Concepts' Information Bottleneck Models

Karim Galliamov[1] Syed M Ahsan Kazmi[2] Adil Khan[3] Adín Ramírez Rivera[4]

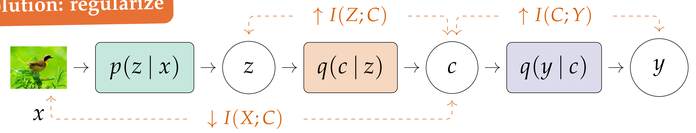
[1]University of Amsterdam [2]University of West of England [3]University of Hull [4]University of Oslo

Goal Interpretable concepts must **not encode irrelevant details** of the raw input, and should **predict labels and representations**.

$$\max I(Z;C) + I(C;Y) - \beta I(X;C)$$

Preserving — Compressing

Our solution: regularize



Bounded CIB: IB_B

Close the gap to the ideal loss

$$\begin{aligned}\mathcal{L}_{\text{CIBM}} \geq &(1-\beta)\ \mathop{\mathbb{E}}_{p(z)}\Big[H(p(c\mid z)) - H\Big(p(c\mid z), q(c\mid z)\Big)\Big] - \mathop{\mathbb{E}}_{p(c)} H\Big(p(y\mid c), q(y\mid c)\Big)\\ &= -\mathcal{L}_{\text{S-CIBM}}\end{aligned}$$

Estimator-based CIB: IB_E

Fix to not lose the data anchor

$$\mathcal{L}_{\text{CIBM}} \geq H(Y) + H(C) - \mathop{\mathbb{E}}_{p(c)} H\Big(p(y\mid c), q(y\mid c)\Big) - \mathop{\mathbb{E}}_{p(z)} H\Big(p(c\mid z), q(c\mid z)\Big) - \beta I(X;C)$$

By treating the entropies of the concepts and the labels as constants, we obtain our loss

$$\mathcal{L}_{\text{E-CIB}} = \mathop{\mathbb{E}}_{p(c)} H\Big(p(y\mid c), q(y\mid c)\Big) + \mathop{\mathbb{E}}_{p(z)} H\Big(p(c\mid z), q(c\mid z)\Big) - \beta(\rho - I(X;C))$$

Concept Leakage

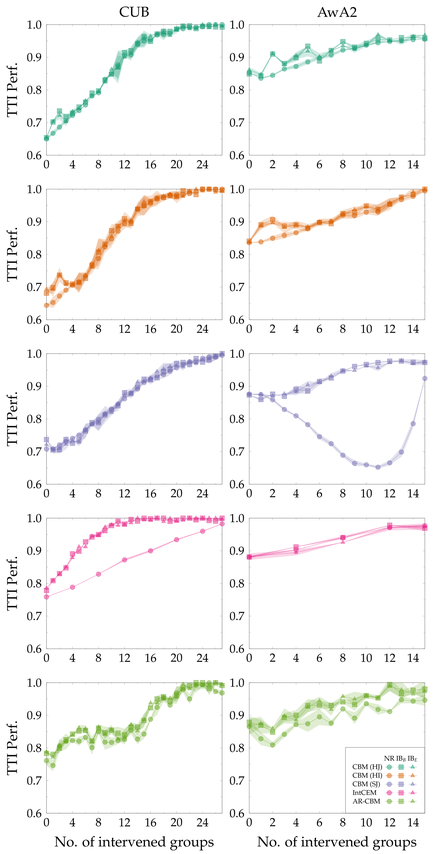
Model	Complete CS OIS	Complete CS NIS	Sel. Drop-out CS OIS	Sel. Drop-out CS NIS	Random Drop-out CS OIS	Random Drop-out CS NIS
CBM (HJ)	4.31 ± 0.57	65.03 ± 1.84	15.71	77.02	13.10 ± 1.20	73.09 ± 2.12
CBM (HJ) +IB_B	2.59 ± 0.65	60.02 ± 1.28	12.67	70.86	10.32 ± 0.57	70.38 ± 2.12
CBM (HJ) +IB_E	2.56 ± 0.47	59.52 ± 1.91	12.47	71.09	10.36 ± 0.80	69.95 ± 1.35
CBM (HI)	3.77 ± 0.58	65.15 ± 2.03	14.27	74.16	12.31 ± 1.04	71.78 ± 1.86
CBM (HI) +IB_B	3.58 ± 0.76	64.95 ± 1.59	14.01	73.16	12.63 ± 1.31	70.89 ± 2.04
CBM (HI) +IB_E	3.69 ± 0.71	64.89 ± 1.51	14.00	73.13	12.60 ± 1.29	70.93 ± 2.09
CBM (SJ)	4.69 ± 0.43	66.25 ± 2.31	16.29	78.39	12.97 ± 0.78	74.19 ± 1.04
CBM (SJ) +IB_B	2.01 ± 0.07	62.32 ± 1.95	13.26	72.80	10.95 ± 1.49	71.94 ± 0.98
CBM (SJ) +IB_E	2.00 ± 0.09	62.31 ± 2.08	13.23	72.85	11.03 ± 1.38	72.03 ± 1.07

