ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΜΕΤΑΠΤΥΧΙΑΚΟ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
MSc IN COMPUTER SCIENCE

DEPARTMENT OF INFORMATICS

M.Sc. IN COMPUTER SCIENCE

M.Sc. Thesis

“Medical Image Labeling and Report Generation”

Vasiliki Kougia

F3321805

Supervisor: Ion Androutsopoulos

Assistant supervisor: John Pavlopoulos

ATHENS, SEPTEMBER 2019

Acknowledgements

I would like to sincerely thank my supervisor Ion Androutsopoulos for the opportunity he gave me to work on this interesting field, his support and his valuable advice. I would also like to express my heartfelt thanks to my second supervisor, John Pavlopoulos for the guidance and encouragement he offered me through our co-operation, as well as the time he invested and his positive energy. In addition, I would like to express my gratitude to the members of the AUEB Natural Language Processing Group for their help and interesting discussions we had and for always cheering me up. Finally, a big thanks goes to my family for their support and especially, my friends for always being there and believing in me.

Abstract

In this thesis, we experiment with medical image tagging and text generation tasks, with the ultimate goal of helping physicians who need to generate reports from medical images. First, we address the task of medical image tagging, which aims to identify medical terms (tags) describing an image. Towards this end, we develop models employing deep learning and retrieval-based methods and apply them to two datasets that consist of X-ray exams. Second, we implement image captioning models in order to produce text that describes the findings present in an image and the final diagnosis. We also examine current datasets and methods and present results of our models evaluated with widely used measures.

Contents

1 Introduction

Medical professionals examine a great amount of medical images daily, e.g., PET/CT scans or radiology images, to conclude to a diagnosis and write their findings as medical reports. This process is time-consuming and in many cases there are not enough experienced clinicians to deal with its difficulties. In order to help the diagnostic process, Natural Language Processing (NLP) and Computer Vision techniques, combined with recent advances in deep learning, can greatly assist the interpretation of biomedical images and generation of reports. Automatic methods can reduce medical errors (e.g., suggesting findings to inexperienced physicians) and benefit medical departments by reducing the cost per exam (Bates et al., 2001; Lee et al., 2017).

There are many tasks applied to medical images that can assist clinicians during medical examinations, like classification (e.g, normal or abnormal), lesion detection, segmentation of affected organs etc. (Litjens et al., 2017). Two tasks that can assist the diagnostic process of describing the findings of an image are medical image tagging and medical image captioning. In medical image tagging the task is to assign medical terms (tags) to an image (Figure 2.1), assisting physicians to focus on interesting image regions (Shin et al., 2016). In medical image captioning, the task is to generate from each medical image a text, which can be a single sentence or paragraphs like the full-text reports written by radiologists (Figure 3.3) describing the findings (Jing et al., 2018). Despite the importance of these two tasks, related resources are not always easily accessible and the methods applied are currently limited.

1.1 Contribution

Recently, there is a growing interest in the automatic analysis of medical images and the generation of diagnostic medical reports from the images (Jing et al., 2018; Liu et al., 2019). However, because of the challenging nature of the biomedical domain, there are many difficulties that need to be addressed, especially regarding the datasets (Oakden-Rayner, 2019) and the evaluation measures of medical image captioning (Kilickaya et al., 2016). We study the datasets available for medical image tagging and captioning and their challenges.

Also, we implement and evaluate models for both tasks and examine their results to draw useful conclusions.

1.2 Outline

The rest of the thesis is organized as follows:

- Chapter 2 describes the medical image tagging task, the datasets, the models we applied and their results.
- In Chapter 3, we examine the task of medical image captioning. We present the models we implemented and analyse their results.
- In Chapter 4, we examine the task of jointly tagging medical images and generating captions from the images and their tags. Again, we discuss the models we implemented and their performance.
- Chapter 5 concludes and proposes directions for future work.

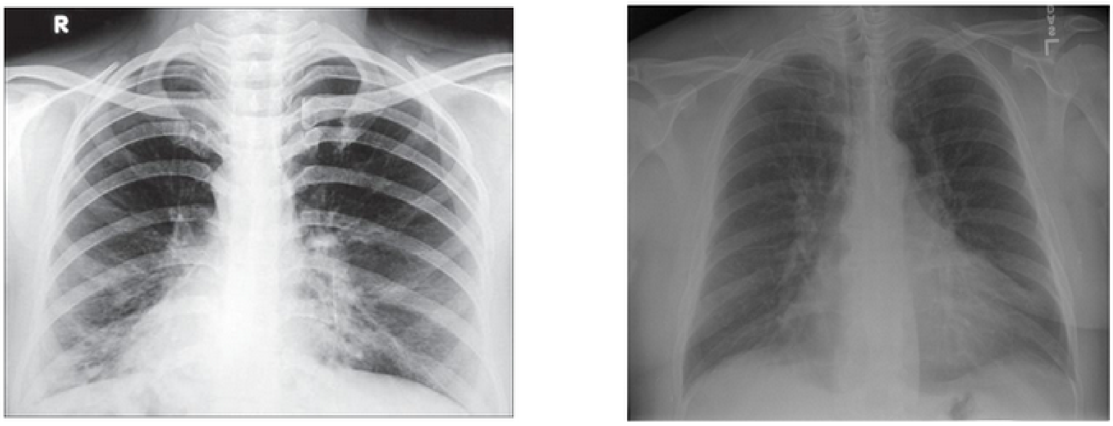
2 Medical image diagnostic tagging

2.1 Task Description

The first step towards the interpretation of medical images is to identify the visualized abnormalities. This is a tagging task, where given images of medical examinations such as X-rays, PET/CT scans etc., we identify tags that describe the findings and assign them to the corresponding image (Figure 2.1). These tags represent keywords from a large and continuously growing list that accurately describes the images. In effect, this is a multi-label classification task, where the labels (classes) are the available terms. In this chapter, we address the medical image tagging problem with multi-label classification methods and retrieval methods in order to decide the presence of each diagnostic tag in each image. This was the aim of the ImageCLEFmed 2019 Caption task (Pelka et al., 2019), in which we participated with four systems, achieving the best results and being ranked 1st, 2nd, 3rd, and 5th, among 60 systems submitted from 10 teams (Kougia et al., 2019b). We describe these systems below and present their results on one more dataset, which we also use later for text generation.

2.2 Datasets

There are many datasets covering a wide range of medical examination images (X-rays, MRIs, PET/CT scans etc.) and body parts (chest, brain, abdomen etc.), which can be used for several tasks like classification (e.g., classify an examination into normal or abnormal), object or lesion detection and segmentation (Litjens et al., 2017). Here, we will briefly describe the most popular and recent datasets annotated either with a number of classes that can be used for standard classification tasks or with a large number of terms representing keywords that describe the images, as shown in Table 2.1. Also, two of the 8 datasets shown in Table 2.1 were released along with a competition (MURA by Rajpurkar et al. (2017a) and CheXpert by Irvin et al. (2019)) and the three ImageCLEF datasets were created for the ImageCLEFmed Caption tasks (Eickhoff et al., 2017; de Herrera et al., 2018; Ionescu et al., 2019). These competitions encourage the implementation of new systems in order to develop an automatic method and assist the diagnostic process



Concepts		Tags	
CUI	**UMLS Term**	**MTI**	**Manual**
C0043299	x-ray procedure	Cardiomegaly	Cardiomegaly
C1548003	radiograph		Pulmonary Congestion
C0817096	thoracics		
C1962945	radiogr		

Figure 2.1: Images from the ImageCLEFmed 2019 dataset (left) and the IU X-ray dataset (right) and their assigned tags.

followed by clinicians.

ChestXray14

X-rays is the most commonly performed examination and often radiologists are called to examine a large number of them daily (Wang et al., 2017; Irvin et al., 2019; Jing et al., 2018). Chest X-rays in particular, are important for the detection of pneumonia and other thoracic diseases. ChestX-ray14 (Wang et al., 2017) was the first publicly available dataset of the size of hundreds of thousands and is until today a popular dataset used in many publications (Rajpurkar et al., 2017b, 2018; Wang et al., 2018).[1] It consists of 112,120 chest X-rays, each one classified into one or more of 14 disease classes. These labels were extracted from the diagnostic reports using term extraction tools (merged

[1]ChestX-ray14 was initially called ChestX-ray8 and contained 8 disease classes.

results from the DNorm (Leaman et al., 2015) and Metamap (Aronson and Lang, 2010) tools).

However, ChestX-ray14 suffers from a main problem that is encountered when using automatic ways to extract tags or classes (silver annotations) in order to create an annotated medical dataset. Term extraction methods are not 100% accurate and can result to false or ambiguous labels, so it is necessary to have a radiologist check the tags (ground truth) of the dataset (Oakden-Rayner, 2019), either by visual review of the images or by reading the original reports. Another good practice that helps to face this problem is to provide a test set that is manually labeled by clinical experts (gold annotations). This way, models trained on the inaccurate data with the silver annotations will show their weaknesses when tested on the test set with the gold annotations. The authors of ChestX-ray14 only evaluated their extraction method on a different set of images, which were manually tagged, collected from Open Access Biomedical Image Search Engine (OpenI).[2] So, they did not actually evaluated the silver annotations of their own dataset, which as later reviews show (Oakden-Rayner, 2019; Rajpurkar et al., 2017b) were inaccurate. Also, they did not provide their own manually labeled test set.

MURA, CheXpert and MIMIC-CXR

The following datasets were manually evaluated by radiologists in order to assess the accuracy of their labels. All the datasets, even MURA, which is manually labeled have a test set with additional labels from certified radiologists.[3] The MURA dataset (Rajpurkar et al., 2017a) was released by a team of the Stanford Machine Learning Group and consists of 14,863 studies from 7 different locations: elbow, finger, forearm, hand, humerus, shoulder, and wrist. Each study contains one or more radiographic images and is classified by a radiologist as normal or abnormal. CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019) are the most recently released datasets with a large amount of chest X-rays. Both were labeled by applying the custom rule-based labeler created by the authors of CheXpert (Irvin et al., 2019), called CheXpert labeler, to the diagnostic

[2] https://openi.nlm.nih.gov/

[3] However, the manually labeled test sets are not publicly available. In some cases, results on the hidden test set can be obtained after submitting executable code to the dataset competition.

reports with 14 classes (13 abnormalities and a "No Finding" tag).[4] For each label the decision can be blank for unmentioned, 0 for negative, -1 for uncertain, and 1 for positive.

