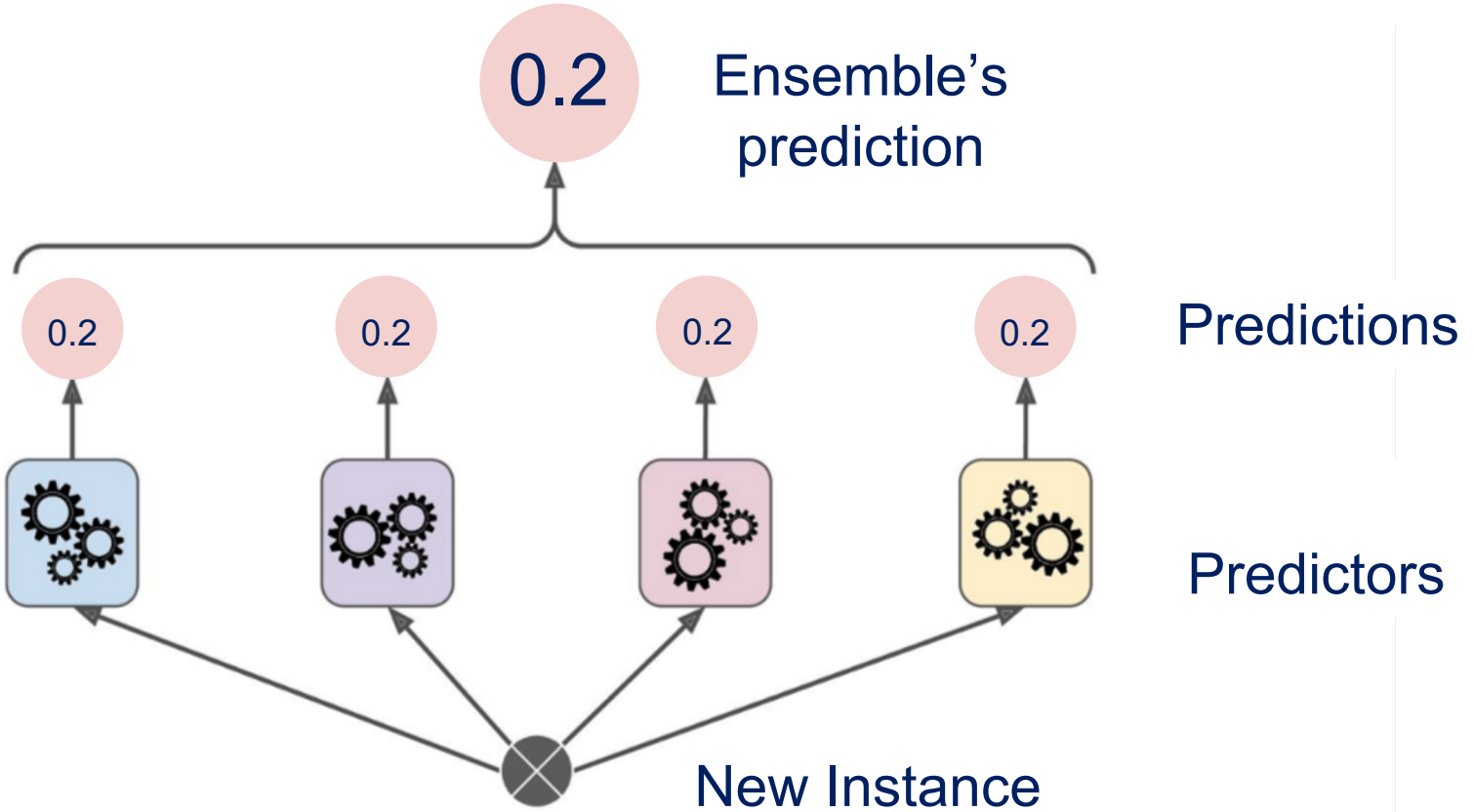
DICE:
Diversity In Deep Ensembles
Via
Conditional Redundancy Adversarial Estimation

