L^2)	0.0559
19	knn-orb-all-2018-04-24_1620	ORB	k-NN(L^2)	0.0314
21	knn-ip-faae-all-2018-04-27_1512	F-AAE	k-NN(cosine)	0.0280
22	knn-faae-all-2018-04-26_0933	F-AAE	k-NN(L^2)	0.0272

where similarity metrics were not previously learned, which is the case for the Euclidean distance between features.

Like in logistic regression, the training set was split into two portions 90%/10%, leaving the latter for fine tuning and validation. The Euclidean (L_2) distance was mainly used for determining the similarity between data points. However, as an attempt to exploit the hyperspherical vector space in the features emerging from the AAE and the F-AAE, we have also tested cosine similarity for these representations. For this particular metric, features were linearly normalized to unit norm, so that the internal product could be employed by the index. Faiss [197] was used for the vector similarity search.

A modified form of the k-nearest neighbors (k-NN) algorithm was used: a positive prediction for a given label was made when at least one of the captured neighbors is positive for that label, and false otherwise. This tweak makes the algorithm much more sensitive to very sparse classification labels. The hyperparameter k was exhaustively searched in $1, 2, 3, 4, 5$ to yield the highest F_1 score against the validation set. Analogous to the threshold in logistic regression, the optimal k was used to build the predictions on the testing set.

6.4 Results and Discussion

The methods described were combined into a total of nine official submissions from the team, as listed in Table 6.9. Additional details regarding each run are provided further below, as separate tables. *Rank* is the final position out of all graded runs from this year's participations in the task. The Kind of representation and classifier type is specified. *Kind* corresponds to the feature extractor used. *Classifier* is either *linear(C)* for logistic regression on the C most frequent concepts, or *k-NN(M)* for similarity search with the given metric. All F_1 scores stated are the micro F_1 scores of each sample averaged across the corresponding set (validation or testing).

Table 6.10: Results obtained from the submissions to the ImageCLEF 2018 concept detection task with logistic regression.

Kind	Target Concepts	Optimal Thres.	Val. Precision	Val. Recall	Val. F_1	**Test** F_1
AAE	500	0.1	0.216225	0.291748	0.248372	0.110176
AAE	2,500	-	-	-	-	0.108229
	1,000	0.05	0.107612	0.100736	0.104060	-
	1,000	0.025	0.079174	0.089273	0.083921	-
ORB	500	0.1	0.213660	0.254010	0.232094	0.097769
F-AAE	500	0.075	0.182600	0.147899	0.163428	0.082475

Table 6.10 shows the validation metrics and accompanying final score of the runs based on logistic regression, including which of the tested thresholds yielded the highest score. The *validation F_1 score*, which was obtained from evaluating the model against the validation set, only assumes the existence of the respective target concepts (i.e. the 500 most frequent in most runs). Nonetheless, these metrics were assumed to be acceptable for an objective comparison among local runs, and have indeed defined the same ranking order as the final *Test F_1 score*.

The adversarial auto-encoder resulted in better quality features than the ORB bags of words or the flipped adversarial auto-encoder. Although the optimal thresholds were low, representations with more descriptive power required less threshold lowering. The second row shows the final outcome of merging the three portions of the multi-label classifiers together (500 + 1,000 + 1,000). Locally, they were evaluated independently. The extended label set classifiers in this case, albeit covering more biomedical concepts, were less informative on the rarer labels, which ended up crippling the final score.

The work by Lipton, Elkan, and Narayanaswamy [204] provide some relevant insights on classifiers aiming to maximize F_1 score. First, that the optimal threshold in a probabilistic classifier which maximizes the score s will be $\max \mathbb{E}_{p(y|s)}[\frac{F_1}{2}]$. Notably, the thresholds identified in our experiments would usually be not far from this optimal threshold, yielding a slightly lower score than half the obtained validation F_1 score. The same work presents an algorithm to further fine-tune these thresholds to attain a more ideal operating point. However, given the higher risks of biasing the thresholds and uninformative classifier observation (from the difficuly of classifying rare concepts), we chose to avoid overfitting the validation set by selecting a few thresholds within the interval known to contain the optimal threshold.

With the k-nearest neighbors algorithm, five runs were submitted Table 6.11. It is understandable that logistic regression has a strong vantage point over this method, since it learns a linear vector space that is more ideal for classification, whereas k-NN is restricted to a very limited set of fixed metrics. The resulting validation scores were significantly lower, but were closer to the final score against the testing set. It is also worth emphasizing that cosine similarity over the features of the AAE and the F-AAE resulted in slightly better

Table 6.11: Results obtained from the submissions to the ImageCLEF 2018 concept detection task with similarity search.

Kind	Metric	k	Val. Precision	Val. Recall	Val. F_1	**Test** F_1
AAE	cosine	2	0.065085	0.120610	0.073711	0.056958
AAE	L^2	2	0.062775	0.116644	0.071168	0.055936
ORB	L^2	4	0.023813	0.080327	0.032593	0.031376
F-AAE	cosine	3	0.065085	0.069767	0.031575	0.027978
F-AAE	L^2	3	0.021505	0.062078	0.028007	0.027188

metrics, meaning that normalizing all representations to unit norm was beneficial.

The low performance obtained with the F-AAE in both classification algorithms was theoretically justified when describing the nature of this model in Section 6.3.1. However, these results were as low as to suggest that the model would still greatly benefit from better training. From observing generated samples side by side with real images, we have noticed that certain image categories were very underrepresented to non-existent in the generated samples. Mode collapse is one of the major issues that may arise in GAN training, and is still heavily tackled in recent literature.

Although our team has attained the highest mean F_1 scores for this edition of the task, the potential of other methods presented by the remaining participating teams is not to be disregarded. The ImageSem team also exhibited interesting methods and results [205]. A similarity metric between concepts was employed to produce a set of 459 clusters containing similar concepts, named “representation CUIs”. Their best F_1 score, of 0.092849, was obtained with a pre-trained *Inception-v3* network, fine-tuned for classifying images into ten of the representation CUIs most likely for the image to have. ImageSem also experimented a retrieval approach, using LIRE [206] for feature extraction and search similarity, coupled with Latent Dirichlet Allocation (LDA) for selecting the concepts from the images retrieved by the query image. The IPL team, in a similar fashion to the previous year, employed a variation of k-nearest neighbors classification after extracting dense SIFT BoVWs and generalized quad-tree-based bags of colors (named QBoC) from the images, obtaining a score of 0.0509 [207]. After fusing both feature sets, the concepts of an image are predicted based on a score defined by the distance between other samples and previously defined semantic groups for the concepts.

We highlight that certain techniques presented in these participations could be applied in our methods without significant changes to the pipeline. Namely, grouping concepts into representation CUIs can be very effective against the presence of redundant concepts, as was the case in this data set, and can significantly narrow down the problem of multi-label classification to a smaller set of labels, at the potential risk of grouping some concepts that should be independently discriminated. Multi-label classification by taking the top N "most likely" labels makes an inaccurate assumption that images have approximately the same number of concepts. When considering the full list of concepts present in the training set, we calculated a standard deviation of 496.5 in the number of concepts per sample. A multi-label classification through logistic regression and adjusted thresholds can be made on top of representation CUIs instead, thus better adaptating to a variable number of positive labels. In the similarity search technique, any of the image retrieval techniques from the ImageSem team and the k-NN algorithm from the IPL team could be used as an alternative to the search method presented above in Section 6.3.3.

6.5 Final Remarks

This chapter presents representation learning as a useful source of medical image description algorithms. It takes unsupervised representation learning techniques from state-of-the-art, while facing them against a more traditional bags of visual words approach. The methods were evaluated with the biomedical concept detection task of the ImageCLEF 2017 and 2018 editions of caption prediction challenge. We have tested the hypothesis that a powerful image descriptor can contribute to efficient concept detection with some level of certainty, without observing the original image. Results are presented for six different approaches, where two of them rely on visual key-point extraction and description algorithms, and other two of them are based on generative adversarial networks. Overall, these methods have significantly outperformed our previous participation and are comparable with other techniques in the challenge.

As identified in previous work and hereby ensured, it is possible to obtain high quality representations with modern deep learning approaches, in contrast with previously popular computer vision methods such as the SIFT bags of visual words.