Dataset	Images	Cases	Classes	Tags
ChestX-ray14	112,120 X-rays	-	14	-
MURA	40,561 X-rays	14,863	Binary	-
IU X-Ray	7,470 X-rays	3,955	-	MTI & Manual
ICLEF2017	184,614 Medical Images	-	-	20,463 UMLS CUIs
ICLEF2018	232,305 Medical Images	-	-	111,155 UMLS CUIs
ICLEF2019	80,786 Radiology Images	-	-	5,528 UMLS CUIs
MIMIC-CXR	371,920 X-rays	227,943	14	-[5]
CheXpert	224,316 X-rays	65,240	14	-

Table 2.1: Datasets containing medical images (X-rays, pathology images etc.) and classes or tags. A case represents a medical exam that contains one or more images and one overall diagnosis.

ImageCLEF Concept Detection datasets

The ImageCLEF 2017, 2018 and 2019 datasets were the datasets of the corresponding ImageCLEF concept detection tasks (Eickhoff et al., 2017; de Herrera et al., 2018; Pelka et al., 2019). In ImageCLEFmed Caption 2017 (Eickhoff et al., 2017) and 2018 (de Herrera et al., 2018), the datasets were noisy. They included generic and compound images, covering a wide diversity of medical images; there was also a large total number of concepts (111,155) and some of them were too generic and did not appropriately describe the images (Zhang et al., 2018). They also included captions, which we will describe later in Chapter 3.

The ImageCLEFmed Caption 2019 dataset is a subset of the Radiology Objects in COntext (ROCO) dataset (Pelka et al., 2018). It consists of medical images extracted from open access biomedical journal articles of PubMed Central,[6] with compound and non-radiology images filtered out using a Convolutional Neural Network (CNN) model. The CNN model was trained using the noisy ImageCLEF datasets of 2017 and 2018. Also, the total number of Unified Medical Language System (UMLS) concepts was reduced to 5,528, with 6 concepts assigned to each training image on average. Each image was extracted along with its caption. The caption was processed using QuickUMLS (Soldaini and Goharian,

[4] https://github.com/stanfordmlgroup/chexpert-labeler
[5] There are tags present in the reports, but we are not aware of the statistics.
[6] https://www.ncbi.nlm.nih.gov/pmc/

2016) to produce the gold UMLS concept unique identifiers (CUIs). An image can be associated with multiple CUIs and each CUI is accompanied by its corresponding UMLS term (Figure 2.1).

IU X-ray

The first dataset with publicly available reports was IU X-ray (Demner-Fushman et al., 2015) with access provided through the OpenI Search Engine.[7] The dataset contains 3,955 radiology reports, which correspond to 7,470 frontal and lateral chest X-rays. Each report consists of four sections (indication, comparison, findings and impression, which we will describe in Chapter 3) and two types of tags (Figure 2.1). First, there are manual tags that were assigned by two trained coders and comprise Medical Subject Heading (MeSH)[8] and RadLex[9] terms. Second, the Medical Text Indexer (MTI) was used to extract automated tags from the 'findings' and 'impression' sections of the reports, which represent only MeSH terms.[10] The MTI labeler does not handle negation, so the authors used MetaMap (Aronson and Lang, 2010) to detect tags that showed negation and discarded them.

Discussion

The major difference of the medical image tagging datasets from the general image datasets, such as ImageNet (Deng et al., 2009) and MSCOCO (Lin et al., 2014), is the presence of cases, which can comprise more than one image, but only one medical report (Table 2.1). Since the images of a case do not have an independent output, the actual size of these datasets is the number of different cases which is much smaller than the number of the existing images. Using each image of the same case as a different instance (image-based approach) would be wrong because the diagnosis was conducted by the radiologist by looking at all the case's images (Shin et al., 2016; Jing et al., 2018). Hence, we cannot assume findings of the report are present in every image. To address this challenge a simplistic approach is to first obtain the probability distributions over the tags from the classifier for each image of the case separately and then decide the labels for the case

[7] https://openi.nlm.nih.gov/
[8] https://goo.gl/iDvwj2
[9] http://www.radlex.org/
[10] https://ii.nlm.nih.gov/MTI/

by averaging the obtained probabilities (Rajpurkar et al., 2017a). This method has the disadvantage that a tag detected in only one image may not be chosen. In a more recent approach, Irvin et al. (2019) propose to choose the tags with the largest probabilities per image. Another solution would be to extract one embedding for every image of the case and then add or average them to obtain one embedding that will represent the whole case and will be fed to the classifier. This approach was used by Li et al. (2018) for medical image captioning, which we will examine in the next chapter. We chose to train our models using the image-based approach for the training and validation sets, because of its simplicity, despite its limitations. For the tuning of thresholds during inference and the final prediction of the tags on the test set we use an image-based as well as a case-based approach and compare both. We describe our approach in more detail in Section 2.5.

2.3 Related Work

Deep learning methods are widely used for classification tasks in the medical domain. Usually CNNs pre-trained on ImageNet (Deng et al., 2009) are used as image encoders followed by a Feed Forward Neural Network (FFNN) that serves as a classifier (Esteva et al., 2017; Rajpurkar et al., 2017b, 2018). However, ImageNet contains images that are photographs of various scenes and are very different from medical images, so the CNN models are fine-tuned to achieve better results. Esteva et al. (2017) fine-tuned the Inception V3 CNN model to classify skin lesions into malignant or benign, achieving results close to the ones predicted by dermatologists. This showed that CNNs pre-trained on ImageNet can be used in medical imaging tasks when fine-tuned and perform well, despite the differences between general and medical images.

CheXNet (Rajpurkar et al., 2017b) also follows a deep learning approach to classify X-rays of the ChestX-ray 14 dataset (Wang et al., 2017) to 14 labels of thoracic diseases. Rajpurkar et al. (2017b) uses DenseNet-121 (Huang et al., 2017) to encode images, adding a FFNN to assign one or more of the 14 classes to each image. The authors evaluated the predicted results with the F1 metric and reported state of the art results. In a subsequent work, Rajpurkar et al. (2018) presented CheXNeXt which consisted of an ensemble of 10 networks with the same architecture as CheXNet. First, an ensemble from multiple CheXNet networks is used to relabel the ChestX-ray dataset in to order to correct its false labels. Then, the networks are trained again, now on the relabeled data and an ensemble of the 10 best is used for the final predictions.

Retrieval methods have also been used for the medical image tagging task. The ImageCLEF Concept Detection tasks aim to detect abnormalities in medical images. The systems that participated in the competition in 2017, 2018 and 2019 (de Herrera et al., 2018; Pelka et al., 2019) employed deep learning as well as retrieval methods. In 2017, the top 10 systems belonged to the same team (Valavanis and Stathopoulos, 2017) and used retrieval methods that outperformed all the other deep learning systems. Valavanis and Stathopoulos (2017) experimented with several ways to represent the images (localized compact features, bag of visual words (BoVW) and bag of colors (BoC)) and used k-NN to retrieve the most similar training images for each test image. Then, they gave a score to each of the concepts of the k images based on their frequency in the k images or by using the Random Walk with Restart (RWR) algorithm (Wang et al., 2006). Finally, they assigned the concepts with highest scores to the test image. In 2018, the first ranked team (Pinho and Costa, 2018) used an adversarial auto-encoder for unsupervised feature learning, while the second ranked team (Zhang et al., 2018) used a multi-label classification method with the Inception V3 CNN as the encoder. Both teams also used retrieval methods, but they achieved lower results with the retrieval methods. In 2019, we submitted four systems to the Concept Detection subtask (Kougia et al., 2019b). The first ranked system we submitted followed a deep learning approach. We were the only team that used a retrieval method, which outperformed all other teams and was only slightly worse than our winning deep learning system.

2.4 Models implemented in this thesis

2.4.1 DenseNet-121 Encoder + 1-NN Image Retrieval

This system is based on a retrieval approach, which we have also used in our previous work on medical image captioning (Kougia et al., 2019a). Given a test image, the 1-NN returns the caption of the most similar training image, using a CNN encoder to map each image to a dense vector. Here, it returns the tags of the most similar training image.

To encode the images we use the DenseNet-121 (Huang et al., 2017), a CNN with 121 layers, where all layers are directly connected to each other improving information flow and avoiding vanishing gradients. We start with DenseNet-121 pre-trained on ImageNet (Deng et al., 2009) and fine-tune it on our medical dataset to achieve better results.[11] The fine-tuning process was performed as when training DenseNet-121 in our CheXNet-based system described in Section 2.4.3, including data augmentation. The images are rescaled to 224×224 and normalized with the mean and standard deviation of ImageNet to match the requirements of DenseNet-121 and how it was pre-trained on ImageNet. We extract the image embeddings from the last average pooling layer, which is the last layer before the classifier.

First, we obtain the image embeddings of all training images. Then, given a test image we compare its image embedding with the ones of the training images, using cosine similarity. We retrieve the most similar training image and assign its tags to the test image.

2.4.2 DenseNet-121 Encoder + k-NN Image Retrieval

We now extend the previous 1-NN baseline to retrieve the k-most similar training images and use their concepts, as follows. The encoder is the same DenseNet-121 CNN that was described in Section 2.4.1. The same encoding process is used to obtain the image embeddings, but with a different tag assignment method. Given a test image (Fig. 2.2), we again use the fine-tuned DenseNet-121 to obtain the image's embedding. We then retrieve the k training images with the highest cosine similarity (computed on image embeddings) to the test image, and return the r concepts that are most frequent among the concepts of the k images. We set r to the average number of concepts per image of the particular k retrieved images. In particular cases where the dataset contains images with no abnormalities, if more than half of the retrieved images are normal, then the test image is classified as normal as well. We tune the value of k in the range from 1 to 200 using the validation set.

[11] We used the implementation of https://keras.io/applications/#densenet.