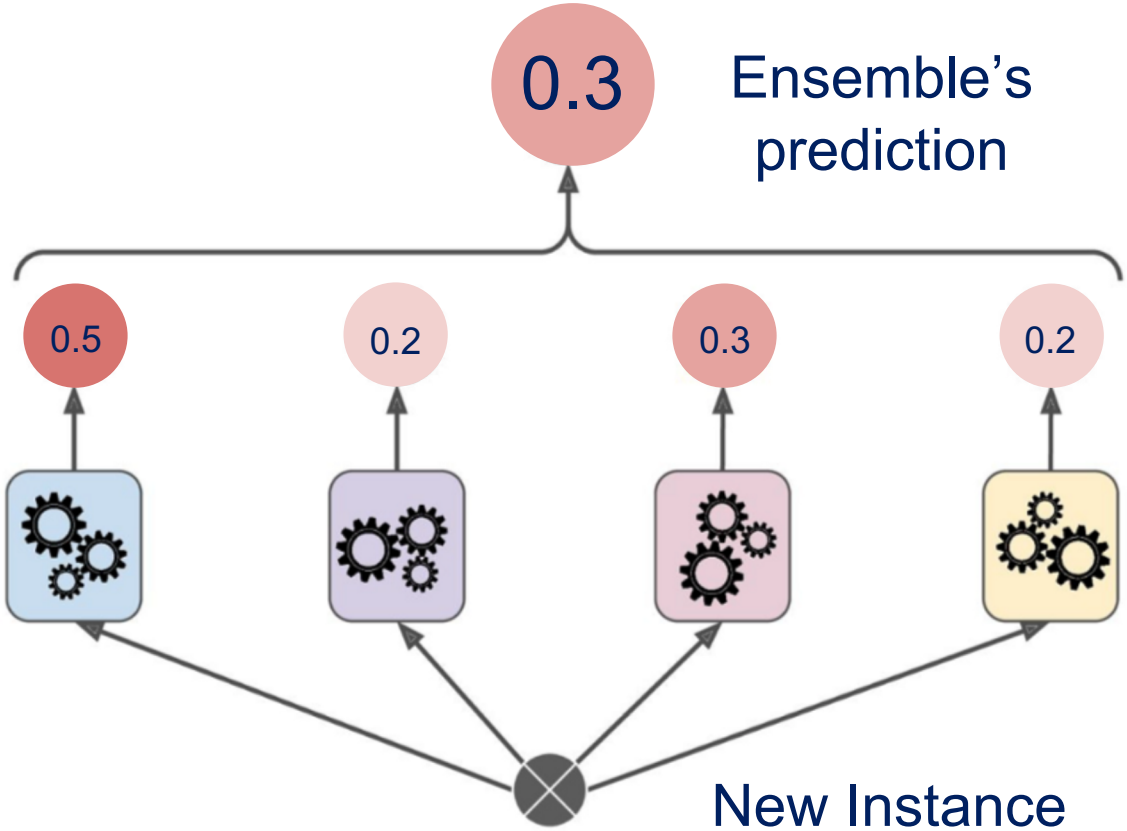
Alexandre Ramé and Matthieu Cord

ICLR 2021

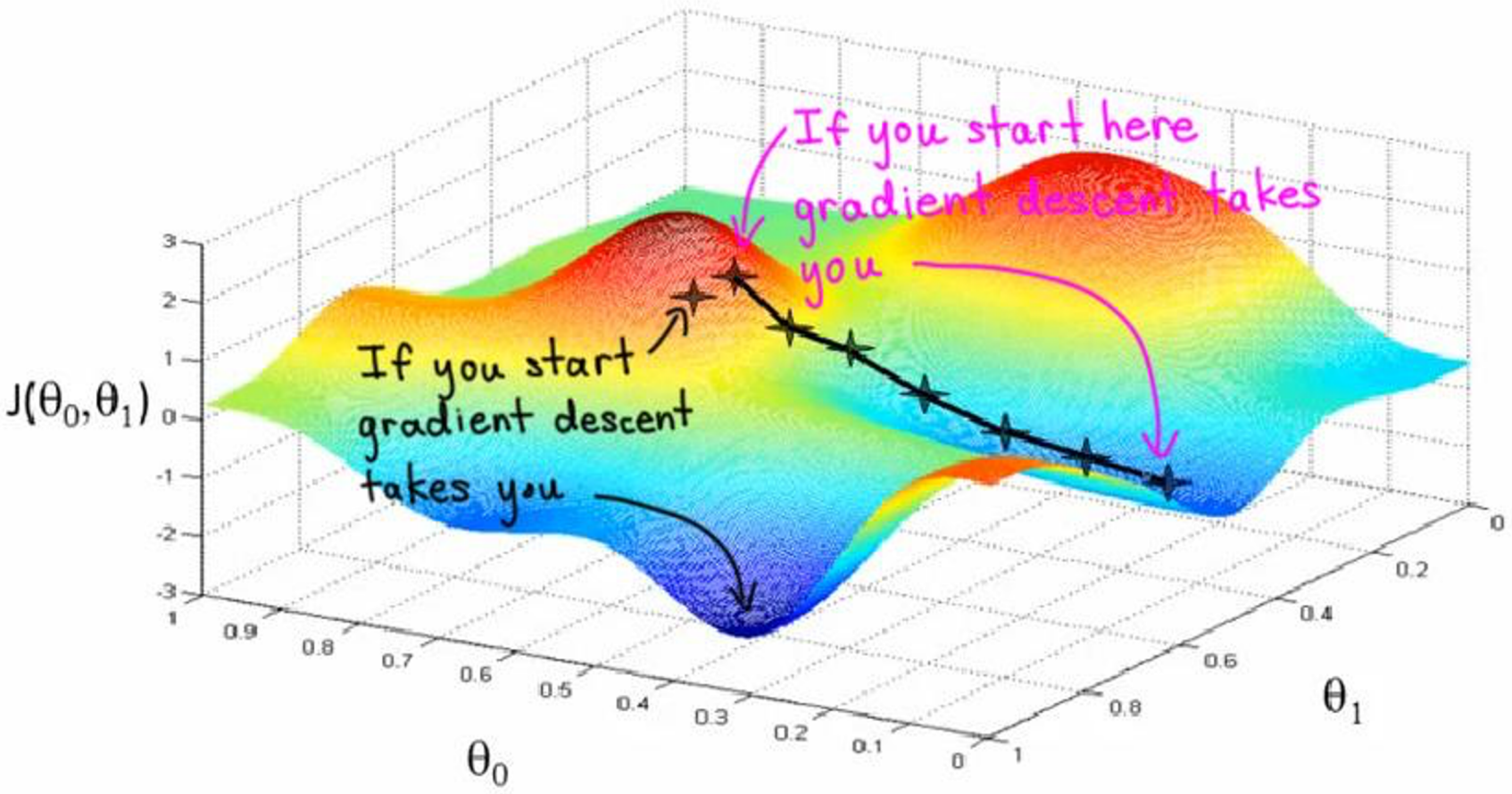
Ensemble


0.2
Ensemble's prediction
0.2
0.2
0.2
0.2
Predictions
Predictors
New Instance

Motivation: diversity in ensemble

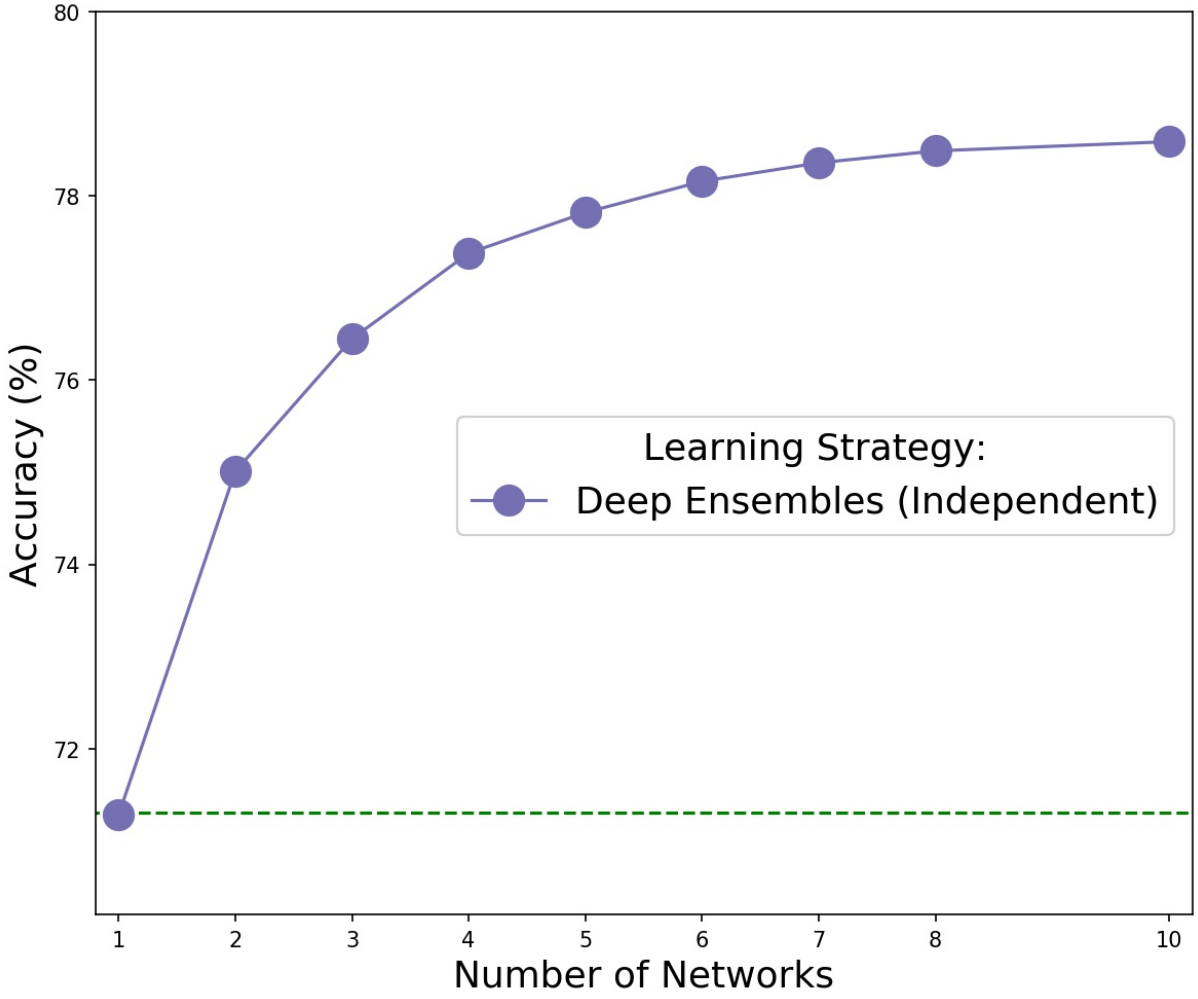


Deep ensembles: randomness in initializations



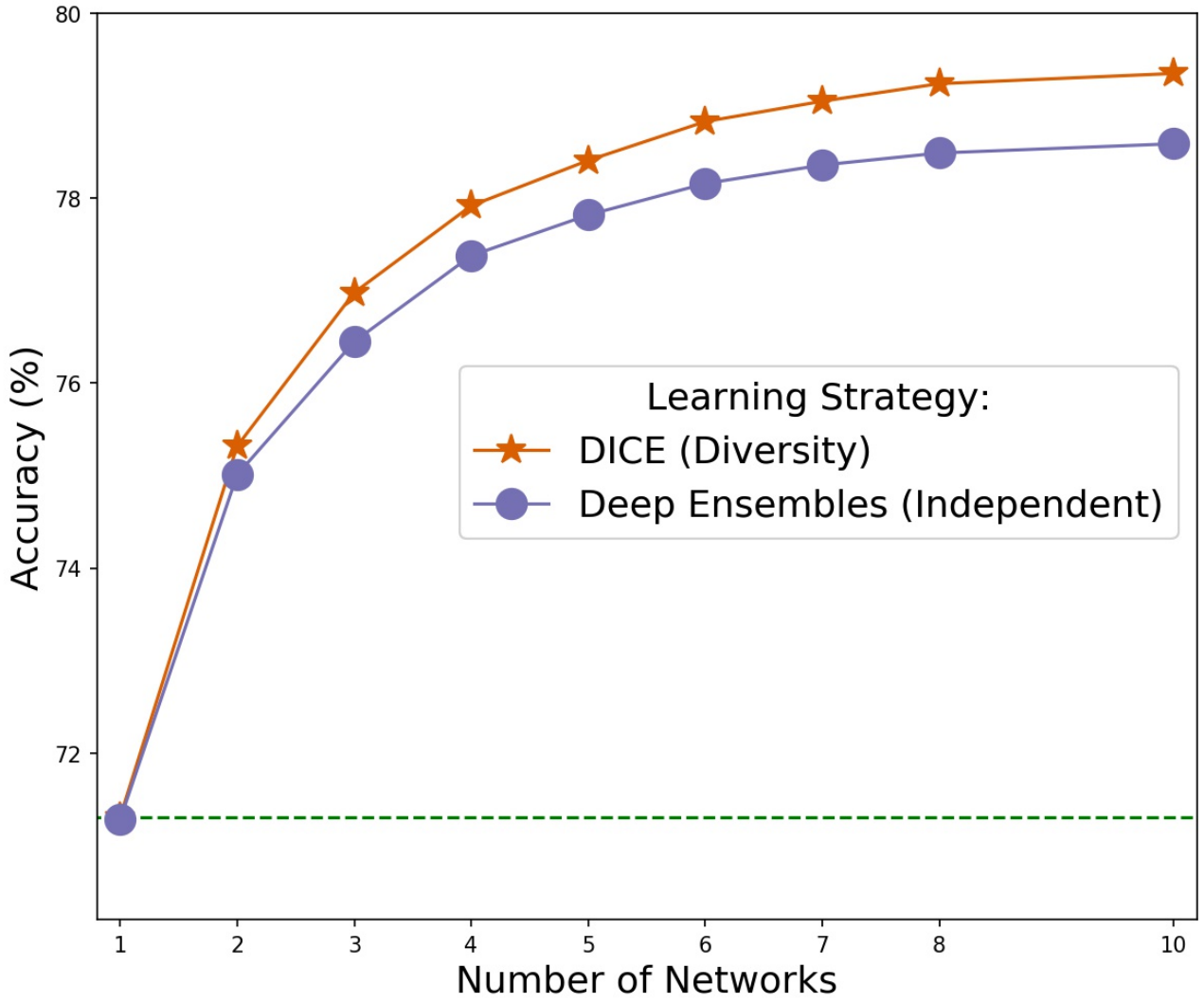
[1] Simple and scalable predictive uncertainty estimation using deep ensembles. Lakshminarayanan *et al.*, in *NeurIPS* 2017.
[2] Deep Ensembles: A Loss Landscape Perspective. Fort *et al.*, 2019.

