ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Bachelor Thesis
in
Computer Science

Medical Image Tagging

Foivos Charalampakos

Supervisors: John Pavlopoulos, Ion Androutsopoulos, Vasiliki Kougia

September 2021

Foivos Charalampakos

Medical Image Tagging

September 2021

Supervisor: John Pavlopoulos, Ion Androutsopoulos, Vasiliki Kougia

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics

Natural Language Processing Group

Athens, Greece

Abstract

Medical image classification is a very challenging and interesting task that has seen incredible advancements with the use of deep learning approaches. This thesis addresses the task of multi-label medical image classification, called medical image tagging. The task aims to assign medical terms (tags) describing possible findings in an image with the ultimate goal of helping clinicians in the diagnostic process. Towards this goal, we develop models employing several deep learning architectures for encoding the images and combine them with classification-based and retrieval-based methods in order to classify the images. We experiment with two datasets, one consisting of radiology images of several modalities (e.g., MRI scans, angiography etc.) and one that consists of chest X-rays.

Περίληψη

Η κατηγοριοποίηση ιατρικών εικόνων είναι ένα δύσκολο και ενδιαφέρον αντικείμενο που έχει γνωρίσει μεγάλη πρόοδο με τη χρήση μεθόδων βαθιάς μάθησης. Αυτή η εργασία πραγματεύεται την ταξινόμηση ιατρικών εικόνων σε πολλές κατηγορίες, που ονομάζεται επισήμανση ιατρικών εικόνων (medical image tagging). Η εργασία στοχεύει στον εντοπισμό ιατρικών όρων που περιγράφουν πιθανά ευρήματα σε μια εικόνα ασθενούς με απώτερο σκοπό να βοηθήσει τους ιατρούς στη διαγνωστική διαδικασία. Για αυτόν τον σκοπό, αναπτύσσουμε μοντέλα που χρησιμοποιούν αρχιτεκτονικές βαθιάς μάθησης για την κωδικοποίηση των εικόνων και τις συνδυάζουμε με μεθόδους ταξινόμησης και μεθόδους ανάκτησης πληροφορίας προκειμένου να κατηγοριοποιήσουμε τις εικόνες. Τα πειράματα περιλαμβάνουν δύο σύνολα δεδομένων, ένα που αποτελείται από ακτινολογικές εικόνες πολλών κατηγοριών (π.χ., αξονικές τομογραφίες, αγγειογραφίες κ.ά.) και ένα που αποτελείται από ακτινογραφίες θώρακα.

Acknowledgements

I would like to thank all the people that I cooperated with and I am grateful for the efforts they made in order for this work to be completed. First of all, I would like to sincerely thank my supervisor and professor, Ion Androutsopoulos, who gave me the opportunity to work with him in this very interesting research field that covers my interests. Next, I would like to thank my other two supervisors, John Pavlopoulos, for his restless offer of advice and ideas regarding the content of this thesis, and Vasiliki Kougia, for her contributions regarding technical issues and for her kind help and suggestions that helped me better understand notions needed for my work. Thank you all three for guiding me through my first steps in academic research. I would also like to thank Vasilis Karatzas for his help and the great cooperation we had. Last but not least, I would like to wholeheartedly express my gratitude to my family, friends and past professors that believed in me and supported me throughout the years. I would not have gotten to where I am today without them.

Contents

1 Introduction

Medical professionals often subject their patients to various screening examinations that are important for a correct diagnosis. Thus, they examine a large amount of medical images like MRI scans, X-rays and PET/CT scans in order to be more accurate about their diagnosis [Wil+12]. In many cases, examining medical images can be time consuming and may also be a difficult task for inexperienced clinicians. In order to help the diagnostic process regarding medical imaging, Computer Vision (CV) techniques can be used and assist professionals in the interpretation of medical images. Additionally, recent advancements in deep learning have led to increasingly accurate computer vision models that can prove very useful in automating medical image examination as well as more tasks in the biomedical domain.

There are several tasks regarding medical imaging that can assist clinicians during medical examinations, like classification (binary, multi-class or multi-label) [Sha+21], lesion detection [Kij+19], segmentation [Hes+19] etc. This thesis discusses a form of multi-label medical image classification called medical image tagging. In medical image tagging, the goal is to identify one or more medical terms (tags) that that correctly describe possible diagnostic findings based on given image, and assign them to the image. Medical image tagging is a very interesting task that can constitute a first step in the creation of medical reports in a way that the tags predicted could be used to create coherent text. This text generation task is called medical image captioning and its goal is to automatically create text reports that describe findings in medical images [KPA19a]. Furthermore, tagging can accelerate the process of diagnosis by indicating possible findings present in a medical image and help clinicians to identify them more easily. Towards this direction, part of the work of this thesis was done for the participation of the AUEB NLP Group[1] in ImageCLEFmed 2021 Caption's concept detection sub-task [Pel+21]. The competition aimed to the development of systems that can perform medical image tagging. We participated in the competition, addressing the problem with the use of deep learning architectures for image representation in combination with multi-label classifiers and retrieval-based methods. Based on the intuition that patients with similar conditions exhibit similar medical screening results, non-parametric retrieval methods combined with image similarity metrics can be used with relative success. Following our previous success [KPA19b; Kar+20], we managed to rank first in the campaign achieving best results with the use of a retrieval-based system, showcasing the usefulness of the approach [Cha+21]. We will

[1] http://nlp.cs.aueb.gr/

describe these systems and present their results as well as additional work done using another dataset.

Thesis Structure

The rest of this thesis is organized as follows:

Chapter 2

Chapter 2 constitutes an introduction to (medical and non-medical) image classification and the approaches that have been used throughout the years in the literature.

Chapter 3

In Chapter 3, we present the models that were implemented during this thesis.

Chapter 4

Chapter 4 discusses the data that was used for our experiments as well as the details of our experimentations.

Chapter 5

Chapter 5 concludes and proposes directions for future work.

2 Background and Related Work

2.1 Background

Image classification is the task of classifying an image into one or more categories by identifying certain objects of interest that are present in (or missing from) the image. In general, we have three main families:

- **Binary classification:** We need to find whether an image does or does not belong in a certain category. For example, we care to find whether a picture contains a dog or not. Another example, in the medical domain, is to find whether an exam is normal or abnormal [Raj+17b; KKK18].

- **Single-label multi-class classification:** We need to find whether an image belongs to a certain category but we have to choose from several others. The categories are mutually exclusive. For example, we need to find whether an animal image contains a dog or a cat. In the medical domain, we would want to find if a chest X-ray contains a condition such as cardiomegaly etc. [Zon+21; Lee+20].

- **Multi-label multi-class classification:** In this problem, an image can belong to several categories at the same time. For example, we could find that an image contains both a dog and a cat at the same time. Concerning the medical domain, an exam can contain multiple conditions that should be identified [Irv+19].

The classification process involves an algorithm that extracts *features* from the pixels of the image. These features represent different properties of the image that will help the algorithm classify it correctly. Features are divided in low-level (e.g., edges of different directions, shapes, textures) and high-level (e.g., ears, eyes etc.). For example, a cat classifier should be able to extract higher-level features such as the eyes, ears and the tail of the animal in order to classify it. In the past, traditional computer vision and machine learning algorithms relied on hand-crafted image filters and features in order to work well [KVB88]. The disadvantage of this approach was that feature engineering became extremely difficult as it required extensive input tuning. The rise of deep learning revolutionized image classification and the computer vision field in general [Dea20], as contemporary neural networks learn features present in the image in an automated way during their training

process, thus removing the need for preprocessing the data in order to derive them by hand.

2.2 Deep Classification Pipeline

Image classification can be seen as the following high-level process: a neural network receives an input image as a three-dimensional matrix where the size of the first two dimensions corresponds to the length and width of the images in pixels and the third dimension is the number of the color channels (in most cases, we have 3 RGB channels). The image is passed through the layers of the network, which outputs the probability of the image belonging to a certain category.

In more detail, image classification is a supervised learning problem, meaning that we define a set of target classes (categories or objects to identify in an image), and train a network to recognize them using labeled example images. The objective of the training is to minimize a loss function that shows us how well the network guesses. The network is trained using the back-propagation algorithm [RHW86] and updates its weights in order to decrease the loss in every iteration and thus learn more meaningful image features. Therefore, the goal of these networks is to extract useful features from the images which are transformed into vector encodings (representations), also called image embeddings. Because of this, these networks are called image encoders. Many image encoder architectures have been studied during the previous years and we will take a brief look at them in the next section [Hua+17; TL19; He+16; Dos+21; Ros57].

In order to train the network, we choose the loss function depending on the problem we want to solve. When it comes to binary classification, the most common approach is the following: the final layer of the network contains one output neuron with a sigmoid activation function which outputs a probability in $[0, 1]$. The network is trained using the binary cross-entropy loss function. When we have a multi-class classification problem, the final layer contains as many neurons as the categories we want to identify and a softmax activation function that converts the logits to a vector of (normalized) probabilities in $[0, 1]$. The network is trained using the categorical cross-entropy loss function. This approach can also be used with binary classification where the final layer will contain only two neurons (one for each class). Finally, when we are dealing with a multi-label classification problem, the final layer again contains as many neurons as the categories present in our data but we use the sigmoid activation function and the binary cross-entropy loss function. In essence, the network acts as n binary classifiers where $n =$ number of distinct categories.

2.3 Architectural evolution

In this section, we will briefly describe the three main architectures of neural image encoders that have been used throughout the years.

