Concept Bottleneck Models

Pang Wei Koh | Thao Nguyen | Yew Tang | Steve Mussmann | Emma Pierson | Been Kim | Percy Liang

Presenter - Aakash Agrawal

Background

Pang Wei Koh

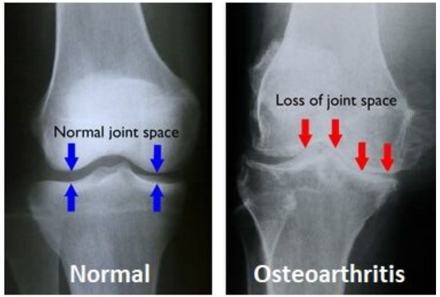
Percy Liang

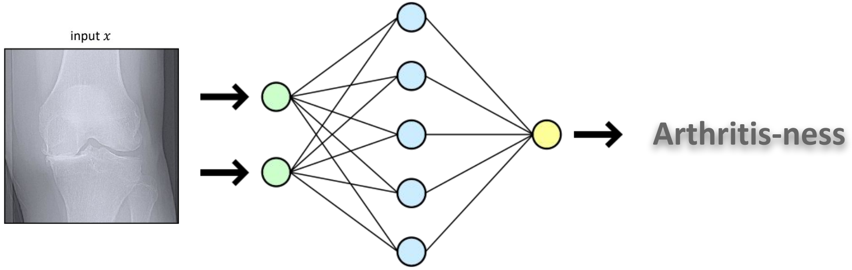
- Reliable Machine Learning
  - Better understanding and adapting foundational models
  - How do we make models more reliable and trustworthy?
- Building foundational models from the first principles
  - How do we attribute model predictions back to training data[1]?

Motivation

Arthritis Severity

- Indicators:
  - Joint space
  - Bone spur
  - Calcification
  - ....



