TextCAVs:
Debugging Vision Models Using Text

iMIMIC @MICCAI 2024

Angus Nicolson, Yarin Gal, Alison Noble

Interpretability

- *The ability to explain or present in terms understandable to a human*

- A valid aim, but as a goal in of itself, this is difficult to optimise or to measure
- Instead, let's measure how useful interpretability tools are at performing specific tasks, e.g.
  - Improve user trust
  - Improve user performance
  - Debug a model
  - Discover harmful biases

Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv 2017

Interpretability

- *The ability to explain or present in terms understandable to a human*

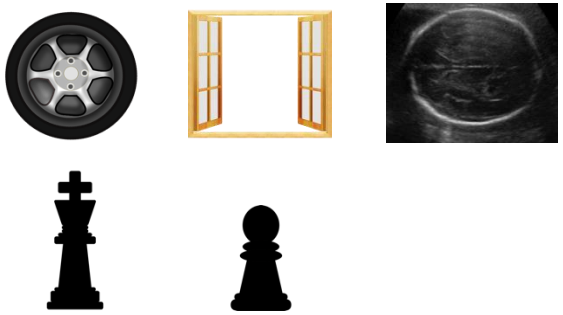
- A valid aim, but as a goal in of itself, this is difficult to optimise or to measure
- Instead, let's measure how useful interpretability tools are at performing specific tasks, e.g.
  - Improve user trust
  - Improve user performance
  - **Debug a model**
  - **Discover harmful biases**

Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. arXiv 2017

Testing with Concept Activation Vectors (TCAV)

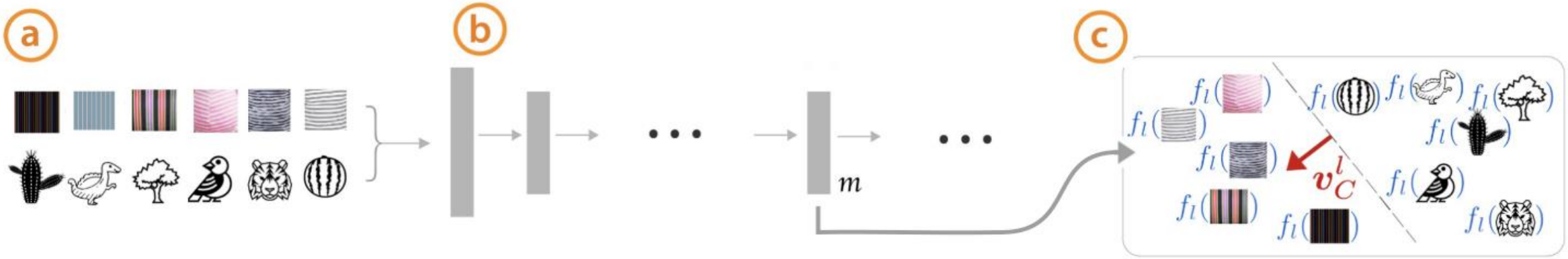
- Explains deep learning models in terms of concepts
- Concepts can be of a variety of different types and are defined by probe datasets of concept examples

  - Textures
    - Striped, spotty, banded
  - Colour
    - Red, blue, bright, dark
  - Basic shapes
    - Circles, squares, triangles
  - Context dependent objects
    - Wheels, windows, the shape of the sylvian fissure
  - Other
    - Chess tactics (e.g. king safety, passed pawns)



T. McGrath, et. al. Acquisition of Chess Knowledge in AlphaZero. *PNAS*, 2022.
B. Kim, et. al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Concept activation vectors (CAVs)



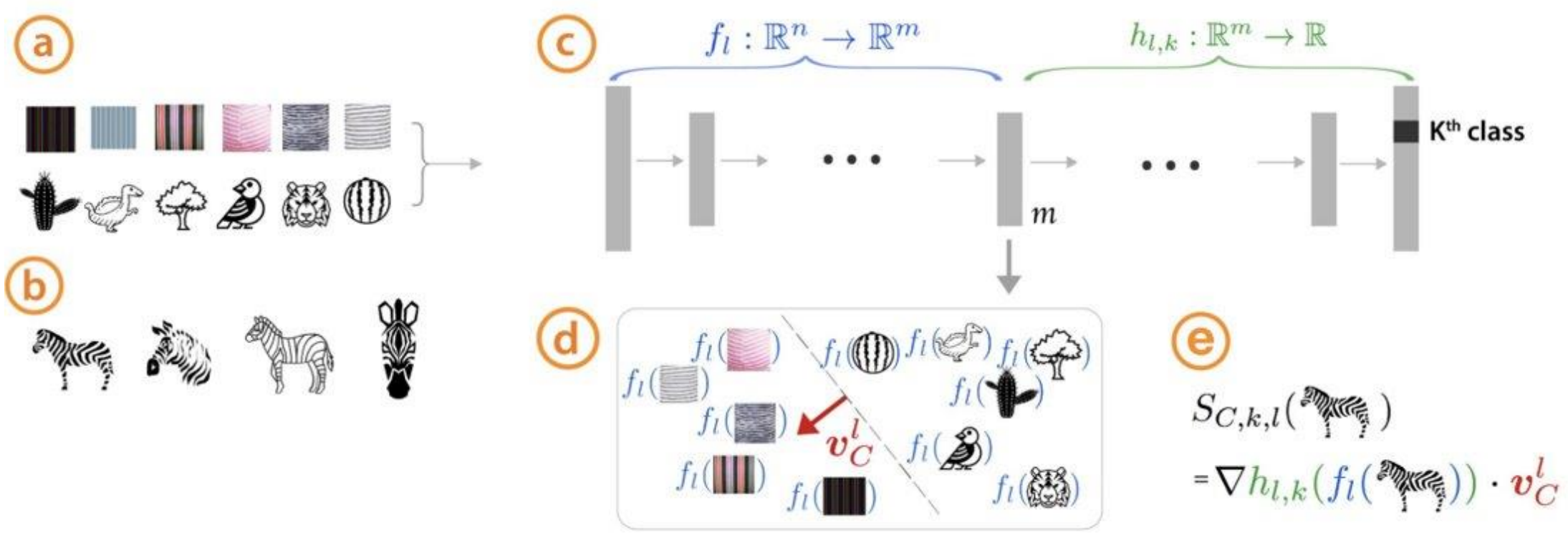
A) Example images representing a concept

B) Activations extracted from sub-layer of the network

C) Train a linear classifier with your concept/random images as the two classes. CAV is the vector orthogonal to the decision boundary (in activation space).