Multi-Layer Perceptrons

At first, traditional MLP networks [Ros57; Ros63] were used for image processing. These networks are composed of multiple linear layers followed by non-linear activation functions and are stacked together (see Fig. 2.1). Each neuron is connected with every neuron from the previous layer. Such architectures do not take into account the spatial structure of images, treating pixels which are far apart in the same way as pixels that are close together. Furthermore, the full connectivity between nodes result in computational intractability when dealing with higher resolution images (e.g., a layer receiving a 1000×1000-pixel image with 3 color channels will have 3 million ($1000 * 1000 * 3$) weights).

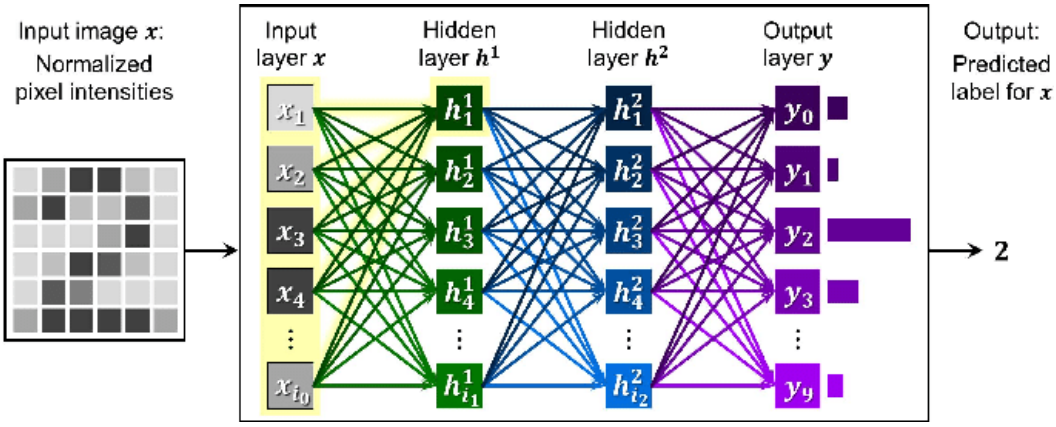


Fig. 2.1: An MLP for handwritten digit classification. There are four layers (the input layer, two hidden layers and the final/output layer). The final layer receives the image representation vector (i.e., the output vector of the previous layer) as an input and outputs a probability distribution showing the probability of the input image belonging in each category. Figure taken from [Che+20b].

.

Convolutional Neural Networks

A breakthrough in building image processing models came when CNNs [LeC+89] were discovered. These networks consist of convolutional layers that apply learnable filters using the *convolution operation* by sliding over the image and pooling layers that downsample

the convolved features and provide translational invariance. CNNs are variants of MLPs that, following the concept of human receptive fields, exploit spatially local correlation by enforcing a sparse local connectivity pattern between neurons of adjacent layers. Therefore, each neuron is connected to only a small region of the input volume and not to all of it like in MLPs (see Fig. 2.2). The extent of this connectivity is called the *receptive field* of the neuron. This architecture ensures that the learned convolutional filters produce the strongest response to a spatially local input pattern. Stacking many convolutional and pooling layers together, leads to non-linear filters that become increasingly global (i.e. responsive to a larger region of pixel space) as the depth of the network increases so that it first creates representations of small parts of the input, then from them assembles representations of larger areas.

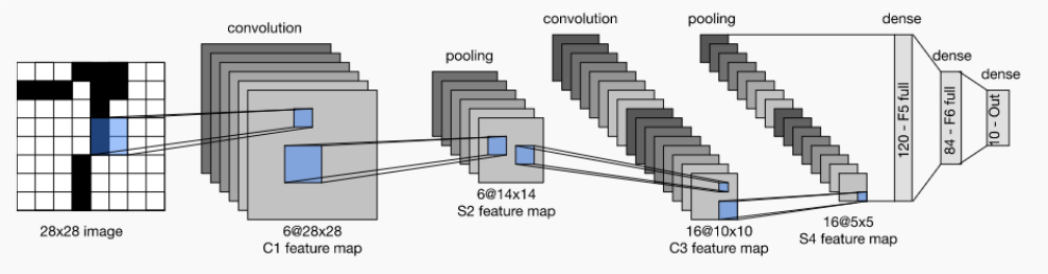


Fig. 2.2: A CNN for handwritten digit classification. There are many stacked convolutional and pooling layers that help the network extract global features. We can see that the representation is sent to an MLP which will output probabilities regarding the classification categories. Figure taken from [Zha+21b].

Vision Transformers

In recent years, Transformer networks [Vas+17; Dev+19] have dominated the field of Natural Language Processing (NLP) producing state of the art results in several language related tasks. Due to their success, Transformers have begun to enter the field of computer vision [Dos+21; Bel+19; Ram+19] as an alternative to CNN encoders. Visual Transformers (ViT) [Dos+21] receive images divided in patches and they handle them in parallel, like NLP Transformers receive word tokens (see Fig. 2.3). Usually, in NLP, the input is a set of vectors (called word embeddings) that are either pre-trained or produced by learnable linear layers. Likewise, in ViT, the image patches are encoded by a learnable linear layer and the output vector is fed to the Transformer encoder. When applied to images, the multi-head self-attention (the main component of the Transformer architecture) acts as a generalization of convolutional layers [CLJ20]. Furthermore, Transformers are able to learn complex **spatial relationships** between high level image features (like ears, eyes etc.), something that CNNs are unable to do. More specifically, CNNs are highly effective at

recognising features e.g., facial: eyes, ears, a nose etcHowever, CNNs struggle to learn the spatial relationships between these sub-components and other details such as the number of each of these components that should be present. As long as the sub-components are present, CNNs are likely to classify the image as containing a face. However, a face that contains multiple ears or multiple eyes will also be classified as a face. CNNs have no concept of the number or relative position of these features. In other words, CNNs are good feature detectors but bad at the relationship between features. [1]

ViT can be used with CNNs as well, if we want to decrease the computational resources that we need. The CNN can be used to process lower level features in the image that will be fed to the Transformer encoder. In addition, more recent works explore whether the use of the attention mechanism in the Transformer architecture is critical, by replacing the multi-head attention layers [Vas+17] with feed-forward MLPs [Liu+21; Tol+21; Tou+21] and achieving comparable results.

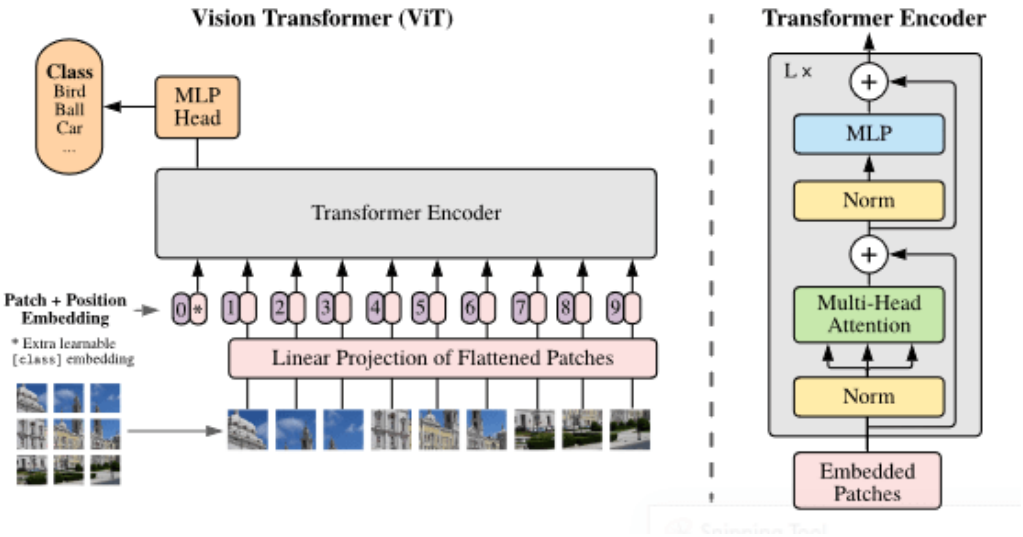


Fig. 2.3: A high-level view of ViT. An image is divided in equal patches and is fed to the encoder in order to create the image embedding. The encoder is a Transformer network as described in [Vas+17]. The image embedding is fed to an MLP in order to be classified. Figure taken from [Dos+21].

2.4 Related Work

[1] https://techblog.ezra.com/transformers-in-medical-computer-vision-643b0af8fc41

2.4.1 Generic Image Classification

Deep learning methods are extensively used in order to solve image classification tasks of various schemes (binary, multi-class etc.). Current approaches heavily rely on *transfer learning*, a method that allows us to use networks that have already been pre-trained with a large dataset in order to achieve better results. Neural networks are trained using a very large dataset, usually ImageNet [Den+09], and we use the pre-trained weights in our task either by adding a linear classifier on top of the learned (frozen) features which we train with our target dataset or by fine-tuning the whole pre-trained network along with our classifier. This approach has enabled the use of several complex architectures [He+16; TL19; Hua+17] which have been trained on ImageNet and produce great results in a plethora of tasks.

The current state-of-the-art (SoTA) results on the ImageNet dataset [Den+09] (90.45% Top-1 accuracy) are achieved by a huge ViT model that consists of 1,843 million parameters [Zha+21a]. It is an up-scaled version of the original ViT model [Dos+21] and is pre-trained on a huge dataset of natural images. Additionally, Chen et al. [CHG21] mention that ViT can achieve good results without needing very large image datasets to pre-train on.

As for the SoTA CNN network, EfficientNet-L2 [TL19] achieves 90.2% Top-1 Accuracy on the ImageNet dataset when iteratively trained with a student-teacher approach [Pha+21] (a form of knowledge distillation [HVD15]) using a large unlabeled dataset.

