CLIP It Right: Connecting Medical Images with Meaning

9 min read · Just now

S Sarah KM

A Hands-on Workshop into Multimodal AI with BiomedCLIP

How do we teach machines to read an MRI like a radiologist and describe it like a doctor?

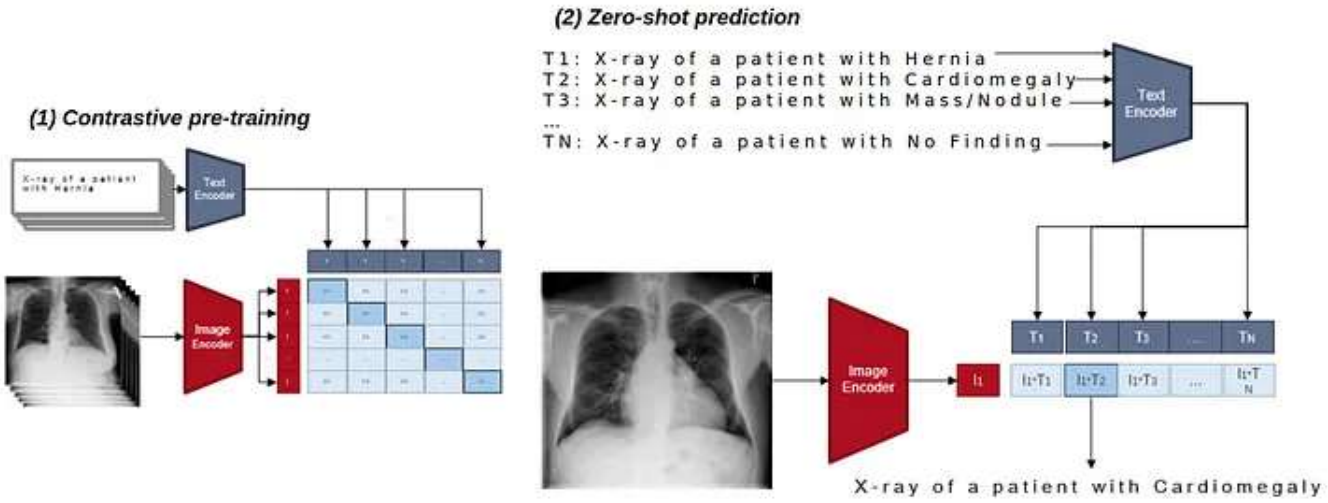
Welcome to this hands-on introduction to **multimodal AI in medicine**, where you learn how AI understands and combines clinical data (text and medical images).

In this workshop, we'll explore and experiment with a real multimodal AI model, ***BiomedCLIP-PubMedBERT***, which links clinical reports with medical images (e.g. chest X-rays) in a shared semantic space for predictions, retrieval, and classification tasks.

In this Workshop you will learn:

- ***How AI "understands" text***
  Keywords: tokenization, token embeddings, text encoder, text embeddings
- ***How AI "sees" images***
  Keywords: vision/image encoder, image embeddings
- ***How AI connects text & images to make predictions***
  Keywords: shared embedding space, zero-shot classification, retrieval, correlation-based prediction

We will walk through the CLIP pipeline (below) step by step to help you understand how each submodule works.



Before you continue, you can find a detailed hands-on Tutorial here:

Google Colab

Edit description

colab.research.google.com

We really want you to have the best experience possible, so we'd love you to have a good look through the provided notebook to help you get to know multimodal AI and what's important.

