Interpretable Neural Networks for Computer Vision: Clinical Decisions that are Computer-Aided, not Automated

Cynthia Rudin
Professor of Computer Science
Duke University

Right for the wrong reasons (Clever Hans)

Clever Hans performing in 1904

Why interpretability?

- Confounding: Right for the wrong reasons (Clever Hans)
- High-stakes decisions: Should patient get a biopsy?
- Troubleshooting: Can’t second-guess the reasoning process of a black box.
    - Will it work if I switch equipment?
    - Will it work for all types of patients?
    - Can I check if it’s working on my current patient?
    - Is the information that I fed into the system correct?
- Responsibility: It is the *doctor*’s responsibility to make a good decision. (Isn’t it?)

The use of black box models makes all of these much worse.

Black box models turn computer-aided decisions into automated decisions.

”Explaining” deep NN’s with saliency maps doesn’t work

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps		 “Explanation”	

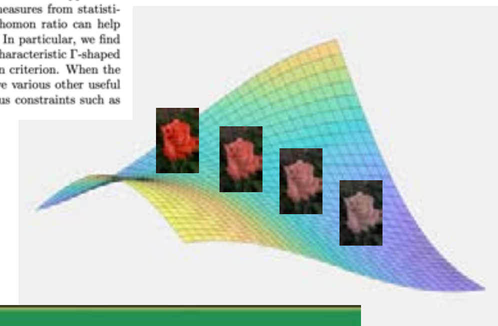
Do you trust the network now?

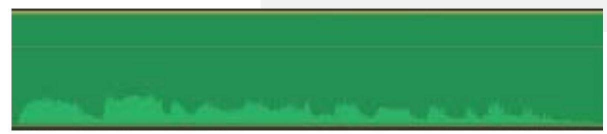
Lots of work in radiology on attention maps now…

Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.





Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Problem spectrum

Very sparse models (trees, scoring systems)

With minor pre-processing, all methods have similar performance

Neural networks

Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

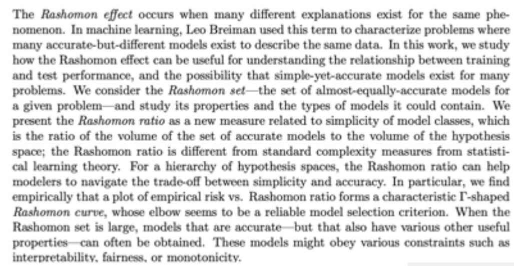
- pixels/voxels, words, a bit of a sound wave

Problem spectrum

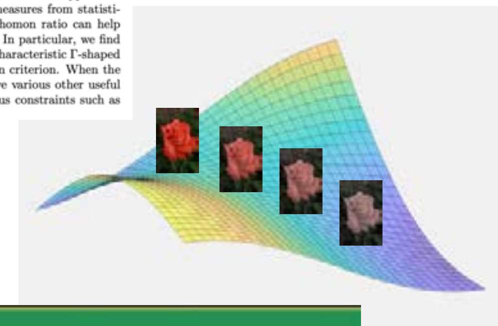
age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

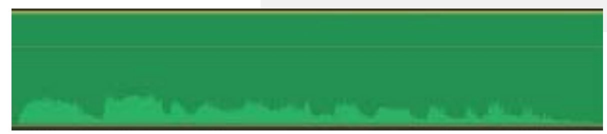
Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data



The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.





Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

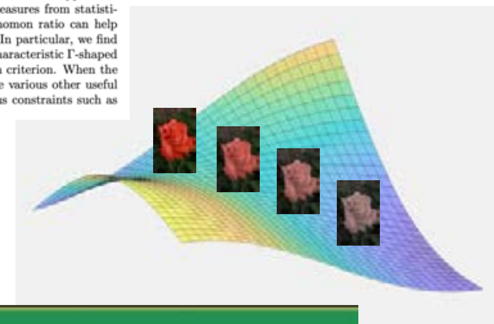
Problem spectrum

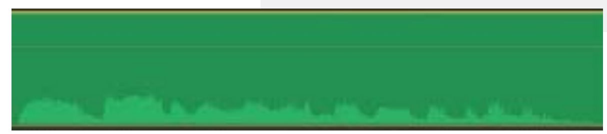
age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.





Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

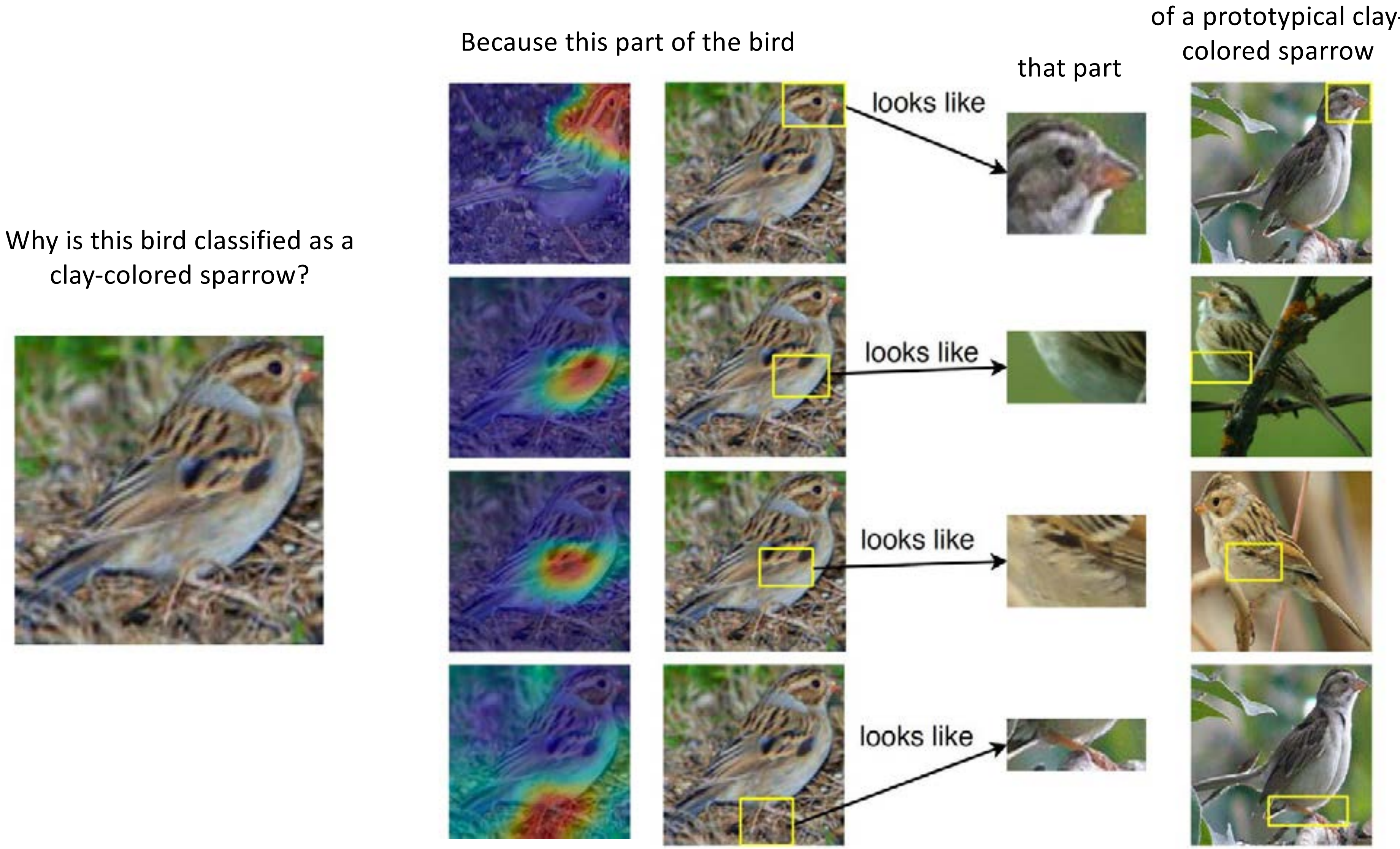
For tabular data: decision trees, linear/additive models/scoring systems
For raw data: interpretable neural networks

models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are inter-pretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

There has been an increasing trend in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives. Many of not. There is a spectrum between fully transparent models (where we understand how all the variables are jointly related to each other) and models that are lightly constrained in model form (such as models

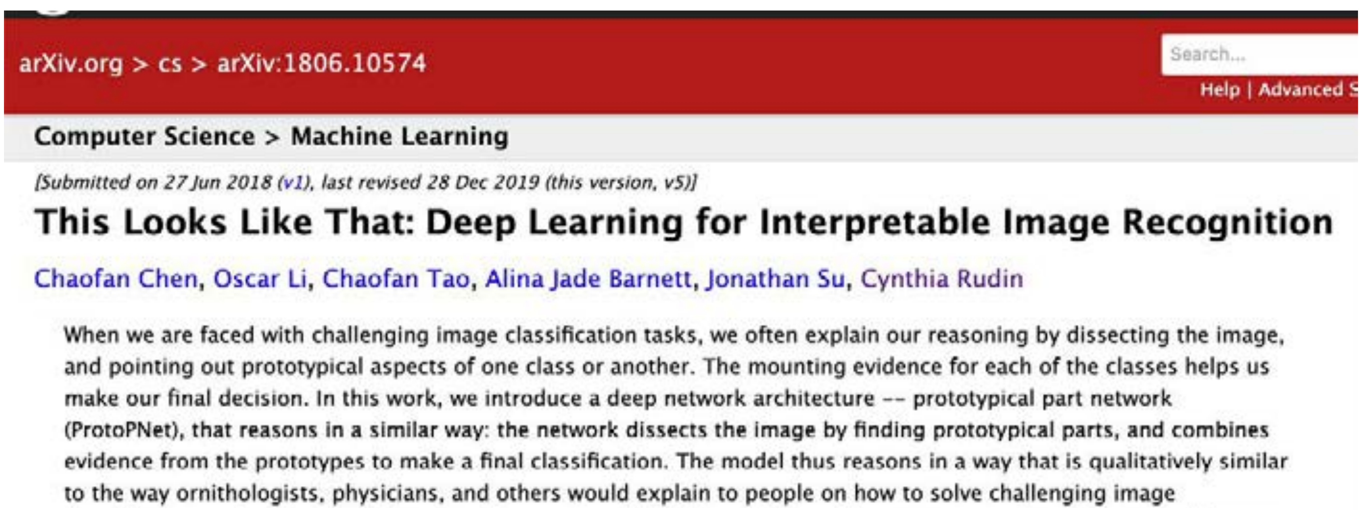
But what is an interpretable neural network?

Approach 1: A neural network that does case-based reasoning


Why is this bird classified as a clay-colored sparrow?
Because this part of the bird
looks like
looks like
looks like
looks like
that part
of a prototypical clay-colored sparrow

“*This* Looks Like *That:* deep learning for interpretable image recognition”

NeurIPS 2019 (spotlight)


arXiv.org > cs > arXiv:1806.10574

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 27 Jun 2018 (v1), last revised 28 Dec 2019 (this version, v5)]

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, Cynthia Rudin

When we are faced with challenging image classification tasks, we often explain our reasoning by dissecting the image, and pointing out prototypical aspects of one class or another. The mounting evidence for each of the classes helps us make our final decision. In this work, we introduce a deep network architecture -- prototypical part network (ProtoPNet), that reasons in a similar way: the network dissects the image by finding prototypical parts, and combines evidence from the prototypes to make a final classification. The model thus reasons in a way that is qualitatively similar to the way ornithologists, physicians, and others would explain to people on how to solve challenging image

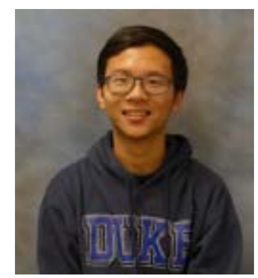
- Adds a “prototype” layer to a black box, forces the network to do case-based reasoning.
- Prototypes are learned during training.