2.4.2 Medical Image Classification

Medical image classification is a more complex task as objects of interest (e.g., conditions) lie only in particular areas of the image and are more difficult to be identified. Furthermore, labeled medical data is more difficult to find than natural images data and there are several different medical modalities (e.g., ophthalmology images, x-ray chest images etc.) so transfer learning (using a medical dataset that contains all modalities for pre-training) is not easy.

Despite the large difference between natural and medical images, image encoders pre-trained on ImageNet [Den+09] and fine-tuned on the target medical dataset have achieved results close to predictions from clinicians [EKN17]. Following this direction, Rajpurkar et. al [Raj+17a] proposed CheXnet that uses a DenseNet-121 [Hua+17] image encoder pre-trained on ImageNet [Den+09] and fine-tuned on the ChestX-ray 14 dataset [Wan+17] to classify chest X-rays to 14 different conditions. Following the work of [Raj+17a], Kougia et al. [KPA19b; KPA20] as well as Karatzas et al. [Kar+20] managed to acquire the 1st position at the ImageCLEFmed concept detection task on 2019 [Ion+19] and 2020 [Ion+20] respectively. They used a DenseNet-121 as the image encoder along with a Feed-Forward Neural Network (FFNN) as the classifier. Their work also included ensembles of different deep learning models.

Retrieval-based methods have also proved useful in medical image classification. In 2017,

Valavanis and Stathopoulos [VS17] used the k-NN algorithm [FH89] and acquired the 2nd position at ImageCLEFmed 2017 concept detection task [Eic+17]. They did not use any deep learning model as they experimented with methods such as bag of colors (BoC) and bag of visual words (BoVW) to represent the images. Kougia et al. [KPA19b; KPA20] acquired the 3rd position at the ImageCLEFmed 2019 concept detection task by implementing a neural retrieval-based method. They used k-NN to fetch the k most similar training images for each test image using image embeddings produced by a DenseNet-121 encoder and scaled-dot product similarity. Additionally, Sonker et. al [Son+20] acquired the 2nd position in the 2020 task [Ion+20] using k-NN on ResNet-101 [He+16] embeddings. In 2021, with work done as part of this thesis, we acquired the 1st position at the 2021 ImageCLEFmed concept detection task [Pel+21] using an ensemble of 1-NN retrievers that utilized several CNN image encoders [Cha+21].

A more recent advancement in medical image classification involves the use of self-supervised learning. Self-supervised learning is a form of unsupervised learning that is used to learn representations of the data without the use of labels by solving a pre-training task. Several pre-training tasks have been proposed [DGE15; NF16; Wu+18] and the most successful until now is the instance discrimination task [Che+20a; CH21; He+20]. In this setting, visual encoders are trained to project similar images (i.e., images that depict the same content) close together in the embedding space and dissimilar images far away. For example, an NLP counterpart is the famous *word2vec* [Mik+13]. This form of self-supervised learning (when instance discrimination is the pre-training task) is called *contrastive learning* as the network is trained to "contrast" instances. This is achieved with the use of a contrastive loss function which calculates the similarity between different image samples. Due to the scarcity of human-annotated medical datasets, contrastive learning can be used in order to create more robust and domain-related image representations compared to those created by ImageNet pre-training by leveraging unlabeled medical data that is more easily accessible. Azizi et al. [Azi+21] utilized SimCLR [Che+20a] and the CheXpert dataset (without its labels) [Irv+19] for contrastive pre-training and reported big improvements compared to supervised ImageNet [Den+09] pre-training. Furthermore, they proposed an algorithm that made use of exam cases found in medical datasets for contrastive pre-training. Similar to this work, Sowrirajan et al. [Sow+21] used MoCo [He+20] for the contrastive pre-training to improve Chest X-ray [Irv+19] classification performance. Additionally, Zhang et al. [Zha+20] proposed a multi-modal contrastive learning pipeline where they jointly train an image and a text encoder to contrast images with text reports. The objective was to maximize the agreement between true image-text pairs versus random pairs between the image and text modalities using a bidirectional contrastive objective. They reported improvements compared to ImageNet pre-training and image-only contrastive pre-training algorithms [CH21; Azi+21; Sow+21] in several medical datasets. They used ResNet-50 [He+16] as their image encoder, ClinicalBERT as their text encoder [Als+19] and the MIMIC-CXR dataset [Joh+19] for the pre-training.

3 Implemented methods

This section will describe the systems that were implemented for the work of this thesis. All methods used image encoders based on deep neural networks combined with an Feed Forward Neural Network (FFNN) or a retrieval-based component.

3.1 Deep learning-based approaches

3.1.1 CNN encoder + FFNN

This system is based on previous work from Kougia et al. [KPA19b; KPA20; Raj+17a] and Karatzas et al. [Kar+20]. It uses a CNN image encoder and a single dense layer on top comprised of $|C|$ neurons, where C is the number of the possible tags, that acts as the classifier. We experimented with several modern CNN architectures (such as ResNet-50 [He+16], DenseNet-121 [Hua+17]) that were pre-trained on ImageNet [Den+09] and fine-tuned them on the target medical dataset. In a fully-supervised setting, we fine-tuned the model using labeled images that were rescaled to 224×224 and normalized with the mean and standard deviation of ImageNet to match the requirements of the encoder and how it was pre-trained. We trained the model by minimizing the binary cross entropy loss (as we experimented on multi-labeled datasets) and used Adam [KB15] as the optimizer with its default hyper-parameters. Following Kougia et al. [KPA19b], we used early stopping on the validation set with patience of 3 epochs and decayed the learning rate by a factor of 10 with patience of 1 epoch. At inference, we predicted the tags for each image using their probabilities, as estimated by the last fully connected layer of the classifier. For each tag, we assign it to the test image if its output probability exceeds a threshold t. Instead of using the default value $t = 0.5$, we tuned the decision threshold using the validation data. Similar to Karatzas et al. [Kar+20], we also experimented with ensembles using set operations on the different predicted tags. We trained five classifiers and the two best performing were kept. We combined their predicted tags using either the UNION or the INTERSECTION operation.

Additionally, wanting to improve the representations learned by the image encoder in order to achieve better performance, we also experimented with a pre-training step that utilized supervised contrastive learning [Kho+20], introducing it to the biomedical domain. Supervised contrastive learning is a special case of contrastive learning where the labels in the data are employed in order to provide more information about similar instances. In

the self-supervised setting, there are no labels present in the data. The loss is calculated as such: for each anchor image in an augmented batch, an augmented sample is used as the positive (similar) example and all the others samples in the batch are used as negatives (dissimilar). The loss takes the following form [Che+20a]:

$$L_1 = -\sum_{i \in I} \log \frac{exp(sim(z_i, z_j)/\tau)}{\sum\limits_{\alpha \in A(i)} exp(sim(z_i, z_\alpha)/\tau)} \tag{3.1}$$

where z_i is the anchor image, z_j is the positive example, sim denotes cosine similarity, I is the augmented batch, $A(i) \equiv I \setminus \{i\}$, z_a is a negative example and τ is a temperature hyper-parameter for restricting the range of similarity scores [Che+20a].
In supervised contrastive learning, we utilize label information in order to use a different sample of the same class as the positive instead of generating one through the anchor. The loss takes the following form [Soh16]:

$$L_2 = -\sum_{i \in I} \log \frac{exp(sim(z_i, z_k)/\tau)}{\sum\limits_{\alpha \in A(i)} exp(sim(z_i, z_\alpha)/\tau)} \tag{3.2}$$

where z_k is now a positive example originating from a different sample image with the same class as the anchor.

For this experiment, we used a ResNet-50 [He+16] initialized with ImageNet weights as our image encoder. The network mapped each image to a dense representation vector and after normalizing it, we passed it through a projection MLP [Che+20a]. The output of the projection was used in the calculation of the pre-training loss. We further pre-trained the model following the pipeline presented in [Kho+20] while using the supervised contrastive loss with a temperature value of $\tau = 0.1$ and Adam as the optimizer. At the end of the pre-training, we discarded the projection MLP and used the (frozen) backbone encoder as a feature extractor by adding a dense layer on top. We then trained the top layer similarly to the fully-supervised setting that was mentioned above.

3.1.2 ViT encoder + FFNN

In this system, we replaced the CNN image encoder with a ViT architecture [Dos+21] (see Section 2.3). We used the ViT encoder pre-trained on ImageNet provided by the original paper [Dos+21], attached a fully-connected layer, and fine-tuned them together on the medical dataset using the supervised setting that was described in the previous section. To create the output of the encoder (i.e., the final representation to be fed in the classifier), we considered two approaches:

- **Classification token:** The authors of the original paper used a learnable embedding, called classification token, which was prepended to the input sequence of the linearly encoded patches (see Fig. 2.3). This embedding served as the representation of the whole input image.

- **Using the outputs of the encoder:** ViT is comprised of a stack of Transformer encoders. We used the outputs of the final encoder and flattened them to a single vector and then passed it to the linear classifier.