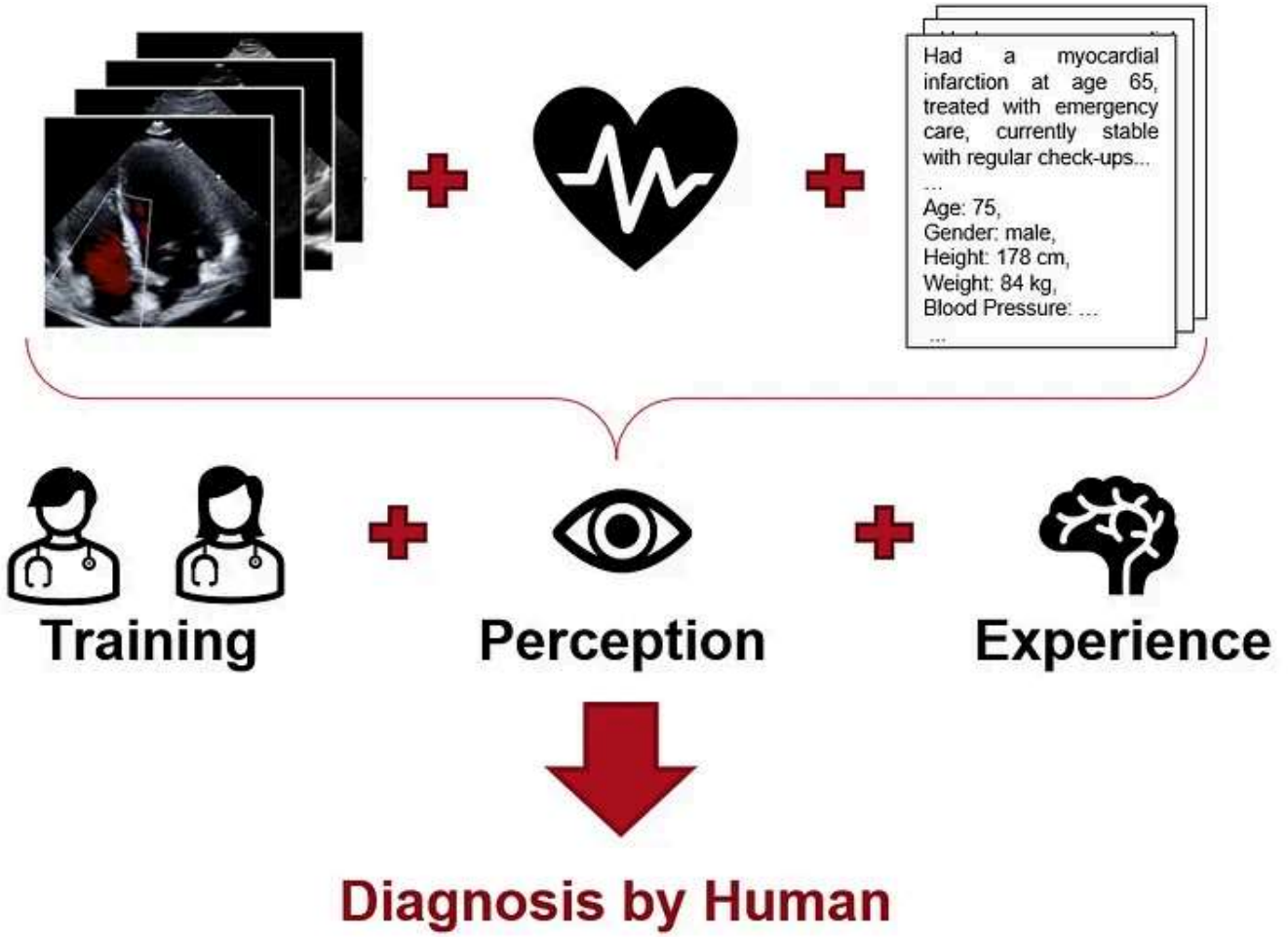
1. Introduction

Whether you're a clinician making a diagnosis or a data scientist building a model, you know that one data point is rarely enough. A clinician wouldn't base any decision on a single echocardiogram or ECG signal.

In medicine, decision making relies on combining diverse information sources:

- Medical imaging (e.g. echo, X-ray, MRI)
- Clinical reports and notes
- Lab values and vitals
- Waveforms (e.g. ECG, EEG)

This rich, layered approach is called *multimodal reasoning*.



So as you think multimodally — so should AI.

Why should AI make a decission based only on a single image?

Modern multimodal AI aims to mirror the way humans reason: By combining diverse information sources to improve **accuracy**, **interpretability**, and **generalizability**.

1.1 What Is Multimodal AI?

Multimodal AI is designed to integrate diverse data types from multiple sources into a **shared representation space**, allowing the AI to understand, compare an reason across them.

Think of it like a 'universal language' for data, where text, images, signals are all represented in the same way as **embeddings**. This shared landscape is called ***shared embedding space***.

By learning shared embeddings, a multimodal model can:

- Match a radiology report to the correct image (Which we will do today)

- Retrieve similar prior cases or reports from a large database
- Predict diagnosis

without task-specific training.

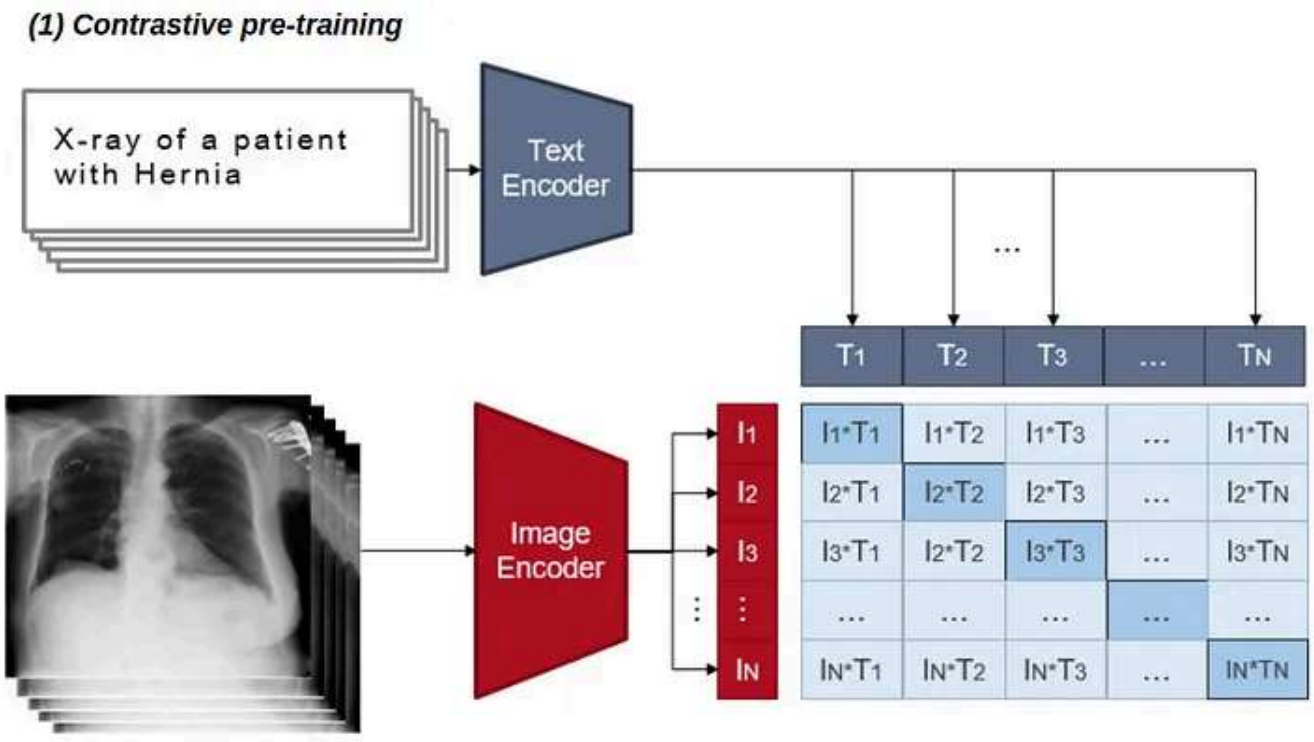
This is called ***zero-shot learning***, because the model generalizes to new tasks it wasn't explicitly trained on.

How Does It Work?

The recent most powerful method for this is **Contrastive Language-Image Pre-training (CLIP)**.

CLIP trains two neural networks:

- A ***text encoder*** that transforms text inputs into a single vector, representing its semantic content (***token embedding***).
- An ***image encoder*** does the same by transforming an image into a vector representing its visual content (***image embedding***).



The models are trained so that the vectors corresponding to semantically similar text-image pairs are close together in the shared vector space, while those corresponding to dissimilar pairs are far apart.