Oscar

Chaofan

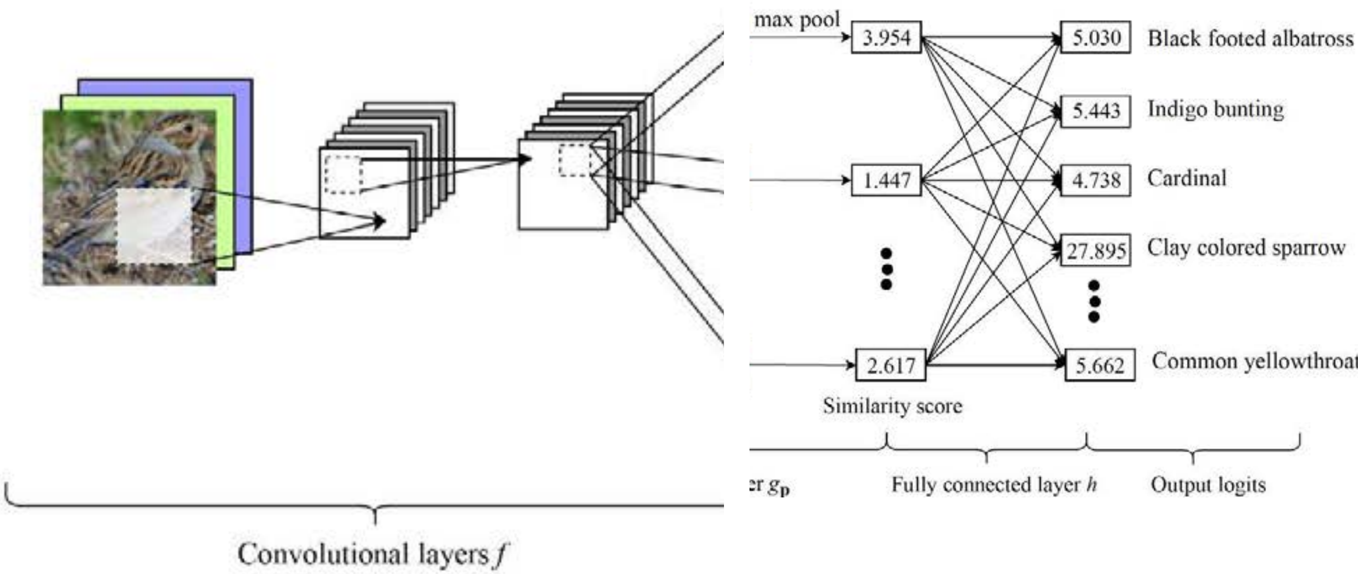
Alina

Jonathan

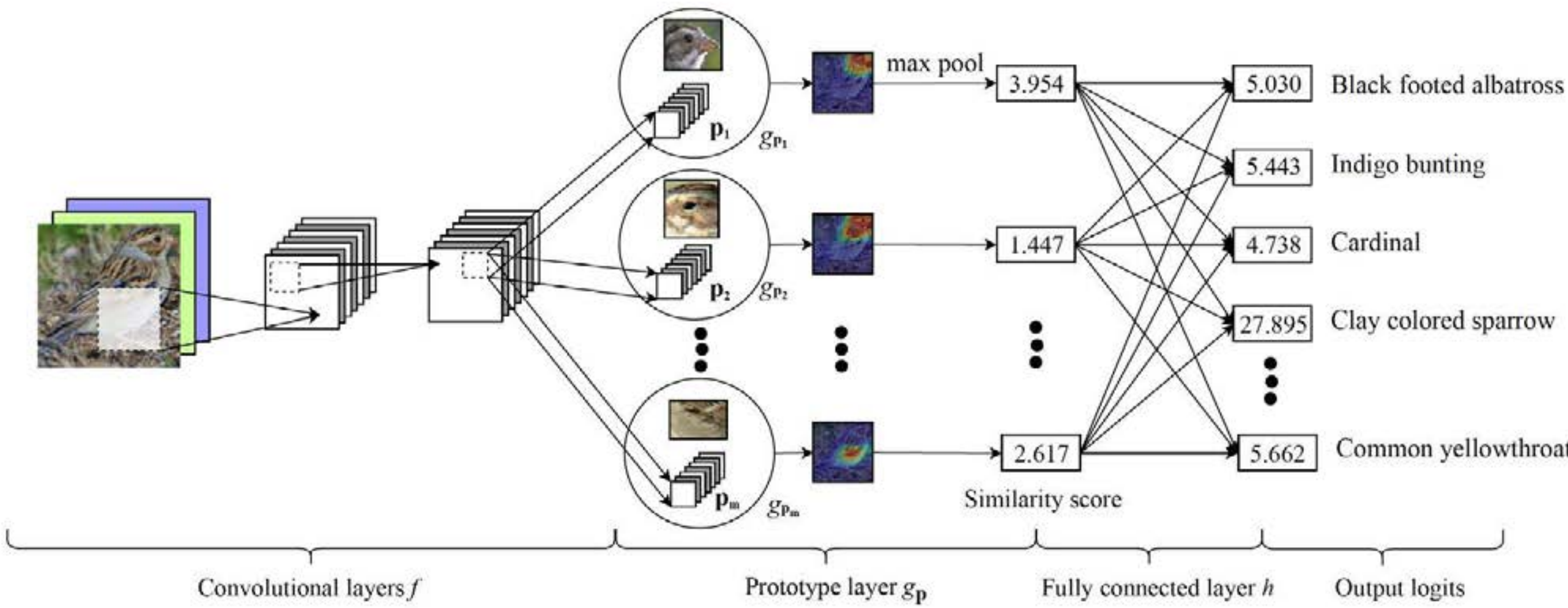


Daniel

Take any “standard” black box CNN…

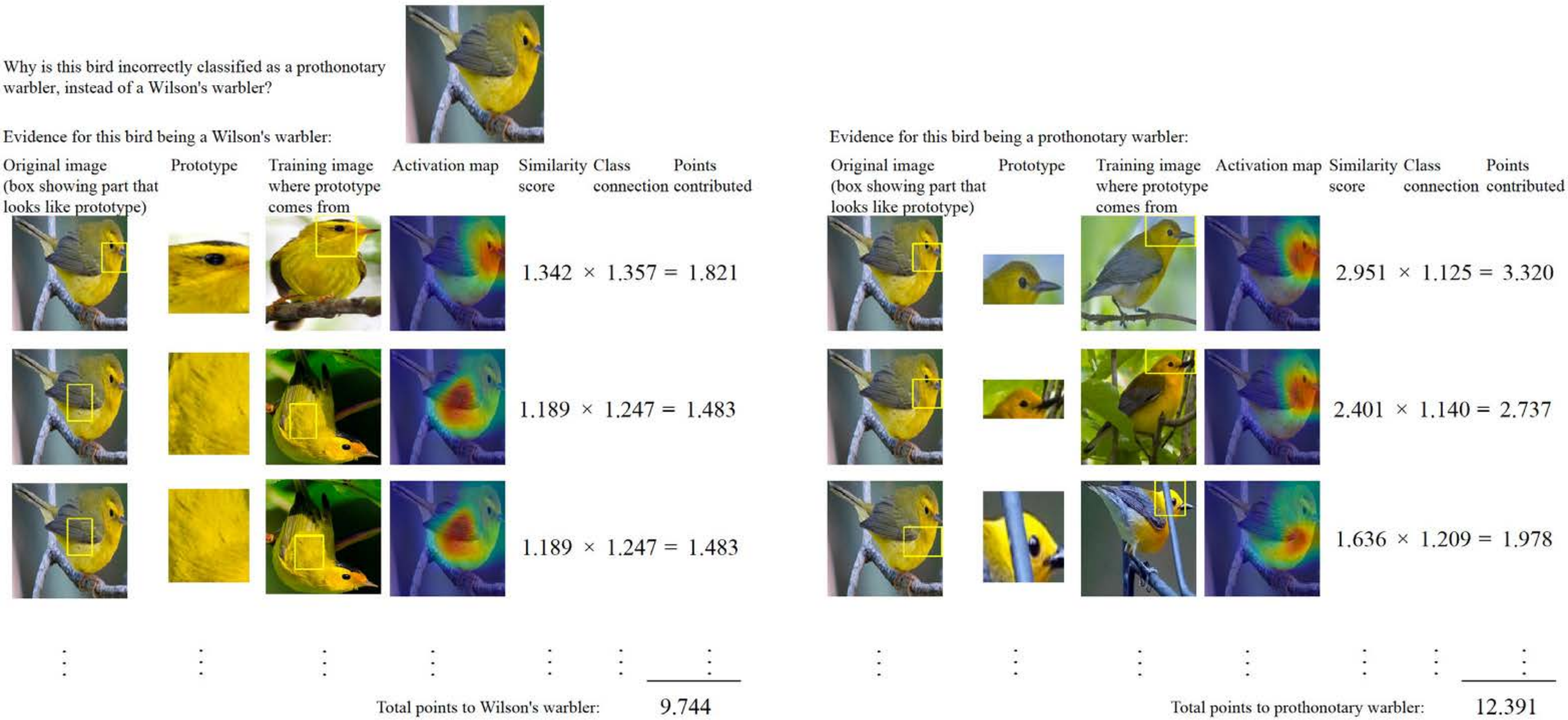


And transform it to be interpretable

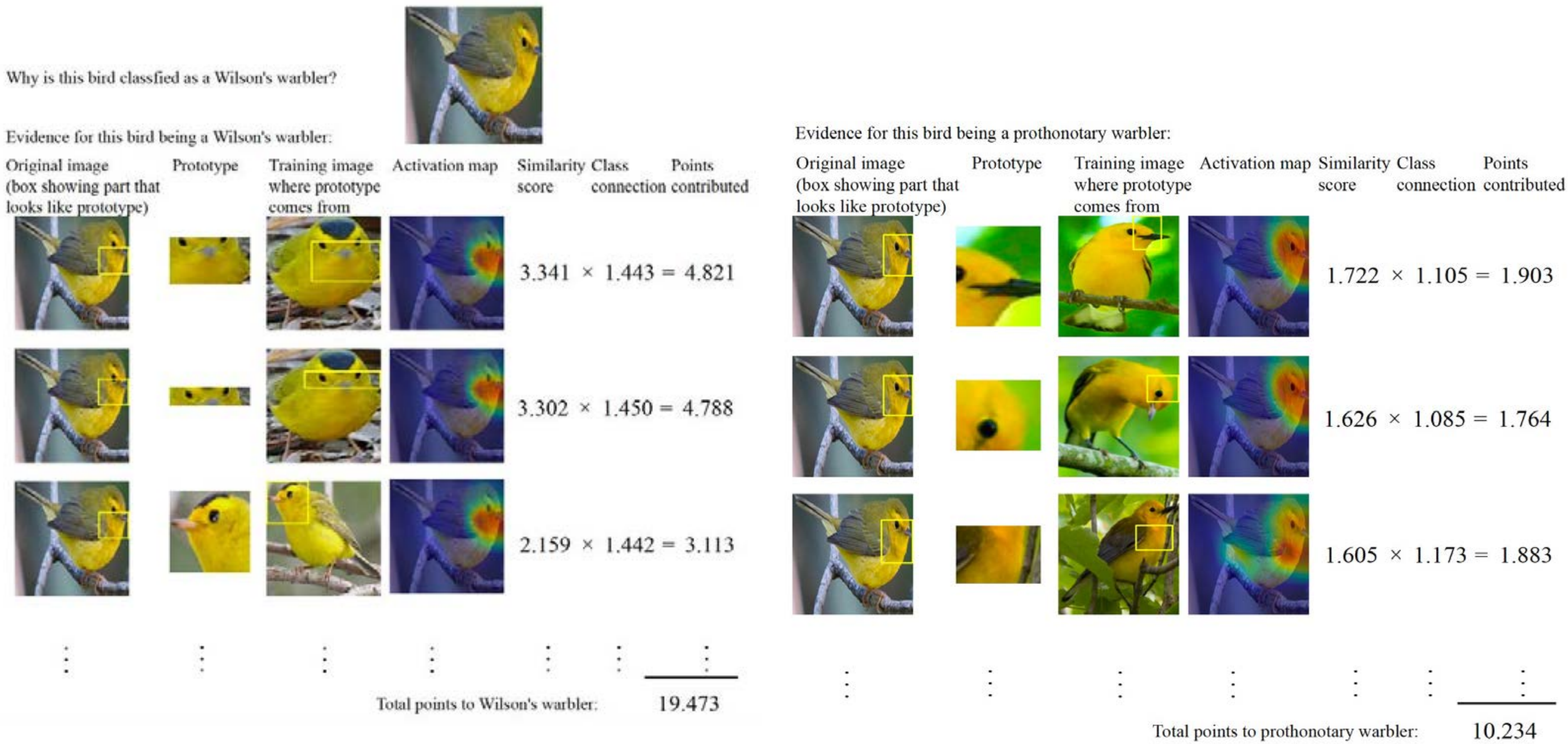



Why is this bird classfied as a red-bellied woodpecker?
Evidence for this bird being a red-bellied woodpecker:
Original image (box showing part that looks like prototype)
Prototype
Training image where prototype comes from
Activation map
Similarity score
Class connection
Points contributed
6.499 × 1.180 = 7.669
4.392 × 1.127 = 4.950
3.890 × 1.108 = 4.310
Total points to red-bellied woodpecker: 32.736
Evidence for this bird being a red-cockaded woodpecker:
Original image (box showing part that looks like prototype)
Prototype
Training image where prototype comes from
Activation map
Similarity score
Class connection
Points contributed
2.452 × 1.046 = 2.565
2.125 × 1.091 = 2.318
1.945 × 1.069 = 2.079
Total points to red-cockaded woodpecker: 16.886

Base model: DenseNet161



Base model: VGG-16



CUB-200

- 200 classes of birds
- Original black box accuracy between 74.6% (VGG16) and 82.3% (Res34).
- Interpretable model's accuracy between 76.1% (VGG) and 80.2% (Dense121). Combining several interpretable networks together yields 84.8%, and still yields an interpretable model.

So even for computer vision, we can still have an interpretable model of the same accuracy as a black box.

Where might this be useful?