3.2 Retrieval-based approaches

3.2.1 Image encoder + k-NN retrieval

This system is the same as the one presented in [KPA19a; Kar+20]. We experimented with CNN and ViT encoders similar to the ones described in the previous section. For each training image, we extracted its image embedding (encoding) from the last global average pooling layer of the encoder (or the classification token in the case of ViT) and stored it in an offline dictionary. Given a test image, we obtained its embedding (the same way we created the training embeddings) and then retrieved the top k most similar training images according to their cosine similarity with the test image (calculated on image embeddings) and assigned the r most frequent tags present in them to the test image. We tuned the two hyper-parameters using the validation set. For k, we experimented with values between 1 and 200. For r, we used values between 1 and 10 and also considered two additional functions:

- The average number of concepts in the k images:

$$r = \frac{1}{k} \sum_{i=1}^{k} n_i \tag{3.3}$$

- A weighted sum which depended on the cosine similarities of the retrieved images:

$$r = \sum_{i=1}^{k} \frac{\cos(g, g_i)}{\sum_{j=1}^{k} \cos(g, g_j)} \cdot n_i \tag{3.4}$$

where:

n_i = the number of ground-truth concepts of the i-th retrieved image
g = the test image
g_i = the i-th closest training image
$\cos(g, g_i)$ = the cosine similarity between images g and g_i

In special cases where the dataset contains images that depict healthy patients, if the most frequent tag in the retrieved images is a tag that declares that no condition is present in the image (e.g., "no finding", "normal" etc.), then the test image is classified only with this tag.

3.2.2 1-NN ensemble

This system is an extension of the one that was described in [Kou19] and was also used for biomedical captioning [KPA19a]. Similar to k-NN (see Section 3.2.1) where we considered multiple neighbors, 1-NN returns the tags of the single most similar training image, given a test image. We used several CNN image encoders (e.g., DenseNet-121 [Hua+17], ResNet-50 [He+16], EfficientNet-B0 [TL19]) in order to form the ensemble. We again calculated the image embeddings for all the training images and for each test image, for each encoder, we retrieved the training image with the highest cosine similarity resulting in a total of $|E|$ images, one retrieved per encoder, where E equals the number of encoders used. For the tags assignment, we considered four different strategies:

- **MAJORITY VOTING:** For each tag, we counted how many of the E images (one retrieved per encoder) are associated with it. If the tag is present in the majority (i.e., $|tag| \geq \frac{|E|}{2}$), we assign it to the test image.

- **UNION OF INTERSECTIONS:** We considered the INTERSECTION of the tags retrieved by a subset of encoders (usually the encoders that shared the same architecture) and then we used the UNION of the above INTERSECTIONS.

- **INTERSECTION OF UNIONS:** This is the reversed calculation of the previous strategy. We considered the UNION of the tags retrieved by a subset of encoders and then used the INTERSECTION of the UNIONS.

- **UNION OF UNIONS:** We used the UNION of the tags of the images retrieved by a subset of encoders and then used the UNION of the UNIONS.

Similar to k-NN, for datasets with "normal" images, if the most frequent tag in the retrieved images is a tag that declares that no condition is present in the image (e.g., "no finding", "normal" etc.), then the test image is classified only with this tag.

3.3 Hybrid approaches

3.3.1 Similarity-based decision

This system is loosely based on the work done by Li et al. [Li+18] for biomedical image captioning. It is a hybrid method that uses a 1-NN retrieval component and a classifier. We used an image encoder in order to obtain the image embeddings for all the training images. During inference, we retrieved the training image with the highest cosine similarity to the test image embedding and applied a similarity threshold s. We tuned the value of s in the interval $[0.65, 1]$ with $step = 0.01$. If the similarity exceeded s, the tags of the neighbor were assigned to the query image. If not, we fed the test image to the (already trained) classification component and assigned the predicted tags to it.

4 Experiments

4.1 Datasets

There is a variety of datasets that contain medical images and are available in the literature. They cover a wide range of medical examinations (X-rays, PET/CT scans etc.) and body parts (brain, eye etc.) and have been used for several tasks (most notably classification, detection and segmentation) [Lit+17]. Furthermore, recent works [Joh+19; Irv+19; Wan+17; Raj+17b] have introduced several **large-scale**, labeled, medical datasets to the community (although most of them are only comprised of chest X-rays). The MIMIC-CXR dataset [Joh+19] also includes textual reports written by doctors. Two of these datasets (CheXpert [Irv+19] and MURA [Raj+17b]) were released along with a competition that invited participants to develop medical classification systems using deep learning. Moreover, the ImageCLEFmed datasets were collected and distributed for the ImageCLEFmed Caption competition tasks [Eic+17; Her+18; Ion+19; Ion+20; Pel+21]. The aim of such competitions is to encourage the implementation of new systems in order to develop automated methods to assist the diagnostic process and therefore help clinicians in their daily workflow.
We will briefly describe the ImageCLEFmed 2021 Caption [Pel+21] and IU X-ray [Dem+15] datasets that consist of of medical images annotated with multiple medical terms (tags).

4.1.1 ImageCLEFmed Caption 2021

In 2021, the ImageCLEFmed Caption task focused on medical image understanding based on the visual perception of the images and medical text data. Such data is medical captions or UMLS® Unique Identifiers (CUIs) paired with each image [Pel+21]. This year, the competition also included a Caption Prediction sub-task [Her+18; Eic+17], which was not included in the two previous years.
In order to make the task more realistic, the organizers used real radiology images annotated by medical doctors in contrast with previous years where the dataset originated from biomedical articles of the PMC Open Access[1] [Rob01]. This year's dataset was the same with the one used in the ImageCLEFVQA task [Aba+21] and stemmed from the MedPix[2] database. The organizers stated that only cases where the diagnosis was made based on the image were selected and their corresponding text annotations were used as a basis for the

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[2] https://medpix.nlm.nih.gov/home

extraction of the concepts and captions. Furthermore, the dataset included images divided in 7 sub-classes. Each sub-class was indicated only with a corresponding set of concepts that showed the image acquisition technique. Instead of distributing the dataset into 7 different categories like last year, the above way was chosen [Pel+21; Ion+20; Kar+20]. All images in the dataset were accompanied by the unique identifier (CUI) of the UMLS [Bod04] tags (here called concepts). An image can be associated with multiple CUIs (see Figure 4.1)

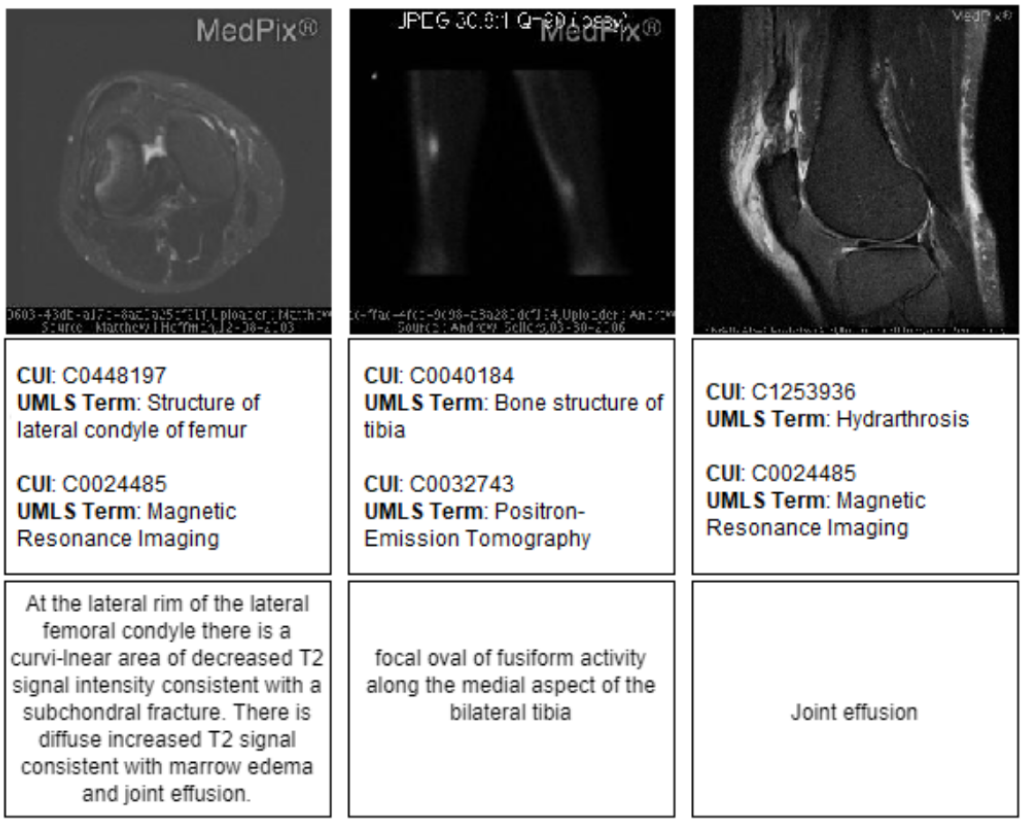


Fig. 4.1: Three images of the ImageCLEFmed 2021 dataset (1st row) with their corresponding tags (2nd row) and captions (3rd row). Figure taken from [Cha+21].

The number of unique concepts was reduced once again compared to previous years. There were 111,156 possible concepts in 2018 [Zha+18], 5,528 in 2019 [KPA19b], 3,047 in 2020 [Kar+20] and 1,585 in 2021. The average number of concepts assigned to each image was 3.48. The minimum number of concepts assigned to an image was 1 (found in 10 images), and the maximum was 12 (found in 2 images). The distribution of the assigned concepts can be seen in Figure 4.2. Furthermore, the dataset is highly imbalanced (e.g., 632 concepts appear only one time, 1 concept appears 1,400 times) and there is no concept appearing in every image, unlike in the last year's dataset. [Cha+21; Kar+20]. Table 4.1 shows the 5 most frequent concepts of the training set and how many training images they were assigned to[3].

