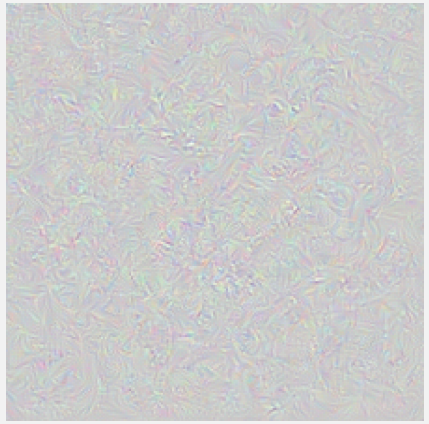
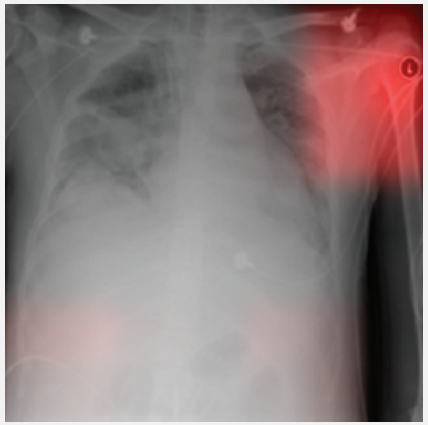
SELF-SUPERVISION ON IMAGES AND TEXT REDUCES RELIANCE ON VISUAL SHORTCUT FEATURES

Anil Palepu, Andrew Beam

FULLY-SUPERVISED DNNS LEARN SHORTCUTS





Article: Super Bowl 50

Paragraph: "*Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.*"

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"

Original Prediction: John Elway

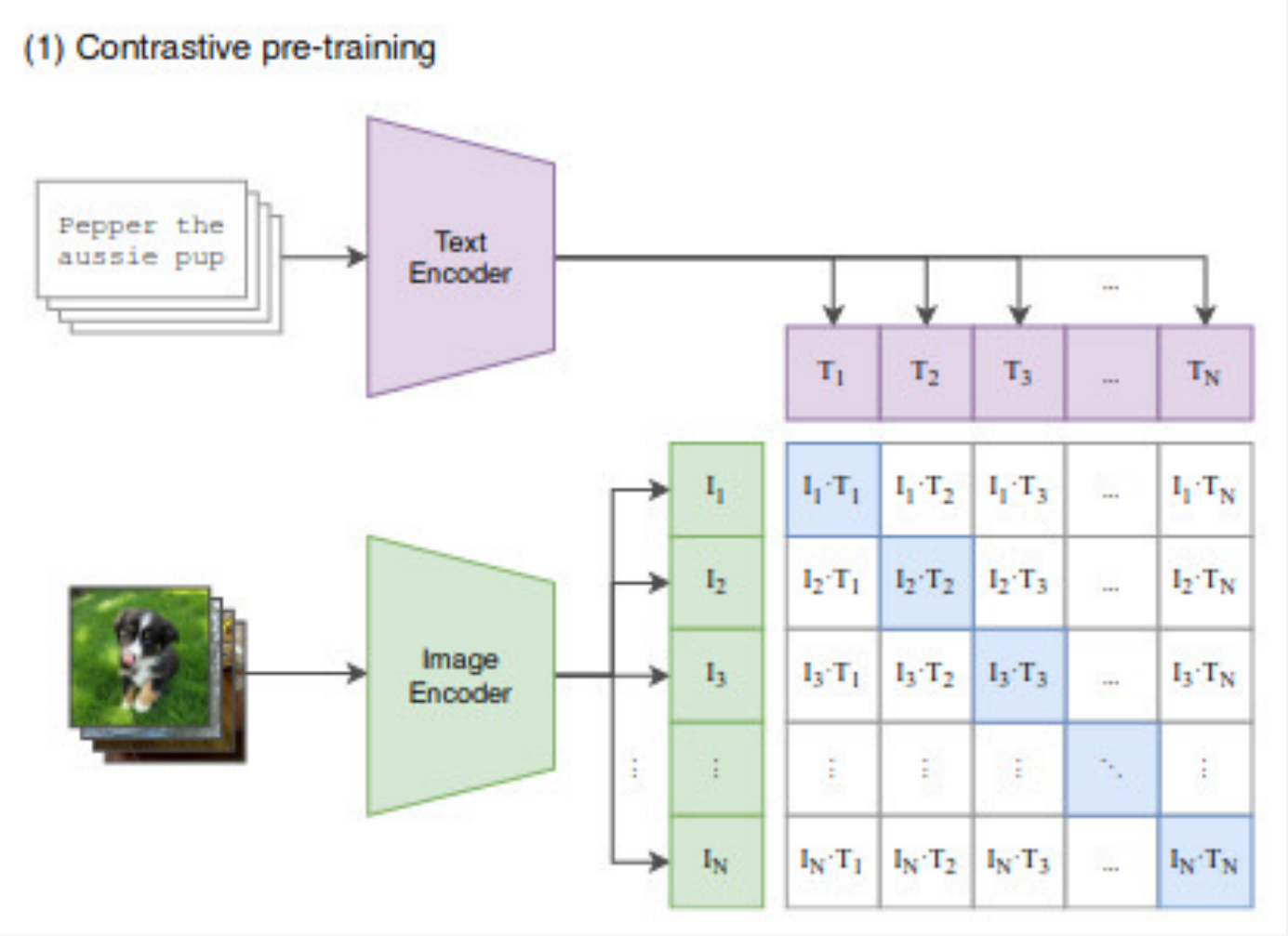
Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