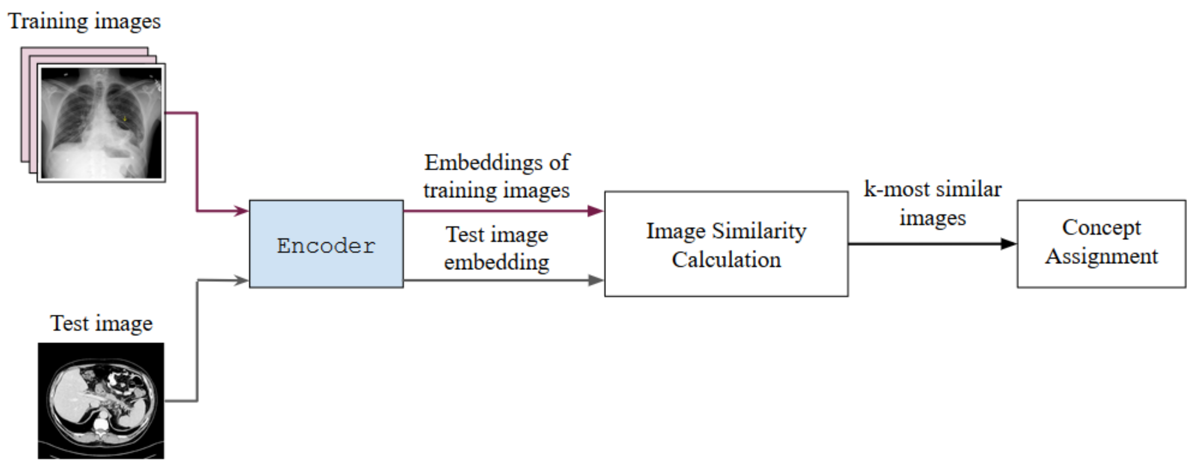


Figure 2.2: Illustration of how our DenseNet-121 encoder and k-NN image retrieval system works at test time.

2.4.3 CheXNet-based, DenseNet-121 Encoder + FFNN

This system, which is based on CheXNet (Rajpurkar et al., 2017b), achieved the best results in ImageCLEFmed Caption 2019. We re-implemented CheXNet in Keras[12] using the DenseNet-121 CNN encoder and adding an FFNN to assign one or more of the tags of a particular dataset to each image.

All images are again rescaled to 224×224 and normalized using the mean and standard deviation values of ImageNet. Also, during training, the images are augmented by applying random horizontal flip. After fine-tuning DenseNet-121 on our dataset, image embeddings are again extracted from the last average pooling layer. In this system, however, the image embeddings are then passed through a dense layer with one output for each tag and sigmoid activations to produce a probability per tag. We trained the model by minimizing the binary cross entropy loss. We used Adam (Kingma and Ba, 2014) with its default hyper-parameters, early stopping on the validation set, and patience of 3 epochs. We also decayed the learning rate by a factor of 10 when the validation loss stopped improving.

At test time, we predict the tags for each test image using their probabilities, as estimated by the trained model. For each tag, we assign it to the test image if the corresponding predicted probability exceeds a threshold t. We use the same t value for all the tags, which resulted from tuning on the validation set.

[12] https://keras.io/

2.4.4 Based on Jing et al., VGG-19 Encoder + FFNN

This system is based on the work of Jing et al. (2018). They presented an encoder-decoder model to generate tags and medical reports from medical images. We fully describe the model of Jing et al. (2018) in Chapter 3. It uses a VGG-19 (Simonyan and Zisserman, 2014) image encoder, a multi-label classifier to produce tags from the images, and a hierarchical Long Short-Term Memory (LSTM) network that generates texts by attending on both image and tag embeddings. Here, we describe a simplified version of the first part of Jing et al.'s model, the part that performs multi-label image classification that we implemented in Keras.

Again, we rescale the images to 224×224 and normalize them using the mean and standard deviation of ImageNet. We feed the resulting images to the VGG-19 CNN, which has 19 layers and uses small kernels of size 3×3. We used VGG-19 pre-trained on ImageNet.[13] The output of the last fully connected layer of VGG-19 is then given as input to a dense layer with a softmax activation to obtain a probability distribution over the tags. The model is trained using categorical cross entropy, which is calculated as:

$$E = -\sum_{i=1}^{|C|} y_{true,i} \log_2(y_{pred,i}) \tag{2.1}$$

where C is the number of tags, y_{true} is the ground truth binary vector of a training image, and y_{pred} is the predicted softmax probability distribution over the tags C for the training image. Categorical cross entropy sums loss terms only for the gold concepts of the image, which have a value of 1 in y_{true}. When using softmax and categorical cross-entropy, usually y_{true} is a one-hot vector and the classes are mutually exclusive (single-label classification). To use softmax with categorical cross entropy for multi-label classification, where y_{true} is binary but not necessarily one-hot, the loss is divided by the number of gold labels (true concepts) (Gong et al., 2014; Mahajan et al., 2018). Jing et al. (2018) achieve this by dividing the ground truth binary vector y_{true} by its L1 norm, which equals the number of

[13] https://keras.io/applications/#vgg19

gold labels. Hence, the categorical cross-entropy loss is computed as follows:

$$E = -\sum_{i=1}^{|C|} \frac{y_{true,i}}{\parallel y_{true} \parallel_1} \log_2(y_{pred,i}) = -\frac{1}{M}\sum_{j=1}^{M} \log_2(y_{pred,j}) \tag{2.2}$$

where M is the number of gold labels (true tags) of the training image, which is different per training image. In this model, the loss of Eq. 2.2 achieved better results on the development set, compared to binary cross entropy with a sigmoid activation per tag. We used the Adam optimizer with initial learning rate 1e-5 and early stopping on the validation set with patience 3 epochs. We compute the average number of gold tags per training image and assign this number of tags to each given test image, selecting the ones with the highest probability scores.

2.5 Experimental Results

For the classification task we experimented with the models described above (DenseNet + 1-NN, DenseNet + k-NN, DenseNet + FFNN and VGG-19 + FFNN) on the IU X-ray and ImageCLEFmed 2019 datasets. The evaluation of the systems was conducted by computing the F1 scores on each test image (in effect comparing the binary ground truth vector y_{true} to the binary predicted tags vector y_{pred}) and then averaging over all test images (Figure 2.3). In cases where a test image has no gold tags and no predicted tags assigned, we set the F1 score to 1. That is, because if an image has no tags we assume that there is no abnormality present, so the correct choice for a model would be to not assign any abnormality tag. If the model succeeds then the maximum F1 score is assigned. Instead, if the model falsely assigns a tag then the F1 score is set to zero.

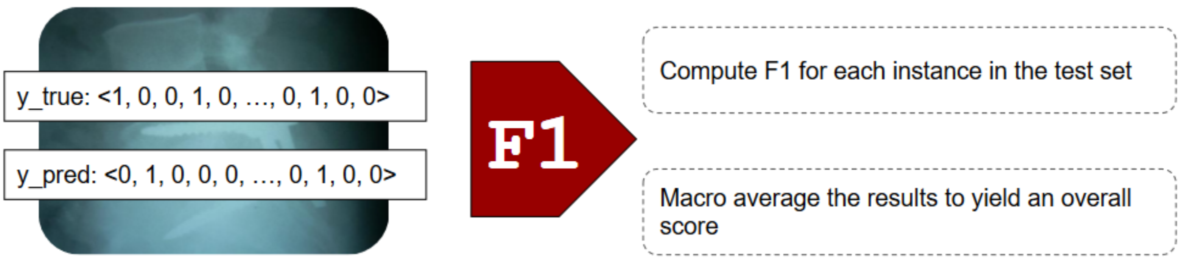


Figure 2.3: Example showing how the F1 evaluation works.

ImageCLEFmed Caption 2019

The provided training set of the ImageCLEFmed Caption 2019 contains 56,629 images, the validation set 14,157 and the remaining 10,000 images were used for testing. The gold concepts of the test set were not available so we randomly selected 20% of the training images (11,326 images along with their gold concepts) to serve as a development set, on which we evaluate our models during development. The models we used to produce the submitted results were trained on the entire training set. The validation set was used for hyper-parameter tuning and early stopping.

In the ImageCLEF2019 dataset all images are assigned at least one tag. In this dataset the concept of cases is not present as in other medical datasets (Table 2.1), since the images are randomly extracted from biomedical articles. Also, there is a large number of tags, many of which are non-informative e.g., 'improved', 'start', 'travel', so there are no tags that clearly state abnormalities or the absence of them as in IU X-ray. Table 2.2 reports the evaluation results of each system on the development set we created, since the ground truths of the test set are not available. The results on the hidden test set of the competition are also included.

IU X-ray

Since IU X-ray contains full diagnostic reports along with tags, we use it for our experiments in classification and text generation. During pre-processing we found that 104 reports contained no image and 25 reports were missing both the 'findings' and 'impression' sections, thus they could not be used for text generation. We discarded these reports, which led us to 3,826 reports with 7,430 images. We also observed that 1,726 reports did not contain MTI tags. From these, 1,354 were manually tagged as normal, which means that the MTI system failed to assign the abnormalities present in the remaining 372 reports. We performed experiments both with MTI and manual tags to compare the results and the quality of the two types of tags. There are 571 unique MTI tags and 1,679 manual tags (including the 'normal' tag). We used 80% of the dataset for training, 10% for validation and 10% for testing, splitting with respect to cases.

To deal with the fact that IU X-ray consists of cases that contain multiple images and only one report (see Section 2.2), we follow two approaches: an image-based, where each image of a case is an independent instance with the correct labels of the case replicated to all the images of the case and a case-based, where each instance is an entire case (all of its images) with the correct labels of the case. During training we follow the image-based approach, while for the evaluation we use both. In the image-based evaluation each image of a case is a separate test instance that is evaluated independently. The y_{true} binary vector of each image is the one of the case it belongs to and is compared to the y_{pred} binary vector predicted from a model in order to compute the F1 score (Figure 2.3). In the case-based evaluation, first the model predicts the tags for each image separately. Then, we take the union of the abnormality tags the model assigned and use them to produce the y_{pred} that is compared with the y_{true} of the case. In systems where we perform tuning (DenseNet + FFNN and DenseNet + k-NN) we also used case-based evaluation during tuning.

We perform classification on both types of tags present: MTI and manual (Table 2.2). Also, for the cases of the IU X-ray that are normal and have no assigned tags, the corresponding images have a y_{true} binary vector that contains only zeros (Figure 2.3).

