


Unsupervised Analysis of Chest Radiograph and Radiology Impressions

TEAM:

LAHARI BONI

JAYATHA CHANDRA





Project Idea

- ▶ Chest x-rays provide crucial visual information to diagnose health conditions. But fully understanding these scans takes specialized medical skill and effort. We want to make this knowledge more accessible.
- ▶ Our project explores predictive analytics classifying scans combined with text summarization of visual patterns. Together, this pipeline aims to replicate manual review more efficiently. We cluster images based on common anatomy, map new scans to these groups via algorithms, and generate summaries reflecting assigned clusters. Effectively we transform image insights into accessible data.

- Chest x-rays provide crucial visual information to diagnose health conditions. But fully understanding these scans takes specialized medical skill and effort. We want to make this knowledge more accessible.
- Our project explores predictive analytics classifying scans combined with text summarization of visual patterns. Together, this pipeline aims to replicate manual review more efficiently. We cluster images based on common anatomy, map new scans to these groups via algorithms, and generate summaries reflecting assigned clusters. Effectively we transform image insights into accessible data.

Dataset Description:

- ▶ The dataset is an extensive collection of chest X-rays sourced from Indiana University's publicly available medical archives.
- ▶ Dataset includes two primary components: over 7000 radiographic images and two CSV files – one containing the reports and the other detailing the image projections.

Dataset Pre-Processing:

- ▶ Checked for missing values in CSV files and imputed with empty strings and removed missing data as the medical data is sensitive.
- ▶ Merged the files through aggregation for clustering, text analysis and added a new column called 'caption' based on impression and findings columns.
- ▶ Used the preprocessed caption column for image captioning.

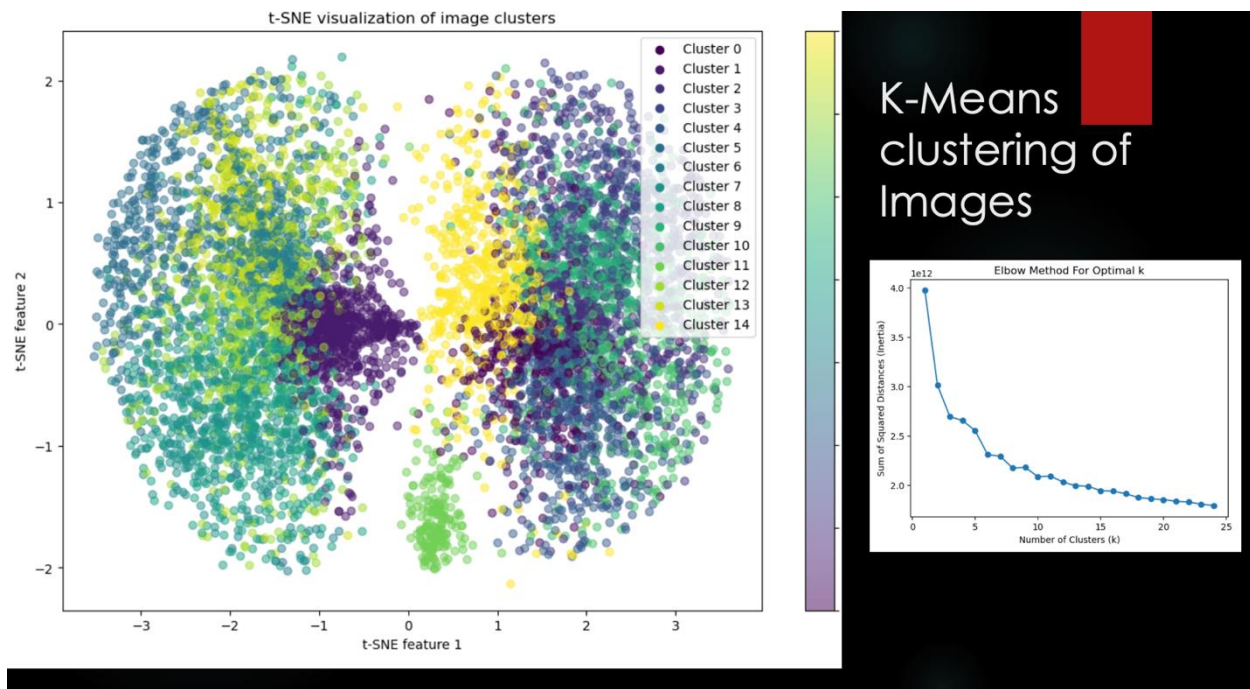
Dataset Description:

The dataset under consideration is an extensive collection of chest X-rays sourced from Indiana University's publicly available medical archives. The X-ray images are digital scans that provide visual insights into the thoracic cavity, showcasing various anatomical and pathological conditions. Accompanying these images are textual medical reports, meticulously detailing the findings and impressions of radiologists.