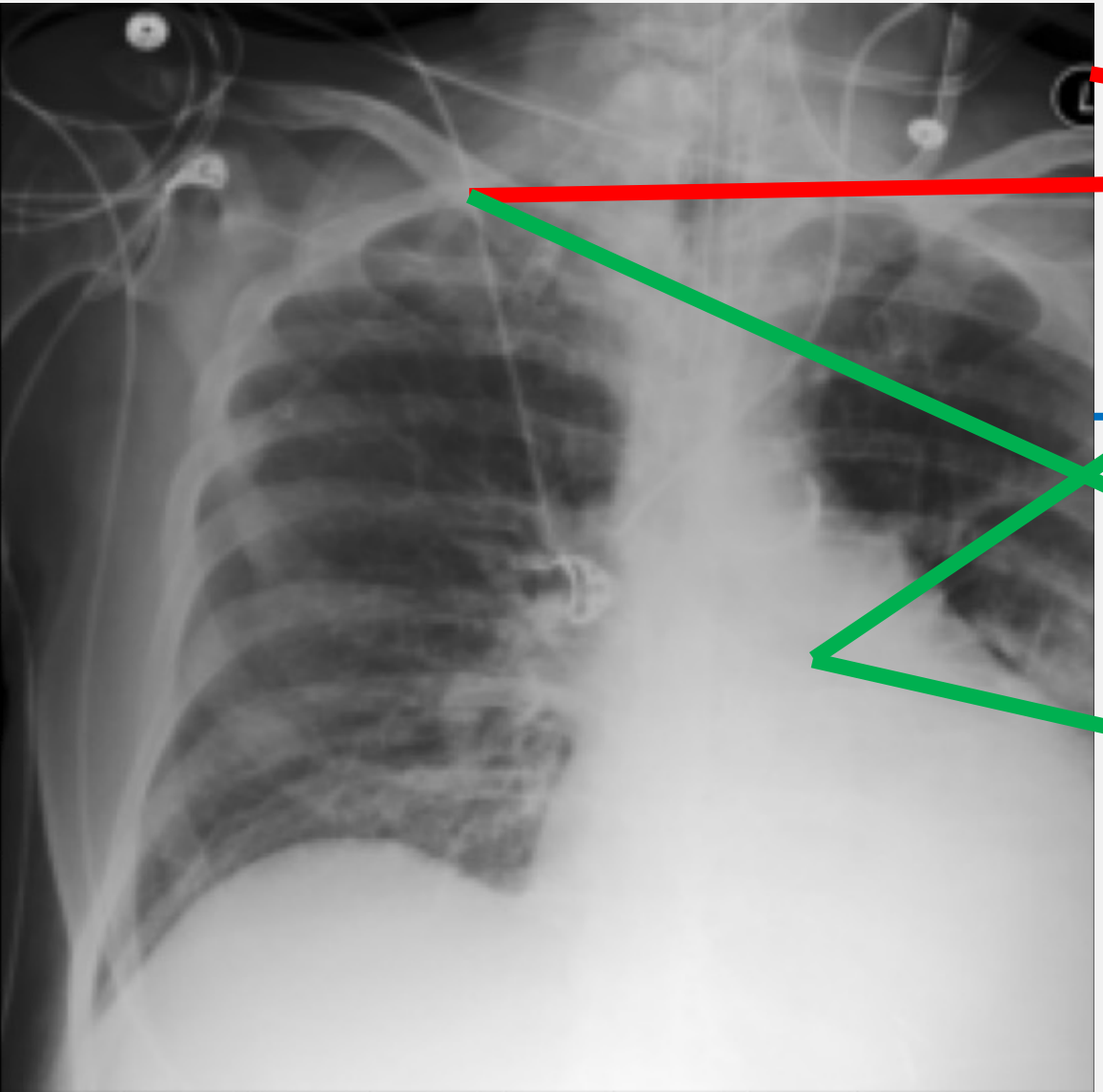
Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11):665–673, 2020

CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

- The CLIP architecture jointly trains a vision and text encoder
- It has been shown to have impressive zero shot performance on a variety of retrieval and classification tasks
- We hypothesize that pretraining with this architecture may reduce shortcut reliance



LABEL-SUPERVISED VS TEXT-SUPERVISED FEATURES



Cardiomegaly

Text concepts (A right subclavian catheter line was inserted)

Text concepts (The cardiac silhouette is enlarged)

Label-Supervised

Shortcut features can be learned if associated with the label of interest in the training data

Text-Supervised

Now incentivizing learning specific & diverse clinical concepts simultaneously.

Hypothesis: True clinical features will be prioritized over shortcuts

DATASETS

- Training/Validation: MIMIC-CXR-JPG
  - Images: Random augmentations
  - Radiology Reports: Findings & Impression
- Fine-tuning/Test: CheXpert

Clinical Labels (Shortcut):

Atelectasis (A)
Cardiomegaly (C)
Consolidation (N)
Edema (E)
Pleural Effusion (P)

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

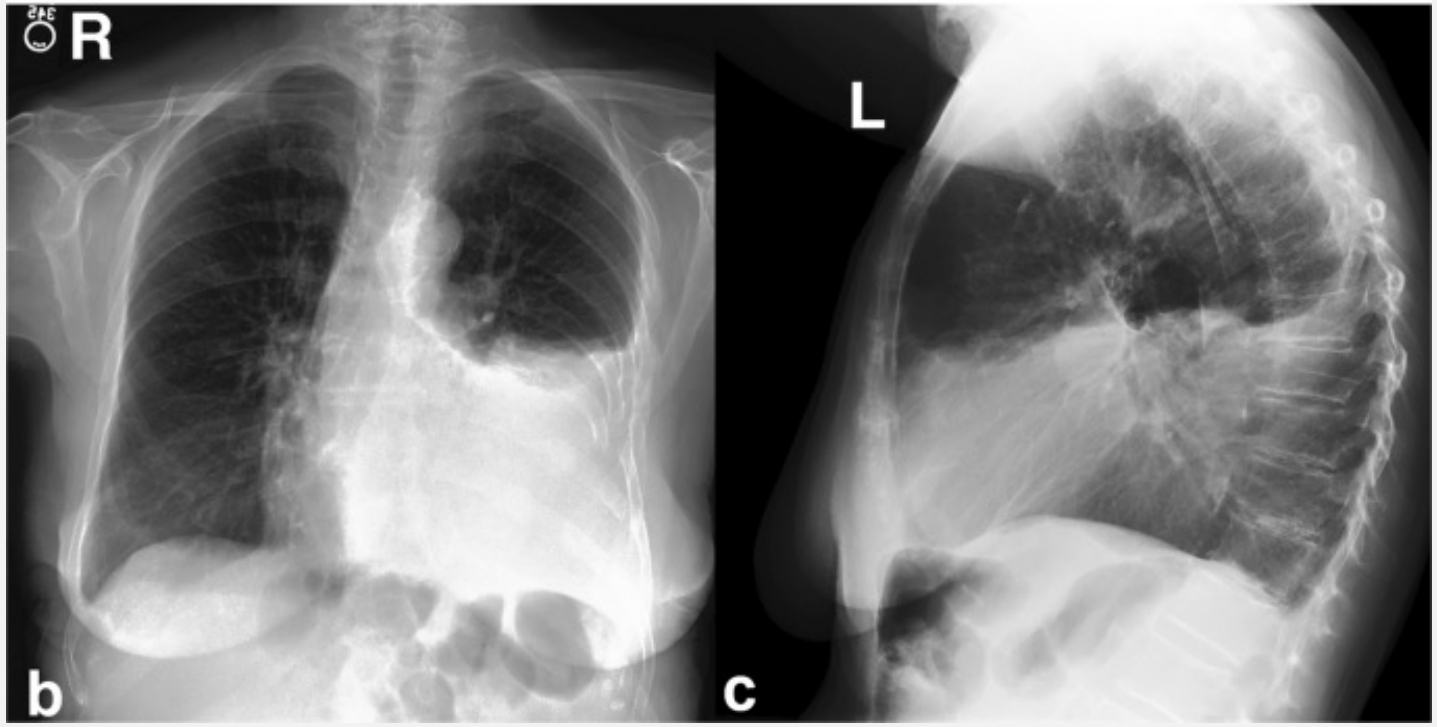
FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion

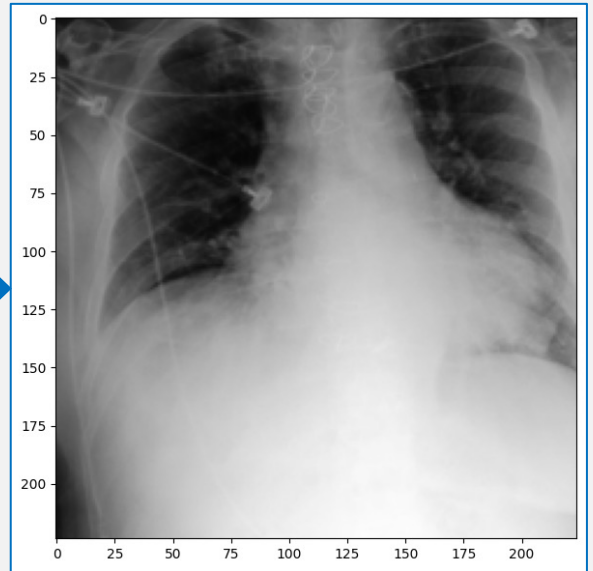
a

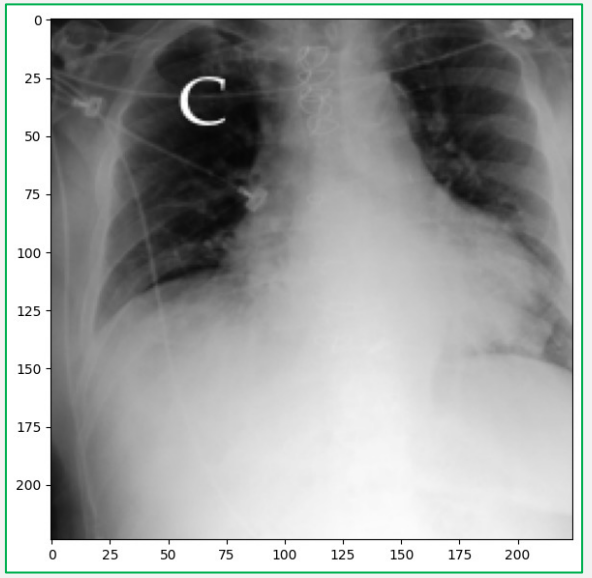


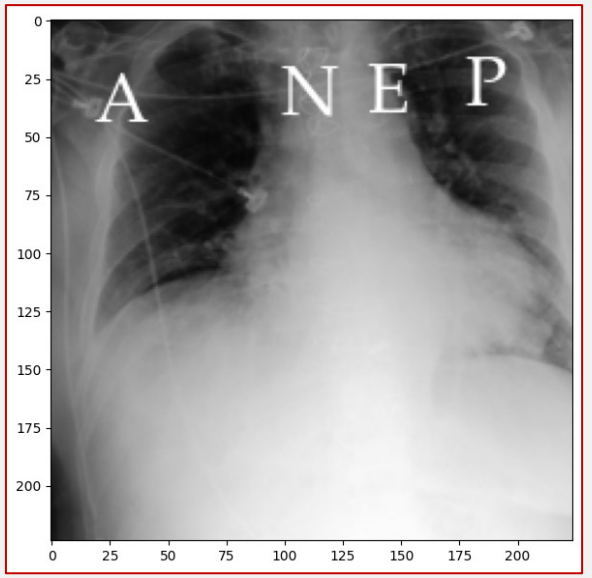
SHORTCUT GENERATION (TRAIN TIME)

Yes (p = 0.9)
Add watermarks?
No (p = 0.1)

Yes (p = 0.9)
Use "correct" shortcuts?
No (p = 0.1)







CXR only positive for Cardiomegaly (C)