Results

Table 2.2 shows the F1 scores achieved by our models on both datasets. In the ImageCLEFmed Caption 2019 dataset the DenseNet + FFNN (CheXNet-based) system outperforms all the other models on both the development and the test set.[14] However, the simple image retrieval system (DenseNet + k-NN) achieves very competitive results.[15]

In IU X-ray, DenseNet + FFNN system once again outperforms all the other systems on both types of tags and both types of evaluations. The main difference from the ImageCLEF 2019 results is that all the systems have now much higher results, but the DenseNet + k-NN system is not very close to the best CheXNet-based one as before. The much higher overall results are probably due to the reduced number of tags; 571 MTI tags and 1,679 manual tags of IU X-ray against the 5,528 tags of ImageCLEF2019. Also, the VGG-19 + FFNN system has very low results on this dataset due to the fact that

[14] The tuned value of the probability threshold for the DenseNet + FFNN system was 0.16.
[15] After tuning the resulted value of k was 199.

System	ICLEF2019		IU X-ray			
	Dev	Test	Image-based		Case-based	
			MTI	Manual	MTI	Manual
S1	0.138	-	0.244	0.190	0.166	0.140
S2	0.257	0.274	0.421	0.319	0.387	0.272
S3	**0.260**	**0.282**	**0.487**	**0.370**	**0.491**	**0.365**
S4	0.248	0.264	0.091	0.055	0.101	0.056

Table 2.2: Results of our systems DenseNet + 1-NN (S1), DenseNet + k-NN (S2), DenseNet + FFNN (S3) and VGG-19 + FFNN (S4) on the ImageCLEFmed Caption 2019 and IU X-ray datasets. We trained the S3 and S4 models three times and report the average F1 score of the results of these three models. We do not have the results of the DenseNet + 1-NN baseline on the test set, since it was not submitted in the competition. For IU X-ray we perform image-based and case-based evaluation on both types of tags (MTI and manually extracted tags).

there are normal images with a ground truth vector of all zeros. This system is trained using categorical cross entropy loss, so when the y_{true} binary vector contains only zeros then the loss will be zero (Section 2.4.4), causing the system not to be properly trained. Hence, systems that use this method for multi-label classification do not work correctly for datasets that contain images with no gold assigned tags.

Generally the classification using the manual tags has lower results than when using the MTI tags. This may be due to the fact that the manual tags are much more specific. 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.). This way, a system may correctly identify the abnormality but assign a tag that fails to describe its position correctly. This will have a zero F1 score, even though the abnormality was correct. However, the MTI tags cannot be trusted because they were produced in an automatic way, which was not evaluated by humans.

For the image-based and case-based evaluation we observe that the latter leads to lower results for the image retrieval systems (DenseNet + 1-NN and DenseNet + k-NN), while the deep learning systems are not significantly affected.

3 Medical image diagnostic text generation

3.1 Task Description

The next step, after having identified abnormalities in medical images, is to produce a medical report describing these abnormalities and concluding to a diagnosis. This process is similar to the image captioning task that has been applied to general images (Vinyals et al., 2015, 2016; Xu et al., 2015; Krause et al., 2017). Image captioning is the task of generating captions that describe the objects of an image and the relationships between them, as shown in Figure 3.1. This task has attracted a lot of interest and many methods have been suggested and successfully applied, achieving state of the art results (Anderson et al., 2018; Liu et al., 2018).

Figure 3.1: Example of general image captioning produced by the Show and Tell model (Vinyals et al., 2016).

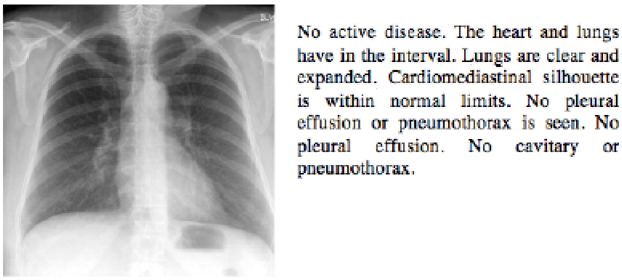


Figure 3.2: Example of medical image captioning produced by the model of Jing et al. (2018).

Recently, image captioning technology has been applied to medical images in order to produce a report: a diagnosis that describes the condition of the patient (Figure 3.2). The publications addressing this task, which we define as medical image captioning (Kougia et al., 2019a) are currently limited, but there is a growing interest since it can be helpful to medical experts.

3.2 Datasets

While there are many labeled medical datasets publicly available as mentioned in Chapter 2, the medical reports of their images are, in most cases, not available. There are only 4 datasets that provide reports, as shown in Table 3.1, which can be single sentences describing a diagnosis based on the images or longer structured medical reports (Figure 3.3). We do not include in Table 3.1 datasets that were used in publications of medical image captioning but were not publicly available, like the BCIDR dataset used by Zhang et al. (2017a,b).

ImageCLEF Caption Dataset

The ImageCLEF caption dataset was released in 2017 for the Image Concept Detection and Caption Prediction task (Eickhoff et al., 2017; de Herrera et al., 2018). The dataset consists of 184,614 medical images and their captions, extracted from biomedical articles of PubMed Central (PMC).[16] The organizers used an automatic method, based on a medical image type hierarchy (Müller et al., 2012), to classify the 5.8M extracted images as clinical or not and also discard compound ones (e.g., images consisting of multiple X-rays), but their estimation was that the overall noise in the dataset would be as high as 10% or 20% (Eickhoff et al., 2017). In 2018, the ImageCLEF caption organizers employed a Convolutional Neural Network (CNN) to classify the same 5.8M images based on their type and to extract the non-compound clinical ones, leading to 232,305 images along with their respective captions (de Herrera et al., 2018). Although they reported that compound images were reduced, they noted that noise still exists, with non-clinical images present (e.g., images of maps). Additionally, a wide diversity between the types of the images has been reported (Liang et al., 2017). The length of the captions varies from 1 to 816 words (Su et al., 2018; Liang et al., 2017). Only 1.4% of the captions are duplicates (associated with more than one image), probably due to the wide image type diversity. The average caption length is 21 words and the vocabulary size is 157,256. However, in 2019 the Caption Prediction task was not included in the ImageCLEFmed Caption challenge and the captions of the images were not public (see details about the ImageCLEFmed Caption

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

2019 dataset in Section 2.2).

PEIR Gross

The PEIR Gross dataset consists of images from the Pathology Education Informational Resource (PEIR) digital library, which is a public access image database for use in medical education.[17] Jing et al. (2018), who were the first to use images from this database, employed 7,442 teaching images of gross lesions (i.e., visible to the naked eye) from 21 PEIR pathology sub-categories, along with their associated captions.[18] In this dataset, each image is associated with one descriptive sentence. However, PEIR Gross seems less related to the task, since its images are photographs of medical incidents rather than images of an examination.

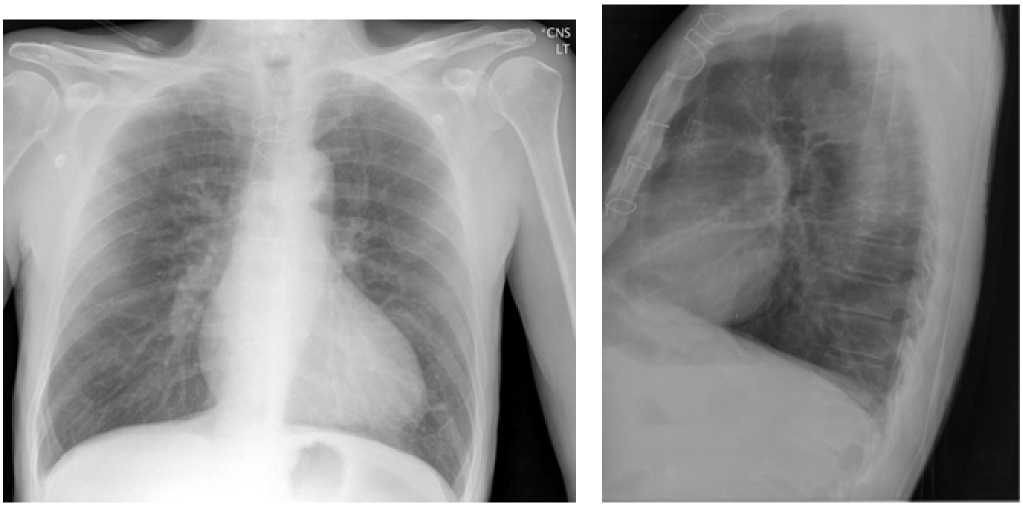
IU X-ray

The IU X-Ray collection (Demner-Fushman et al., 2015) provided by the Open Access Biomedical Image Search Engine (OpenI) contains frontal and lateral chest X-rays and reports that consist of the following sections: Comparison (previous information about the patient), Indication (symptoms, reasons of examination), Findings (the radiology observations), Impression (the final diagnosis) and tags (Figures 3.3 and 2.1). A system would ideally generate the 'findings' and 'impression' sections, possibly concatenated (Jing et al., 2018). We also follow this approach concatenating these sections.

As mentioned in Chapter 2, IU X-ray contains 3,955 radiology reports associated with 7,470 X-rays. After removing the 104 reports that contained no image and the 25 reports that were missing both 'findings' and 'impression' sections, the remaining reports are 3,826 with 7,430 images. We also used the same split, with respect to cases, using 80% of the dataset for training, 10% for validation and 10% for testing. During text pre-processing we removed punctuation and numbers, which resulted to 1820 unique words. To create the vocabulary for out text generation models we kept only words with frequencies higher than 3 in the training set, corresponding to 1,268 words. The maximum length of captions in the training set is 226 and the average length is 30. We split each report to sentences

[17] http://peir.path.uab.edu/library/

[18] PEIR pathology contains 23 sub-categories, but only 22 contain a gross sub-collection (7,443 images in total). We observe that one image was not included by Jing et al. (2018).



COMPARISON: Chest radiograph from XXXX, XXXX.
INDICATION: Chest pain.
FINDINGS: The cardiac silhouette is borderline enlarged. Otherwise, there is no focal opacity. Mediastinal contours are within normal limits. There is no large pleural effusion. No pneumothorax.
IMPRESSION: Borderline enlargement of the cardiac silhouette without acute pulmonary disease.

Figure 3.3: Example of a case from the IU X-ray dataset. It consists of a frontal and lateral chest X-ray and a report with four sections: Comparison, Indication, Findings and Impression.

using the sentence splitter of NLTK.[19] We observe that the average number of sentences per report is 5, while each sentence consists of 6 words on average.

