Drop-Bottleneck

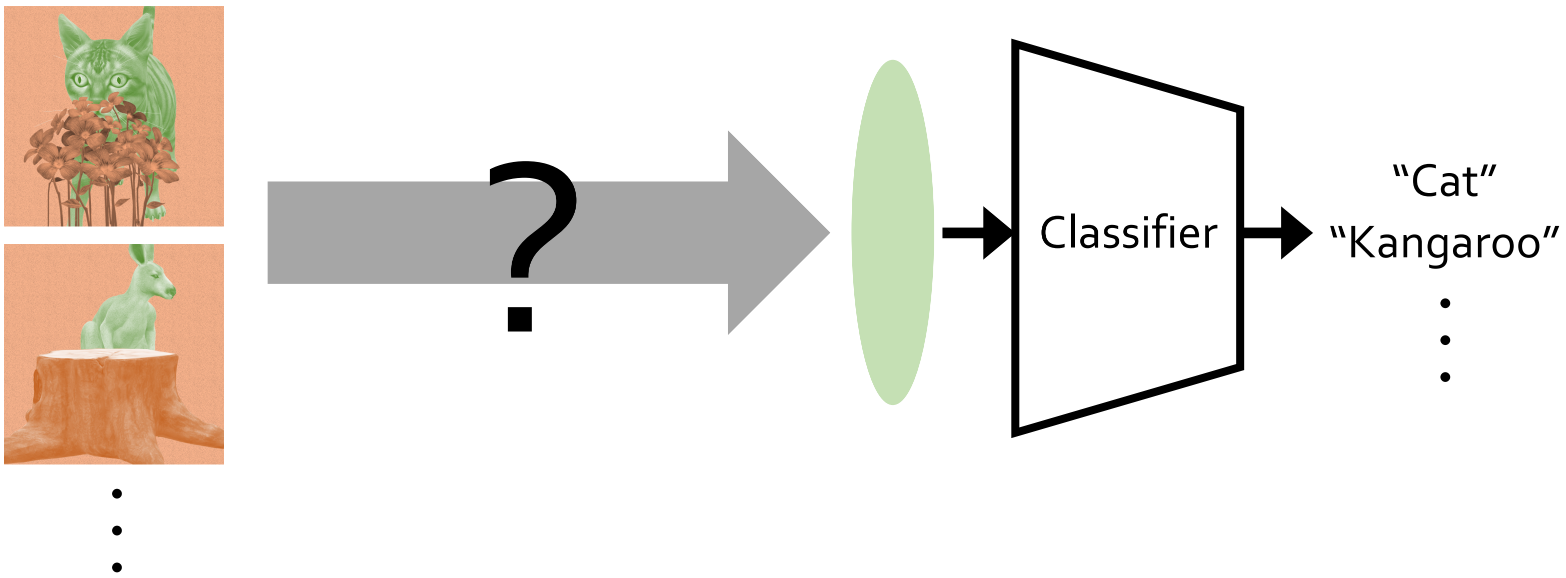
Learning Discrete Compressed Representation for Noise-Robust Exploration

ICLR 2021

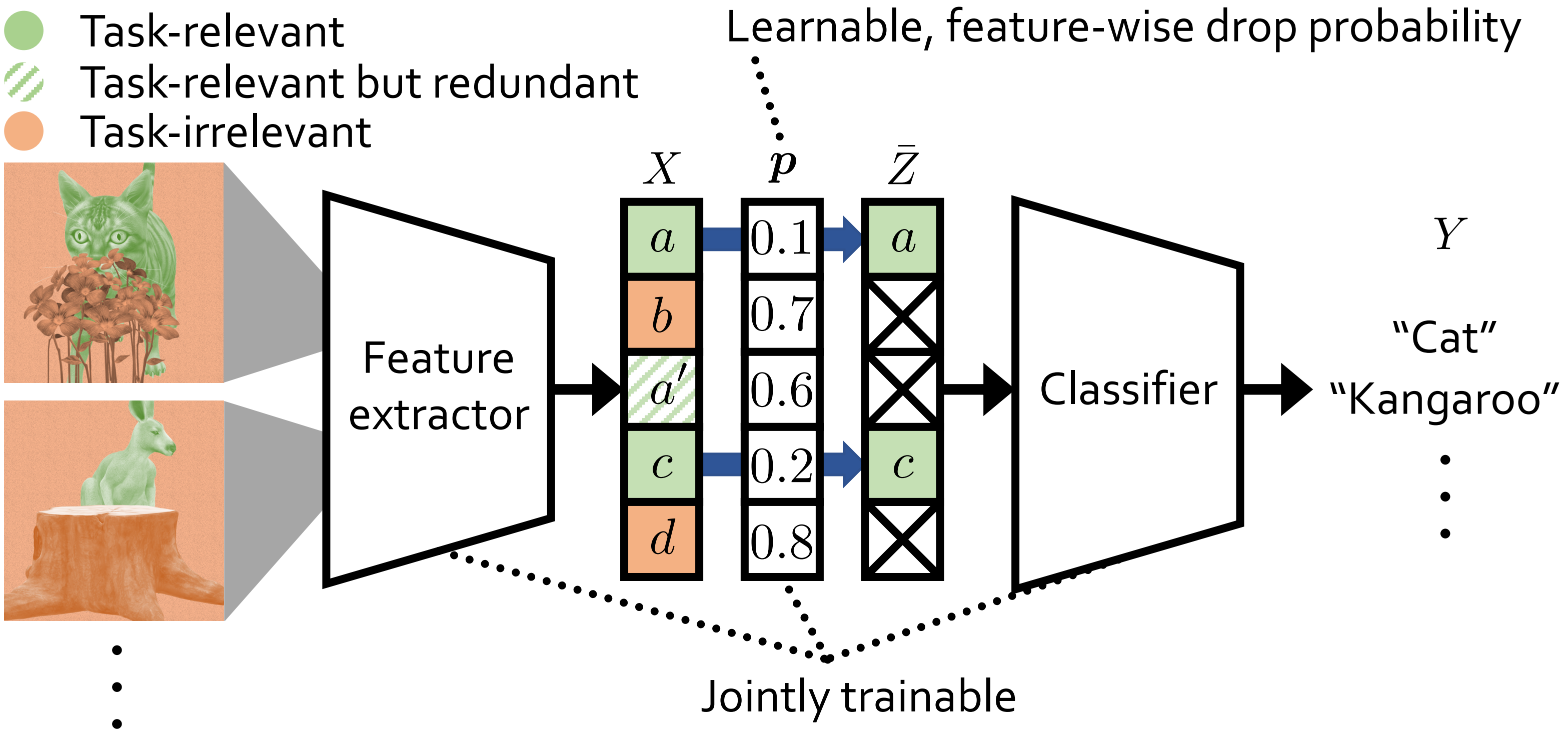
Jaekyeom Kim, Minjung Kim, Dongyeon Woo & Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