MODELS DIFFER IN TRAINING DATA & ARCHITECTURE

1) Real CNN &
2) Shortcut CNN

- ResNet-50 classifier, supervised training with clinical labels

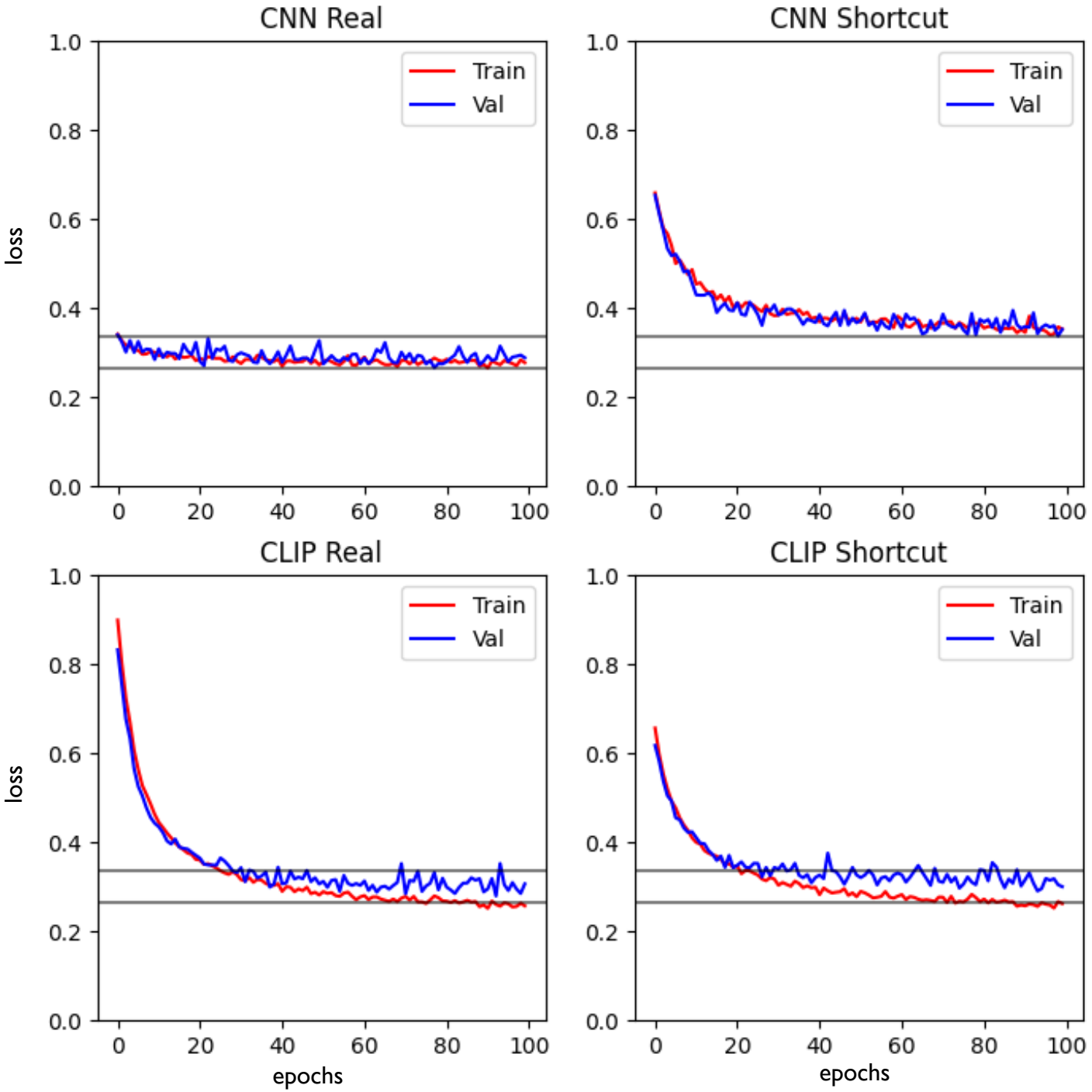
3) Real CLIP &
4) Shortcut CLIP

- Identical ResNet-50 classifier, pretraining with domain-specific BERT and clinical text

-All models were fine-tuned for classification with 1% of real CheXpert training data (no watermarks)
-Then evaluated on various versions of CheXpert test data

FINE-TUNING CURVES SHOW CNN FAILS TO UNLEARN SHORTCUTS

- Fine-tuning data:
  - Train = 1 % CheXpert (2235 CXRs)
  - Val = 0.2% CheXpert (447 CXRs)
  - Same hyperparameters
- Horizontal lines are drawn at the best val loss for the shortcut/real CNN
- **The Shortcut CNN fails to reach a comparable loss in the same time**

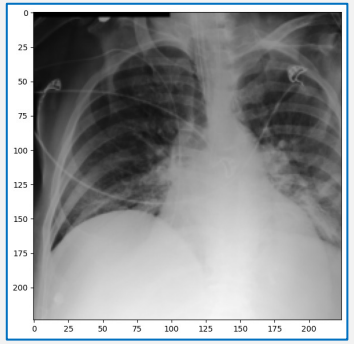


SHORTCUT AND ADVERSARIAL GENERATION (TEST TIME)