Deep neural networks are known to be computationally intensive to train, but the recent breakthroughs and open technologies made these methods more accessible and practical. The convergence of an auto-encoder is relatively easy to attain, as the reconstruction loss constitutes a single major objective that can be optimized to a plateau. This is not the case for GANs, the nature of which imposes two adversarial objectives, thus bringing additional concerns for training stability. In particular, the results obtained with the BiGANs suggest that the model would benefit from better training. GANs can empirically provide good results, but the additional complexity, the difficulty of convergence, and the possibility of mode collapse can significantly cripple their performance in representation learning. Nonetheless, these issues are already a high focus of attentions at this time, and will likely lead to substantial improvements in GAN design and training.

It is also important that these approaches are augmented with non-visual information. In particular, a medical imaging archive should take the most advantage of the available data beyond pixel data.

This chapter presents the methods used to obtain unsupervised representations for medical images in literature. We show that the use of deep learning methods can surpass more traditional representations, such as the bags of visual words, in terms of descriptive power. These representations were evaluated by treating the concept detection as a multi-label classification problem, and attained a best mean F_1 score of 0.1102 with logistic regression, ranking first in the 2018 edition of the concept detection task. A score of 0.0570 was also attained with parameterless vector searching alone. No external sources of evidence were used for any of the presented methods.

On the other hand, these results may not seem to provide a substantial jump when compared to the initial iteration of the ImageCLEF caption challenge. For instance, the best run of the 2017 concept detection task which did not rely on any external resources had a mean F_1 score of 0.1436 [199], and the use of a pre-trained neural network had achieved a score of 0.1583 [198]. Granted, the scores are not directly comparable due to a variety of factors which could influence the performance of concept detection. As a new and disjoint set of medical images and concepts, the quality of the latest data set was slightly improved with the exclusion of multi-figures, and the overall size of the training set was increased by roughly 28%. On the other hand, the number of different concept identifiers presented in the training set's ground truth increased significantly, which may also make the detection task more difficult. The sample-averaged F_1 score for evaluating these solutions is certainly preferable over the macro F_1 score, which would have skewn the scores heavily due to the excessive weight applied to the rare labels [204]. Nevertheless, we find that most concepts in the set do not have enough images with a positive label for a valuable classifier, and that the obtained performance measurements are within the expected range of scores in this task, as the classified labels are very varied and scarce.

Multiple roads to future work can be outlined from the ImageCLEF 2018 caption challenge.

- Generative adversarial networks still make a hot topic, but it is likely to bring remarkable breakthroughs in feature learning. With the emergence of promising techniques for improving the quality and training process of GANs to this purpose, they should likewise be considered for this task and potentially other similar problems.
- There is an open opportunity to learn richer representations with semi-supervised learning, by taking advantage of the concepts availble in the training set. The original paper on the adversarial auto-encoder contemplates more than one form of incorporating label information in the regularization process [183], and the currently known approaches to semi-supervised GANs are much more diverse at the time of writing [208, 209].
- We do not exclude the possibility of attaining better scores with other classification algorithms, potentially those which are more adapted to an extreme multi-label classification scenario. The decisions made in this work were based on the desire to test

the representation's descriptiveness without learning another complex feature space.

- Finally, we find of significant importance that past efforts in the ImageCLEF challenges make a smooth transition to future editions, so as to obtain more competitive baselines and enable new research groups to join in with less technical debt. The tools built for the ImageCLEF 2018 caption challenge in this work were released on GitHub as open source software[4], thus contributing to this cause.

---

[4] `https://www.github.com/bioinformatics-ua/imageclef-toolkit`

Chapter 7

Conclusion

Information systems have become part of our daily lives, in such ways, and in such an intensity, that were once unpredictable. This is no exception in healthcare, where increasingly higher magnitudes of imaging studies are produced daily and the concept of PACS has been an essential foundation towards improved tools and new use cases in medical imaging. By admitting a multimodal perspective of PACS archives, medical images that were once in a "dormant" state can be explored at a higher level than the typical capabilities of a traditional archive in plain compliance to the DICOM standard.

It is a statement of this thesis that existing systems are very likely to shift into these new paradigms in the future: users should be able to perform content-based queries at will and obtain meaningful results in the process; through the analysis of indexed medical images, an archive will support searching for keywords that would otherwise not be recorded in meta-data; and that representation learning methods will play a valuable role in the aforementioned goals, as they yield powerful visual features for Content-based Medical Image Retrieval. The final outcomes translate to enhanced computer systems in medical imaging, assisting medical staff in saving lives.

7.1 Achievements and Limitations

The following major contributions, enumerated next, can be individually reasoned about: (1) an extensible multimodal search engine for medical imaging repositories; (2) a framework for automated labeling of medical images in a PACS archive was designed and implemented, enriching the archive with additional keywords that were previously unavailable in searching; and (3) the proposal of modern feature learning methods for medical image understanding, emphasizing its potential in concept detection and CBMIR. However, the culmination of this work lies in the consolidation of all three proposals, where its integration is technically feasible through extensible medical imaging archives such as Dicoogle, and its application is worthwhile by virtue of the latest advances in information retrieval and representation learning.

In Chapter 4, a multimodal search engine was designed, developed, and integrated in an open source PACS archive. It was designed to accommodate state-of-the-art query fusion algorithms and to be easily extendable with multiple query providers, thus fulfilling a framework for clinical practice and image retrieval benchmarking scenarios. The functionalities presented at the end of the chapter are not to be disregarded, namely: (1) queries based on ROIs; (2) a more intuitive form of user-guided relevance feedback; (3) the challenges of series-level query; and (4) built-in query expansion. While the last point does not pose many concerns considering the the state-of-the-art in information retrieval, the other three are more specific to the subject of medical image retrieval, and each represent a set of additional requirements that should eventually become part of this approach.

Chapter 5 establishes an automated labeling system as an intuitive extension to the medical imaging archive. The intent of this thesis is that such a mechanism becomes a common mechanism of a PACS archive's search application. Moreover, open source software components were made available to facilitate the integration process. This ideal does not emerge without its hindrances, that will hopefully be overcome with additional lines of research. The concern of having enough classifiers to make the mechanism useful might be a pertinent one. As suggested by Shin *et al.* [210], one must give at least as much importance to collecting data sets as to designing new machine learning models. As such, crowd-sourcing initiatives to provide and annotate medical data sets should be considered in this process. The resulting classification models will need to undertake attentive validation. With machine learning models, their quality can only be as good as the quality of the training data itself: in a diagnosis department, a wide diversity of data can avoid misclassification in underrepresented cases, such as in severely deformed organs. Prediction thresholds may also have to be tweaked depending on the seriousness of false negatives or false positives. In a CADx system, it is preferable for a medic to examine multiple images or regions of an unsure detection, than to create false positives that make the expert overlook an important cue [211]. On the other hand, a classifier for a fully automated diagnosis or other important clinical findings should show both high precision and recall [147].

Chapter 6 proposes the use of modern feature learning methods for medical images. This work shows that, in a scenario of biomedical concept detection, unsupervised learning methods for feature extraction are preferable over traditional visual feature extraction algorithms. Unsupervised representation learning techniques were tested on two different data sets of the concept detection task, leading to the highest mean F_1 score of the ImageCLEF 2018 concept detection task, obtained with the features learned by an adversarial auto-encoder. As a consequence, methods based on GANs provide exciting prospects for future developments in representation learning, even when they are only used for the purpose of regularization. The low scores typically obtained from ImageCLEF caption (the highest in 2018 being 0.1102) are a consequence of the very noisy data sets in the task. A more precise evaluation of the methods may be attained in the future as the quality of the data sets improve. In the mean time, having these solutions available as open source software will enable the scientific community to evaluate these methods across future iterations of challenges in medical image understanding. Not only that, the community may wish to implement more of the deep learning techniques currently known and experiment with other classifiers on top of these representations. Considering that new variations of GANs have been published at a fast pace since the inception of the concept, the community is likely to appreciate the effort of building fair comparisons between them, under the scope of concept detection or similar tasks.

One might also question about the challenges of medical image retrieval from a user-oriented perspective, and whether this search would come across as usable by the target audience. A recent study [131] contains outcomes and conclusions which are in line with the work presented here.

1. The multimodal search engine for medical imaging archives was designed to meet certain requirements of medical image retrieval systems, in line with the set of algorithms still presented as state of the art.
2. With automated labeling, one can indeed outline a few potential use cases that may become part of a practitioner's workflow. While this solution has shown evidence of its usefulness, an empirical assessment of user experience with the use of these methods by experts makes another worthy line of research.
3. The use of feature learning instead of other feature extraction algorithms does not require changes in user experience, but would influence the quality of the results received on each query.