B. Kim, et. al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

Testing with CAVs (TCAV)



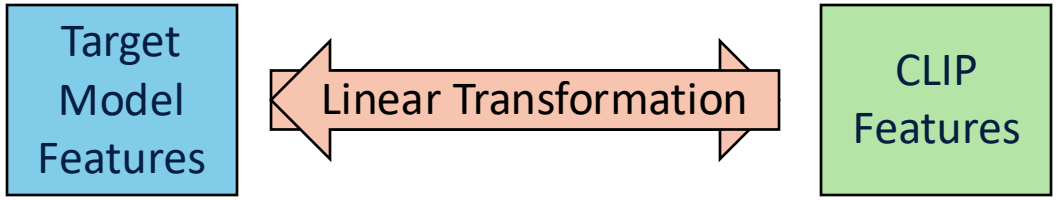
A) Example images representing a concept
B) Example images of target class
C) Activations extracted from sub-layer of the network

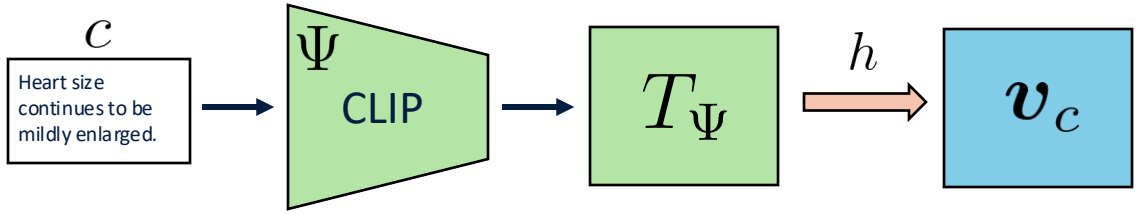
D) Train a linear classifier. CAV is the vector orthogonal to the decision boundary
E) Directional derivative

B. Kim, et. al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

TextCAVs

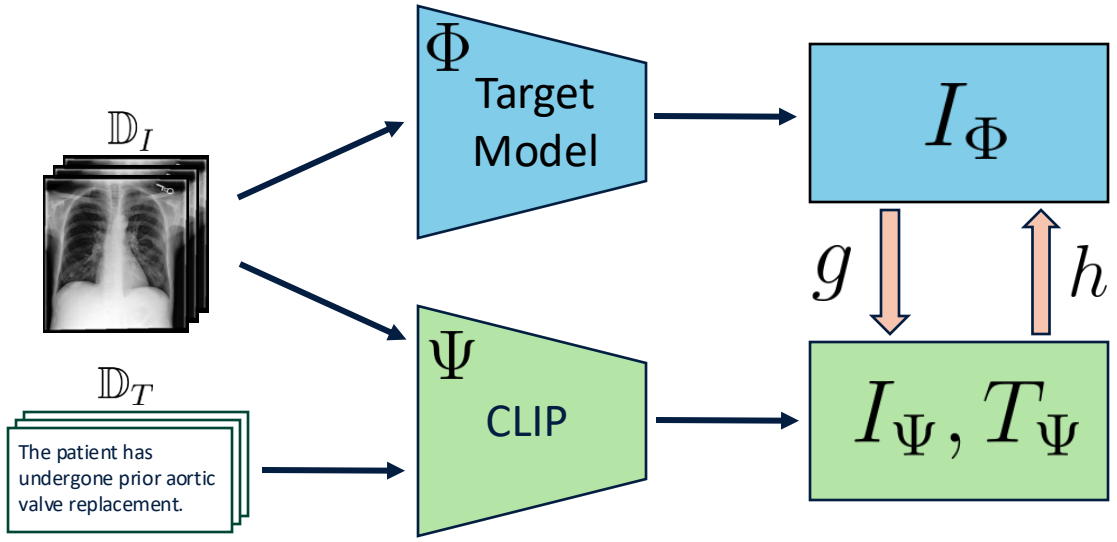
- We build on zero-shot classification methods which convert the features of a target model to the features of a vision-language model like CLIP
- But if we reverse the linear layer…
  - Map text encodings from CLIP → target model feature space
  - Use these features as CAVs
  - A single forward pass to create a new CAV





Yuksekgonul M, Wang M, Zou J. Post-hoc Concept Bottleneck Models. arXiv; 2023
Moayeri M, Rezaei K, Sanjabi M, Feizi S. Text2Concept: Concept Activation Vectors Directly From Text. CVPR 2023
Shipard J, Wiliem A, Thanh KN, Xiang W, Fookes C. Zoom-shot: Fast and Efficient Unsupervised Zero-Shot Transfer of CLIP to Vision Encoders with Multimodal Loss. arXiv; 2024

TextCAVs – Training h



Reconstruction Loss

$$\mathcal{L}_{mse} = ||h(I_\Psi) - I_\Phi||^2 + ||g(I_\Phi) - I_\Psi||^2$$

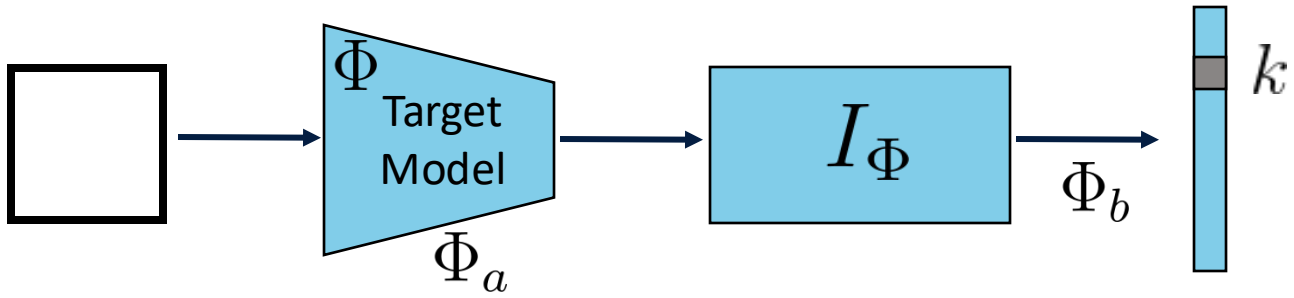
Cycle Loss

$$\mathcal{L}_{cyc} = ||h(g(I_\Phi)) - I_\Phi|| + ||g(h(I_\Psi)) - I_\Psi|| + ||g(h(T_\Psi)) - T_\Psi||$$

