Deep Deterministic Uncertainty: A New Simple Baseline

THU-PM-362


Jishnu Mukhoti * [1,2]


Andreas Kirsch * [1]

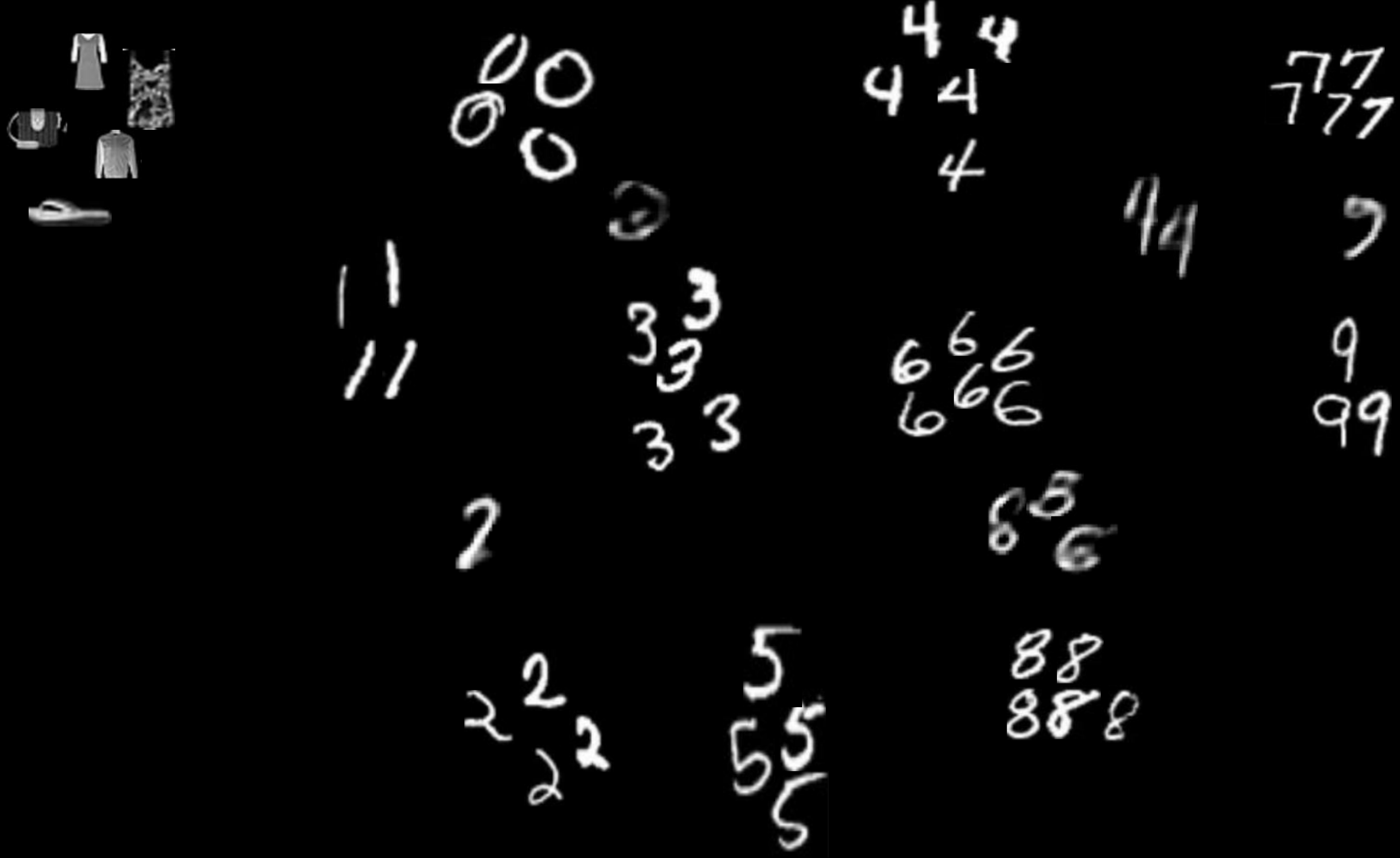

Joost van Amersfoort [1]


Philip H.S. Torr [2]

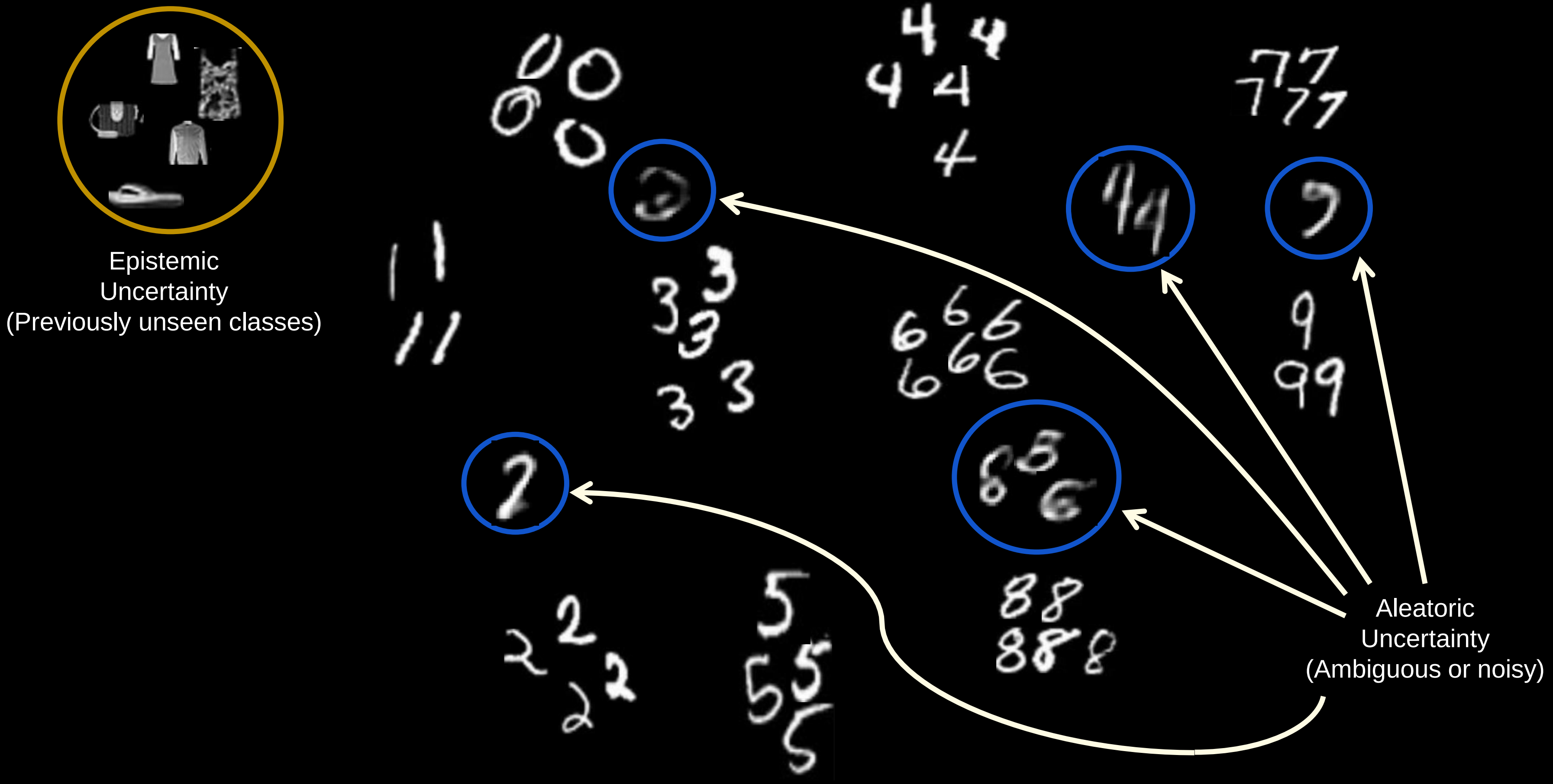

Yarin Gal [1]