Overview



Overview



Discretely compressed representations
via information bottleneck (IB) framework!

Motivation

- Information bottleneck (IB) framework*

$$\text{minimize} - \underbrace{I(Z;Y)}_{\text{prediction}} + \beta \underbrace{I(Z;X)}_{\text{compression}}$$

* Tishby et al., 2000, The information bottleneck method

Motivation

- Information bottleneck (IB) framework*

$$\text{minimize} - \underbrace{I(Z;Y)}_{\text{prediction}} + \beta \underbrace{I(Z;X)}_{\text{compression}}$$

- IB method that provides
  - o non-stochastic compressed representations for stability and consistency
  - o increased practical efficiency as the result of compression

* Tishby et al., 2000, The information bottleneck method

Motivation

- Information bottleneck (IB) framework*

$$\text{minimize} - \underbrace{I(Z;Y)}_{\text{prediction}} + \beta \underbrace{I(Z;X)}_{\text{compression}}$$

- IB method that provides
  - o non-stochastic compressed representations for stability and consistency
  - o increased practical efficiency as the result of compression
- Prior IB methods (e.g. VIB†) lack the properties.

* Tishby et al., 2000, The information bottleneck method

† Alemi et al., 2017, Deep Variational Information Bottleneck

Drop-Bottleneck (DB)

- Our approach (Drop-Bottleneck): IB method that *discretely* drops irrelevant features with joint feature learning

Drop-Bottleneck (DB)

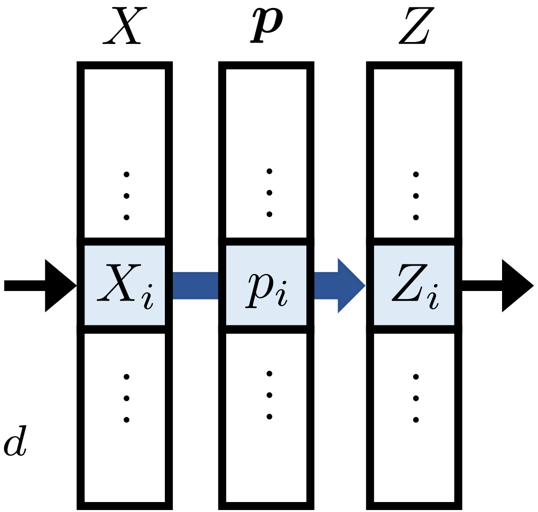
- Our approach (Drop-Bottleneck): IB method that *discretely* drops irrelevant features with joint feature learning

- DB defines representation $Z = C_{\boldsymbol{p}}(X)$ as

$$Z_i = b \cdot \mathrm{Bernoulli}(1 - p_i) \cdot X_i,$$

$$\text{for } b = \frac{d}{d - \sum_k p_k}$$

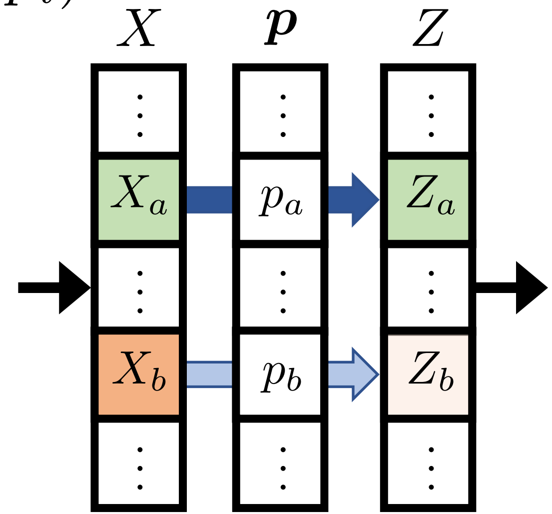
learning feature-wise drop probability $\boldsymbol{p} \in [0, 1]^d$



Objective and Training of DB

- We derive and minimize upper bound of compression term as

$$
\begin{aligned}
I(Z;X) \leq \hat{I}(Z;X) &= \textstyle\sum_{i=1}^{d} I(Z_i;X_i) \\
&\approx \textstyle\sum_{i=1}^{d} H(X_i)(1-p_i)
\end{aligned}
$$



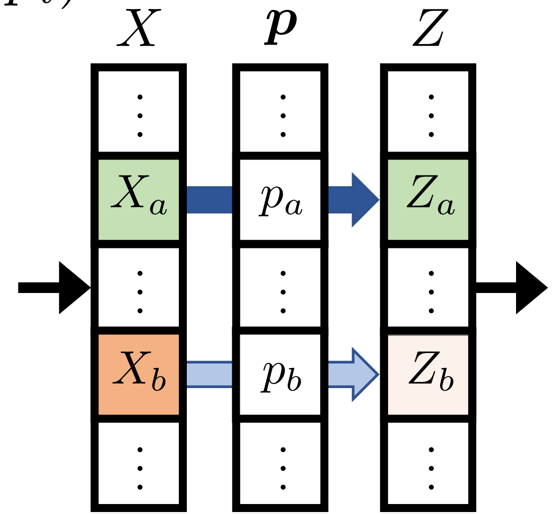
Objective and Training of DB

- We derive and minimize upper bound of compression term as

$$I(Z;X) \le \hat{I}(Z;X) = \sum_{i=1}^{d} I(Z_i;X_i)$$
$$\approx \sum_{i=1}^{d} H(X_i)(1-p_i)$$

