Contrastive Learning in the Medical Domain

For Visual Question Answering and Phenotyping

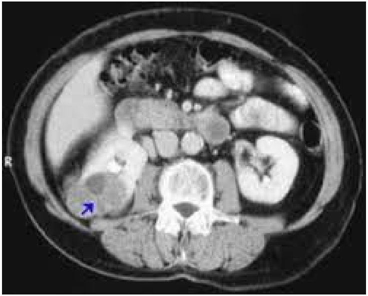
Chase Mathis

SIBMI Student at HMS

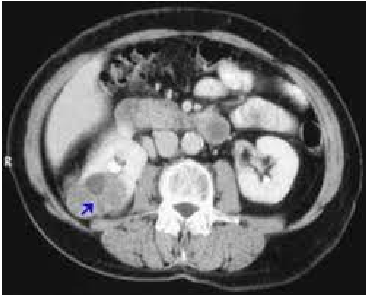
August 9, 2023

Overview

Long Term Goals



Long Term Goals

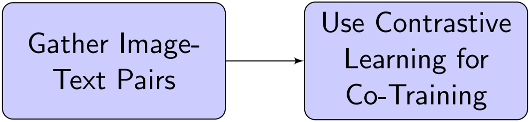


$\rightarrow$ Recurrence

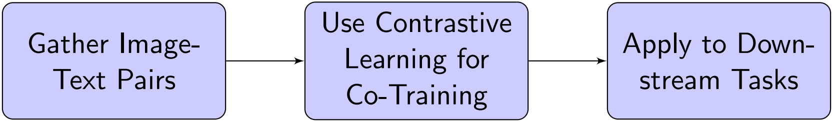
Developing a 2D Algorithm

Gather Image-Text Pairs

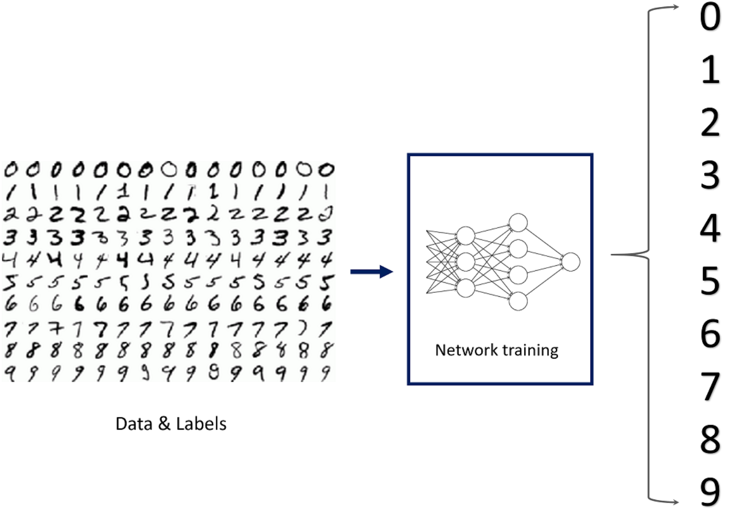
Developing a 2D Algorithm



Developing a 2D Algorithm

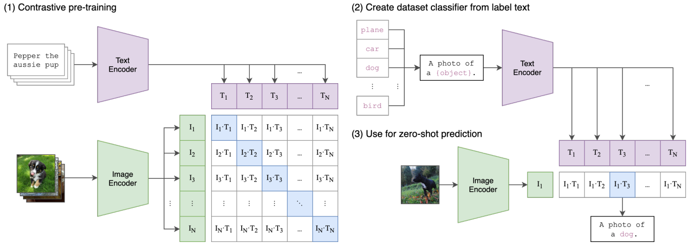


Contrastive Learning 101



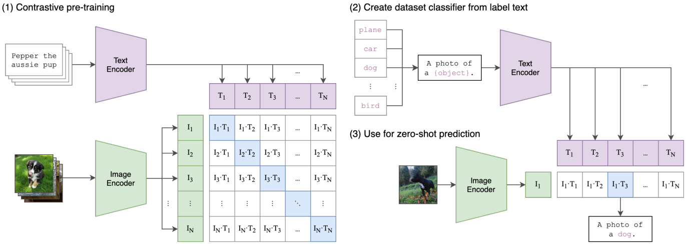
Neural networks predict a class based on an image input.

