DECOUPLING CONCEPT BOTTLENECK MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept Bottleneck Model (CBM) is a kind of powerful white-box neural network, which maps the input data onto a group of high-level concepts, and then utilizes the mapped concepts to implement downstream tasks. Equipped with CBM, human can understand why the model makes such predictions through the corresponding concepts. However, high-level concepts in real-world senarios often suffer a troublesome collection and commonplace insufficiency. Thus, it's often intractable for the collected concepts to contain sufficient information to conduct and interpret the downstream tasks. In this paper, we theoretically prove the dilemma of concept/label distortion in CBM due to the mixture of explicit/implicit information. To alleviate the trade-off between concept-distortion and label-distortion, we propose a Decoupling Concept Bottleneck Model (DCBM), which decouples the heterogeneous information into explicit and implicit concepts. Extensive experiments on two real-world datasets, including CUB and Derm7pt, expose the success in the alleviation of distortions, where DCBM achieves state-of-the-art results in both concept classification and label identification tasks. Moreover, DCBM can perform comprehensive human-computer interactions within a decoupling regime. On the one hand, we can edit the predicted explicit concept values, as well as the simulated explicit concept values decoupled from implicit ones, to promote the final prediction. On the other hand, when the model makes wrong label identification, DCBM can trace back the corresponding cause automatically through a concept fine-tuning method. These two methods only need few additional training costs, while performance can be promoted significantly.

1 INTRODUCTION

Figure 1: Forcible concept/label distortion in CBM. (a) is the main pipeline in CBM, which expects an ideal condition with all explicit concepts as in (b). However, due to the insufficiency of labeled concepts in most real-world scenarios, various latent concepts are implicit to us, like that in (c). Thus, the implicit concepts in CBM have to be mixed into explicit ones in order to adapt the multiple objectives.

Our Contributions

- To our best knowledge, we are the first to prove the concept-distortion and label-distortion in widely-used Concept Bottleneck Model (CBM), especially when high-level concepts are insufficient.
- We propose Decoupling Concept Bottleneck Model (DCBM), a novel concept-based framework, to alleviate the distortions, where the transparency and interaction of AI system can still be preserved.
- We have done multiple experiments to illustrate the efficiency of DCBM, which achieves state-of-the-art results in both concept classification and label identification tasks. In particular, the concept/label distortion in DCBM can be alleviated conspicuously, compared with CBM.
- Under a decoupling regime, DCBM affords to perform novel comprehensive human-computer interactions, including a forward intervention process and a backward rectification process. These steps, consuming few additional training costs, help us ascertain the

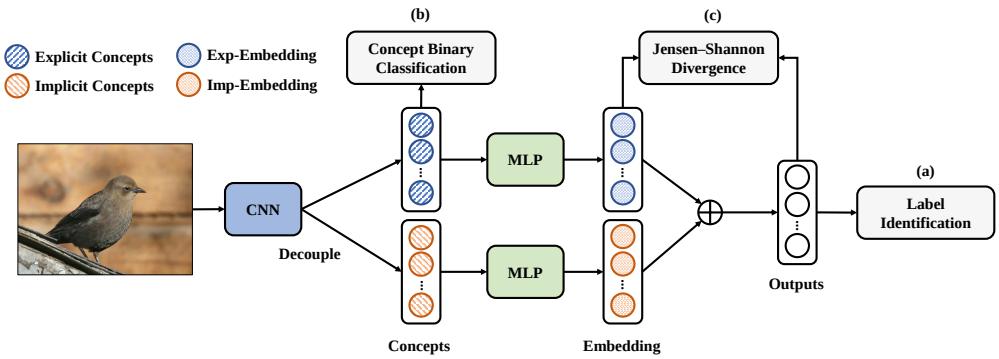

Figure 2: Pipeline of our work. 'Exp-Eembedding' and 'Imp-Eembedding' represent the embeddings of explicit and implicit concepts, respectively. We jointly consider three objectives, including: a) Label identification, to optimize the label prediction accuracy; b) Concept binary classification, to optimize the concept prediction accuracy; c) JS divergence, to restrict Exp-Eembedding not to be far away from the outputs.

