FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Lukáš Chaloupský

Automatic generation of medical reports from chest X-rays

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: Mgr. Rudolf Rosa, Ph.D.

Study programme: Computer Science

Study branch: Software and Data Engineering

Prague 2022

First of all, I would like to thank my supervisor Mgr. Rudolf Rosa, Ph.D. for all his time, guidance and valuable advices he gave me while working on this thesis. I would also like to thank my parents for their unlimited support and patience during my studies.

Title: Automatic generation of medical reports from chest X-rays

Author: Bc. Lukáš Chaloupský

Institute: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Rudolf Rosa, Ph.D., Institute of Formal and Applied Linguistics

Abstract: This thesis deals with the problem of automatic generation of medical reports in the Czech langugage based on the input chest X-ray images using deep neural networks. The first part deals with the analysis of problem itself including comparison of existing solutions from several common points of view. In order to interpret medical images in the Czech language we present a fine-tuned a Czech GPT2 model specialized on medical texts based on the original pre-trained English GPT2 model along with its evaluation. In the second part the created Czech GPT2 is used for training neural network model for generating medical reports. The training was conducted on freely available data along with data pre-processing and their adjustment for the Czech language. Furthermore the model results are discussed and evaluated using standard metrics for natural language processing to determine the performance.

Keywords: natural language processing, image captioning, x-ray, medical report generation

Contents

Introduction

In hospital, inspecting the X-rays and writing a corresponding medical reports is a hard work that requires experienced specialized doctors, of which there are not many. A great deal of people visit hospitals daily and X-rays are taken for many of them. Automatic interpretation of X-ray image has a great potential to improve health care and it could be particulary helpful to doctors in order to distinguish serious cases from the ordinary ones and overall accelerate and improve their work.

Automatic generation of radiology reports is a subset of a general problem called Image Capitioning, i.e. generation of overall textual captions to input images. Image captioning is a combination of Natural Language Processing and Computer Vision areas, experiencing a lot of progress in the last years. Most often the Image Captioning problem is solved using Deep Learning techniques. The specifity of this subset is that we do not want to generate just a general caption of the image, but the exact description of all findings contained in the given medical image. There were done multiple studies for this task in other languages but none in the Czech language.

Deep learning by its very nature has wide range of uses in a medical sector as it can capture complex relations in any kind of data with excellent performance results. Nevertheless in the medical environment the accuracy of predictions is crucial in order to determine the final diagnosis. Therefore, we should not consider the models as such as something that is unmistakably true, but as an auxiliary tool that should help doctors to examine X-rays.

Inasmuch as it is not so challenging to detect fractures on the limbs, this area is less interesting than others which have a variety of diverse possible problems. One of these areas is chest for which there exists multiple freely accessible datasets containing full textual medical reports. However, all these available datasets have one common downside, they are not in the Czech language. The natural question arises, where do we obtain these much needed data? We have to face and solve this core problem in our thesis.

Goals

First of all, we will take a closer look at the problem itself. This includes breaking down the problem and analyzing all its parts individually together with presenting possible existing alternatives for each part.

Our first goal is to fine-tune a language model directly for the Czech language. The language model will be specialized directly to medical texts in order to capture the essence of the problem. However, ahead of the medical specialization, we want to fine-tune a general czech language model. Fine-tuning will be based on the original English GPT2 model presented in Radford et al. [2019].

Finally, we want to utilize our fine-tuned language model for training neural network model interpreting chest X-rays images and generating corresponding medical textual reports to them in the Czech language. This section also involves the overall data preparation directly for the Czech language. In addition, the training will be done in multiple setups. All possibilities will be evaluated with the purpose of determination of their final performance.

Thesis structure

In the very first chapter we present a detailed description of our problem. Every aspect of our problem is introduced and all existing solutions or possibilities are discussed with their pros and cons. Moreover we introduce there some of the important related works.

Following chapter is dealing with the design of solution to our problem, with all reasosings and decisions made. This includes not only the final neural network model, but also the language model fine-tuning and data preparation.

All experiments done with our models take their part in the third chapter. Describing all used scripts and differents setups together with data variatns.

Whole fourth chapter is then dedicated to evaluation of experiments done in the preceding part. Furthermore our models will be compared to the performance of other existing solutions.

Finally, in epilog we discuss what we have accomplished in the thesis, what the resulting consenquences are and what the future possibilities are.

1. Problem Analysis

This chapter deals with the overall analysis of the problem itself. In the very beginning we present the definition of the problem. Every aspect of the problem is further discussed in detail along with a comparison of possible solutions. Moreover, the next section of the chapter describes data we work with and their alternatives. The final part of this chapter presents some of the important related works.

1.1   Methods of generation

First of all, we will present posssible approaches for image caption generation. Inasmuch as almost all solutions for image captioning based on neural networks are using an encoder-decoder architecture as their core with further minor or major adjustments we will discuss just this architecture. Encoder serves for the purposes of extracting visual features from the input images into intermmediate vector representations for the decoder. The decoder is subsequently fed with these visual feature vectors and decodes them token by token into the natural language text.

Moreover, over the past years the encoder-decoder architecute has been enhanced using the attention mechanism. The problem of encoder-decoder architecture without any other mechanism is that all the information about the image have to be encoded globally inside a single vector. However not all the information about the image is relevant for the caption generation. For this reason the attention mechanism was introduced. Instead of a single vector, the set of feature vectors representing the spatial information of the input image is used. This allows us to dynamically focus only on specific areas during the generation. The overall nature of the mechanism is derived from the way our brains concentrace on the images.