[3]We used the REST API (https://documentation.uts.nlm.nih.gov/rest/home.html) of the UMLS Metathesaurus (uts.nlm.nih.gov/home.html) to map each CUI to its term

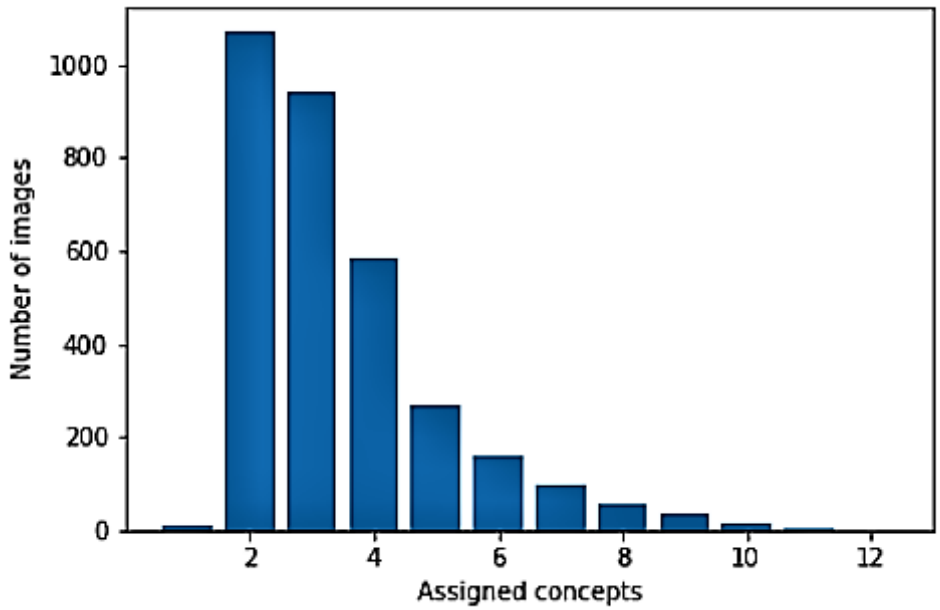


Fig. 4.2: Distribution of assigned concepts in the ImageCLEFmed 2021 dataset.

A small training set of 2,756 images and a validation set of 500 images were provided. A separate test set consisted of 444 images and the concepts for the test images were unknown. For our experiments, we merged the provided training and validation sets and used 10% of the merged data as our validation set, and another 10% of the merged data as our development set in which we evaluated the performance of our models. The remaining 80% served as the training set.

Tab. 4.1: The 5 most frequent concepts (CUIs) in the training set of ImageCLEFmed Caption 2021 and how many training images they are assigned to.

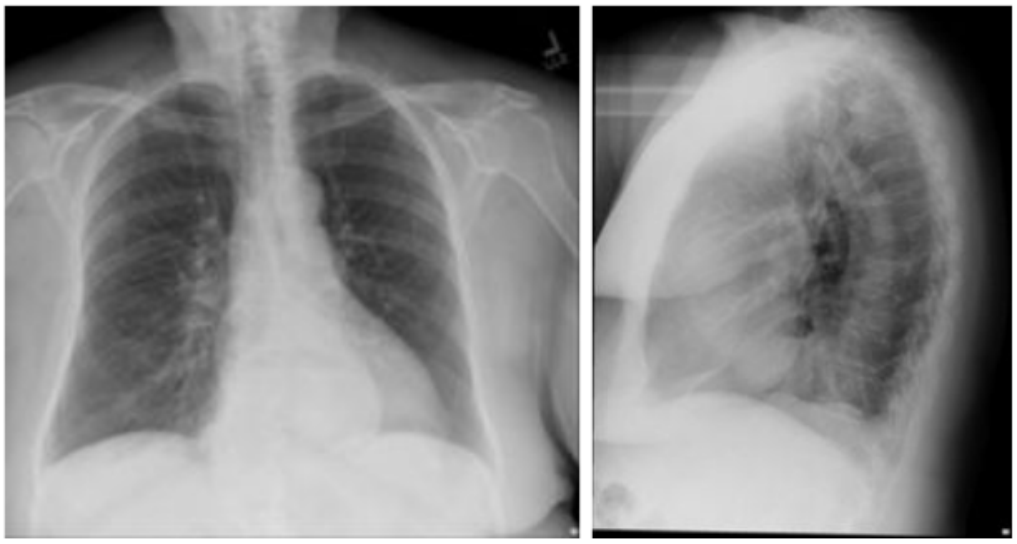
CUI	UMLS term	Images
C0040398	TOMOGRAPHY, EMISSION-COMPUTED	1,400
C0024485	MAGNETIC RESONANCE IMAGING	796
C1306645	PLAIN X-RAY	627
C0041618	ULTRASONOGRAPHY	373
C0009924	CONTRAST MEDIA	283

4.1.2 IU X-ray

The IU X-ray dataset [Dem+15] is the first chest X-ray dataset with publicly available reports and is publicly available through the Open Access Biomedical Image Search Engine (OpenI)[4]. The authors used special software [FM08] to de-identify the texts and images of the collected data and discarded patient cases where patient information was still able to be identified. This pre-processing resulted in 3,955 radiology reports (one per patient) and 7,470 chest X-rays. Each report consists of four sections ('INDICATION', 'COMPARISON', 'FINDINGS' and 'IMPRESSION'). These four sections are usually used by clinicians as a standardized document structure template. The 'FINDINGS' section describes the imaging characteristics of a body structure or function that can have a clinical impact. The 'COMPARISON' section contains previous information about the patient, often from preceding medical exams. The 'IMPRESSION' usually summarizes the most important findings and interprets their clinical value while 'INDICATION' conveys the medical reason for the patient to be subjected to the examination (e.g., symptoms) [Pav+21].
Each report is also associated with two types of tags that stem from the report text. For the first type of tags, two (medically) trained coders used the 'IMPRESSION' and 'FINDINGS' sections to associate each report with a number of tags (we will refer to them as manual tags). They are Medical Subject Heading (MeSH)[5] terms, supplemented with Radiology Lexicon terms (RadLex)[6]. Each annotation tag referred to a pathology, a foreign body or transplant, anatomy (human body parts), signs (imaging observations), or attributes (object or disease characteristics). Furthermore, there are no tags for negated terms (e.g., 'no signs of cardiomegaly') or uncertain findings. However, there is a 'No indexing' tag that denotes that the report could not be labeled. For the second type, tags were automatically extracted from the 'IMPRESSION' and 'FINDINGS' sections of each report using the Medical Text Indexer (MTI) [MYA13]. The resulting tags are called 'MTI encoding' and we will refer to them as 'MTI tags'. MTI tags are single words or terms (e.g., 'Hiatus'), while manual tags may combine anatomical site and type (e.g., 'Hiatal/large') [Pav+21].

We discarded 104 reports that contained no image and 25 reports that were missing both the 'FINDINGS' and 'IMPRESSION' sections thus they did not contain tags. The result was 3,826 reports with 7,430 images. There were also 1,726 reports that did not contain MTI tags. IU X-ray is a dataset which contains X-rays that depict not only exams with medical conditions but also healthy exams. 1,354 from the 1,726 reports that did not contain MTI tags were manually tagged as normal from the two coders. When a report is 'normal', it does not contain any MTI tags. This means that the MTI system failed to assign the abnormalities present in the remaining 372 reports [Kou19]. We performed experiments both with MTI and manual tags. There are 571 unique MTI tags and 1,679 manual tags

[4] https://openi.nlm.nih.gov/
[5] https://goo.gl/iDvwj2
[6] http://www.radlex.org/



FINDINGS: No change. No visible active cardiopulmonary disease. Both lungs remain clear and expanded. Heart and pulmonary XXXX are normal. No change in the large hiatus hernia.

SYSTEM TAGS: hiatus hernia / Hernia, Hiatus

HUMAN TAGS: Hernia, Hiatal/ large

Fig. 4.3: Two images from an X-ray examination in the IU X-ray dataset. We can see the corresponding human-authored 'FINDINGS' section, the manual tags that were extracted by humans (referred to as 'human tags') and the automatically extracted MTI tags (referred to as 'system tags'). The diagnostic caption refers to 'no change' with respect to a previous examination, but there is no link to information and images of the previous examination in this particular dataset. Figure taken from [Pav+21].

(including the 'normal' tag). The average number of manual tags assigned to each report is 2.07 and the average number of MTI tags is 1.75. The minimum number of manual tags assigned to a report is 1 (found in 2233 reports), and the maximum is 12 (found in 4 reports). As for the MTI tags, the minimum number of tags assigned to a report is 0 (indicating that the report is normal - found in 1751 reports), and the maximum is 17 (found in 2 reports). The average number of assigned concepts in the manual setting is 2.05 while in the MTI setting is 2.18. The distribution of the assigned tags (for both types) can be seen in Figure 4.4. Additionally, we encounter imbalance as the majority of reports are labeled as normal. The number of possible tags is again very large and many tags appear very few times (e.g., in the manual tags setting, there are 1035 tags appearing only 1 time and in the MTI setting, there are 229 tags appearing 1 time). Tables 4.2 and 4.3 shows the 5 most frequent tags (for manual and MTI tags respectively) and how many training images they were assigned to.

Tab. 4.2: The 5 most frequent manual tags in IU X-ray and how many reports they are assigned to.

Tag	Reports
'normal'	1,379
'lung/hypoinflation'	238
'lung/hyperdistention'	159
'cardiomegaly'	139
'cardiomegaly/mild'	129

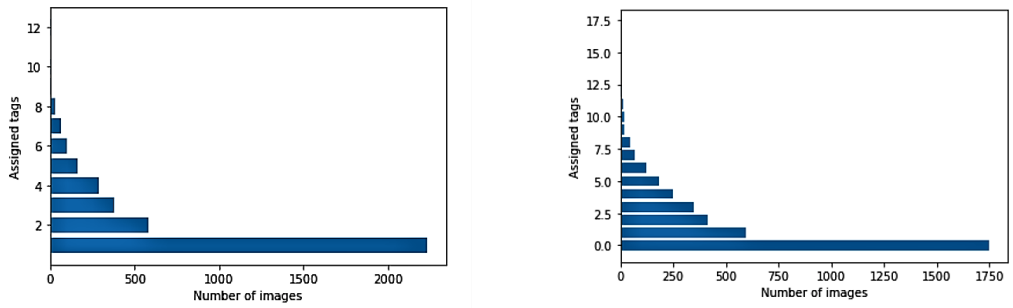
Tab. 4.3: The 5 most frequent MTI tags in IU X-ray and how many reports they are assigned to.