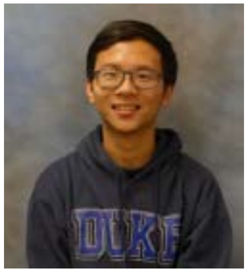
Mammography

- Breast cancer is a leading cause of death in the USA (Kochanek et al 2020)
- Hundreds of thousands of cases diagnosed in the USA each year (> 300K in 2019), causing tens of thousands of deaths each year.
- Mammography is the *hardest task* in all of radiology (Moss, 2020)
  - Radiologists miss ~1/5 of breast cancers
  - half of women getting an annual mammogram over 10 years will have a false positive.
  - up to 3/4 of biopsies come back as benign, i.e., potentially unnecessary surgeries

Our team

Alina Barnett

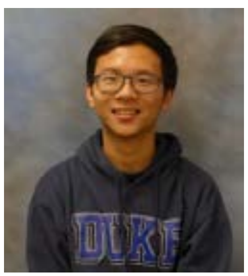
Chaofan Chen



Daniel Tao

Our team

Alina Barnett



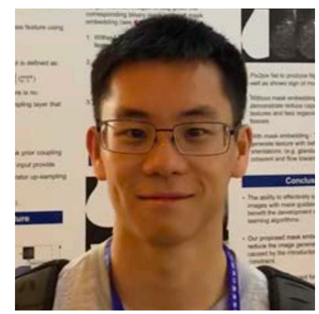
Daniel Tao

Chaofan Chen

Several AI algorithms have FDA approval for radiology!!

Joseph Lo
Professor of Radiology

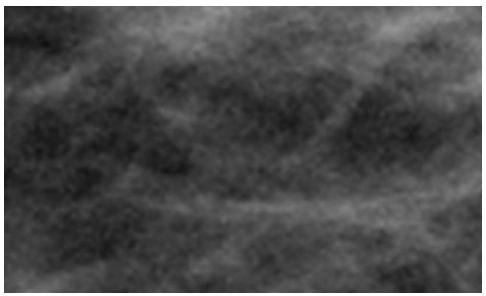
Fides Regina Schwartz, M.D.



Yinhao Ren, PhD student

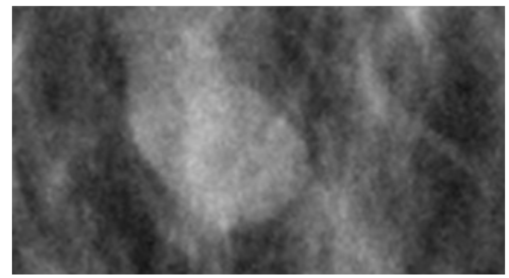
Two different problems

Breast lesion detection



Black box says:
There’s no lesion in this image.

Whether to order a biopsy for a lesion?



Black box says:
Don’t get a biopsy.

Why is AI mammography hard?

- AI radiology is hard
- Mammography is just really, really hard.
- No data.
- Confounding: Right for the wrong reasons → hard to deal with.
- How to design a system that would actually be helpful?
  - Malignancy vs. Benign is NOT the right problem to solve!

Why is AI mammography hard?

- AI radiology is hard
- Mammography is just real
- No data.
- Confounding: Right for the
- How to design a system th
    - Malignancy vs. Benign

HealthImaging

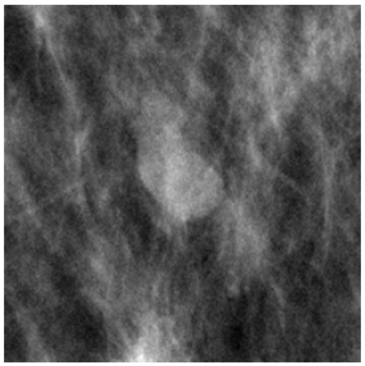
INSIGHTS IN IMAGING & INFORMATICS

Algorithm's 'unexpected' weakness raises larger concerns about AI's potential in broader populations

Matt O'Connor | April 05, 2021 | Artificial Intelligence

A new investigation revealed "unexpected" shortcomings when using a federally cleared artificial intelligence tool to detect intracranial hemorrhages. The findings pushed researchers to call for more standardization when evaluating AI-based clinical decision support platforms.

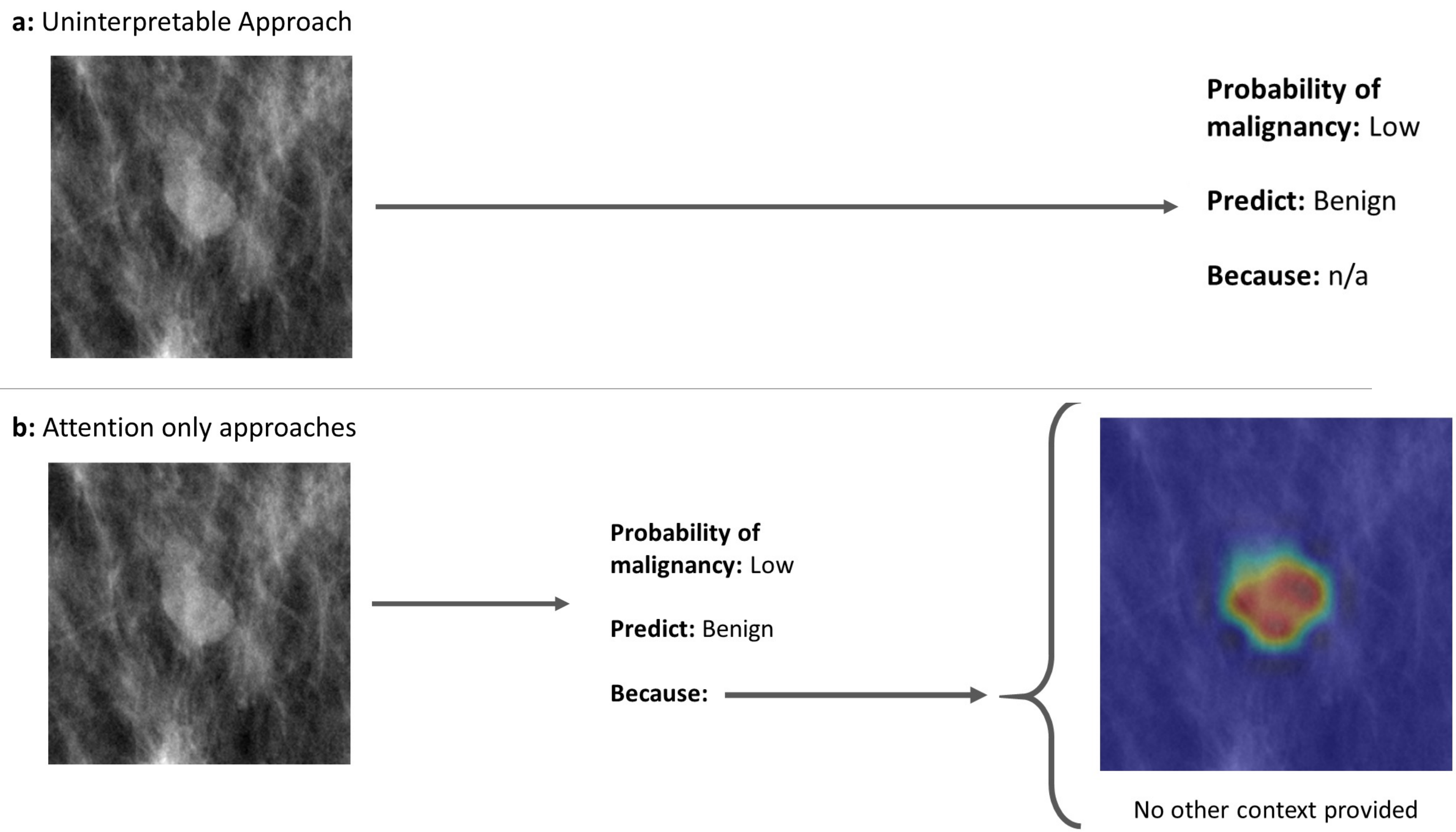
a: Uninterpretable Approach



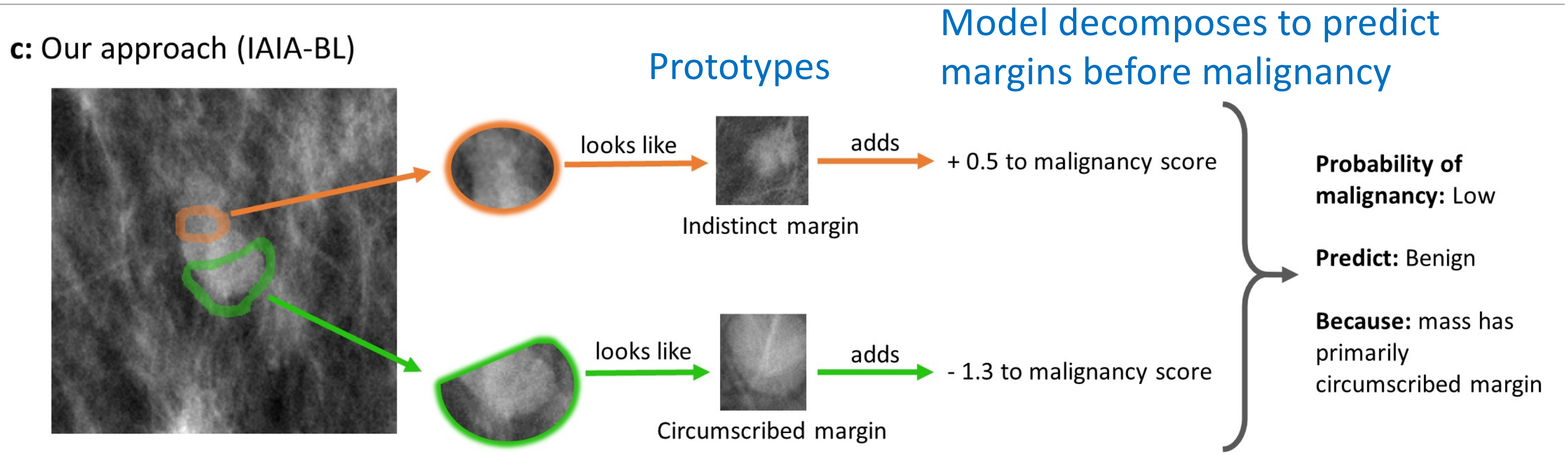
Probability of malignancy: Low

Predict: Benign

Because: n/a


a: Uninterpretable Approach
Probability of malignancy: Low
Predict: Benign
Because: n/a
b: Attention only approaches
Probability of malignancy: Low
Predict: Benign
Because:
No other context provided

Interpretable AI algorithm for Breast Lesions (IAIA-BL)



Data Availability

- Public data availability is abysmal
- Low-quality images, outdated equipment, inconsistent labeling
- Some wanted us to hand over IP…

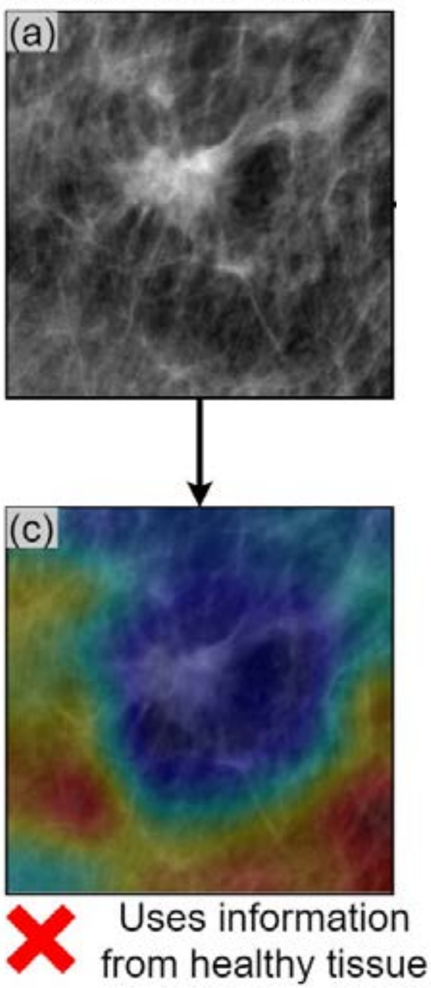
Data

- From Duke!
- 1136 digital screening mammogram images of masses in the breast from 484 patients at Duke University hospitals.

1136?! 484?!