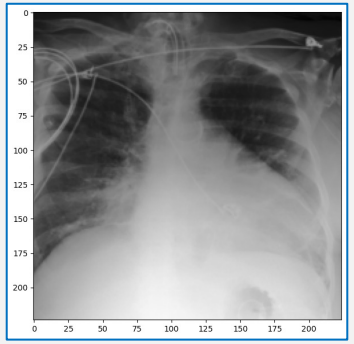
SELECTIVELY ADDING SHORTCUTS TO AID OR CONFUSE THE MODELS

No watermarks

Atelectasis+



Atelectasis-

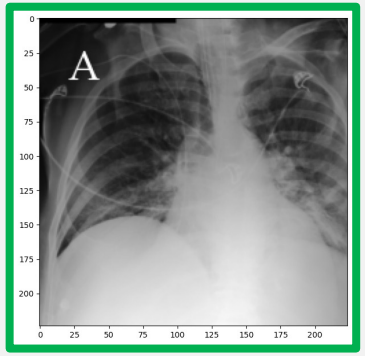


1 version

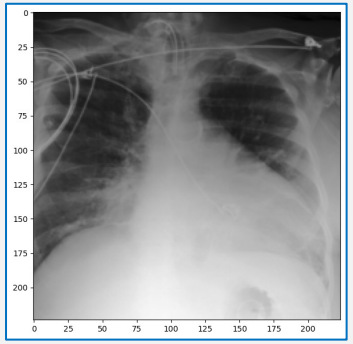
Real Test

Label-pos images have the shortcuts

Atelectasis+



Atelectasis-

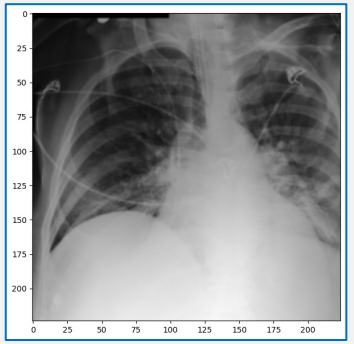


5 versions (1 per label)