- Z is not differentiable w.r.t. $\boldsymbol{p}$

  ⇨ use Concrete relaxation of Bernoulli*



* Maddison et al., 2017, The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables

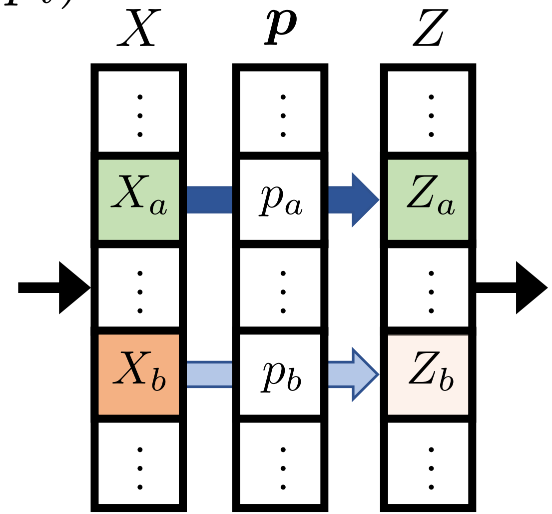
Objective and Training of DB

- We derive and minimize upper bound of compression term as

$$I(Z;X) \le \hat{I}(Z;X) = \sum_{i=1}^{d} I(Z_i;X_i)$$
$$\approx \sum_{i=1}^{d} H(X_i)(1-p_i)$$

- Z is not differentiable w.r.t. $\boldsymbol{p}$
  ⇨ use Concrete relaxation of Bernoulli*

- Allows joint training with feature extractor that outputs X, via prediction term (e.g. using Deep InfoMax†)

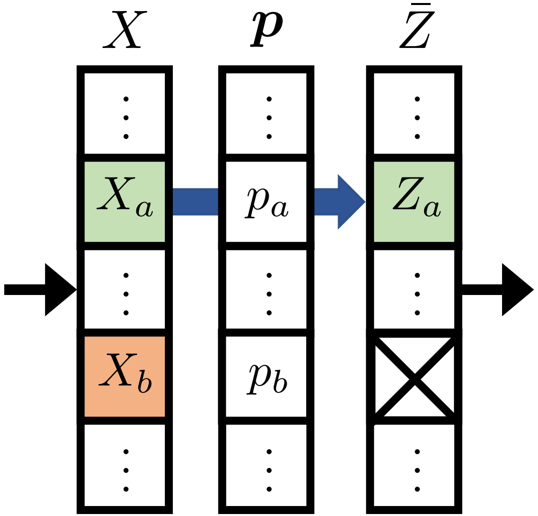


* Maddison et al., 2017, The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables
† Hjelm et al., 2019, Learning deep representations by mutual information estimation and maximization

Deterministic Compressed Representations

- We define deterministic compressed representation $\bar{Z} = \bar{C}_{\boldsymbol{p}}(X)$ as

$$\bar{Z}_i = \bar{b} \cdot \mathbb{1}(p_i < 0.5) \cdot X_i, \quad \text{for } \bar{b} = \frac{d}{\sum_k \mathbb{1}(p_k < 0.5)}$$

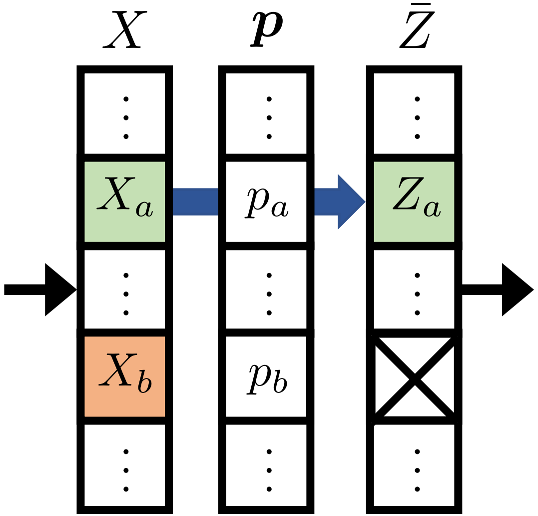


Deterministic Compressed Representations

- We define deterministic compressed representation $\bar{Z} = \bar{C}_{\boldsymbol{p}}(X)$ as

$$\bar{Z}_i = \bar{b} \cdot \mathbb{1}(p_i < 0.5) \cdot X_i, \quad \text{for } \bar{b} = \frac{d}{\sum_k \mathbb{1}(p_k < 0.5)}$$

- Useful for inference tasks that require consistent representations

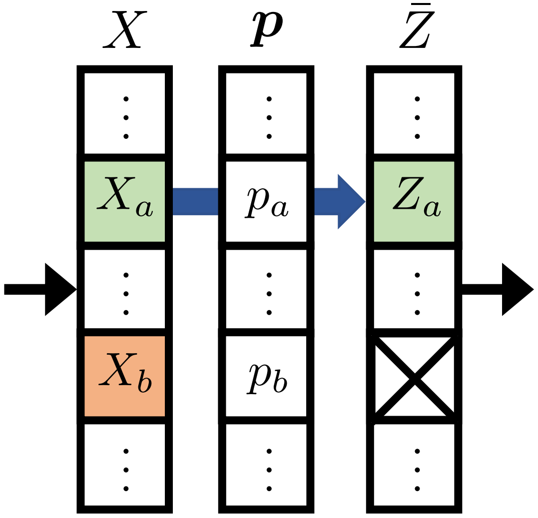


Deterministic Compressed Representations

- We define deterministic compressed representation $\bar{Z} = \bar{C}_{\boldsymbol{p}}(X)$ as

$$\bar{Z}_i = \bar{b} \cdot \mathbb{1}(p_i < 0.5) \cdot X_i, \quad \text{for } \bar{b} = \frac{d}{\sum_k \mathbb{1}(p_k < 0.5)}$$