Attention mechanism for image captioning was first introduced in the Xu et al. [2015] work using the Bahdanau attention proposed in Bahdanau et al. [2014]. After extractions of visual feature vectors h_j from image, we want to assign a weight α_{ij} to each of this vectors indicating the relevance of the image position j when generating output at position i. This results in the context vector c_i, a dynamic weigthed combination of h_j features, that is presented as another input to the decoder. All attention calculations are defined as follows:

$$c_i = \sum_{j=1}^{t} \alpha_{ij} h_j \tag{1.1}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{t} \exp(e_{ik})} \tag{1.2}$$

$$e_{ij} = a(s_{i-1}, h_j) \tag{1.3}$$

where s_{i-1} is a previous hidden state and $a(s_{i-1}, h_j)$ is an attention model.

1.1.1 Encoder

Encoder is responsible for extracting high-level visual features of the input image into one or more visual feature vectors. Moreover, the encoder can include also additional parts for another independent purposes, e.g. detection of objects in the image and their classification, which may be further passed to the decoder as a separate or an extra input. We will describe some of the most used neural network architectures utilized as an encoder.

1.1.1.1 Convolutional neural networks

Convolutional neural networks (CNN)[1] are the most leveraged type of neural networks for the purposes of image analysis. They are applying 2D convolution filters for all positions in the image with the shared weights allowing to detect similar patterns independently on the position. Each layer of filters will downsize the dimensions of the previous layer and increases the number of filters. By gradually applying filters, the network is learning more complex patterns.

Instead of training convolutional neural networks from scratch, already pre-trained CNN models are often used and possiblly further fine-tuned to a specific area. These models have been trained on a vast amount of data thanks to which they learned to recognize a relevant features in the lower layers applicable to any kind of images. In contrast, training from scratch is costly and takes a much longer time. Among the used pre-trained CNN model architectures are for example ResNet[He et al., 2016], VGGnet[Simonyan and Zisserman, 2014], DenseNet[Huang et al., 2017] or EfficientNet[Tan and Le, 2019].

1.1.1.2 Transformers

With the Vaswani et al. [2017], a new neural network archtecture called tranformers was introduced. The transformers are using multi-head self-attention mechanism to compute the relations between the elements in the sequences. Although the transformers were originally designated for NLP tasks, they can be also utilized for other domains like images due to their robustness. Vision transformers (ViT), presented in Dosovitskiy et al. [2020], is an adaptation of transformers encoder for image classification without any CNNs. The core of the encoder is identical to the one in the original transformer. Nevertheless, the input image is represented as a sequence of patches for which their embeddings are computed together with the positional encodings as the input, as we can see in the Figure 1.1.

1.1.2 Decoder

The second main part of the network is the decoder which serves as a language model for generating corresponding captions to input images. During the generation process the visual features of the image from the encoder and the already generated text are taken into account for the predtiction of the next token or word. In the process of generation, the attention mechanism (described above in

---

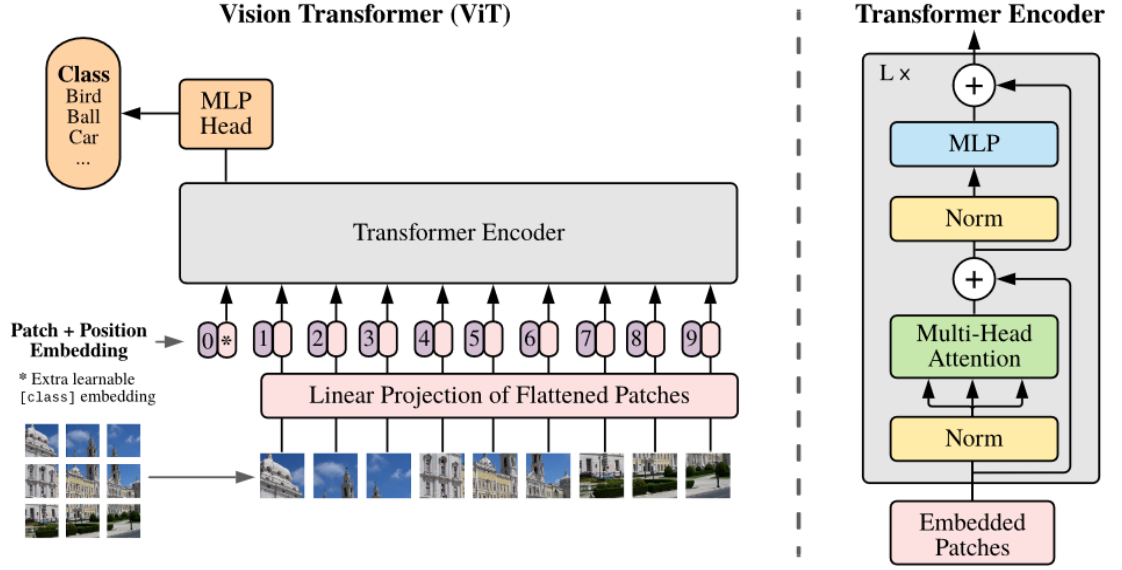[1]https://en.wikipedia.org/wiki/Convolutional_neural_network

Figure 1.1: Visual Transformer architecture from Dosovitskiy et al. [2020].

the Chapter 1.1) is almost always incorporated in order to focus only on the substantial particular areas. The following subsections presents common approaches used as the backbones for the decoder part of the network.