The Indiana University Chest X-Ray dataset is a significant collection encompassing over 7,000 chest X-ray images, each accompanied by detailed medical reports. The dataset includes two primary components: the radiographic images and two CSV files – one containing the reports and the other detailing the image projections.

Data Pre-Processing:

- **Checked for missing and null values:** Imputed with empty strings and removed missing data as it is sensitive to manipulate.
- **Merging data files:** Using aggregation we merged data from both CSV files which was used for Clustering images, text analysis and text summarization.
- **Column created:** Created a new column named 'caption' which was used for Image Captioning.



Process of creating K-means clustering on the Chest X-Rays images:

The images in the dataset are converted into a 2D array using NumPy.

Cluster Size Determination:

- A range of values for the number of clusters (k _values) is chosen (from 1 to 24 in this case).
- The code iterates through each value of k , fits a MiniBatchKMeans clustering model to the data.
- The sum of squared distances of samples to their closest cluster center (inertia) is computed for each k value.

Elbow Method:

- The inertia values for different values of k are plotted on a graph.
- The elbow of the curve is typically considered as a point where the rate of decrease of inertia slows down. So, we choose this at $k=15$ value.

Image Clustering with KMeans and VGG16 Features:

Feature Extraction:

- The code uses the VGG16 model, pre-trained on ImageNet, for feature extraction. The model is configured to exclude the top layer (fully connected layer) to retain spatial information.
- Images are preprocessed and fed through the VGG16 model to obtain features.
- Features are flattened to create a 2D array (**features_flattened**).

Clustering:

- KMeans clustering is applied to the flattened feature vectors to group images into clusters. For all the clusters, labels are assigned.

Visualization: The clusters are visualized using t-SNE visualization as shown in the slide.

Text Summarization of Reports

```
Cluster 0 Summary:
There is persistent mild to moderate cardiomegaly . Multiple thoracic deformities due to osteoporosis . No active disease. No acute pulmonary disease . No pneumothorax . No destructive lesions of t

Cluster 1 Summary:
Cardiomegaly with low lung volumes which are grossly clear . Low lung volumes with right basilar atelectasis . No acute cardiopulmonary abnormality identified . No active disease . No suspicious appe

Cluster 2 Summary:
Heart size borderline enlarged for technique, tracheostomy tube tip approximately 4.7 cm above the carina . Minimal degenerative changes of the thoracic spine . Mild bronchovascular crowding without

Cluster 3 Summary:
Left lower lobe, superior segment, airspace consolidation, radiographic appearance most typical for pneumonia . Left parapneumonic pleural effusion. No acute radiographic cardiopulmonary process. Sta

Cluster 4 Summary:
Heart size within normal limits, minimal aortic ectasia/tortuosity . No definite pleural effusion seen, no pneumothorax . Metastatic disease is possible . No acute disease. No acute cardiopulmonary f

Cluster 5 Summary:
No acute or active cardiac, pulmonary or pleural disease . No evidence of active disease . Bullous emphysema and interstitial fibrosis with no acute cardiopulmonary findings . Negative chest radiogr

Cluster 6 Summary:
The patient was diagnosed with pulmonary interstitial edema and bilateral pleural effusions . The patient had no active disease, no acute disease. No evidence of active disease . No signs of acute co

Cluster 7 Summary:
Follow up to resolution recommended with followup chest, abdomen and pelvis with contrast for further evaluation . Cardiomegaly with no focal airspace disease . No evidence of active disease. No evid

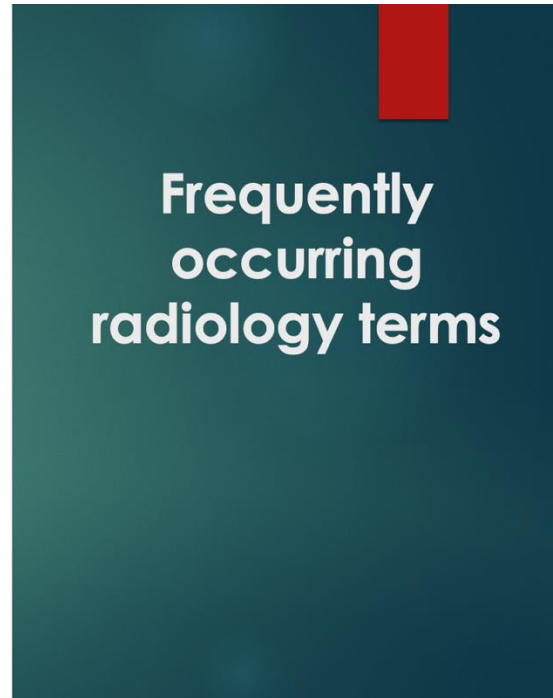
Cluster 8 Summary:
...

Cluster 14 Summary:
Left-sided biventricular cardiac pacemaker . Leads appear intact . No evidence of pneumothorax . Stable right-sided cardiac generator with right atrial and right ventricular leads, sternotomy .
```