MIMIC-CXR

Recently, the full text reports of the original image-tags pairs of the MIMIC-CXR dataset (Johnson et al., 2019) were made available. When the original MIMIC-CXR paper was published the reports were not available, so the paper does not mention details and statistics about the reports. Liu et al. (2019) are the first that used this dataset for medical image captioning and they report that there are 206,563 cases, each associated with a report that has the same structure as the IU X-ray reports (Comparison, Indication, Findings and Impression). This is the first dataset with hundreds of thousands of images, whose full-text reports have been made available and can be used for diagnostic text

[19] https://www.nltk.org/api/nltk.tokenize.html

generation.

Dataset	Images	Tags	Texts
MIMIC-CXR	473,057 X-rays	-	206,563 reports
ICLEF Caption	232,305 medical images	UMLS concepts	232,305 sentences
IU X-ray	7,470 X-rays	Manual & MTI tags	3,955 reports
PEIR Gross	7,443 teaching images	top TF-IDF caption words	7,443 sentences

Table 3.1: Medical datasets suitable for captioning.

Discussion

We observe that datasets of medical image captioning suffer from a major shortcoming, which is the large imbalance of classes. Most of the images report no findings. Even when there are findings (e.g., a disease), the wide range of possible diseases makes some of them appear rarely in the training data. For example, in the IU X-ray dataset the majority of cases are normal reporting 'No acute cardiopulmonary abnormality', which makes these words appear very often and others rather rarely. This could lead to reduced generalization power of the models trained over such datasets. The very specific vocabulary and limited diversity in the sentences causes the same captions to be generated frequently.

In medical image captioning the challenge of having medical cases with multiple images but only one overall diagnosis, has not been addressed in most of the existing approaches. The most common method is to use each image separately replicating the output diagnosis (Shin et al., 2016; Wang et al., 2018; Jing et al., 2018) (the image-based approach we also described in Chapter 2). As mentioned in Chapter 2, this is not correct because the findings described in the diagnosis may not be present in every image. Li et al. (2018) were the only ones who addressed this issue by feeding two images from every case of the IU X-ray dataset to a CNN encoder to extract two image embeddings. Then, they averaged these two embeddings and used the resulted embedding for the text generation. There are cases in IU X-ray with one or more than two images, but the authors do not mention these, so we assume they have been excluded. The approach we followed for case-based classification in Chapter 2, where we first predict the tags for each image and then concatenate and assign them to the case, cannot be applied here, since the output must now be a fluent text. This issue needs to be further researched, so for this thesis we adopt the image-based approach for training and evaluation of caption generation.

3.3 Related Work

Initial Approaches

A first attempt towards medical image captioning was made by Varges et al. (2012), who employed Natural Language Generation to assist medical professionals turn cardiological findings (e.g., from diagnostic imaging procedures) into fluent and readable textual descriptions. From a different perspective, Schlegl et al. (2015) used both the image and the textual report as input to a CNN, trained to classify images with the help of automatically extracted semantic concepts from the textual report. Kisilev et al. (2015a,b) employed a radiologist to mark an image lesion, and a semi-automatic segmentation approach to define the boundaries of that lesion. Then, they used structured Support Vector Machines (Tsochantaridis et al., 2004) to generate semantic tags, originating from a radiology lexicon, for each lesion. In a later work, Kisilev et al. (2016) used a CNN to detect Regions of Interest (ROIs) in the image and then fully connected layers to assign to each ROI predefined features describing abnormalities. Then, the assigned features were integrated to template sentences to form a caption.

Since 2016, inspired by deep learning methods and in particular the encoder-decoder architecture, which was successfully applied to general image captioning, a few publications have introduced models that generate diagnostic reports for medical images (Table 3.2). Shin et al. (2016) were the first that implemented the encoder-decoder architecture for medical images, to create a model that generates reports in the form of annotations. They used the IU X-ray dataset and a Network in Network (Lin et al., 2013) or GoogLeNet (Szegedy et al., 2015) as the encoder of the images, obtaining better results with GoogLeNet. The encoder was pre-trained to predict (from the images) 17 classes, corresponding to MeSH terms that were frequent in the reports and did not co-occur frequently with other MeSH terms. An LSTM or Gated Recurrent Unit (GRU) (Cho et al., 2014) was used as the Recurrent Neural Network (RNN) decoder to generate image descriptions from the image encodings. In a second training phase, the mean of the RNN's state vectors (obtained while describing each image) was used as an improved representation of each training image. The original 17 classes that had been used to pre-train the CNN were replaced by 57 finer classes, by applying k-means clustering to the improved vector representations of the training images. The CNN was then retrained to predict the 57 new classes and this led to improved BLEU (Papineni et al., 2002) scores for the overall CNN-RNN system. The generated descriptions, however, were not evaluated by humans. Furthermore, the generated descriptions were up to 5 words long and looked more like bags of terms (e.g., 'aorta thoracic, tortuous, mild'), rather than fluent coherent reports.

Attention-based Approaches

Subsequent work pointed out the importance of highlighting the findings described in the report on the image to make the diagnosis easily interpretable (Zhang et al., 2017b; Jing et al., 2018; Wang et al., 2018). To achieve that, attention mechanisms were incorporated to their proposed models. Zhang et al. (2017b) were the first to employ an attention mechanism in medical image to text generation, with the MDNet model.[20] They used a dataset with pathological bladder cancer images to generate their reports, which have the form of a small paragraph explaining the findings (BCIDR Dataset). MDNet used ResNet (He et al., 2016) for image encoding, but extending its skip connections to address vanishing gradients. The image representation acts as the starting hidden state of a decoder LSTM, enhanced with an attention mechanism over the image. (During training, this attention mechanism is also employed to detect diagnostic labels.) The decoder is cloned to generate a fixed number of sentences, as many as the symptom descriptions in the diagnostic paragraph, by taking as input the type of symptom description of the corresponding sentence. This model performed slightly better than a state of the art generic image captioning model (Karpathy and Fei-Fei, 2015) in most evaluation measures.

Jing et al. (2018) implemented an encoder-decoder architecture with attention. Their encoder (VGG-19 (Simonyan and Zisserman, 2014)), which we described in Section 2.4.4 is used to encode and extract equally sized patches from each image, where each patch is a 'visual' feature vector. A Multi-Layer Perceptron (MLP) is then fed with the visual feature vectors of each image (representing its patches) and predicts terms from a pre-determined term vocabulary. The word embeddings of the predicted terms of each image are treated as 'semantic' feature vectors representing the image. The decoder, which produces the text,

[20] Zhang et al. had introduced earlier TandemNet (Zhang et al., 2017a), which also used attention, but for medical image classification. TandemNet could perform captioning, but the authors considered this task as future work, that was addressed with MDNet.

is a hierarchical RNN, consisting of a sentence-level LSTM and a word-level LSTM. The sentence-level LSTM produces a sequence of embeddings, each specifying the information to be expressed by a sentence of the image description (acting as a topic). For each sentence embedding, the word-level LSTM then produces the words of the corresponding sentence, word by word. More precisely, at each one of its time-steps, the sentence-level LSTM of Jing et al. examines both the visual and the semantic feature vectors of the image. Following previous work on image captioning, that added attention to encoder-decoder approaches (Xu et al., 2015; You et al., 2016; Zhang et al., 2017b), an attention mechanism (an MLP fed with the current state of the sentence-level LSTM and each one of the visual feature vectors of the image) assigns attention scores to the visual feature vectors, and the weighted sum of the visual feature vectors (weighted by their attention scores) becomes a visual 'context' vector, specifying which patches of the image to express by the next sentence. Another attention mechanism (another MLP) assigns attention scores to the semantic feature vectors (that represent the terms of the image), and the weighted sum of the semantic feature vectors (weighted by attention) becomes the semantic context vector, specifying which terms of the image to express by the next sentence. At each time-step, the sentence-level LSTM considers the visual and semantic context vectors, produces a sentence embedding and updates its state, until a stop control instructs it to stop. Given the sentence embedding, the word-level LSTM produces the words of the corresponding sentence, again until a special 'stop' token is generated. Jing et al. showed that their model outperforms models created for general image captioning with visual attention (Vinyals et al., 2015; Donahue et al., 2015; Xu et al., 2015; You et al., 2016).

Wang et al. (2018) adopted an approach similar to that of Jing et al. (2018), using a ResNet-based CNN to encode the images and an LSTM decoder to produce image descriptions, but their LSTM is flat, as opposed to the hierarchical LSTM of Jing et al. (2018). Wang et al. also demonstrated that it is possible to extract additional image features from the states of the LSTM, much as Jing et al. (2018), but using a more elaborate attention-based mechanism, combining textual and visual information. Wang et al. experimented with the same OpenI dataset that Shin et al. and Jing et al. used. However, they did not provide evaluation results on OpenI and, hence, no direct comparison can be made against the results of Shin et al. and Jing et al. Nevertheless, focusing on experiments that generated paragraph-sized image descriptions, the results of Wang et al. on the (not publicly available) ChestX-ray14 dataset (e.g., BLEU-1 0.2860, BLEU-2 0.1597) are much worse than the OpenI results of Jing et al. (e.g., BLEU-1 0.517, BLEU-2 0.386), possibly because of the flat (not hierarchical) LSTM decoder of Wang et al.

Recent Approaches

Gale et al. (2018) argued that existing medical image captioning systems fail to produce a satisfactory medical diagnostic report from an image, and to provide evidence for a medical decision. They focused on classifying hip fractures in pelvic X-rays, and argued that the diagnostic report of such narrow medical tasks could be simplified to two sentence templates; one for positive cases, including 5 placeholders to be filled by descriptive terms, and a fixed negative one. They used DenseNet (Huang et al., 2017) to get image embeddings and a two-layer LSTM, with attention over the image, to generate the constrained textual report. Their results, shown in Table 3.2, are very high, but this is expected due to the extremely simplified and standardized ground truth reports. (Gale et al. report an improvement of more than 50 BLEU points when employing this assumption.) The reader is also warned that the results of Table 3.2 are not directly comparable, since they are obtained from very different datasets.

Li et al. (2018) developed a hybrid model that combines text generation and retrieval. First, the image embeddings are fed to a sentence decoder of stacked RNN layers that produces an embedding for each sentence of the report. For each of these embeddings that represent sentences, an agent, which is trained using reinforecement learning decides for each sentence if it will be generated or retrieved from a database of template sentences. Li et al. (2018) are the first to use one embedding for both images of each medical case by averaging the embeddings of each image. However, they do not mention cases that have one or more than two images.