TextCAVs - Explanations

- We can measure the sensitivity of the model to a concept by calculating the directional derivative
  - i.e., the similarity between A CAV, $\boldsymbol{v}_c$, and the gradient of the logit output with respect to the activations, $\nabla \Phi_{b,k}$.
- If Φ_b is linear, then its gradient does not depend on the activations, so we can calculate the directional derivative without any images

$$S_{c,k}(\boldsymbol{x}) \propto \nabla \Phi_{b,k}(\Phi_{a,k}(\boldsymbol{x})) \cdot \boldsymbol{v}_c$$



B. Kim, et. al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

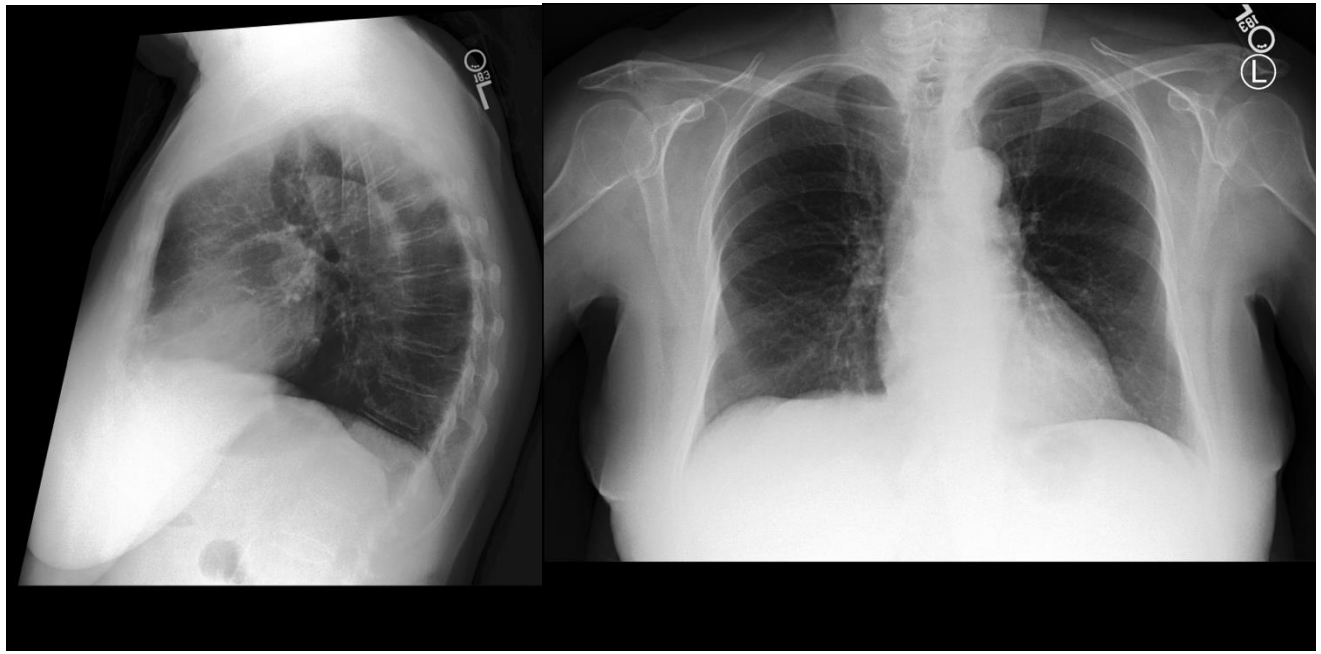
ImageNet

- A quick sense-check
  - Do the explanations look plausible for ImageNet?
- We used a LLM to generate a list of concepts and then ranked the concepts by directional derivative

bullfrog	**albatross**	**orangutan**	**bucket**	**cellphone**
american bullfrog	gannet	orangutan	crab buckets	mp3 player
green frog	seagull	howler monkey	diaper pail	phone
boreal toad	sea eagle	macaque	bucket	phone case
western toad	shearwater	tarsier	laundry basket	memory card
frog	gull	great ape	watering can	walkman
musk turtle	white-tailed eagle	long-nosed monkey	flower pot	cordless phone
snapping turtle	petrel	gibbon	cooking pot	bluetooth
toad	merganser	gorilla	dustbin	smartwatch
terrapin turtle	wading bird	langur	fishing basket	card reader

MIMIC-CXR

- We demonstrate that TextCAVs can be used in a medical domain
- MIMIC-CXR: Chest X-Rays with associated clinical reports
- 14 different classes, each assigned based off the clinical reports



FINAL REPORT

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___F with chest pain // ?pna

COMPARISON: ___.