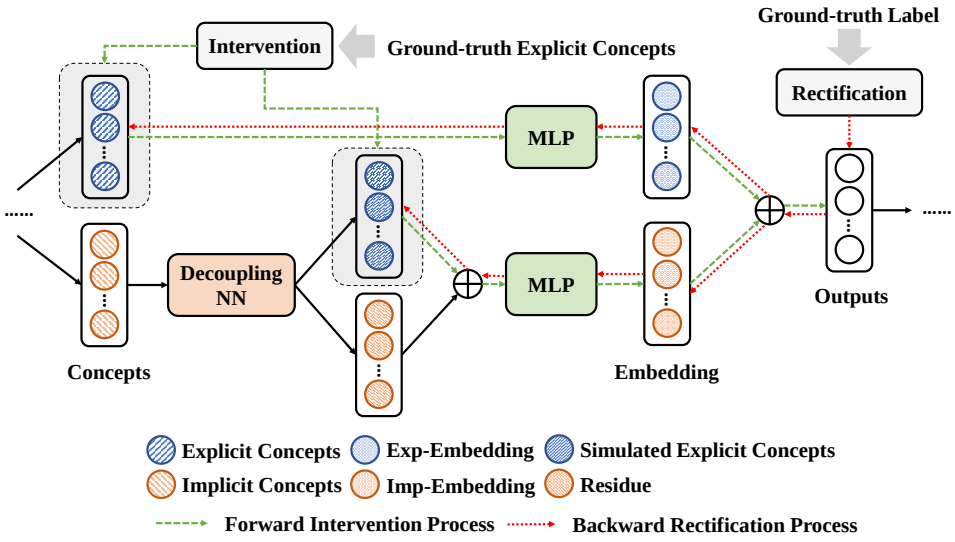

Figure 3: Forward Intervention and Backward Rectification. Implicit concepts are decoupled into two parts, i.e., simulated explicit concepts and residues, to extract the information from implicit concepts that can be expressed linearly by explicit concepts. In the process of forward intervention, we both intervene explicit concepts and simulated explicit concepts through ground-truth explicit concepts, to promote the label prediction accuracy, while in the process of backward rectification, DCBM can automatically trace back to the cause through a concept fine-tuning method.

cause of wrong predictions more clearly and have magnificant promotions in concept/label prediction accuracy.

Table 1: Brief summary of the referred datasets in this paper. Here # means 'the number of'. Note that the splitting of train-test-validation is in line with (Wah et al., 2011; Kawahara et al., 2018), respectively.

DATASET	#CONCEPT	#CLASS	#TRAIN	#VAL	#TEST
Derm7pt (Kawahara et al., 2018)	8	2	688	322	636
CUB (Wah et al., 2011)	112				
CUB%50	56	200	4796	1198	5794
CUB%20	22				

2 Related Work

2.1 Concept Bottleneck Model

3 Preliminary

4 Decoupling Concept Bottleneck Model

5 Experiments

5.1 Dataset

We consider two commonly-used datasets, including bird species and dermoscopic images. Both of them express high-level concepts to enable us to provide interpretability analysis.

CUB (Wah et al., 2011): Caltech-UCSD Birds-200-2011, which we call CUB in the following text, is an extension of CUB-200 (Welinder et al., 2010), collecting a total of 200 bird species with 11,788 bird photographs. CUB comprises 312 bird attributes including wing shape, back pattern, eye color, etc. Since the concepts are noisy and sparse, we leave only 112 binary concepts instead, following the work (Koh et al., 2020). This task aims to classify each bird photograph out of 200 different species. Thanks to the splitting standard of train/test/validation dataset (Wah et al., 2011), we only need to follow this given setting.