ImageCLEF Caption Prediction Tasks

In addition to the above publications, the ImageCLEF Caption Prediction subtask ran successfully for two consecutive years (Eickhoff et al., 2017; de Herrera et al., 2018) but was not included in 2019. Participating systems (see Table 3.3) used image similarity

Method	Dataset	B1	B2	B3	B4	MET	ROU
Shin et al. (2016)	IU X-ray	78.5	14.4	4.7	0.0	-	-
Zhang et al. (2017b)	BCIDR	91.2	82.9	75.0	67.7	39.6	70.1
Wang et al. (2018)	Chest X-ray 14	28.6	15.9	10.3	7.3	10.7	22.63
	IU X-ray	-	-	-	-	-	-
Jing et al. (2018)	IU X-ray	51.7	38.6	30.6	24.7	21.7	44.7
	PEIR Gross	30.0	21.8	16.5	11.3	14.9	27.9
Zhang et al. (2017a)	BCIDR	-	-	-	-	-	-
Gale et al. (2018)	Frontal Pelvic X-rays	91.9	83.8	76.1	67.7	-	-
Han et al. (2018)	Lumbar MRIs	-	-	-	-	-	-

Table 3.2: Medical image captioning methods, the datasets they use and the most common evaluation measures: BLEU-1/-2/-3/-4 (B1, B2, B3, B4), METEOR (MET) and ROUGE (ROU). Zhang et al. (2017b) and Wang et al. (2018) used the Rouge-L version. Zhang et al. (2017a) and Han et al. (2018) did not provide any evaluation results of the generated reports. Wang et al. (2018) provided evaluation results only on Chest X-ray 14. While, as mentioned before, the ChestX-ray 14 dataset cannot be used for captioning because its reports are not publicly available, it seems that Wang et al. (2018) had access to the reports since they created the dataset.

to retrieve images similar to the one to be described, then aggregating the captions of the retrieved images; or they employed an encoder-decoder architecture; or they simply classified each image based on UMLS concepts and then aggregated the respective UMLS 'semantic groups'[21] to form a caption. Liang et al. (2017) used a pre-trained VGG CNN encoder and an LSTM decoder, similarly to Karpathy and Fei-Fei (2015). They trained three such models on different caption lengths and used an SVM classifier to choose the most suitable decoder for the given image. Furthermore, they used a 1-Nearest Neighbor method to retrieve the caption of the most similar image and aggregated it with the generated caption. Zhang et al. (2018), who achieved the best results in 2018, used the Lucene Image Retrieval software (LIRE) to retrieve images from the training set and then simply concatenated the captions of the top three retrieved images to obtain the new caption. Abacha et al. (2017) used GoogLeNet to detect UMLS concepts and returned the aggregation of their respective UMLS semantic groups as a caption. Su et al. (2018) and Rahman (2018) also employed different encoder-decoder architectures.

[21] https://goo.gl/GFbx1d

Team	Year	Approach	BLEU
Liang et al.	2017	ED+IR	26.00
Zhang et al.	2018	IR	25.01
Abacha et al.	2017	CLS	22.47
Su et al.	2018	ED	17.99
Rahman	2018	ED	17.25

Table 3.3: Top-5 participating systems at the ImageCLEF Caption Prediction competition, ranked based on average BLEU (%), the official evaluation measure. Systems used an encoder-decoder (ED), image retrieval (IR), or classified UMLS concepts (CLS). We exclude 2017 systems employing external resources, which may have seen test data during training (Eickhoff et al., 2017). 2018 models were limited to use only pre-trained CNNs.

3.4 Models implemented in this thesis

3.4.1 Nearest Neighbor

This is a baseline we implemented for medical image captioning (Kougia et al., 2019a) and also used for medical image tagging (see Section 2.4.1) modified to retrieve tags. This method is based on the intuition that similar biomedical images have similar diagnostic captions; this would also explain why image retrieval systems perform well in biomedical image captioning (Table 3.3). We use ResNet-18[22] to encode images, and cosine similarity to retrieve similar training images. The caption of the most similar retrieved image is returned as the generated caption of a new image. This baseline can be improved by employing an image encoder trained on biomedical images, such as X-rays (Rajpurkar et al., 2017b).

3.4.2 BlindRNN

Our first model is a baseline that we call BlindRNN. It is a flat RNN generating text as a language model that accepts a sequence of words and produces the next word (decoder part of Figure 3.4, without the image embedding). It consists of an LSTM layer with 512 units and a dense layer on top that outputs a probability distribution over the words of the vocabulary. We keep in the vocabulary words that occur more than three times in the training set and set the maximum caption length to 50 words. We insert a special

[22] https://goo.gl/28K1y2

start and end token in the start and the end of each caption respectively and pad each caption up to the maximum length. During training, BlindRNN takes as input only the gold caption, which serves as the ground truth. We use pre-trained biomedical word2vec embeddings that are frozen during training.[23] The model is trained using categorical cross entropy loss computed on the output probabilities of the dense layer and the ground truth binary vector representing the true next word. We use the Adam optimizer and set the initial learning rate to 0.0001. Early stopping is also used with a patience of three epochs.

During inference, for each test instance we start by feeding the start token to the LSTM and obtain the probabilities for the next word. We employ two ways to choose among these probabilities. First, we experiment with a greedy method in which we choose the word with the highest probability, and second we use a sampling method in which a word is sampled from the probability distribution. The sampled word is also given as input to the LSTM in order to produce the next word. This process continues until the end-of-caption token is produced or the maximum caption length is reached.

3.4.3 Encoder-Decoder

In this model we extend the previous BlindRNN to a full encoder-decoder architecture based on the Show and Tell model of Vinyals et al. (2015) (Figure 3.4).[24] The BlindRNN now serves as a decoder of the image embedding extracted from the encoder. For the encoder we use the DenseNet + FFNN (CheXNet-based) system we have implemented for classification (see Section 2.4.3) and has been fine-tuned on the IU X-ray training set. We extract each image embedding from the last average pooling layer and unlike Show and Tell, which uses the image only as the initial state of the LSTM, we feed it to the decoder at each time step, as shown in Figure 3.4 with the unrolled LSTM. The model is not trained end-to-end; we keep the encoder frozen and train the decoder as described in the previous section.

We also use the same inference process as in BlindRNN, but now we give as input to the LSTM the test image embedding along with the corresponding word at each time step.

[23] https://archive.org/details/pubmed2018_w2v_200D.tar.

[24] Our implementation of the Encoder-Decoder model is based on this tutorial: https://machinelearningmastery.com/develop-a-caption-generation-model-in-keras/

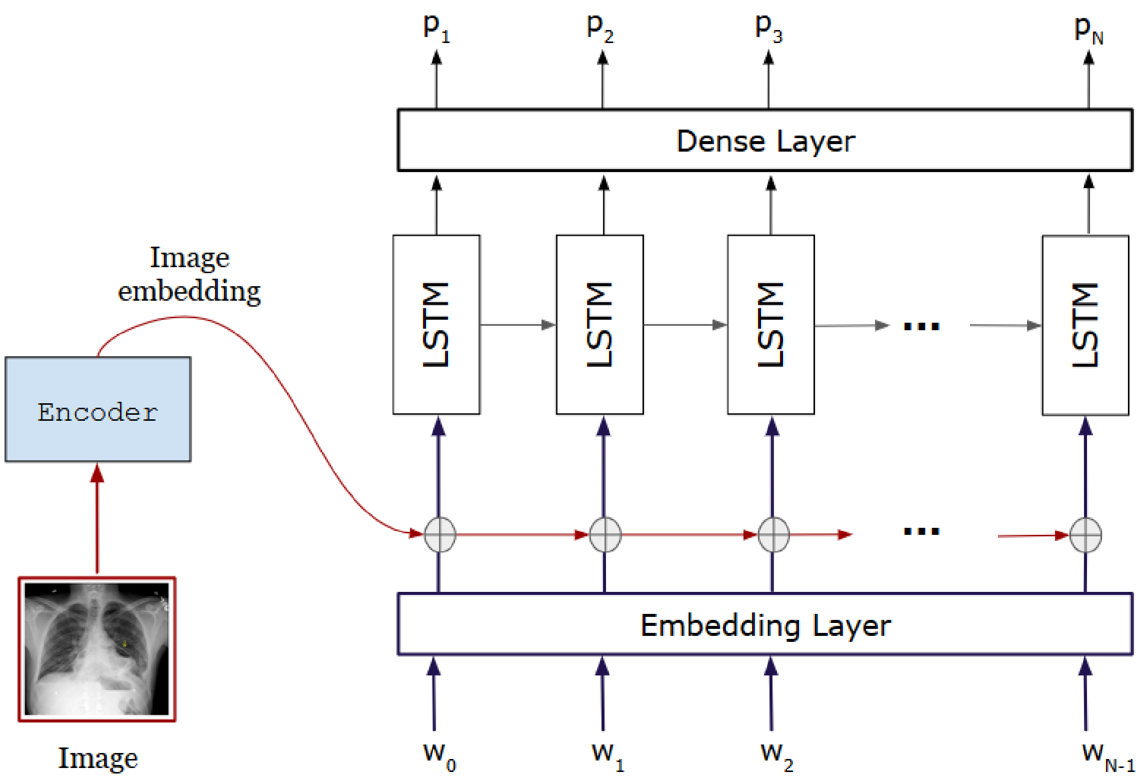


Figure 3.4: Illustration of how our encoder-decoder model works during training, where the LSTM is shown unrolled into time. This model is based on the Show and Tell model (Vinyals et al., 2015), with the difference that the image is not only used as the initial state of the LSTM. The image embedding is extracted from the last average pooling layer of the encoder and is concatenated at each time step with the word embedding of the current word.

3.5 Experimental Results

Evaluation Measures

The most common evaluation measures in medical image captioning are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), which originate from machine translation and summarization. The more recent CIDER measure (Vedantam et al., 2015), which was designed for general image captioning (Kilickaya et al., 2016), has been used in only two medical image captioning works (Zhang et al., 2017b; Jing et al., 2018). SPICE (Anderson et al., 2016), which was also designed for general image captioning (Kilickaya et al., 2016), has not been used in any medical image captioning work we are aware of.

