MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231208

An Image Retrieval Pipeline in a Medical Data Integration Center

Ka Yung CHENG[a,1], Santiago PAZMINO[a], Björn BERGH[a], Markus LANGE-HEGERMANN[b] and Björn SCHREIWEIS[a]
[a] *Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany*
[b] *inIT—Institute Industrial IT, OWL University of Applied Sciences and Arts, Lemgo, Germany*

Abstract. Medical images need annotations with high-level semantic descriptors, so that domain experts can search for the desired dataset among an enormous volume of visual media within a Medical Data Integration Center. This article introduces a processing pipeline for storing and annotating DICOM and PNG imaging data by applying Elasticsearch, S3 and Deep Learning technologies. The proposed method processes both DICOM and PNG images to generate annotations. These image annotations are indexed in Elasticsearch with the corresponding raw data paths, where they can be retrieved and analyzed.

Keywords. Medical image retrieval, data lake, DICOM, deep learning, elasticsearch

1. Introduction

The implementation of a biomedical imaging data platform requires both storing a vast volume of images and linking to existing images in other specialized department systems. This leads to difficulties in retrieving and searching for specific imaging data, due to the accuracy and computational time of finding a suitable image. In addition, there are different image formats, such as DICOM and all kinds of non-DICOM, lacking searchable image annotations and metadata. Furthermore, some images may be stored repeatedly and cause redundant storage and work. Hence, there is also a pressing demand to annotate medical images with high-level semantic descriptors [1].

2. Methods

We propose a processing pipeline to store and annotate both DICOM and PNG images. As part of our ETL pipeline [2], we first store the raw medical images in the data storage S3. Then, the event streaming platform Kafka notifies the dataflow processing system Apache Nifi. Nifi streams image data and communicates over a Rest API with a keyword generator, to and insert both generated image information with a S3 key (file path) into the search engine Elasticsearch (ES). The keyword generator is designed to annotate DICOM and PNG data based on Flask: 1) For DICOM images, the ingest component extracts and indexes the standard DICOM header. Over 100 DICOM images of 7

[1] Corresponding Author: Ka Yung Cheng, email: KaYung.Cheng@uksh.de.

anonymized test patients in the research project TOMAS [3] are extracted and indexed in ES.; 2) Alternatively, for unannotated PNG images, the component uses a "Deep Learning (DL) keyword generator" to generate appropriate instance-level image captions. Common deep CNNs are trained in a supervised way for this classification task, based on the IRMA dataset [4], containing around 7,300 X-ray images corresponding to the top 20 classes.

3. Results

The analyses of DICOM image tags in the ES, like "Modality (0008,0060)" and "Derivation Description (0008,2111)", identified potential problems to train a deep learning model base on our clinical data. For annotating with DL classification results, PNG images can also be annotated with an average precision 0.62 (F1 Measure of 0.52) of IRMA top 20 classes and 0.6 (F1 Measure of 0.55) for their anatomical segments, even the dataset contains images with optical characters and surgical frames.

4. Discussion

Data Lake with the use of Elasticsearch and S3 can not only handle healthcare messages but also tackle more file types - like DICOM and PNG files in this article. The imbalance of the dataset and inhomogenes settings affected the classification performance. However, better DL models or algorithms might be considered for further retrieval tasks. For instance, Content-Based Image Retrieve (CBIR) [4] is known as query by image content and returns images that are most similar to the query image.

5. Conclusions

A comprehensive retrievable medical image storage pipeline, including storing, feature indexing, and data searching, is designed and evaluated. Elasticsearch and S3 are powerful tools to retrieve any type of multi-media files in modern data lake solutions. We need to further improve our DL keyword generator to devise a reproducible way to integrate common DL technology and generate retrievable keywords for ES. Our next steps will be collecting, cleansing, training and evaluating with more data.

References

[1] Chen W, Liu Y, Wang W, Bakker EM, Georgiou T, Fieguth P, Liu L, Lew MS. Deep learning for instance retrieval: A survey. IEEE Trans Pattern Anal Mach Intell. 2022 Nov;1-20, doi: 10.1109/TPAMI.2022.3218591.

[2] Cheng KY, Pazmino S, Schreiweis B. ETL processes for integrating healthcare data - Tools and architecture patterns. Stud Health Technol Inform. 2022 Nov;299:151-6, doi: 10.3233/SHTI220974.

[3] Gundlach JP, Schmidt S, Bernsmeier A, Günther R, Kataev V, Trentmann J, Schäfer JP, Röcken C, Becker T, Braun F. Indication of liver transplantation for hepatocellular carcinoma. J Clin Med. 2021 Mar;10(6):1155, doi: 10.3390/jcm10061155.

[4] Zhang S, Zhi L, Zhou T. Medical image retrieval using empirical mode decomposition with deep convolutional neural network. Biomed Res Int. 2020 Dec;2020:6687733, doi: 10.1155/2020/6687733.