> The result? A model that can *connect meaning across modalities,* just like you do while looking at an image with the report in mind

1.2 Recent Work

Here is a small overview at how this approach is already being used in state-of-the-art approaches. You can find a more comprehensive overview in the survey paper *CLIP in Medical Imaging: A Survey* from *Zhao et al.*, 2025.

BiomedCLIP-PubMedBERT
Zhang et al. 2025

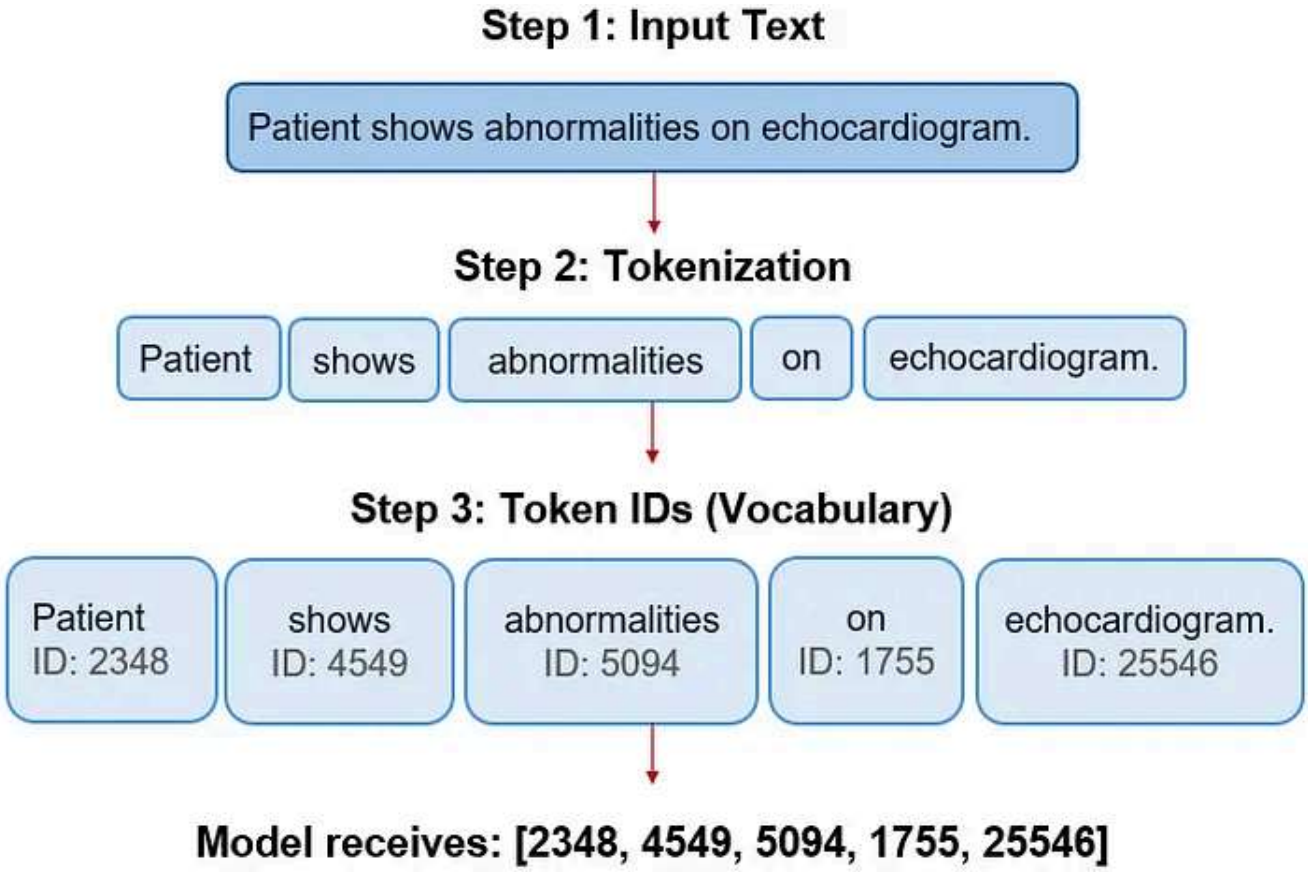
- Trained on millions of biomedical image–text pairs
- Enables zero-shot classification, retrieval, and label-free learning
- Foundation model for healthcare-specific language-vision tasks

2. How AI Sees Text

2.1 Tokenization

Before a model can *understand* a clinical text, it must first break it into smaller understandable pieces. This process is called ***tokenization,*** which is the foundation of how language models process and represent text.

Tokens are the building blocks of meaning, representing small text units numerically.