(* - joint first authors)
1 - Oxford Applied and Theoretical Machine Learning Group (OATML), University of Oxford
2 - Torr Vision Group (TVG), University of Oxford

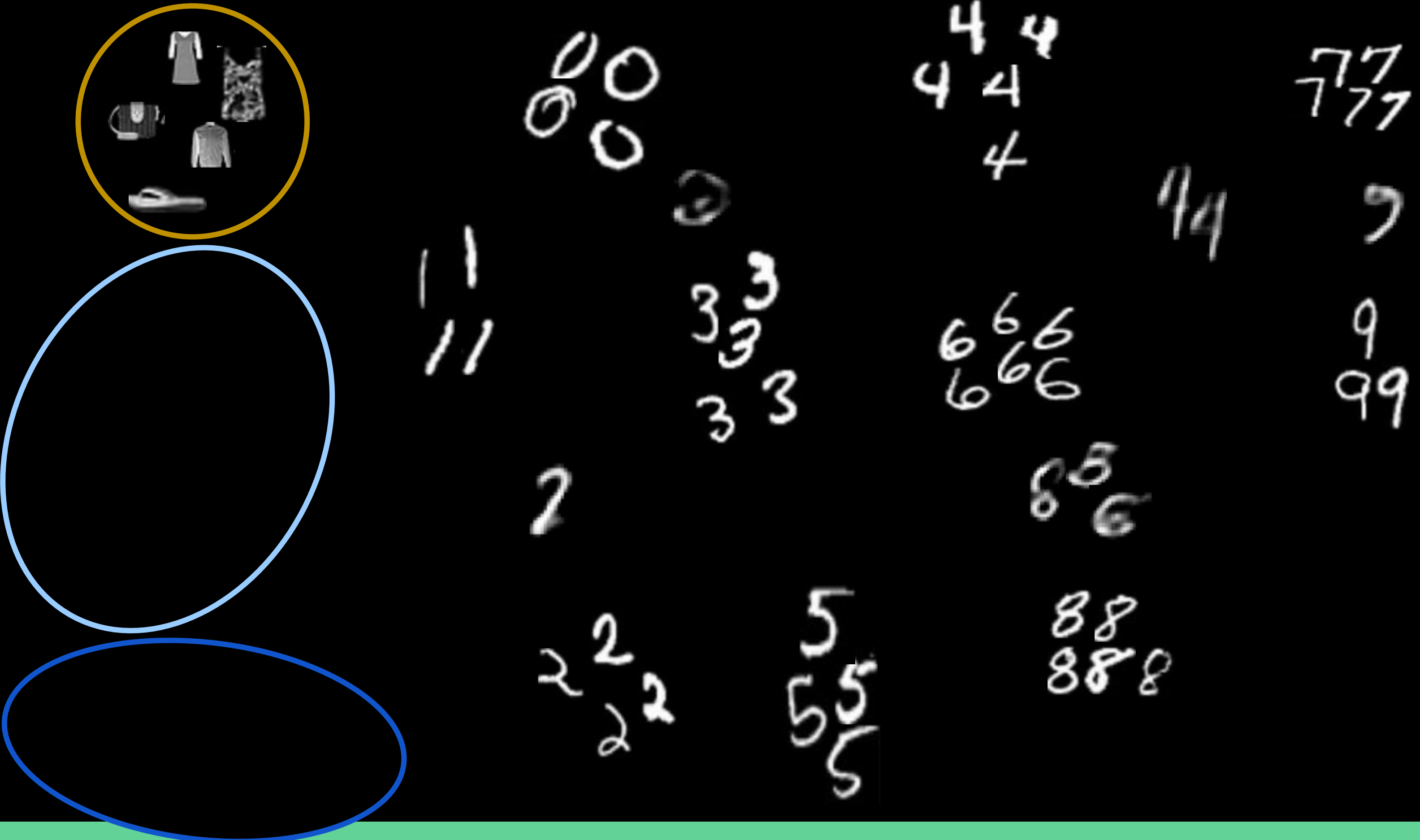
Epistemic & Aleatoric Uncertainty



Epistemic & Aleatoric Uncertainty

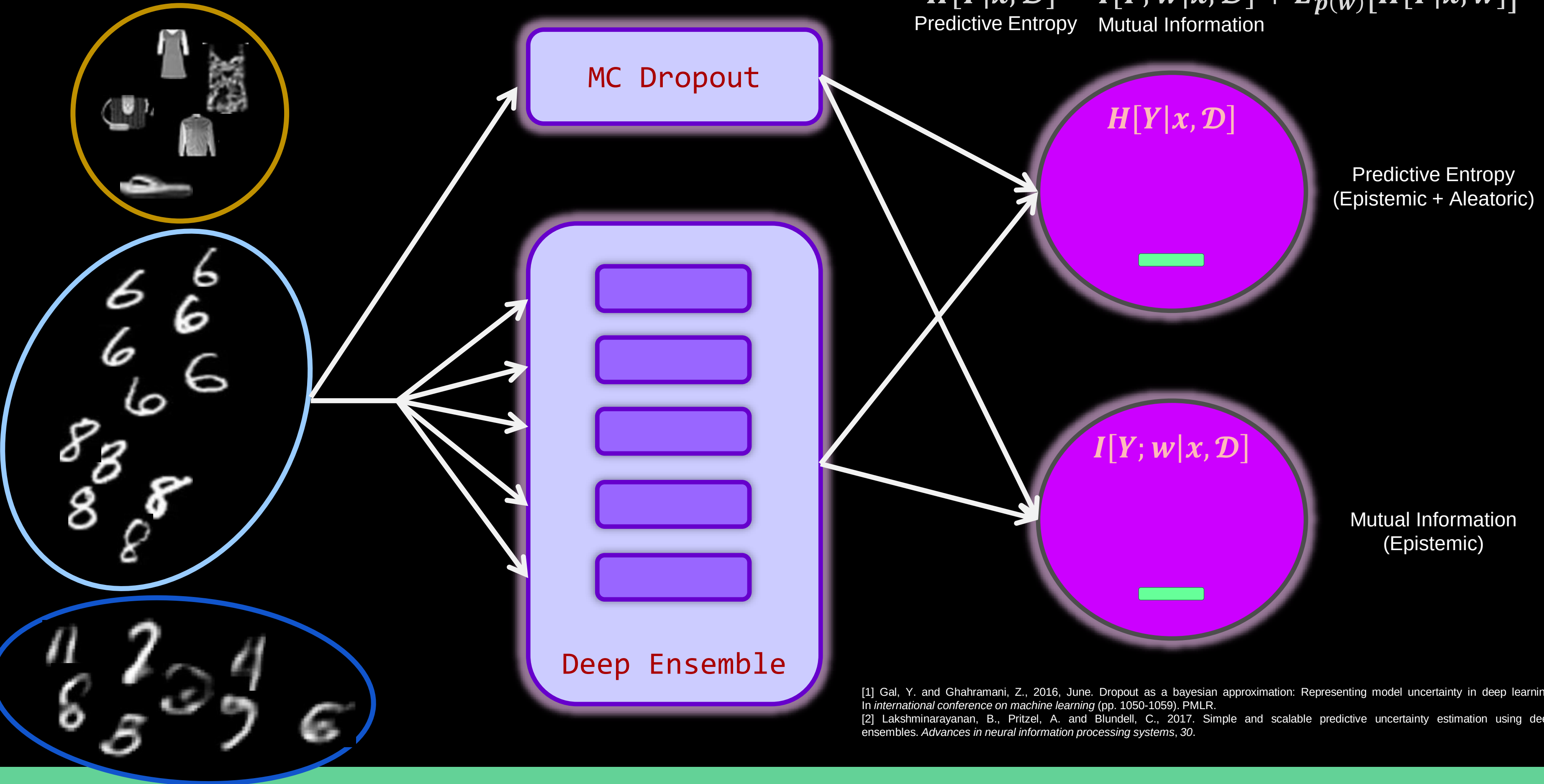


Current State-of-the-art