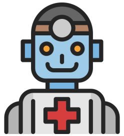


End-to-End model

End-to-end Models

input x

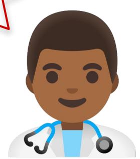
uniteractive



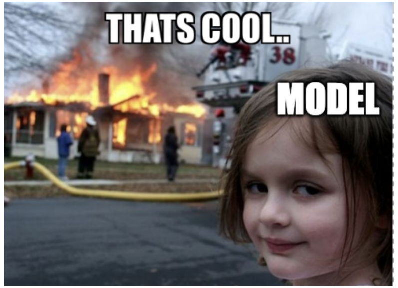
End-to-End model

No Arthritis

Looks like arthritis

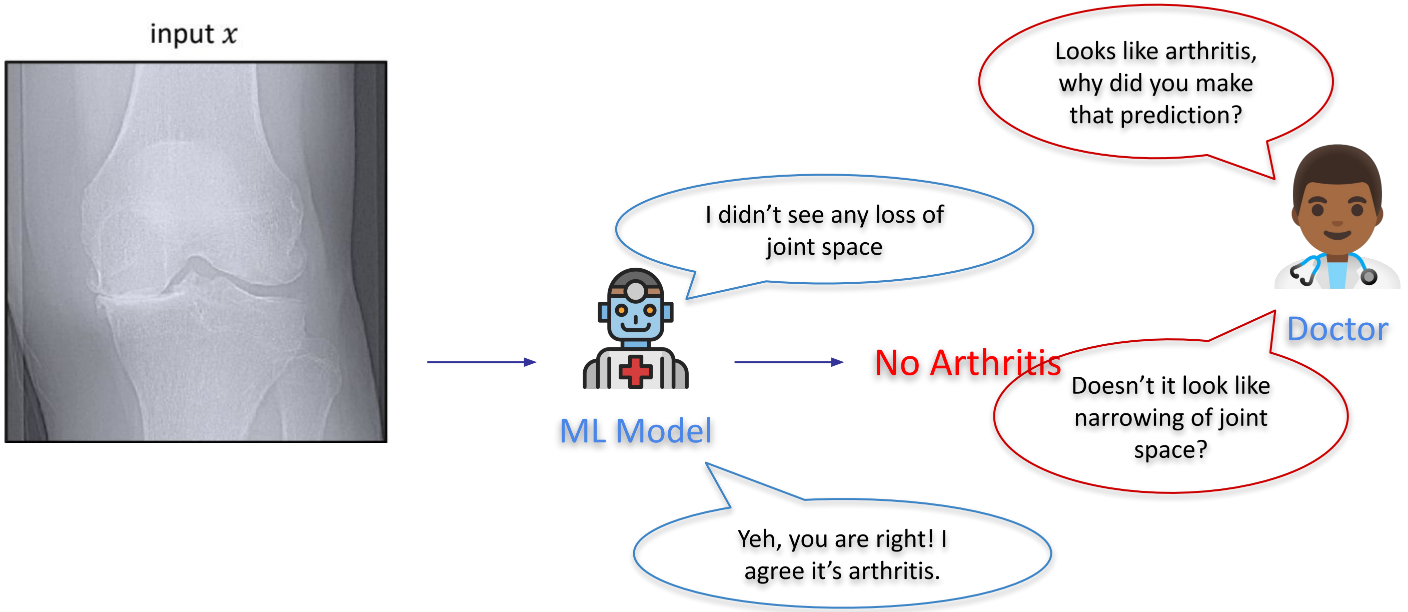


Doctor



Dataset: https://nda.nih.gov/oai (Osteoarthritis Initiative)

Ideal: Interact via high-level concepts

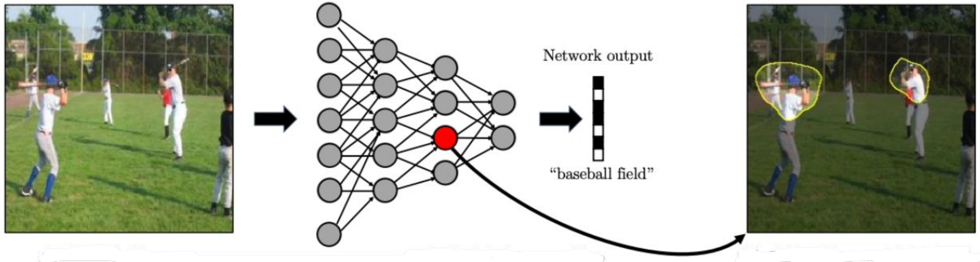


Prior Works using high-level concepts

Post-hoc

Network Dissection **Dau et. al.** 2017

- **Goal**: quantifying the interpretability of individual neurons or channels in CNN
- Use a large, densely labeled dataset of visual concepts (objects, parts, textures, colors, etc.) and measure **correlation** of concepts with individual neurons



Concept is not a part of the training!!

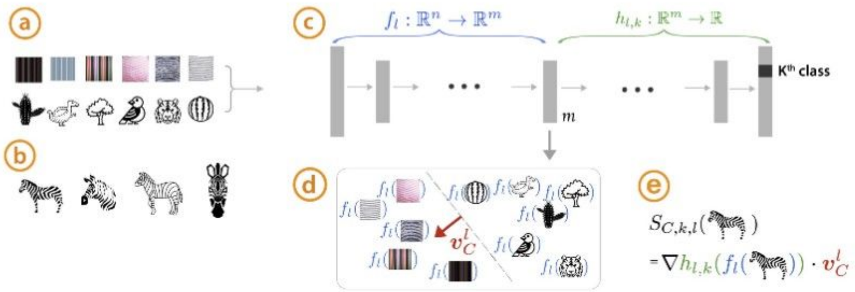
https://netdissect.csail.mit.edu/

Prior Works using high-level concepts

Post-hoc

TCAV - Testing with Concept Activation Vectors **Kim et. al.** 2018

- **Goal**: Analyse how much a concept (eg. bone spur) was important for a prediction
- **Step 1**: Train a linear classifier to distinguish activations of concept examples vs. random examples.
- **Step 2**: These linear classifiers produce **concept activation vectors** that can measure how sensitive the model's predictions are to changes in the concept.



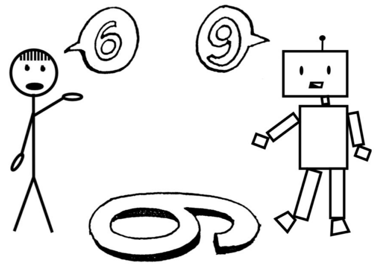
Concept is not a part of the training!!

https://arxiv.org/abs/1711.11279

Prior Perspectives

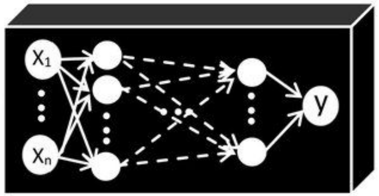
ML models are difficult to interact with