1.1.2.1 Recurrent neural networks

Recurrent neural networks (RNN)[2] are a category of neural networks designated for sequence data processing. Context of the previous part of the sequence can influence the subsequent elements due to the RNN cell's internal memory (hidden state). Each time step RNN cell computes its activation function from current input and hidden state producing updated hidden state as output. For language modeling tasks, the last generated token or word is given to the RNN as the next input until the entire text is produced. The autoregressiveness of the RNNs is the reason why they are ideal for use as a language model. The two most used types of RNN cells are LSTM[Hochreiter and Schmidhuber, 1997] and GRU[Cho et al., 2014], however the LSTM cells are more utilized as the decoder because they can remember longer sequences. Moreover, we can combine them in a hierarchical manner to capture more complex structures in the generated text. Nevertheless, the downside of the RNNs is the training time due to their sequential nature. The whole sentence must pass through RNN token by token and cannot be parallelized.

1.1.2.2 GPT2

Just as in the case of encoder, with the advent of transformers presented in the Vaswani et al. [2017] paper, the deoder part of transformers started to be used as the language model for image captioning task for their great results in the natural language processing (NLP)[3] tasks. One of the advantages of the

---

[2]https://en.wikipedia.org/wiki/Recurrent_neural_network
[3]https://en.wikipedia.org/wiki/Natural_language_processing

transformers against the previously used recurrent neural networks is the loss of the need for sequential processing during the training. Due to the fact that the processing of the whole caption can be done in parallel, the entire training process is significatnly accelerated. Moreover, the transformers can capture longer ranges dependencies in the text, as they process the sentence as a whole instead of sequentially by words.

One of the state-of-the-art autoregressive language models using transformers as their backbone is OpenAI GPT-2 model from the Radford et al. [2019] paper, which outperforms other language models on many NLP tasks. It was trained on a massive English dataset called WebText (introduced in the same article) containing a total of 40 GB of raw text. The resuling model is able to generate large coherent texts. Furthermore, it can be fine-tuned to a different domain or to a completely different language.

1.2   Data

In previous part we talked about possible methods of generation. Another crucial aspect we need to discuss are data, which are a basic building block of our thesis. This part focuses on the analysis of the data we used in our thesis, but also on their alternatives.

In order to solve our task and train neural network we need to get dataset containing the X-rays images along with their textual descriptions and optionally some other attributes of the examined X-rays. Moreover, the fundamental feature we need is that the data must be in the Czech language.

1.2.1   Existing datasets

Medical environment provides a plenty of diverse potential problems, which can be researched. As already mentioned, in this thesis we focus specifically on the X-ray images. Because it is not so hard to detect fractures on the limbs, this area is not as interesting as others. One area that is rich in its diversity is the chest. As a result, this area is explored the most and therefore there exists multiple datasets with full textual mecidal reports. In the following section we describe some of them.

Apart from the datasets described below, other datasets with similar type are being used with the aim of solving our task. Amongst them belong datasets such as ImageCLEFmed Caption[Rückert et al., 2022], PadChest[Bustos et al., 2020], BCIDR[Zhang et al., 2017] and PEIR Gross[Jing et al., 2017]. Moreover, except for datasets containing textual reports there exist a lot of other datasets worth mentioning containing different kind of information for each X-ray. These include, for example, CheXpert[Irvin et al., 2019], VinDr-CXR[Nguyen et al., 2020], ChestX-ray8[Wang et al., 2017] and its expanded version ChestX-ray14.

1.2.1.1 Indiana University chest X-ray

Indiana University chest X-Ray dataset has become a standard in the field of medical report generation, it was presented in the Demner-Fushman et al. [2015] paper. This dataset is an open source collection of pairs of chest X-rays and their corresponding semi-stuctured textual radiology reports, which is freely availble on the web[4] without any additional requirements. We have a choice if we want to download just reports or images and in either PNG or DICOM format. The entire dataset consists of 7470 chest X-ray images that cover not only the frontal (PA[5]) view, but also the lateral (side) one. These images corresponds to a total of 3995 patient's medical text reports.

Figure 1.1 shows an example from the Indiana University chest X-ray dataset. Each dataset pair is carefully de-identified in order to remove any personal information. The text of the report is semi-structured in up to 5 sections. The most important sections are *impression*, where the overall diagnosis is stated, *findings* section describing the details of examination and *tags* which are of two types - manual and automatic. Manual tags were annotated manually using MeSH[6] and RadLex[7] codes, automatic were encoded from the reports using the MTI indexer. The rest of the sections are *indication* and *comparison*.

The disadvantage of this dataset is that it is relatively small. On the other hand, it is a quite clean and manually checked dataset containing also additional information about images in a form of tags described above.
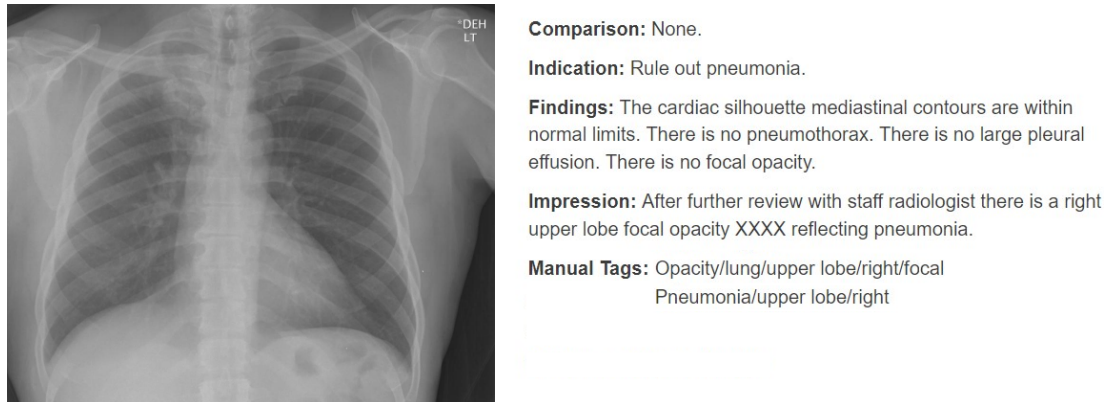

Comparison: None.