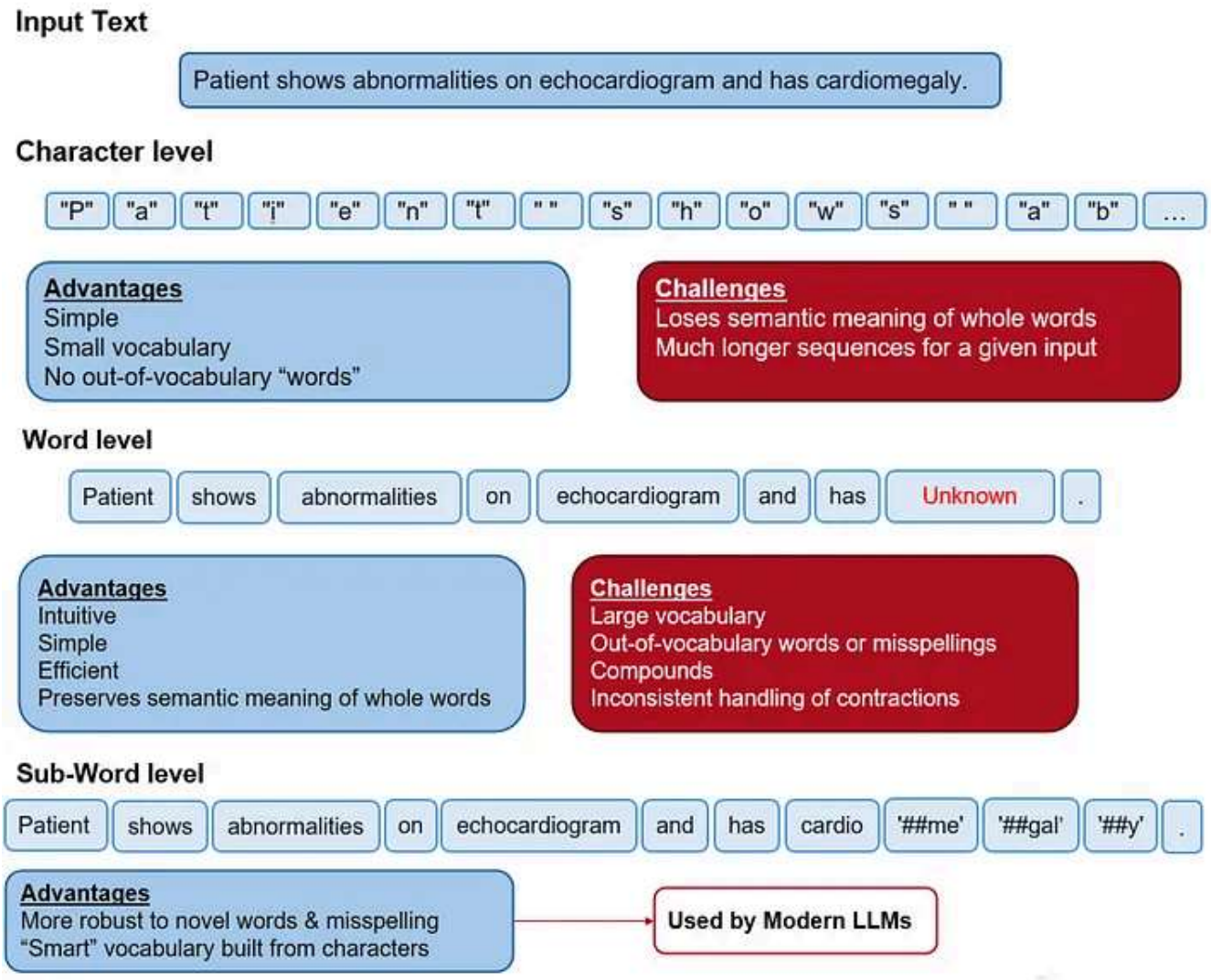


But what is the best way to tokenize your sentence?

Tokens might be:

- characters
- words
- sub-word combinations (like prefixes or syllables)

The different tokenization strategies have different trade-offs in vocabulary size, efficiency, and handling of rare or misspelled words:



> *Why does this matter?* Tokenization affects how the model represents meaning and how to handle things like rare words, misspellings or medical terminology.

📄 Many modern Large Language Models, like BERT (Bidirectional Encoder Representations from Transformers), use sub-word tokenization. In this workshop, we use PubMedBERT, a biomedical version of BERT trained on clinical and research text from PubMed.

Text:
The unicorn heart was examined using cardiac MRI scans

	Token	Token ID
0	[CLS]	2
1	the	1680
2	unic	20803
3	##orn	4610
4	heart	3196
5	was	1734
6	examined	2904
7	using	2019
8	cardiac	3443
9	mri	4668
10	scans	8115
11	[SEP]	3

You may recognised the special tokens `[CLS]` and `[SEP]`. These have a specific structural and runctional roles in Sentence Transformer models:

- `[CLS]` ('classification'): Appears at the beginning of the input sequence. Its final hidden state often serve as a summary representation of the entire sequence.
- `[SEP]` ('separator'): Marks boundaries between sentences, especially when processing pairs of sentences. Useful in tasks like semantic similarity or QA. Two sentences might be concatenated with a `[SEP]` token between them, allowing the model to distinguish where one senctence ends and the next begins.

So when encoding a single sentence, you might format it as `[CLS] sentence [SEP]`, while a sentence pair would be `[CLS] sentence1 [SEP] sentence2 [SEP]`.

2.2 Token Embedding

Now that the text has been broken down into smaller pieces, how does the model process it?

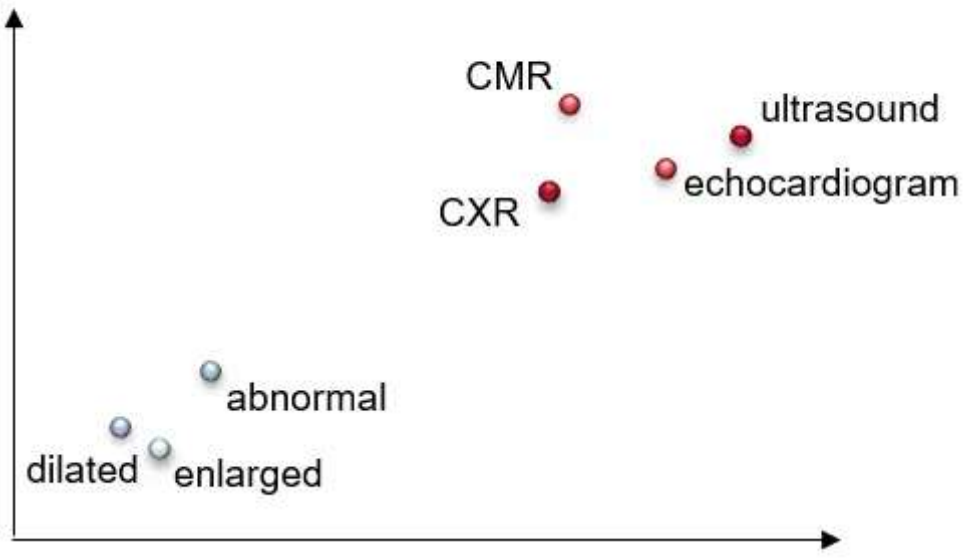
Once a sentence is tokenized, the model transforms each token into an n-dimensional vector, also known as a ***token embedding***. These high-dimensional representations capture both the meaning and context of each token.

For example, words with similar meanings tend to occur in similar contexts:

- The patient's **echocardiogram** showed an *abnormal* LV.
- The patient's **CMR** showed a *dilated* LV.

In both cases, the words **echocardiogram** and **CMR** appear in similar contexts, as do the terms *dilated* and *abnormal*.