- direct interventions not possible



Trade-off between interpretability and predictive accuracy

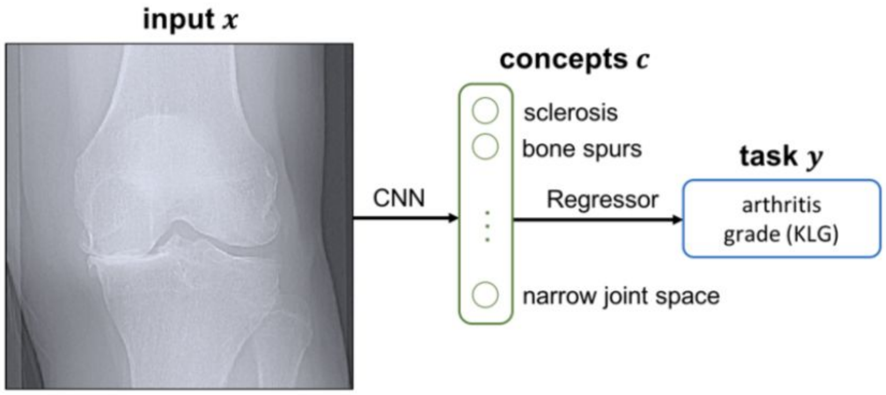
High-level concepts only meant for post-hoc interpretation



Neural Networks are black boxes, lacking transparency

Concept Bottleneck Models

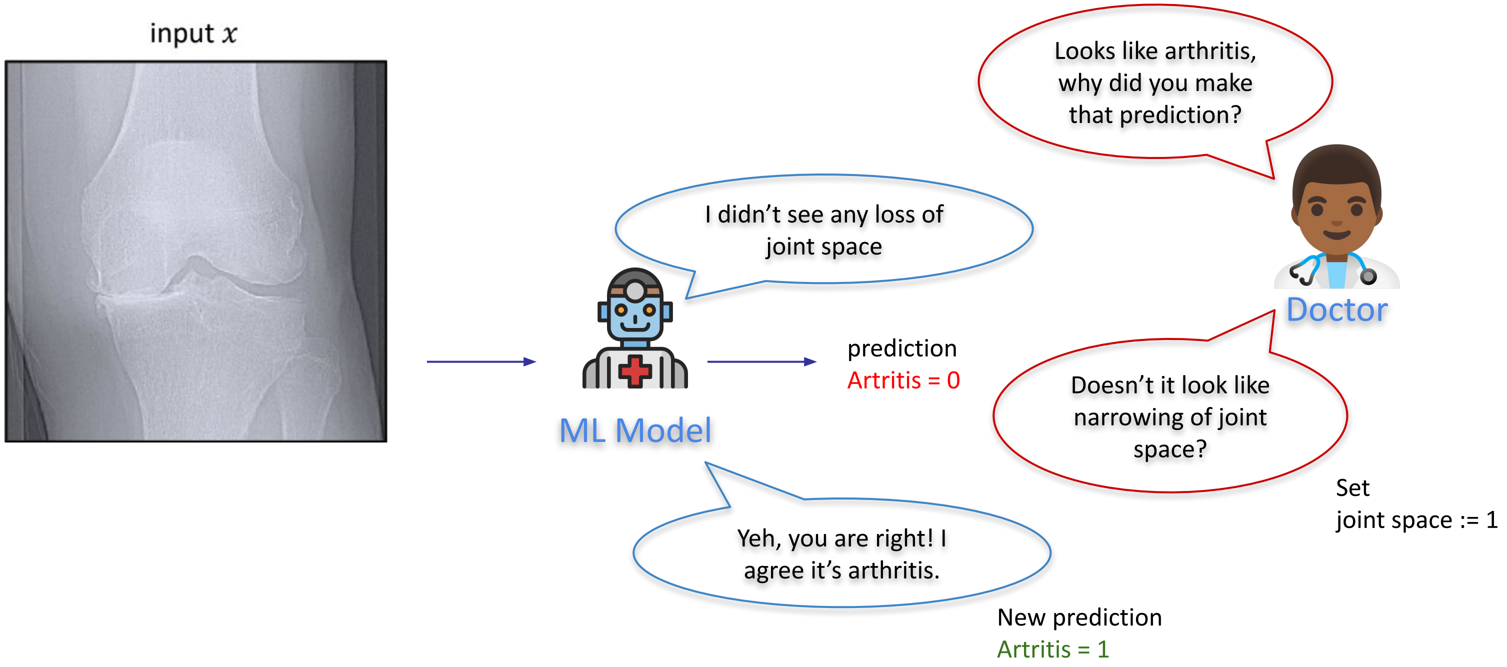
*Use concept **explicitly** as a part of the training!!*