Deep ensembles success



[1] Simple and scalable predictive uncertainty estimation using deep ensembles. Lakshminarayanan *et al.*, in *NeurIPS* 2017.
[2] Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. Ashukha *et al.*, in *ICLR* 2020.
[3] On Power Laws in Deep Ensembles. Lobacheva *et al.*, in *NeurIPS* 2020.

DICE: beyond independent strategy



Previous ensemble strategies

How to increase diversity ?

❖ Stochasticity in data: bagging

[1] Bagging predictors. Leo Breiman, in *Machine Learning*, 1996.

[2] Why Are Bootstrapped Deep Ensembles Not Better. Nixon *et al.*, in *NeurIPS* workshop 2020.

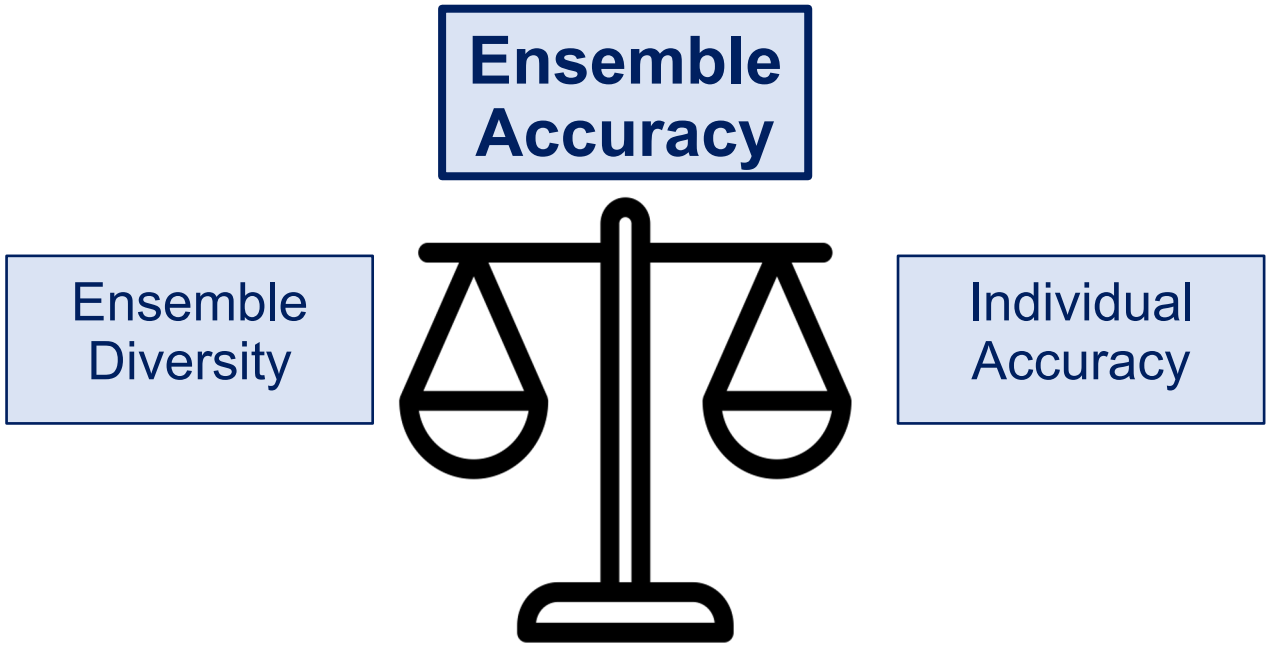
❖ Regularization in predictions

[3] Diversity with cooperation: Ensemble methods for few-shot classification. Dvornik *et al.*, in *ICCV*, 2019.

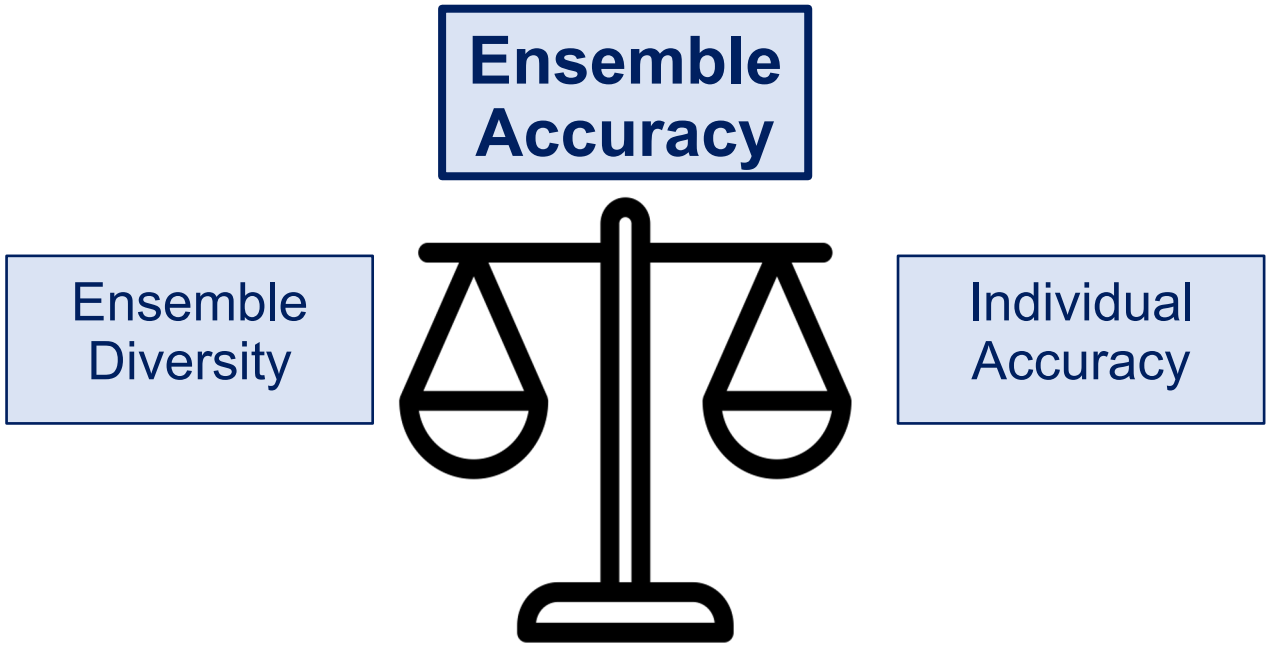
[4] Improving adversarial robustness via promoting ensemble diversity. Pang *et al.*, in *ICML* 2019.