Indication: Rule out pneumonia.

Findings: The cardiac silhouette mediastinal contours are within normal limits. There is no pneumothorax. There is no large pleural effusion. There is no focal opacity.

Impression: After further review with staff radiologist there is a right upper lobe focal opacity XXXX reflecting pneumonia.

Manual Tags: Opacity/lung/upper lobe/right/focal
Pneumonia/upper lobe/right

Figure 1.2: Sample from the Indiana University Chest X-ray dataset.

1.2.1.2 MIMIC-CXR v2.0.0

MIMIC-CXR v2.0.0 is another dataset consisting of full semi-structured medical textual reports against corresponding chest X-rays that was presented in the Johnson et al. [2019a] paper. As the previous dataset, it is openly available on the web[8]. In order to get access to the dataset, we have to go through registra-

---

[4]https://openi.nlm.nih.gov/faq#collection
[5]Posterior-Anterior
[6]https://www.nlm.nih.gov/mesh/meshhome.html
[7]http://radlex.org/
[8]https://physionet.org/content/mimic-cxr/2.0.0/

tion and verification steps. The verification phase includes completion of CITI[9] *Data or Specimens Only Research* course for *Human Subject Research*. Moreover we need somebody trustworthy as a reference to confirm the authenticity of our identity. After the verification we get access to all datasets in the same repository.

The dataset consists of 377,110 X-ray images in the DICOM[10] format connected to a total of 227,835 radiology reports for 65,379 patients. Each report is structured into multiple different sections. In order to satisfy legal requirements, entire dataset is automatically de-identified to remove any protected health information[11]. Similalry to the previous dataset the essential two sections of each report are *impression* and *findings*. There also exists older MIMIC-CXR-JPG[12] dataset, presented in the Johnson et al. [2019b] paper. This is an older version of MIMIC-CXR v2.0.0 dataset consisting of the exactly same images, only in JPG format, but each image is assigned 14 labels indicating the presence of the category in the report instead of its textual form. Each category has assigned either a *1*, *0* or *-1* label with the meaning *positively mentioned*, *negatively mentioned* or *uncertain*. The labels were determined from the reports utilizing the CheXpert[Irvin et al., 2019] and the NegBio[Peng et al., 2018] open-source labelers.

The advantage of this dataset is its vast number of samples. Moreover, as described above, we can get additional information in a form of categories to every image. Nevertherless the textual reports carry some noise in them in the form of grammatical mistakes and incorrect formatting. We face these issues in the Chapter 2.3.2.

---

[9]https://about.citiprogram.org/series/human-subjects-research-hsr/
[10]https://www.dicomstandard.org/
[11]https://en.wikipedia.org/wiki/Protected_health_information
[12]https://physionet.org/content/mimic-cxr-jpg/2.0.0/

FINAL REPORT

EXAMINATION: CHEST (PA AND LAT)

INDICATION: History: ___F with dyspnea

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

FINDINGS: Heart size remains mild to moderately enlarged. The aorta is tortuous and diffusely calcified. Mediastinal and hilar contours are otherwise unchanged. Previous pattern of mild pulmonary edema has essentially resolved. Mild atelectasis is seen in the lung bases without focal consolidation. Blunting of the costophrenic angles bilaterally suggests trace bilateral pleural effusions, not substantially changed in the interval. No pneumothorax is present.

IMPRESSION: Interval resolution of previously seen mild pulmonary edema with trace bilateral pleural effusions.

Figure 1.3: Sample from the MIMIC-CXR dataset.

1.2.2 Czech data

All freely available datasets presented in the previous part have one common downside, namely they are not in the Czech language. As a part of elaboration of this thesis an intesive communication with real czech hospitals and other possible sources of real data took place. The goal of this communication was to create the very first open czech dataset of this kind. Processing of this kind of data would mean not only preparing the data into suitable format but also it would include proper anonymization of any personal information about the patients within the data.

However, inasmuch as the authentic patients data from hospitals are subject to strict privacy rules and we are not employees of any hospital, the institutions

decided that they cannot provide the data in any way without the concious permission of patients given before the examination. With this result we need to find a different way how to obtain this much needed czech data.

1.2.3 Translators

In the previous sections we discovered that there is no dataset in the Czech language for our problem and there is no easy way how to get acces to the real data in order to build one. The only thing left is to create a new artificial dataset using an automatic translation. We will compare different freely accesible translators and choose the right one for our needs.

1.2.3.1 DeepL

At the moment, DeepL[13] translator provides the finest available translations beating even the ones from Google Translate. Moreover, it has freely usable web application and REST API. However, the main drawback of the DeepL translator is that its REST API is highly limited - only 500 000 characters per month can be translated for free. Furthermore, any translation above this limit is costly and thus this path is not appropriate for translating large textual datasets. One way to get around this problem is to use their internal REST API used specifically for the web application, which is free to use. We investigated and implemented this potential way in our thesis and further experimented how much it can be used, but unfortunately even this internal REST API is strictly limited for only tens of consecutive[14] translations making it unusable for out needs.