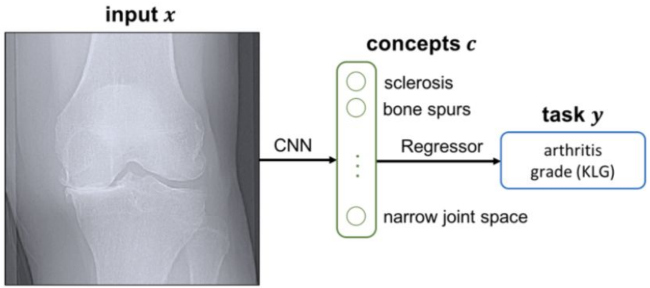
- Interpretability
- Predictability
- intervenability

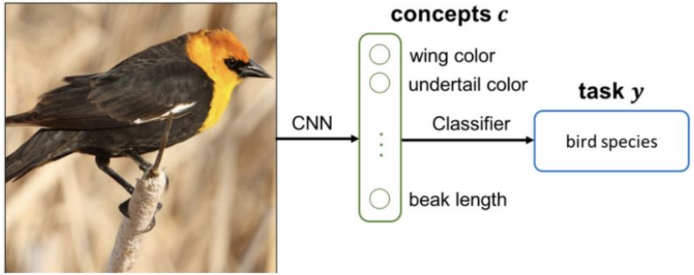
Incorporate pre-defined, high-level concepts into the learning procedure

Interacting via high-level concepts



Training Regimen





Training Methods	Representation
Independent	$x \rightarrow c \;\&\; c \rightarrow y$
Sequential	$x \rightarrow c \textit{ then } \hat{c} \rightarrow y$
Joint (λ > 0)	$x \rightarrow c \rightarrow y$
Standard (control, λ = 0)	$x \rightarrow y$

$$\mathbf{L} = \arg\min_{f,g} \sum_i \left[L_Y(f(g(x^{(i)})); y^{(i)}) \; + \; \sum_j \lambda L_{C_j}(g(x^{(i)}); c^{(i)}) \right]$$

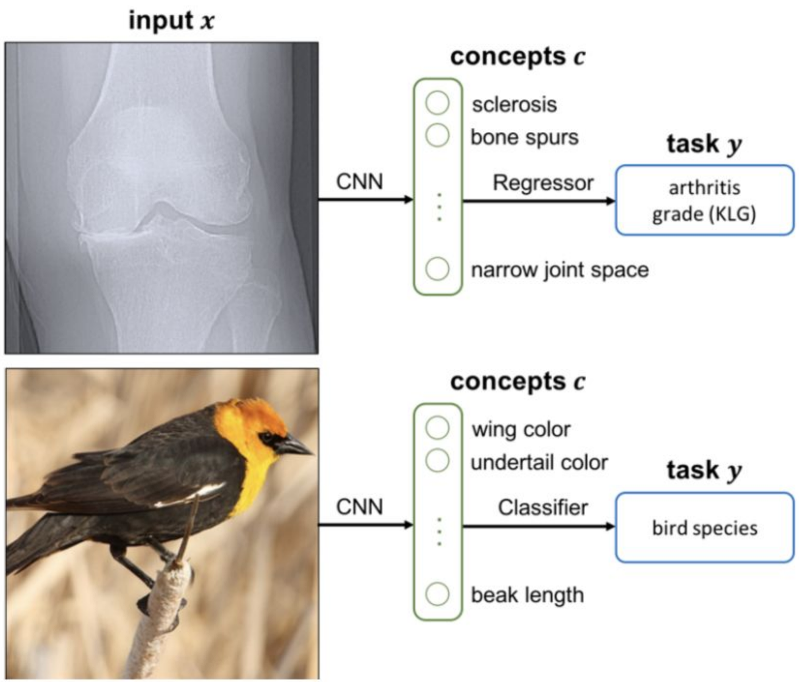
λ = low → prioritise labels

λ = high → prioritise concepts

Data Sources and Task

- X-ray grading (OAI)
  - **Task**: Regression. Given an x-ray image, predict the KLG grade (4 values, higher grades indicating more arthritis severity).
  - Instance level concepts - k = 10 (bone spur, joint space, calcification, etc.).