Current State-of-the-art

$$\underbrace{H[Y|x,\mathcal{D}]}_{\text{Predictive Entropy}} = \underbrace{I[Y;w|x,\mathcal{D}]}_{\text{Mutual Information}} + E_{p(w)}\big[H[Y|x,w]\big]$$



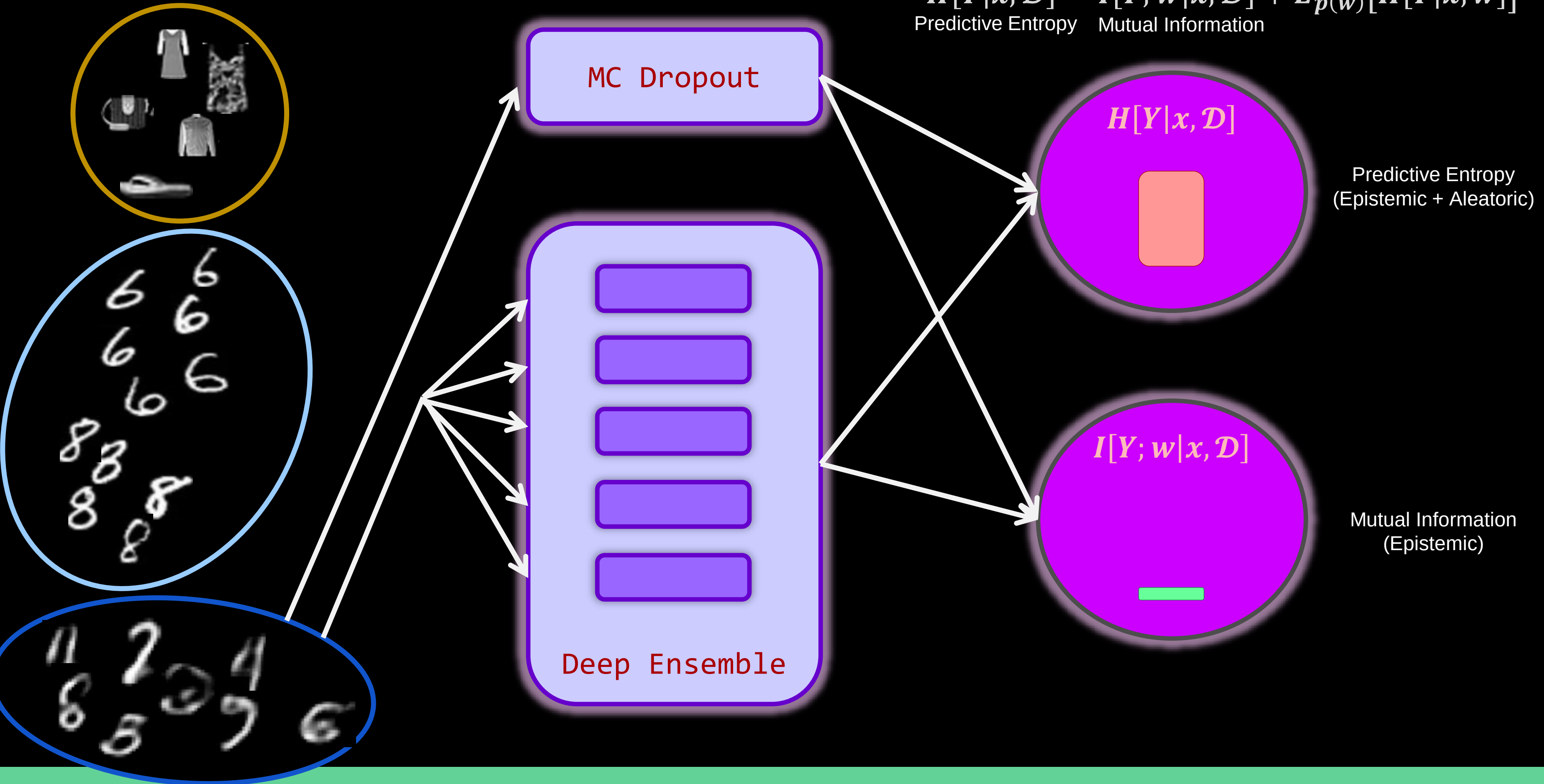
[1] Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

[2] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems, 30*.