- Useful for inference tasks that require consistent representations



- Provides feature dimensionality reduction at inference time

Exploration in RL Environments with DB

- We train $\boldsymbol{p}$ and f_ϕ using DB and Deep InfoMax with
$$X = f_\phi(S'), \qquad Z = C_{\boldsymbol{p}}(X), \qquad Y = C_{\boldsymbol{p}}(f_\phi(S))$$
for transitions (S, A, S')

Exploration in RL Environments with DB

- We train $\boldsymbol{p}$ and f_ϕ using DB and Deep InfoMax with
$$X = f_\phi(S'), \qquad Z = C_{\boldsymbol{p}}(X), \qquad Y = C_{\boldsymbol{p}}(f_\phi(S))$$
for transitions (S, A, S')

- $I(Z;Y) = I(C_{\boldsymbol{p}}(f_\phi(S')); C_{\boldsymbol{p}}(f_\phi(S)))$ encourages compressed representations of S and S' predictable about each other

Exploration in RL Environments with DB

- We train $\boldsymbol{p}$ and f_ϕ using DB and Deep InfoMax with
$$X = f_\phi(S'), \qquad Z = C_{\boldsymbol{p}}(X), \qquad Y = C_{\boldsymbol{p}}(f_\phi(S))$$
for transitions (S, A, S')

- $I(Z;Y) = I(C_{\boldsymbol{p}}(f_\phi(S')); C_{\boldsymbol{p}}(f_\phi(S)))$ encourages compressed representations of S and S' predictable about each other

- We keep episodic memory* $M = \{\bar{C}_{\boldsymbol{p}}(f_\phi(s_1)), \ldots, \bar{C}_{\boldsymbol{p}}(f_\phi(s_{t-1}))\}$ and quantify novelty of s_t using Deep InfoMax's discriminator

* Savinov et al., 2019, Episodic Curiosity through Reachability

Experiments: Exploration in Noisy Environments

- Three noisy-TV settings: ImageAction, Noise, NoiseAction*
- Environments: DMLab, VizDoom

Method	VizDoom						DMLab					
	Sparse			Very Sparse			Sparse			Very Sparse		
	IA	N	NA	IA	N	NA	IA	N	NA	IA	N	NA
PPO	0.00	0.00	0.00	0.00	0.00	0.00	8.5	11.6	9.8	6.3	8.7	6.1
+ ICM	0.00	0.50	0.40	0.00	0.73	0.20	6.9	7.7	7.6	4.9	6.0	5.7
+ EC	–	–	–	–	–	–	13.1	18.7	14.8	7.4	13.4	11.3
+ ECO	0.21	0.70	0.33	0.19	0.79	0.50	18.5	28.2	18.9	16.8	26.0	12.5
+ **Ours**	**0.90**	**1.00**	**0.99**	**0.90**	**1.00**	**0.90**	**30.4**	**32.7**	**30.6**	**28.8**	**29.1**	**26.9**

* Savinov et al., 2019, Episodic Curiosity through Reachability

Experiments: Exploration in Noisy Environments

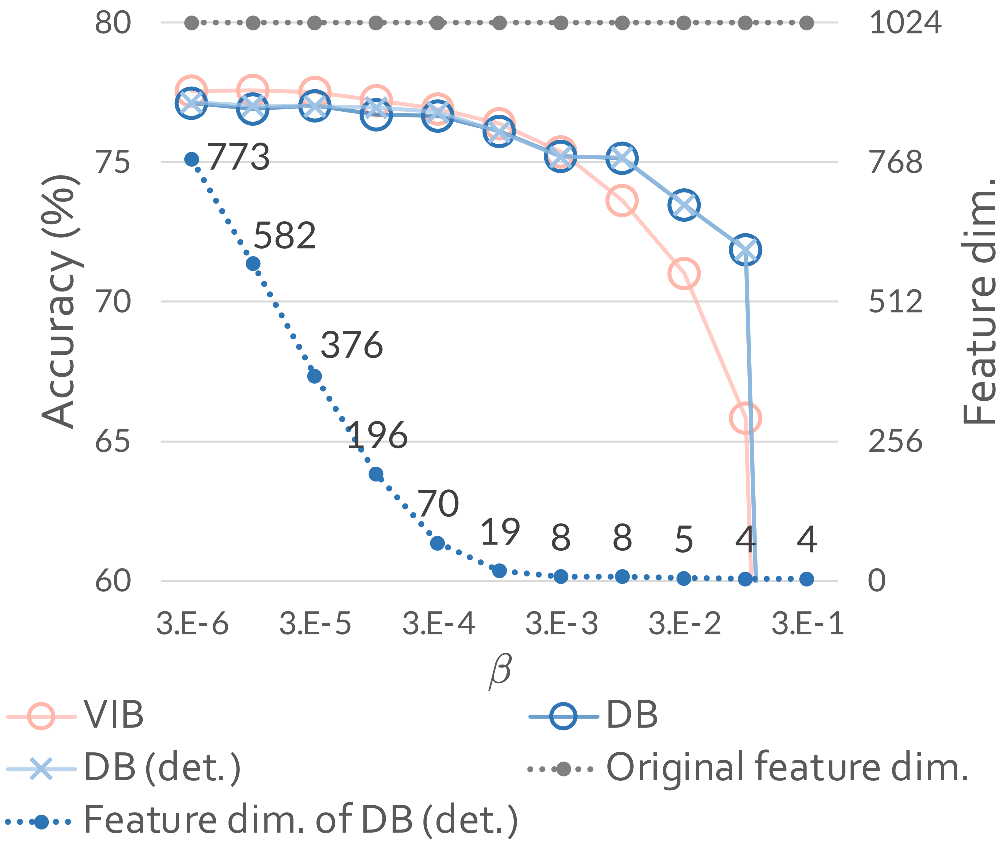
- Three noisy-TV settings: ImageAction, Noise, NoiseAction*
- Environments: DMLab, VizDoom
- SOTA in all tasks (average episodic sum of rewards; higher is better)