Derm7pt (Kawahara et al., 2018): Seven-Point Checklist Dermatology dataset, abbreviated as Derm7pt, is a dataset comprising 1,011 groups of images. Each group contains a clinic photograph and a dermoscopic image, respectively. Due to the lack of Seborrheic Keratosis (SK) samples, we follow the setting of the work (Lucieri et al., 2020), only considering Melanoma (MEL) and Naevi (NV) ones with a total of 823 groups. Derm7pt contains several authoritative concepts (Kittler et al., 2016) including Pigmented Networks, Streaks, etc. On account of the concept absence and interpretability on Melanoma detection, we leave 8 malignancy-associated binary concepts, including Blue Whitish Veil, Pigmented Networks, and Regression Structures, as (Yuksekgonul et al., 2022) did. This task is a binary classification problem, judging whether an image is a melanoma or naevi. Similarly, the train/test/validation indexes are already given as well.

Note that CUB succeeds to have a large group of concepts, while Derm7pt fails. However, most datasets we have investigated typically have a small number of concepts (Nevitt et al., 2006), or have to be extracted (Tschandl et al., 2018; Rotemberg et al., 2021; Krizhevsky et al., 2009; Fong & Vedaldi, 2018) by other trained models (Lucieri et al., 2020; Abid et al., 2021). This phenomenon is understandable, since high-quality and abundant labeled concepts are labour-intensive and infrequent. This bring us a new question — What if CUB also has an insufficient group of concepts? And then, how will CBM perform under this condition? To validate our assumption, we randomly drop 50% and 80% concepts in CUB, respectively, and form two new datasets named **CUB%50** and **CUB%20**. We display a brief summary of the referred datasets in Table 1.

5.2 CONCEPT/LABEL DISTORTION IN CBM

We report that there exists a forcible concept/label distortion in CBM.

6 CONCEPT CONTRIBUTION ANALYSIS

How much does the black-box concepts affect?

7 FORWARD INTERVENTION AND BACKWARD RECTIFICATION

8 CONCLUSION

9 CITATIONS WITHIN THE TEXT

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis using `\citet{}` (as in "See Hinton et al. (2006) for more information."). Otherwise, the citation should be in parenthesis using `\citep{}` (as in "Deep learning shows promise to make progress towards AI (Bengio & LeCun, 2007).").

The corresponding references are to be listed in alphabetical order of authors, in the REFERENCES section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

10 FIGURES

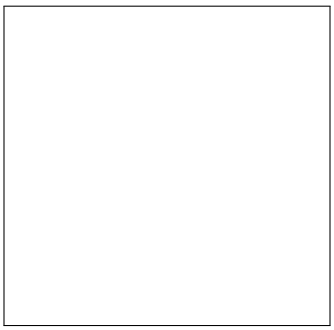

Figure 4: Sample figure caption.

11 TABLES

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 2.

Numbers and Arrays

Table 2: Sample table title

PART	DESCRIPTION
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
A	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor A
$\mathsf{A}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathbf{X}} y$	Tensor containing derivatives of y with respect to $\mathbf{X}$
$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x
$D_{\mathrm{KL}}(P \| Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\frac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\mathrm{condition}}$	is 1 if the condition is true, 0 otherwise

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully explaining model mistakes using conceptual counterfactuals. *arXiv preprint arXiv:2106.12723*, 2021.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738, 2018.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

Harald Kittler, Ashfaq A Marghoob, Giuseppe Argenziano, Cristina Carrera, Clara Curiel-Lewandrowski, Rainer Hofmann-Wellenhof, Josep Malvehy, Scott Menzies, Susana Puig, Harold Rabinovitz, et al. Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the international society of dermoscopy. *Journal of the American Academy of Dermatology*, 74(6):1093–1106, 2016.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–10. IEEE, 2020.

M Nevitt, D Felson, and Gayle Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1, 2006.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. URL https://openreview.net/forum?id=HAMeOIRD_g9.

A APPENDIX

You may include other additional sections here.