→ reduced individual performances

Trade off: ensemble diversity vs. individual accuracy



Our intuition: diversity in features



Same predictions but for different reasons

=> diversity in features

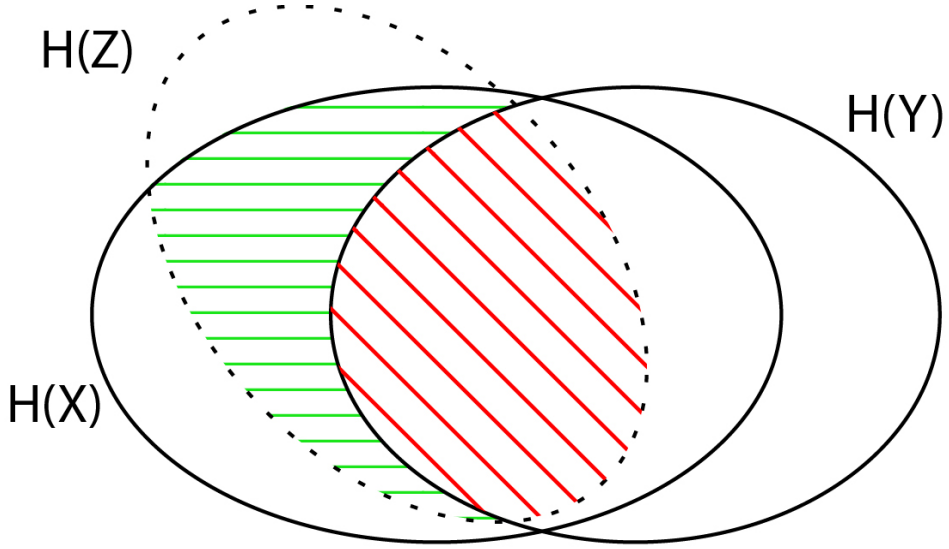
Information bottleneck theory

Nobody knows what entropy really is.

John Van Neumann to *Claude Shannon*

Background: information bottleneck

Information in features Z:
- Sufficient to predict Y
=> $I(Y;Z) \uparrow$
- Compressed wrt X
=> $I(X;Z|Y) \downarrow$



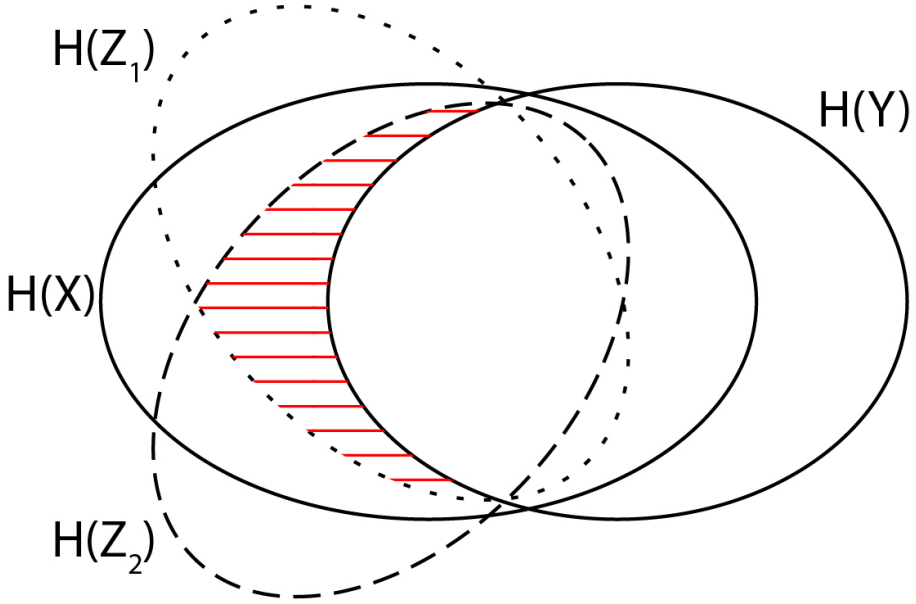
$$CEB(Z_i) = -I(Y;Z_i) + \frac{1}{\beta} I(X;Z_i|Y)$$

[1] The Information Bottleneck Method. Tishby, 1999.
[2] Deep variational information bottleneck. Alemi *et al.*, in *ICLR* 2017.
[3] The Conditional Entropy Bottleneck. Fischer, 2020.

DICE: information bottleneck in deep ensembles

Information in features:
- Sufficient
- Compressed

Thus, not redundant
=> $I(Z_1; Z_2|Y) \downarrow$



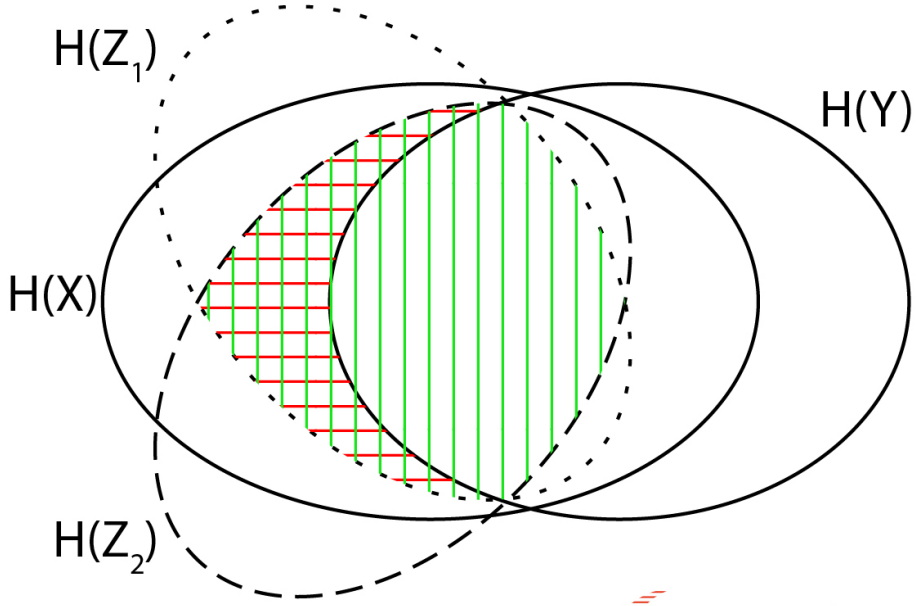
$$DICE = min_{Z_1,Z_2} CEB(Z_1) + CEB(Z_2) + \delta I(Z_1; Z_2|Y)$$

(for $M = 2$ networks)

DICE: information bottleneck in deep ensembles

Information in features:
- Sufficient
- Compressed

Thus, not redundant
=> $I(Z_1; Z_2|Y) \downarrow$



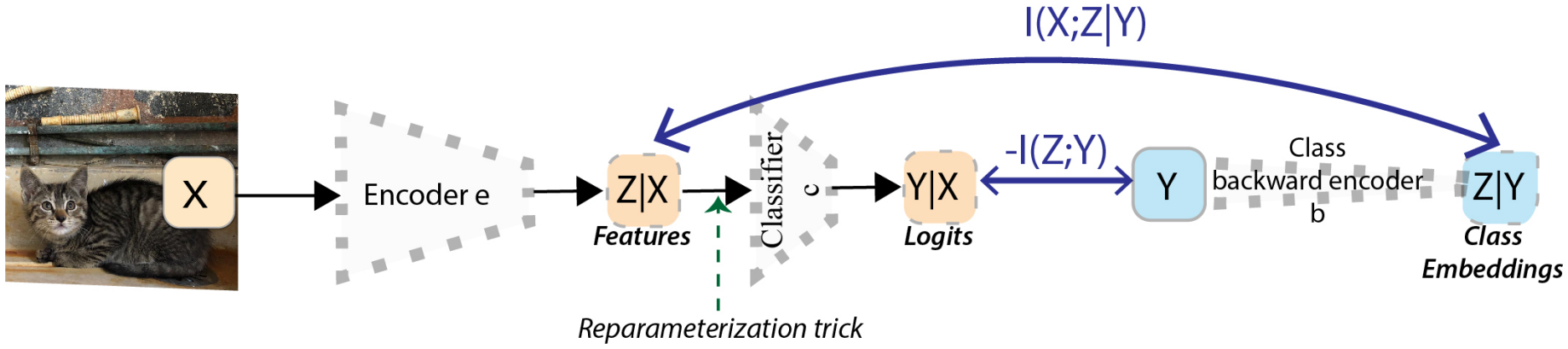
$$DICE = min_{Z_1,Z_2} CEB(Z_1) + CEB(Z_2) + \delta I(Z_1; Z_2|Y)$$

Conditioning on label Y:
- Reduce spurious correlations
- Independent bias, but not independent features

