Critical Review: "Concept Bottleneck Models" by Koh et al.

Aakash Agrawal
A69034394
aaa015@ucsd.edu
University of California San Diego

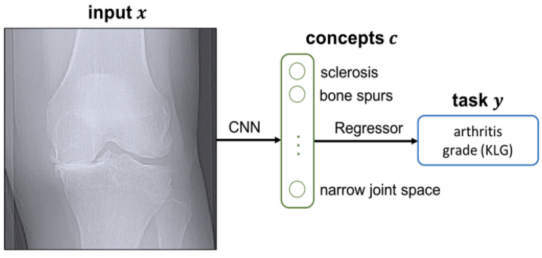
Paper: https://arxiv.org/abs/2007.04612

A. **Executive Summary**

The gist of the paper is that the authors aim to make foundational models like neural networks more accessible and easier to understand. To do this, the authors revisit the concept bottleneck model (CBM), which lets practitioners use their domain knowledge to integrate high-level concepts into the learning process. Although concept bottleneck models require dense concept annotations for training, they offer many advantages of intervenability and interpretability while also attaining a high predictive accuracy.

The authors propose three training methods for learning CBMs: Independent, Sequential, and Joint, which mainly differ in their formulation of loss function. They suggest that any neural network can be converted into a CBM by resizing an intermediate layer, allowing it to leverage the benefits of CBMs. To illustrate this, they present two case studies: one for regression and another for classification, demonstrating CBMs' predictive performance and interpretability. The authors **claim** that CBMs deliver competitive accuracy compared to unrestricted end-to-end neural networks.

The authors then discuss the trade-off between task accuracy and concept accuracy, noting that as the number of interventions increases, task error tends to decrease. This is particularly evident in the case of Independent CBMs, which rely on true concepts for training the frontend. As a result, during test-time interventions, Independent CBMs do not experience distribution shifts and see a more significant reduction in error as compared to Joint and Sequqential CBMs. Finally, they present an experiment demonstrating the robustness of CBMs to background and covariate shifts compared to other end-to-end models, emphasizing the many advantages of using CBMs.



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

Fig 1 left: CBMs incorporate pre-defined, high-level concepts into the learning procedure. Fig 1 right: different training regimens of CBM.

B. **Relationship to other papers and ideas discussed in class**

I would like to establish some key connections to a few other papers discussed in the class. For each of these papers, firstly, I provide a brief gist of the underlying idea and then showcase how CBMs relate to the idea.

- **Principles of Explanatory Debugging to Personalise Interactive Machine Learning** by Kulesza et al. Paper: https://dl.acm.org/doi/10.1145/2678025.2701399

  The paper talks about the idea of enabling users to better control the models they are building, an idea that closely resonates with CBMs. Explanatory debugging allows for two-way interaction and exchange of information between end users and ML systems. Although the paper uses the Multinomial Naive Bayes (**MNB**) algorithm, we can easily replace these with **CBMs**, and all the functionalities still hold.

  The paper outlines several principles of explanatory debugging, grouped under the broader concepts of Explainability and Correctability. Here's how CBM aligns with these principles:

  Explainability Principles:
  - **Be iterative, Be Sound**: We can disclose the activations of the Concept Bottleneck Layer (CBL) that are used to make final predictions to the users; this assists users in interacting with the model and building better mental models of their own.
  - **Be complete**: Similar to **EluciDebug**, we can showcase all of the concepts available to the model for making predictions.
  - **Don't Overwhelm**: To avoid overwhelming, when dealing with a large number of concepts, we can use the top sets of concepts the model is using to make predictions. If the user is interested, they should be able to see and modify other concepts as well.

  Correctability Principles:
  - **Be actionable:** CBM lets users add, remove, or modify concepts, allowing users to input their intuition into the model.
  - **Be reversible**: users can revert the concept which would allow for the correction of mistakes (modifications making the prediction worse).
  - **Honoring user feedback:** CBM allows users to give concept-level feedback and lets the model use this updated info for its final predictions.

  Hence, CBMs clearly fit and relate to the ideas and Principles of Explanatory Debugging, allowing practitioners to personalize their ML systems.