Current State-of-the-art

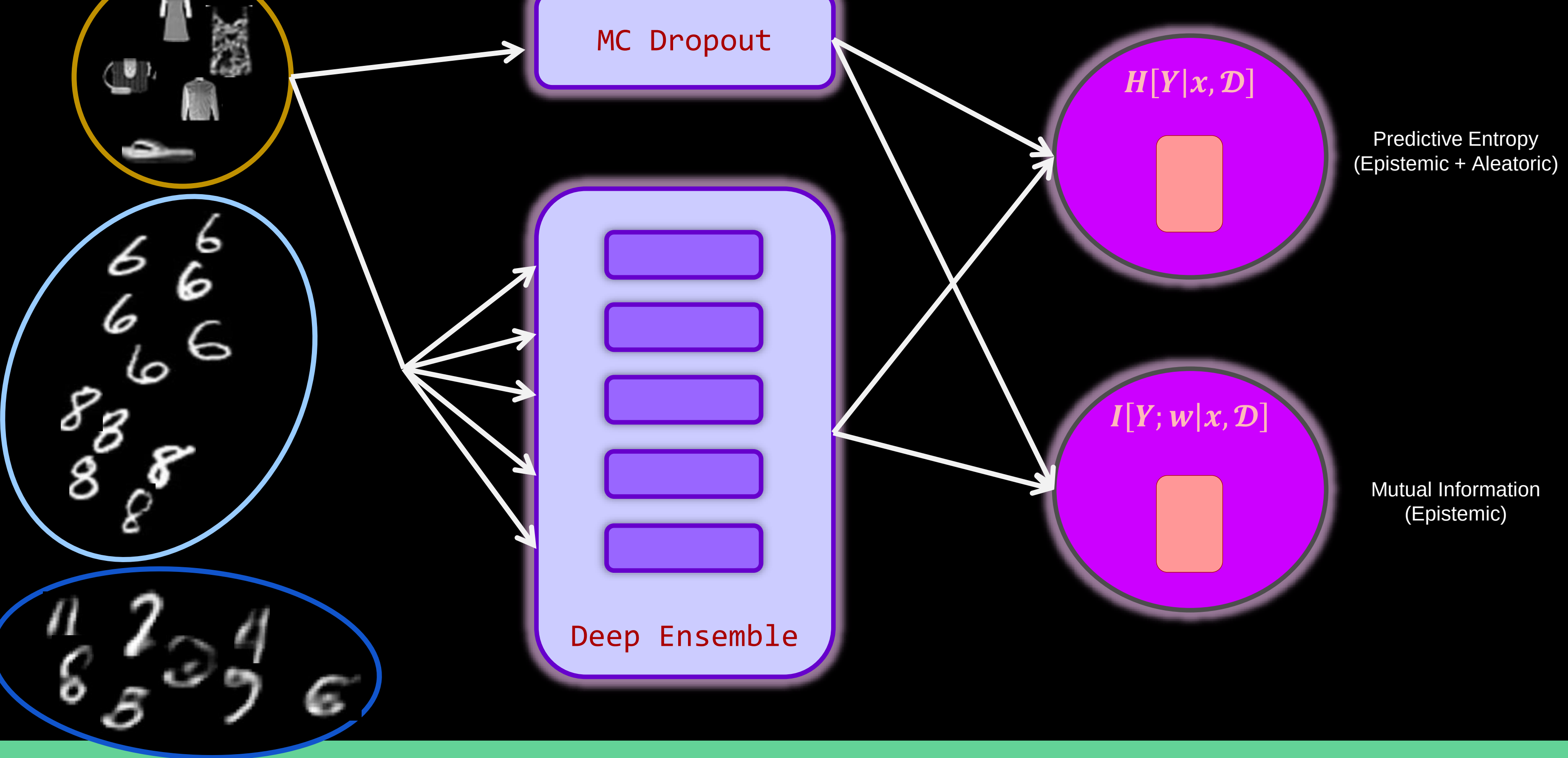
$$H[Y|x,\mathcal{D}] = I[Y;w|x,\mathcal{D}] + E_{p(w)}\left[H[Y|x,w]\right]$$

Predictive Entropy   Mutual Information



Current State-of-the-art

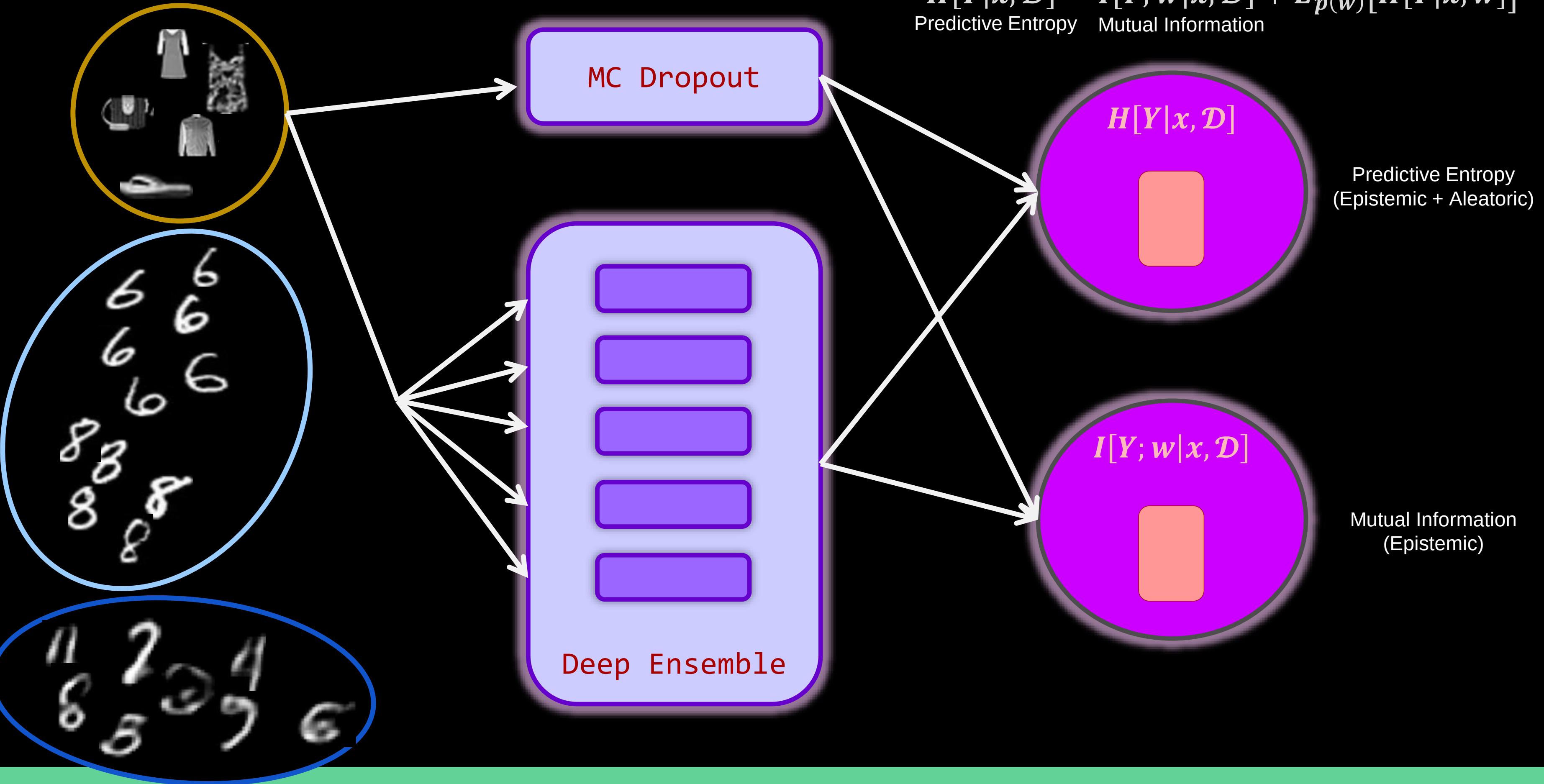
$$\underbrace{H[Y|x,\mathcal{D}]}_{\text{Predictive Entropy}} = \underbrace{I[Y;w|x,\mathcal{D}]}_{\text{Mutual Information}} + E_{p(w)}\big[H[Y|x,w]\big]$$