Method	VizDoom						DMLab					
	Sparse			Very Sparse			Sparse			Very Sparse		
	IA	N	NA	IA	N	NA	IA	N	NA	IA	N	NA
PPO	0.00	0.00	0.00	0.00	0.00	0.00	8.5	11.6	9.8	6.3	8.7	6.1
+ ICM	0.00	0.50	0.40	0.00	0.73	0.20	6.9	7.7	7.6	4.9	6.0	5.7
+ EC	–	–	–	–	–	–	13.1	18.7	14.8	7.4	13.4	11.3
+ ECO	0.21	0.70	0.33	0.19	0.79	0.50	18.5	28.2	18.9	16.8	26.0	12.5
+ **Ours**	**0.90**	**1.00**	**0.99**	**0.90**	**1.00**	**0.90**	**30.4**	**32.7**	**30.6**	**28.8**	**29.1**	**26.9**

* Savinov et al., 2019, Episodic Curiosity through Reachability

Experiments: Comparison with Variational IB (VIB)

- ImageNet classification accuracy and feature dimensionality of VIB* and DB (+ deterministic) equipped with Inception-ResNet-v2

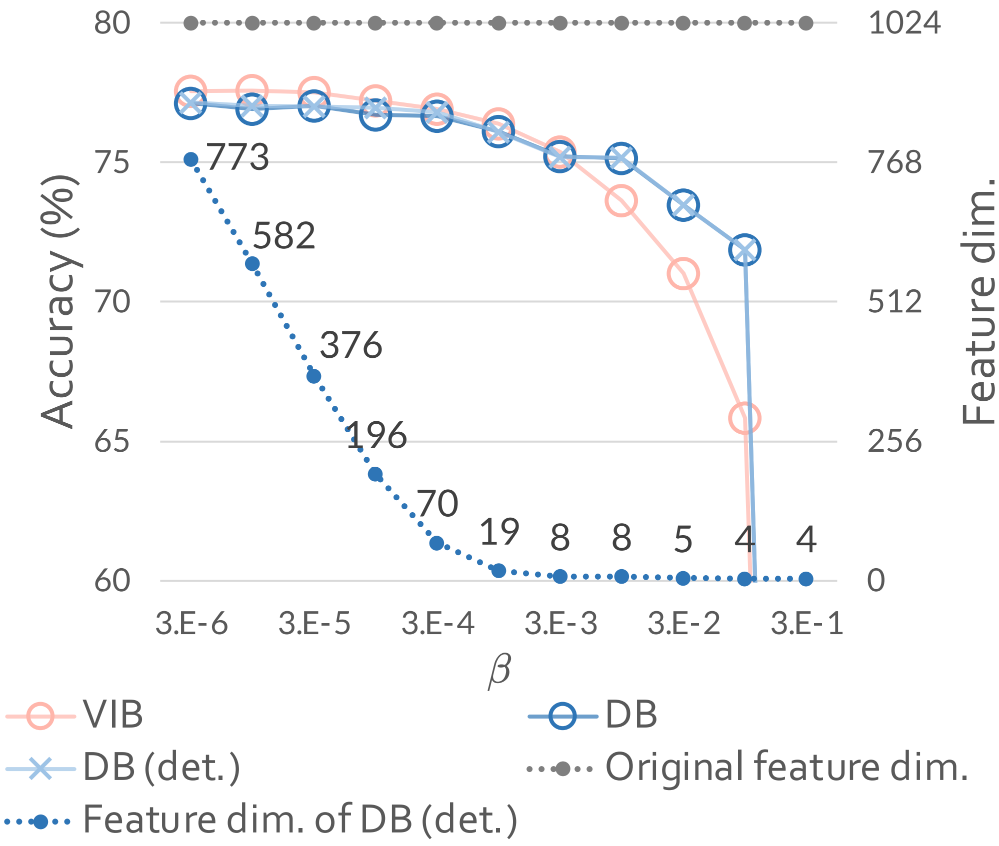


* Alemi et al., 2017, Deep Variational Information Bottleneck

Experiments: Comparison with Variational IB (VIB)

- ImageNet classification accuracy and feature dimensionality of VIB* and DB (+ deterministic) equipped with Inception-ResNet-v2

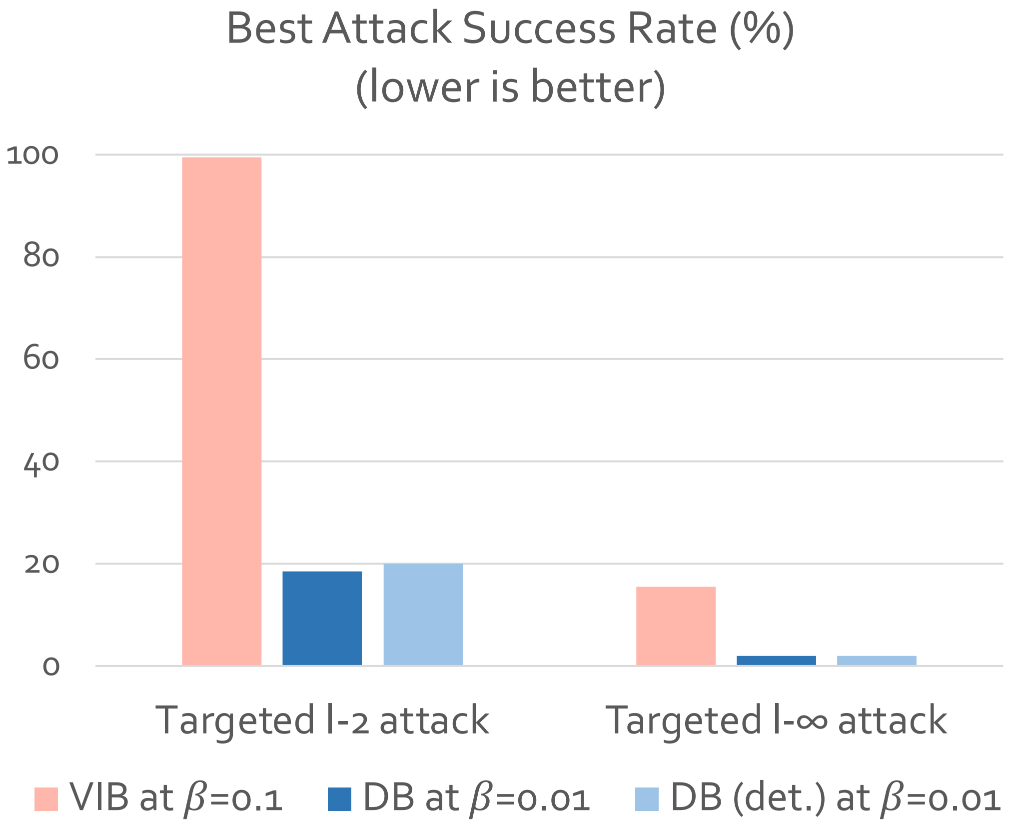
- DB's deterministic representation achieves accuracy $\geq 75\%$ using only 8 features