The objective of the token embeddings is to represent the context of the words. We would expect **echocardiogram** and **CMR** to be closer together, while **CMR** and *dilated* serve different roles and are expected to be farther apart.



> The goal is for the model to learn embeddings that reflect *semantic similarity*.

2.3 Text Feature Embedding

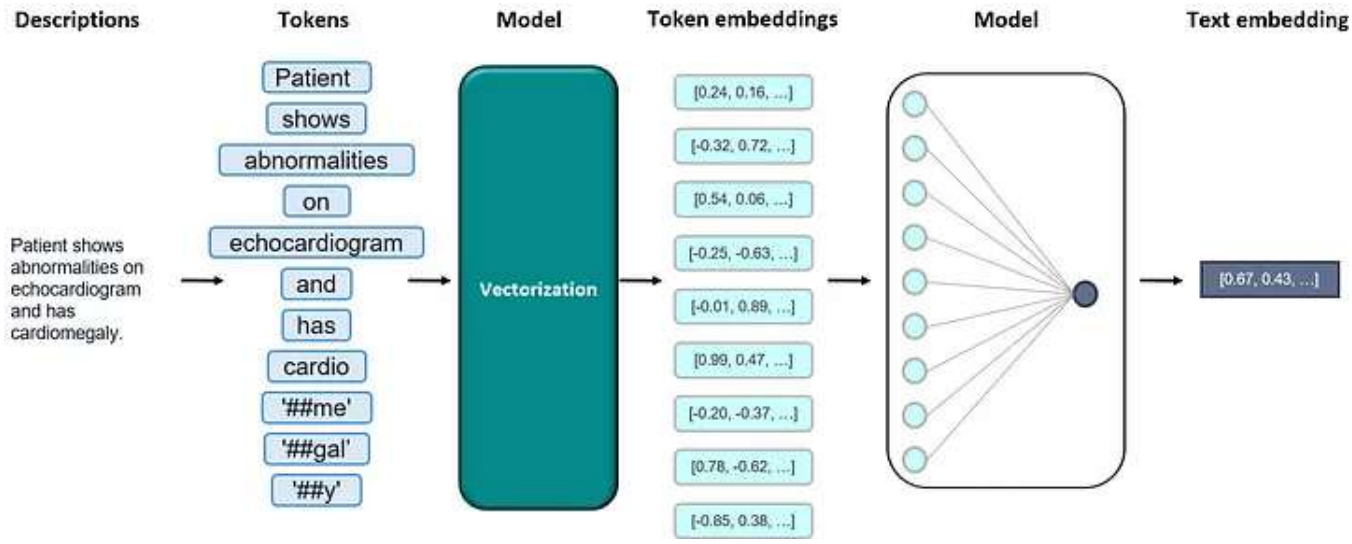
While individual tokens carry meaning, language becomes powerful when you think about the whole sentence or text. Some words might be more relevant than others, or can have totally different meanings in different settings. The meaning of a sentence is more than the sum of its words, it only becomes apparent when the words are combined.

Take this example:

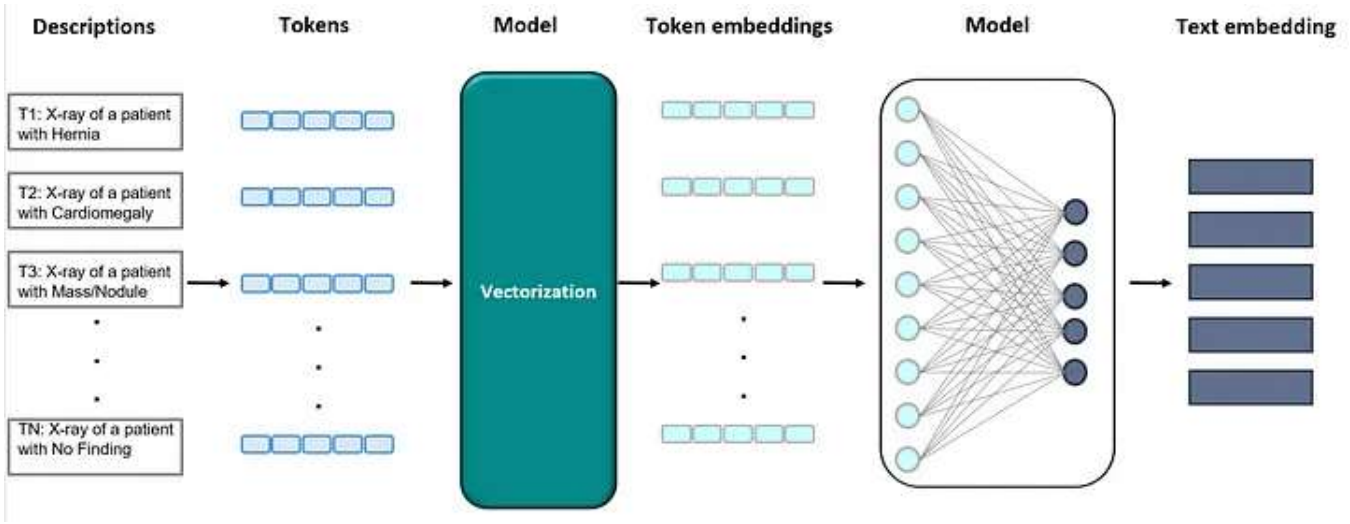
- *The duck is swimming on the lake.* ➞ "duck" = animal

- *You have to duck to go through the door.* ➞ “duck” = action

Despite the identical word, the context shifts the meaning entirely. To account for this, the next step is to summarise the token embeddings of a sentence into a single ***text embedding*** that reflects its meaning.



We do this for several sentences in order to create a ***text embedding space***. By comparing these embeddings, the model can identiy semantically similar sentences, even if they use different words.