Contrastive Learning 101



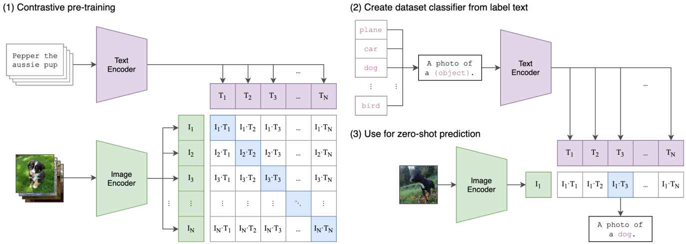
- Contrastive Learning focuses on placing pairs of data close together in a hyper-dimensional space, rather than predicting classes.

Contrastive Learning 101



- Contrastive Learning focuses on placing pairs of data close together in a hyper-dimensional space, rather than predicting classes.
- It involves an image encoder f : Image $\to \mathbb{R}^e$ and a text encoder g : Text $\to \mathbb{R}^e$.

Contrastive Learning 101

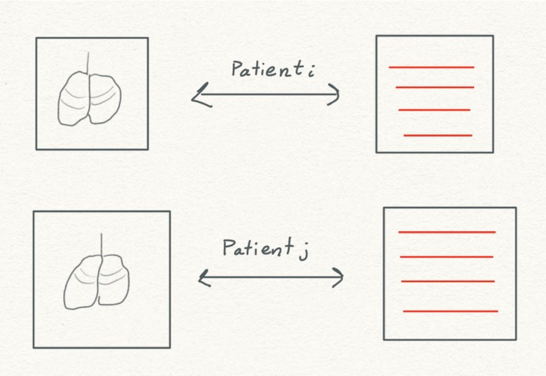


- Contrastive Learning focuses on placing pairs of data close together in a hyper-dimensional space, rather than predicting classes.
- It involves an image encoder f : Image $\to \mathbb{R}^e$ and a text encoder g : Text $\to \mathbb{R}^e$.
- Encoders allow us to preform many downstream tasks.

Contrastive Learning Video

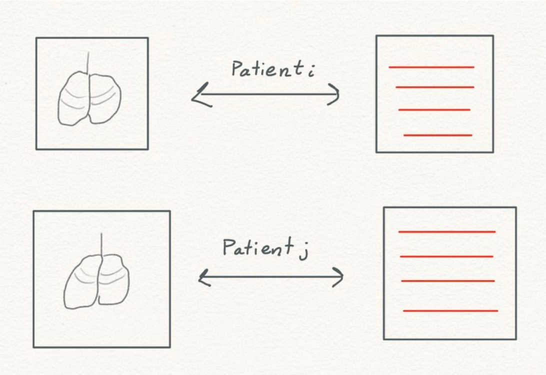
Video

MedCLIP Pretraining



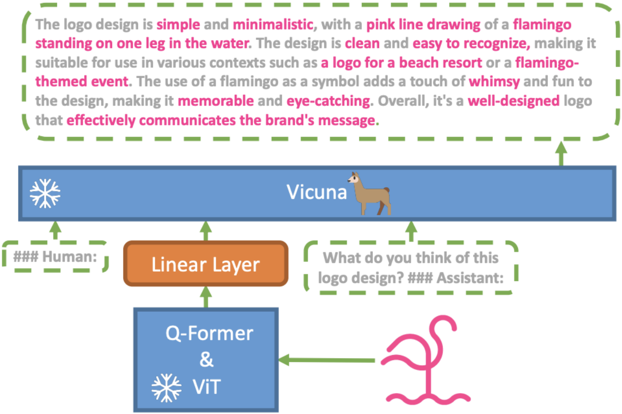
- We used the MedCLIP algorithm (Wang et al. 2022).

MedCLIP Pretraining



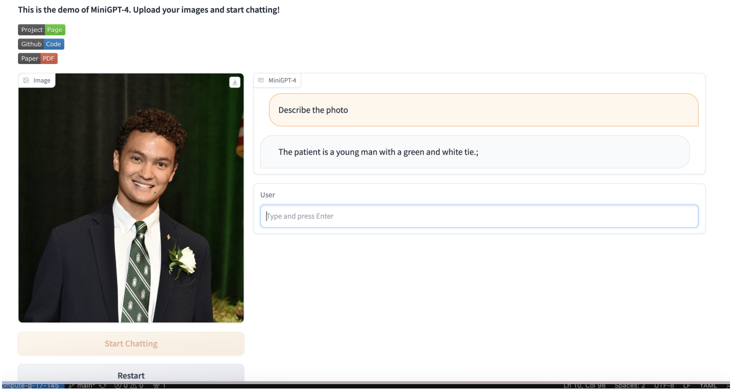
- We used the MedCLIP algorithm (Wang et al. 2022).
- This algorithm differs from the classic CLIP model, as it treats image text pairs more broadly.