Transforming DICE into a Tractable Loss

$$DICE = min_{Z_1, Z_2} CEB(Z_1) + CEB(Z_2) + \delta I(Z_1; Z_2 | Y)$$

VCEB: variational approximation loss



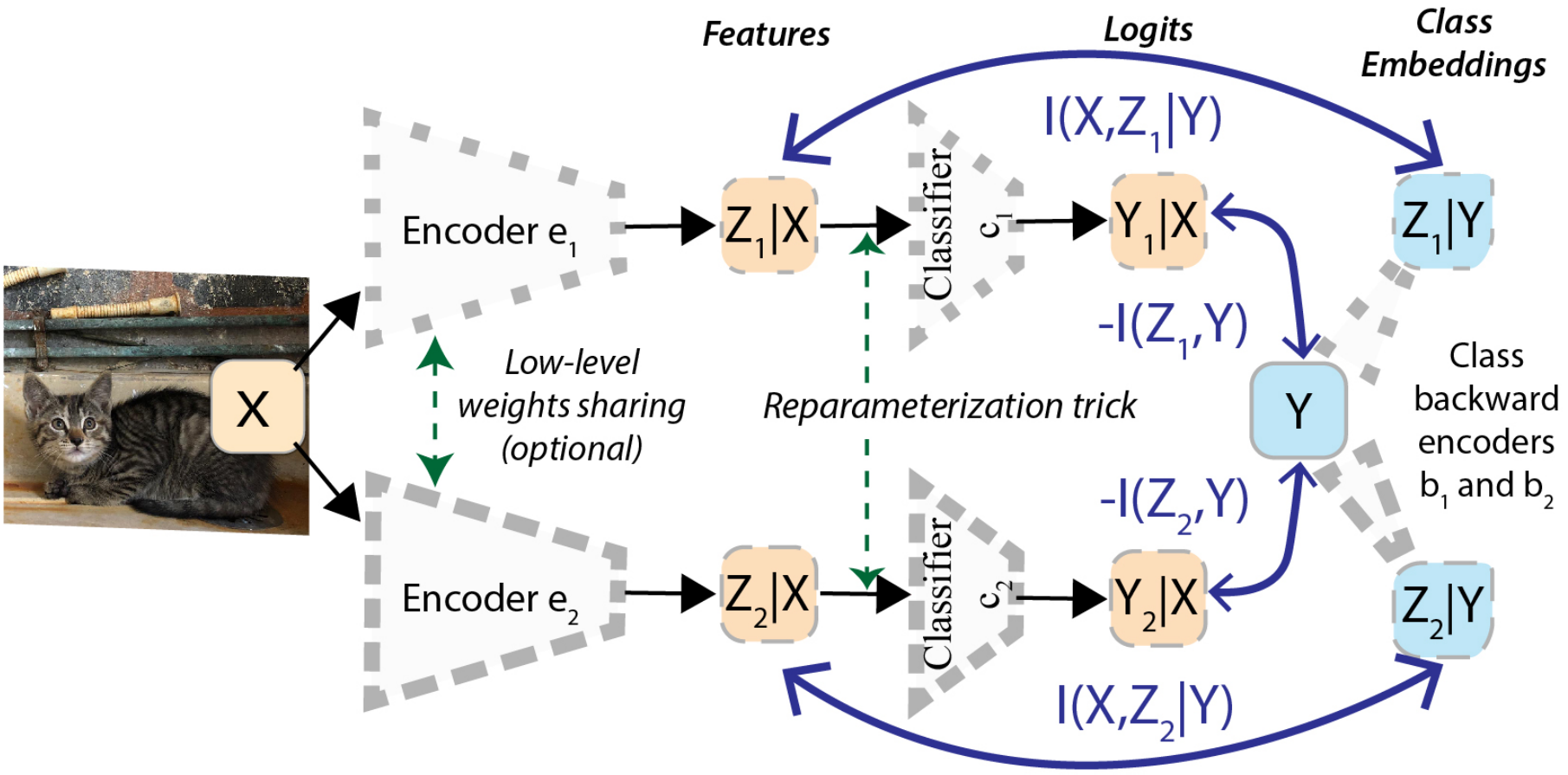
$$CEB(Z_i) \leq VCEB(\{e_i, c_i, b_i\})$$
$$\approx \frac{1}{N} \sum_{n=1,..,N} \left[\frac{1}{\beta} D_{KL}\left(e_i(z|x^n) \,\|\, b_i(z|y^n)\right) - \mathbb{E}_\varepsilon log\left(c_i\left(y^n|e_i(x^n,\varepsilon)\right)\right)\right]$$

Where:

- e_i distribution encoder of input x
- c_i classifier that targets class y
- b_i class y backward encoder

[1] The Conditional Entropy Bottleneck. Fischer, 2020.

Architecture: ensemble



Step 1: classification with conditional entropy bottleneck

Transforming DICE into a Tractable Loss

$$DICE = min_{Z_1, Z_2} CEB(Z_1) + CEB(Z_2) + \delta I(Z_1; Z_2 | Y)$$

No Markov properties between Z_1, Z_2 and Y
=> No variational approximation

Neural estimation of conditional redundancy

Donsker-Varadhan representation:

$$I(Z_1; Z_2|Y) = \mathbb{E}_{t\sim P(Z_1,Z_2,Y)}[f^*(t)] - log\{\mathbb{E}_{t'\sim P(Z_1,Y)P(Z_2|Y)}[exp(f^*(t'))]\}$$

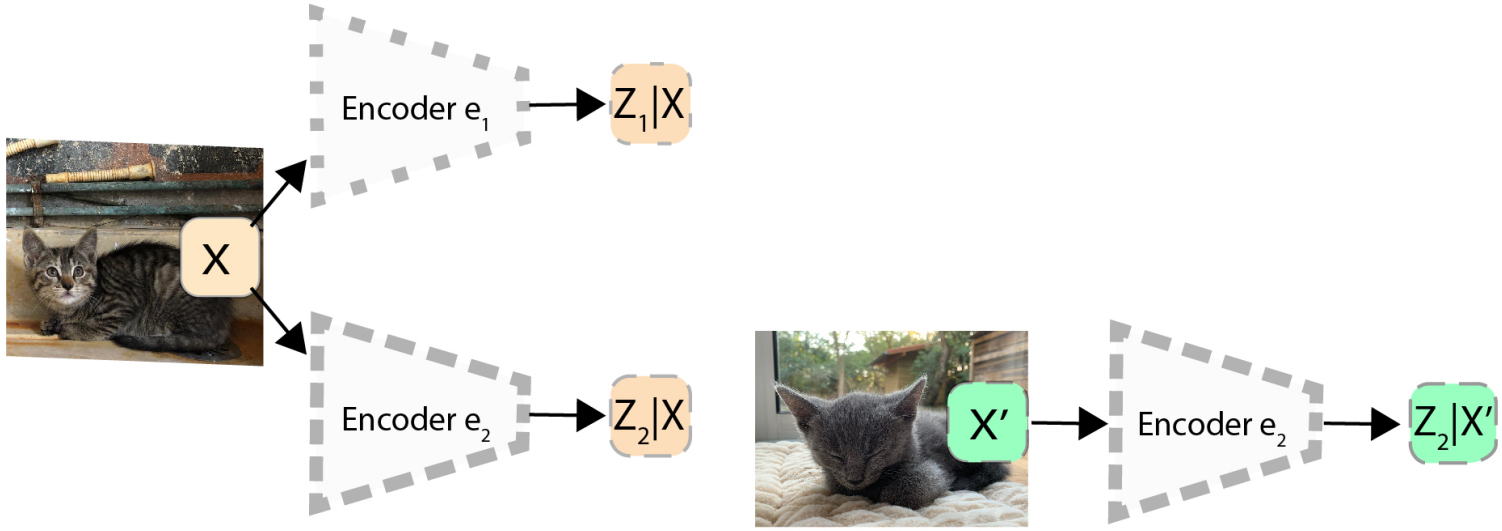
with f^* the pointwise likelihood ratio:

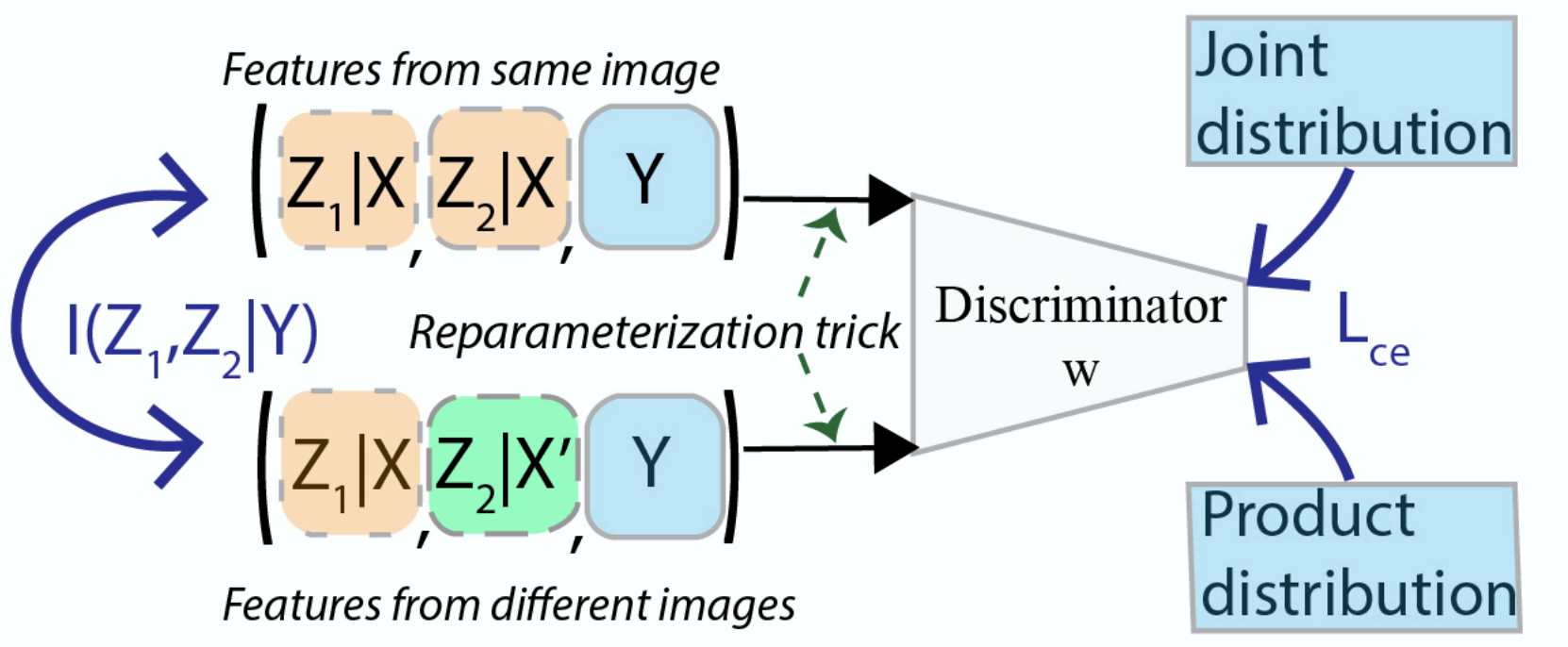
$$f^*(z_1, z_2, y) = \frac{p(z_1, z_2, y)}{p(z_1, y)p(z_2|y)} \approx \frac{w(z_1, z_2, y)}{1 - w(z_1, z_2, y)}$$

and w a neural network trained as a discriminator.

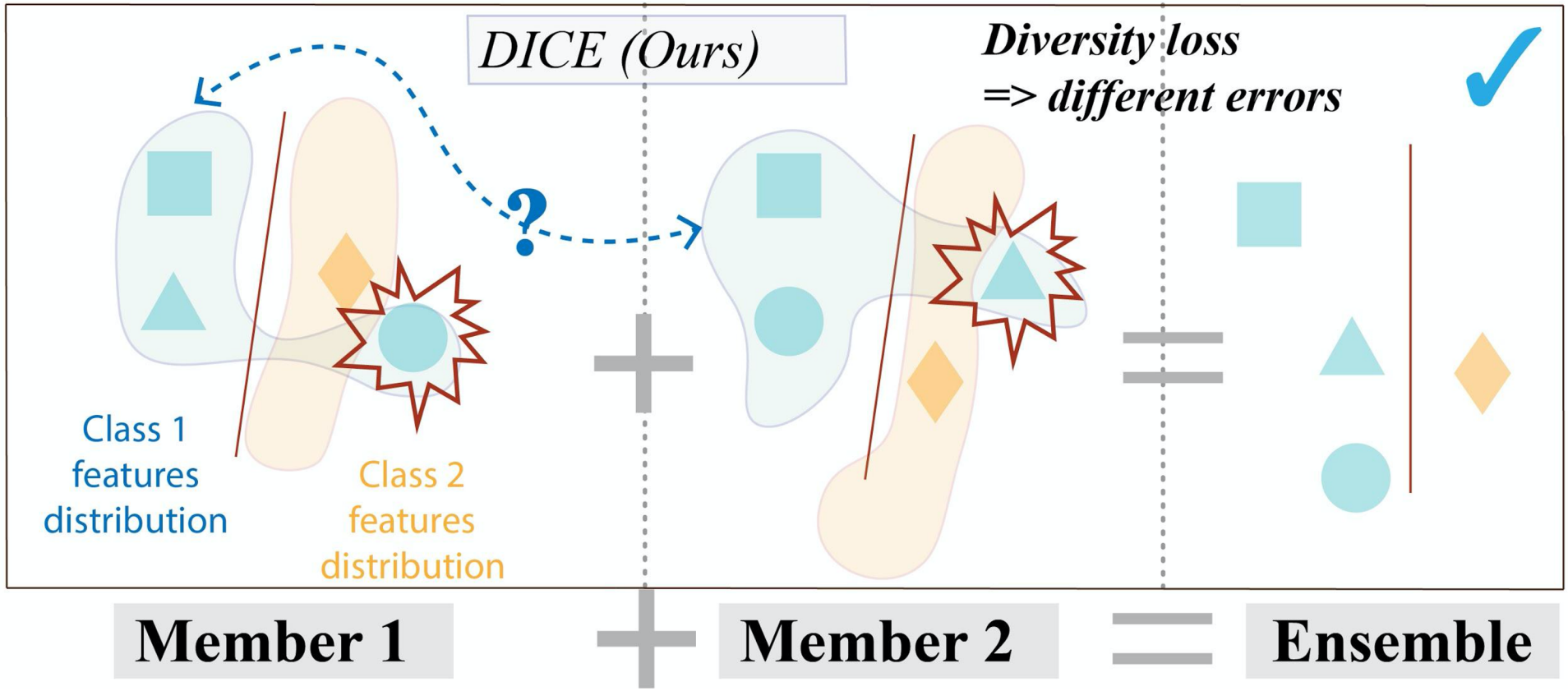
Batch:	Tuple of two features extracted from:
Joint distribution	The same image
Product distribution	Two different images from same label Y

[1] Asymptotic evaluation of certain markov process expectations for large time. Donsker *et al.*, 1975.
[2] Mutual information neural estimation. Belghazi *et al.*, in *NeurIPS* 2016.

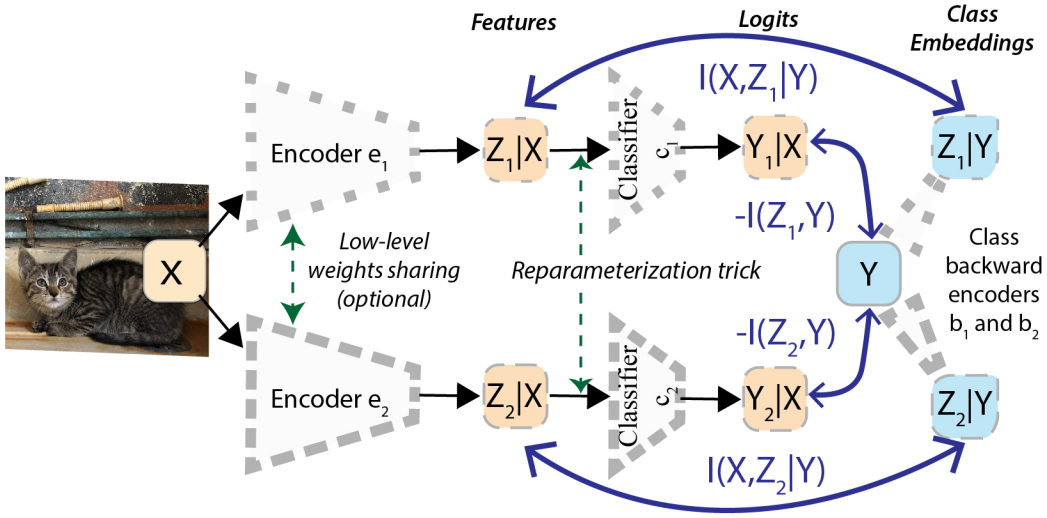

Encoder e_1
$Z_1|X$
X
Encoder e_2
$Z_2|X$
X′
Encoder e_2
$Z_2|X'$


Features from same image
$Z_1|X$
$Z_2|X$
Y
$I(Z_1,Z_2|Y)$
Reparameterization trick
Discriminator
w
L_{ce}
$Z_1|X$
$Z_2|X'$
Y
Features from different images
Joint distribution
Product distribution