* Alemi et al., 2017, Deep Variational Information Bottleneck

Experiments: Comparison with Variational IB (VIB)

- Comparison of VIB and DB's robustness to targeted l-2 and l- ∞ adversarial attacks*

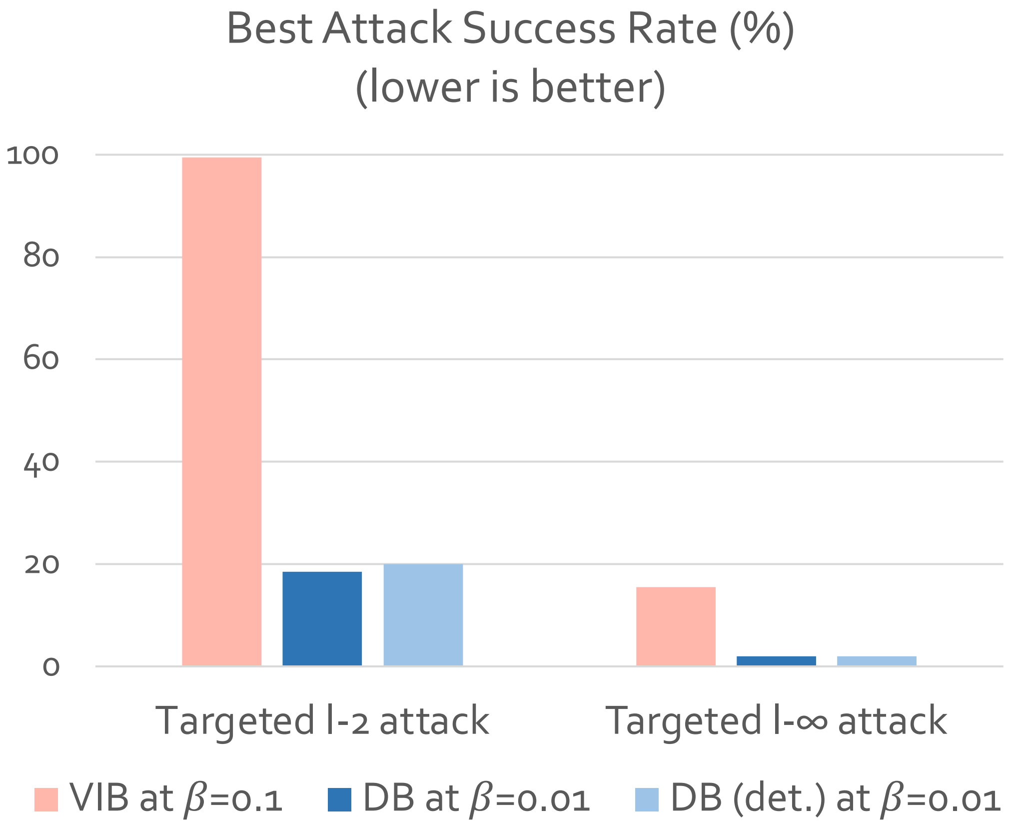


* Carlini & Wagner, 2017, Towards Evaluating the Robustness of Neural Networks

Experiments: Comparison with Variational IB (VIB)

- Comparison of VIB and DB's robustness to targeted l-2 and l- ∞ adversarial attacks*

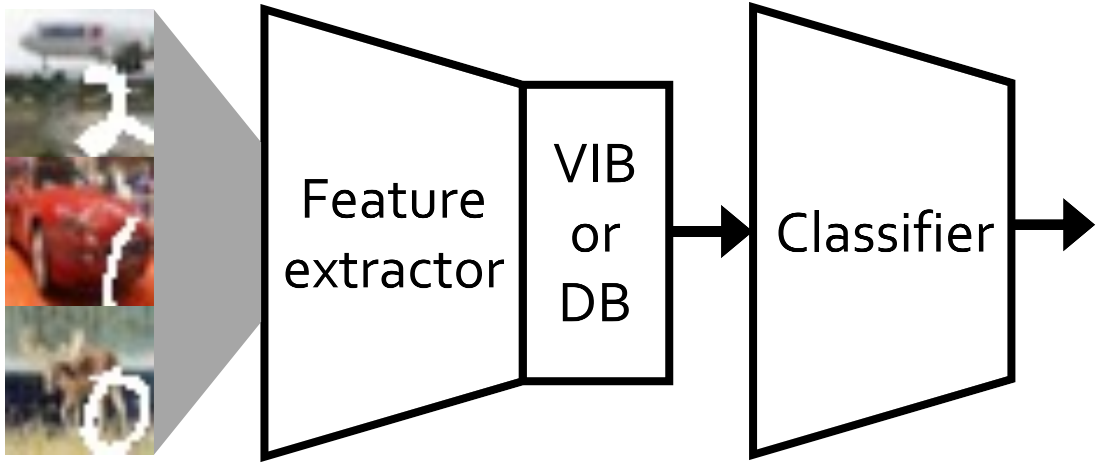
- DB (+ deterministic) shows superior adversarial robustness to VIB