Downstream Task 1: Radiology Report Q&A



- We used the MiniGPT4 model for VQA (Zhu et al. 2023).
- To align the model with medical imaging, we made **three** changes to the model.

First Change



Using a Vicuna LLM fine tuned on discharge notes, the model behaved more like a clinician

Issues Continue

MiniGPT4 generated text:
"The lungs are clear with no evidence of pulmonary edema . There is no evidence of cardiomegaly or pericardial effusion . There is no pleural effusion or pneumothorax ... **The patient is wearing a T - shirt ... The patient's creatinine level is elevated at 1.6 mg/dL**, which is higher than the normal range of 0.6-1.3 mg/dL for females.

Second Change

- The miniGPT4 **frozen** ViT was trained using general images.

Second Change

- The miniGPT4 **frozen** ViT was trained using general images.
- A dog, tree, boat, plane, etc.

Second Change

- The miniGPT4 **frozen** ViT was trained using general images.
- A dog, tree, boat, plane, etc.
- Using the CLIP framework, we can train a vision encoder to recognize diseases.

Second Change

- The miniGPT4 **frozen** ViT was trained using general images.
- A dog, tree, boat, plane, etc.
- Using the CLIP framework, we can train a vision encoder to recognize diseases.
- We train this ViT on radiology specific images-text pairs and then re-insert it into the framework

Third Change

- General Image-Text contrastive learning assumes simple descriptions of the following "MadLib" style:
- A photo of a {**thing**} that is {**color**} doing {**activity**} with {**background description**} in the background.

Third Change

- General Image-Text contrastive learning assumes simple descriptions of the following "MadLib" style:
- A photo of a {**thing**} that is {**color**} doing {**activity**} with {**background description**} in the background.
- General text descriptions are simple and short.

Third Change

- General Image-Text contrastive learning assumes simple descriptions of the following "MadLib" style:
- A photo of a {**thing**} that is {**color**} doing {**activity**} with {**background description**} in the background.
- General text descriptions are simple and short.
- Radiology reports and complex and long.

Third Change

- General Image-Text contrastive learning assumes simple descriptions of the following "MadLib" style:
- A photo of a {**thing**} that is {**color**} doing {**activity**} with {**background description**} in the background.
- General text descriptions are simple and short.
- Radiology reports and complex and long.
- We split radiology report descriptions into sub-parts.

Third Change

- General Image-Text contrastive learning assumes simple descriptions of the following "MadLib" style:
- A photo of a {**thing**} that is {**color**} doing {**activity**} with {**background description**} in the background.
- General text descriptions are simple and short.
- Radiology reports and complex and long.
- We split radiology report descriptions into sub-parts.
- Randomly sample categories of the radiology report, then train.

Previous Work

- XRay-GPT and MedFlamingo Moor et al. 2023 work in this area of VQA.

Previous Work

- XRay-GPT and MedFlamingo Moor et al. 2023 work in this area of VQA.
- XRay-GPT trains the linear layer and LLM, without touching the ViT.

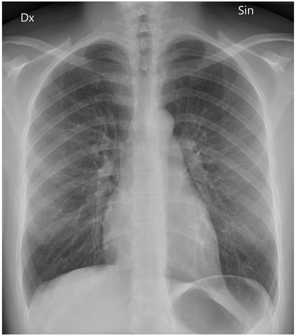
Previous Work

- XRay-GPT and MedFlamingo Moor et al. 2023 work in this area of VQA.
- XRay-GPT trains the linear layer and LLM, without touching the ViT.
- MedFlamingo is more general, training on a huge collection of data.
- MedFLamingo uses 8 A100's to train for over 6.75 days and has 1.3B trainable parameters.

Previous Work

- XRay-GPT and MedFlamingo Moor et al. 2023 work in this area of VQA.
- XRay-GPT trains the linear layer and LLM, without touching the ViT.
- MedFlamingo is more general, training on a huge collection of data.
- MedFLamingo uses 8 A100's to train for over 6.75 days and has 1.3B trainable parameters.
- Our algorithm uses one A100 for $\sim$ 10 hours of training and has only $\sim$ 3M trainable parameters