- Bird Identification (CUB)
  - **Task**: Classification. Given a bird image, classify it into correct bird species (200 bird species).
  - Class level concepts - k = 112 (wing color, beak shape, etc.).



Results

- Label (Task) Accuracy
  - Competitive accuracy with standard e2e models
- Concept Accuracy
  - Better compared to SENN and post-hoc analysis methods like TCAV

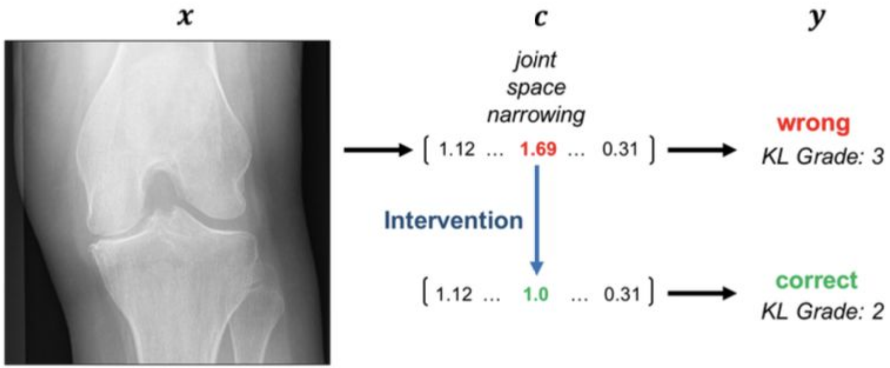
No Interventions

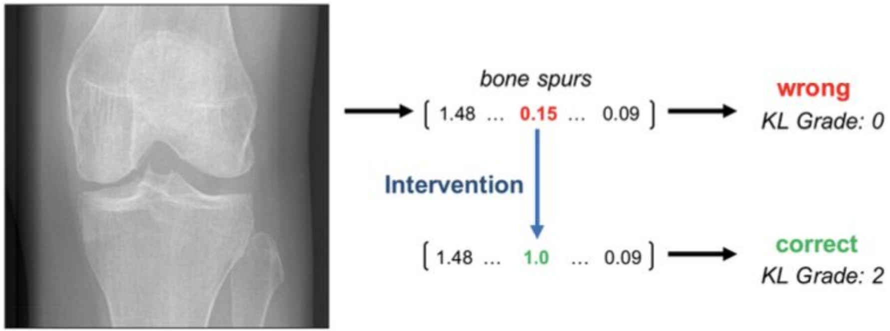
		X-rays (y RMSE)	X-rays (c RMSE)	Birds (y error)	Birds (c error)
Concept bottleneck models	Independent	0.44	0.53	0.24	0.03
	Sequential	0.42	0.53	0.24	0.03
	Joint	0.42	0.54	0.20	0.03
	Standard (no concepts)	0.44	--	0.18	--

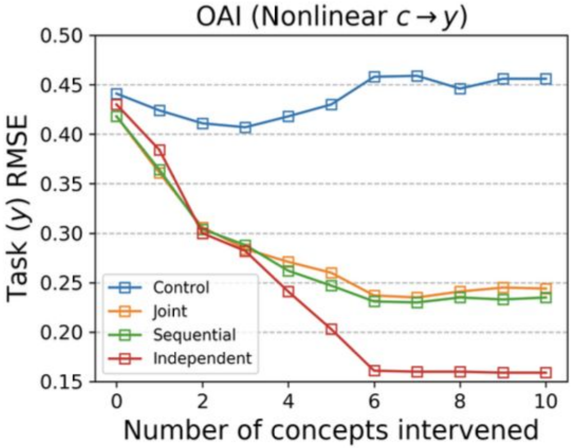
SENN - https://people.csail.mit.edu/davidam/docs/SENN.pdf