7.2 Future Directions

Other than the open topics raised above, this thesis reveals additional insights on what can be investigated next that would contribute to the field. One potentially relevant line of research lies on devising a standard for multimodal information retrieval. As previously mentioned in Section 4.3.5, MRML should be regarded here as prior art that did not succeed into becoming a settled standard. This thesis makes an appeal for a new attempt at building a common ground for multimedia retrieval services and protocols, with emphasis on compatibility for the World Wide Web, and preferably guided by the FAIR Data principles [212].

Future work in feature learning ought to consider semi-supervised learning as a means of building more descriptive representations from known categories and other annotations. There is an interesting translation to a typical PACS, where DICOM meta-data could be embedded into the representation learned, or just guide the training process to produce more discriminative features. Subsequently, these representations would be evaluated in a medical information retrieval scenario, as well as with other data sets in the medical imaging domain.

◆

References

[1] E. Pinho, T. M. Godinho, F. Valente, and C. Costa, “A Multimodal Search Engine for Medical Imaging Studies”, *Journal of Digital Imaging*, vol. 30, no. 1, pp. 39–48, 2017, ISSN: 1618727X. DOI: `10.1007/s10278-016-9903-z`. [Online]. Available: `http://dx.doi.org/10.1007/s10278-016-9903-z`.

[2] E. Pinho and C. Costa, “Automated Anatomic Labeling Architecture for Content Discovery in Medical Imaging Repositories”, *Journal of Medical Systems*, 2018.

[3] ——, “Unsupervised Learning for Concept Detection in Medical Images: A Comparative Analysis”, *Applied Sciences*, vol. 8, no. 8, 2018. DOI: `10.3390/app8081213`. arXiv: `1805.01803`. [Online]. Available: `http://www.mdpi.com/2076-3417/8/8/1213`.

[4] J. M. Silva, E. Pinho, E. Monteiro, J. F. Silva, and C. Costa, “Controlled searching in reversibly de-identified medical imaging archives”, *Journal of biomedical informatics*, vol. 77, pp. 81–90, 2018.

[5] E. Pinho, F. Valente, and C. Costa, “A PACS-oriented Multimodal Search Engine”, in *Computer Assisted Radiology and Surgery*, Heidelberg, Germany, 2016, S12.

[6] E. Pinho and C. Costa, “Extensible Architecture for Multimodal Information Retrieval in Medical Imaging Archives”, in *Signal Image Technology & Internet Based Systems*, K. Yetongnon, A. Dipanda, R. Chbeir, G. D. Pietro, and L. Gallo, Eds., Naples, Italy: IEEE Comput. Soc., 2016, pp. 316–322.

[7] E. Pinho, J. F. Silva, J. M. Silva, and C. Costa, “Towards Representation Learning for Biomedical Concept Detection in Medical Images: UA. PT Bioinformatics in ImageCLEF 2017”, in *Working Notes of Conference and Labs of the Evaluation Forum*, Dublin, Ireland: CEUR-WS.org, 2017. [Online]. Available: `http://ceur-ws.org/Vol-1866/paper%7B%5C_%7D149.pdf`.

[8] J. F. Silva, J. M. Silva, E. Pinho, and C. Costa, “3D-CNN in drug resistance detection and tuberculosis classification”, in *CLEF2017 Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS. org< http://ceur-ws. org>(September 11-14 2017)*, 2017.

[9] E. Pinho and C. Costa, “Comparative analysis of unsupervised representation learning methods for concept detection in medical images”, in *Computer Assisted Radiology and Surgery*, Berlin, Germany, 2018.

[10] ——, “Feature Learning with Adversarial Networks for Concept Detection in Medical Images: UA.PT Bioinformatics at ImageCLEF 2018”, in *CLEF2018 Working Notes*, Avignon, France: CEUR-WS.org $<$http://ceur-ws.org$>$, 2018.

[11] J. M. Silva, A. Guerra, J. F. Silva, E. Pinho, and C. Costa, “Face De-Identification Service for Neuroimaging Volumes”, in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, Karlstad, Sweden: IEEE, 2018.

[12] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, “A review of content-based image retrieval systems in medical applications—clinical benefits and future directions”, *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2004, ISSN: 1386-5056. DOI: `http://dx.doi.org/10.1016/j.ijmedinf.2003.11.024`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1386505603002119`.

[13] C. A. Roobottom, G. Mitchell, and G. Morgan-Hughes, “Radiation-reduction strategies in cardiac computed tomographic angiography”, *Clinical radiology*, vol. 65, no. 11, pp. 859–867, 2010.

[14] B. Myers, *U.S. medical imaging informatics industry reconnects with growth in the enterprise image archiving market*, 2012. [Online]. Available: `http://www.frost.com/prod/servlet/press-release.pag?docid=268728701` (visited on 05/22/2015).

[15] C.-Y. Wu, H.-Y. Hu, L. Chen, N. Huang, Y.-J. Chou, *et al.*, “Investigating the utilization of radiological services by physician patients: a population-based cohort study in Taiwan”, *BMC health services research*, vol. 13, no. 1, p. 284, 2013.

[16] D. W. Lee and F. Levy, “The sharp slowdown in growth of medical imaging: An early analysis suggests combination of policies was the cause”, *Health Affairs*, vol. 31, no. 8, pp. 1876–1884, 2012, PMID: 22842655. DOI: `10.1377/hlthaff.2011.1034`. eprint: `https://doi.org/10.1377/hlthaff.2011.1034`. [Online]. Available: `https://doi.org/10.1377/hlthaff.2011.1034`.

[17] A. B. Shinagare, I. K. Ip, S. K. Abbett, R. Hanson, S. E. Seltzer, *et al.*, “Inpatient imaging utilization: Trends of the past decade”, *American Journal of Roentgenology*, vol. 202, no. 3, W277–W283, 2014, ISSN: 0361803X. DOI: `10.2214/AJR.13.10986`.

[18] J. Kim, A. Kumar, T. W. Cai, and D. D. Feng, “Multi-Modal Content Based Image Retrieval in Healthcare: Current Applications and Future Challenges”, *New Technologies for Advancing Healthcare and Clinical Practices*, pp. 44–59, 2011.

[19] H. K. Huang, “Short history of PACS. part I: USA.”, *European journal of radiology*, vol. 78, no. 2, pp. 163–76, May 2011, ISSN: 1872-7727. DOI: `10.1016/j.ejrad.2010.05.007`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0720048X1000207X`.

[20] B. Mansoori, K. K. Erhard, and J. L. Sunshine, “Picture archiving and communication system (PACS) implementation, integration & benefits in an integrated health system”, *Academic radiology*, vol. 19, no. 2, pp. 229–35, Feb. 2012, ISSN: 1878-4046. DOI: `10.1016/j.acra.2011.11.009`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1076633211005587`.

[21] K. Häyrinen, K. Saranto, and P. Nykänen, “Definition, structure, content, use and impacts of electronic health records: A review of the research literature”, *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, 2008, ISSN: 13865056. DOI: `10.1016/j.ijmedinf.2007.09.001`.

[22] L. A. B. Silva, R. Pinho, L. S. Ribeiro, C. Costa, and J. L. Oliveira, “A centralized platform for geo-distributed PACS management”, *Journal of Digital Imaging*, vol. 27, no. 2, pp. 165–173, 2014.

[23] J. Philbin, F. Prior, and P. Nagy, “Will the next generation of PACS be sitting on a cloud?”, *Journal of Digital Imaging*, vol. 24, no. 2, pp. 179–183, 2011.

[24] L. A. B. Silva, C. Costa, and J. L. Oliveira, “A PACS archive architecture supported on cloud services”, *International journal of computer assisted radiology and surgery*, vol. 7, no. 3, pp. 349–358, 2012.

[25] L. M. Kaufman, “Data security in the world of cloud computing”, *Security & Privacy, IEEE*, vol. 7, no. 4, pp. 61–64, 2009.

[26] C. Viana-Ferreira, S. Matos, and C. Costa, “Long-term prefetching for cloud medical imaging repositories”, *Studies in Health Technology and Informatics*, vol. 210, pp. 1028–1030, 2014.

