Residual Embedding Regression

Chase Mathis

March 5, 2024

Overview

1. The Project at a High Level
2. The algorithm
3. Results

Background

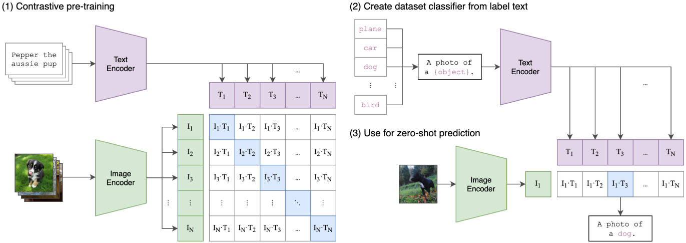

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
T_1 T_2 T_3 ... T_N
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure: OpenAI's CLIP algorithm is a self-supervised image-text learning method [2]

Bioinformatic Applications of CLIP: MedCLIP

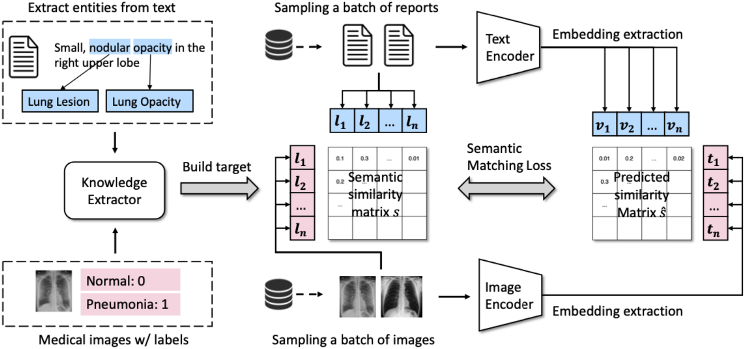


Figure: The MedCLIP algorithm [2]

Bioinformatic Applications of CLIP: BioVIL -T

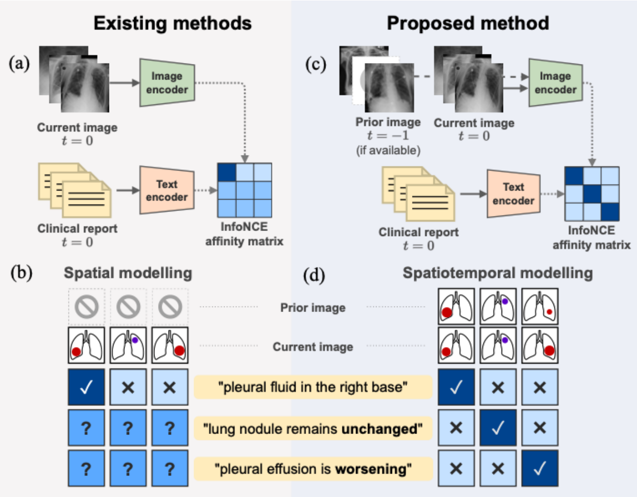


Figure: The BioVIL-T algorithm [1]

Residuals in Statistics

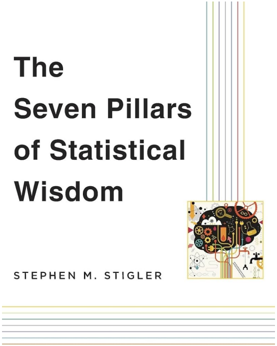
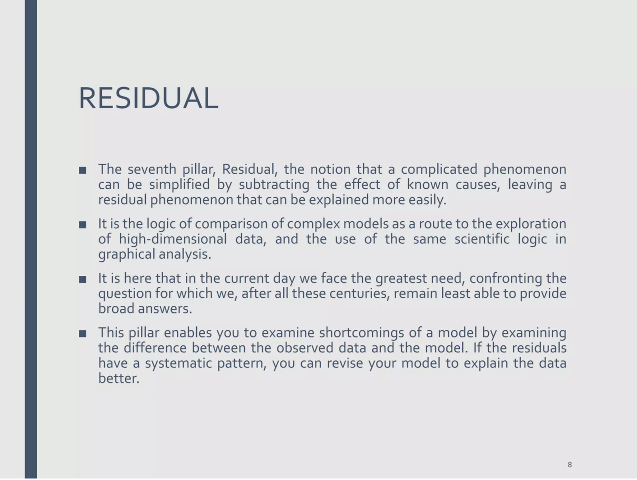