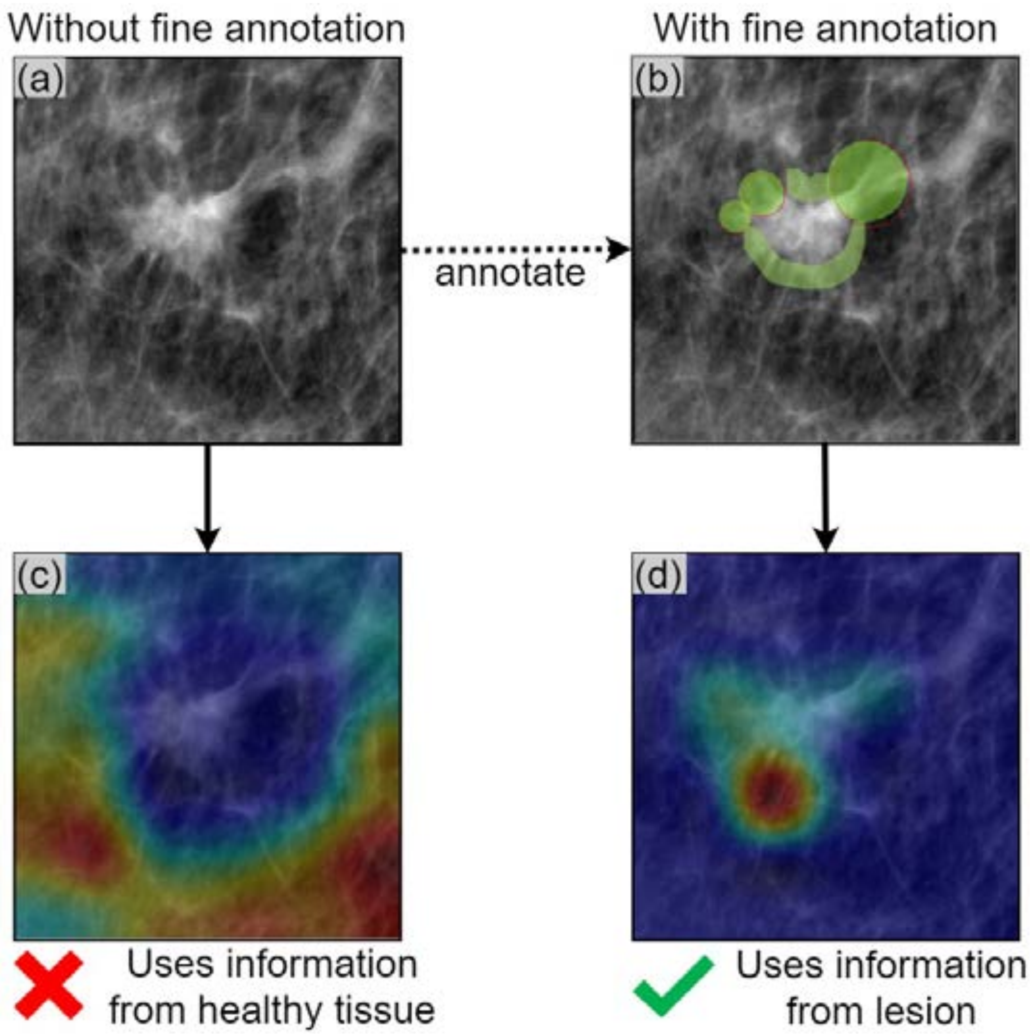
125 masses with spiculated margin
220 with indistinct margin
41 with microlobulated margin (didn't use)
579 with obscured margin (didn't use)
171 with circumscribed margin

Let's do it!



Clever Hans performing in 1904

black box accuracy = “interpretable” accuracy



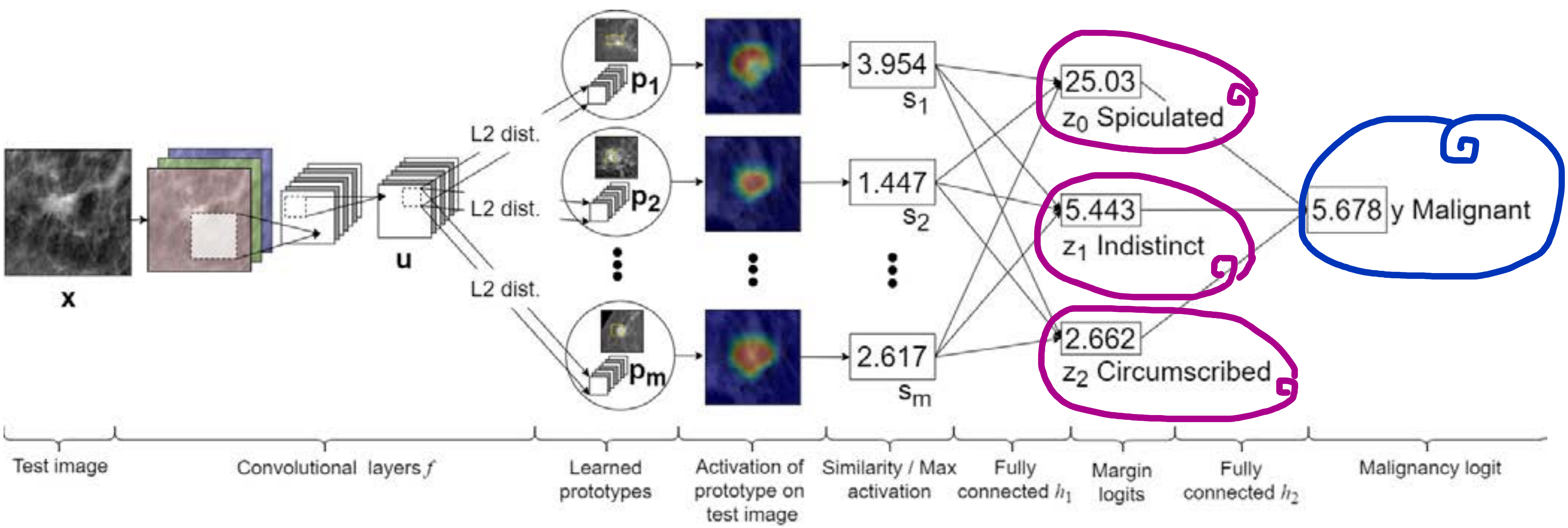
Fine annotation

- Radiologists add fine labels for only 37 images (9%)
- We generalized ProtoPNet to handle mixed labelling: fine-grained attention labels and standard labels

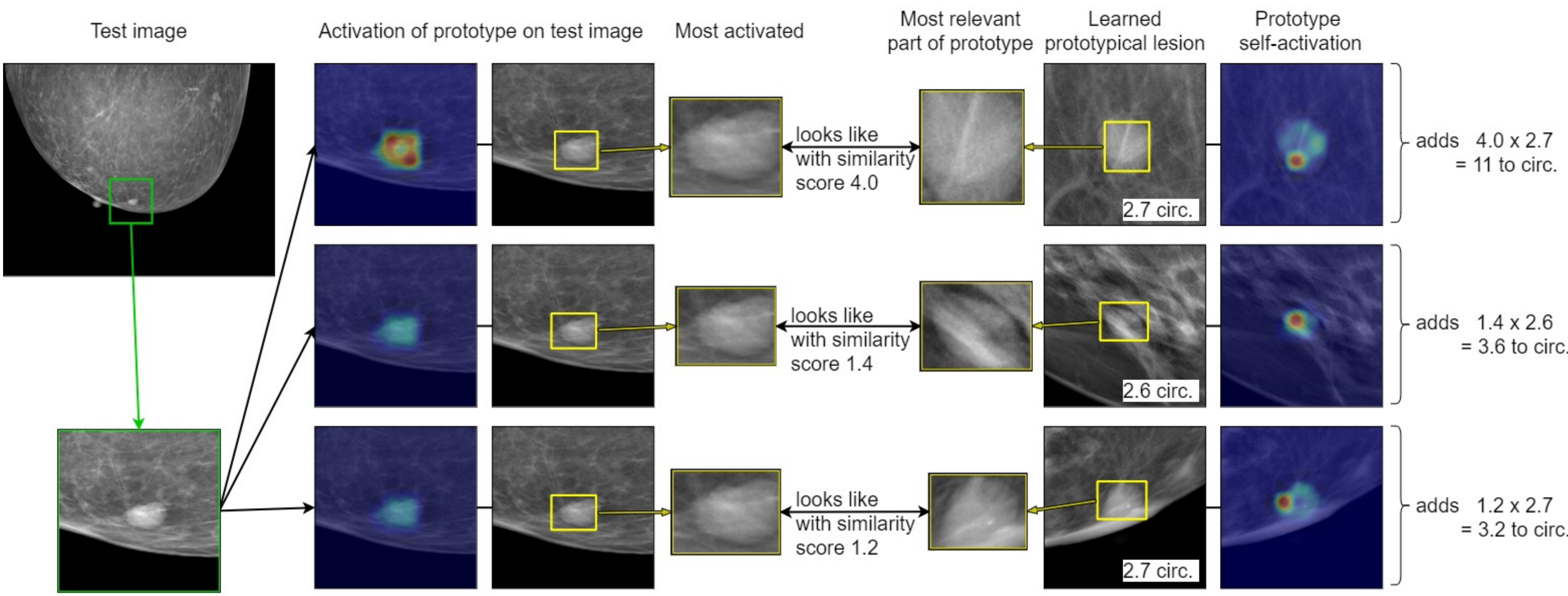
black box accuracy = interpretable accuracy

IAIA-BL Architecture

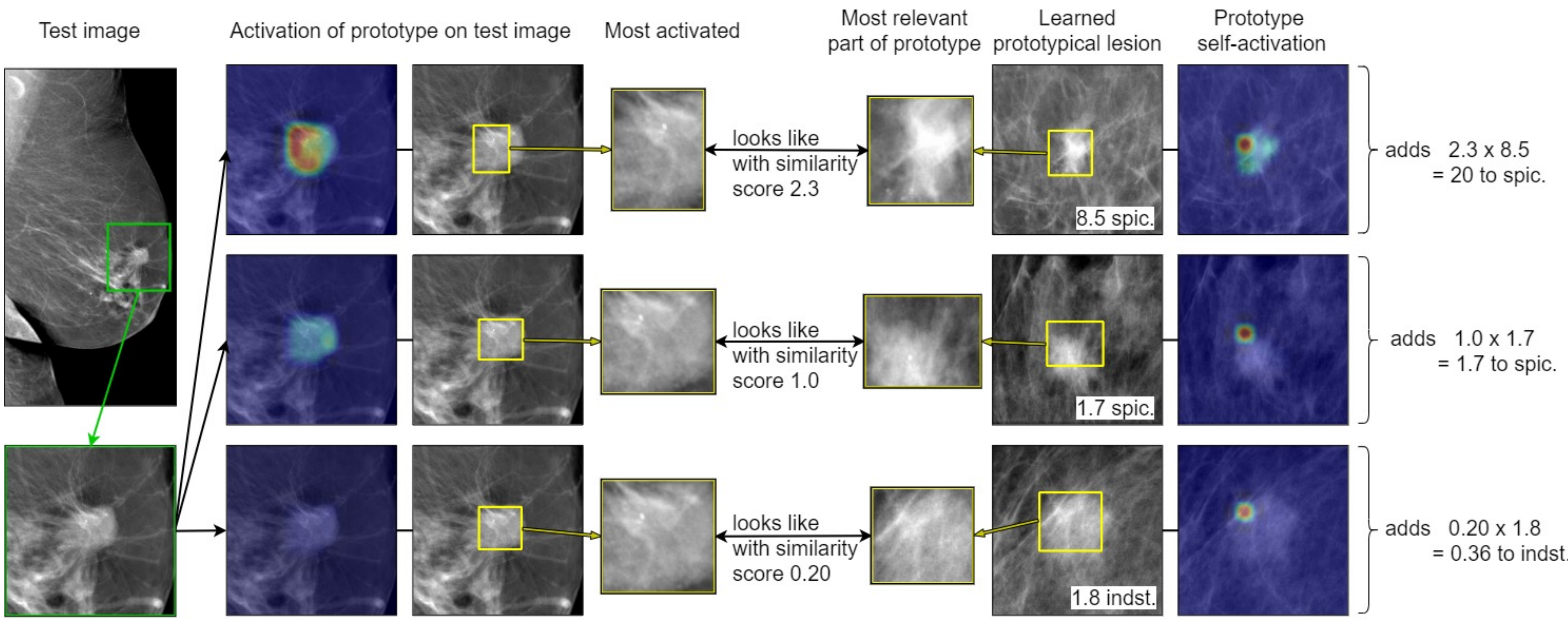
Each classifier is a generalized ProtoPNet, Combined model is linear



Example: why correctly classified as circumscribed



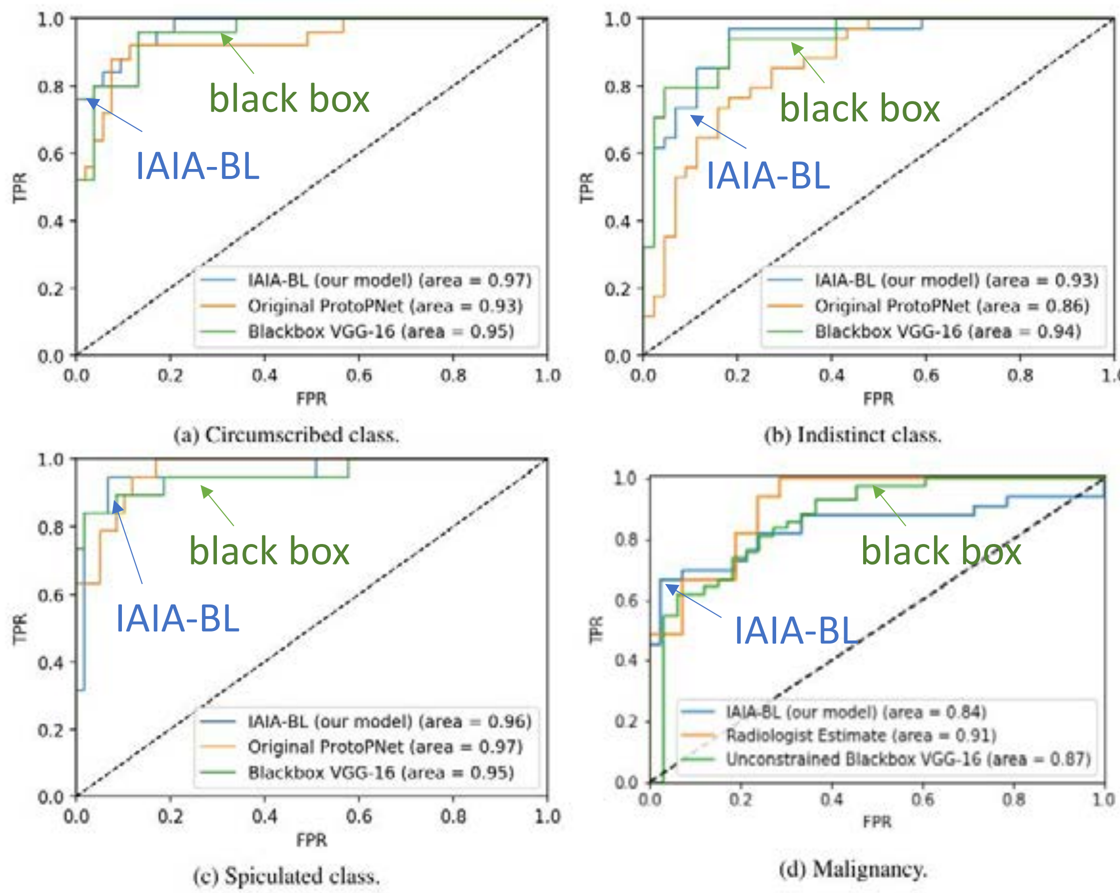
Example: why correctly classified as spiculated



Preliminary results

- Performance *as good or better* than uninterpretable
- (Uninterpretable gets to use confounding info!)