[27] E. L. Siegel and D. S. Channin, “Integrating the Healthcare Enterprise: A Primer: Part 1. Introduction 1”, *Radiographics*, vol. 21, no. 5, pp. 1339–1341, 2001.

[28] P. Seifert, “IHE Radiology Technical Framework Supplement XDS-Ib Integration Profile”, 2009.

[29] L. S. Ribeiro, C. Viana-Ferreira, J. L. Oliveira, and C. Costa, “XDS-I outsourcing proxy: ensuring confidentiality while preserving interoperability”, *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1404–1412, 2014.

[30] A. A. Twair, W. C. Torreggiani, S. M. Mahmud, N. Ramesh, and B. Hogan, “Significant savings in radiologic report turnaround time after implementation of a complete picture archiving and communication system (PACS)”, *Journal of Digital Imaging*, vol. 13, no. 4, pp. 175–177, 2000. [Online]. Available: `http://link.springer.com/article/10.1007/BF03168392`.

[31] M. Uffmann and C. Schaefer-Prokop, “Digital radiography: the balance between image quality and required radiation dose”, *European journal of radiology*, vol. 72, no. 2, pp. 202–208, 2009.

[32] L. A. B. Silva, L. Ribeiro, M. Santos, C. Costa, and J. L. Oliveira, “Screening radiation exposure for quality assurance.”, *Studies in Health Technology and Informatics*, vol. 205, pp. 622–626, 2014. [Online]. Available: `http://europepmc.org/abstract/MED/25160261`.

[33] F. H. B. Binkhuysen and E. R. Ranschaert, “Teleradiology: evolution and concepts”, *European journal of radiology*, vol. 78, no. 2, pp. 205–209, 2011.

[34] National Electrical Manufacturers Association (NEMA), *Digital imaging and communications in medicine (DICOM) standard*, Rosslyn, VA, USA, 2018. [Online]. Available: `http://www.dicomstandard.org`.

[35] ——, *Digital imaging and communications in medicine (DICOM) standard, PS3.1 – introduction and overview*, Rosslyn, VA, USA, 2018. [Online]. Available: `http://www.dicomstandard.org`.

[36] ——, *Digital imaging and communications in medicine (DICOM) standard, PS3.3 – information object definitions*, Rosslyn, VA, USA, 2018. [Online]. Available: `http://www.dicomstandard.org`.

[37] O. S. Pianykh, *Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide.* Springer Science & Business Media, 2009.

[38] National Electrical Manufacturers Association (NEMA), *Digital imaging and communications in medicine (DICOM) standard, PS3.6 – data dictionary*, Rosslyn, VA, USA, 2018. [Online]. Available: `http://www.dicomstandard.org`.

[39] R. Noumeir, “Benefits of the DICOM structured report”, *Journal of Digital Imaging*, vol. 19, no. 4, pp. 295–306, 2006, ISSN: 0897-1889. DOI: `10.1007/s10278-006-0631-7`. [Online]. Available: `http://dx.doi.org/10.1007/s10278-006-0631-7`.

[40] C. P. Langlotz, “RadLex: A new method for indexing online educational materials”, *Radiographics*, vol. 26, no. 6, pp. 1595–1597, 2006, ISSN: 0271-5333. DOI: `10.1148/rg.266065168`.

[41] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records", in *Proceedings of the 2006 ACM symposium on Applied computing*, ACM, 2006, pp. 235–239.

[42] National Electrical Manufacturers Association (NEMA), *Digital imaging and communications in medicine (DICOM) standard, PS3.4 – service class specifications*, Rosslyn, VA, USA, 2018. [Online]. Available: `http://www.dicomstandard.org`.

[43] ——, *Digital imaging and communications in medicine (DICOM) standard, PS3.18 – web services*, Rosslyn, VA, USA.

[44] F. Valente, L. A. B. Silva, T. M. Godinho, and C. Costa, "Anatomy of an Extensible Open Source PACS", *Journal of Digital Imaging*, vol. 29, no. 3, pp. 284–296, Jun. 2016, ISSN: 0897-1889. DOI: `10.1007/s10278-015-9834-0`.

[45] C. Costa, F. Freitas, M. Pereira, A. Silva, and J. L. Oliveira, "Indexing and retrieving DICOM data in disperse and unstructured archives", *International Journal of Computer Assisted Radiology and Surgery*, vol. 4, no. 1, pp. 71–77, 2009.

[46] C. Costa, C. Ferreira, L. Bastião, L. Ribeiro, A. Silva, *et al.*, "Dicoogle - an open source peer-to-peer PACS", *Journal of Digital Imaging*, vol. 24, no. 5, pp. 848–856, 2011.

[47] L. Bastiao Silva, L. Beroud, C. Costa, J. L. Oliveira, *et al.*, "Medical imaging archiving: A comparison between several NoSQL solutions", in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*, IEEE, 2014, pp. 65–68.

[48] S. Wang, W. Pavlicek, C. C. Roberts, S. G. Langer, M. Zhang, *et al.*, "An automated DICOM database capable of arbitrary data mining (including radiation dose indicators) for quality monitoring", *Journal of Digital Imaging*, vol. 24, no. 2, pp. 223–233, 2011.

[49] M. Santos, L. Bastião, C. Costa, A. Silva, and N. Rocha, "DICOM and clinical data mining in a small hospital PACS: A pilot study", in *ENTERprise Information Systems*, Springer, 2011, ch. 16, pp. 254–263. DOI: `10.4018/978-1-466-3667-5.ch016`.

[50] S. Al-Janabi, A. Huisman, and P. J. Van Diest, *Digital pathology: Current status and future perspectives*, Jul. 2012. DOI: `10.1111/j.1365-2559.2011.03814.x`. [Online]. Available: `http://doi.wiley.com/10.1111/j.1365-2559.2011.03814.x`.

[51] T. M. Godinho, R. Lebre, L. A. B. Silva, and C. Costa, "An efficient architecture to support digital pathology in standard medical imaging repositories", *Journal of biomedical informatics*, vol. 71, pp. 190–197, 2017.

[52] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review", *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[53] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning.* MIT press Cambridge, 2016, vol. 1.

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[55] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting”, *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[56] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, *et al.*, “SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <1MB Model Size”, *arXiv*, pp. 1–5, 2016, ISSN: 0302-9743. DOI: `10.1007/978-3-319-24553-9`. arXiv: `1602.07360`. [Online]. Available: `http://arxiv.org/abs/1602.07360`.

[57] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era”, *arXiv preprint arXiv:1707.02968*, vol. 1, 2017, ISSN: 15505499. DOI: `10.1109/ICCV.2017.97`. arXiv: `1707.02968`.

[58] Z. C. Lipton and J. Steinhardt, “Troubling trends in machine learning scholarship”, *arXiv preprint arXiv:1807.03341*, 2018.

[59] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”, Mar. 2016. arXiv: `1603.04467`. [Online]. Available: `http://arxiv.org/abs/1603.04467`.

[60] M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, and S. C. Strother, “Machine learning in medical imaging”, *Signal Processing Magazine, IEEE*, vol. 27, no. 4, pp. 25–38, 2010.

[61] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, “Computer-aided detection and diagnosis of breast cancer with mammography: recent advances”, *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 236–251, 2009.

[62] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval”, vol. 1, no. c, A. C.-B. E. Salas, Ed., pp. 1–18, 2009, ISSN: 13864564. DOI: `10.1109/LPT.2009.2020494`. arXiv: `05218657199780521865715`. [Online]. Available: `http://dspace.cusat.ac.in/dspace/handle/123456789/2538`.

[63] J. Pérez, M. Arenas, and C. Gutierrez, “Semantics and complexity of SPARQL”, *ACM Transactions on Database Systems (TODS)*, vol. 34, no. 3, p. 16, 2009.

[64] X. S. Zhou, S. Zillner, M. Moeller, M. Sintek, Y. Zhan, *et al.*, “Semantics and CBIR: A medical imaging perspective”, in *International Conference on Content-based Image and Video Retrieval*, ser. CIVR ’08, New York, NY, USA: ACM, 2008, pp. 571–580, ISBN: 978-1-60558-070-8. DOI: `10.1145/1386352.1386436`. [Online]. Available: `http://doi.acm.org/10.1145/1386352.1386436`.

[65] L. Hollink, G. Schreiber, J. Wielemaker, B. Wielinga, *et al.*, “Semantic annotation of image collections”, in *Knowledge capture*, 2003, pp. 41–48.