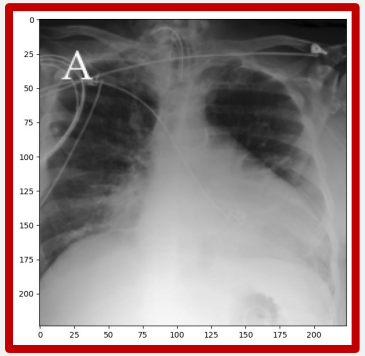
Shortcut Test

Label-neg images have the shortcuts

Atelectasis+

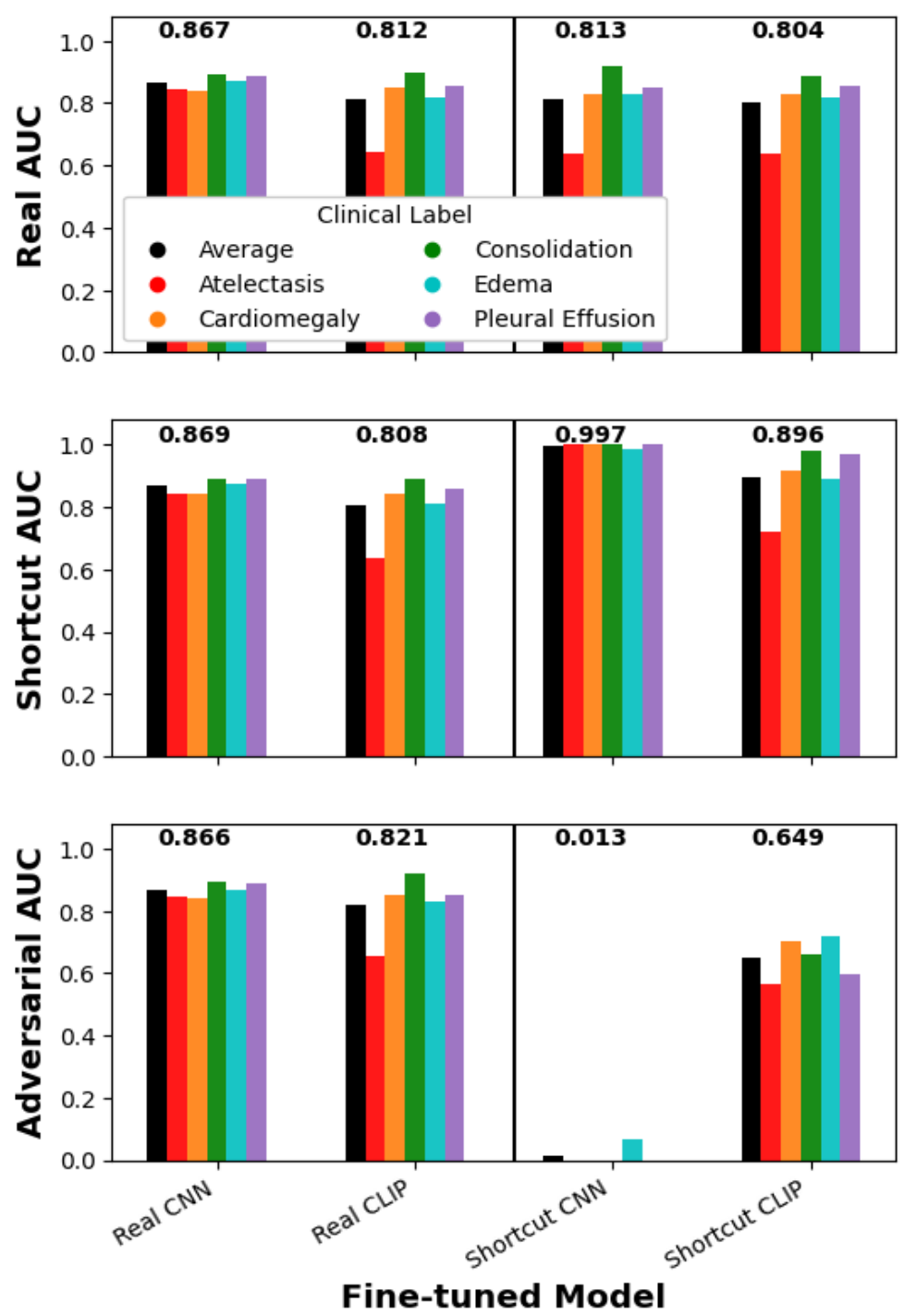


Atelectasis-



5 versions (1 per label)