Text Summarization of Reports:

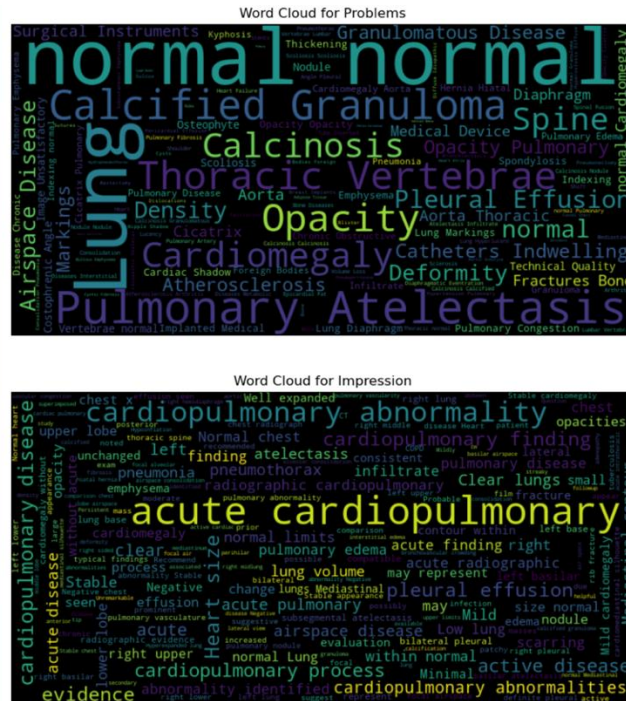
The summarization is performed using a Hugging Face transformers library, pre-trained BART model, and the resulting summaries are printed for further analysis. In our case, the automated summarization of clusters of reports is completed through a systematic process as shown below:

- First, it iterates over clusters using the `grouped_reports` variable, effectively organizing and segregating the input reports based on their assigned clusters.
- Subsequently, for each cluster, the code retrieves the corresponding reports. The `summarize_reports` function is then invoked to generate a concise summary, capturing the essential information from the set of reports within the cluster.
- Finally, the code prints the cluster number along with its associated summary, offering a comprehensible overview of the key insights encapsulated within each report cluster. This approach streamlines the summarization process, providing an efficient means to distill valuable information from diverse sets of reports.



- **Merging data:** The findings and impression column data are combined for analysis.
- **Tokenization:** The text is broken down into smaller units called tokens.
- **Counting Frequencies:** After preprocessing, the frequency of each token is counted. The number of times a token appears in the text is recorded.
- **Identifying Most Common Terms:** The terms with the highest frequencies are considered the most common. These terms are often key indicators of the prevalent themes or topics within the text.
- **Visualization:** To better understand the distribution, the results are visualized using Word Cloud.