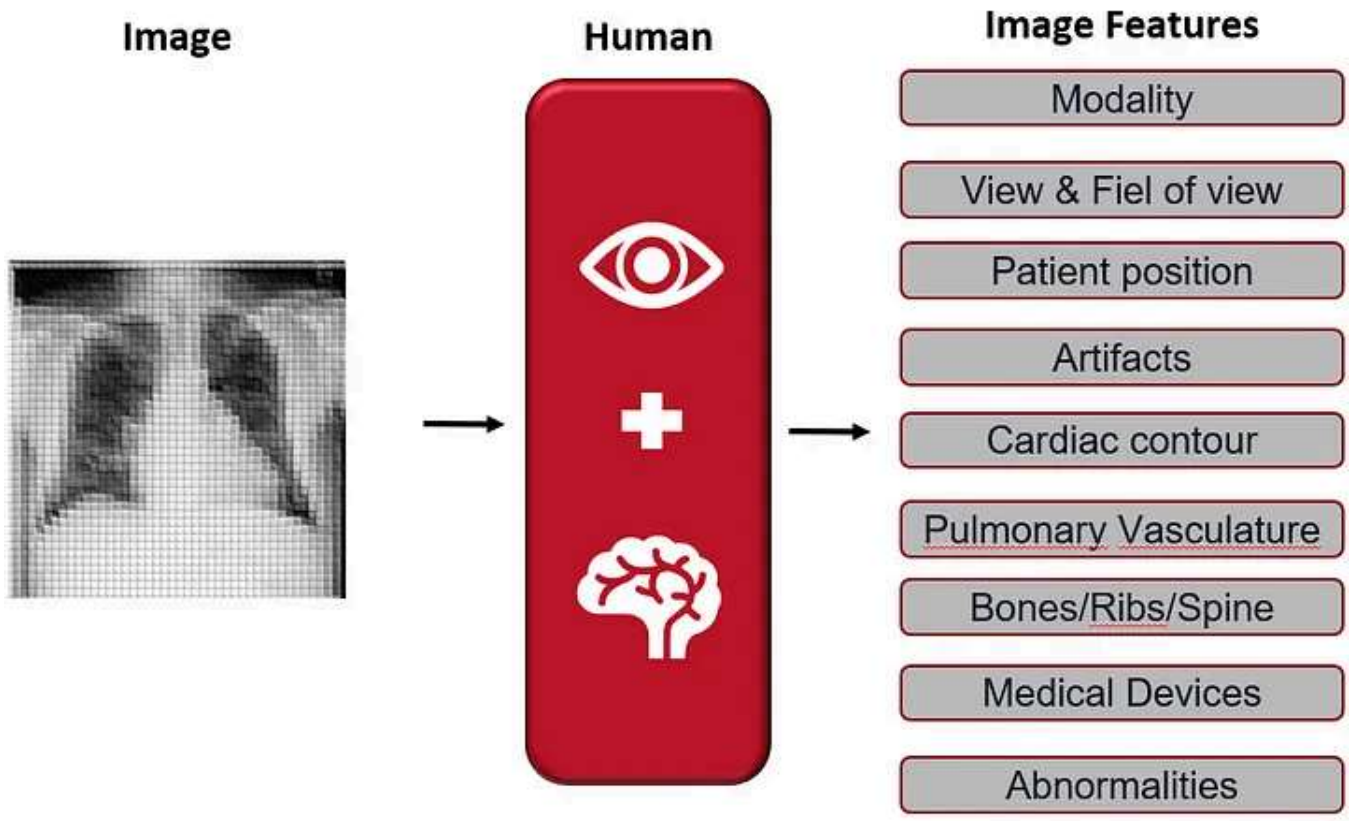


Now that we understand how AI understands text, it’s time to look at image processing.

3. How AI Sees Images

When you look at a chest X-ray, you don’t examine every pixel one by one. Instead, you recognize global and local patterns in the image. You can identify the modality,

the contour of the heart, artefacts, abnormalities or devices.



We want AI to process images in a similarly intelligent way: Not by memorizing raw pixels, but by learning meaningul visual that help it to understand what it “sees”.

3.1 Image Embedding

Before 2021, most AI systems in medical imaging used *Convolutional Neural Networks (CNNs)* to process images. CNNs are effective for local pattern recoginition, such as edges or textures. However, they are not suitable fo capturing global relationships across an image.

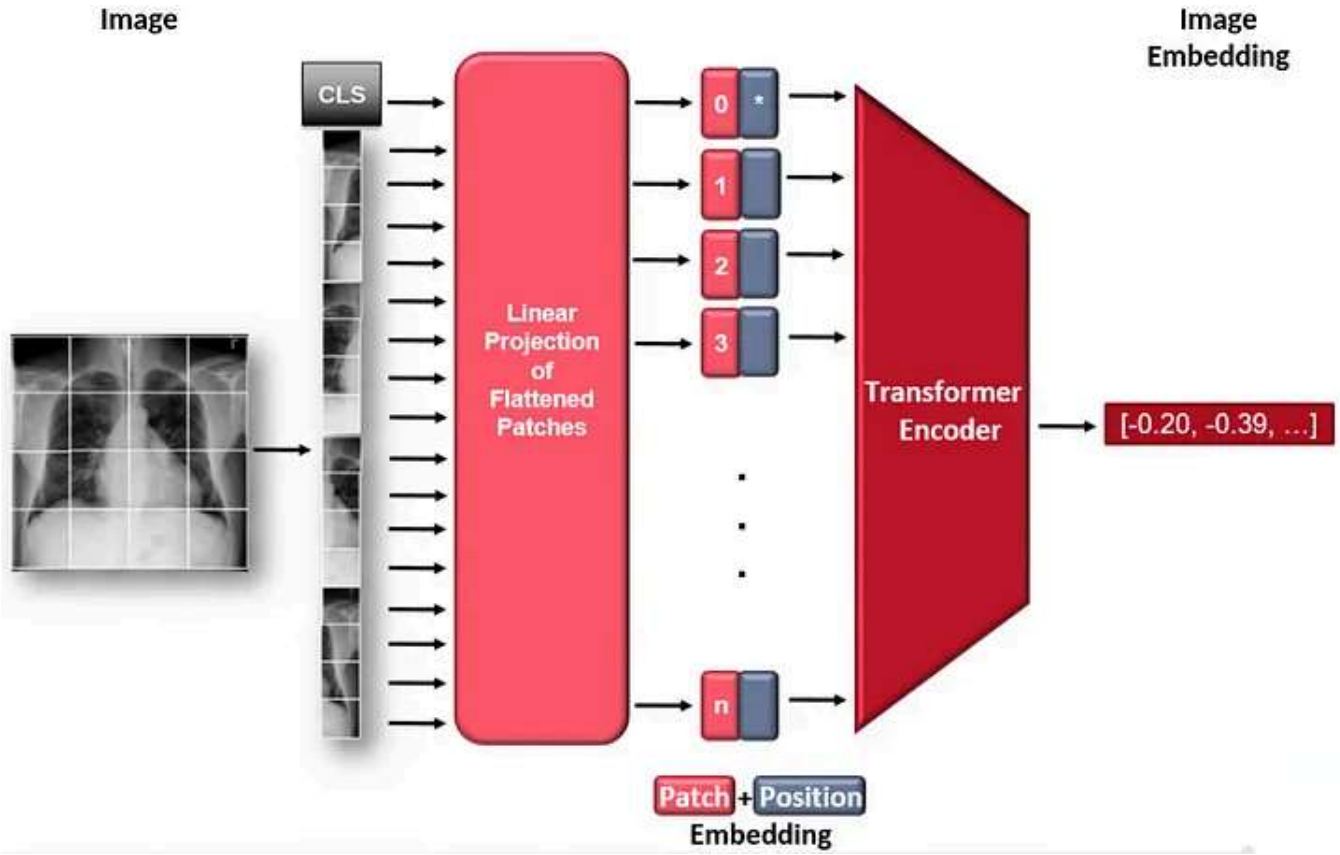
Then came the ***Vision Transformer (ViT)***. Inspired by the Transformer archtecture in NLP, ViT treats an image not as a 2D grid of pixels, but as a **sequence of patches**, much like tokens in text.