TCAV - https://arxiv.org/abs/1711.11279

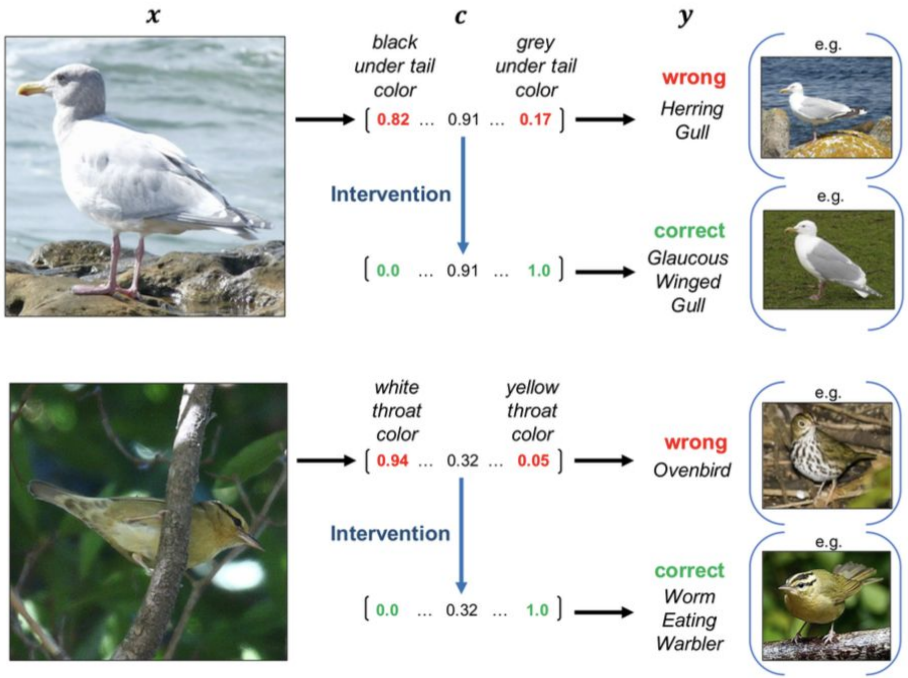
Interventions (OAI)

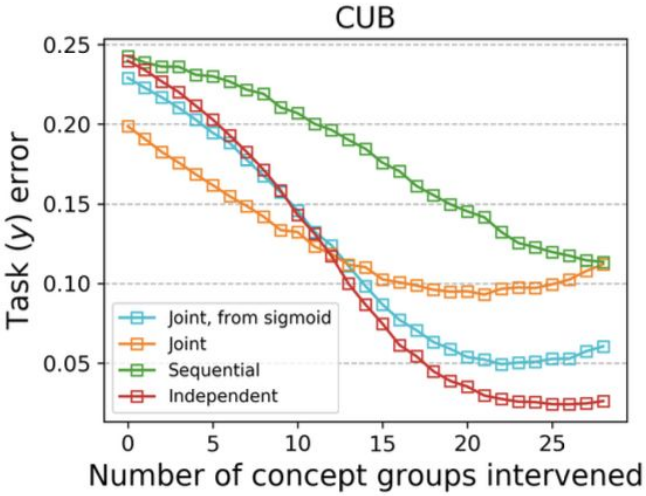






Interventions (CUB)





Intervene on concept groups

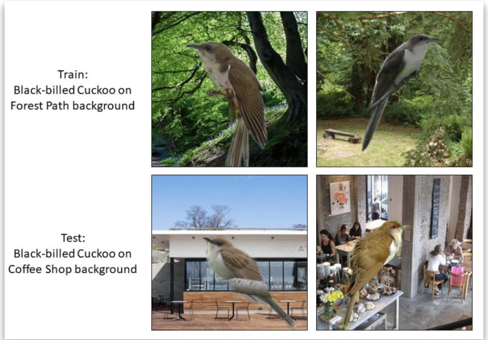
Strengths

- Incorporating the goals of **interpretability**, **predictability**, and **intervenability** in a single model, which is quite critical for high stakes environments like medicine.

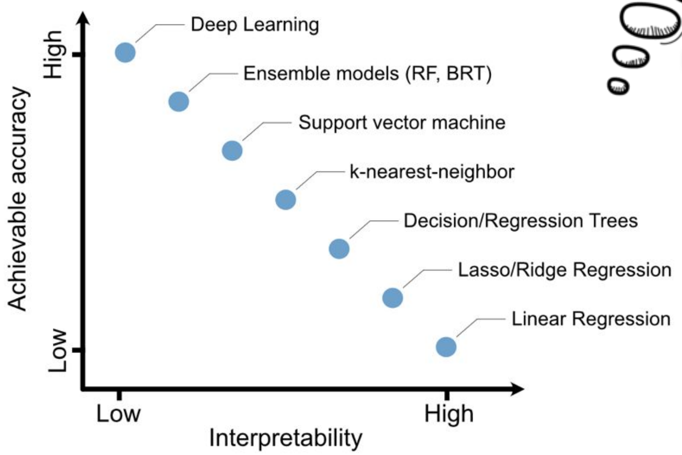
- Any Neural Net model can be adapted to a CBM and exploit the benefits they offer.