1.2.3.2 Google Translate

Google Translate[15] has become already de facto standard in the world of machine translation and it is the most used freely accessible language translation service in the world. In terms of quality, the translations are still great although little bit worse than those from DeepL. The web application is free of any charge and anybody can use it as much as he needs. Nevertheless, just as in the case of DeepL, their REST API services are limited and translation of anything above that limit is expensively charged. For these reasons, as in the previous case, we must find another way.

1.2.3.3 CUBBITT

Machine Translation[16] is an extensive area of research, as a result of which there exist many other projects and academic papers nowadays. One of them is CUBBITT[17] translator, which was developed at our faculty. The whole system is presented and described in detail in the Popel et al. [2020] paper.

---

[13]https://www.deepl.com/translator

[14]REST API calls are delayed from each other for some time, otherwise the service is blocked immediately

[15]https://translate.google.com/

[16]https://en.wikipedia.org/wiki/Machine_translation

[17]https://lindat.mff.cuni.cz/services/translation/

CUBBITT translator provides translations which are comparable to the ones from DeepL and Google Translate services. As other mentioned translators it provides an openly available web application for machine translation. Moreover and most importantly it provides REST API that is completely unlimited in text volume and free to use without any additional charges. These are the reasons why we will utilize CUBBITT in our thesis as a translator to create our artificial dataset.

On the other hand, CUBBITT has not support for auto-correcting input text compared to above mentioned services. Moreover, there are some patterns in the text which CUBBITT cannot translate at all or translates them incorrectly. These problems complicates our situation as the data from hospitals carry some natural noise in them. We face these complications in Chapter 2.3.2.

1.3   Related work

The last section of this chapter is dedicated to the description and comparison to some of the related works that solves identical or similar problem as we do.

The most significant related work is Alfarghaly et al. [2021] inasmuch as we base our thesis on it. This paper focuses on the identical problem as we do, only in the English language. The proposed solution uses a pre-trained and further fine-tuned Chexnet model, presented in Rajpurkar et al. [2017], as a visual features encoder and distilGPT2 language model[Radford et al., 2019] as a decoder which is additionally conditioned on the visual features and predicted tag's word2vec[Mikolov et al., 2013] embeddings. For training the neural network the Indiana University chest X-ray dataset, described in more detail in Chapter 1.2.1.1, is used. Figure 1.3 shows examples of the model outputs.

Another paper making use of transformers is Chen et al. [2020]. The visual features of images are extracted using pre-trained convolutional neural network and they are further passed to the transformer encoder outputting hidden states, that are further presented to the transformer decoder for the report generation. However, the decoder architecture contains special memory module and also enhances the layer normalization. The memory module serves for memorization of text patterns which occur in the similar images inasmuch as they can further help for generating the report. We can see its effect in the Figure 1.4. Indiana University chest X-ray and MIMIC-CXR v2.0.0 datasets (see Chapter 1.2.1 for more details) were used as training data.

Yuan et al. [2019] paper proposes a hiearchical encoder-decoder architecture as we can see in the Figure 1.4 for the purpose of generating textual reports. Pairs of frontal and lateral X-ray images are used as an input to the network instead of single images, as the authors claim that the images should be complementary to each other instead of being processed independently. The RestNet-152 model pretrained on the CheXpert[Irvin et al., 2019] dataset is utilized as the encoder with three outputs used later - the global and local features of the images, predticted observations and medical concepts. The decoder is hierarchical LSTM decoder
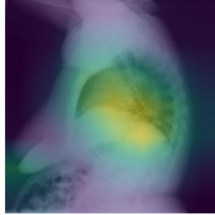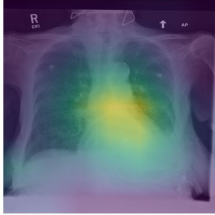
	Normal	Accurate	Missing Details	False
Ground Truth	Negative. The cardiomediastinal silhouette is normal in size and contour. No focal consolidation, pneumothorax or large pleural effusion. normal.	Stable cardiomegaly with mild pulmonary interstitial edema. Unchanged cardiomegaly negative for pneumothorax or focal consolidation. No large effusion. Mildly prominent interstitial opacities.	Minimal left basilar atelectasis versus infiltrate. Low lung volumes. Normal cardiomediastinal contours. Low lung volumes with minimal left basilar opacities. No pneumothorax or pleural effusions.	Prominent hiatal hernia. left basilar opacity compatible pleural effusion and atelectasis. Right pleural effusion. no pulmonary edema / overt chf identified. stable senescent mediastinal contour.
Prediction	No evidence of active Disease. The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	Cardiomegaly and mild interstitial pulmonary edema. Moderate cardiomegaly bibasilar and perihilar interstitial opacities. No pneumothorax	No acute cardiopulmonary abnormality. The lungs are clear, and without focal air space opacity. The cardiomediastinal silhouette is normal in size and contour. There is no mediastinal contour, large pleural effusion, pneumothorax, or focal airspace consolidation.	Interval left subclavian central venous catheter with tip approximating the high svc. No evidence of pneumothorax. Generalized heart size and mediastinal contours appear within normal limits. Atherosclerotic changes of the aorta. Moderate degenerative changes of the thoracic spinea.

Figure 1.4: Examples of generated medical reports from Alfarghaly et al. [2021].

comprising of the two parts: *sentence decoder* and *word decoder*. The sentence decoder takes visual features and generates hidden state for each sentence, which are along with the predicted medical concepts presented to the word decoder in order to generate the report. There are many other papers using a hierarchical architecture, e.g. Huang et al. [2019].