Current State-of-the-art

$$H[Y|x,\mathcal{D}] = I[Y;w|x,\mathcal{D}] + E_{p(w)}\left[H[Y|x,w]\right]$$

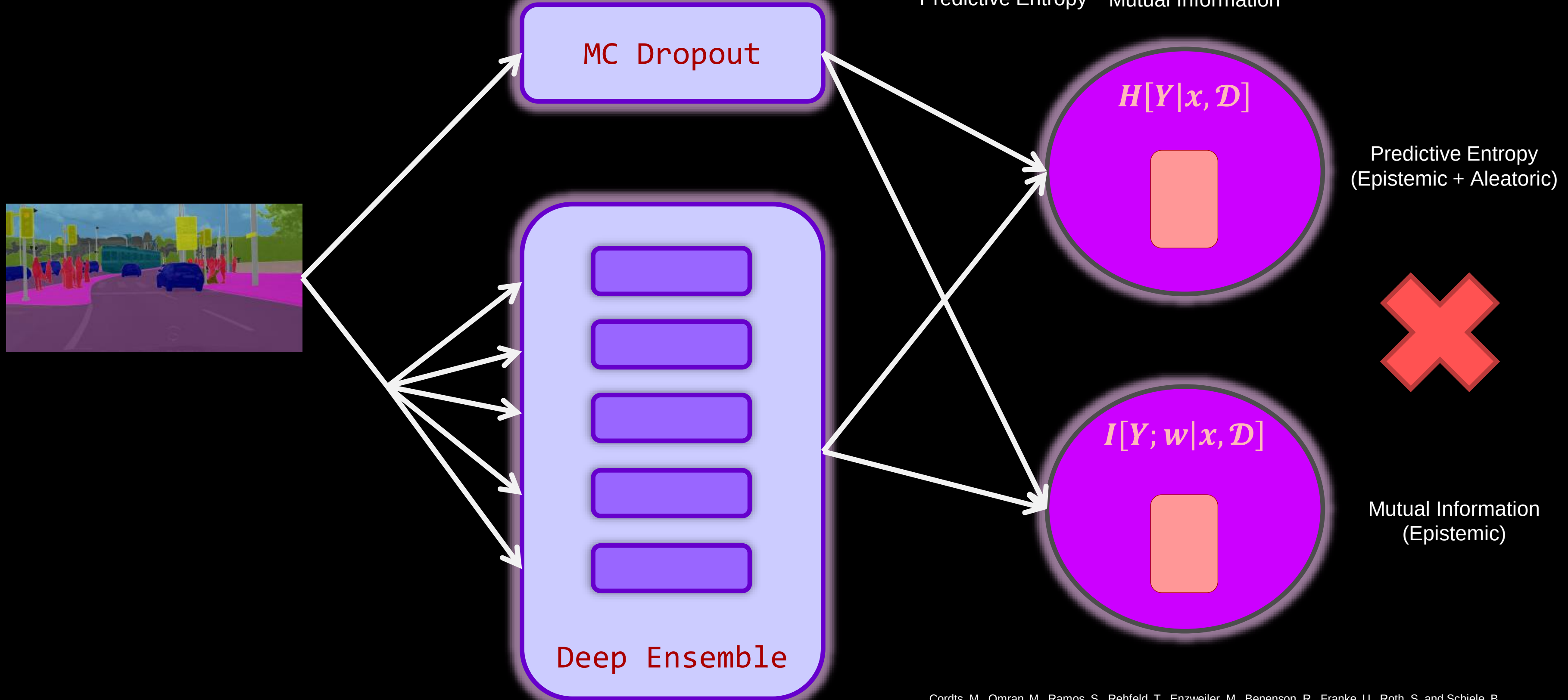
Predictive Entropy Mutual Information



Current State-of-the-art

$$H[Y|x,\mathcal{D}] = I[Y;w|x,\mathcal{D}] + E_{p(w)}\big[H[Y|x,w]\big]$$

Predictive Entropy   Mutual Information

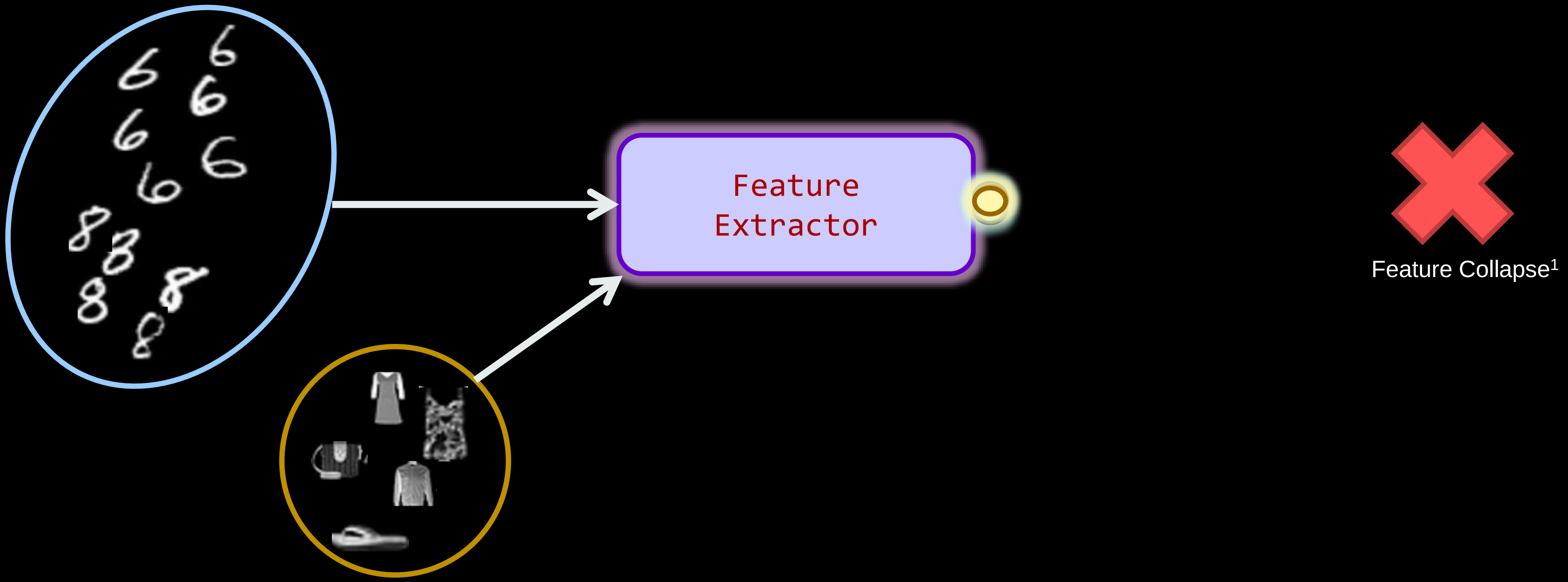


Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

Single Forward Pass Uncertainty

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor



Van Amersfoort, J., Smith, L., Teh, Y.W. and Gal, Y., 2020, November. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