[66] W. Wei and P. M. Barnaghi, “Semantic support for medical image search and retrieval”, en, in *Biomedical Engineering 2007: Proceedings of the 5th IASTED International Conference on Biomedical Engineering*, Sep. 2007. [Online]. Available: http://epubs.surrey.ac.uk/470687/3/Barnaghi%7B%5C%%7D7B%7B%5C_%7D%7B%5C%%7D7DSemantic%20support.pdf%20http://epubs.surrey.ac.uk/470687/3/Barnaghi%7B%5C_%7DSemantic%20support.pdf.

[67] H. J. Lowe, I. Antipov, W. Hersh, and C. A. Smith, “Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation and knowledge-based retrieval.”, in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, 1998, pp. 882–886, ISBN: 1531-605X.

[68] O. Bodenreider, “The unified medical language system (UMLS): Integrating biomedical terminology”, *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[69] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, *et al.*, “The national cancer institute’s thesaurus and ontology”, *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 1, pp. 75–80, 2003.

[70] A. C. of Radiology, A. C. of Radiology. Committee on Diagnostic Coding Index, A. C. of Radiology. Committee on Diagnostic Coding Index, and Thesaurus, *Index for radiological diagnoses: including diagnostic ultrasound.* American College of Radiology, 1986.

[71] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, 2001. arXiv: 1204.6441.

[72] D. Brickley and R. V. Guha, “Resource Description Framework (RDF) Schema Specification 1.0: W3C Candidate Recommendation 27 March 2000”, 2000.

[73] S. Bechhofer, “OWL: Web ontology language”, in *Encyclopedia of Database Systems*, Springer, 2009, pp. 2008–2009.

[74] J. Palotti, A. Hanbury, H. Müller, and C. E. Kahn Jr, “How users search and what they search for in the medical domain”, *Information Retrieval Journal*, pp. 1–36, 2016.

[75] E. Bellon, M. Feron, T. Deprez, R. Reynders, and B. Van den Bosch, “Trends in PACS architecture”, *European Journal of Radiology*, vol. 78, no. 2, pp. 199–204, 2011, ISSN: 0720048X. DOI: 10.1016/j.ejrad.2010.05.025.

[76] T. L. Kunil, Shi-Kuo Chang, T. L. Kunil, *et al.*, “Pictorial data-base systems”, *Computer*, vol. 14, no. 11, pp. 13–21, 1981, ISSN: 0018-9162. DOI: 10.1109/C-M.1981.220243.

[77] N.-S. Chang and K.-S. Fu, “Query-by-pictorial-example”, *Software Engineering, IEEE Transactions on*, no. 6, pp. 519–524, 1980.

[78] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, *et al.*, "Query by image and video content: The QBIC system", *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[79] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age", *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008, ISSN: 03600300. DOI: `10.1145/1348246.1348248`.

[80] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, *et al.*, "Content-Based Image Retrieval in Radiology: Current Status and Future Directions", *Journal of Digital Imaging*, vol. 24, no. 2, pp. 208–222, 2011, ISSN: 0897-1889. DOI: `10.1007/s10278-010-9290-9`. [Online]. Available: `http://dx.doi.org/10.1007/s10278-010-9290-9`.

[81] P. Ghosh, S. Antani, L. R. Long, and G. R. Thoma, "Review of medical image retrieval systems and future directions", in *24th International Symposium on Computer-Based Medical Systems (CBMS)*, Jun. 2011, pp. 1–6, ISBN: 978-1-4577-1188-6. DOI: `10.1109/CBMS.2011.5999142`.

[82] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval", in *International Conference on Computer Vision Systems*, Springer, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 312–322. DOI: `10.1007/978-3-540-79547-6_30`. [Online]. Available: `http://link.springer.com/10.1007/978-3-540-79547-6%7B%5C_%7D30`.

[83] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[84] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features", in *European conference on computer vision*, Springer, 2006, pp. 404–417.

[85] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF", in *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2011, pp. 2564–2571.

[86] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints", in *Workshop on statistical learning in computer vision, ECCV*, Prague, vol. 1, 2004, pp. 1–2.

[87] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos", in *Ninth {IEEE} International Conference on Computer Vision*, IEEE, 2003, pp. 1470–1477.

[88] C. Wengert, M. Douze, and H. Jégou, "Bag-of-colors for improved image search", in *ACM International Conference on Multimedia*, ACM, 2011.

[89] R. A. Blechschmidt, R. Werthschutzky, and U. Lorcher, "Automated CT image evaluation of the lung: A morphology-based concept", *IEEE Transactions on Medical Imaging*, vol. 20, no. 5, pp. 434–442, May 2001, ISSN: 0278-0062. DOI: `10.1109/42.925296`.

[90] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity search: the metric space approach.* Springer Science & Business Media, 2006, vol. 32. [Online]. Available: `http://www.nmis.isti.cnr.it/amato/similarity-search-book/`.

[91] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval", *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[92] C. Beecks, M. S. Uysal, and T. Seidl, "Efficient k-nearest neighbor queries with the signature quadratic form distance", *Proceedings - International Conference on Data Engineering*, pp. 10–15, 2010, ISSN: 10844627. DOI: `10.1109/ICDEW.2010.5452772`.

[93] X. Guorong, C. Peiqi, and W. Minhui, "Bhattacharyya distance feature selection", in *International Conference on Pattern Recognition*, IEEE, vol. 2, 1996, pp. 195–199.

[94] P. Welter, T. M. Deserno, B. Fischer, R. W. Günther, and C. Spreckelsen, "Towards case-based medical learning in radiological decision making using content-based image retrieval", *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 68, 2011, ISSN: 1472-6947. DOI: `10.1186/1472-6947-11-68`. [Online]. Available: `https://doi.org/10.1186/1472-6947-11-68`.

[95] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals", *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593–601, 2001, ISSN: 0167-8655. DOI: `http://dx.doi.org/10.1016/S0167-8655(00)00118-5`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0167865500001185`.

[96] J. Kalpathy-Cramer, A. G. S. de Herrera, D. Demner-Fushman, S. Antani, S. Bedrick, *et al.*, "Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at ImageCLEF 2004–2013", *Computerized Medical Imaging and Graphics*, vol. 39, pp. 55–61, 2015.

[97] H. Müller, T. Deselaers, T. M. Deserno, J. Kalpathy–Cramer, E. Kim, *et al.*, "Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks", in *Advances in Multilingual and Multimodal Information Retrieval*, Springer, 2008, pp. 472–491.

[98] B. Ionescu, H. Müller, M. Villegas, A. G. S. de Herrera, C. Eickhoff, *et al.*, "{Overview of ImageCLEF 2018}: Challenges, Datasets and Evaluation", in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ser. Proceedings of the Ninth

International Conference of the CLEF Association (CLEF 2018), Avignon, France: {LNCS} Lecture Notes in Computer Science, Springer, 2018.

[99] F. Valente, C. Costa, and A. Silva, "Dicoogle, a PACS featuring profiled content based image retrieval", *Public Library of Science One*, vol. 8, no. 5, e61888, 2013, ISSN: 19326203. DOI: `10.1371/journal.pone.0061888`.

[100] G. Bradski *et al.*, "The OpenCV library", *Doctor Dobbs Journal*, vol. 25, no. 11, pp. 120–126, 2000.

[101] H. Muller, X. Zhou, A. Depeursinge, M. Pitkanen, J. Iavindrasana, *et al.*, "Medical Visual Information Retrieval: State of the Art and Challenges Ahead", in *IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2007, pp. 683–686, ISBN: 1424410177.

[102] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey", *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010, ISSN: 09424962. DOI: `10.1007/s00530-010-0182-0`.

[103] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, *et al.*, "Automated multi-modality image registration based on information theory", in *Information processing in medical imaging*, vol. 3, 1995, pp. 263–274.

[104] A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?", *Proceedings of the IEEE*, vol. 4, no. 96, pp. 541–547, 2008.

[105] A. Jaimes, M. Christel, S. Gilles, R. Sarukkai, and W.-Y. Ma, "Multimedia Information Retrieval: What is it, and why isn't anyone using it?", in *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, ser. MIR '05, New York, NY, USA: ACM, 2005, pp. 3–8, ISBN: 1-59593-244-5. DOI: `10.1145/1101826.1101829`. [Online]. Available: `http://doi.acm.org/10.1145/1101826.1101829`.