FINDINGS:

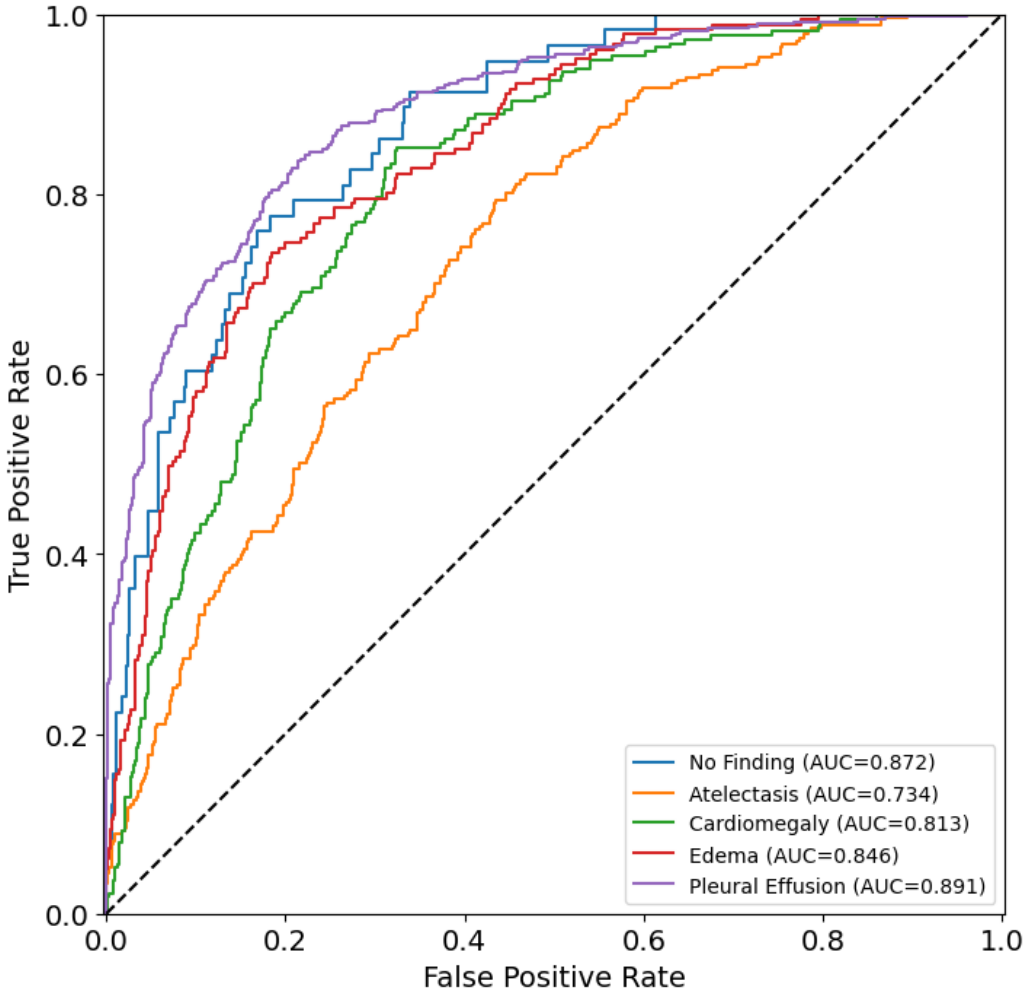
PA and lateral views of the chest provided. Lung volumes are somewhat low. There is no focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen.

IMPRESSION:

No acute intrathoracic process.

Model Training

- We trained a ResNet50 on a 5-way multi-label classification task:
  - No Finding
  - Atelectasis (collapsed lung)
  - Cardiomegaly (enlarged heart)
  - Edema (fluid in the lungs)
  - Pleural effusion (fluid between the lungs and chest wall)
- Mean AUC: 0.831
- Mean Acc: 81.7 %



Which concepts?

- Extract each sentence from the “FINDINGS” and “IMPRESSION”
- Use a random subset of 5000 sentences to obtain a wide variety
- In future work we’d like to use a handcrafted list/interactive session with a radiologist

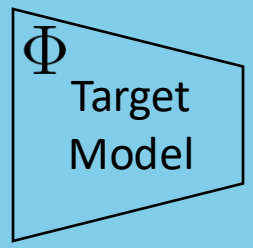
```
                                FINAL REPORT
EXAMINATION:  CHEST (PORTABLE AP)

INDICATION:  ___ year old woman with CNS lymphoma,  // assess for pleural
effusion prior to giving methotrexate

COMPARISON:  ___ at 15:58

FINDINGS:

Again seen is the indwelling right-sided catheter, with tip over distal SVC.
In addition, there is a new right-sided PICC line, with tip overlying the
right atrium.  No pneumothorax detected.

Inspiratory volumes are low and the right hemidiaphragm remains elevated, with
opacity at the right base, similar to prior.  Minimal patchy opacity in the
retrocardiac region is improved slightly.  No gross effusion is detected on
this AP view.  No definite change in the cardiomediastinal silhouette.

Focal opacity the left upper zone represent artifact due to overlap of the
first anterior and fifth posterior left ribs.

IMPRESSION:

No gross effusion detected on either side, but smaller posterior effusions
would not be apparent on this film.  If clinically indicated, a lateral view
could help for further assessment of posterior fusions.

Continued opacity at the right lung base, similar prior.  This is new compared
with ___, but similar the most recent prior study.  This most
likely represents atelectasis, but amount in appropriate clinical setting, an
infectious consolidation could have similar appearance.

Mild patchy opacity at the left base is improved compared with ___.

New right PICC line tip overlies the right atrium.  Clinical correlation is
requested.
```

Example explanations

- Some are clearly linked to the class
  - "The lungs are clear"
  - "increasing atelectasis"
  - "Heart size continues to be mildly enlarged"
- Others are not
  - "There is a fracture of the upper most sternal wire, unchanged."
  - "Nasogastric tube extends below the hemidiaphragm and out of view"

No Finding	Atelectasis	Cardiomegaly
The lungs are clear and the cardiac, mediastinal, and hilar contours are normal.	Nasogastric tube extends below the hemidiaphragm and out of view.	Marked cardiac enlargement as before and unchanged position of previously described metallic prosthesis of porcine type.
Normal chest radiograph with unremarkable appearance of the lung parenchyma and normal appearance of the heart and the mediastinal and hilar contours.	Interval placement of a basilar right sided pleural space pigtail catheter with improved small right pleural effusion and right medial lung base atelectasis.	Heart size continues to be mildly enlarged.
The trachea is slightly deviated to the right by the aortic knob, which is ill-defined.	Worsening of the left retrocardiac opacity likely secondary to increasing atelectasis and/or effusion.	The patient has undergone prior aortic valve replacement.
This could represent a granuloma or possibly a bone island in the rib itself.	There is persistent elevation of the left hemidiaphragm with evidence of Bochdalek hernia seen at the left lower hemithorax.	Dense retrocardiac opacity which could represent effusion, atelectasis, consolidation or a combination thereof.
There is a fracture of the upper most sternal wire, unchanged.	Stable opacification of the mid and lower right lung consistent with large loculated pleural effusions and adjacent atelectasis.	The heart continues to be enlarged with mild to moderate CHF.