- **Do users benefit from Interpretable Vision?** by Sixt et al.
  Paper: https://arxiv.org/abs/2204.11642

  In this paper, the authors explore the idea of simply using model predictions for interpretability as opposed to applying interpretable ML methods. They note that automatically discovered concepts tend to perform worse than the baseline (which compares different inputs and outputs), mainly due to confusion caused by spatially overlapping features in explanations based on these concepts. Here, **CBMs** can provide a compelling alternative as they encode concepts (leg position, color, shape, rotation, etc) in a distinct, non-overlapping way, enabling "**counterfactual explanations**". By changing these concepts or characteristics from peaky to stretchy, the system can help gain insights into which characteristics are relevant to the system.

- **Do Feature Attribution Methods Correctly Attribute Features?** by Zhou et al. Paper: https://arxiv.org/abs/2104.14403

  Throughout our course, we came across many explanation methods that were shown to be affected by spurious correlations. Similarly, this paper highlights how attribution methods like **Saliency Maps**, **Attention mechanisms,** and **Rationale models** struggle with robustness and performance. CBMs present an interesting case in this setup, given their robustness to spurious correlations and covariate shifts. It will be interesting to see how they perform using the **dataset modification technique** for evaluating feature attributions.

Additionally, both Prof Lipton in The Mythos of Interpretability and Prof Rudin in Stop explaining black box models, critique post hoc explanation methods, advocating for the use of inherently interpretable models instead. In this context, Concept Bottleneck Models offer an intriguing case for exploration.

B.1. **Contributions of CBM to the interpretability/explainability literature**
CBMs present a completely different class of methods in the realm of interpretability and explainability. They bust the prior myths around deep learning models being mere black boxes and lacking transparency. They also challenge the idea that high-level concepts are only useful for post-hoc interpretation. Additionally, they push back against the notion that machine learning models are hard to interact with. The authors also showcase that the notion of trade-off between interpretability and predictive accuracy is flawed, and it is possible to achieve these competing objectives in a single model.

B.2. **Why was it important for us to read this paper?**
Yes, I believe this was a valuable paper to include in the syllabus. It exposes students to the latest trends in interpretable machine learning, specifically introducing them to a class of techniques known as inherently interpretable models. These models offer a distinct approach compared to the more commonly used post-hoc explanations and other data interpretation methods like exploratory data analysis (EDA).

C.1. **Key Strengths**

1. The paper is the first of its kind to attempt to combine the goals of **interpretability**, **predictability**, and **intervenability** in a single model, which is crucial for many high-stakes scenarios in healthcare, finance, and many other fields.
2. The applications demonstrated (OAI, CUB) are thorough and seem convincing, clearly demonstrating the usefulness of the proposed model. The paper provides a concise case study showcasing the **robustness** of CBMs to spurious correlations like background shifts. CBMs also enable **counterfactual explanations** such as *if the model did not think the joint space was too narrow, would it have predicted severe arthritis?*
3. Any deep learning model can be transformed into a Concept Bottleneck Model and exploit the benefits they offer simply by resizing one of the intermediate layers to match the number of concepts and adding a loss to facilitate learning of that layer.
4. Facilitates **human-model interaction** via interventions while achieving performance superior to either humans or the model alone.
5. CBMs can aid in model **debugging** by allowing the inspection of intermediate concepts before the final prediction. This enforces trust by ensuring that model behavior aligns with the domain knowledge of experts and their expectations.
6. CBMs corroborate data science **ethics** and **fairness** by offering transparent, human understandable reasoning for their decisions, which is crucial for ethical use.

