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

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

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.

Medical image interpretation is a subset of a general problem called Image Capitioning, i.e. generation of 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 mecidal 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 Czech language. The language model will be specialized directly to medical texts in order to capture the essence of the problem. 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 reports to them in multiple setups. All possibilities will be evaluated to determine 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 Problem definition

1.2 Methods of generation

1.3 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.

1.3.1 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.

1.3.1.1 Indiana University chest X-ray dataset

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 stuctured textual radiology reports, which is freely availble on the web[1] 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

[1] `https://openi.nlm.nih.gov/faq#collection`

(PA[1]) 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 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[1] and RadLex[1] 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 clean and manually checked dataset containing also additional information about images in a form of tags described above.

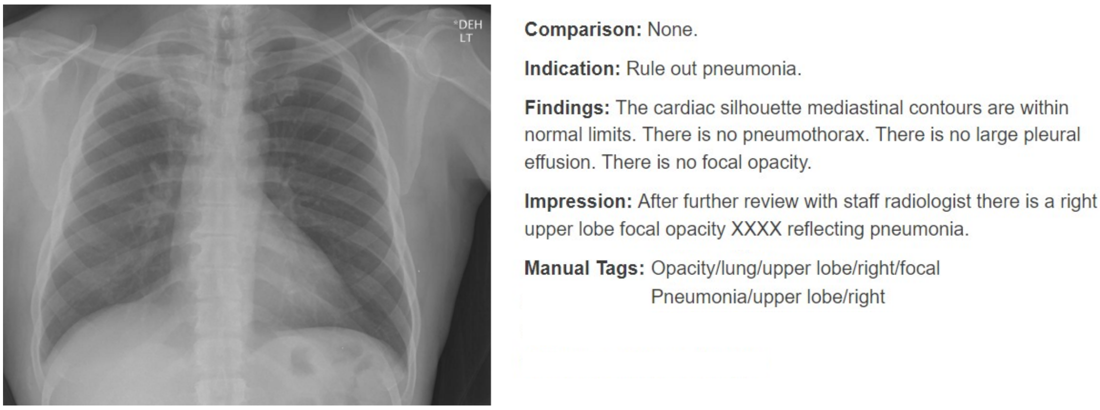

Figure 1.1: Sample from the Indiana University Chest X-ray dataset.

1.3.1.2 MIMIC-CXR

MIMIC-CXR dataset description

[1] Posterior-Anterior
[1] `https://www.nlm.nih.gov/mesh/meshhome.html`
[1] `http://radlex.org/`

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 1.2: Sample from the MIMIC-CXR dataset.

1.3.2 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.

1.3.3 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.

1.3.3.1 DeepL

At the moment, DeepL[1] 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[2] translations making it unusable for out needs.

1.3.3.2 Google Translate

Google Translate[3] 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.

1.3.3.3 CUBBITT

Machine Translation[4] 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[5] translator, which was developed at our faculty. The whole system is presented and described in detail in the Popel et al. [2020] paper.

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

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

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

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

[5] 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 X.

1.4 Language models

1.4.1 GPT2

1.5 Related work

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

2. Title of the second chapter

2.1 Title of the first subchapter of the second chapter

2.2 Title of the second subchapter of the second chapter

Conclusion

Bibliography

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`.

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.

List of Figures

List of Tables

List of Abbreviations

A. Attachments

A.1 First Attachment