We report results on BLEU-1, -2, -3, -4, METEOR and ROUGE-L (Table 3.4). BLEU is the most common measure (Papineni et al., 2002). It measures word n-gram overlap between the generated and the ground truth caption. A brevity penalty is added to penalize short generated captions. BLEU-1 considers unigrams (i.e., words), while BLEU-2, -3, -4 consider bigrams, trigrams, and 4-grams respectively. The average of the four variants was used as the official measure in the ImageCLEF Caption task. METEOR (Banerjee and Lavie, 2005) extended BLEU-1 by employing the harmonic mean of precision and recall (F-score), biased towards recall, and by also employing stemming (Porter stemmer) and synonymy (WordNet). To take into account longer subsequences, it includes a penalty of up to 50% when no common n-grams exist between the machine-generated description and the reference. ROUGE-L (Lin et al., 2013) is the ratio of the length of the longest common subsequence between the machine-generated description and the reference human description, to the size of the reference (ROUGE-L recall); or to the generated description (ROUGE-L precision); or a combination of the two (ROUGE-L F-measure). We note that several ROUGE variants exist, based on different n-gram lengths, stemming, stopword removal, etc., but ROUGE-L is the most commonly used variant in medical image captioning so far.

The main challenge for the text generation models in the medical field is caption diversity. Medical reports have a limited vocabulary and the same sentences occur many times. We observe that in the 5,930 training images there are 2,438 distinct captions, while in the 758 images of the test set there are only 354. This causes the text generation models to overfit and output the same caption for each test image or with only slight changes. To evaluate the ability of the models to generate different captions we report the diversity in the produced test captions (Table 3.4).

Results

Table 3.4 shows the results of our systems using the measures mentioned above. The medical reports we use in our models (concatenated 'impression' and 'findings' sections) comprise multiple sentences. So, for the flat RNNs we implemented (BlindRNN and Encoder-Decoder), we experimented with two approaches; first, we consider the caption as one unified sentence and second, we insert a special sentence delimiter token that shows when a new sentence starts within the whole caption. The sentence delimiter helps the systems produce more accurate captions and achieve better results, leading to improvements of at least two points in all the measures (Table 3.4). In the case of the Encoder-Decoder, the sentence delimiter also improves the diversity of the generated captions.

Also, we follow two inference approaches: greedy and sampling. In the greedy approach we choose the word with the highest probability to be generated at each time step. In the sampling approach the next word is randomly sampled from the probability distribution over the vocabulary produced by the model. As we see (Table 3.4) the sampling method yields much lower results than the greedy one in all the captioning measures, but achieves high caption diversity.

In the greedy inference method, the BlindRNN model produces only one caption, which is expected since at the inference phase it always takes as input the start token. This baseline actually produces the most frequent sentences for each test instance and performs better that the 1-NN baseline. On the other hand, the Encoder-Decoder model has much higher diversity since it takes information from each image. Overall, the system with the best results is the Encoder-Decoder, in particular the one that uses a sentence delimiter. However, the difference with the simple BlindRNN is not large, showing that the use of images did not help as much as expected.

Model	**S.D.**	Inference	B1	B2	B3	B4	Met	Rou	Diversity
1-NN	-	Greedy	28.2	15.2	9.0	5.6	12.7	21.1	**204**
BlindRNN	-	Greedy	33.2	20.6	13.6	9.5	14.1	27.0	1
BlindRNN	✓	Greedy	**37.9**	22.9	15.5	11.2	16.7	30.9	1
Encoder-Decoder	-	Greedy	33.6	20.9	13.8	9.7	14.3	27.3	54
Encoder-Decoder	✓	Greedy	37.1	**23.5**	**16.2**	**11.7**	**17.2**	**32.0**	103
BlindRNN	-	Sampling	27.1	14.7	8.4	4.9	11.5	21.0	**733**
BlindRNN	✓	Sampling	**31.7**	17.7	10.6	6.6	14.1	24.5	723
Encoder-Decoder	-	Sampling	28.1	15.8	9.6	6.1	12.5	23.2	697
Encoder-Decoder	✓	Sampling	31.4	**18.2**	**11.2**	**7.2**	**14.9**	**26.9**	722

Table 3.4: Results of the text generation systems on IU X-ray using greedy and sampling inference. We report results on the most common caption evaluation measures: BLEU-1/-2/-3/-4 (B1, B2, B3, B4), METEOR (MET) and ROUGE (ROU), as well as the diversity, which shows the number of different captions produced. For the BlindRNN and the Encoder-Decoder models we follow two approaches: with and without the special sentence delimiter (S.D.).

4 Diagnostic tagging for diagnostic captioning

The tasks we described in the previous chapters are similar in that they both aim to detect abnormalities in medical images and describe them. Medical image tagging achieves this by assigning keywords to the images, while medical image image captioning by producing text. In this chapter we examine if these two tasks can help each other to achieve better performance by incorporating information from one task to another. Towards this end, we experiment with two systems, which are based on the systems we discussed in the previous chapters and we fully explain below.

4.1 Models implemented in this thesis

4.1.1 Encoder-Decoder + frozen Visual Classification (VC)

This model is based on the Encoder-Decoder model described in Section 3.4.3, which performs medical image captioning. This captioning model consists of the DenseNet-121 CNN and a flat LSTM decoder. Here, in order to incorporate tagging to the captioning we add the classifier (FFNN layer of DenseNet-121 + FFNN model as described in Section 2.4.3) we already trained on IU X-ray to the captioning model. First, the image is fed to the DenseNet-121 CNN and we extract its image embedding from the last average pooling layer. Then, the image embedding is given as input to both the classifier and the decoder. The classifier assigns tags to the image and the decoder produces one word of the text at each time step. The decoder is the same as in the Encoder-Decoder model of Chapter 3, but is now trained with the sum of the captioning and the classification loss. This means that the decoder is trained with an extra loss value depending on the tags assigned to each image. The DenseNet-121 encoder stays frozen and also, in this approach the weights of the FFNN classifier are not updated during training.

4.1.2 Encoder-Decoder + trainable Visual Classification (VC)

In this approach, we again incorporate tagging in the Encoder-Decoder captioning model. As described above, the FFNN layer of our CheXNet-based model (Section 2.4.3) is added to the captioning model in order to assign tags to the images. The difference in this model is that we perform multitask learning by training both the FFNN classifier and the decoder. The DenseNet-121 CNN is again frozen and we only use it to extract the image representation. The classifier, which is already trained for medical image tagging is now jointly trained with the decoder. The total model loss is the sum of the classification and captioning losses and is used to train all the model's components, except for the CNN.

4.2 Results

In this section we present the results of the models described above, in order to conclude on how the tagging and captioning can assist each other. For the experiments in this chapter we adopt the sentence delimiter approach, since it yields the best results for our image captioning models (Table 3.4). To evaluate the results of the models, we use the same evaluation measures as in the previous chapters (Sections 2.5 and 3.5), while for the captioning models we experiment with different ways of inference.

Table 4.1 shows the results of our models in the captioning task, which will give us an insight on how tagging affects medical image captioning. As we see, the simple Encoder-Decoder using the greedy inference achieves overall the best results in the evaluation measures. However, the sampling method yields the highest diversity, especially for the Encoder-Decoder + frozen VC model. Low diversity of the generated captions is not correct, compared to the diversity of the gold captions, meaning that models just generate frequent words that may not describe the correct diagnosis and the fact that the captioning evaluation measures do not capture this, is a weakness. In our experiments the sampling approach seems to be the only one overcoming this challenge and achieving high diversity. Observing the results for each inference method separately we observe that the greedy inference method in our Encoder-Decoder + VC models does not work, since it leads to only one generated caption with the same results as the BlindRNN. In the sampling inference approach the Encoder-Decoder + VC models achieves not only higher diversity, but also better results in most metrics. Overall, the addition of the tagging task to captioning seems to yield better results and especially better diversity, depending on the inference method. The Encoder-Decoder + trainable VC model that operates by performing multitask learning could be further improved by applying different weights to the classification and captioning losses (losses were equally weighted in this work).

Model	Inference	B1	B2	B3	B4	Met	Rou	Diversity
BlindRNN	Greedy	**37.9**	22.9	15.5	11.2	16.7	30.9	1
E-D	Greedy	37.1	**23.5**	**16.2**	**11.7**	**17.2**	**32.0**	**103**
E-D + frozen VC	Greedy	**37.9**	22.9	15.5	11.2	16.7	30.9	1
E-D + trainable VC	Greedy	**37.9**	22.9	15.5	11.2	16.7	30.9	1
BlindRNN	Sampling	31.7	17.7	10.6	6.6	14.1	24.5	723
E-D	Sampling	31.4	18.2	11.2	7.2	14.9	**26.9**	722
E-D + frozen VC	Sampling	**33.9**	**19.5**	**12.0**	**7.6**	**15.3**	26.5	**731**
E-D + trainable VC	Sampling	**33.9**	19.4	11.9	**7.6**	15.2	25.8	730

Table 4.1: Results of the medical image captioning models using the sentence approach for greedy and sampling inference. The best results per inference approach are shown.

In the other direction, of how captioning affects classification, Table 4.2 shows the classification results of the Encoder-Decoder + trainable VC model along with the results of our other tagging models (Section 2.5). The DenseNet-121 + FFNN (CheXNet-based) system achieves the best performance. The Encoder-Decoder + trainable VC model consists of the CheXNet-based system and a captioning component, but it does not lead to any improvement over the best tagging method of Section 2.4.3 (DenseNet-121 + FFNN). That leads us to the conclusion that captioning does not help tagging. Another approach that we plan to experiment with is to use the captions generated from a captioning model to extract additional tags from the captions, by applying a medical labeler, e.g., the MTI labeler.

System	**F1**
DenseNet-121 + FFNN	**0.487**
E-D + trainable VC	0.482
DenseNet-121 + k-NN	0.421
DenseNet-121 + 1-NN	0.244
VGG-19 + FFNN	0.091

Table 4.2: Results of our tagging systems on IU X-ray using the MTI tags, following the image-based evaluation approach (see Discussion paragraph of Section 2.2).