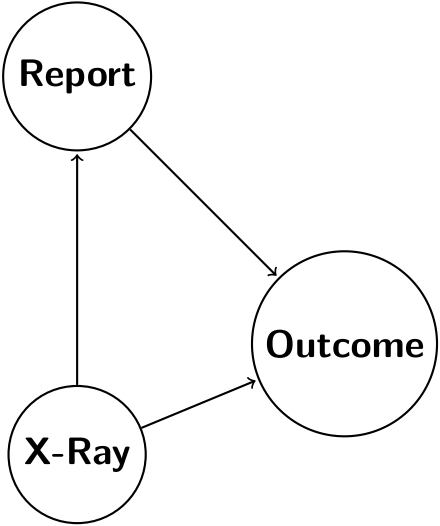
Figure: Here, Statistician Stephen Stigler lists Residuals and their analysis as one of the seven pillars of statistics



RESIDUAL

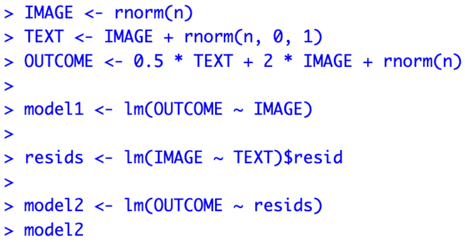
- The seventh pillar, Residual, the notion that a complicated phenomenon can be simplified by subtracting the effect of known causes, leaving a residual phenomenon that can be explained more easily.
- It is the logic of comparison of complex models as a route to the exploration of high-dimensional data, and the use of the same scientific logic in graphical analysis.
- It is here that in the current day we face the greatest need, confronting the question for which we, after all these centuries, remain least able to provide broad answers.
- This pillar enables you to examine shortcomings of a model by examining the difference between the observed data and the model. If the residuals have a systematic pattern, you can revise your model to explain the data better.

Residual for Direct Effect Modeling

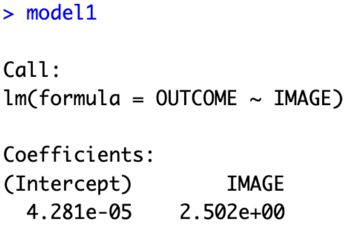


- We hypothesize that the $X - ray$ impacts the future outcome, independently of the radiology report
- Clearly, the X-ray affects the Radiology Report
- The expert radiologist can predict future status

Numeric Simulations



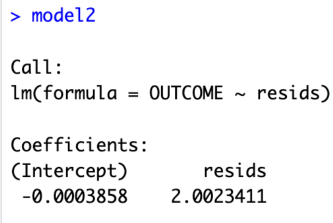
```
> IMAGE <- rnorm(n)
> TEXT <- IMAGE + rnorm(n, 0, 1)
> OUTCOME <- 0.5 * TEXT + 2 * IMAGE + rnorm(n)
>
> model1 <- lm(OUTCOME ~ IMAGE)
>
> resids <- lm(IMAGE ~ TEXT)$resid
>
> model2 <- lm(OUTCOME ~ resids)
> model2
```



```
> model1

Call:
lm(formula = OUTCOME ~ IMAGE)

Coefficients:
(Intercept)        IMAGE
  4.281e-05    2.502e+00
```



```
> model2

Call:
lm(formula = OUTCOME ~ resids)

Coefficients:
(Intercept)       resids
 -0.0003858    2.0023411
```

Motivation Summary

Motivation

Ultimately, we believe that there is information in the image *not* in the text. We want to model only this information.

Setting up the Algorithm

- Use pre-trained image encoder and text encoder
- We experimented with MedCLIP [3] and Bio-VIL-T [1]
- Ultimately, MedCLIP's algorithm focused on zero-shot prediction, whereas Bio-VIL-T focused on inference.
- Define image-encoder f : Image $\mapsto \mathbb{R}^e$
- Define text-encoder g : Tokens $\mapsto \mathbb{R}^e$
- Define our connector-model $h : \mathbb{R}^e \mapsto \mathbb{R}^e$
- Define our image embeddings, text embeddings as $\mathbf{i}, \mathbf{w}$

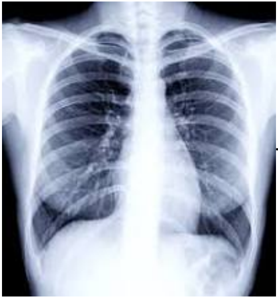
Then, we train h such that

$$\min \|h(\mathbf{w}) - \mathbf{i}\|$$

And after training h, define $\mathbf{r}$ as

$$\mathbf{r} = \hat{h}(\mathbf{w}) - \mathbf{i}$$

Algorithm Diagram



$$\begin{bmatrix} | \\ \mathbf{i} \\ | \end{bmatrix} - \begin{bmatrix} | \\ \hat{h}(\mathbf{w}) \\ | \end{bmatrix} = \begin{bmatrix} | \\ \mathbf{r} \\ | \end{bmatrix}$$

There is new disease on the right $\longrightarrow \begin{bmatrix} | \\ \mathbf{w} \\ | \end{bmatrix}$