Sources of noise



The target model is sensitive to unexpected concepts

h

Errors in the feature conversion between the CLIP and target models

$\nabla \Phi_{b,k}$

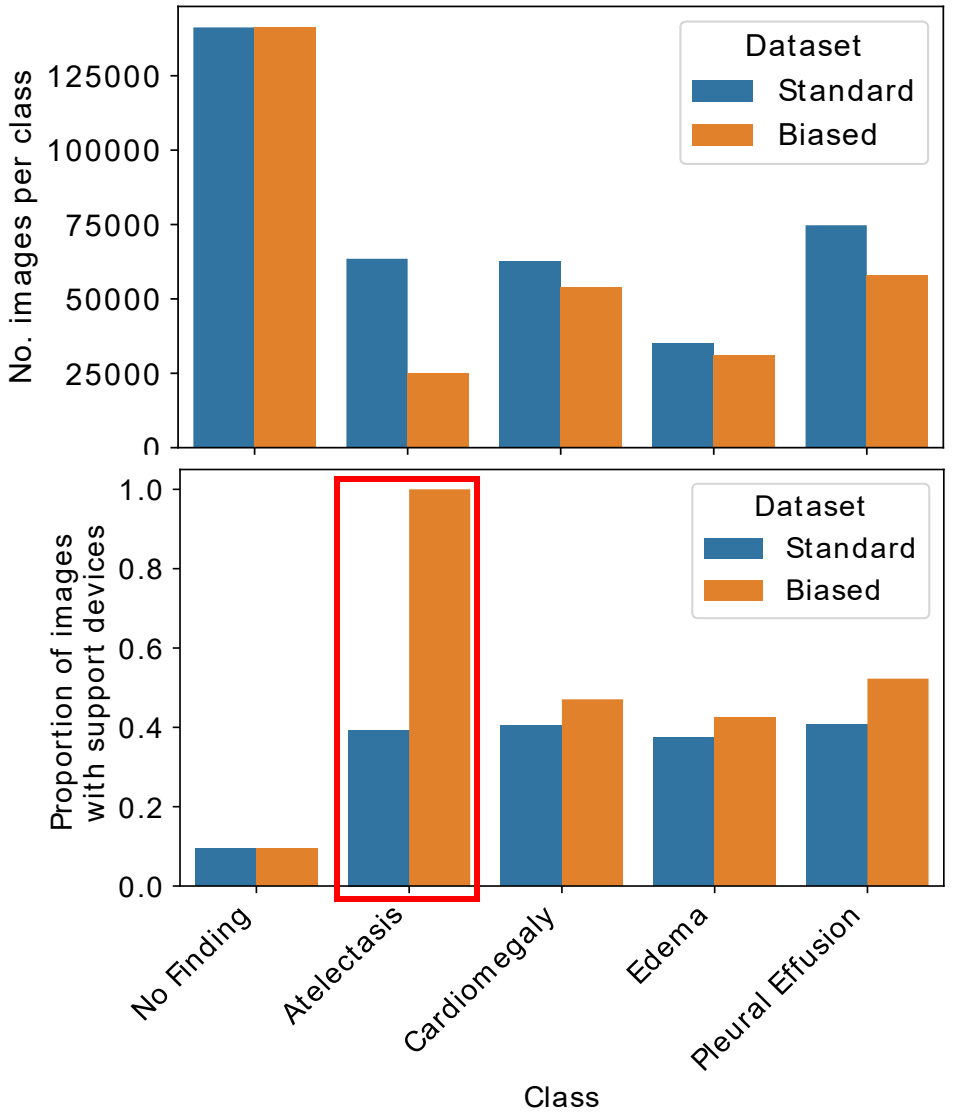
The inherent noise in gradient vectors

This is what we want to measure!