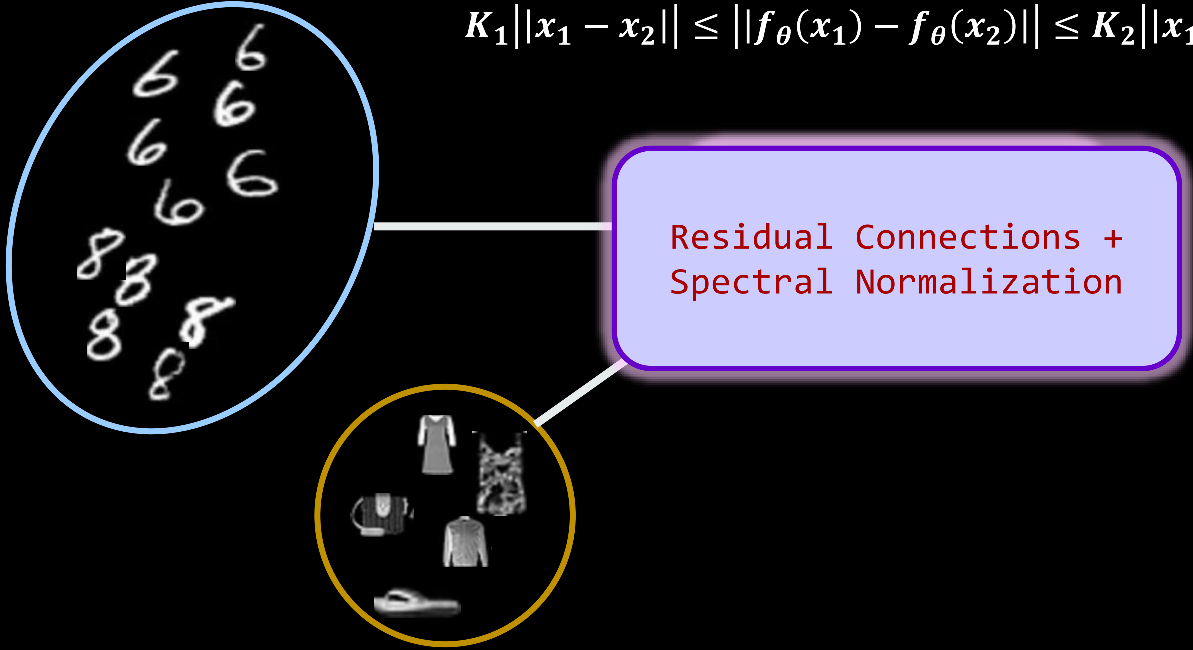
Single Forward Pass Uncertainty

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor

Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$



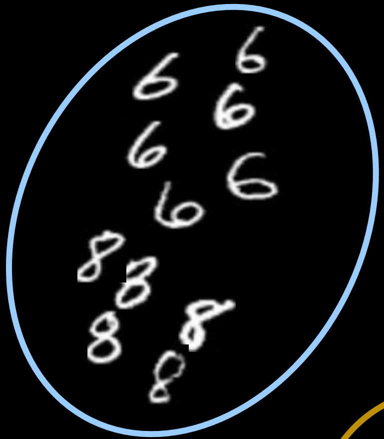
Single Forward Pass Uncertainty

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

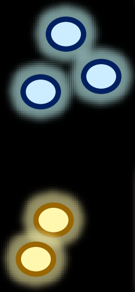
Requirement 1: A sensitive & smooth feature extractor

Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \le ||f_\theta(x_1) - f_\theta(x_2)|| \le K_2||x_1 - x_2||$$



Sensitive + Smooth Feature Extractor



Radial Basis Function (RBF):DUQ

Gaussian Process(GP): SNGP

However, DUQ and SNGP require changes in the training setup, DUQ cannot scale to large number of classes, SNGP has a number of hyper-parameters to fine-tune and neither of them explicitly model epistemic and aleatoric uncertainty.

Van Amersfoort, J., Smith, L., Teh, Y.W. and Gal, Y., 2020, November. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T. and Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, *33*, pp.7498-7512.

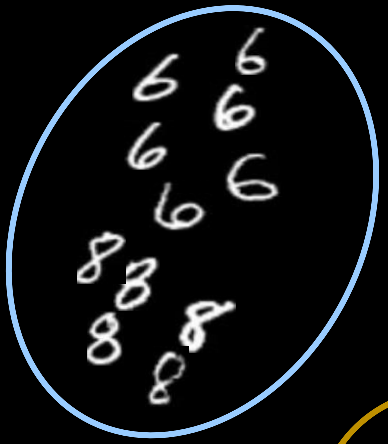
Single Forward Pass Uncertainty

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

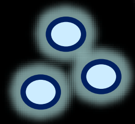
Requirement 1: A sensitive & smooth feature extractor

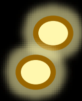
Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$



Sensitive + Smooth Feature Extractor





Radial Basis Function (RBF):DUQ

Gaussian Process(GP): SNGP

However, DUQ and SNGP require changes in the training setup, DUQ cannot scale to large number of classes, SNGP has a number of hyper-parameters to fine-tune and neither of them explicitly model epistemic and aleatoric uncertainty.

Van Amersfoort, J., Smith, L., Teh, Y.W. and Gal, Y., 2020, November. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T. and Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, *33*, pp.7498-7512.

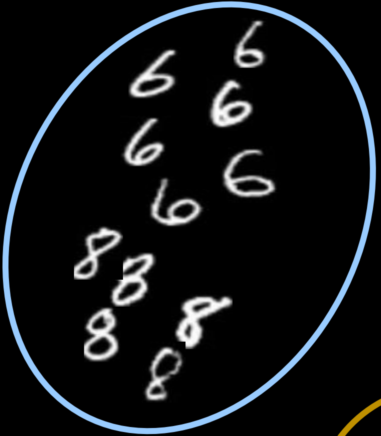
Deep Deterministic Uncertainty (DDU)

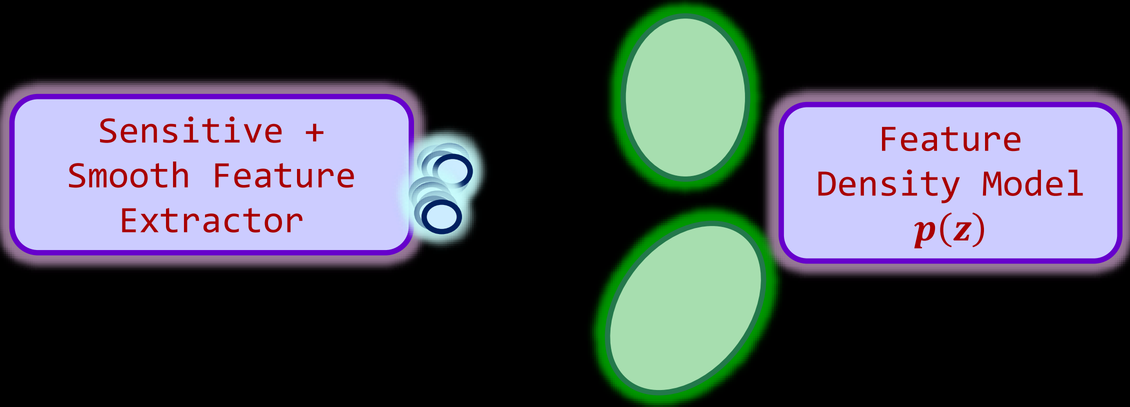
Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor

Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$





Requirement 2: A density model in the feature space

$$p(z) = \sum_c p(z|y = c)p(y = c)$$

A simple GDA with $p(z|y = c)$ modelled using a single mean and covariance matrix is good enough.