ROC Curves



Why is AI mammography hard?

- Mammography is just really, really hard.
- No data. <Thank you Joseph>
- Confounding: Right for the wrong reasons → hard to deal with.
- How to design a system that would actually be helpful?
  - Malignancy vs. Benign is NOT the right problem to solve!

Fine annotation

Decompose the problem.
Interpretable models for each part
Case-based reasoning

What we did NOT do:
- black box + saliency
- malignant vs. benign only

Approach 2: Neural Disentanglement

arXiv.org > cs > arXiv:2002.01650

Computer Science > Machine Learning

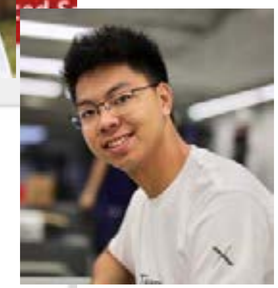
[Submitted on 5 Feb 2020 (v1), last revised 19 Oct 2020 (this version, v4)]

Concept Whitening for Interpretable Image Recognition

Zhi Chen, Yijie Bei, Cynthia Rudin

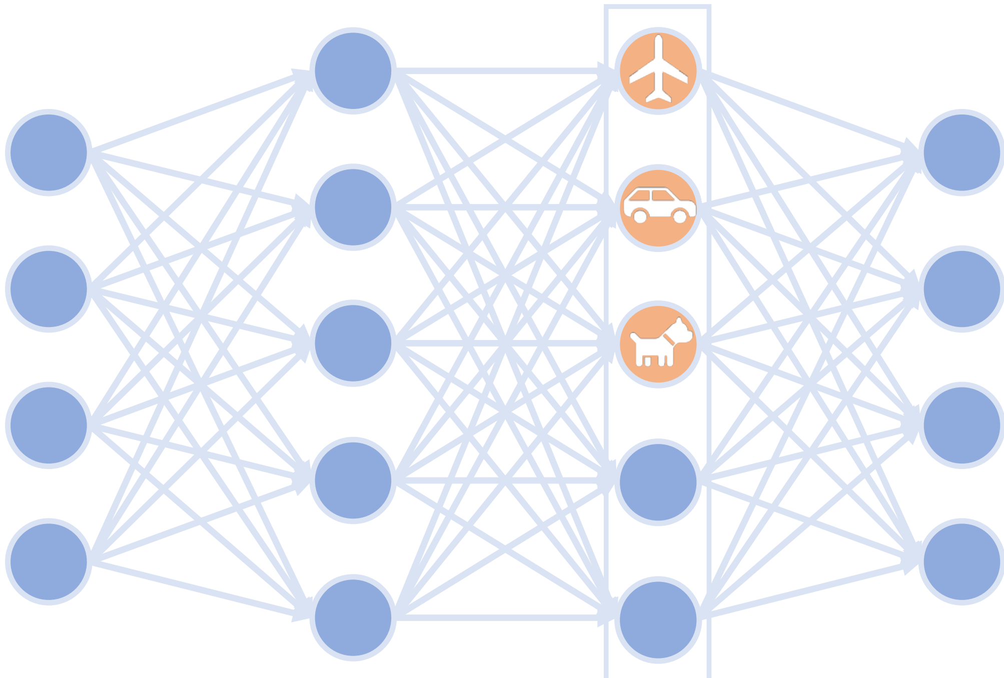
What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning is undoubtedly important, but the calculations of neural networks are very challenging to understand. Attempts to see inside their hidden layers can either be misleading, unusable, or rely on the latent space to possess properties that it may not have. In this work, rather than attempting to analyze a neural network posthoc, we introduce a mechanism, called concept whitening (CW), to alter a given layer of the network to allow us to better understand the computation leading up to that layer. When a concept whitening module is added to a CNN, the axes of the latent space are aligned with known concepts of interest. By experiment, we show that CW can provide us a much clearer understanding for how the network gradually

Zhi

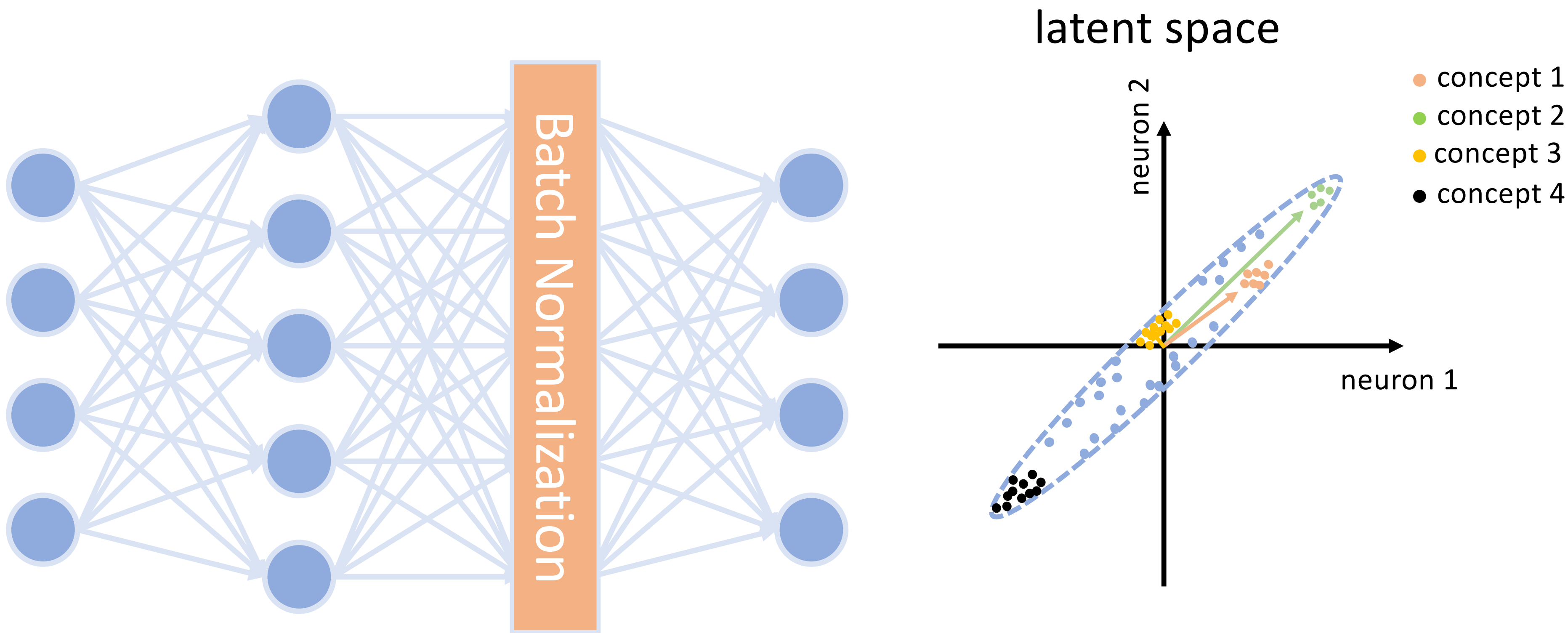


Webster

Nature Machine Intelligence, 2020

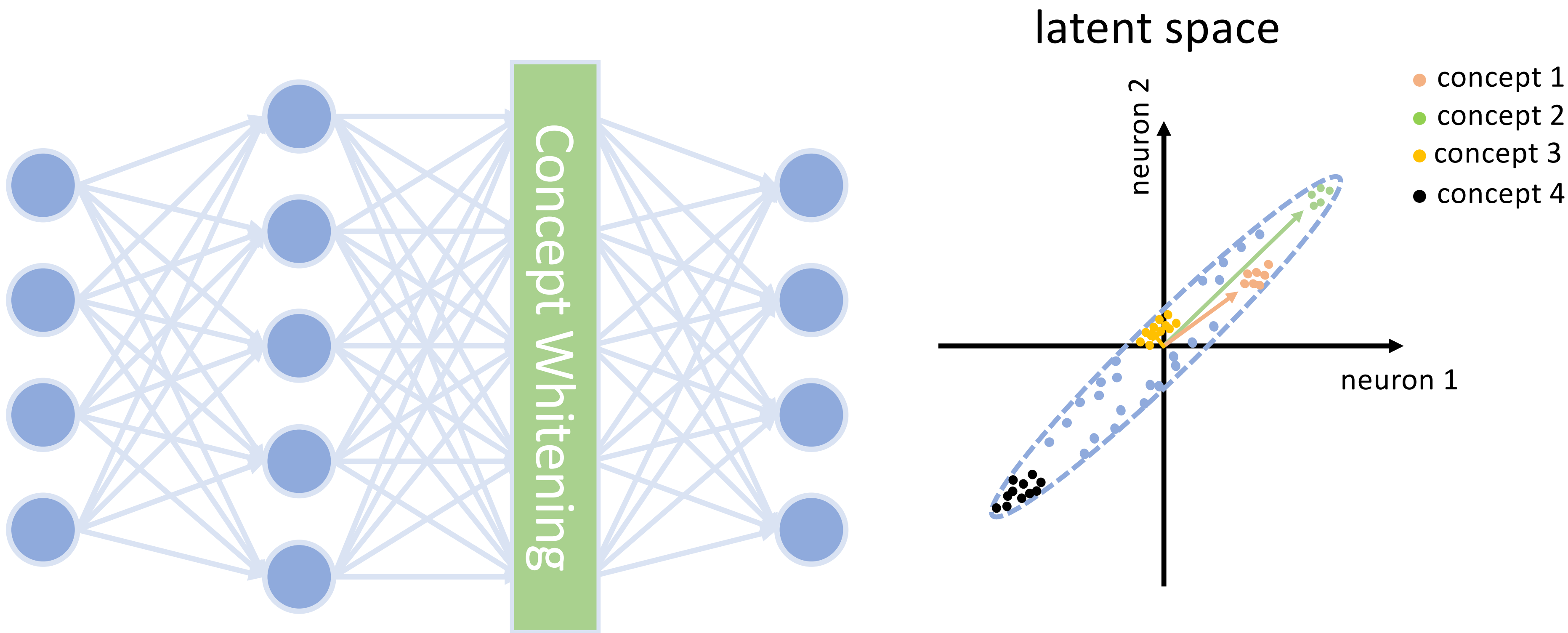


CNN's are not naturally disentangled

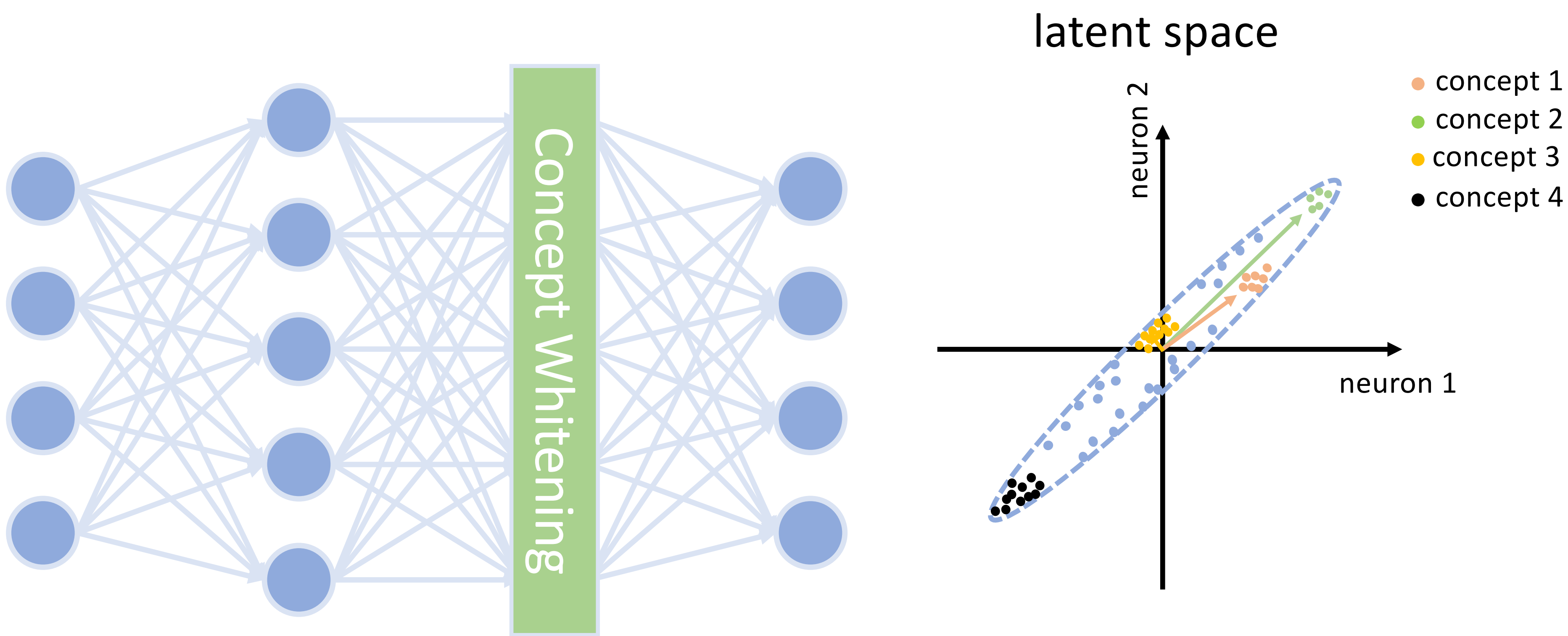


Consider the latent space of a Batch Norm layer

Create a vector pointing towards each concept. They are not naturally orthonormal.