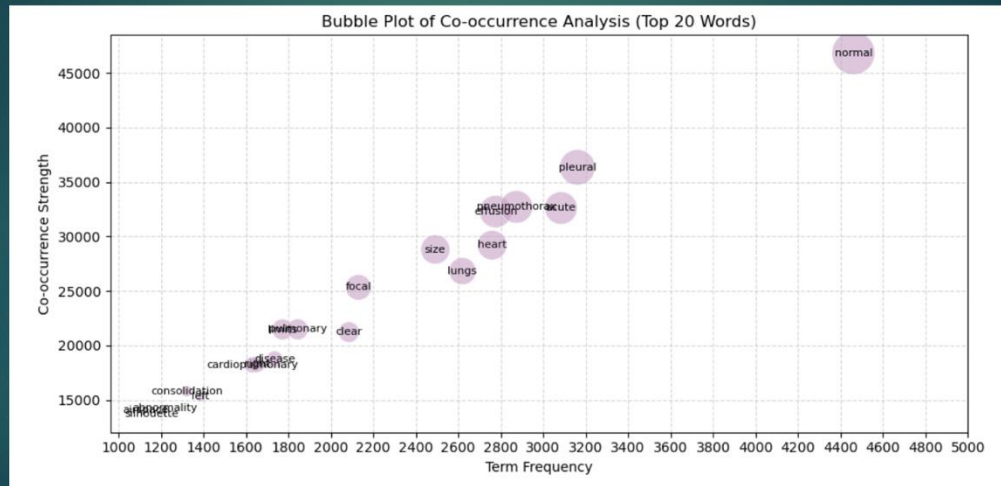
Correlation between Problem and impression



Process of Correlation analysis between the 'Problems' and 'Impression' from medical report:

- **Handling Missing Values:** Initial preprocessing involves addressing missing values in these columns by replacing them with empty strings, considering the sensitive nature of medical data.
- **Textual Representation:** Text data is transformed into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This captures the importance of terms within the context of each document.
- **Cosine Similarity Computation:** The cosine similarity between TF-IDF vectors of 'Problems' and 'Impression' is computed, generating a similarity matrix. This quantifies the textual correlation between the two features.
- **Mean Cosine Similarity:** The mean cosine similarity across all pairs is calculated, providing a quantitative measure of the overall textual correlation between 'Problems' and 'Impression.'
- **Statistical Analysis (One-Sample t-test):** A one-sample t-test is conducted to assess if the mean similarity significantly differs from a null hypothesis mean of 0, indicating no correlation. The resulting p-value is compared with a predefined significance level (alpha).
- **Statistical Inference:** Based on the p-value comparison, the analysis concludes whether there is a statistically significant correlation between the textual content of 'Problems' and 'Impression,' providing insights into their relationship.
- **Visualization:** Word Cloud shows the distribution of terms and their importance in the textual correlation analysis.

Co-occurrence analysis of findings and impression

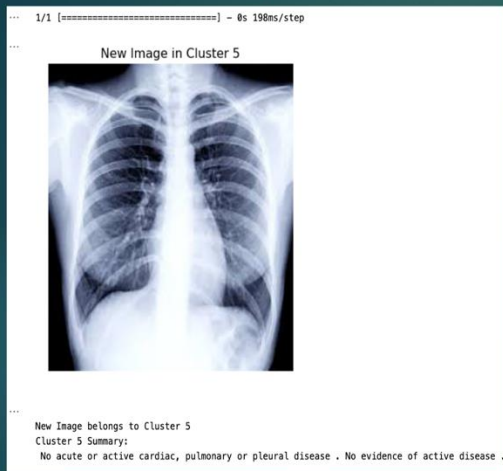


The below steps explain co-occurrence analysis of findings and impression columns in the medical reports:

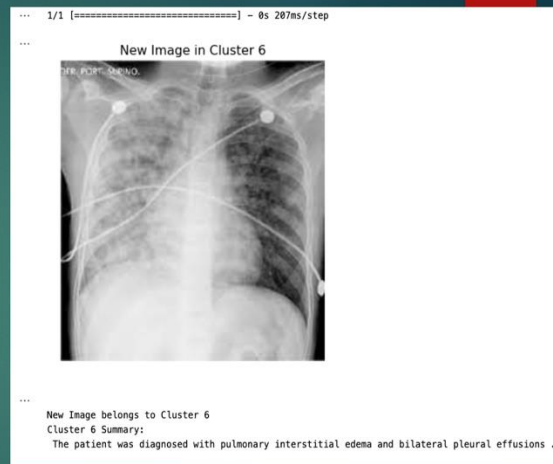
- **Text Concatenation:** The 'findings' and 'impression' columns are concatenated into a single 'text' column, creating a consolidated representation of each medical report.
- **Count Vectorization:** The CountVectorizer is employed to convert the textual data into a term-document matrix, where each row corresponds to a medical report and each column corresponds to a unique term. We limited the analysis to the top 20 most frequent terms to focus on the most relevant information.
- **Co-occurrence Matrix Construction:** The CountVectorizer output is a term-document matrix represented as a sparse matrix. The co-occurrence matrix is computed by transposing the term-document matrix (X^T) and multiplying it with the original matrix (X). The resulting matrix, `co_occurrence_matrix`, captures the frequency of co-occurrence between terms. Each element `co_occurrence_matrix[i, j]` denotes how often term `i` and term `j` appear together across all documents.
- **Diagonal Values Adjustment:** To focus solely on inter-term relationships, the diagonal values of the co-occurrence matrix (self-co-occurrence) are set to 0. This ensures that terms do not contribute to their own co-occurrence strength.