Here’s how ViT works:

1. **Patch Embedding** The image is divided into non-overlapping patches. Each patch is flattened and linearly projected into a fixed-size vector, forming a patch embedding.

2. **Positional Encoding** Transformers are order-agnostic. This means, that they do not memorize the position of the image patch. Therefore, positional encodings are added to retain the spatial structure of the image.
3. **Transformer Encoding** The Patch and Positional embeddings are passed through a standard Transformer encoder, using self-attention to learn the relationships between patches in the image.
4. **Image Embedding (CLS Token)** A special `[CLS]` token is prepended to the patch sequence, similar to the `[CLS]` token in text embeddings. Its final hidden state is used as the summary representation of the entire image.

This final hidden state token is the ***image embedding***. This compact, high-dimensional vector captures the semantic content of the image.



📄 ViT were developed by a Team from Google in 2021. Their work was published in An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

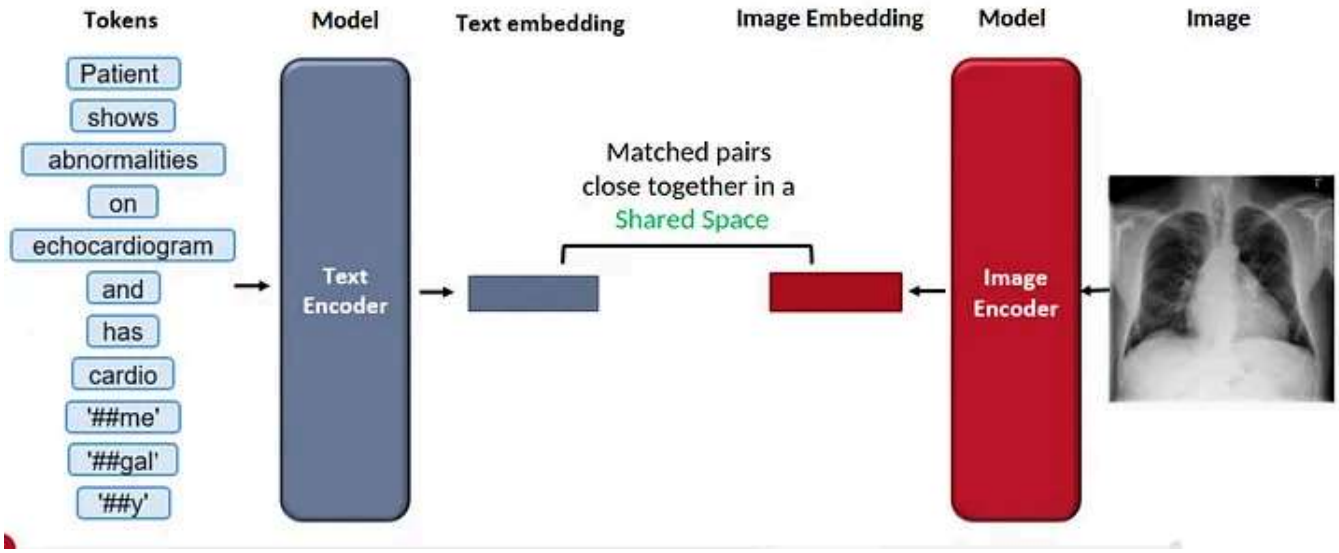
4. Speaking the Same Language

So far, we've explored ho AI processes text, via tokenization and text embeddings, and images, via ViT and image embeddings.

But how do we get them to "talk" to each other?

For this, we need to project both text and image embeddings into a ***shared embedding space***. In this "place" different modalities (like clinical reports and X-rays) can be directly compared.