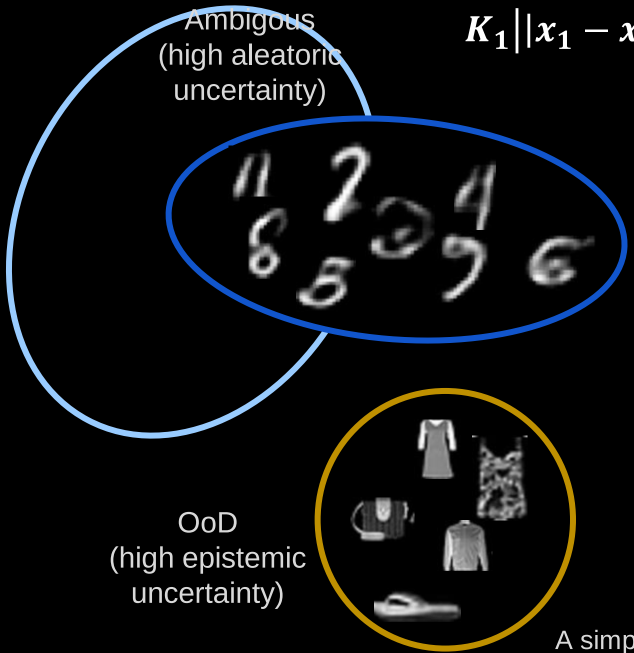
Deep Deterministic Uncertainty (DDU)

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor

Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$



Sensitive +
Smooth Feature
Extractor

Feature
Density Model
$p(z)$

Requirement 2: A density model in the feature space

$$p(z) = \sum_c p(z|y=c)p(y=c)$$

A simple GDA with $p(z|y=c)$ modelled using a single mean and covariance matrix is good enough.

Deep Deterministic Uncertainty (DDU)

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor

Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$

Ambigous
(high aleatoric
uncertainty)

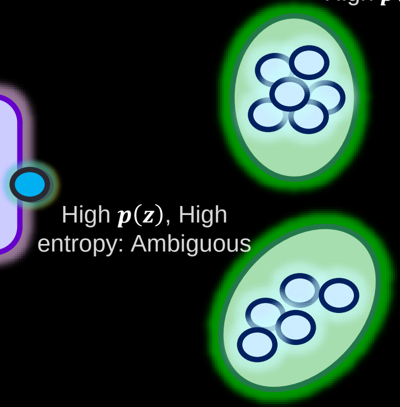
OoD
(high epistemic
uncertainty)

Sensitive +
Smooth Feature
Extractor

High $p(z)$, low entropy: iD

Low $p(z)$: OoD

High $p(z)$, High
entropy: Ambiguous



Requirement 2: A density model in the feature space

$$p(z) = \sum_c p(z|y=c)p(y=c)$$

A simple GDA with $p(z|y=c)$ modelled using a single mean and covariance matrix is good enough.

Deep Deterministic Uncertainty (DDU)

Motivation: Have a deterministic single forward pass model which can quantify uncertainty.

Requirement 1: A sensitive & smooth feature extractor

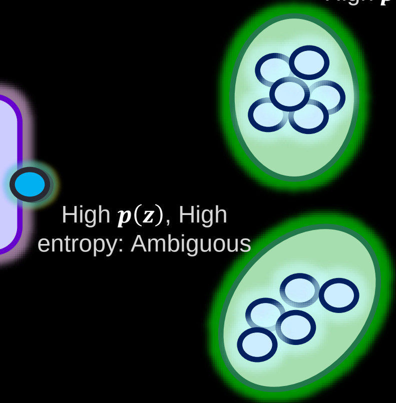
Bi-Lipschitz constrained feature space

$$K_1||x_1 - x_2|| \leq ||f_\theta(x_1) - f_\theta(x_2)|| \leq K_2||x_1 - x_2||$$

Ambigous
(high aleatoric
uncertainty)

Sensitive +
Smooth Feature
Extractor

High $p(z)$, low entropy: iD



Low $p(z)$: OoD

High $p(z)$, High
entropy: Ambiguous

Requirement 2: A density model in the feature space

$$p(z) = \sum_c p(z|y=c)p(y=c)$$

OoD
(high epistemic
uncertainty)

A simple GDA with $p(z|y=c)$ modelled using a single mean and covariance matrix is good enough.

DDU on OoD Detection

Table 1. *OoD detection performance of different baselines using a Wide-ResNet-28-10 architecture with the CIFAR-10 vs SVHN/CIFAR-100/Tiny-ImageNet and CIFAR-100 vs SVHN/Tiny-ImageNet dataset pairs averaged over 25 runs. SN:* Spectral Normalisation, *JP:* Jacobian Penalty. The best deterministic single-forward pass method and the best method overall are in bold for each metric.

Train Dataset	Method	Penalty	Aleatoric Uncertainty	Epistemic Uncertainty	Accuracy (↑)	ECE (↓)	AUROC SVHN (↑)	AUROC CIFAR-100 (↑)	AUROC Tiny-ImageNet (↑)
CIFAR-10	Softmax	-	Softmax Entropy	Softmax Entropy	95.98 ± 0.02	**0.85 ± 0.02**	94.44 ± 0.43	89.39 ± 0.06	88.42 ± 0.05
	Energy-based [46]	-		Softmax Density			94.56 ± 0.51	88.89 ± 0.07	88.11 ± 0.06
	DUQ [65]	JP	Kernel Distance	Kernel Distance	94.6 ± 0.16	1.55 ± 0.08	93.71 ± 0.61	85.92 ± 0.35	86.83 ± 0.12
	SNGP [45]	SN	Predictive Entropy	Predictive Entropy	**96.04 ± 0.09**	1.8 ± 0.1	94.0 ± 1.3	91.13 ± 0.15	89.97 ± 0.19
	DDU (ours)	**SN**	**Softmax Entropy**	**GDA Density**	95.97 ± 0.03	**0.85 ± 0.04**	**97.86 ± 0.19**	**91.34 ± 0.04**	**91.07 ± 0.05**
	5-Ensemble [40]	-	Predictive Entropy	Predictive Entropy	**96.59 ± 0.02**	**0.76 ± 0.03**	97.73 ± 0.31	**92.13 ± 0.02**	90.06 ± 0.03
				Mutual Information			97.18 ± 0.19	91.33 ± 0.03	90.90 ± 0.03
					Accuracy (↑)	ECE (↓)	AUROC SVHN (↑)		AUROC Tiny-ImageNet (↑)
CIFAR-100	Softmax	-	Softmax Entropy	Softmax Entropy	80.26 ± 0.06	4.62 ± 0.06	77.42 ± 0.57		81.53 ± 0.05
	Energy-based [46]	-		Softmax Density			78 ± 0.63		81.33 ± 0.06
	SNGP [45]	SN	Predictive Entropy	Predictive Entropy	80.00 ± 0.11	4.33 ± 0.01	85.71 ± 0.81		78.85 ± 0.43
	DDU (ours)	**SN**	**Softmax Entropy**	**GMM Density**	**80.98 ± 0.06**	**4.10 ± 0.08**	**87.53 ± 0.62**		**83.13 ± 0.06**
	5-Ensemble [40]	-	Predictive Entropy	Predictive Entropy	**82.79 ± 0.10**	**3.32 ± 0.09**	79.54 ± 0.91		82.95 ± 0.09
				Mutual Information			77.00 ± 1.54		82.82 ± 0.04

Table 2. *OoD detection performance of different baselines using ResNet-50, Wide-ResNet-50-2 and VGG-16 architectures on ImageNet vs ImageNet-O [26].* Best AUROC scores are marked in bold.