- Facilitates human-model interaction via **interventions** while achieving **performance** superior to either humans or the model alone.

- CBMs are more **robust** to background and covariate shifts.



Tradeoff between Interpretability & Accuracy?



Do we need to trade?

Concept Bottleneck Model

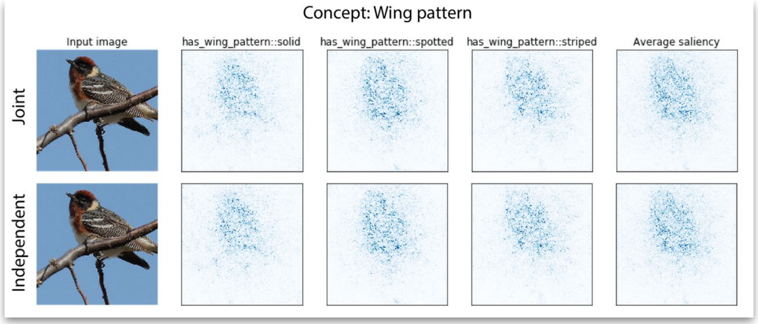
Do Concept Bottleneck Models learn as intended?

- Concepts might not align with meaningful input space representations
  - No meaningful mapping of inputs to concepts

- There is **no concept bottleneck**. Concept predictor also encodes class label in Joint CBMs.
  - **Corruptions** in the concept predictions.

- **Intervention breaks** - using incorrect concept predictions to predict the correct output.

- Interpretability is an **illusion**



LABEL LEAKAGE !!

Independent bottleneck 

- Interpretability
- Predictability
- Intervenability

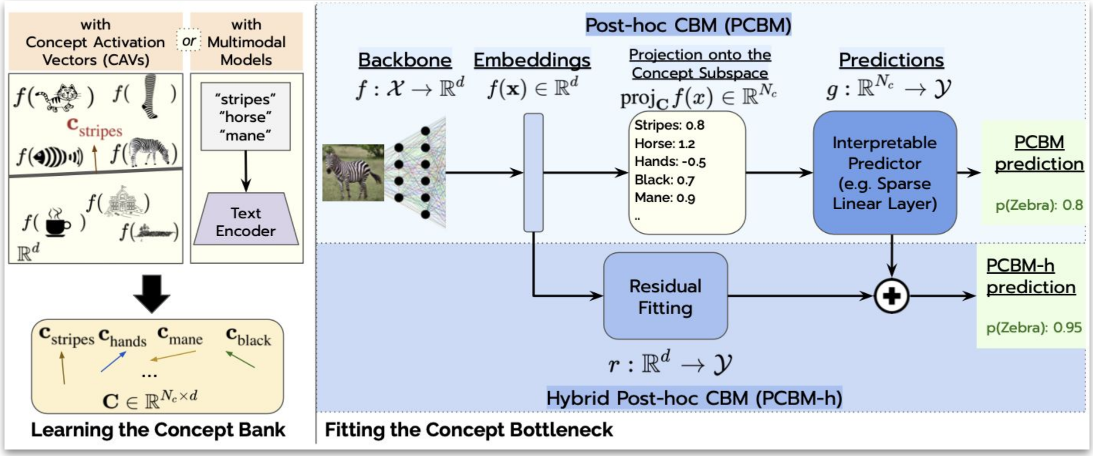
*Margeloiu, Andrei, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik and Adrian Weller. "**Do Concept Bottleneck Models Learn as Intended?**"* https://arxiv.org/pdf/2105.04289 (ICLR 21)

Limitations

- Information Leakage
- Interventions doesn’t work
- Each prediction needs to be reviewed and monitored manually
- Dense Concept Annotations
  - Lack of datasets with concept annotations
  - Annotations are costly

- Potential performance drops as compared to end-2-end NN
- Global Interventions

Post-hoc concept bottleneck models (P-CBMs)