[106] M. U. Bokhari and F. Hasan, "Multimodal Information Retrieval: Challenges and Future Trends", *International Journal of Computer Applications*, vol. 74, no. 14, pp. 9–12, 2013.

[107] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data", *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1025–1039, 2013, ISSN: 08971889. DOI: `10.1007/s10278-013-9619-2`.

[108] T. M. Deserno, M. O. Güld, B. Plodowski, K. Spitzer, B. B. Wein, *et al.*, "Extended Query Refinement for Medical Image Retrieval", *Journal of Digital Imaging*, vol. 21, no. 3, pp. 280–289, 2008, ISSN: 0897-1889. DOI: `10.1007/s10278-007-9037-4`. [Online]. Available: `http://dx.doi.org/10.1007/s10278-007-9037-4`.

[109] R. Schaer, D. Markonis, and H. Müller, "Architecture and applications of the parallel distributed image search engine (ParaDISE)", *FoRESEE, Stuttgart, Germany*, 2014.

[110] O. A. Jiménez-del-Toro, A. Hanbury, G. Langs, A. Foncubierta-Rodrıguez, and H. Müller, "Overview of the VISCERAL retrieval benchmark 2015", 2015.

[111] C. E. Kahn Jr and C. Thao, "GoldMiner: A radiology image search engine", *American Journal of Roentgenology*, vol. 188, no. 6, pp. 1475–1478, 2007.

[112] M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, *et al.*, "BioText search engine: Beyond abstract search", *Bioinformatics*, vol. 23, no. 16, pp. 2196–2197, 2007.

[113] A. Depeursinge and H. Müller, "Fusion Techniques for Combining Textual and Visual Information Retrieval", in *ImageCLEF*, ser. The Information Retrieval Series, H. Müller, P. Clough, T. Deselaers, and B. Caputo, Eds., vol. 32, Springer Berlin Heidelberg, 2010, ch. 1, pp. 95–114, ISBN: 978-3-642-15180-4. DOI: `10.1007/978-3-642-15181-1_6`. eprint: `9907372v1`. [Online]. Available: `http://dx.doi.org/10.1007/978-3-642-15181-1_6`.

[114] T. Berber and A. Alpkoçak, "DEU at ImageCLEFmed 2009: Evaluating re–ranking and integrated retrieval model", in *Working Notes of the 2009 CLEF Workshop, Corfu, Greece (September 2009)*, 2009.

[115] T. Hofmann, "Probabilistic latent semantic indexing", in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1999, pp. 50–57.

[116] A. Foncubierta-Rodrıguez, A. de Herrera, and H. Müller, "Medical image retrieval using bag of meaningful visual words: Unsupervised visual vocabulary pruning with PLSA", in *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, ACM, 2013, pp. 75–82.

[117] Y. Cao, S. Steffey, H. Jianbiao, D. Xiao, C. Tao, *et al.*, "Medical Image Retrieval: A Multimodal Approach", *Cancer Informatics*, 2015.

[118] M. E. Renda and U. Straccia, "Web Metasearch: Rank vs. Score Based Rank Aggregation Methods", in *Proceedings of the 2003 ACM Symposium on Applied Computing*, ser. SAC '03, New York, NY, USA: ACM, 2003, pp. 841–846, ISBN: 1-58113-624-2. DOI: `10.1145/952532.952698`. [Online]. Available: `http://doi.acm.org/10.1145/952532.952698`.

[119] D. F. Hsu and I. Taksa, “Comparing rank and score combination methods for data fusion in information retrieval”, *Information Retrieval*, vol. 8, no. 3, pp. 449–480, 2005, ISSN: 13864564. DOI: `10.1007/s10791-005-6994-4`.

[120] E. A. Fox and J. A. Shaw, “Combination of multiple searches”, *NIST SPECIAL PUBLICATION SP*, p. 243, 1994.

[121] M. Jović, Y. Hatakeyama, F. Dong, and K. Hirota, “Image retrieval based on similarity score fusion from feature similarity ranking lists”, *Fuzzy Systems and Knowledge Discovery*, pp. 461–470, 2006.

[122] A. Mourão, F. Martins, and J. Magalhães, “Multimodal medical information retrieval with unsupervised rank fusion”, *Computerized Medical Imaging and Graphics*, vol. 39, pp. 35–45, 2015, ISSN: 1879-0771. DOI: `10.1016/j.compmedimag.2014.05.006`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/24909951`.

[123] J. Villena-Román, S. Lana-Serrano, and J. C. González-Cristóbal, “MIRACLE at ImageCLEFmed 2007: Merging textual and visual strategies to improve medical image retrieval”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5152 LNCS, pp. 593–596, 2008, ISSN: 03029743. DOI: `10.1007/978-3-540-85760-0-74`.

[124] J. H. Lee, “Analyses of multiple evidence combination”, in *ACM SIGIR Forum*, ACM, vol. 31, ACM, Dec. 1997, pp. 267–276, ISBN: 0-89791-836-3. DOI: `10.1145/278459.258587`. [Online]. Available: `http://dl.acm.org/citation.cfm?id=278459.258587`.

[125] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, “Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods”, in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’09, New York, NY, USA: ACM, 2009, pp. 758–759, ISBN: 978-1-60558-483-6. DOI: `10.1145/1571941.1572114`. [Online]. Available: `http://doi.acm.org/10.1145/1571941.1572114`.

[126] D. Markonis, M. Holzer, S. Dungs, A. Vargas, G. Langs, *et al.*, “A survey on visual information search behavior and requirements of radiologists”, *Methods of Information in Medicine*, vol. 51, no. 6, p. 539, 2012.

[127] O. Jimenez-del-Toro, S. Otálora, M. Atzori, and H. Müller, “Deep multimodal case–based retrieval for large histopathology datasets”, in *International Workshop on Patch-based Techniques in Medical Imaging*, Springer, 2017, pp. 149–157.

[128] A. Benczúr, I. Bíró, M. Brendel, K. Csalogány, B. Z. Daróczy, *et al.*, “Cross-modal retrieval by text and image feature biclustering”, CLEF, 2007.

[129] H. Müller, T. Pun, and D. Squire, "Learning from user behavior in image retrieval: Application of market basket analysis", *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 65–77, 2004.

[130] H. Müller, A. G. S. de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. K. Antani, *et al.*, "Overview of the ImageCLEF 2012 Medical Image Retrieval and Classification Tasks.", in *CLEF (online working notes/labs/workshop)*, 2012, pp. 1–16.

[131] D. Markonis, R. Schaer, and H. Müller, "Evaluating multimodal relevance feedback techniques for medical image retrieval", *Information Retrieval Journal*, pp. 1–13, 2016.

[132] T. Syeda-Mahmood, F. Wang, D. Beymer, A. Amir, M. Richmond, *et al.*, "AALIM: Multimodal mining for cardiac decision support", in *Computers in Cardiology*, IEEE, vol. 34, IEEE, Sep. 2007, pp. 209–212, ISBN: 978-1-4244-2533-4. DOI: `10.1109/CIC.2007.4745458`. [Online]. Available: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4745458`.

[133] H. Müller, "Medical multimedia retrieval 2.0", *Yearb Med Inform*, vol. 47, no. 1, pp. 55–63, 2008.

[134] A. Mourão and F. Martins, "NovaMedSearch: A multimodal search engine for medical case-based retrieval", in *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013, pp. 223–224.

[135] A. Hanbury and H. Müller, "Khresmoi–multimodal multilingual medical information search", *MIE village of the future*, 2012.

[136] A. Widmer, R. Schaer, D. Markonis, and H. Müller, "Gesture Interaction for Content–based Medical Image Retrieval", in *Proceedings of International Conference on Multimedia Retrieval*, ACM, 2014, p. 503.

[137] D. Markonis, R. Donner, M. Holzer, T. Schlegl, S. Dungs, *et al.*, "A Visual Information Retrieval System for Radiology Reports and the Medical Literature", in *Multimedia modeling conference*, 2014. [Online]. Available: `http://publications.hevs.ch/index.php/attachments/single/621`.

[138] D. Markonis, R. Schaer, A. García Seco de Herrera, and H. Müller, "The Parallel Distributed Image Search Engine (ParaDISE)", pp. 1–23, 2017. arXiv: `1701.05596`.