* Carlini & Wagner, 2017, Towards Evaluating the Robustness of Neural Networks

Experiments: Removal of Task-Irrelevant Info

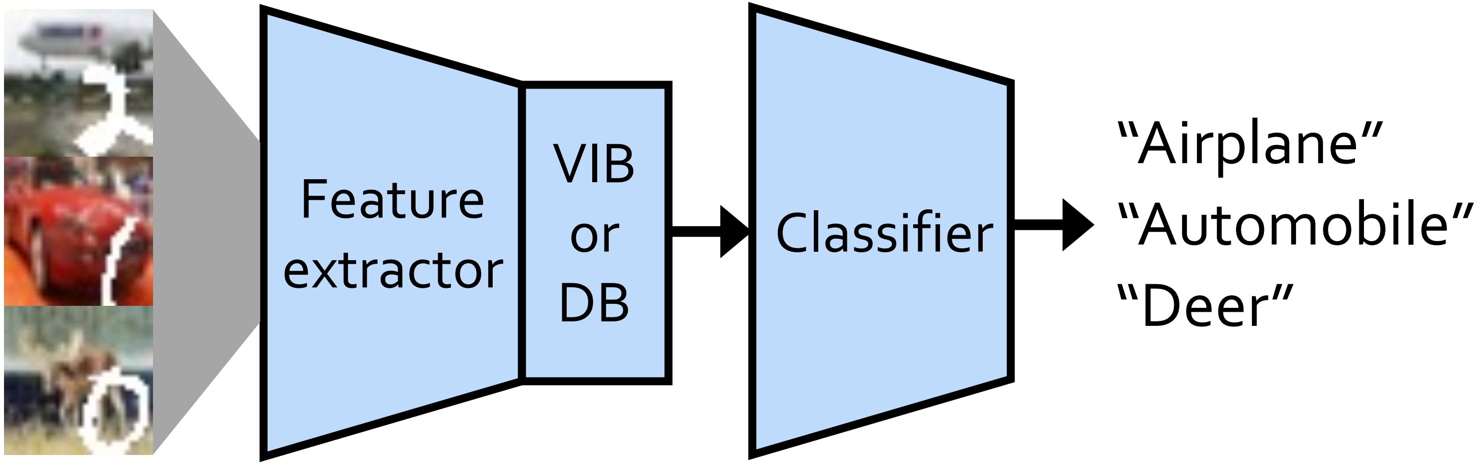
• Testing task-irrelevant information removal using Occluded CIFAR*



* Achille & Soatto, 2018, Information Dropout: Learning Optimal Representations Through Noisy Computation

Experiments: Removal of Task-Irrelevant Info

- Testing task-irrelevant information removal using Occluded CIFAR*

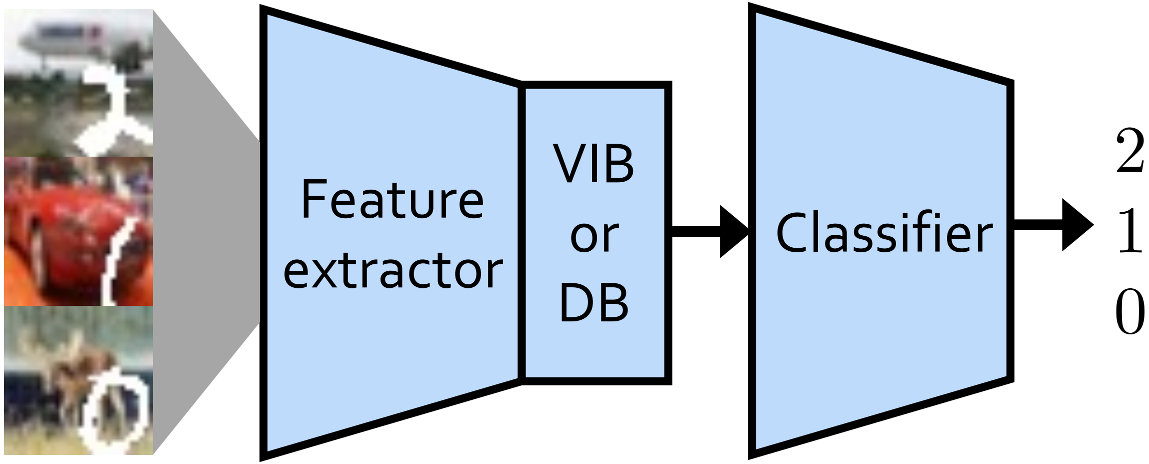


Phase 1: full training with primary (CIFAR) labels

* Achille & Soatto, 2018, Information Dropout: Learning Optimal Representations Through Noisy Computation