Adversarial Test



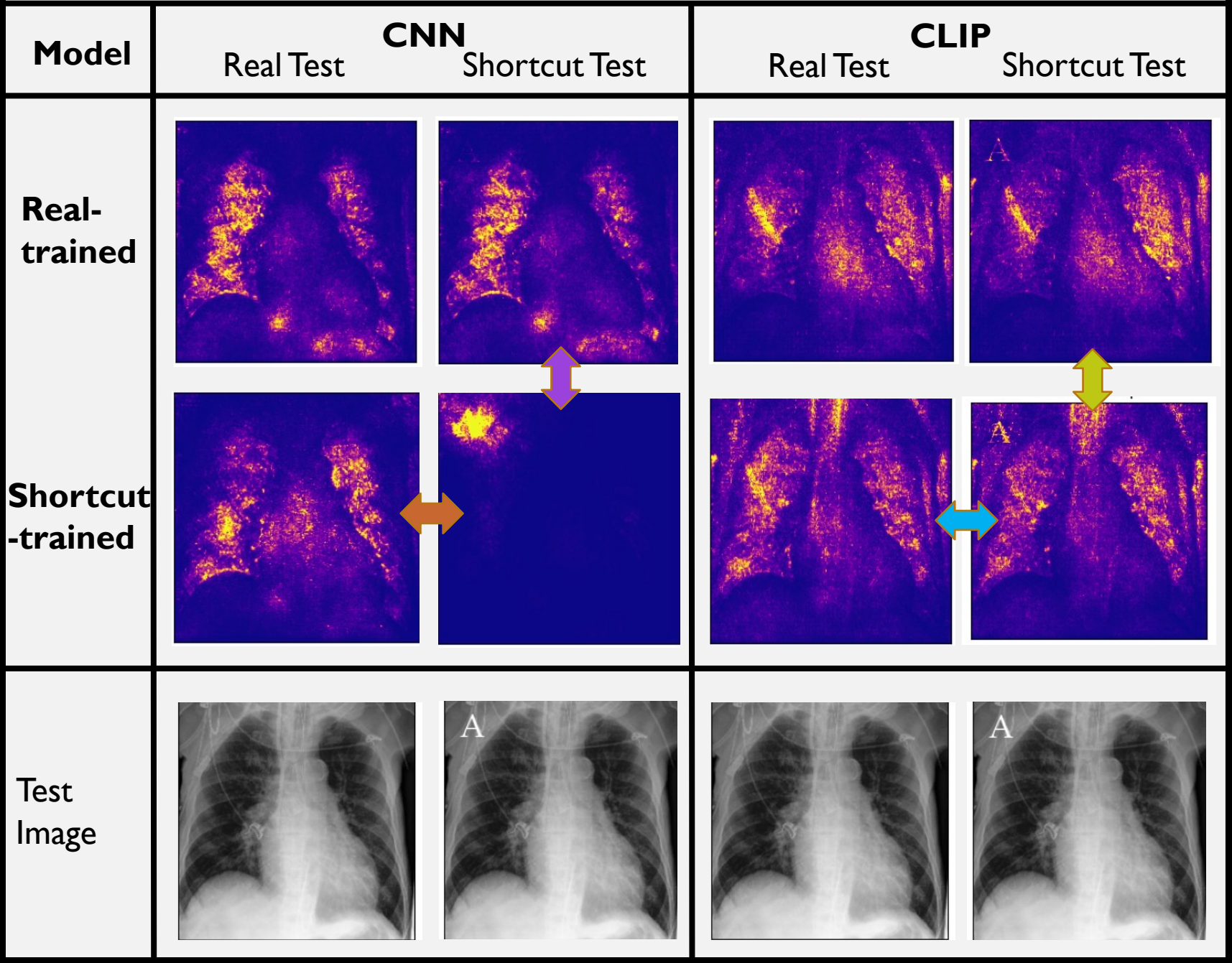
AUCS OF EACH MODEL

Real-trained models are consistent regardless of testing data

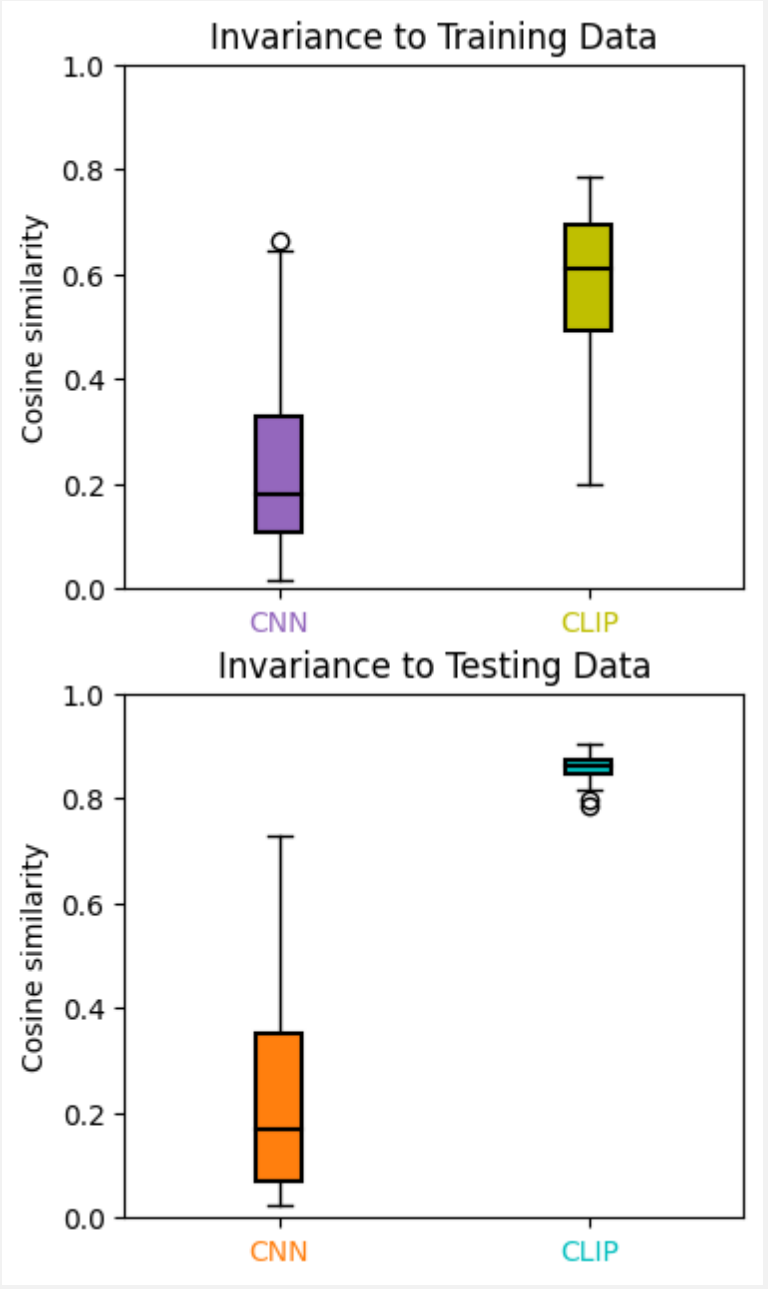
Shortcut CNN is almost perfect on shortcut data and fails completely on adversarial test

Shortcut CLIP is less affected by shortcut/adversarial data

(a) Example Integrated Gradient Maps



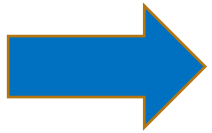
(b) Integrated Gradient Consistency



CONCLUSION

Unlike the shortcut CLIP architecture, the shortcut CNN model has:

- Higher val. loss on real images after fine-tuning
- Near-perfect/near-zero test-AUC on shortcut/adversarial test sets respectively
- Drastically distinct integrated gradient maps when a shortcut is present



Self-supervision during pretraining = less shortcut-reliant model

QUESTIONS?

- Anil Palepu
- Email: apalepu@mit.edu

- Dr. Andrew Beam
- Email: andrew_beam@hms.harvard.edu