D. Smilkov, et al., Smoothgrad: removing noise by adding noise. arXiv:1706.03825 (2017)

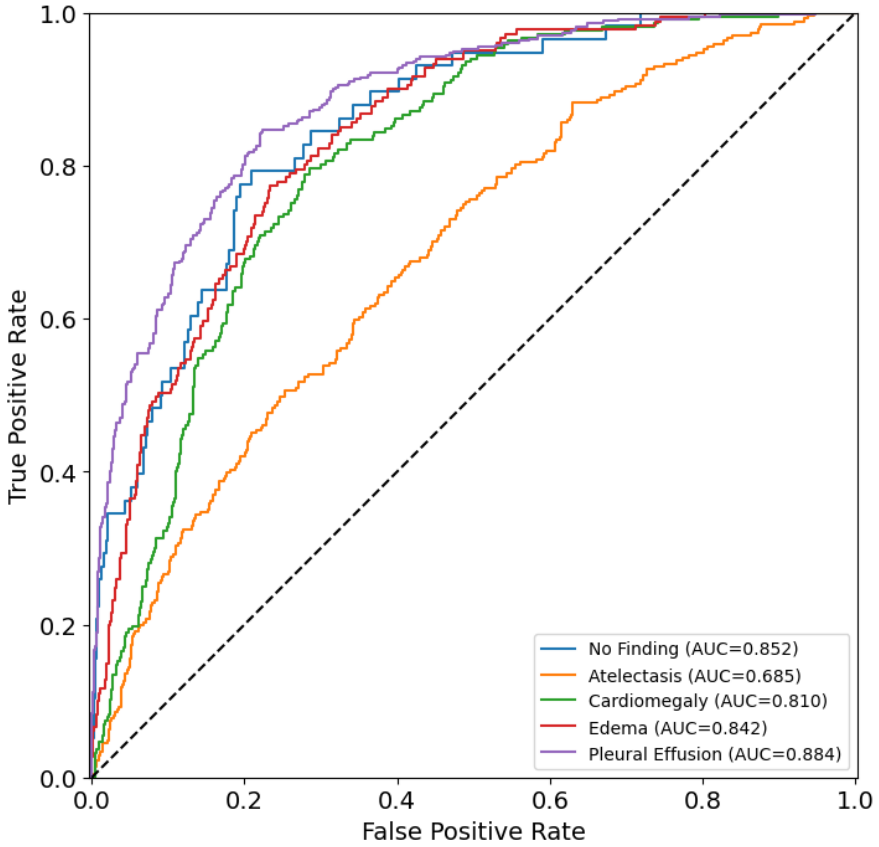
Biased Dataset

- How useful is TextCAVs for model debugging?
- We induced a dataset bias in MIMIC-CXR so that all participants with Atelectasis had a support device
  - (by removing all participants with Atelectasis and no Support Device)

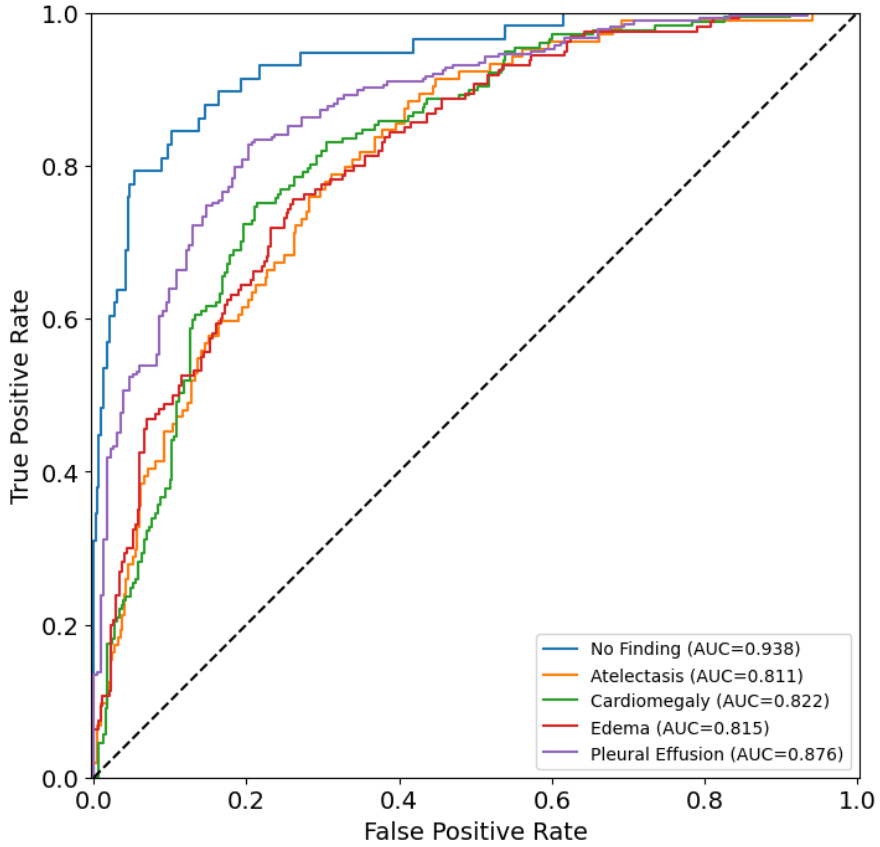


Biased Model

Standard Test Set



Biased Test Set



Biased explanations?

- All the top-5 explanations refer to Support Devices, and none refer to Atelectasis
- A clear change in model explantions, when model behaviour changes
- This indicates that TextCAVs can be used to debug models
- But it's not very quantitative...

No Finding	Atelectasis	Cardiomegaly
Bronchial wall thickening is minimal.	ET and NG tubes positioned appropriately.	If cardiomegaly persists, the presence of a pericardial effusion could be excluded with echocardiography.
Hilar and mediastinal contours are otherwise normal.	ET tube, nasogastric tube, Swan-Ganz catheter, and midline drains are all in standard placements.	Worsening heart failure in the context of chronic atelectasis.
This could represent a granuloma or possibly a bone island in the rib itself.	Nasogastric tube extends below the hemidiaphragm and out of view.	The patient has undergone prior aortic valve replacement.
No discrete solid pulmonary nodule are concerning mass.	Impella LVAD and transvenous atrioventricular pacer leads unchanged in their respective positions.	Moderate-to-severe cardiomegaly and stigmata of previous mitral valve repair noted.
There is a fracture of the upper most sternal wire, unchanged.	Nasogastric tube has been placed that extends well into the stomach.	The heart remains moderately enlarged and the aorta remains unfolded and tortuous.

Concept Relevance Score

- The concept relevance score (CRS) is simply the proportion of the top-N (N=50) TextCAVs that are related to the class
- As an example, a sentence was labelled as related to Edema if:
  - it directly diagnosed the class
    - *Worsening cardiogenic pulmonary edema*
  - or if the class was implied
    - *bilateral parenchymal opacities*
    - *there is alveolar opacity throughout much of the right lung*
- We also labelled the atelectasis TextCAVs on if they referred to support devices
  - Standard: 0.26 (13/50)
  - Biased: 0.88 (44/50)

Model	Standard		Biased		
Metric	AUC	CRS	AUC	AUC*	CRS
No Finding	0.87	0.74	0.85	0.94	0.76
Atelectasis	0.73	0.56	0.68	0.81	0.04
Cardiomegaly	0.81	0.94	0.81	0.82	0.90
Edema	0.85	0.90	0.84	0.81	0.80
Pleural Effusion	0.89	1.00	0.88	0.88	1.00
Mean	0.83	0.83	0.81	0.85	0.70

Summary

- We introduce TextCAVs, an interpretability method that can be used to provide textual explanations for image-based deep learning models
- Once two linear layers have been trained, new explanations can be generated with minimal compute and no imaging data
- We demonstrate that TextCAVs produce reasonable explanations for both natural images and chest X-ray imaging
- We show that TextCAVs can be used to debug models, finding dataset and model biases

