Self-Supervision on Images and Text Reduces Reliance on Visual Shortcut Features

Anil Palepu, Dr. Andrew Beam

Introduction:

- Fully-supervised deep neural networks often learn shortcuts, or decision rules based on spurious association during training that are not useful in broader testing and deployment settings.
- Recent self-supervised models like CLIP have demonstrated an ability to jointly train image and text encoders. We hypothesized this type of pretraining would yield a more shortcut resilient vision model.

Data:

- Data: MIMIC_CXR (train/val) and CheXpert (fine-tune/test).
- Clinical Labels: Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion
- Shortcut data: Generated synthetic shortcut data from chest x-ray images by randomly watermarking shortcuts associated with the labels

Methods:

- Trained the following four models: Real CNN, Real CLIP, Shortcut CNN, Shortcut CLIP.
  - Real vs Shortcut = Type of training data used
  - CNNs - supervised training on clinical labels
  - CLIP models - contrastive training w/ radiology reports and the BiomedVLP-CXR-BERT text encoder
- Classification heads added to CLIP models to create 4 identical CNN architectures that differed only in training
- Fine-tuned each model on the same 1% of CheXpert train set using identical hyperparameters

Evaluations:

- For each label, we generated a *shortcut* dataset (watermarks on all label-positive images) and *adversarial* dataset (watermarks on label-negative images). We computed AUCs of the models on the *real*, *shortcut*, and *adversarial* test datasets
- We computed integrated gradient maps for each model and compared consistency between maps

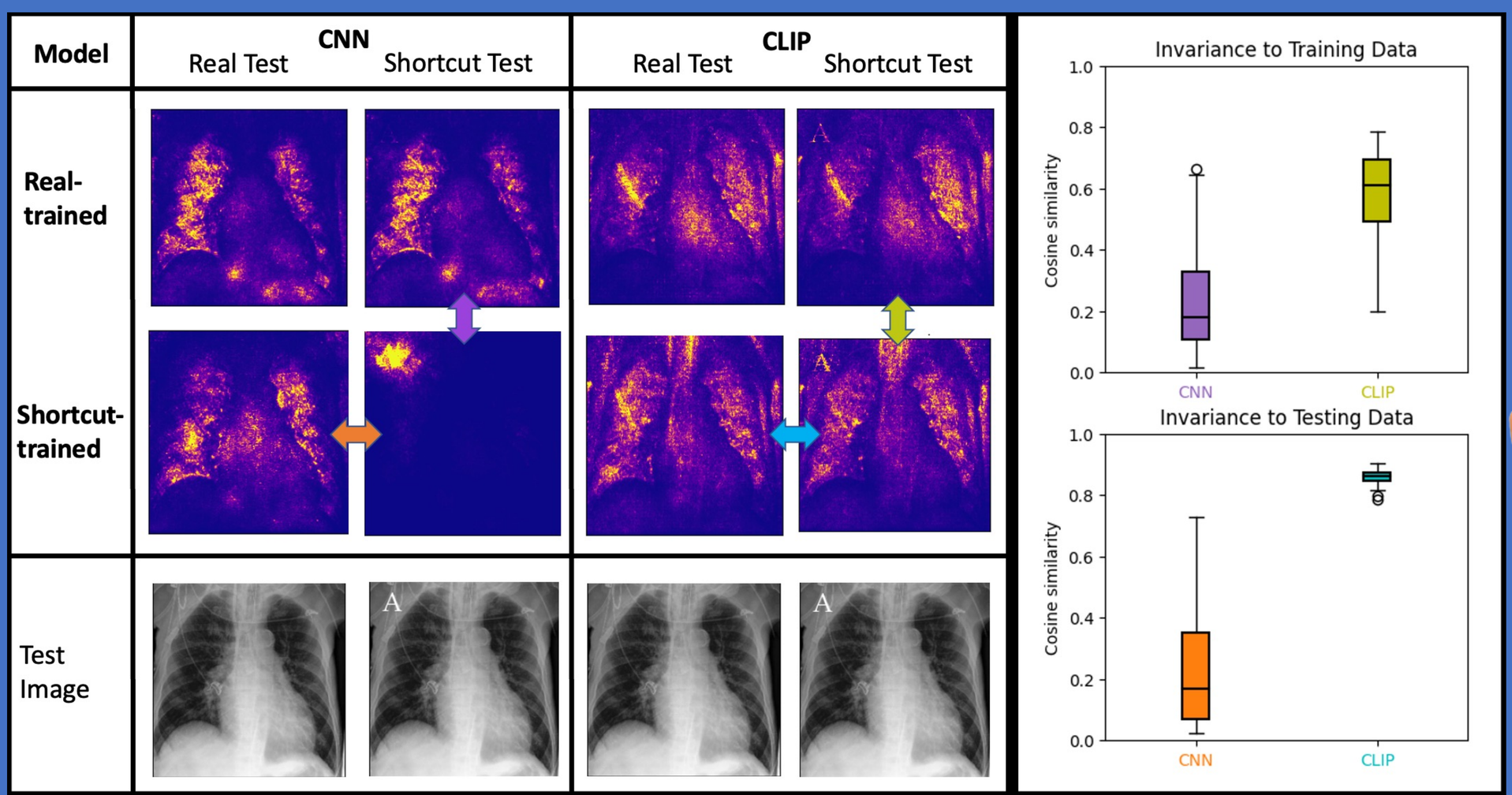
Results:

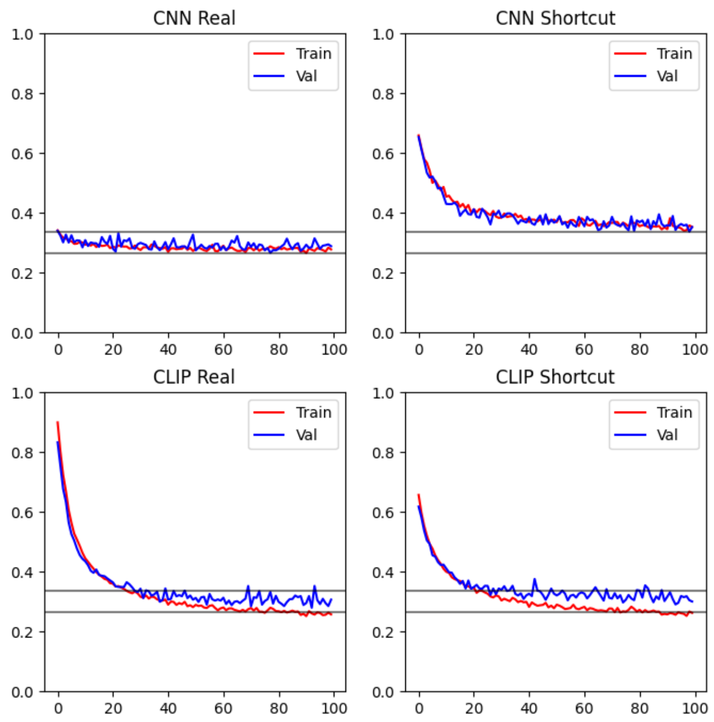
- Even after fine-tuning on real data, the shortcut-trained CNN fails to unlearn shortcut dependence
- The shortcut-trained CNN model is heavily reliant on shortcuts, failing completely on adversarial test data. The shortcut-trained CLIP model is relatively resilient, though still affected.
- Qualitatively, the integrated gradient maps suggest that the shortcut trained CNN focuses heavily on the watermarks when they are present.
- The integrated gradients are far more consistent for CLIP, regardless of if there were shortcuts present in the training data or in the test images.

Conclusions:

- The self-supervision provided by radiology reports promoted learning of clinically relevant features beyond the watermarked shortcuts, and lack of this self-supervision yielded a more “stubborn” and shortcut-reliant model.

Self-supervision with **text** **improves** CNN **robustness** to visual **shortcut features**

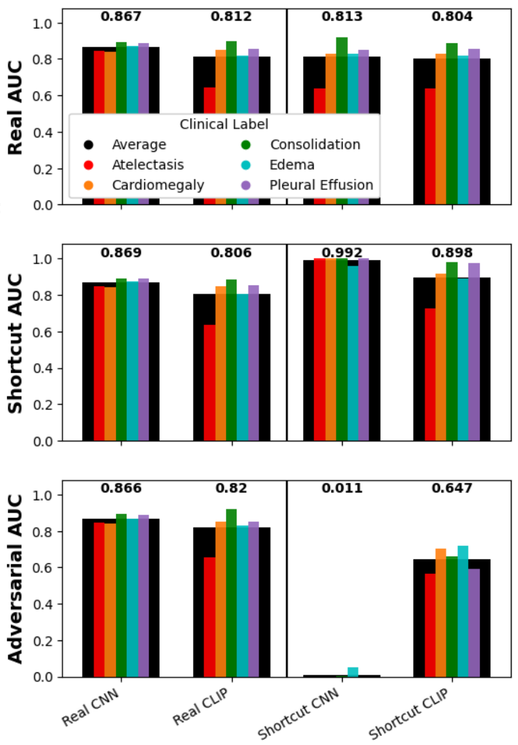




Above: Integrated gradients for an example image show shortcut CNN focuses on the watermark. The boxplots demonstrate CLIP models have more consistency regardless of shortcuts in train/test data

Left: Shortcut-trained CNN model fails to achieve comparable loss even after fine-tuning on real data

Right: Shortcut-trained CNN model gets near perfect accuracy on shortcut test data and near-zero on adversarial shortcuts. Shortcut-trained CLIP is relatively resilient.



beamlab

Contact us!

Anil Palepu
apalepu@mit.edu

Dr. Andrew Beam
andrew_beam@
hms.harvard.edu