5 Conclusions and future work

5.1 Conclusions

In this thesis we addressed two tasks: medical image tagging and captioning, which aim to assist the diagnostic process followed by medical experts when they examine medical images (e.g., X-rays, MRIs, PET/CT scans etc.). Some of the available annotated medical datasets that can be used for classification were presented. We focused on two datasets (ImageCLEFmed Caption 2019 and IU X-ray) which are suitable for the tagging task and performed this task using the four systems we implemented. The systems can be distinguished into retrieval-based and deep learning-based. The best performing system is based on deep learning and consists of a DenseNet-121 followed by an FFNN as described in the work of Rajpurkar et al. (2017b). However, a retrieval-based method achieved competitive results in one of the datasets as well.

The second task we addressed was the medical image captioning. For this task we examined the available datasets and their challenges. We presented two baselines: a simple image retrieval method that returns the caption of the most similar image and a language model that does not look at the image and, hence, always generates the same caption. Then, we implemented an Encoder-Decoder model, which consists of the DenseNet-121 CNN that encodes the image and a flat LSTM decoder that generates the text. We also experimented with various inference methods in order to achieve the highest diversity of generated captions. In the last chapter we examined how the two tasks, image tagging and caption generation, can affect each other by adding a classifier to the Encoder-Decoder model. The tagging results did not improve when trained along with the captioning model. The captioning results on the other hand, showed some improvement when the tagging task was added to captioning.

5.2 Future Work

In future work, we plan to experiment more with models that perform classification for a large number of tags like the ones extracted from medical reports and create a good classification component that can be used to improve captioning. We also plan to experiment with more complex models for medical image captioning. One important component to be added to the Encoder-Decoder model is visual attention, which can help the decoder focus on a specific part of the image when generating each word. Another improvement is to change the decoder to a hierarchical RNN, which performs better for generating paragraphs with many sentences, like the medical reports. A different approach we intend to examine is image retrieval and ways we use it to achieve better results for both tagging and captioning, based on the hypothesis that similar images will have similar diagnoses.

References

A. B. Abacha, A. G. S. de Herrera, S. Gayen, D. Demner-Fushman, and S. Antani. NLM at ImageCLEF 2017 Caption Task. In *CLEF CEUR Workshop*, Dublin, Ireland, 2017.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision*, pages 382–398, Amsterdam, Netherlands, 2016.

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and Top-down Attention for Image Captioning and VQA. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, Salt Lake City, UT, USA, 2018.

A. R. Aronson and F.-M. Lang. An overview of MetaMap: Historical Perspective and Recent Advances. *Journal of the American Medical Informatics Association*, 17(3): 229–236, 2010.

S. Banerjee and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA, 2005.

D. W. Bates, M. Cohen, L. L. Leape, J. M. Overhage, M. M. Shabot, and T. Sheridan. Reducing the Frequency of Errors in Medicine using Information Technology. *Journal of the American Medical Informatics Association*, 8(4):299–308, 2001.

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014.

A. G. S. 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, September 10-14 2018. CEUR-WS.org <http://ceur-ws.org>.

D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami Beach, FL, USA, 2009.

J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2625–2634, Boston, MA, USA, 2015.

C. Eickhoff, I. Schwall, A. G. S. de Herrera, and H. Müller. Overview of ImageCLEFcaption 2017 - the Image Caption Prediction and Concept Extraction Tasks to Understand

Biomedical Images. In *CLEF2017 Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland, September 11-14 2017. CEUR-WS.org <http://ceur-ws.org>.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639):115–118, 2017.

W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer. Producing Radiologist-Quality Reports for Interpretable Artificial Intelligence. *arXiv:1806.00340*, 2018.

Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep Convolutional Ranking for Multilabel Image Annotation. In *International Conference on Learning Representations*, 2014.

Z. Han, B. Wei, S. Leung, J. Chung, and S. Li. Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193, Granada, Spain, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, Las Vegas, NV, USA, 2016.

G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4700–4708, Honolulu, HI, USA, 2017.

B. Ionescu, H. Müller, R. Péteri, Y. D. Cid, V. Liauchuk, V. Kovalev, D. Klimuk, A. Tarasau, A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Demner-Fushman, D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, M. Lux, C. Gurrin, O. Pelka, C. M. Friedrich, A. G. S. de Herrera, N. Garcia, E. Kavallieratou, C. R. del Blanco, C. C. Rodríguez, N. Vasillopoulos, K. Karampidis, J. Chamberlain, A. Clark, and A. Campello. ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, September 9-12 2019. LNCS Lecture Notes in Computer Science, Springer.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031*, 2019.

B. Jing, P. Xie, and E. Xing. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2577–2586, Melbourne, Australia, 2018.

A. E. Johnson, T. J. Pollard, S. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv:1901.07042*, 2019.

A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 3128–3137, Boston, MA, USA, 2015.

M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-Evaluating Automatic Metrics for Image Captioning. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 199–209, Valencia, Spain, 2016.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014.

P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul. From Medical Image to Automatic Medical Report Generation. *IBM Journal of Research and Development*, 59(2):1–7, 2015a.

P. Kisilev, E. Walach, S. Y. Hashoul, E. Barkan, B. Ophir, and S. Alpert. Semantic Description of Medical Image Findings: Structured Learning Approach. In *British Machine Vision Conference*, pages 1–11, Swansea, UK, 2015b.

P. Kisilev, E. Sason, E. Barkan, and S. Y. Hashoul. Medical Image Description Using Multi-task-loss CNN. In *Deep Learning and Data Labeling for Medical Applications*, pages 121–129, Athens, Greece, 2016.

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*, pages 26–36, Minneapolis, MN, USA, 2019a.

V. Kougia, J. Pavlopoulos, and I. Androutsopoulos. AUEB NLP Group at ImageCLEFmed Caption 2019. In *CLEF2019 Working Notes. CEUR Workshop Proceedings*, pages 9–12, Lugano, Switzerland, 2019b.

J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325, Honolulu, HI, USA, 2017.

R. Leaman, R. Khare, and Z. Lu. Challenges in Clinical Natural Language Processing for Automated Disorder Normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015.

J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim. Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

Y. Li, X. Liang, Z. Hu, and E. P. Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *Neural Information Processing Systems (NIPS)*, pages 1530–1540, Montreal, Canada, 2018.

S. Liang, X. Li, Y. Zhu, X. Li, and S. Jiang. ISIA at the ImageCLEF 2017 Image Caption Task. In *CLEF CEUR Workshop*, Dublin, Ireland, 2017.

C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out ACL Workshop*, pages 74–81, Barcelona, Spain, 2004.

M. Lin, Q. Chen, and S. Yan. Network in Network. *arXiv:1312.4400*, 2013.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland, 2014.

G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. V. D. Laak, B. V. Ginneken, and C. I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi. Clinically Accurate Chest X-Ray Report Generation. *arXiv:1904.02633*, 2019.

X. Liu, H. Li, J. Shao, D. Chen, and X. Wang. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *European Conference on Computer Vision*, pages 353–369, Munich, Germany, 2018.

D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the Limits of Weakly Supervised Pretraining. In *European Conference on Computer Vision*, pages 181–196, Munich, Germany, 2018.

H. Müller, J. Kalpathy-Cramer, D. Demner-Fushman, and S. Antani. Creating a Classification of Image Types in the Medical Literature for Visual Categorization. In *Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, San Diego, CA, USA, 2012.

L. Oakden-Rayner. Exploring large scale public medical image datasets. *arXiv:1907.12720*, 2019.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, 2002.

O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In *MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*, pages 180–189, Granada, Spain, 2018.

O. Pelka, C. M. Friedrich, A. G. S. de Herrera, and H. Müller. Overview of the ImageCLEFmed 2019 Concept Prediction Task. In *CLEF2019 Working Notes*, volume ISSN 1613-0073 of *CEUR Workshop Proceedings*, Lugano, Switzerland, September 09-12 2019. CEUR-WS.org <http://ceur-ws.org/Vol-2380/>.

E. Pinho and C. Costa. Feature Learning with Adversarial Networks for Concept Detection in Medical Images: UA.PT Bioinformatics at ImageCLEF 2018. In *CLEF2018 Working Notes. CEUR Workshop Proceedings*, Avignon, France, 2018.

M. M. Rahman. A Cross Modal Deep Learning Based Approach for Caption Prediction and Concept Detection by CS Morgan State. In *CLEF2018 Working Notes. CEUR Workshop Proceedings*, Avignon, France, 2018.

P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, et al. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv:1712.06957*, 2017a.

P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv:1711.05225*, 2017b.

P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, et al. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLOS Medicine*, 15(11):1–17, 2018.

T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs. Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks. In *Information Processing in Medical Imaging*, pages 437–448, Isle of Skye, UK, 2015.

H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to Read Chest X-rays: Recurrent Neural Cascade Model for Automated Image Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, Las Vegas, NV, USA, 2016.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*, 2014.

L. Soldaini and N. Goharian. QuickUMLS: A Fast, Unsupervised Approach for Medical Concept Extraction. In *MedIR workshop*, 2016.

Y. Su, F. Liu, and M. P. Rosen. UMass at ImageCLEF Caption Prediction 2018 Task. In *CLEF2018 Working Notes. CEUR Workshop Proceedings*, Avignon, France, 2018.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, MA, USA, 2015.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *ICML*, pages 104–114, Banff, Alberta, Canada,, 2004.

L. Valavanis and S. Stathopoulos. IPL at ImageCLEF 2017 Concept Detection Task. In *CLEF CEUR Workshop*, Dublin, Ireland, 2017.

S. Varges, H. Bieler, M. Stede, L. C. Faulstich, K. Irsig, and M. Atalla. SemScribe: Natural Language Generation for Medical Reports. In *LREC*, pages 2674–2681, Istanbul, Turkey, 2012.

R. Vedantam, Z. C. L. Zitnick, and D. Parikh. CIDEr: Consensus-Based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, Boston, MA, USA, 2015.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, Boston, MA, USA, 2015.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2016.

C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image Annotation Refinement Using Random Walk with Restarts. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 647–650, Santa Barbara, CA, USA, 2006.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 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*, pages 2097–2106, Honolulu, HI, USA, 2017.

X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058, Quebec City, Canada, 2018.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, pages 2048–2057, Lille, France, 2015.

Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image Captioning with Semantic Attention. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 4651–4659, Las Vegas, NV, USA, 2016.

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, 2018.

Z. Zhang, P. Chen, M. Sapkota, and L. Yang. TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 320–328, Quebec City, Canada, 2017a.

Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436, Honolulu, HI, USA, 2017b.