Residual Regression

Given a small dataset of future progression based off Mimic-CXR [1], we see if the residuals can predict future outcome.
Train a classifier neural network model

$$c : \mathbb{R}^e \mapsto \{0, 1, 2\}$$

such that we can predict the outcome of a patient (worsening, stable, improving) from their previous residual information. [1]

[1] Because of the class imbalance, I used cross-entropy loss with specific weights based off prior class imbalance

Residual Heatmap

- We can analyze the residual heatmap with cosine similarity
- Further, given our trained classifier model c, we can generate heatmaps for interpretability and inference

Patch Embeddings vs. Image Embeddings

- I've been careless with the word "embedding"
- However, "image embeddings" and "patch embeddings" are not always the same
    - **Image Embeddings:** Represent the entire image as a high-dimensional vector. These embeddings capture global features of an image and are typically used in image classification, object detection, and more.
    - **Patch Embeddings:** Break down the image into smaller segments or patches and then represent each patch as a vector. These embeddings are useful for tasks requiring fine-grained analysis, such as identifying specific objects within an image or understanding texture and patterns.

Patch Embeddings vs. Image Embeddings

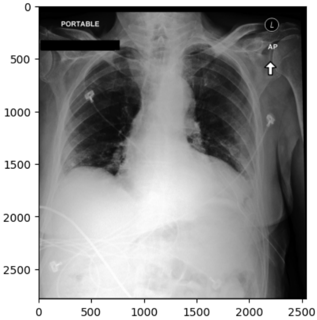


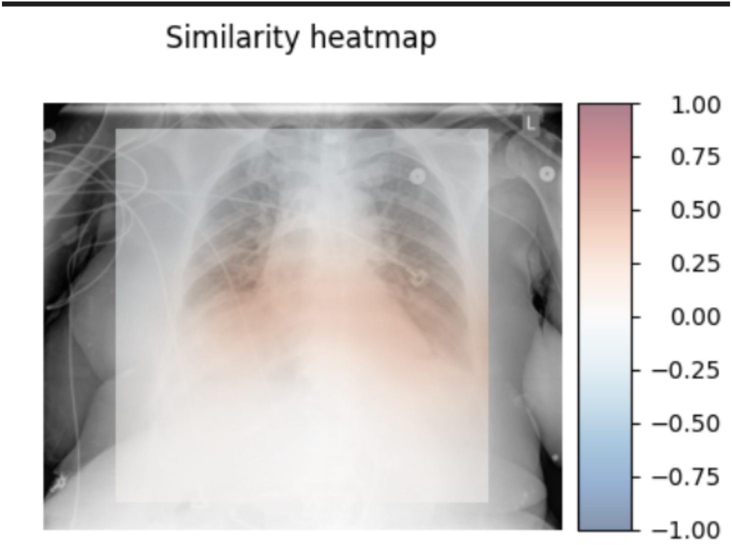
Figure: “The patient has Consolidation on their left lung”

- MedCLIP’s algorithm used image and text embeddings for zero-shot prediction [3]
- The image has an embedding **i** and the text has an embedding **w**.
- We can predict the disease of a patient by finding which disease has the highest cosine similarity

How does BioVil-T Generate Heatmaps?

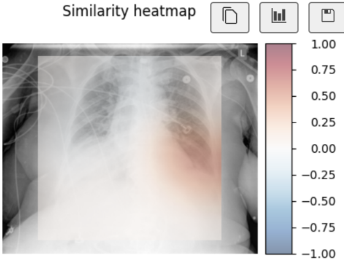
Patch Embedding Heatmap Video [click]

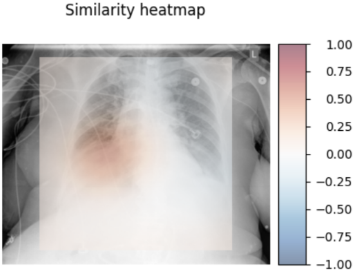
Bio-Vil Caption Results: Heart Location



Bio-Vil Caption Results: Left vs Right Location

Disease is seen on the left, right respectively





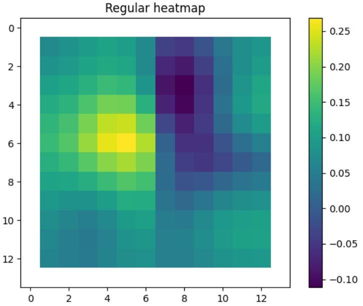
Residual Regression Results

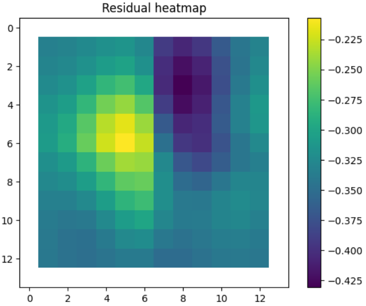
Research Question

Can a residual of the image embedding and phenotype $\hat{f}(\mathbf{w}) - \mathbf{i}$ produce the same heatmap as the full image embedding and phenotype?