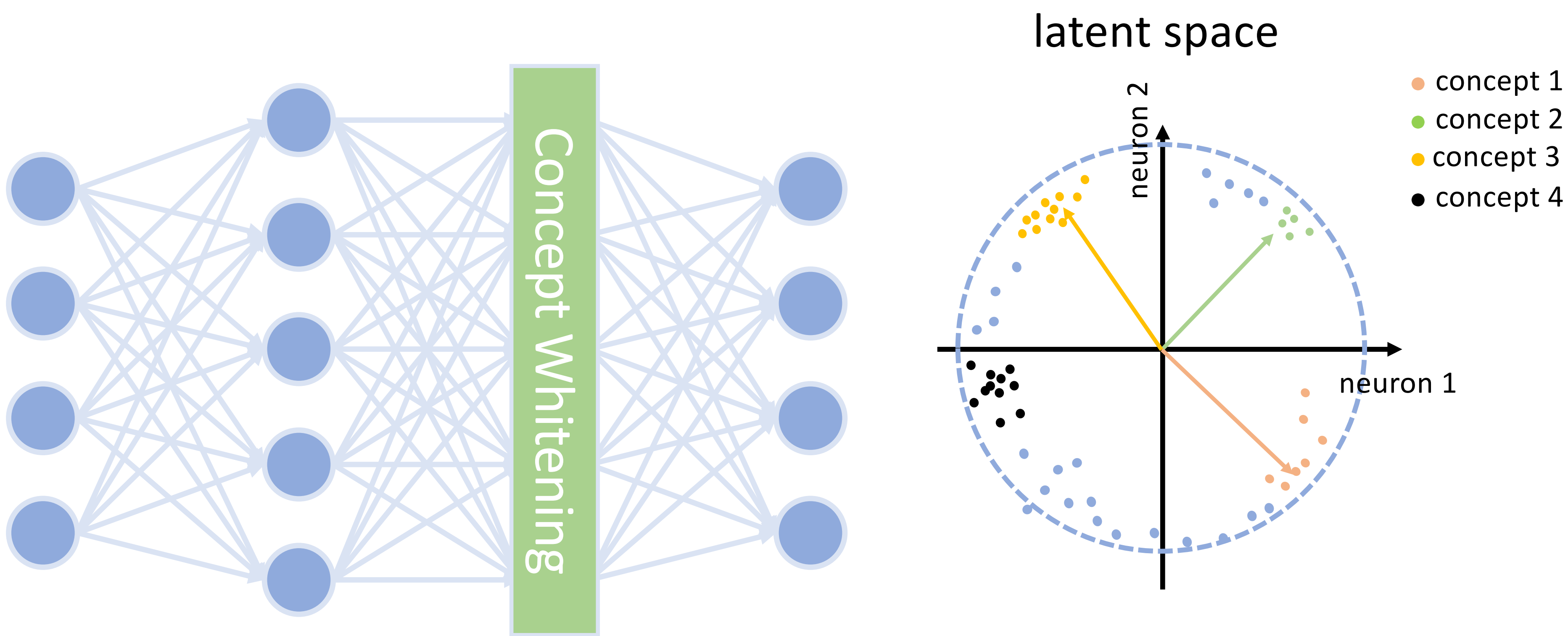


Concept Whitening
latent space
concept 1
concept 2
concept 3
concept 4
neuron 2
neuron 1



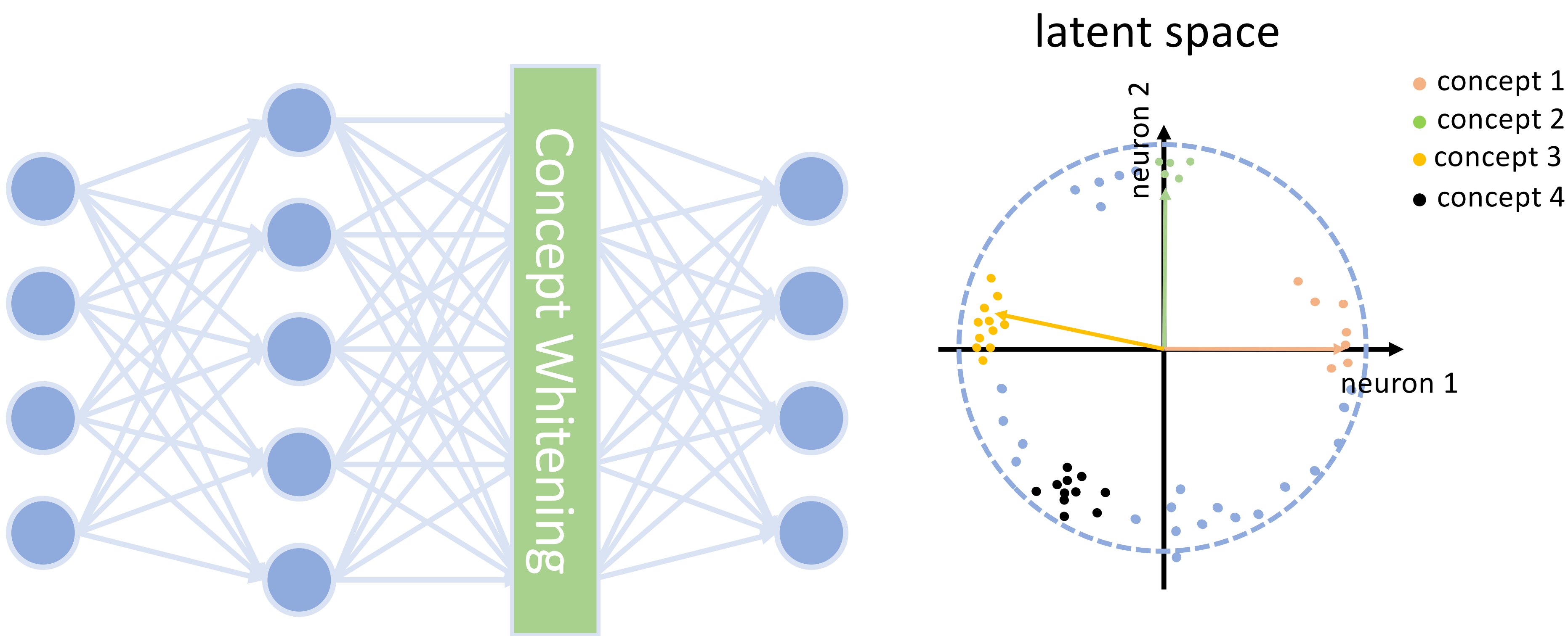
When a Concept Whitening module is added to a CNN,

- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest



When a Concept Whitening module is added to a CNN,

- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest

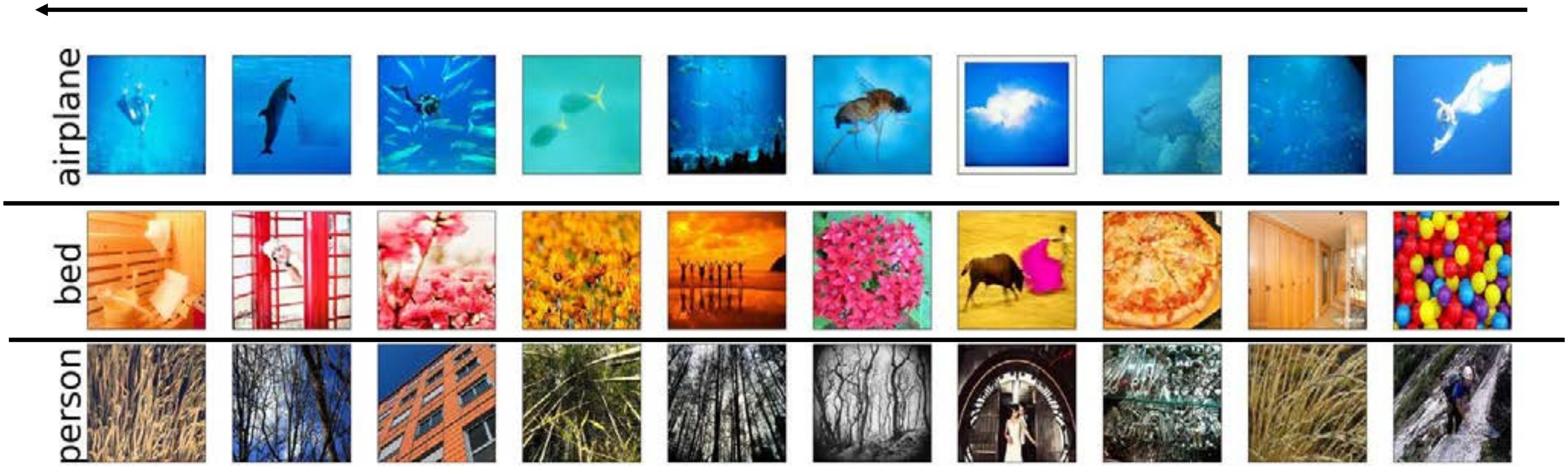


When a CW module is added to a CNN,

- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest

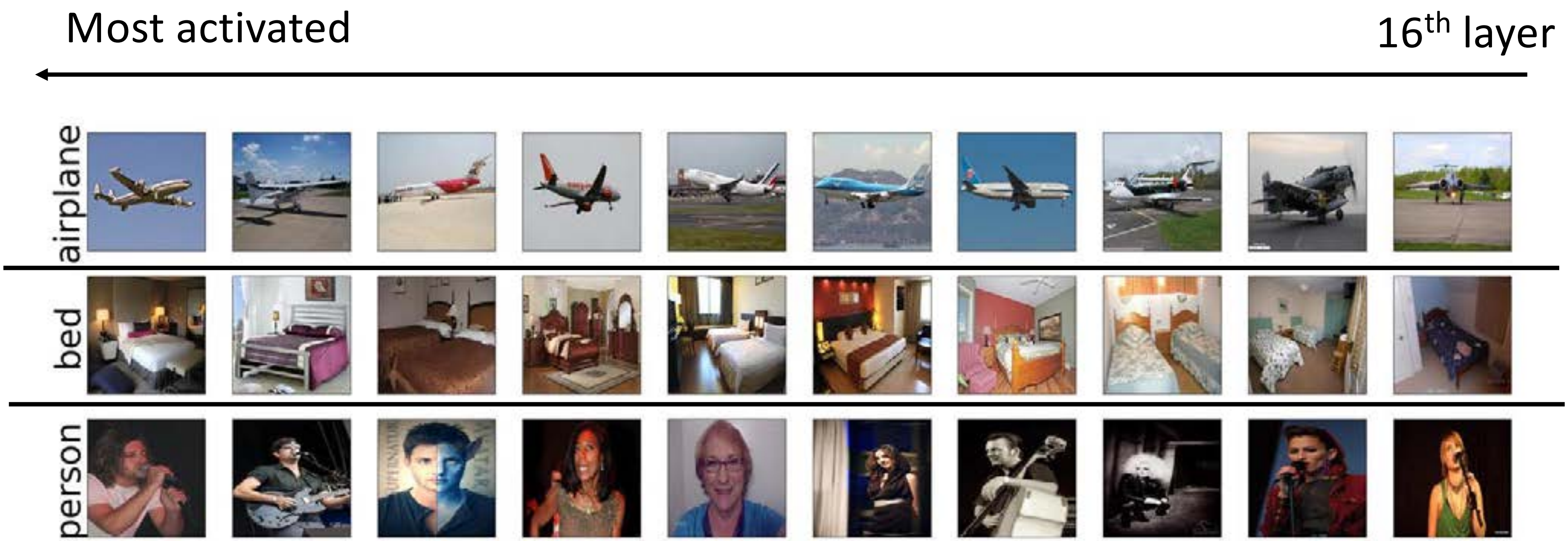
Most activated

2nd layer

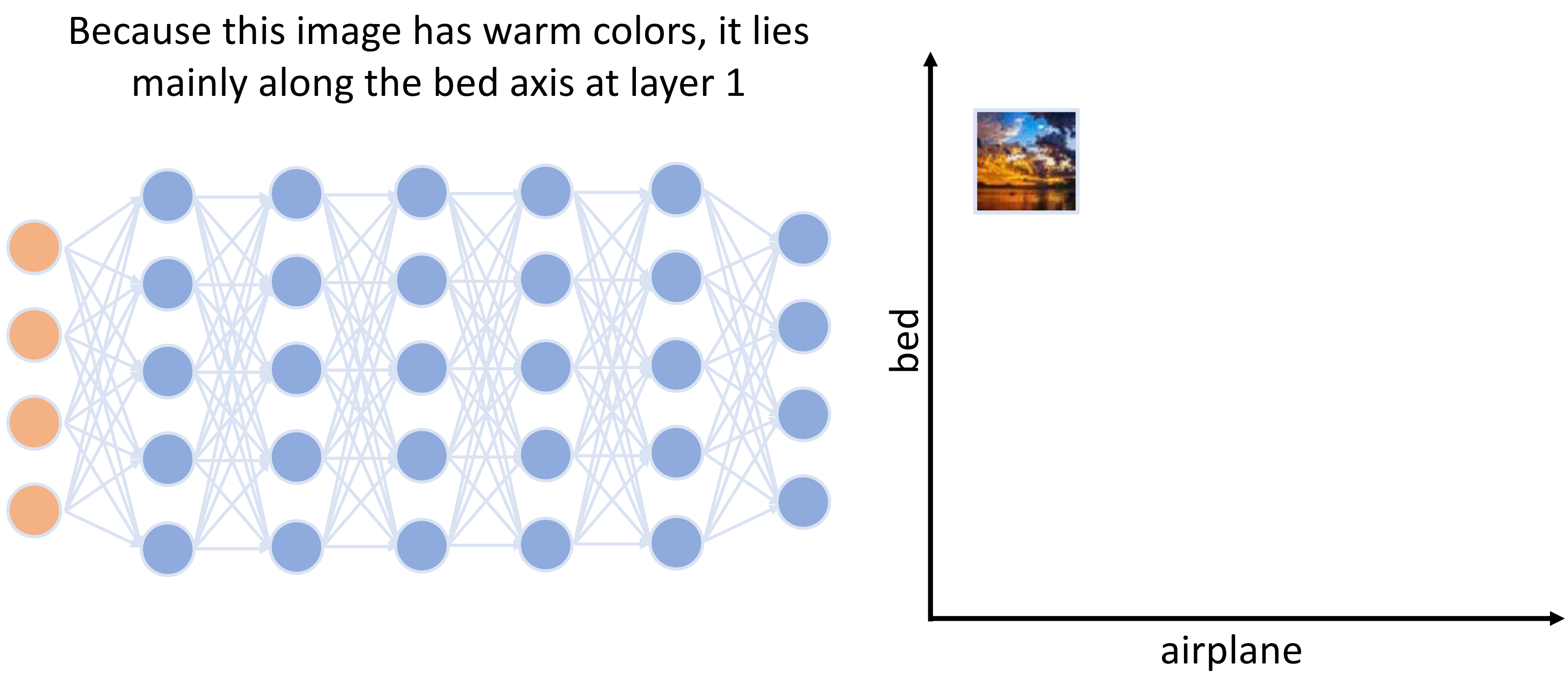


When CW is added to different layers…

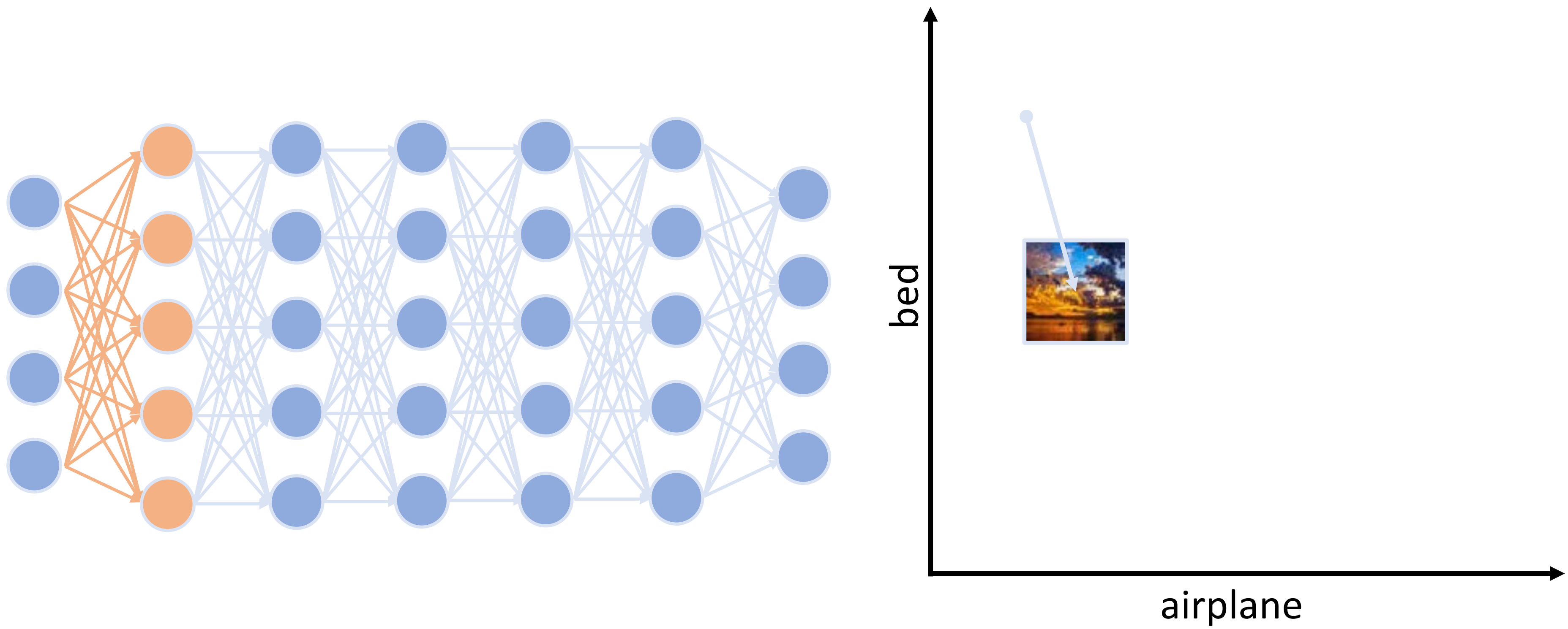
In earlier layers, color and texture information related to the concepts are represented along the axes



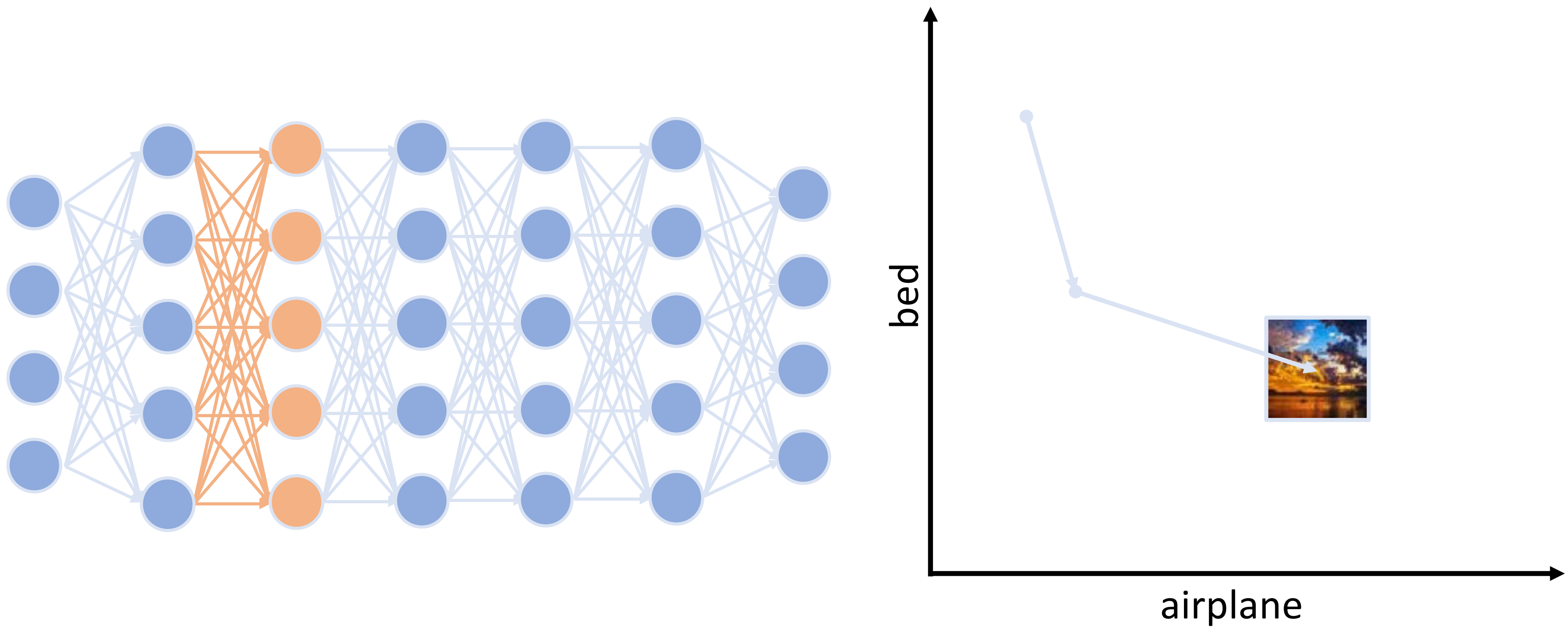
Concept information now lies along the axes.



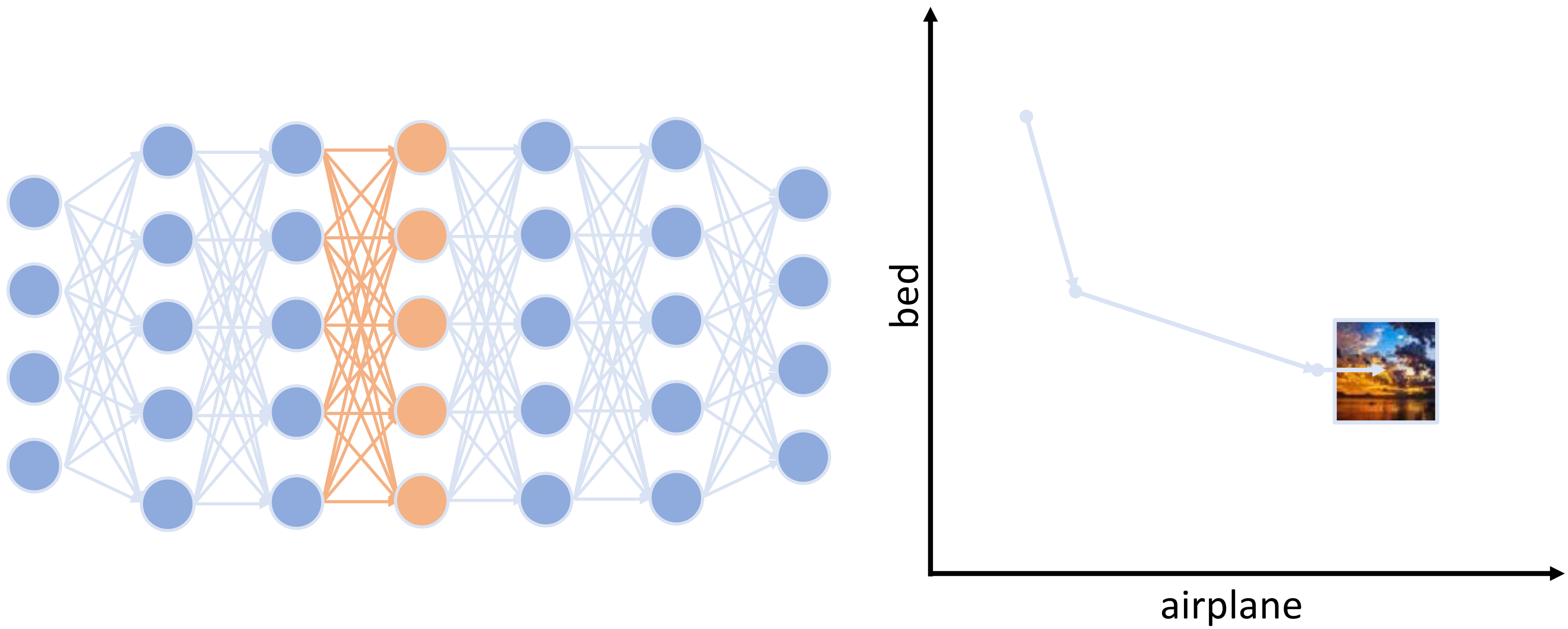
See how an image travels through the layers