Model	Complete CS OIS	Complete CS NIS	Sel. Drop-out CS OIS	Sel. Drop-out CS NIS	Random Drop-out CS OIS	Random Drop-out CS NIS
IntCEM	8.74 ± 0.30	75.41 ± 3.83	20.85	80.19	18.31 ± 0.09	76.56 ± 2.00
IntCEM +IB_B	6.15 ± 0.20	69.96 ± 2.22	17.22	76.67	14.07 ± 0.43	72.60 ± 2.40
IntCEM +IB_E	6.07 ± 0.23	71.28 ± 1.87	18.04	75.75	13.98 ± 0.29	73.89 ± 1.99
AR-CBM	3.90 ± 0.27	62.30 ± 1.52	14.16	63.40	12.58 ± 0.86	60.86 ± 1.32
AR-CBM + IB_B	2.84 ± 0.26	59.84 ± 1.48	10.96	59.66	10.17 ± 0.48	56.31 ± 1.33
AR-CBM + IB_E	2.72 ± 0.16	59.96 ± 0.98	11.09	60.77	11.01 ± 1.05	57.23 ± 0.79
ProbCBM	4.30 ± 0.10	64.22 ± 1.04	16.01	76.92	13.81 ± 0.21	75.01 ± 0.86
ProbCBM + IB_B	2.43 ± 0.38	60.29 ± 2.02	13.04	72.76	10.29 ± 0.50	70.86 ± 1.87
ProbCBM + IB_E	2.57 ± 0.52	59.81 ± 1.74	13.05	73.08	11.15 ± 0.62	71.13 ± 2.09