[139] M. M. Rahman, D. You, M. S. Simpson, S. K. Antani, D. Demner-Fushman, *et al.*, "Multimodal biomedical image retrieval using hierarchical classification and modality fusion", *International Journal of Multimedia Information Retrieval*, vol. 2, no. 3, pp. 159–173, 2013.

[140] M. Montague and J. A. Aslam, "Relevance Score Normalization for Metasearch", in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ser. CIKM '01, New York, NY, USA: ACM, 2001, pp. 427–433, ISBN: 1-58113-436-3. DOI: `10.1145/502585.502657`. [Online]. Available: `http://doi.acm.org/10.1145/502585.502657`.

[141] W. Müller, H. Müller, S. Marchand-Maillet, T. Pun, D. M. Squire, *et al.*, "MRML: A communication protocol for content-based image retrieval", in *International Conference on Advances in Visual Information Systems*, Springer, Springer Berlin Heidelberg, 2000, pp. 300–311, ISBN: 978-3-540-40053-0. DOI: `10.1007/3-540-40053-2_27`. [Online]. Available: `http://link.springer.com/10.1007/3-540-40053-2_27`.

[142] D. Markonis, M. Holzer, F. Baroz, R. L. R. De Castaneda, C. Boyer, *et al.*, "User-oriented evaluation of a medical image retrieval system for radiologists", *International Journal of Medical Informatics*, vol. 84, no. 10, pp. 774–783, 2015, ISSN: 18728243. DOI: `10.1016/j.ijmedinf.2015.04.003`.

[143] A. Rosenthal, A. Gabrielian, E. Engle, D. E. Hurt, S. Alexandru, *et al.*, "The TB Portals: An open-access, web-based platform for global drug-resistant tuberculosis data sharing and analysis", *Journal of Clinical Microbiology*, JCM–01 013, 2017.

[144] J. Faruque, C. F. Beaulieu, J. Rosenberg, D. L. Rubin, D. Yao, *et al.*, "Content-based image retrieval in radiology: analysis of variability in human perception of similarity", *Journal of Medical Imaging*, vol. 2, no. 2, p. 25 501, 2015.

[145] J. J. Rocchio, "Relevance feedback in information retrieval", *The SMART retrieval system: experiments in automatic document processing*, pp. 313–323, 1971.

[146] M. O. Guld, M. Kohnen, D. Keysers, H. Schubert, B. B. Wein, *et al.*, "Quality of DICOM header information for image categorization", in *SPIE*, vol. 4685, 2002, pp. 280–287.

[147] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential", *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198–211, 2007, ISSN: 08956111. DOI: `10.1016/j.compmedimag.2007.02.002`. [Online]. Available: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1955762`.

[148] A. H. T. Le, B. Liu, and H. K. Huang, "Integration of computer-aided diagnosis/detection (CAD) results in a PACS environment using CAD–PACS toolkit and DICOM SR", *International Journal of Computer Assisted Radiology and Surgery*, vol. 4, no. 4, pp. 317–329, Jun. 2009, ISSN: 1861-6410. DOI: `10.1007/s11548-009-0297-y`. [Online]. Available: `http://link.springer.com/10.1007/s11548-009-0297-y`.

[149] V. Dicken, B. Lindow, L. Bornemann, J. Drexl, A. Nikoubashman, *et al.*, “Rapid image recognition of body parts scanned in computed tomography datasets”, *International Journal of Computer Assisted Radiology and Surgery*, vol. 5, no. 5, pp. 527–535, 2010.

[150] X. Zhou, N. Kamiya, T. Kara, H. Fujita, R. Yokoyama, *et al.*, “Automated recognition of human strucure from torso CT images”, in *International Conference on Visualization, Imaging, and Image Processing*, 2004, pp. 584–589.

[151] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, *et al.*, “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest”, *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.

[152] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, *et al.*, “A survey on deep learning in medical image analysis”, *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, ISSN: 13618423. DOI: `10.1016/j.media.2017.07.005`. arXiv: `1702.05747`.

[153] H. R. Roth, C. T. Lee, H.-C. Shin, A. Seff, L. Kim, *et al.*, “Anatomy-specific classification of medical images using deep convolutional nets”, in *IEEE 12th International Symposium on Biomedical Imaging*, 2015, pp. 101–104, ISBN: VO -. DOI: `10.1109/ISBI.2015.7163826`. arXiv: `1504.04003`.

[154] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, “An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification”, *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, p. 1, 2016, ISSN: 2168-2194. DOI: `10.1109/JBHI.2016.2635663`.

[155] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum, and J. Mongan, “High-throughput classification of radiographs using deep convolutional neural networks”, *Journal of Digital Imaging*, vol. 30, no. 1, pp. 95–101, Feb. 2017, ISSN: 1618-727X. DOI: `10.1007/s10278-016-9914-9`. [Online]. Available: `http://link.springer.com/10.1007/s10278-016-9914-9`.

[156] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation”, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Springer International Publishing, 2015, pp. 234–241. DOI: `10.1007/978-3-319-24574-4_28`. [Online]. Available: `http://link.springer.com/10.1007/978-3-319-24574-4_28`.

[157] Y. Wang, Y. Qiu, T. Thai, K. Moore, H. Liu, *et al.*, “A two-step convolutional neural network based computer-aided detection scheme for automatically segmenting adipose tissue volume depicting on CT images”, *Computer Methods and Programs in Biomedicine*, vol. 144, pp. 97–104, Jun. 2017, ISSN: 0169-2607. DOI: `10.1016/J.CMPB.2017.03.017`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0169260716310744`.

[158] T. M. Lehmann, B. E. Wein, D. Keysers, M. Kohnen, and H. Schubert, "A monohierarchical multiaxial classification code for medical images in content-based retrieval", in *IEEE International Symposium on Biomedical Imaging*, IEEE, 2002, pp. 313–316.

[159] M. Riegler, K. Pogorelov, S. L. Eskeland, P. T. Schmidt, Z. Albisser, *et al.*, "From Annotation to Computer-Aided Diagnosis", *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 13, no. 3, pp. 1–26, May 2017, ISSN: 15516857. DOI: `10.1145/3079765`. [Online]. Available: `http://dl.acm.org/citation.cfm?doid=3104033.3079765`.

[160] M. K. Kundu, M. Chowdhury, and S. Das, "Interactive radiographic image retrieval system", *Computer Methods and Programs in Biomedicine*, vol. 139, pp. 209–220, 2017, ISSN: 0169-2607. DOI: `https://doi.org/10.1016/j.cmpb.2016.10.023`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S0169260716301766`.

[161] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, *et al.*, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015", *Medical physics*, vol. 44, no. 5, pp. 2020–2036, 2017.

[162] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *arXiv preprint arXiv:1502.03167*, pp. 1–11, 2015, ISSN: 0717-6163. DOI: `10.1007/s13398-014-0173-7.2`. arXiv: `1502.03167`. [Online]. Available: `http://arxiv.org/abs/1502.03167`.

[163] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[164] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization", in *International Conference on Learning Representations*, 2015. [Online]. Available: `https://arxiv.org/pdf/1412.6980.pdf`.

[165] L. A. B. Silva, C. Costa, and J. L. Oliveira, "Semantic search over DICOM repositories", in *IEEE International Conference on Healthcare Informatics*, Sep. 2014, pp. 238–246, ISBN: 978-1-4799-5701-9. DOI: `10.1109/ICHI.2014.41`. [Online]. Available: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7052496`.

[166] C. Kurtz, A. Depeursinge, S. Napel, C. F. Beaulieu, and D. L. Rubin, "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications.", *Medical Image Analysis*, vol. 18, no. 7, pp. 1082–100, Oct. 2014, ISSN: 1361-8423. DOI: `10.1016/j.media.2014.06.009`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1361841514001030`.

[167] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013, ISSN: 0162-8828. DOI: `10.1109/TPAMI.2013.50`.

[168] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, *et al.*, “Deep learning for health informatics”, *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017, ISSN: 2168-2194. DOI: `10.1109/JBHI.2016.2636665`. [Online]. Available: `http://ieeexplore.ieee.org/document/7801947/`.

[169] L. Van Der Maaten, E. Postma, and J. Van Den Herik, “Dimensionality Reduction : A Comparative Review”, *October*, vol. 10, pp. 1–35, 2009, ISSN: 0169328X. DOI: `10.1080/13506280444000102`. [Online]. Available: `http://www.uvt.nl/ticc`.