C.2. **Key Weaknesses**

1. CBMs require densely **annotated concepts** and labels, which are very difficult to procure and are a very **expensive** prospect. Additionally, there are only a few datasets with high-quality concept annotations, which limits their usability and adoption.
2. Each prediction requires **manual review** and monitoring, making test-time interventions a very time-consuming process.
3. CBMs allow interventions for only a single input at a time, known as **local interventions**. They don't support global interventions, which would involve changing the model's behavior by editing the model itself.
4. Empirical results fail to support the claim that CBMs can achieve task accuracies comparable to or better than standard models, undermining their appeal for high-stakes applications.
5. Does not talk about the **scalability** of CBMs. For example, in cell-type classification, a good set of concepts might be very large. This can pose challenges for both interpretability and test-time intervenability.
6. No meaningful mapping of inputs to concepts: Margeloiu et al., in their paper "**Do Concept Bottleneck Models Learn as Intended?**" used saliency maps to show that concepts in a CBM might not align with meaningful input space representations.
7. **Information Leakage**: Since the concept predictor (aka backend) also encodes/learns class labels in the case of the Joint and Sequential CBMs, this leads to corruptions in concept predictions. Hence, there is **no bottleneck.**
8. **Interventions don't work**: This information (or label) leakage implies that CBMs are using incorrect predictions to predict the correct output. Hence, interventions also break.

Margeloiu et al. assert that ***Independent CBM*** might be the only variant of CBM that caters to all the goals of **interpretability**, **predictability**, and **intervenability**.

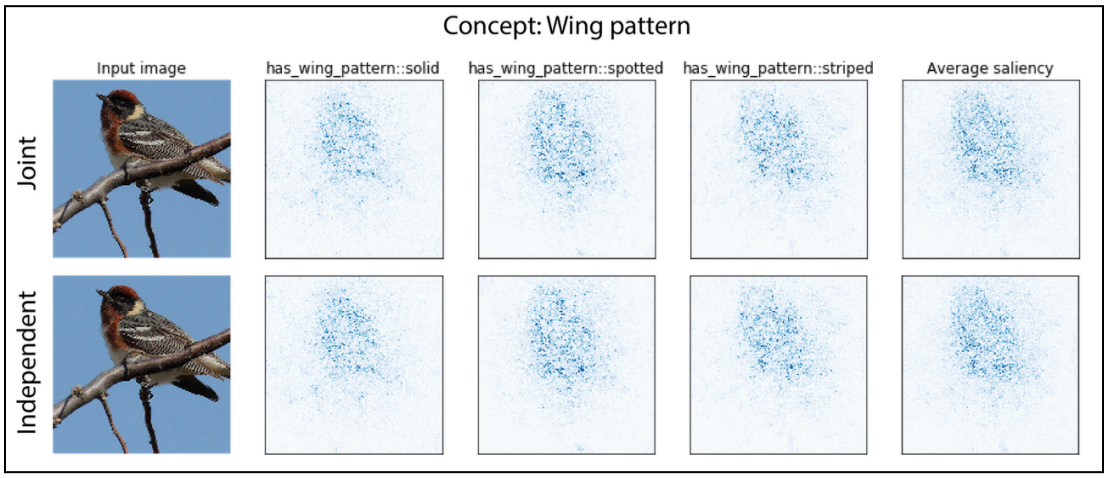
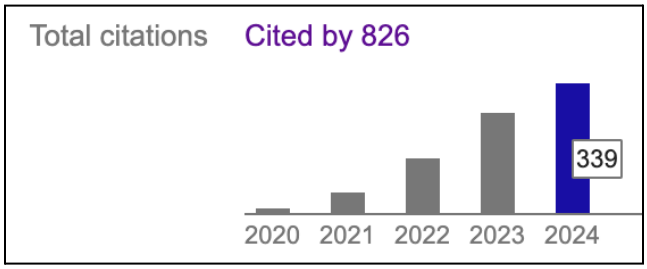


Fig 2: Saliency Map output of the joint and independent CBMs for the concept "wing pattern." The concepts attend to the entire bird image.

D. <u>Impact & Significance in the IML Domain</u>

The paper has received **826** citations as of today, reflecting its substantial influence in the field.



- (🚀) This work has become a **launchpad** for recent advances in interpretable ML research, particularly in the development of inherently interpretable models.
- (🤝) CBMs are getting a lot of traction, with ongoing efforts to address the existing limitations, which include label leakage, costly annotations, etc. They are also increasingly being integrated across various fields in data science, including NLP, LLMs, and Computer Vision.
- (⚖️) CBMs are gaining attention in the context of **data science ethics** and **fairness**, offering a transparent and accountable approach to model decisions, which is crucial for ethical applications in areas like healthcare and finance.