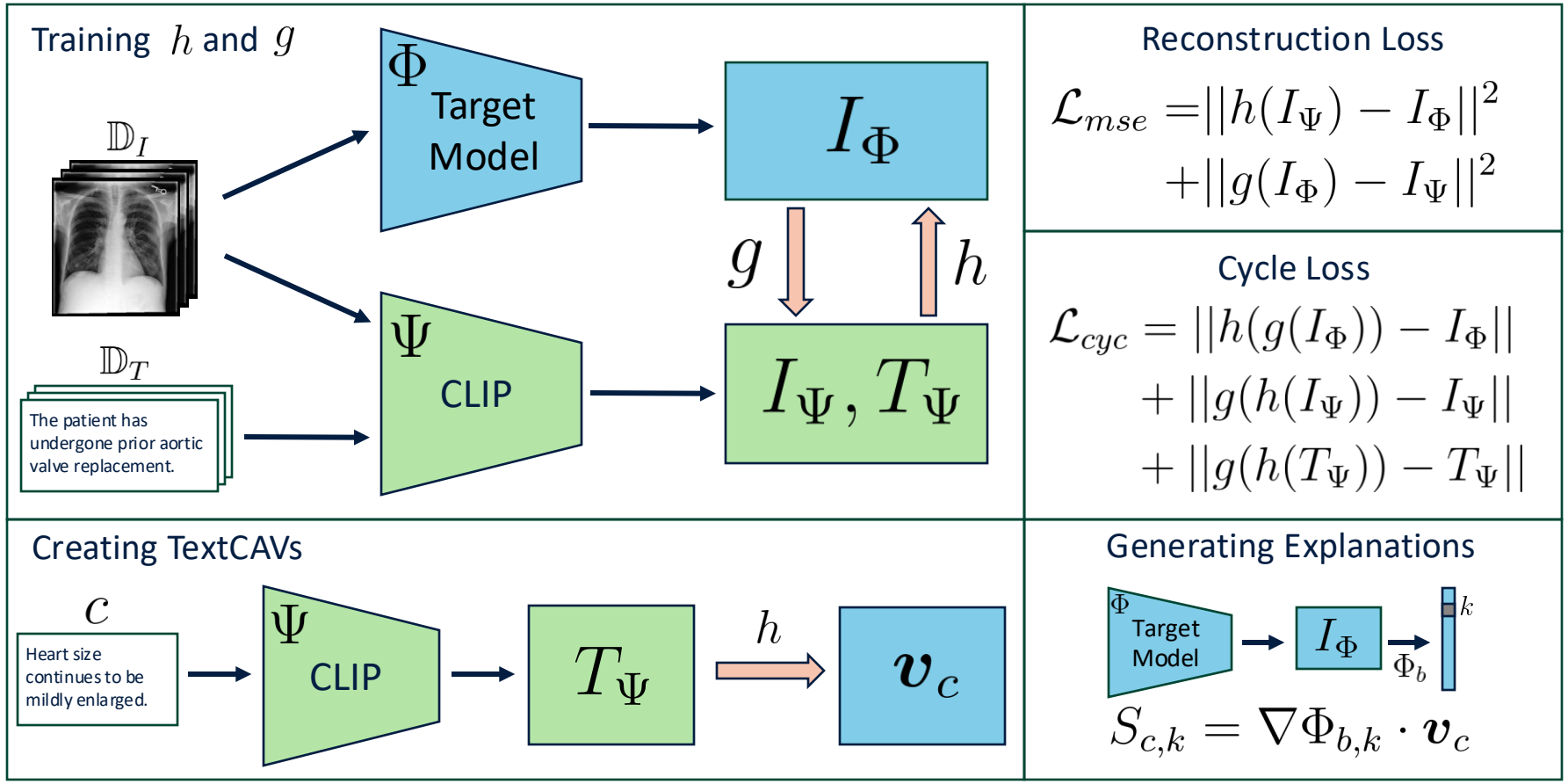
Acknowledgments

Prof. Alison Noble

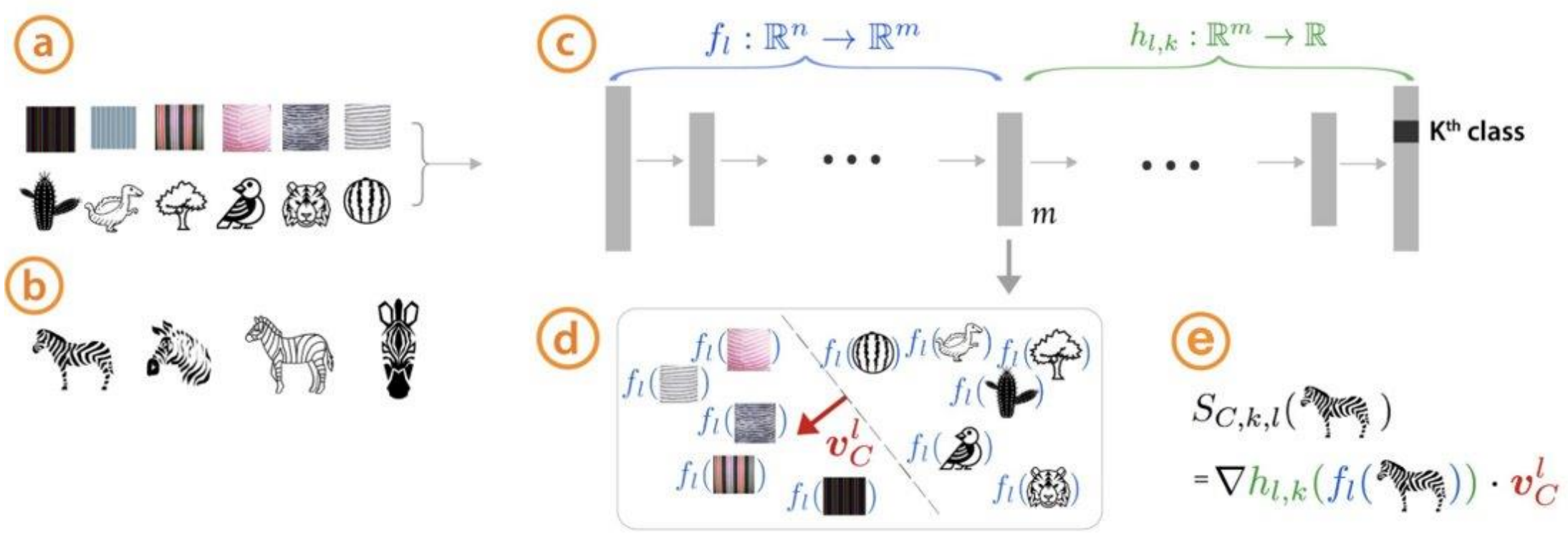
Prof. Yarin Gal

Questions?