[170] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications”, *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[171] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms”, *Advances in Neural Information Processing Systems*, vol. 19, no. 2, pp. 801–808, 2006, ISSN: 10495258. DOI: `10.1.1.69.2112`. arXiv: `arXiv:1506.03733v1`.

[172] X. Wangming, W. Jin, L. Xinhai, Z. Lei, and S. Gang, “Application of Image SIFT Features to the Context of CBIR”, in *International Conference on Computer Science and Software Engineering*, IEEE, 2008, pp. 552–555, ISBN: 978-0-7695-3336-0. DOI: `10.1109/CSSE.2008.1230`. [Online]. Available: `http://ieeexplore.ieee.org/document/4722680/`.

[173] I. Dimitrovski, D. Kocev, I. Kitanovski, S. Loskovska, and S. Džeroski, “Improved medical image modality classification using a combination of visual and textual features”, *Computerized Medical Imaging and Graphics*, vol. 39, pp. 14–26, 2015.

[174] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks”, in *Advances in neural information processing systems*, 2007, pp. 153–160.

[175] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-a. Manzagol, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion”, *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[176] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes”, pp. 1–14, 2014, ISSN: 0004-6361. DOI: `10.1051/0004-6361/201527329`. arXiv: `1703.06211`. [Online]. Available: `http://arxiv.org/abs/1703.06211`.

[177] I. J. Goodfellow, J. Pouget-abadie, M. Mirza, B. Xu, D. Warde-farley, *et al.*, “Generative Adversarial Nets”, pp. 1–9, 2014. arXiv: `arXiv:1406.2661v1`.

[178] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, *et al.*, "Improved Techniques for Training GANs", Jun. 2016. arXiv: `1606.03498`. [Online]. Available: `http://arxiv.org/abs/1606.03498`.

[179] L. Mescheder, A. Geiger, and S. Nowozin, "Which Training Methods for GANs do actually Converge?", in *International Conference on Machine learning (ICML)*, 2018.

[180] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning", *arXiv preprint arXiv:1605.09782*, May 2016. arXiv: `1605.09782`. [Online]. Available: `http://arxiv.org/abs/1605.09782`.

[181] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", *arXiv preprint arXiv:1511.06434*, pp. 1–16, 2016. arXiv: `arXiv:1511.06434v2`. [Online]. Available: `http://arxiv.org/abs/1511.06434`.

[182] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, *et al.*, "Adversarially Learned Inference", Jun. 2016. arXiv: `1606.00704`. [Online]. Available: `http://arxiv.org/abs/1606.00704`.

[183] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial Autoencoders", Nov. 2015. arXiv: `1511.05644`. [Online]. Available: `http://arxiv.org/abs/1511.05644`.

[184] Z. Li, X. Zhang, H. Müller, and S. Zhang, "Large-scale Retrieval for Medical Image Analytics: A Comprehensive Review", *Medical Image Analysis*, 2017. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S136184151730138X`.

[185] O. Jimenez-del-Toro, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, *et al.*, "Analysis of histopathology images: From traditional machine learning to deep learning", in *Biomedical Texture Analysis*, Elsevier, 2018, pp. 281–314.

[186] W. Sun, B. Zheng, and W. Qian, "Automatic feature learning using multichannel ROI based on deep structured algorithms for computerized lung cancer diagnosis", *Computers in biology and medicine*, vol. 89, pp. 530–539, 2017.

[187] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning", *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, Jul. 2016, ISSN: 0018-9294. DOI: `10.1109/TBME.2015.2496253`.

[188] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records.", *Scientific reports*, vol. 6, no. April, p. 26 094, 2016, ISSN: 2045-2322. DOI: `10.1038/srep26094`. arXiv: `arXiv:1401.4290v2`. [Online]. Available: `http://www.nature.`

com/articles/srep26094%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27185194%7B%5C%%7D5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4869115.

[189] H. I. Suk, S. W. Lee, and D. Shen, "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis", *Brain Structure and Function*, vol. 220, no. 2, pp. 841–859, 2015, ISSN: 18632661. DOI: 10.1007/s00429-013-0687-3. arXiv: NIHMS150003.

[190] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative adversarial networks for noise reduction in low-dose CT", *IEEE transactions on medical imaging*, vol. 36, no. 12, pp. 2536–2545, 2017.

[191] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, *et al.*, "Medical image synthesis with context-aware generative adversarial networks", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 417–425.

[192] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review", *arXiv preprint arXiv:1809.07294*, 2018.

[193] J. P. Cohen, M. Luck, and S. Honari, "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation", *arXiv preprint arXiv:1805.08841*, 2018.

[194] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization", Jul. 2016. arXiv: 1607.06450. [Online]. Available: http://arxiv.org/abs/1607.06450.

[195] C. Eickhoff, I. Schwall, A. de Herrera, and H. Müller, "Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images", *CLEF working notes, CEUR*, 2017.

[196] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, *et al.*, "Ad click prediction: a view from the trenches", in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1222–1230.

[197] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs", *arXiv preprint arXiv:1702.08734*, 2017, ISSN: 1702.08734. arXiv: 1702.08734. [Online]. Available: http://arxiv.org/abs/1702.08734.

[198] K. Dimitris and K. Ergina, "Concept detection on medical images using Deep Residual Learning Network", in *Working Notes of Conference and Labs of the Evaluation Forum*, Dublin, Ireland: CEUR-WS.org, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper%7B%5C_%7D122.pdf.

[199] L. Valavanis and S. Stathopoulos, "IPL at ImageCLEF 2017 Concept Detection Task", in *Working Notes of Conference and Labs of the Evaluation Forum*, Dublin, Ireland: CEUR-WS.org, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper%7B%5C_%7D144.pdf.

[200] J. Zhang, H. Dang, H. K. Lee, and E.-C. Chang, “Flipped-Adversarial AutoEncoders”, *arXiv preprint arXiv:1802.04504*, 2018.

[201] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, Nov. 2017. arXiv: `1710.10196`. [Online]. Available: `http://arxiv.org/abs/1710.10196`.

[202] M. Arjovsky and L. Bottou, “Towards Principled Methods for Training Generative Adversarial Networks”, *arXiv preprint arXiv:1701.04862*, 2017, ISSN: 0022-538X. DOI: `10.1128/JVI.06350-11`. arXiv: `1701.04862`. [Online]. Available: `http://arxiv.org/abs/1701.04862`.

[203] L. McInnes and J. Healy, “UMAP: Uniform manifold approximation and projection for dimension reduction”, *arXiv preprint arXiv:1802.03426*, 2018.

[204] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, “Thresholding Classifiers to Maximize F1 Score”, *Machine Learning and Knowledge Discovery in Databases*, vol. 8725, pp. 225–239, Feb. 2014. arXiv: `1402.1892`. [Online]. Available: `http://arxiv.org/abs/1402.1892`.

[205] Y. Zhang, X. Wang, Z. Guo, and J. Li, “ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning”, *CEUR Workshop Proceedings*, vol. 2125, 2018, ISSN: 16130073.

[206] M. Lux and S. A. Chatzichristofis, “LIRE: lucene image retrieval: an extensible Java CBIR library”, in *Proceedings of the 16th ACM international conference on Multimedia*, ACM, 2008, pp. 1085–1088.

[207] L. Valavanis and T. Kalamboukis, “IPL at imageCLEF 2018 : A kNN-based Concept Detection Approach”, 2018.

[208] J. T. Springenberg, “Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks”, Nov. 2015. arXiv: `1511.06390`. [Online]. Available: `http://arxiv.org/abs/1511.06390`.

[209] A. Kumar, P. Sattigeri, and P. T. Fletcher, “Semi-supervised Learning with GANs: Manifold Invariance with Improved Inference”, in *Advances in Neural Information Processing Systems*, 2017, pp. 5540–5550. DOI: `arXiv:1705.08850v2`. arXiv: `1705.08850`. [Online]. Available: `http://arxiv.org/abs/1705.08850`.

[210] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, *et al.*, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”, *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[211] R. A. Castellino, “Computer aided detection (CAD): an overview.”, *Cancer imaging : the official publication of the International Cancer Imaging Society*, vol. 5, no. 1, pp. 17–9, 2005, ISSN: 1470-7330. DOI: `10.1102/1470-7330.2005.0018`. [Online]. Available: `http://www.ncbi.nlm.nih.gov/pubmed/16154813%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1665219`.

[212] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship”, *Scientific data*, vol. 3, 2016.