I highlight several prominent works that tackle the limitations of CBMs and explore some areas in data science where CBMs are being increasingly adopted:

D.1. **Efforts to Address CBM Limitations**

1. **Label-free concept bottleneck models** by Oikarinen et al. address the challenge of **costly concept annotation** by using the **GPT-3** model to generate concepts associated with each output class and uses OpenAi's **CLIP** model to learn the concept bottleneck layer. Paper: arXiv:2304.06129.

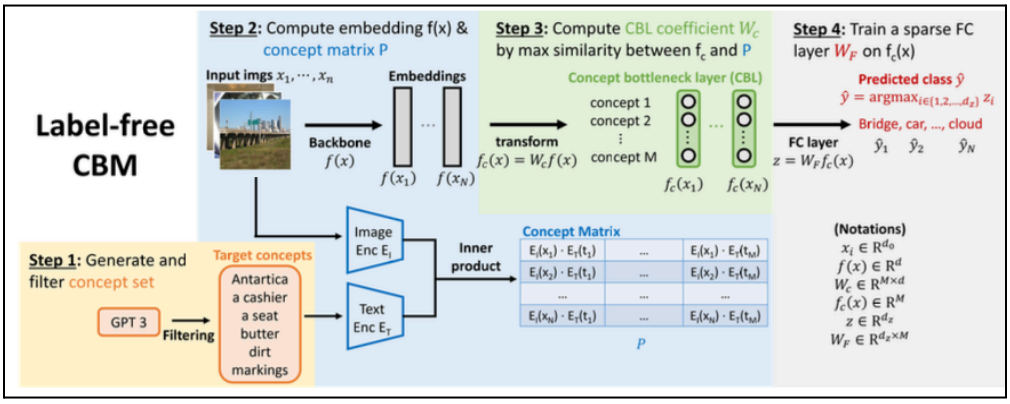


Fig 3: Label Free CBM architecture.

2. **Addressing leakage in concept bottleneck models** by Havasi et al. address the issue of information leakage in the CBMs. Paper: https://openreview.net/forum?id=tglniD_fn9

3. **Classification with Conceptual Safeguards** by Joren et al. requires human intervention only to confirm uncertain concepts when the model abstains. This approach addresses the issue of needing manual review and monitoring for every prediction. Paper: https://openreview.net/pdf?id=t8cBsT9mcg

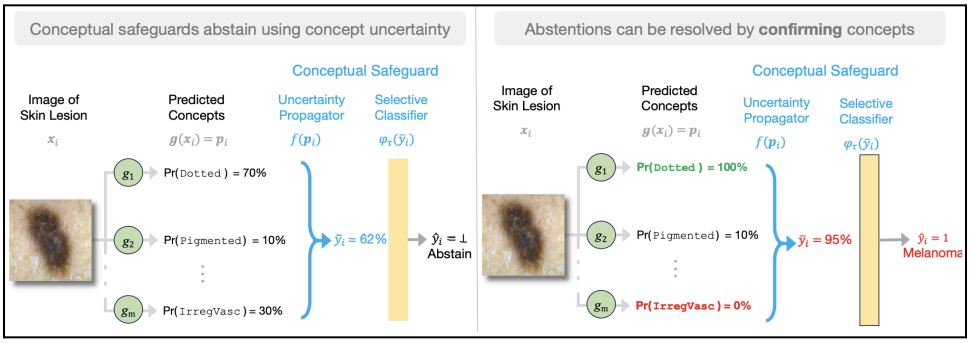


Fig 4: Conceptual safeguard to detect melanoma from an image of a skin lesion. Conceptual Safegaurds have an uncertainty layer that allows the model to predict or abstain based on the confidence of the predicted concepts.