Model	Accuracy (↑)		ECE (↓)		AUROC (↑)				
	Deterministic	3-Ensemble	Deterministic	3-Ensemble	**Softmax Entropy**	**Energy-based Model**	**DDU**	**3-Ensemble PE**	**3-Ensemble MI**
ResNet-50	74.8 ± 0.05	76.01	2.08 ± 0.11	2.07	51.42 ± 0.61	55.76 ± 0.81	**71.29 ± 0.08**	60.3	62.43
Wide-ResNet-50-2	76.75 ± 0.11	77.58	1.18 ± 0.07	1.22	52.71 ± 0.23	57.13 ± 0.4	**73.12 ± 0.19**	60.45	64.81
VGG-16	72.48 ± 0.02	73.54	2.62 ± 0.11	2.59	50.67 ± 0.22	52.04 ± 0.23	54.32 ± 0.14	58.74	**60.56**

Over experiments on multiple OoD detection benchmarks, we find that DDU consistently performs at par with deep ensembles and outperforms DUQ and SNGP.

DDU on Active Learning

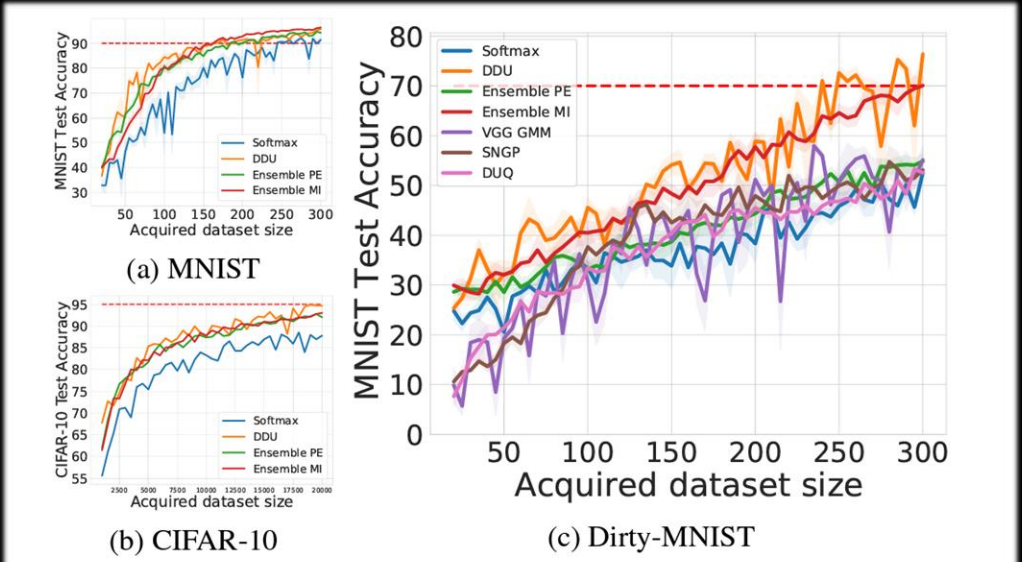
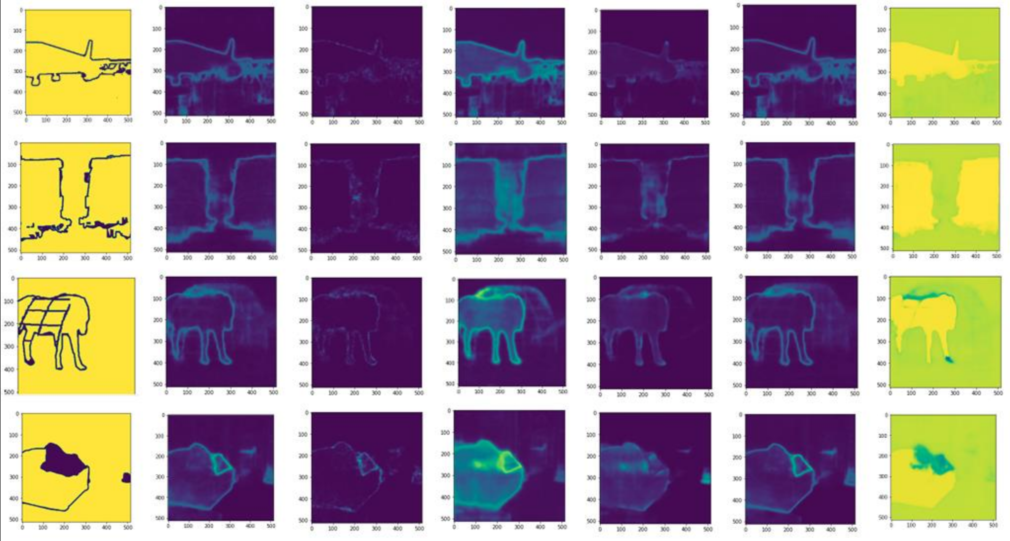


Figure 4. *Active Learning experiments.* Acquired training set size vs test accuracy. DDU performs on par with Deep Ensembles.

DDU's performance improvement is particularly noticeable when there are ambiguous samples in the training set, i.e., when training on Dirty-MNIST instead of MNIST.

DDU on Semantic Segmentation



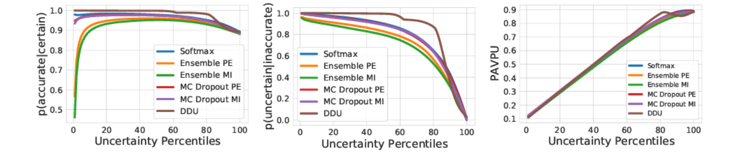


Figure 7. $p(\text{accurate}|\text{certain})$, $p(\text{uncertain}|\text{inaccurate})$ and PAVPU evaluated on PASCAL VOC validation set. DDU outperforms all other baselines.

Table 3. *Pascal VOC val set mIoU and runtime in milliseconds averaged over 10 forward passes.* For MC Dropout, we perform 5 stochastic forward passes.

Baseline	Softmax	MC Dropout	Deep Ensemble	DDU
mIoU	78.53	78.61	78.47	78.53
Runtime (ms)	275.48 ± 1.91	1576.75 ± 1.56	875.87 ± 0.79	$\mathbf{263.83 \pm 2.79}$

DDU’s density particularly captures epistemic uncertainty as is evident from the qualitative samples.
The entropy on the other hand captures aleatoric.
At the same time, DDU also provides the desirable run-time speed benefit over ensembles and MC Dropout.

Thank you!

References

[1] Gal, Y. and Ghahramani, Z., 2016, June. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059). PMLR.

[2] Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

[4] Van Amersfoort, J., Smith, L., Teh, Y.W. and Gal, Y., 2020, November. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

[5] Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T. and Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, *33*, pp.7498-7512.