The last related project we mention in this work is CareBot[18], which is being developed concurrently with our thesis. CareBot is a czech startup founded in 2021 as a reaction to the then ongoing Covid-19 pandemic. As in our work, CareBot focuses its attention mainly on the processing of chest X-rays using neural networks. However, unlike us it does not generate the textual reports for the doctors. Instead of textual reports, it focuses on finding and classifying a total of 15 different types of individual diseases on X-ray images and their subsequent spatial localization. Moreover, they already have a support from the medical environment.
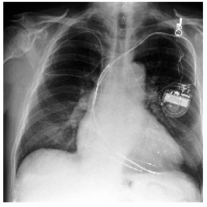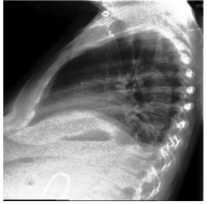
---

[18]https://www.carebot.com/

Figure 1.5: Examples of generated medical reports from Chen et al. [2020].
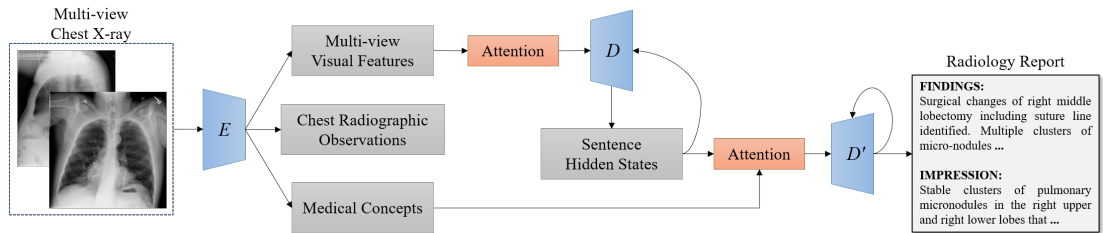

Figure 1.6: Hierarchical architecture from Yuan et al. [2019].
E is encoder, D is sentence decoder and D' is word decoder.

2. Design

This chapter summarizes the overall design of our approach to generating textual medical reports for X-ray images. The problem consists of multiple independent parts we need to deal with. For each of them, we we will present the fundamentals of our solution, along with description of related problematics and decisions made.

2.1 Our approach

As we already mentioned in the Chapter 1.3, the overall solution for the final medical report generation model is based on the Alfarghaly et al. [2021]. We have chosen this approach for multiple reasons. The main reasons to use this work as the backbone for our thesis are following:

1. In the work the state-of-the-art GPT-2 model is utilized as the language model. This gave us a great opportunity as there was none Czech GPT-2 model available at time this thesis began.

2. The encoder is already fine-tuned to extract visual features for specific dataset.

3. All solution source code is freely accessible on the github[1].

As in most works for image captioning, the architecture is encoder-decoder based with an attention mechanism. The high-level solution architecture is depicted in Figure 2.1.
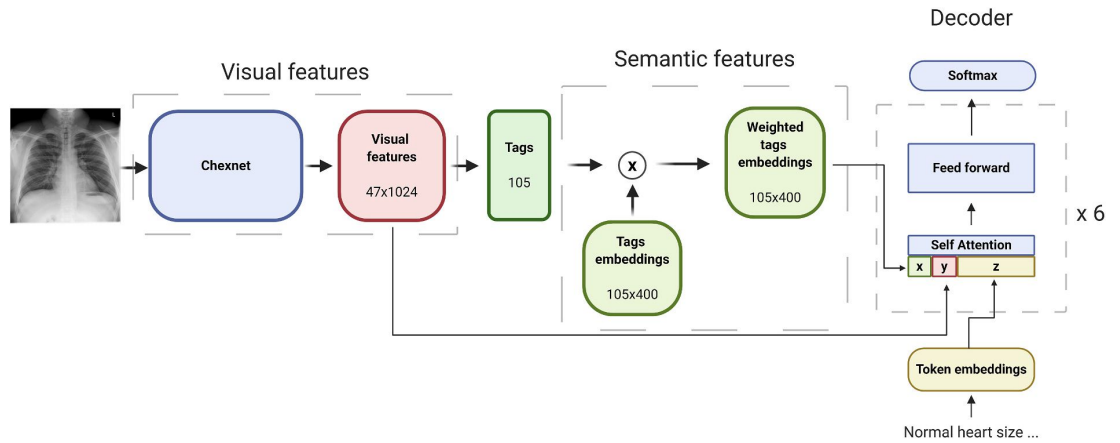

Figure 2.1: Overall architecture used in our solution proposed in Alfarghaly et al. [2021].

---

[1] https://github.com/omar-mohamed/GPT2-Chest-X-Ray-Report-Generation

2.2   Czech GPT-2

The aim of this work is to generate medical reports in the Czech language. In the previous parts, we decided to use the GPT-2 as the language model. However at the time of the beginning of this work, no Czech GPT-2 model was freely available and thus it was essentially necessary to create one. This section describes all the steps needed for fine-tuning the small English GPT-2 to the Czech language. Respectively, we train two versions of the Czech GPT-2 model. One trained on the general Czech textual data and one specialized specifically on Czech medical texts.

2.2.1   Data

In this section we will describe the possible data applicable for training of both the general and medical Czech GPT-2 model together with the decision made about the final data selection and data cleaning.

2.2.1.1   General

For the training of general Czech GPT-2 we have a plenty of data options we can choose from. However, there are important properties of the data we need to sastisfy. As we are transfer learning from English to Czech language we need to have sufficiently large data, so we ensure the GPT-2 will learn properly syntactical and semantical information. Moreover, the data have to be also heterogeneous enough, so the model can capture different types of information and not just for example newspaper articles from a specific area. In order to create a good enough general model we need to meet these criteria.