Tag	Reports
'opacity'	421
'degenerative change'	396
'atelectases'	330
'atelectasis'	321
'scarring'	292

For our experiments, we used 80% of the dataset for training, 10% for validation and 10% for testing, splitting with respect to cases.

Discussion

Something that has to be noted about medical image datasets is the presence of patient cases which can comprise more than one images but only one report (or just a category e.g., normal/abnormal). This is also the case for the IU X-ray dataset. Each report can be associated with one or more images. Using each image of the same case as a different instance (an approach that is called image-based) is not completely correct because firstly, the diagnosis was conducted by the clinician by looking at all the case's images and secondly, findings of the case may not be present in every image. To address these limitations, one can follow a case-based approach. There are different types of case-based approaches. A simple one is to obtain the probability distributions over the classes (or the tags) from the model for each image separately and then average the probabilities and obtain the decision with the risk that a class (or a tag) detected in only one image may not be chosen [Raj+17b]. Another solution would be to to choose the tags with the largest probabilities per image [Irv+19]. A more interesting take is to extract one embedding for every image of the case and then, concatenate, add or average them to obtain a single image embedding that will be the case representation [Li+18].
Despite the above, for our experiments with IU X-ray, we chose to train our models using the image-based approach because of its simplicity. As for the ImageCLEFmed 2021 dataset, it did not have the notion of patient cases.



(a) Manual tags distribution

(b) MTI tags distribution

Fig. 4.4: (a) Distribution of assigned manual tags in the IU X-ray dataset. (b) Distribution of assigned MTI tags in the IU X-ray dataset.

4.2 Evaluation measures

The evaluation of the systems was conducted by computing the F1 scores on each test image (in effect comparing the binary multi-hot ground truth vector y_{true} to the binary multi-hot predicted tags vector y_{pred}) and then macro-averaging over all the test instances.

$$F1^* = \frac{\sum_{i=1}^{n} f1\_score(y_{true}, y_{pred})}{n} \tag{4.1}$$

where $f1\_score$[7] compares the binary vectors y_{true} and y_{pred} and n is the number of the test instances.

This was the official measure of the ImageCLEF competition and we also utilized it for our IU X-ray experiments. For IU X-ray, in cases where a test image has no gold tags (in the MTI setting) and no predicted tags assigned, we set the F1 score to 1. That is, because if an image has no tags, there is no abnormality present, so the correct choice for a model would be to not assign any abnormality tag. If the model falsely assigns a tag then the F1 score is set to zero.

4.3 Experimental results

4.3.1 ImageCLEFmed Caption 2021

We used our development set in order to evaluate our models and submitted those that performed best. To generate predictions for the test set, the training and development sets were merged and a held-out validation set was kept to tune the hyper-parameters. Six of our submissions were based on the 1-NN ensemble system (see Section 3.2.2). We made

[7] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

four submissions, one for each assignment strategy (e.g., majority voting etc.), using an ensemble of 20 image encoders. We will refer to these models as 1-NN@20xS where S denotes the strategy used (S=MJ for MAJORITY VOTING, S=IoU for INTERSECTION OF UNIONS, S=UoI for UNION OF INTERSECTIONS and S=UoU for UNION OF UNIONS). The two additional submissions used three and four encoders, respectively and the MAJORITY VOTING assignment strategy. We will refer to them as 1-NN@3xMJ and 1-NN@4xMJ respectively.
One submission used an ensemble classification system pre-trained with supervised contrastive learning [Kho+20] (see Section 3.1.1). We only submitted the system that drew the UNION of the two seperate models as it performed better than the INTERSECTION operation in our development set. We will refer to this system as 2xCNN+FFNN@U. We used the held-out validation set to tune the decision thresholds of the models. This resulted in the values $t_1 = 0.268$ and $t_2 = 0.304$.
Additionally, submissions were made using the hybrid approach system (see Section 3.3.1). In these submissions, the classification component was the 2xCNN+FFNN@U ensemble, while we also used the second ensemble system with the INTERSECTION (2xCNN+FFNN@I) as the classifier. Both of these methods used 1-NN as the retrieval component. We again used the validation set to tune the similarity threshold. The optimal value turned out to be $s = 0.8$.
Table 4.4 presents the scores of our systems on the development and the official test set, as well their position according to the official ranking.

Tab. 4.4: The results and rankings of our systems on the development and test set.

Approach	F1 Score		Rank
	Development	Test	
1-NN@20xMJ	**61.99**	**50.5**	**1**
1-NN@20xUoI	55.73	45.6	9
1-NN@20xIoU	60.85	49.5	2
1-NN@20xUoU	43.33	34.8	23
1-NN@4xMJ	59.08	49.3	3
1-NN@3xMJ	60.76	49.0	5
2xCNN+FFNN@U	57.24	45.9	8
1-NN/2xCNN+FFNN@U	59.53	46.6	7
1-NN/2xCNN+FFNN@I	59.46	45.1	11
1-NN BASELINE	57.32	–	–
MOST FREQUENT BASELINE	27.92	–	–

Discussion

The number of the unique concepts present in the data was 1,585. Furthermore, the number of the total assigned concepts was 11,361. This fact suggests that many images in the dataset had almost, or completely the same concepts. This may explain why our retrieval-based approaches worked very well in this particular campaign. Furthermore, the reduced (in comparison with previous years) number of unique concepts, probably helped

the systems achieve better results. This is indicated by the increasing score of the winning systems (see Fig. 4.5) and also supported by the low performance of the MOST FREQUENT BASELINE on our development set. This baseline assigned to every test image the same three (as the average number of concepts is 3.48) most frequent concepts of the training data (see Table 4.4). The difference in the score shows that our systems did not assign only the most frequent concepts found in the dataset to the test images. However, due to the great imbalance in the dataset and the objectively large number of unique concepts, models tended to predict only a small fraction of the possible concepts (the concepts present in the label set), yet achieving competitive results and managed to win the campaign. More specifically, the predictions of a single deep learning-based classification model included only 127 (8%) of the possible concepts while the UNION ensemble's predictions included 242 (15.2%), indicating that this form of ensembling can imporve multi-label classification. Our 1-NN baseline predicted a total of 422 (26.6%) different concepts. Furthermore, our winning system predicted 279 (17.6%) of the possible concepts (a reduced number due to majority voting). The common predicted concepts between our 1-NN baseline and the single deep learning-based classifier were only 127, showing that ensembling of the two methods could prove beneficial.

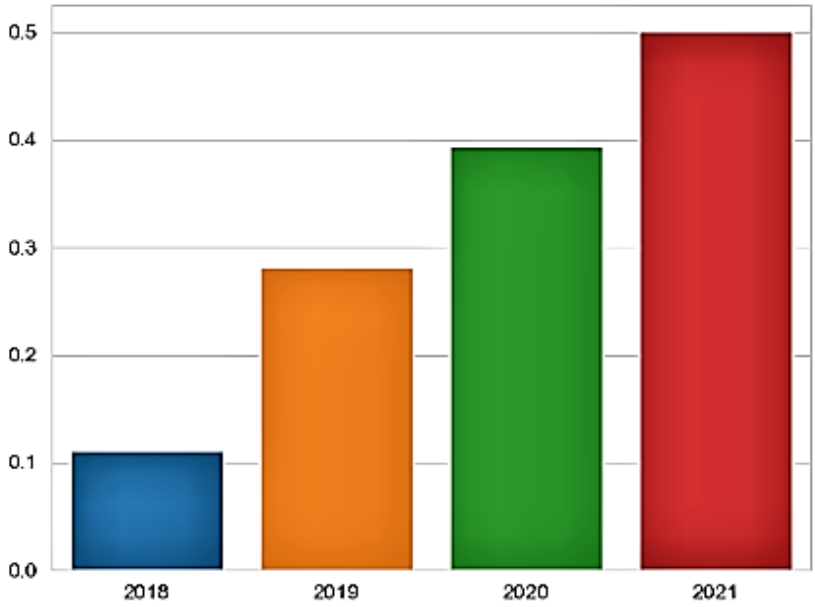


Fig. 4.5: The top F1 score achieved each year in the Concept Detection sub-task of the ImageCLEFmed Caption tasks. Figure taken from [Cha+21].

The value of k ($k = 1$) in our k-NN baseline (see Section 3.2.1) was a result of a tuning process based on the F1 score on the validation data (see Fig. 4.6). Thus, we wanted to improve the performance of the 1-NN baseline by creating the ensemble.
The contrastive pre-training yielded a $\sim 2\%$ improvement in the development F1 score for our deep learning-based approach in comparison with the ImageNet [Den+09] pre-trained encoder. For the pre-training, we created a larger augmented version of the training set

where, for each image, we added four additional variations: we split the image in four horizontal patches and applied Gaussian blur in each patch separately, resulting in four extra noisy images. Furthermore, the pre-training data included random horizontal flip and random rotation.