4. **Post-hoc concept bottleneck models** by Yuksekgonul et al. enable **global** model interventions, which is more efficient than previous approaches that only fixed specific predictions. In addition, they leverage concept annotations from other datasets or use natural language descriptions via multimodal models to automatically generate concepts. Paper: https://arxiv.org/pdf/2205.15480

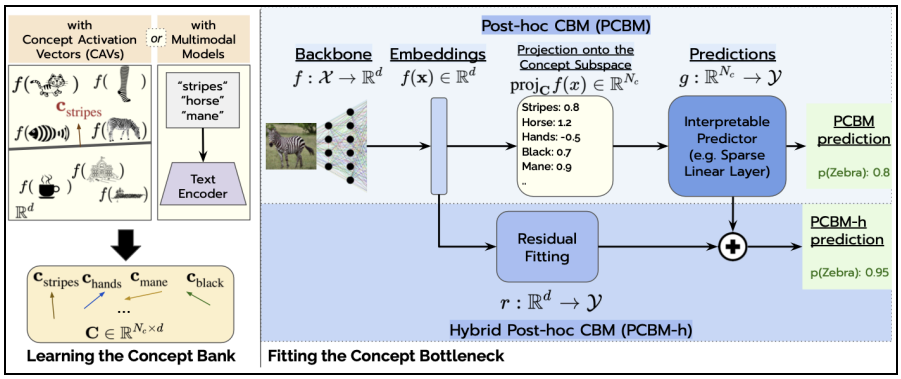


Fig 5: Post-hoc CBM architecture.

D.2. **Increased Adoption of CBMs**

1. NLP: 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).
2. Vision: 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).
3. LLMs: Sun, C.E., Oikarinen, T. and Weng, T.W., 2024. **Crafting large language models for enhanced interpretability**. arXiv preprint arXiv:2407.04307.

E. **Broader Use Cases**

Throughout our course, we've discussed **pairing explanations with specific use cases**, i.e. using interpretability and explainability as a **means to an end**. These use cases include debugging, personalization, recourse, feature selection, etc. Interestingly, CBMs can effectively cater to most of these use cases:

1. While not explicitly stated, CBMs also support effective model **debugging**. CBMs allow practitioners to pinpoint errors or inconsistencies in the model's reasoning using concepts.
2. Another key use case is **actionable recourse**, where CBMs can help users identify high-level concepts that influence predictions, allowing them to understand which factors to change to achieve a desired outcome. This model can be deployed in settings where providing actionable recourse to users is a legal requirement.
3. Additionally, CBMs can generate **counterfactual explanations** by identifying which concept-level changes would lead to a different prediction, offering users a clearer understanding of alternative outcomes.
4. CBMs can also be used for **personalization**, as highlighted in Section B, where they can replace Multinomial Naive Bayes for **explanatory debugging**, thereby allowing for a two-way interaction between users and the models.
5. The main practical use case that the author stresses in this paper is its **application in high-stakes settings**, where collaboration between humans and models is crucial. This model enables practitioners to reason using familiar high-level concepts.

Examples: Practical Use Cases

Here are a few real-life examples based on the above mentioned use cases:

1. The authors mention an application of CBMs in high-stakes settings like **Radiology** for predicting the severity of Arthritis. CBMs can easily be adapted to other similar high-stakes settings like healthcare, medicine, etc, for diagnosis purposes.
2. We can also use CBMs in **finance**. Considering the famous example of credit risk, where it can become a legal requirement to explain the reasons for denying customers credit, CBMs offer clear, high-level concept-based justifications for predictions and provide actionable recourse to users to change their predictions. This makes CBMs highly effective in **legal** and regulatory contexts such as criminal justice.
3. **Personalized** applications: CBMs can help recommend products by explaining the intermediate factors (e.g., "recent purchases," "customer preferences," "seasonal trends") that influence the recommendation, offering more tailored and transparent marketing strategies.

F. Opportunities for Future Research

1. Future research should focus on the **scalability** of CBMs. There can be tasks where the number of concepts can explode, how can users effectively intervene in such settings would be an interesting research problem. The usability of CBMs heavily depends on their ability to interact effectively with human decision-makers.
2. In most cases, CBMs underperform as compared to unrestricted black-box models. **Improving CBM performance** at an architectural level could be an interesting research direction. One approach might involve adding a residual link from x (backend) to y (frontend), though this could reduce intervenability.
3. Research into developing **robust techniques** to prevent information leakage in CBMs is crucial to ensure that any unintended information doesn't get encoded in concept prediction.
4. Recently, many techniques have emerged that leverage open-source models like GPT and CLIP for concept generation. **Quantifying the quality** and relevance of these generated concepts could be an interesting research problem.