Example Outputs



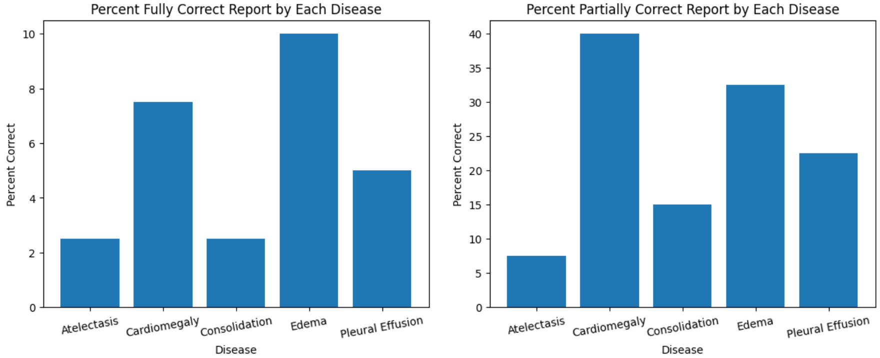
- Cardiomediastinum and pulmonary vasculature appear normal. There is mild interstitial opacity along the left upper lobe apex. The lung bases appear clear.
- "There is no change. The patient has developed mild right pleural effusion compared to ___. No pneumothorax observed."

Results

I used Chexpert's Labeler to classify the output. Irvin et al. 2019

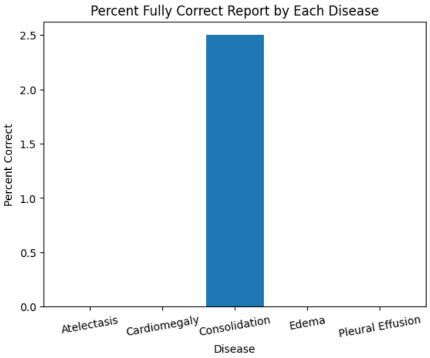
Results

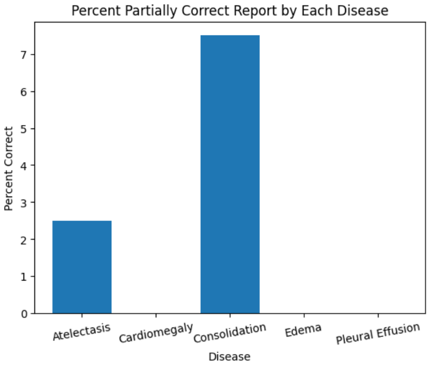
I used Chexpert's Labeler to classify the output. Irvin et al. 2019



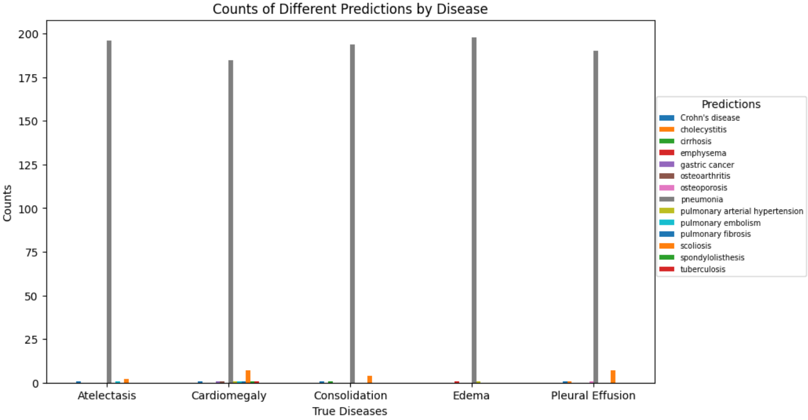
Note: ViT's best zero-shot accuracy was $\sim 45\%$

Comparison to MiniGPT4 without ViT





Comparison to MedFlamingo



Downstream Task 2: Remainder Information

- In statistical classification tasks, we wish for a parsimonious model.

Downstream Task 2: Remainder Information

- In statistical classification tasks, we wish for a parsimonious model.
- Why do we need the image?

Downstream Task 2: Remainder Information

- In statistical classification tasks, we wish for a parsimonious model.
- Why do we need the image?
- We devise a method to extract orthogonal information relevant in the image not represented in the text.

Creating the Input

- We split a report up into unigrams and bigrams.

Creating the Input

- We split a report up into unigrams and bigrams.
- Using the trained text-encoder, we have a list of embeddings $(\mathbf{w_1}, \ldots, \mathbf{w_n})$ of these words or phrases.

Creating the Input

- We split a report up into unigrams and bigrams.
- Using the trained text-encoder, we have a list of embeddings $(\mathbf{w_1}, \ldots, \mathbf{w_n})$ of these words or phrases.

"There is Pneumonia on the left" $\rightarrow$

Creating the Input

- We split a report up into unigrams and bigrams.
- Using the trained text-encoder, we have a list of embeddings $(\mathbf{w_1}, \ldots, \mathbf{w_n})$ of these words or phrases.