In the hybrid approach (see Section 3.3.1), beyond the similarity threshold s, we also experimented with an MLP "decider" network that would decide whether to feed each test image to the retrieval component or to the deep learning classifier. To train this network, we created a binary version of the dataset using the tag similarity between the images. In essence, an image belonged to the "positive" class when another image existed that had been assigned exactly or almost the same concepts (F1 score between the two set of concepts ≥ 0.85). We tuned the decision threshold of this MLP using validation data. However, we did not see any performance improvement in comparison to our similarity threshold approach (56.56% vs. 59.53% F1 score on our development data).

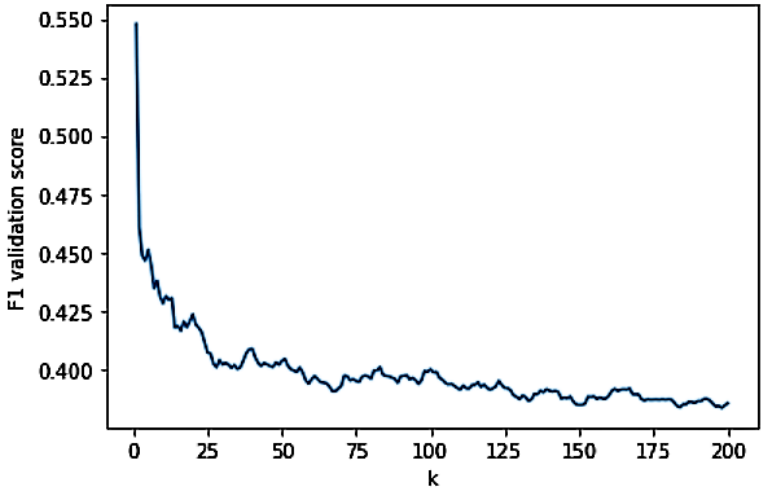


Fig. 4.6: The F1 score on our validation data for different values of the hyper-parameter k in our k-NN retrieval-based system.

4.3.2 IU X-ray

As mentioned, IU X-ray is a dataset that contains patient cases that consist of 2 or more images. For simplicity, we adopted the image-based approach in which we treat each image separately. We used a training/validation/testing split with respect to the cases, meaning that two sets can not contain images from the same patient.

For our deep learning-based classification experiments in the manual setting, we removed the 'normal' tag from the possible concepts for prediction because a model could predict 'normal' and e.g., 'cardiomegaly' for the same image. Thus, we consider an image as normal when there are no tags assigned (like in the MTI setting).

We again experimented with deep learning-based and retrieval-based approaches. Table 4.5 presents the results of our experiments in the test split along with the values of the hyper-parameters of each method.

Tab. 4.5: Results of different approaches in our test split. Value t represents the classification decision threshold, value k is the number of the closest neighbors and variable r is the number of the tags assigned to each image in the k-NN algorithm.

Approach	F1 Score		Hyper-parameters	
	Manual tags	MTI tags	Manual tags	MTI tags
CNN+FFNN	39.17	48.53	$t = 0.35$	$t = 0.2$
CNN+k-NN	39.10	47.61	$k = 126, r = 4$	$k = 113, r =$ average
VIT+FFNN([CLS] EMB.)	38.81	44.2	$t = 0.39$	$t = 0.12$
VIT+k-NN	37.8	46.06	$k = 37, r = 3$	$k = 65, r = 1$
1-NN@20xMJ	20.79	32.62	–	–
MOCO-CNN+FFNN(LINEAR EVAL.)	38.1	46.67	$t = 0.27$	$t = 0.21$
MOCO-CNN+FFNN(END-TO-END)	38.50	43.42	$t = 0.44$	$t = 0.13$
MOST FREQUENT BASELINE	37.8	34.6	–	–

Discussion

Both our deep learning-based and retrieval-based methods failed to predict a large fraction of the possible concepts. For example, in the manual tags setting, the deep learning-based classifier managed to predict only the small number of 26 tags (1.9 %) out of the total 1,350 tags present in the training set. From these 26 tags, the 25 (4.4 %) of them were present in the 564 ground truth tags of the test set. The k-NN approach, which is based on the prediction of the most common tags present in the set, predicted 35 tags (2.5 %) out of the 1,350 training tags and 34 out the 35 were present in the in the 564 ground truth tags of the test set (6.02 %). An interesting point to note is that the retrieval-based method achieves competitive results when compared with the deep learning-based classifier showcasing the usefulness of the approach. However, we can see that the 1-NN ensemble achieved a low performance, indicating that, despite ensembling, tuning the value of k is important. We can also see that the approaches using the ViT encoder achieved competitive results albeit lower compared to their CNN counterparts. In general, the ViT model is more data hungry than CNNs and we are dealing with a relatively small medical dataset. Due to the difference between the two architectures and the lack of the inductive bias that is present in CNN models, ViT requires more data in order to create meaningful image representations. On another note, the experiments yielded better results when the [CLS] embedding was used for the image representation rather than the averaged outputs of the final encoder of the stack.

We also experimented with self-supervised contrastive pre-training using a DenseNet-121 [Hua+17] encoder pre-trained on CheXpert [Irv+19] with the MoCo-CXR algorithm [Sow+21; He+20] [8]. We used two ways of evaluating the pre-trained encoder. The first

[8] We used the checkpoints (in PyTorch (https://pytorch.org)) that can be found in the official GitHub page of the original paper (https://github.com/stanfordmlgroup/MoCo-CXR).

one was the linear evaluation scheme where we froze the encoder, attached a single linear layer on top and trained it as the classifier. This scheme yielded better results than its ImageNet pre-trained counterpart, which had yielded an F1 score of 37.2. The second way of evaluation was the end-to-end scheme, where the backbone encoder and the linear layer were jointly fine-tuned on the target dataset. In our experiments, this approach did not manage to achieve a better score compared to the model that was pre-trained on ImageNet and fine-tuned on the downstream task.

A very interesting note is that, in the manual tags setting, the MOST FREQUENT BASELINE achieves very competitive results by just assigning the 'normal' tag to every test image. This indicates the high imbalance of the dataset and the difficulty the models faced in order to predict a larger number of tags. In general, the classification yields lower results when using the MTI tags than when using the manual tags. This can be explained by the fact that the MTI tags are much fewer and thus provide a more simplified version of the task. However, the MTI tags cannot be trusted completely because they were produced in an automatic way, which was not evaluated by humans. Furthermore, the manual tags are much more specific as each manual tag consists of many findings, where the first one is the main abnormality and the rest describe this abnormality (e.g., its position, severity etc.).

5 Conclusions and future work

5.1 Conclusions

This thesis addressed the task of medical image tagging, a form of multi-label medical image classification. The task aims to assist medical professionals and the diagnostic process when they examine medical images by automatically assigning medical terms (tags) that correspond to the findings of the image. Furthermore, it can constitute a fist step in the automatic generation of medical textual reports. The systems implemented in this thesis were deep learning-based following the recent advancements in medical image analysis. Our approaches also included retrieval-based methods. We focused on two datasets (ImageCLEFmed Caption 2021 and IU X-ray) that were suitable for the task. Based on the notion that similar patient conditions exhibit similar properties depicted in the images, we managed to acquire the first position at the 2021 ImageCLEFmed Caption's concept detection sub-task using a retrieval-based approach that utilized deep neural networks, showcasing its usefulness and how promising it can prove. Further experiments included ensembles of deep learning-based and retrieval-based methods, an implementation of hybrid systems that made use of retrieval components and probabilistic classifiers, as well as exploration of new architectures (e.g., ViT [Dos+21]) for image encoding.

5.2 Future work

In future work, we aim to experiment more with additional retrieval-based methods to see if and how they can also be used to help classification approaches (e.g., retrieval-augmented classification). Moreover, we plan to experiment more with the ViT [Dos+21] architecture and with training strategies such as pre-training with medical datasets instead of ImageNet.

Bibliography

[Aba+21] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, and H. Müller. "Overview of the VQA-MED task at ImageCLEF 2021: Visual question answering and generation in the medical domain". In: *CLEF2021 Working Notes*. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, 2021.

[Als+19] E. Alsentzer, J. Murphy, W. Boag, et al. "Publicly Available Clinical BERT Embeddings". In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA, 2019, pp. 72–78.

[Azi+21] S. Azizi, B. Mustafa, F. Ryan, et al. "Big Self-Supervised Models Advance Medical Image Classification". In: (2021). arXiv: 2101.05224.

[Bel+19] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. "Attention Augmented Convolutional Networks". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 2019, pp. 3285–3294.

[Bod04] O. Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". In: *Nucleic Acids Research* 32 (2004), pp. 267–270.

[CH21] X. Chen and K. He. "Exploring Simple Siamese Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 15750–15758.

[Cha+21] F. Charalampakos, V. Karatzas, V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption Tasks 2021". In: *CLEF2021 Working Notes*. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, 2021, pp. 1184–1200.

[Che+20a] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. 2020, pp. 1597–1607.

[Che+20b] Q. Cheng, J. Kwon, M. Glick, et al. "Silicon Photonics Codesign for Deep Learning". In: *Proceedings of the IEEE* PP (2020), pp. 1–22.

[CHG21] X. Chen, C. Hsieh, and B. Gong. "When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations". In: (2021). arXiv: 2106.01548.

[CLJ20] J. Cordonnier, A. Loukas, and M. Jaggi. “On the Relationship between Self-Attention and Convolutional Layers”. In: *8th International Conference on Learning Representations, ICLR 2020*. Addis Ababa, Ethiopia, 2020.

[Dea20] J. Dean. “1.1 The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design”. In: *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. San Fransisco, CA, USA, 2020, pp. 8–14.

[Dem+15] D. Demner-Fushman, M. Kohli, M. Rosenman, et al. “Preparing a collection of radiology examinations for distribution and retrieval”. In: *Journal of the American Medical Informatics Association : JAMIA* 23 (2015).