Several different datasets were investigated and tested for the training of the general Czech GPT-2 model.

Czech Wikipedia
The first data we used for training the model is the Czech Wikipedia dump[2]. After extraction, the total size of the dataset is approximately 800 MB of raw text. The advantages of this dataset are its easy accessibility and fairly clean data quality. On the other hand, the data are very homogeneous despite the various topics. Each article is written in the general descriptive style. Moreover the data themself are not large enough, the trained model made many both syntactical and semantical mistakes during the text generation.

Balanced Czech National Corpus
Another possibility was to use a balanced version of the Czech National Corpus[Křen et al., 2021] as the original is composed mainly of journalistic articles. The balanced version tries to equalize the amount of data from each category. These categories include *journalism*, *poetry*, *prose*, *educational literature* etc. The major advantage of this dataset is its purity, the texts are syntacticaly correct

---

[2]`https://dumps.wikimedia.org/cswiki/latest/cswiki-latest-pages-articles.xml.bz2`

without any undesirable non-Czech elements and written in the standard Czech language. The dataset does not have any significant downsides and the trained model understood Czech language without any significant ailments. In total, the dataset is comprised of 3,3 GB of raw text.

OSCAR

OSCAR, from the Ortiz Su'arez et al. [2020], is a huge deduplicated multilingual corpus created from the Common Crawl corpus[3] providing data for 166 different languages and available directly in the huggingface datasets library[4]. It consists of the text scraped from websites of very different kinds and thus the data are heteregeneous enough. Moreover, its huge size, as the czech part of the dataset occupies a total of 24 GB of raw text, is another major benefit. On the other hand, because the data are automatically scraped, they carry a noise in them. Besides that, not negligible part of the text are in the non-standard Czech language as the data come from diverse web sources such as forums etc. Nevertheless, the disadvantages are outweighed by the huge size of the corpus and together with following filtering of the text:

1. We take only text that are at least 1200 characters longs as these texts tend to be longer articles written in the standard Czech language instead of advertisements, incomple texts etc.

2. Any text containing control character are filtered out, because the text contains generally undesirable content.

Conclusion

We analyzed various datasets along with their overall properties. Furthermore, the advantages and disadvantages of each were discussed. As a result, we have chosen the **OSCAR** dataset due to its size and heterogeneity. The **Wikipedia** is too small and homogeneous. On the other hand the **Balanced Czech National Corpus** is heterogeneous enough, however it is an almost order of magnitude smaller than **OSCAR**.

2.2.1.2 Medical

Since we have a trained general Czech GPT-2 model from the previous section that already understands a Czech language, the final fine-tuning for medical environment does not require that much data. We need to specialize the model to understand the mecial environment inherently. For this purpose, we use a subset of the UFAL Medical Corpus v. 1.0[5]. These data are further filtered to remove any inappropriate characters, lines and redundant structures. As a result, the data contain a total of 100 MB of raw medical texts. The texts comprise of general medical descriptions, articles and package leaflets for medicines.

---

[3]https://commoncrawl.org/
[4]https://huggingface.co/datasets/oscar
[5]https://ufal.mff.cuni.cz/ufal_medical_corpus

2.2.2 Training

TODO - Popsat postup tréninku podle příspěvku(popis - obecně, trénování tokenizeru, postupné trénování) + learning rate finder(popis, obrázek) + diff lear. rates(popis - 4 skupiny) + ta trénovací křivka(popis - rozdíl oproti klasickýmu, obrázek), popsat ještě rozdíl mezi originállním přístupem a novým(gradual vs full)

2.3 Medical dataset translation

For machine learning in general and NLP tasks especially, the data quality is the alpha and omega of the performance of the final model. Since, as we have already mentioned in the Chapter 1.2.2, we do not have any Czech data directly for the medical examinations of X-rays, to obtain Czech data we must arrange ourselves in a different way. One potential way is to create a new artificial dataset using machine translation of existing datasets. This section discusses the required steps to build a quality dataset using translation.

2.3.1 Translator choice

In the previous part of the text, specifically in the Chapter 1.2.3, we already discussed all possibilities for automatic translation. Our final choice for the translator is the CUBBITT as it provides REST API unlimited in the number of requests and volume.

2.3.2 Preprocessing

The most important part of our machine translation process is the preprocessing of the input text. We already outlined in the Chapter 1.2.1 that the data contain some noise in them as we are dealing with reports in natural language. Moreover, we will use CUBBITT translator, that doest not perfom autocorrection itself and cannot translate some patterns at all as we described in the Chapter 1.2.3.3. For these purposes we incorporate preprocessing before the translation as it would be beneficial to have all texts in standardized form in order to firstly, help CUBBITT with translation to get the report correctly translated, and secondly to ensure that our model receives and process all the data in a indentical report format.

Our preprocessing pipeline encompasses of following procedures. Some of them are dealing with general CUBBITT issues and other with specifics of the medical data.

Line starts

First of all we start with a very simple procedure. We analyze all lines of the report and delete all white characters common for all lines on both ends. The purpose of this modification is to standardize the report format, while preserving its structure.

Anonymous sequences

Inasmuch as the whole datasets are anonymized due to the legal reasons and privacy protection, the reports contain "anonymous sequences", such as "XXXX" or "____", denoting places with original private information about patients. However, these sequences can be attached to surrounding words forming undesirable words. As we already said, the CUBBITT does not auto-correct its input automatically, so these inputs will not be handled in any special way and therefore not translated. For this reason, we separate these sequences to form independent words.