In summary, the co-occurrence matrix provides a quantitative representation of the relationships between terms in the concatenated medical reports. Each entry in the matrix indicates how frequently two terms occur together, shedding light on potential associations and patterns within the text data. This matrix serves as a foundational element for subsequent analyses, such as calculating co-occurrence strength and generating visualizations like the bubble plot

Generating Cluster Summary for new image



Cluster 5 Summary: No acute or active cardiac, pulmonary or pleural disease . No evidence of active disease .



Cluster 6 Summary:
The patient was diagnosed with pulmonary interstitial edema and bilateral pleural effusions .

Generating Cluster Caption/Summary:

The integration of image clustering and natural language processing enriches the analysis of new chest X-Ray images. Upon loading a new medical image, we preprocess it and extract features using VGG16. KMeans clustering is then employed to assign the image to a specific cluster. Subsequently, we retrieve reports from the dataset associated with the identified cluster. To distill complex medical reports, we utilize natural language processing techniques for summarization.

Benefits:

- **Holistic Insight:** By categorizing images into clusters and summarizing associated reports, healthcare professionals gain a holistic view of patient conditions. This holistic insight enhances diagnostic accuracy and facilitates more informed decision-making.
- **Efficient Data Interpretation:** Through automated clustering and summarization, healthcare practitioners can quickly extract key information from reports, saving time and reducing the cognitive load associated with manual data analysis. This efficiency is particularly beneficial in fast-paced healthcare settings, allowing for more effective patient care and resource allocation.

Challenges

1. Dataset Management:

- **Challenge:** Managing over 7000 images from 3000+ patients, each with two types of images (frontal and lateral).
- **Focus:** Accurate image-report pairing, diverse medical conditions, data privacy.

2. Feature Extraction Complexity:

- **Challenge:** Resource-heavy feature extraction from thousands of images using VGG16.
- **Focus:** Need for high computational power and efficient processing methods.

3. Multimodal Data Integration:

- **Challenge:** Merging image features with textual medical reports, handling two image types per patient.
- **Focus:** Effective model design for dual-data integration and interpretation.

4. Text Summarization Hurdles:

- **Challenge:** Difficulties in summarizing medical reports using standard tools like Gensim and NLTK, which were inadequate for the specialized medical text.
- **Focus:** Utilizing the Summa package, which proved more effective in processing complex medical terminologies and structures for precise and contextually relevant summaries

The Challenges we faced during our project implementation and how we over came those are listed below:

1. Dataset Management:

- Challenge: Managing over 7000 images from 3000+ patients, each with two types of images (frontal and lateral).
- Focus: Ensuring accurate pairing of images with their corresponding medical reports.

2. Feature Extraction Complexity:

- Challenge: Extracting features from thousands of images using the resource-intensive VGG16 model.
- Focus: Necessity for substantial computational resources and the development of efficient processing methodologies to manage the high data volume.

3. Multimodal Data Integration:

- Challenge: Integrating image features with textual medical reports, particularly for patients with both frontal and lateral images.
- Focus: Crafting an effective model capable of balancing and interpreting this dual-modality data for accurate analysis.

4. Text Summarization Hurdles:

- Challenge: Difficulties in summarizing medical reports using standard tools like Gensim and NLTK, which were inadequate for the specialized medical text.
- Focus: Utilizing the Summa package, which proved more effective in processing complex medical terminologies and structures for precise and contextually relevant summaries.

Appendix:

Dataset source: <https://openi.nlm.nih.gov/faq#collection>

Project Code: [GitHub](#)