DICE: features unpredictable from each other



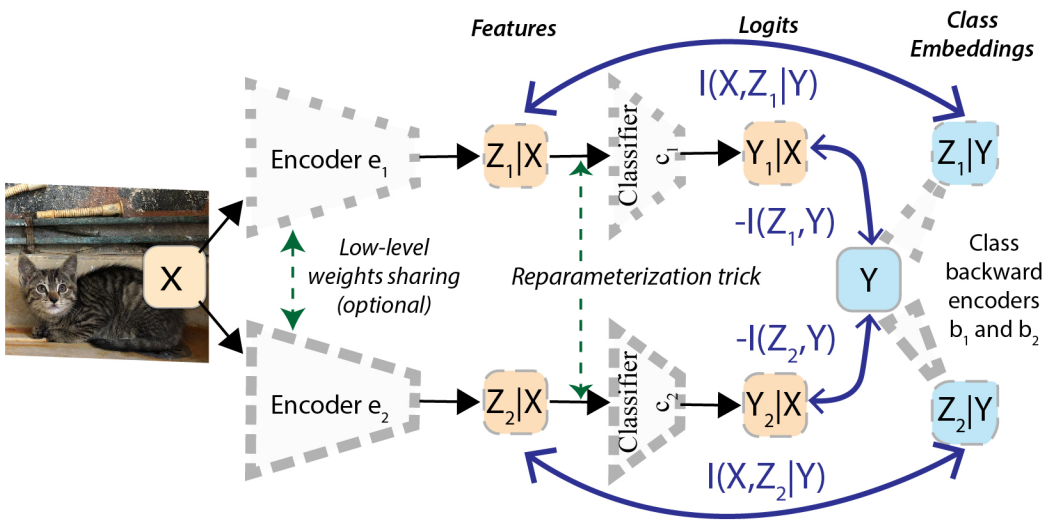
Full training strategy



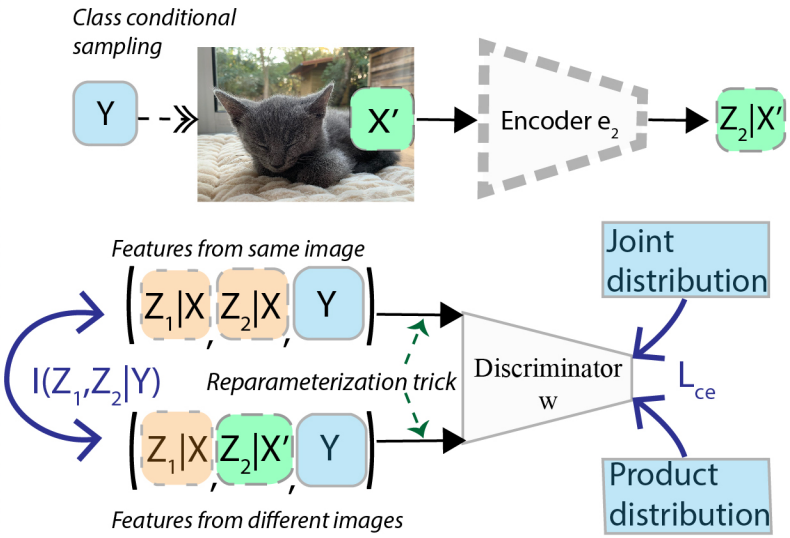
Step 1: classification with conditional entropy bottleneck

$$\mathcal{L}_{\text{DICE}}(\{e_1, c_1, b_1\}, \{e_2, c_2, b_2\}) = VCEB_\beta(\{e_1, c_1, b_1\}) + VCEB_\beta(\{e_2, c_2, b_2\})$$

Full training strategy



Step 1: classification with conditional entropy bottleneck



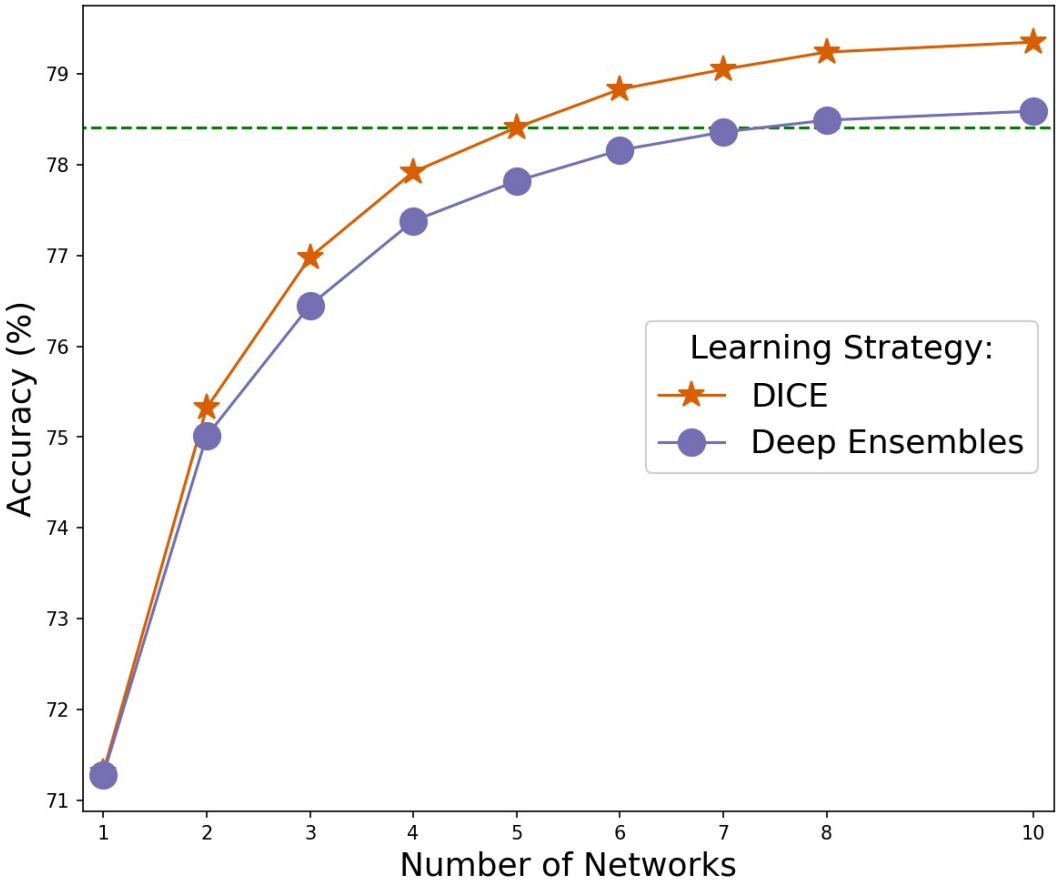
Step 2: diversity with adversarial training

$$\mathcal{L}_{\text{DICE}}(\{e_1, c_1, b_1\}, \{e_2, c_2, b_2\}) = VCEB_\beta(\{e_1, c_1, b_1\}) + VCEB_\beta(\{e_2, c_2, b_2\}) + \delta \mathcal{L}_{DV}^{CR}(e_1, e_2)$$

General case with M members, $M > 2$

$$\mathcal{L}_{\text{DICE}}\left(\{e_i, c_i, b_i\}_{i=1,..,M}\right)$$
$$= \sum_{i=1}^{M} \left[VCEB_\beta(\{e_i, c_i, b_i\}) + \frac{\delta}{M-1} \sum_{j=i+1}^{M} \mathcal{L}_{DV}^{CR}\left(e_i, e_j\right) \right]$$

Main results on CIFAR-100 with ResNet-32



✓ 5 networks with DICE match 7 independent networks

State of the art ensemble on CIFAR-100

Network	Resnet-32			Resnet-110	WRN-28-2	
Architecture	Branch	Branch	Net	Branch	Branch	Net
Size	3	4	4	4	3	3
DE [1]	76.28	76.78	77.38	80.89	78.83	80.01
ADP [2]	76.37	77.21	77.51	81.40	79.21	80.01
CEB [3]	76.36	76.98	77.64	81.17	78.92	80.38
CEBR (Our)	76.72	77.30	77.82	81.55	79.25	80.35
DICE (Our)	**76.89**	**77.51**	**77.92**	**81.93**	**79.59**	**80.55**