Units

Subsequent form of correction that we perform is the separation of numbers and units attached to them. In addition, this also includes general cases, where the number and subsequent word are glued together, while keeping specific medical terms with a similar structure. This is associated with the following step, as the units will not be true-cased properly without this procedure.

True-casing

The most imporant part of the whole preprocessing pipeline is true-casing of the input text. This adjustment is necessary for two essential reasons. CUBBITT has problems with translation of any uppercase texts in general. Capturing the true-case of a text is a complex problem requiring either a large statistical language dictionary or a trained model in order to properly determine the case.

As medical reports are very specific area, the existing solutions for general text true-casing is inapplicable. Medical reports contain a lot of abbreviations and acronyms, which can be often confused with ordinary english words. Training the model for medical true-casing requires even more specific data for a certain domain, because in different contexts the common words can be treated differently. Moreover, obtaining flawless data to cover the entire specific domain is a challenging task. For these reasons, we chose the way of a statistical dictionary. Before the translation of the dataset begins, we create a statistical dictionary from the whole dataset containing the most often used form of every word. We also count with some exceptions, such as headings, that in some datasets can be in uppercase only.

Using the created dictionary, we deal with all uppercase words to assign them the proper form. The results of the true-casing preprocessing are demonstrated in the following examples:

- (1a) "EXAMINATION: CHEST (PORTABLE AP)" →
  "PŘEZKOUŠENÍ: CHEST (PORTABLE AP)"

- (1b) "Examination: Chest (portable AP)" →
  "Vyšetření: Hrudník (přenosný AP)"

- (2a) "SMALL RIGHT PLEURAL ABNORMALITY" →
  "MALÉ PRÁVO PLEURÁLNÍ ABNORMALITY"

- (2b) "Small right pleural abnormality" →
  "Malá pravá pleurální abnormalita"

Paragraphs structure

In some of the medical reports the section headings and corresponding texts do not begin on the same lines. We adjust these situations to a form where each heading and the text belonging to it always starts on the same line, for two reasons. Firstly, we want to normalize the report structure in general and secondly, we want to move the section content as close as possible to the heading, so the translator and even the final model have the context close to each other.

Capitalization

Another part of the preprocessing pipeline is a simple capitalization of each heading and each sentence in the report. This text capitalization process helps CUBBITT not only in the case of medical data, but in general during the translation process of some texts to better understand the boundaries between sentences.

Time

One of the patterns that CUBBITT doest not recognize nor auto-corrects itself in general are times. This applies both, to the specification of hours and minutes, and to the part of the day specification. If any of these parts are incorrectly formatted, the time will we translated incorrectly or not translated at all, and thus we would lose some information or it could damage the fluency of the translated text. For these reasons, we apply preprocessing to normalize all times. We can see the difference in the following examples:

- "Chest radiograph at 1045PM" → "Rentgen hrudníku v 1045PM"

- "Chest radiograph at 10:45 PM" → "Rentgen hrudníku ve 22:45"

White spaces

After all previous procedures, we apply one last very simple final modification, namely, we squash all the whitespace characters inside each line into a single space, while maintaining the format from the very first preprocessing step described above. This step is performed only for normalization purposes.

Lowercasing

The last preprocessing procedure we will mention is lowercasing of the whole text followed by capitalizing the first word of each sentence. This is a separate procedure that was used in the earlier phases of elaboration of this work as CUBBITT has problem with uppercased word as we already mentioned.

3. Experiments

Úvod

3.1 Environment

Popsat prostředí na kterém jsme trénovali - it4i cluster a aic cluster

3.2 Czech GPT-2

Popis toho tréninku - odkaz na předchozí kapitolu. Výsledky z trénování * Normal Czech GPT-2 * Medical Czech GPT-2 Ukázky výsledků z obou Popis scriptů

3.3 Dataset translation

Popsat, jak vypadá ten script, jak funguje a že jsme nakonec překládali jen open-i + Výsledky

3.4 Medical report generation model

Popsat, že jsme použili data z Open-I jako v původním článku, jelikož nám jde hlavně o proof-of-concept + Mimic nemá stejné tagy -¿ muselo by se přetrénovat celý CNN backbone. + Výsledky + Popis toho, jak jsme upravili data do stejného formátu jako v článku, popis těch skriptů je na githubu toho článku

4. Evaluation

4.1 Experiments

4.2 Automatic evaluation

4.2.1 Metrics

4.2.2 Results

4.3 Manual evaluation

4.3.1 Method

4.3.2 Results

4.4 Examples

Conclusion

Bibliography

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23 (2):304–310, 07 2015. ISSN 1067-5027. doi: 10.1093/jamia/ocv080. URL `https://doi.org/10.1093/jamia/ocv080`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, 7:154808–154817, 2019.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019a.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. SYN v9: large corpus of written czech, 2021. URL `http://hdl.handle.net/11234/1-4635`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *arXiv preprint arXiv:2012.15029*, 2020.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.156`.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. Transforming machine translation: a deep learning system reaches news translation quality comparable to human

professionals. *Nature Communications*, 11(4381):1–15, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18073-9. URL https://www.nature.com/articles/s41467-020-18073-9.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Henning Müller, and Christoph M. Friedrich. Overview of ImageCLEFmedical 2022 – caption prediction and concept detection. In *CLEF2022 Working Notes*, CEUR Workshop Proceedings, Bologna, Italy, September 5-8 2022. CEUR-WS.org.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer, 2019.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1   First Attachment