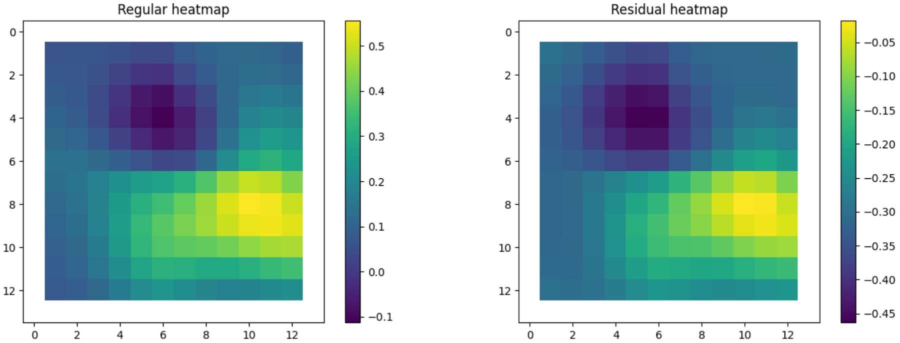
Heatmap Results

The answer is yes!

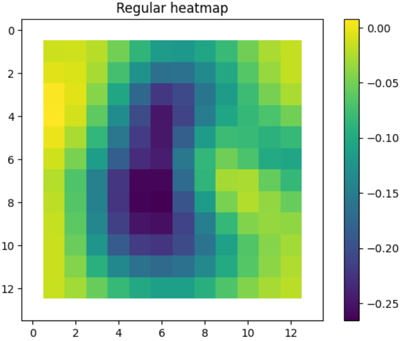


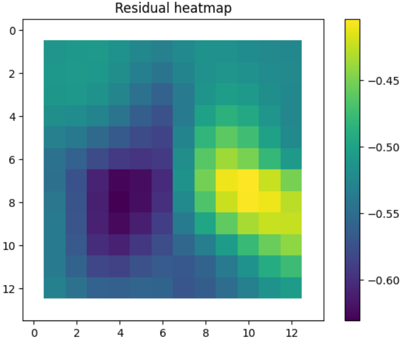


Heatmap Results 2



Heatmap Results 3





The results are sometimes not perfect.

Using Expectation to generate future Heatmap

- Recall our classifier c which maps to an ordinal response variable $\{0, 1, 2\}$
- We train the model to classify using mean pooled image embeddings
- To generate heatmaps, we can run the patch embedding through c and take the expectation or weighted average
- For instance if $c([0.1, 0.3, \dots, 1.5, -0.5]) = [0.2, 0.3, 0.5]$, we would return $1 * 0.3 + 2 * 0.5 = 1.3$

Future Prediction Accuracy

- I achieved high accuracy in predicting future outcomes using only residuals with a classifier Neural Network!
- Now, we leverage the trained model to interpret heatmaps from patch embeddings, providing detailed insights beyond mean-pooled training data.

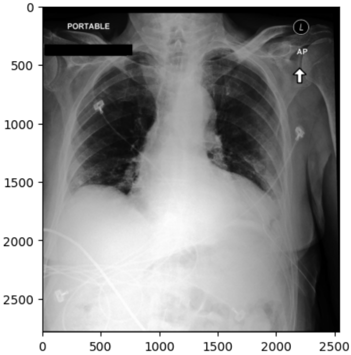


Figure: Old Image

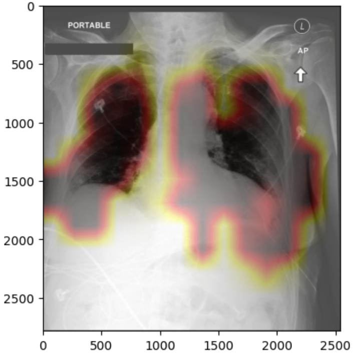


Figure: Same image with Heatmap

Notice how the left side (patient's perspective) is of more concern than the right side. This is interesting because of the future report made by the radiologist indicates:

"...A left lower lobe consolidation a new and although might be secondary to pleural effusion that appears to be increased in the interim, infection is a possibility..."

Discussion

- This algorithm generalizes to find nontrivial things such as disease progression.
- Some future work includes comparing residual embedding results with the full image embedding
- In addition, there are datasets with bounding boxes, so we plan to find the accuracy of heatmaps with that.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay.
Learning to exploit temporal structure for biomedical vision-language processing, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever.
Learning transferable visual models from natural language supervision, 2021.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun.
Medclip: Contrastive learning from unpaired medical images and text, 2022.