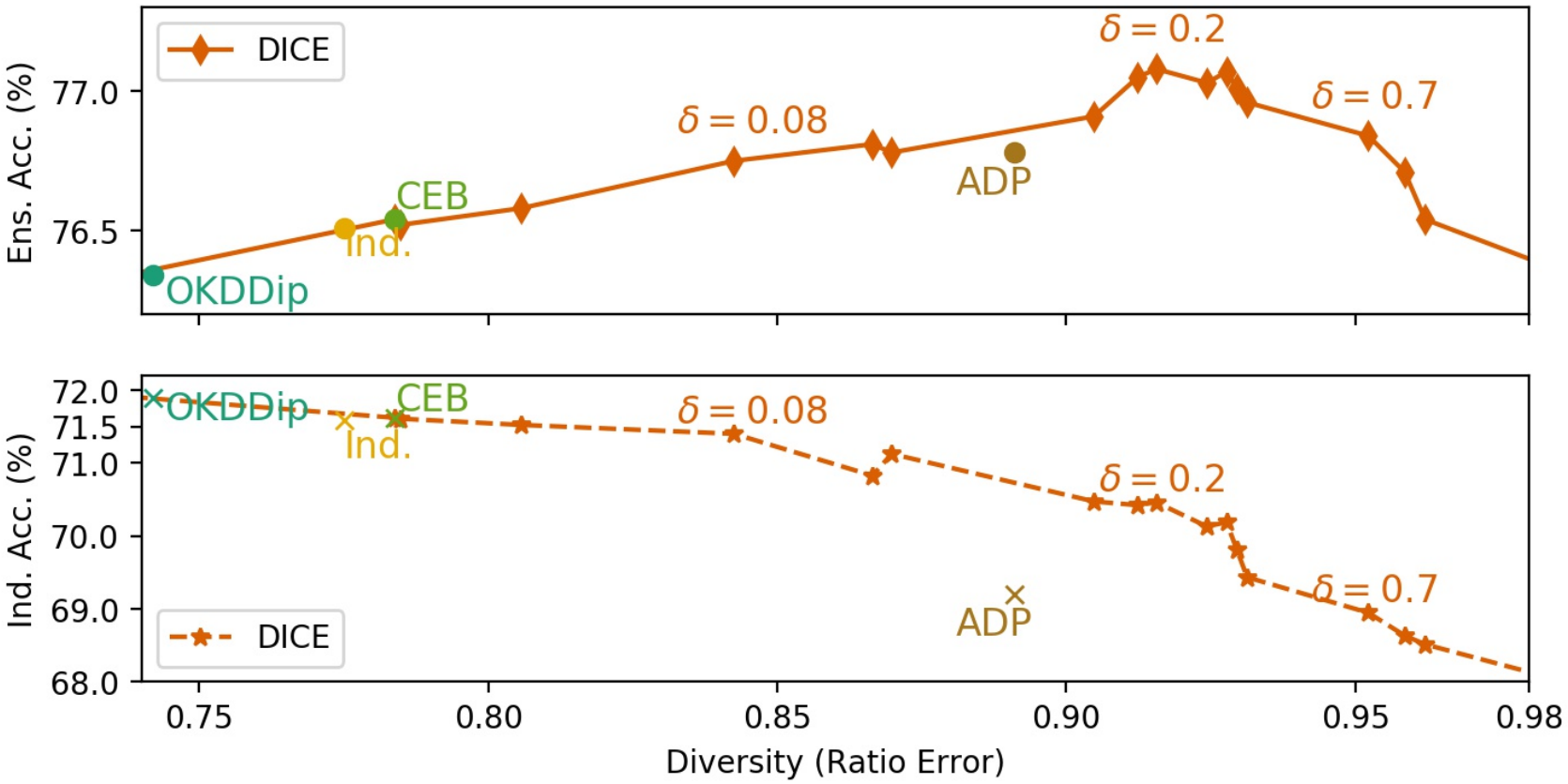
- ✓ $\{+0.52, +0.30, +0.41\}$ for $\{3, 4, 5\}$-branches ResNet-32 wrt. previous sota
- ✓ $\{+0.94, +0.53\}$ for $\{3, 4\}$-branches ResNet-110 wrt. previous sota

[1] Simple and scalable predictive uncertainty estimation using deep ensembles. Lakshminarayanan *et al.*, in *NeurIPS* 2017.

[2] Improving adversarial robustness via promoting ensemble diversity. Pang *et al.*, in *ICML* 2019.

[3] The Conditional Entropy Bottleneck. Fischer, 2020.

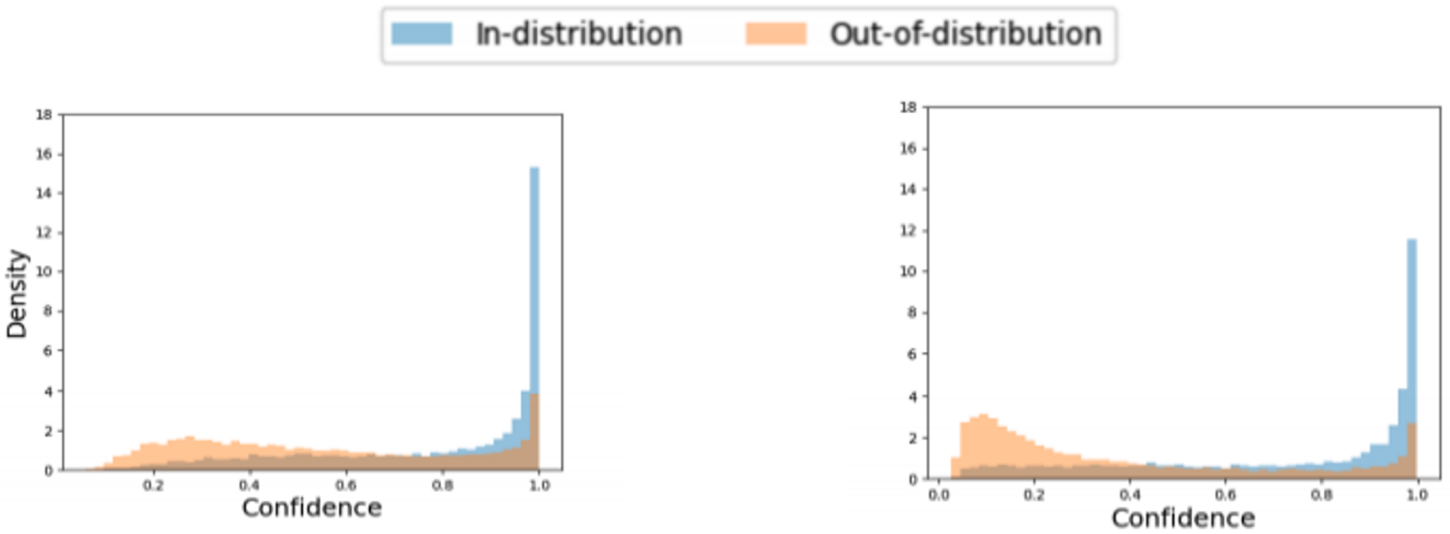
Trade off: ensemble diversity vs. individual accuracy



$$DICE = min_{Z_1,Z_2} CEB(Z_1) + CEB(Z_2) + \delta I(Z_1; Z_2|Y)$$

Improved robustness and OOD detection

Metric (after TS)	1-net	DE	ADP	CEB	CEBR (Our)	DICE (Our)
NLL ↓	10.38	8.10	8.51	8.11	8.05	**7.98**
Brier Score ↓	3.92	3.24	3.27	3.19	3.17	**3.12**



Ind.+TS (74.0)

DICE×w (**77.2**)

Contributions

- Theoretically
  - ✓ Information bottleneck for deep ensembles
  - ✓ Neural estimation of conditional mutual information
  - ✓ New adversarial learning framework

- Empirically
  - ✓ State of the art ensemble on CIFAR-100 and CIFAR-10
  - ✓ Control ensemble diversity vs. individual accuracy trade-off

  => More in paper: uncertainty, calibration, OOD, co-distillation …

DICE:
Diversity In Deep Ensembles
via
Conditional Redundancy
Adversarial Estimation

Merci !

Alexandre Ramé and Matthieu Cord

https://openreview.net/forum?id=R2ZlTVPxOGk