"There is Pneumonia on the left" $\rightarrow$
$\rightarrow$[There, is, Pneumonia, on, the, left, There is, is Pneumonia, Pneumonia on, on the, the left] $\rightarrow$

Creating the Input

- We split a report up into unigrams and bigrams.
- Using the trained text-encoder, we have a list of embeddings $(\mathbf{w_1}, \ldots, \mathbf{w_n})$ of these words or phrases.

"There is Pneumonia on the left" $\rightarrow$
$\rightarrow$[There, is, Pneumonia, on, the, left, There is, is Pneumonia, Pneumonia on, on the, the left] $\rightarrow$

$$\rightarrow \begin{bmatrix} - & w_1 & - \\ & \vdots & \\ - & w_k & - \end{bmatrix}$$

Phenotyping

Given the image embedding for the same patient **i**, we can find the top k phenotypes by computing the inner product (a matrix multiplication).

$$\begin{bmatrix} — & w_1 & — \\ & \vdots & \\ — & w_k & — \end{bmatrix} \begin{bmatrix} | \\ i \\ | \end{bmatrix} = \begin{bmatrix} s_1 \\ \vdots \\ s_k \end{bmatrix}$$

- We find the top k similarities and extract those.
- In our example "There is Pneumonia on the left" if we return for k = 2, (2, 5) $\rightarrow$ (Pneumonia, left) would be our phenotypes

Regression

$$\mathbf{y} \sim f\left(\begin{bmatrix} | \\ i \\ | \end{bmatrix}\right)$$

Regression

$$\mathbf{y} \sim f\left(\begin{bmatrix} | \\ i \\ | \end{bmatrix}\right)$$

$$\mathbf{y} \sim g\left(\left[w_1, \ldots, w_k\right]\right)$$

Regression

$$\mathbf{y} \sim f\left(\begin{bmatrix} | \\ i \\ | \end{bmatrix}\right)$$

$$\mathbf{y} \sim g\left(\left[w_1, \ldots, w_k\right]\right)$$

- Orthogonal: $f(\mathbf{i}) - g(\mathbf{z})$

Regression

$$\mathbf{y} \sim f\left(\begin{bmatrix} | \\ i \\ | \end{bmatrix}\right)$$

$$\mathbf{y} \sim g\left(\left[w_1, \ldots, w_k\right]\right)$$

- Orthogonal: $f(\mathbf{i}) - g(\mathbf{z})$
- Return to the image and interpret the difference by using attention masks.

Results

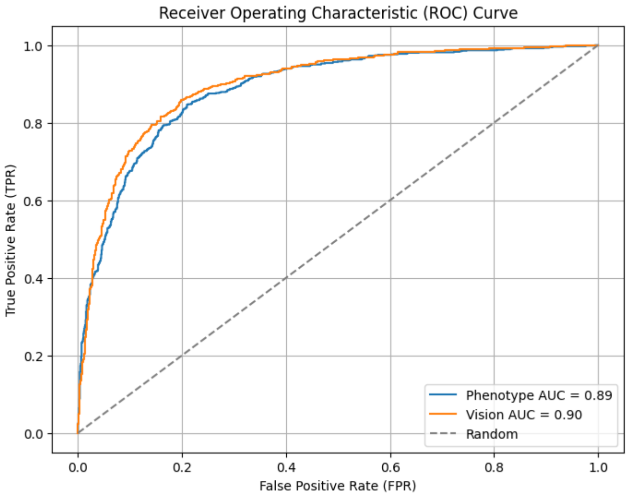


Figure: Healthy/Unhealthy Binary Classification

Conclusion

- We wish to scale this model up for

Conclusion

- We wish to scale this model up for
    - a different response variable

Conclusion

- We wish to scale this model up for
  - a different response variable
  - different radiology modalities

Conclusion

- We wish to scale this model up for
  - a different response variable
  - different radiology modalities
- This project demonstrates

Conclusion

- We wish to scale this model up for
  - a different response variable
  - different radiology modalities
- This project demonstrates
  - how fine-tuning is integral for converting LLM's to the medical domain

Conclusion

- We wish to scale this model up for
  - a different response variable
  - different radiology modalities
- This project demonstrates
  - how fine-tuning is integral for converting LLM's to the medical domain
  - how contrastive learning can be utilized on unstructured data to augment statistical techniques

Acknowledgements

- I want to thank Dr. Junwei Lu, Dr. Tianxi Cai, and Zebin Wang for their guidance and support this summer.
- I also want to thank Dr. Churchill, Jamie, Arya, and all the lecturers for an invaluable summer research experience.

Dr. Junwei Lu

Dr. Tianxi Cai