Experiments: Removal of Task-Irrelevant Info

- Testing task-irrelevant information removal using Occluded CIFAR*



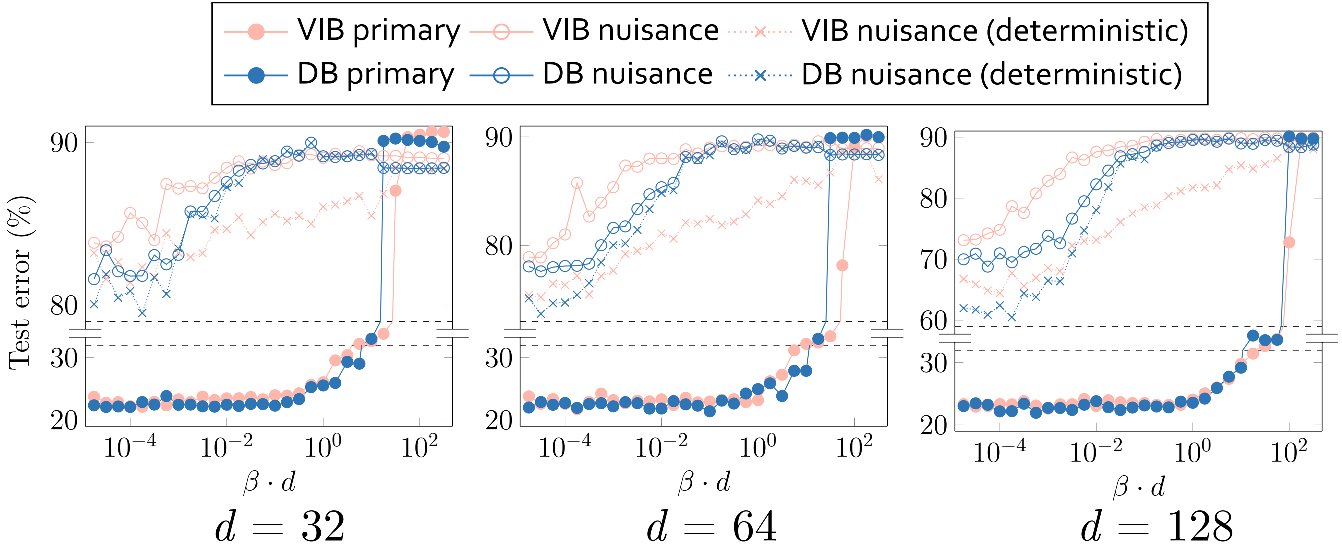
Phase 1: full training with primary (CIFAR) labels

Phase 2: training new classifier only, using nuisance (MNIST) labels

* Achille & Soatto, 2018, Information Dropout: Learning Optimal Representations Through Noisy Computation

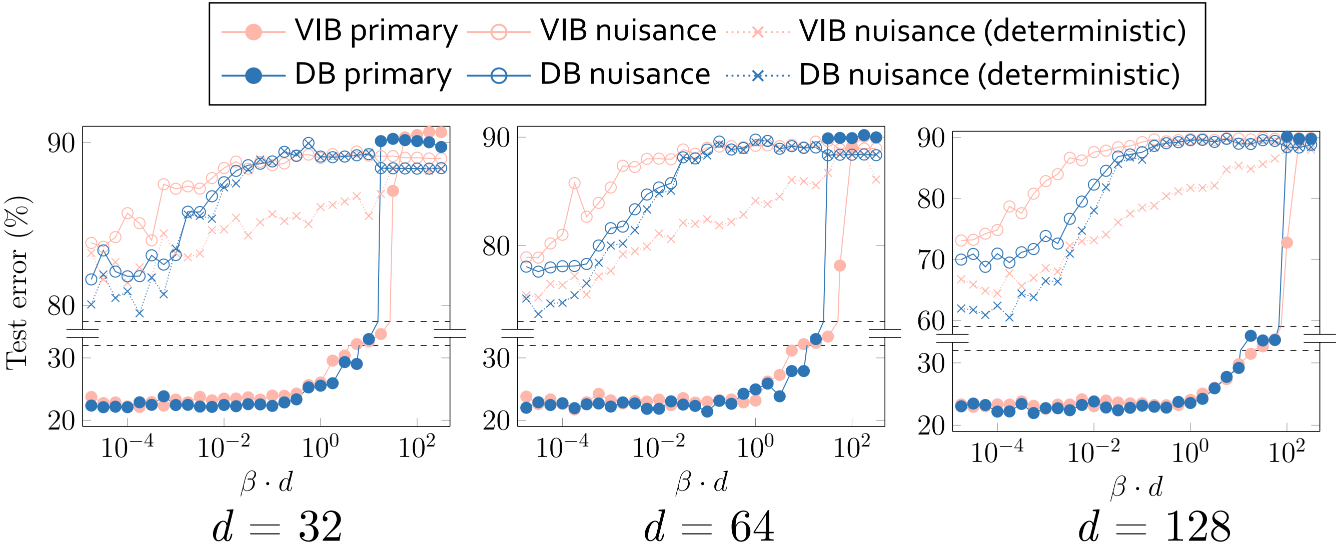
Experiments: Removal of Task-Irrelevant Info

- Nuisance information is removed to maximum first as β grows
  ⇨ controllability over information removal



Experiments: Removal of Task-Irrelevant Info

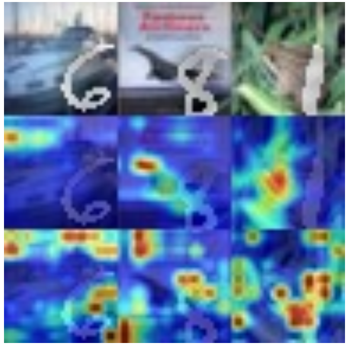
- Nuisance information is removed to maximum first as β grows
  ⇨ controllability over information removal
- Deterministic DB effectively discards irrelevant information as well



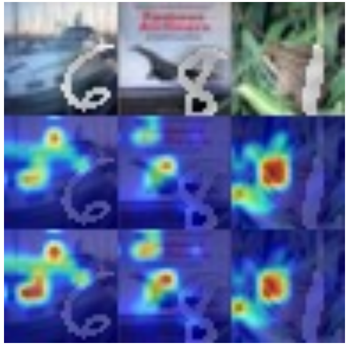
Experiments: Removal of Task-Irrelevant Info

- Nuisance information is removed to maximum first as β grows
  ⇨ controllability over information removal
- Deterministic DB effectively discards irrelevant information as well
- We provide GradCAM* visualization of features used

Raw Image
Primary
Nuisance



Without DB



With DB

* Selvaraju et al., 2017, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Conclusion

- Discrete IB method which jointly learns features and drops task-irrelevant ones
- Provides deterministic representations for inference
- Effective for robustness, feature dimensionality reduction and distilling relevant information
- Achieves SOTA performance in noisy exploration tasks

Thank you

Poster session **6**

May 4, 2021, 5 p.m. (PDT)

https://openreview.net/forum?id=1rxHOBjeDUW

https://vision.snu.ac.kr/projects/db

jaekyeom@snu.ac.kr