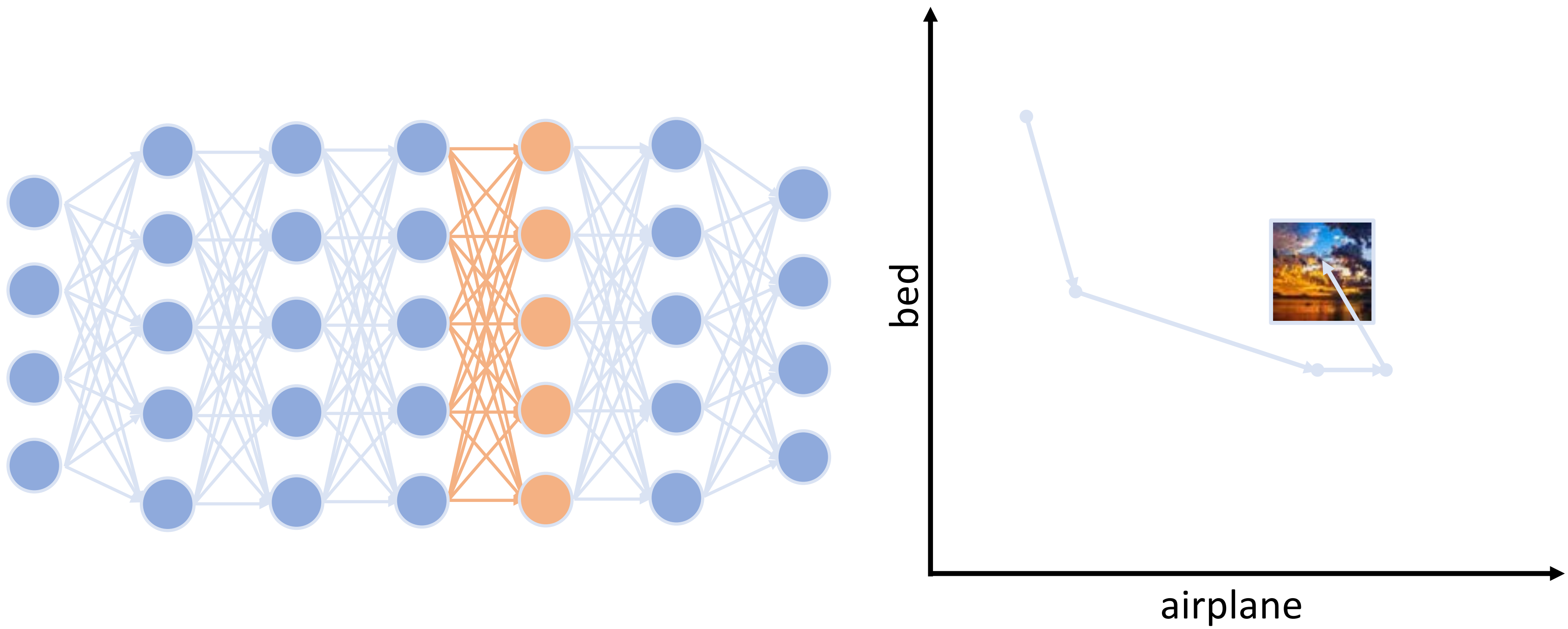
See how an image travels through the layers



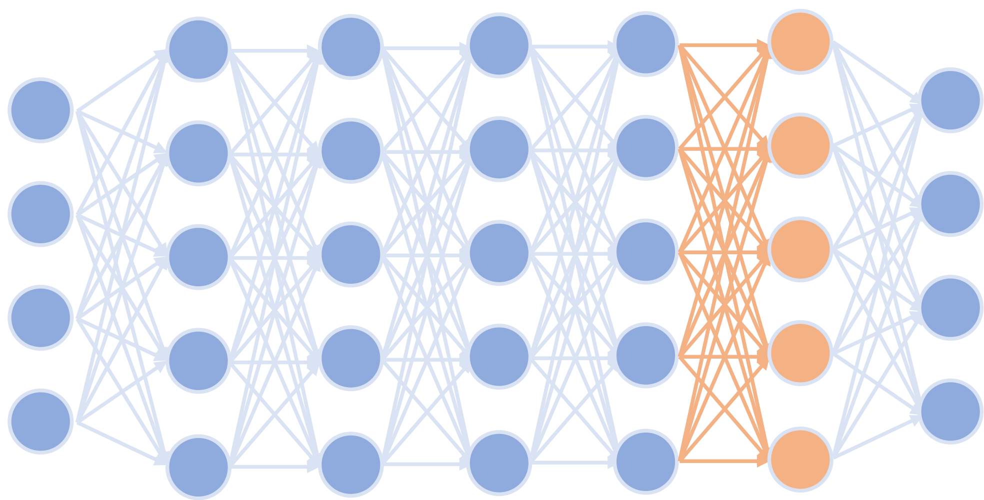
See how an image travels through the layers

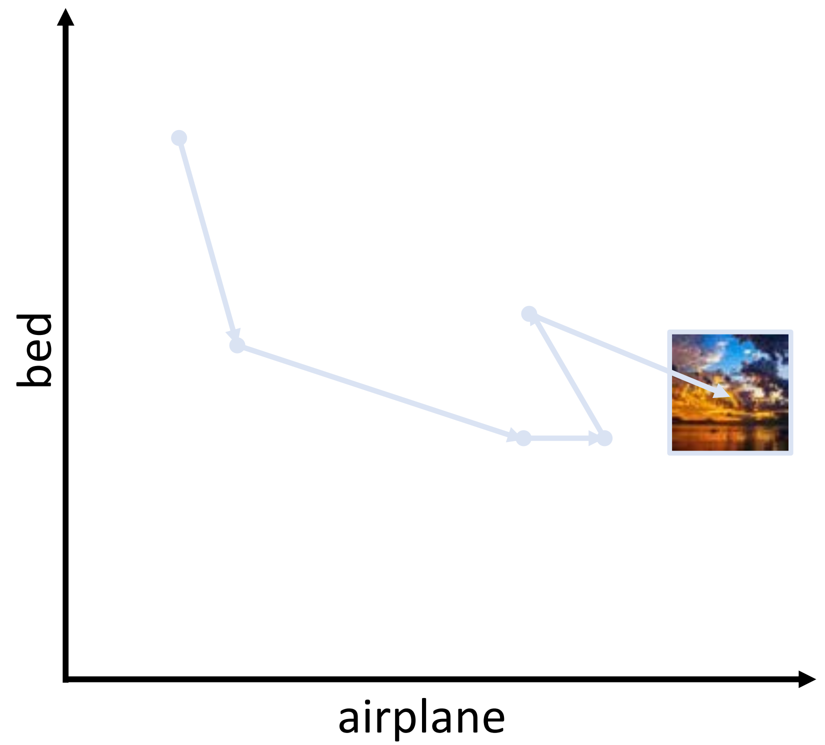


See how an image travels through the layers

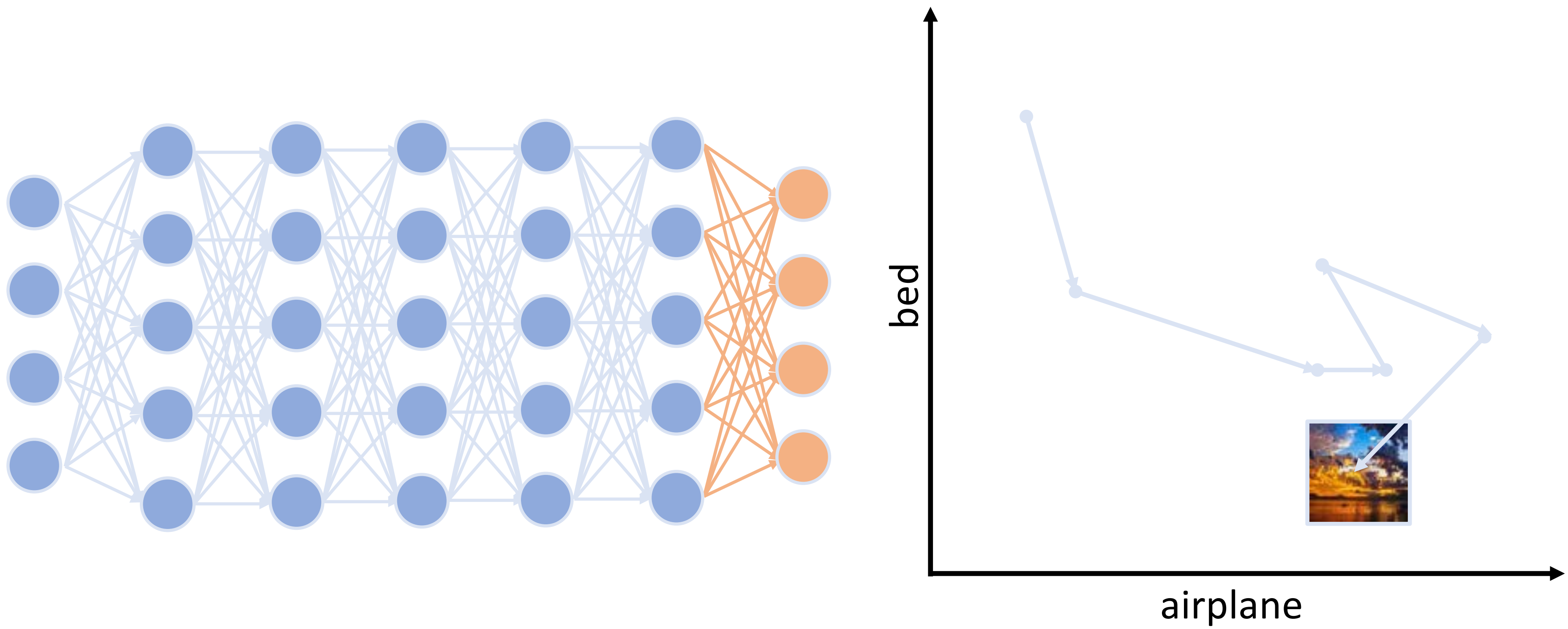


See how an image travels through the layers





See how an image travels through the layers



See how an image travels through the layers

Advantages of CW over BatchNorm

- ➢ No sacrifice in accuracy

  - accuracy is on par with standard CNNs

- ➢ Easy to use

  - warm-start from pretrained model requires only one additional epoch of further training
  - Note: requires training data for the concepts to define the axes

- ➢ Disentangles the latent space

Interpretable deep CNNs for computer vision:

Prototype Network

Case-based reasoning
+
Fine Annotation

Strictly better than saliency

Concept Whitening

Neural
Disentanglement

Strictly better than concept vectors

Take-aways

- There is no scientific evidence supporting a tradeoff between interpretability and accuracy in deep learning.
    - Interpretability helps troubleshoot and helps accuracy

- It is a matter of time until companies try to use black box models for biopsy decisions…

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing prob-

arXiv.org > cs > arXiv:1806.10574

Computer Science > Machine Learning

[Submitted on 27 Jun 2018 (v1), last revised 28 Dec 2019 (this version, v5)]

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, Cynthia Rudin

When we are faced ... and pointing out pr... make our final deci... (ProtoPNet), that rea... evidence from the p... to the way ornitholo... classification tasks. We demonstrate ou... ProtoPNet can achie... ProtoPNets are com... performing deep m... deep models.

arXiv.org > cs > arXiv:2002.01650

Computer Science > Machine Learning

[Submitted on 5 Feb 2020 (v1), last revised 19 Oct 2020 (this version, v4)]

Concept Whitening for Interpretable Image Recognition

Zhi Chen, Yijie Bei, Cynthia Rudin

What does a neural network encode about a concept as we traverse through the layers? Interpretability in mac... unde... later... neur... the r... whit... By e...

arXiv.org > cs > arXiv:2103.12308

Computer Science > Machine Learning

[Submitted on 23 Mar 2021]

IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, Cynthia Rudin

Interpretability in machine learning models is important in high-stakes decisions, such as whether to order a biopsy based on a mammographic exam. Mammography poses important challenges that are not present in other computer vision tasks: datasets are small, confounding information is present, and it can be difficult even for a radiologist to decide between watchful waiting and biopsy based on a mammogram alone. In this work, we present a framework for interpretable machine learning-based mammography. In addition to predicting whether a lesion is malignant or benign, our work aims to follow the reasoning processes of radiologists in detecting clinically relevant semantic features of each image, such as the characteristics of the mass margins. The framework includes a novel interpretable neural network algorithm that uses case-based reasoning for mammography. Our algorithm can incorporate a combination of data with whole image labelling and data with pixel-wise annotations, leading to better accuracy and interpretability even with a small number of images. Our interpretable models are able to highlight the classification-relevant parts of the image, whereas other methods highlight healthy tissue and confounding information. Our models are decision aids, rather than decision makers, aimed at better overall human-machine collaboration. We do not observe a loss in mass margin classification accuracy over a black box neural network trained on the same data.

arXiv.org > cs > arXiv:2103.11251

Computer Science > Machine Learning

[Submitted on 20 Mar 2021]

Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the "Rashomon set" of good models; and (1... Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientis... interested in working in interpretable machine learning.

Review paper

Thanks

All papers are here: https://users.cs.duke.edu/~cynthia/