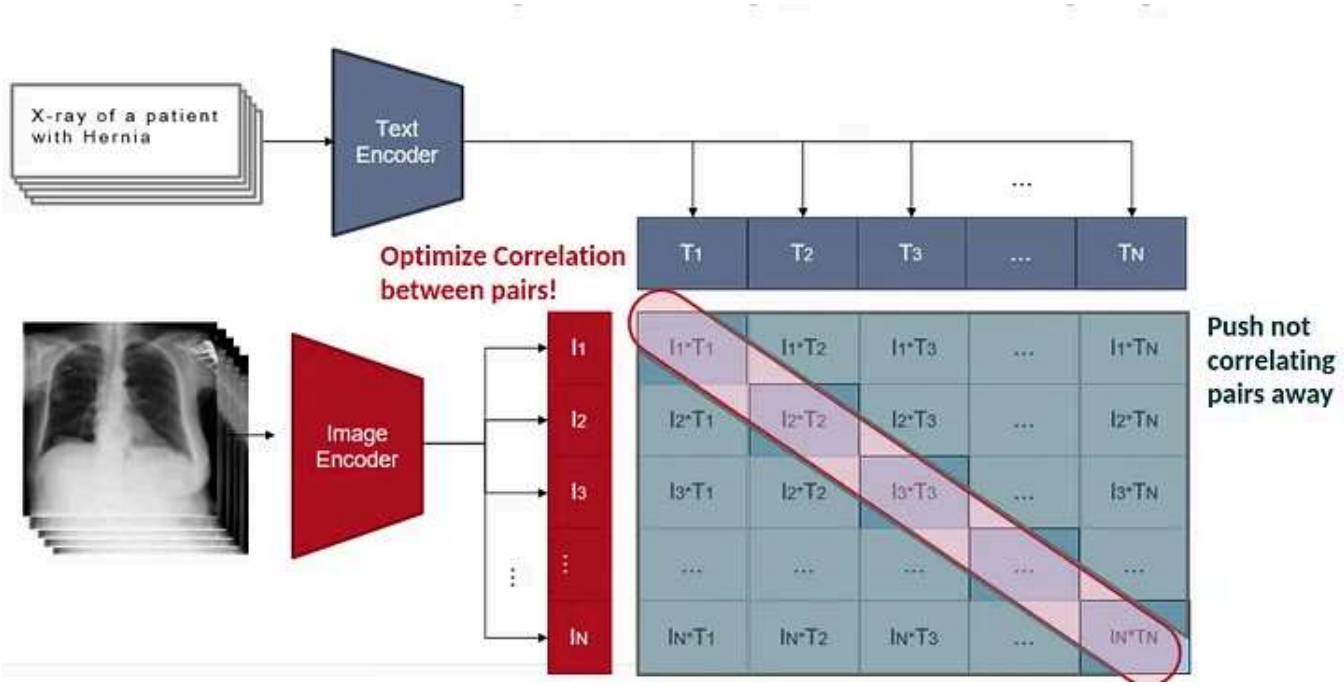
Think of this as a *common language*. A sentence like "the X-ray of a patient shows cardiomegaly" and the matching image are mapped to nearby coordinates, even though they started in completely different formats.



4.1 Multimodal Embedding

This shared space is trained using a ***contrastive learning*** ojective:

- If a report and an image *match*, their embeddings are pulled closer together.
- If they *don't match*, their embeddings are pushed apart.



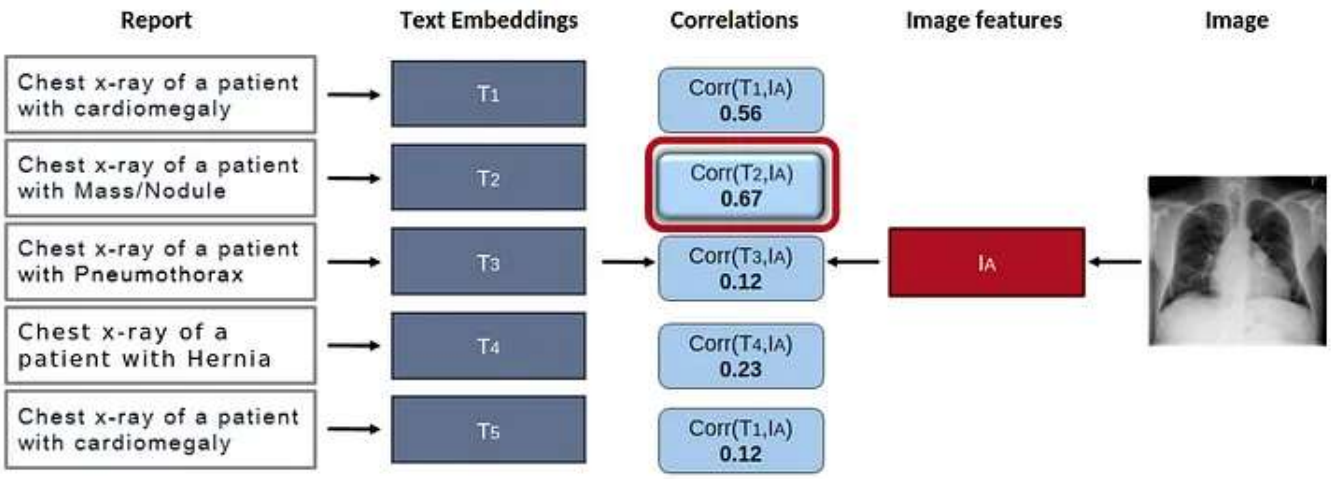
The model learns **semantic alignemnet** between visual and textual inputs. This enbables the model to associate specific clinical descriptions with specific visual patterns, without the need for explicit labels.

4.2 Zero-Shot Classification

Now that both embeddings live in the same space, speaking the same language, we can do something powerful:

Make predictions **without training on labeled examples**. This is called ***zero-shot classification***. It works like this:

- You define a set of possible diagnostic labels as natural-language prompts (e.g., “no abnormality”, “cardiomegaly”, “pleural effusion”).
- Each label is converted into a text embedding.
- For a given image, you comput its image embedding.
- Then you measure similarity (e.g., cosine distance) between the image embedding and each label embedding.
- The **closest match** is selected as the prediciton.



> Why does this matter? It allows clinicians and researches to build new diagnostic pipelines without retraining or annotating thousands of images.

Congratulations!!

You went through this interactive course for multimodal AI in medicine. Hopefully you had fun and are now more familiar with tokenisation, text and vision encoder, and CLIP.

Cheers,

Sarah

Salman

From the Institute for Artificial Intelligence in Cardiovascular Medicine (*Leader:* Prof. Dr. Sandy Engelhardt).

https://www.klinikum.uni-heidelberg.de/chirurgische-klinik-zentrum/herzchirurgie/forschung/ag-artificial-intelligence-in-cardiovascular-medicine

Multimodal Ai | Biomedclip | Tokenization | Transformers

Edit profile

Written by Sarah KM

0 followers · 1 following

No responses yet

Sarah KM

What are your thoughts?

Recommended from Medium