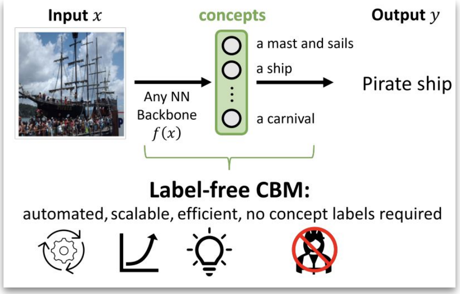


- Use concepts from other datasets, natural language descriptions of concepts via multimodal models
- Allows to change the model behavior via global edits
- Makes CBMs more accessible and expressive in different settings
- When concepts are weak, use Residual model = Blackbox unrestricted model

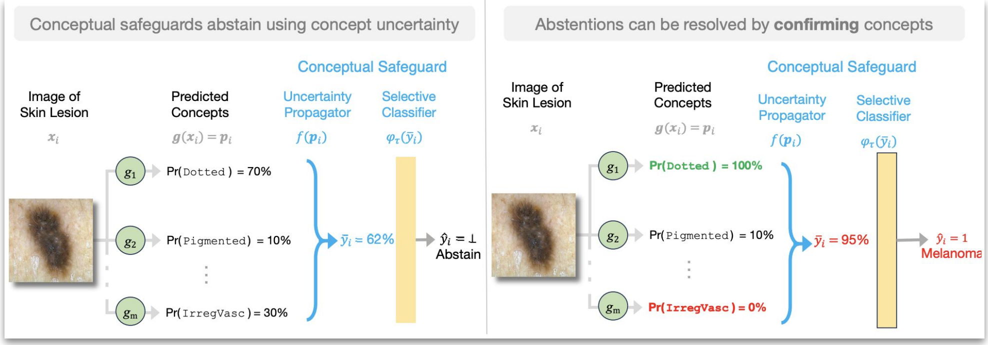
Yuksekgonul, Mert, Maggie Wang and James Y. Zou. "**Post-hoc Concept Bottleneck Models.**" https://arxiv.org/abs/2205.15480 (2022)

Related Works

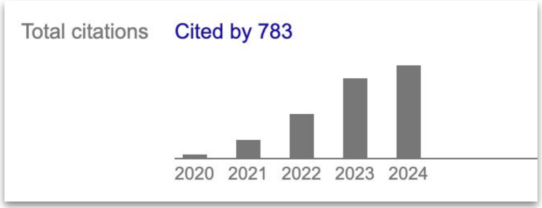
- Oikarinen, T., Das, S., Nguyen, L.M. and Weng, T.W., 2023. ***Label-free concept bottleneck models***. arXiv preprint arXiv:2304.06129. https://arxiv.org/pdf/2304.06129



- Joren, H., Marx, C.T. and Ustun, B., 2023. ***Classification with Conceptual Safeguards.*** In The Twelfth International Conference on Learning Representations. https://openreview.net/pdf?id=t8cBsT9mcg



Citations



- Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C. and Yatskar, M., 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19187-19197).
- Wang, B., Li, L., Nakashima, Y. and Nagahara, H., 2023. Learning bottleneck concepts in image classification. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 10962-10971).
- Oikarinen, T., Das, S., Nguyen, L.M. and Weng, T.W., 2023. Label-free concept bottleneck models. arXiv preprint arXiv:2304.06129.
- Havasi, M., Parbhoo, S. and Doshi-Velez, F., 2022. Addressing leakage in concept bottleneck models. Advances in Neural Information Processing Systems, 35, pp.23386-23397.
- Kim, E., Jung, D., Park, S., Kim, S. and Yoon, S., 2023. Probabilistic concept bottleneck models. arXiv preprint arXiv:2306.01574.

Key Takeaways

- CBMs getting a lot of traction
  - Many limitations in the current state
  - Continued research into improving CBMs

- CBMs look promising in high-stakes environments like medicine. Incorporating the goals of **interpretability**, **predictability**, and **intervenability** in a single model.
- Incorporates pre-defined, **high-level concepts** into the model itself for interpretability as compared to post-hoc analysis.
- Enables better human-machine interaction and trust by allowing experts to intervene

GQs

- Should there be a metric to quantify the **quality** of concepts as well?
- Can CBMs become a **goto** choice of models in a deep learning problem setting? What are the challenges in automating concept extraction for CBMs in domains with **unstructured** data?
- How can CBMs handle situations where concepts themselves may be **noisy** or ambiguous?
- Can CBMs be adapted to **dynamic** environments where concepts may evolve over time?
- How do CBMs handle scenarios where certain concepts are inherently **subjective** or culturally influenced?