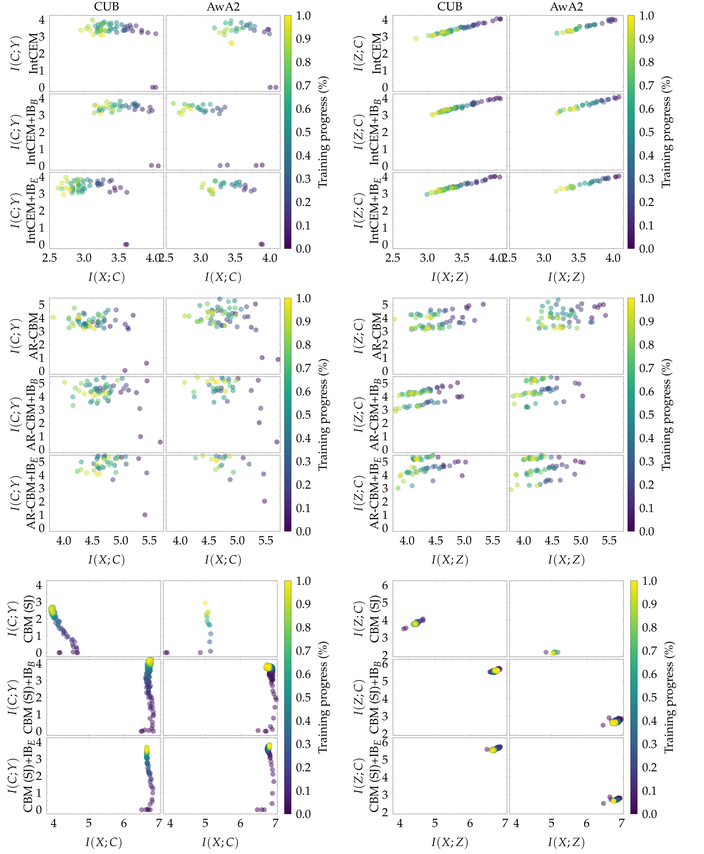
Prediction

Method	Concept	Class
CUB — Black-box	–	0.919±0.002
CBM (HJ)	0.956±0.001	0.650±0.002
CBM (HJ) + IB_B	0.951±0.005	0.654±0.000
CBM (HJ) + IB_E	0.951±0.000	0.655±0.000
CBM (HI)	0.956±0.001	0.644±0.001
CBM (HI) + IB_B	0.954±0.000	0.689±0.000
CBM (HI) + IB_E	0.952±0.000	0.691±0.000
CBM (SJ)	0.956±0.001	0.708±0.006
CBM (SJ) + IB_B	0.961±0.005	0.727±0.000
CBM (SJ) + IB_E	0.958±0.005	0.726±0.000
CBM (SI)	0.956±0.001	0.639±0.001
CBM (SI) + IB_B	0.950±0.002	0.643±0.000
CBM (SI) + IB_E	0.958±0.000	0.646±0.000
ProbCBM	0.956±0.001	0.718±0.005
ProbCBM + IB_B	0.960±0.000	0.741±0.006
ProbCBM + IB_E	0.962±0.000	0.734±0.002
IntCEM	0.954±0.001	0.759±0.002
IntCEM + IB_B	0.957±0.000	0.778±0.000
IntCEM + IB_E	0.959±0.000	0.773±0.000
AR-CBM	0.956±0.002	0.761±0.010
AR-CBM + IB_B	0.961±0.000	0.777±0.008
AR-CBM + IB_E	0.950±0.000	0.774±0.006
AwA2 — Black-box	–	0.893±0.000
CBM (HJ)	0.979±0.000	0.853±0.002
CBM (HJ) + IB_B	0.968±0.000	0.845±0.000
CBM (HJ) + IB_E	0.970±0.000	0.856±0.001
CBM (HI)	0.979±0.000	0.836±0.001
CBM (HI) + IB_B	0.966±0.000	0.823±0.000
CBM (HI) + IB_E	0.976±0.000	0.834±0.000
CBM (SJ)	0.979±0.001	0.876±0.001
CBM (SJ) + IB_B	0.978±0.003	0.878±0.000
CBM (SJ) + IB_E	0.980±0.000	0.874±0.002
CBM (SI)	0.979±0.000	0.852±0.004
CBM (SI) + IB_B	0.967±0.000	0.847±0.005
CBM (SI) + IB_E	0.969±0.000	0.853±0.003
ProbCBM	0.979±0.000	0.880±0.003
ProbCBM + IB_B	0.971±0.001	0.883±0.000
ProbCBM + IB_E	0.973±0.000	0.883±0.000
IntCEM	0.979±0.000	0.884±0.002
IntCEM + IB_B	0.977±0.000	0.884±0.000
IntCEM + IB_E	0.982±0.000	0.886±0.003
AR-CBM	0.979±0.001	0.884±0.006
AR-CBM + IB_B	0.977±0.004	0.876±0.006
AR-CBM + IB_E	0.978±0.004	0.879±0.000

Test-Time Interventions



Information plane dynamics



ICLR 2026 — International Conference on Learning Representations — April, 2026 — Rio de Janeiro, Brazil

ACKNOWLEDGMENTS: Sigma2 NN8104K & SFI Visual Intelligence (RCN 309439) & FRIPRO (grant no. 359216)