[Den+09] Jia Deng, Wei Dong, Richard Socher, et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Miami, FL, USA, 2009, pp. 248–255.

[Dev+19] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, USA, 2019, pp. 4171–4186.

[DGE15] C. Doersch, A. Gupta, and A. A. Efros. “Unsupervised Visual Representation Learning by Context Prediction”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[Dos+21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *9th International Conference on Learning Representations, ICLR 2021*. Vienna, Austria, 2021.

[Eic+17] C. Eickhoff, I. Schwall, A. García Seco de Herrera, and H. Müller. “Overview of ImageCLEFcaption 2017 – Image Caption Prediction and Concept Detection for Biomedical Images”. In: *CLEF2017 Working Notes*. CEUR Workshop Proceedings. Dublin, Ireland: CEUR-WS.org, 2017.

[EKN17] A. Esteva, B. Kuprel, and R. A. Novoa. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542 (2017), pp. 115–118.

[FH89] E. Fix and J. L. Hodges. “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. In: *International Statistical Review / Revue Internationale de Statistique* 57.3 (1989), pp. 238–247.

[FM08] F. Friedlin and C. Mcdonald. “A Software Tool for Removing Patient Identifying Information from Clinical Documents”. In: *Journal of the American Medical Informatics Association : JAMIA* 15 (2008), pp. 601–10.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, 2016, pp. 770–778.

[He+20] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 9726–9735.

[Her+18] A. García Seco de Herrera, C. Eickhoff, V. Andrearczyk, and H. Müller. "Overview of the ImageCLEF 2018 caption prediction tasks". In: *CLEF2018 Working Notes*. CEUR Workshop Proceedings. Avignon, France: CEUR-WS.org, 2018.

[Hes+19] M. H. Hesamian, W. Jia, X. He, and P. Kennedy. "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges". In: *Journal of Digital Imaging* 32 (2019).

[Hua+17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 2017, pp. 2261–2269.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. "Distilling the Knowledge in a Neural Network". In: *NIPS Deep Learning and Representation Learning Workshop*. Montréal, Canada, 2015.

[Ion+19] B. Ionescu, H. Müller, R. Peteri, et al. "ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature". In: 2019, pp. 358–386.

[Ion+20] B. Ionescu, H. Müller, R. Peteri, et al. "ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature, and Internet Applications". In: 2020.

[Irv+19] J. A. Irvin, P. Rajpurkar, M. Ko, et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *AAAI*. 2019, pp. 590–597.

[Joh+19] A. E. W. Johnson, T. Pollard, S. J. Berkowitz, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific Data* 6 (2019), p. 317.

[Kar+20] B. Karatzas, V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption 2020". In: *CLEF2020 Working Notes*. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, 2020.

[KB15] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. San Diego, CA, USA, 2015.

[Kho+20] P. Khosla, P. Teterwak, C. Wang, et al. "Supervised Contrastive Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 18661–18673.

[Kij+19] R. Kijowski, F. Liu, F. Caliva, and V. Pedoia. "Deep Learning for Lesion Detection, Progression, and Prediction of Musculoskeletal Disease". In: *Journal of Magnetic Resonance Imaging* 52 (2019).

[KKK18] M. Kaya, B. A. Karakuş, and S. Karakuş. "Binary Classification of Criminal Tools from the Images of the Case Using CNN". In: *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. 2018, pp. 1–6.

[Kou19] V. Kougia. "Medical image labeling and report generation". MSc thesis (in English), Department of Informatics, Athens University of Economics and Business. 2019.

[KPA19a] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "A Survey on Biomedical Image Captioning". In: *Workshop on Shortcomings in Vision and Language of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, MN, USA, 2019, pp. 26–36.

[KPA19b] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption 2019". In: *CLEF2019 Working Notes*. CEUR Workshop Proceedings. Lugano, Switzerland: CEUR-WS.org, 2019.

[KPA20] V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. "Medical Image Tagging by Deep Learning and Retrieval". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)*. Thessaloniki, Greece, 2020.

[KVB88] N. Kanopoulos, N. Vasanthavada, and R.L Baker. "Design of an image edge detection filter using the Sobel operator". In: *IEEE Journal of Solid-State Circuits* 23.2 (1988), pp. 358–367.

[LeC+89] Y. LeCun, B. Boser, J. S. Denker, et al. "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551.

[Lee+20] J. Lee, J. Kang, S. Park, D. Jang, and J. Lee. "A Multi-Class Classification Model for Technology Evaluation". In: *Sustainability* 12.15 (2020).

[Li+18] Y. Li, X. Liang, Z. Hu, and E. P. Xing. "Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[Lit+17] G. Litjens, T. Kooi, B. E. Bejnordi, et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[Liu+21] H. Liu, Z. Dai, D. R. So, and Q. V. Le. "Pay Attention to MLPs". In: (2021). arXiv: 2105.08050.

[Mik+13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 2013.

[MYA13] J. G. Mork, A. J. Yepes, and A. R. Aronson. "The NLM Medical Text Indexer System for Indexing Biomedical Literature". In: *Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013*. Vol. 1094. CEUR Workshop Proceedings. Valencia, Spain: CEUR-WS.org, 2013.

[NF16] M. Noroozi and P. Favaro. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles". In: *Computer Vision – ECCV 2016*. Amsterdam, NL, 2016, pp. 69–84.

[Pav+21] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail. "Diagnostic Captioning: A Survey". In: *ArXiv* abs/2101.07299 (2021).

[Pel+21] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, et al. "Overview of the ImageCLEFmed 2021 Concept and Caption Prediction Task". In: *CLEF2021 Working Notes*. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, 2021.

[Pha+21] H. Pham, Z. Dai, Q. Xie, M. Luong, and Q. V. Le. "Meta Pseudo Labels". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11557–11568.

[Raj+17a] P. Rajpurkar, J. Irvin, K. Zhu, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: (2017), arXiv:1711.05225.

[Raj+17b] P. Rajpurkar, J. A. Irvin, A. Bagul, et al. "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs." In: *arXiv: Medical Physics* (2017).

[Ram+19] P. Ramachandran, N. Parmar, A. Vaswani, et al. "Stand-Alone Self-Attention in Vision Models". In: *Advances in Neural Information Processing Systems*. Vol. 32. Vancouver, Canada, 2019, pp. 68–81.

[RHW86] D. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[Rob01] R. J. Roberts. "PubMed Central: The GenBank of the published literature". In: *Proceedings of the National Academy of Sciences* 98.2 (2001), pp. 381–382.

[Ros57] F. Rosenblatt. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[Ros63] F. Rosenblatt. "Principles of Neurodynamics. Perceptrons and the theory of brain mechanisms." In: *American Journal of Psychology* 76 (1963), p. 705.

[Sha+21] A. K. Sharma, A. Nandal, A. Dhaka, and R. Dixit. "Medical Image Classification Techniques and Analysis Using Deep Learning Networks: A Review". In: *Health Informatics: A Computational Perspective in Healthcare*. Singapore: Springer Singapore, 2021, pp. 233–258.

[Soh16] K. Sohn. "Improved Deep Metric Learning with Multi-class N-pair Loss Objective". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.

[Son+20] R. Sonker, A. Mishra, P. Bansal, and A. Pattnaik. “Techniques for Medical Concept Detection from Multi-Modal Images”. In: *CLEF2020 Working Notes*. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, 2020.

[Sow+21] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar. “MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models”. In: *Medical Imaging with Deep Learning*. 2021.

[TL19] M. Tan and Q. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, CA, USA, 2019, pp. 6105–6114.

[Tol+21] I. Tolstikhin, N. Houlsby, A. Kolesnikov, et al. “MLP-Mixer: An all-MLP Architecture for Vision”. In: (2021). arXiv: 2105.01601.

[Tou+21] H. Touvron, P. Bojanowski, M. Caron, et al. “ResMLP: Feedforward networks for image classification with data-efficient training”. In: (2021).

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA, 2017, pp. 5998–6008.

[VS17] Leonidas Valavanis and Spyridon Stathopoulos. “IPL at ImageCLEF 2017 Concept Detection Task”. In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*. Vol. 1866. 2017.

[Wan+17] X. Wang, Y. Peng, Le. Lu, et al. “ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, 2017, pp. 3462–3471.

[Wil+12] D. Wilken, X. Baur, L. Barbinova, et al. “What are the benefits of medical screening and surveillance?” In: *European Respiratory Review* 21.124 (2012), pp. 105–111.

[Wu+18] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. “Unsupervised Feature Learning via Non-Parametric Instance Discrimination”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[Zha+18] Y. Zhang, X. Wang, Z. Guo, and J. Li. “ImageSem at ImageCLEF 2018 Caption Task: Image Retrieval and Transfer Learning”. In: *CLEF2018 Working Notes*. CEUR Workshop Proceedings. Avignon, France: CEUR-WS.org, 2018.

[Zha+20] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. “Contrastive Learning of Medical Visual Representations from Paired Images and Text”. In: (2020). arXiv: 2010.00747.

[Zha+21a] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. “Scaling Vision Transformers”. In: (2021).

[Zha+21b] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning*. 2021.

[Zon+21] F. Zong, H. Shenghua, R. Su, W. Xiaowei, and L. Hua. “Deep learning-based multi-class COVID-19 classification with x-ray images”. In: 11598 (2021). Ed. by Cristian A. Linte and Jeffrey H. Siewerdsen, pp. 286–296.

List of Acronyms

MLP Multi-Layer Perceptron

CNN Convolutional Neural Network

ViT Visual Transformer

CV Computer Vision

SoTA State of the art

FFNN Feed Forward Neural Network

NLP Natural Language Processing