TextCAVs


Training h and g
$\mathbb{D}_I$
$\mathbb{D}_T$
The patient has undergone prior aortic valve replacement.
Φ Target Model
Ψ CLIP
I_Φ
I_Ψ, T_Ψ
g
h
Reconstruction Loss
$\mathcal{L}_{mse} = ||h(I_\Psi) - I_\Phi||^2 + ||g(I_\Phi) - I_\Psi||^2$
Cycle Loss
$\mathcal{L}_{cyc} = ||h(g(I_\Phi)) - I_\Phi|| + ||g(h(I_\Psi)) - I_\Psi|| + ||g(h(T_\Psi)) - T_\Psi||$
Creating TextCAVs
c
Heart size continues to be mildly enlarged.
Ψ CLIP
T_Ψ
h
$\boldsymbol{v}_c$
Generating Explanations
Φ Target Model
I_Φ
Φ_b
k
$S_{c,k} = \nabla \Phi_{b,k} \cdot \boldsymbol{v}_c$

Testing with CAVs (TCAV)

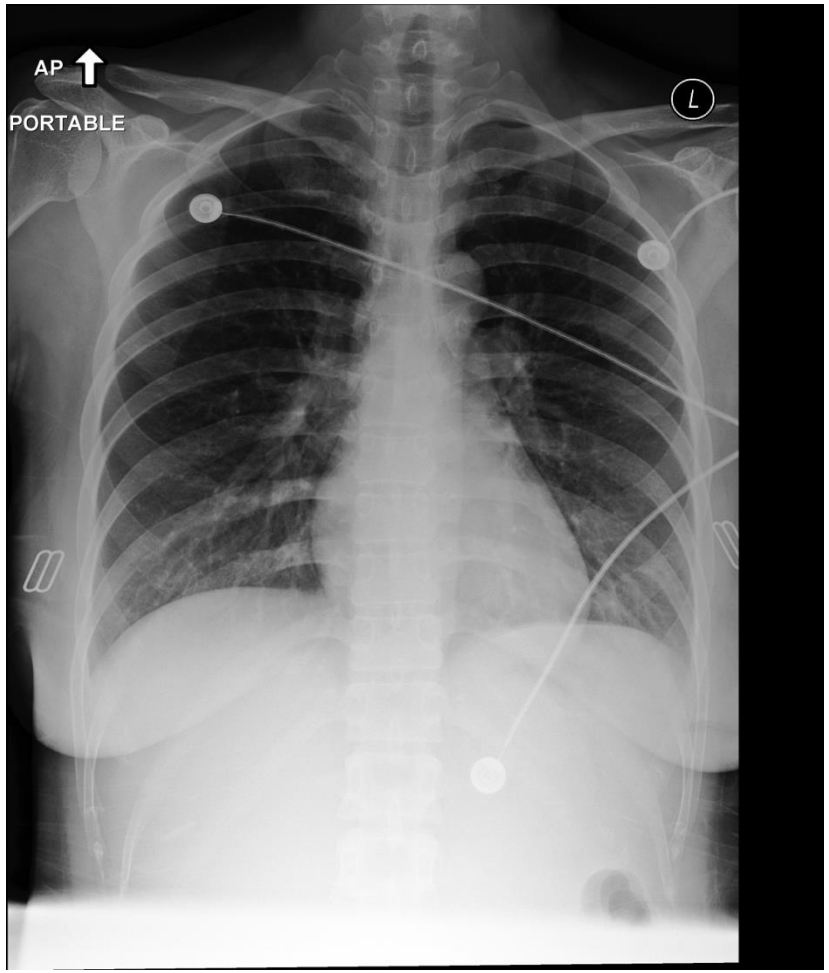


A) Example images representing a concept
B) Example images of target class
C) Activations extracted from sub-layer of the network

D) Train a linear classifier. CAV is the vector orthogonal to the decision boundary
E) Directional derivative

B. Kim, et. al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), ICML, 2018

MIMIC-CXR




```
                                FINAL REPORT
EXAMINATION:  CHEST (PORTABLE AP)

INDICATION:  ___ year old woman with CNS lymphoma,  // assess for pleural
effusion prior to giving methotrexate

COMPARISON:  ___ at 15:58

FINDINGS:

Again seen is the indwelling right-sided catheter, with tip over distal SVC.
In addition, there is a new right-sided PICC line, with tip overlying the
right atrium.  No pneumothorax detected.

Inspiratory volumes are low and the right hemidiaphragm remains elevated, with
opacity at the right base, similar to prior.  Minimal patchy opacity in the
retrocardiac region is improved slightly.  No gross effusion is detected on
this AP view.  No definite change in the cardiomediastinal silhouette.

Focal opacity the left upper zone represent artifact due to overlap of the
first anterior and fifth posterior left ribs.

IMPRESSION:

No gross effusion detected on either side, but smaller posterior effusions
would not be apparent on this film.  If clinically indicated, a lateral view
could help for further assessment of posterior fusions.

Continued opacity at the right lung base, similar prior.  This is new compared
with ___, but similar the most recent prior study.  This most
likely represents atelectasis, but amount in appropriate clinical setting, an
infectious consolidation could have similar appearance.

Mild patchy opacity at the left base is improved compared with ___.

New right PICC line tip overlies the right atrium.  Clinical correlation is
requested.
```