Packed-Ensembles for Efficient Uncertainty Estimation

Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi

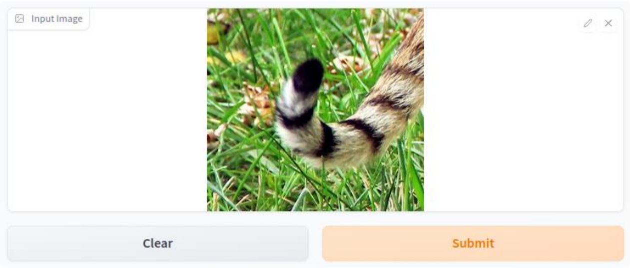
ICLR

ENSTA IP PARIS
cea
valeo.ai
TELECOM Paris IP PARIS

Uncertainty Quantification

From HuggingFace Hub

microsoft/resnet-50



Virtual Pooling

Sam Kieschnick

???

Lindasj22

On overconfidence: Guo et al. 2017

Deep Ensembles
(Lakshminarayanan et al. 2017)

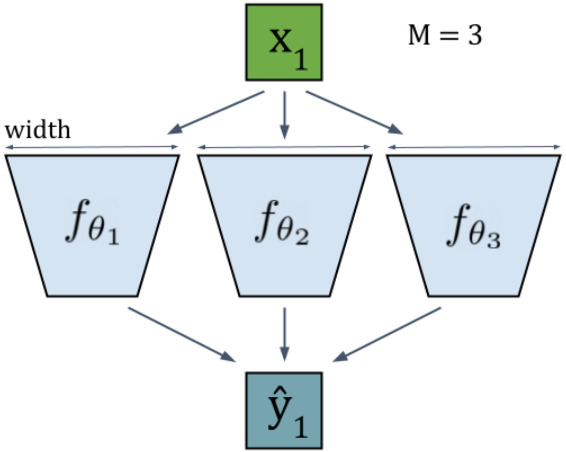


Figure 1 – Deep Ensembles

Cons:

- Number of operations
- Memory storage
- Inference time

Implicit Ensembles
w/ subnetworks

- BatchEnsemble (Wen et al. 2019)
  - Subnetwork-specific parameters
- MIMO (Havasi et al. 2021)
  - Several subnetworks within one network.
- Masksembles (Durasov, et al. 2021)
  - Random masks to disable a subset of parameters at each forward pass.

Cons:

- Multiple forward passes
- Subnetworks not independent

➡ Lottery ticket hyp. (Frankle et al. 2018)

➡ Ensembles of small networks can be as good as medium networks (Lobacheva et al. 2021)

Packed-Ensembles: Seamlessly training ensembles!

Simple & efficient *generalization* of Deep Ensembles

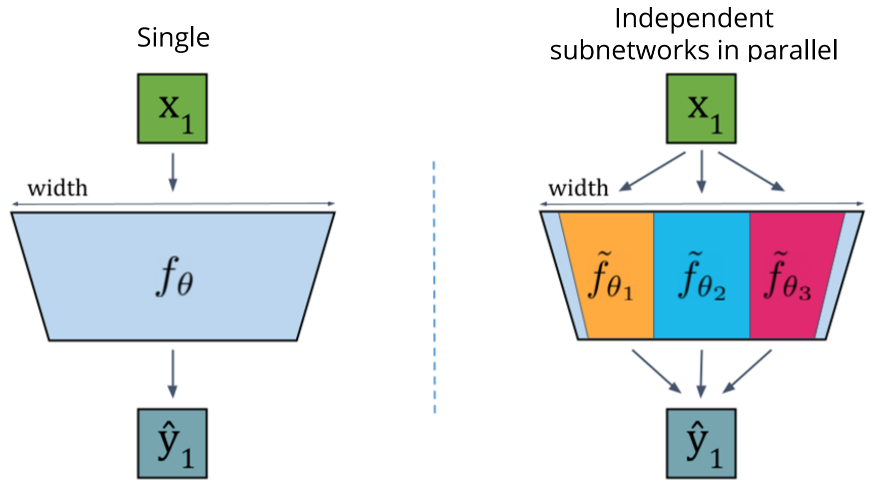


Figure 2 – From a single network to Packed-Ensembles

How can we build a backbone containing independent subnetworks run in parallel?

Grouped convolutions (Krizhevsky et al. 2012)

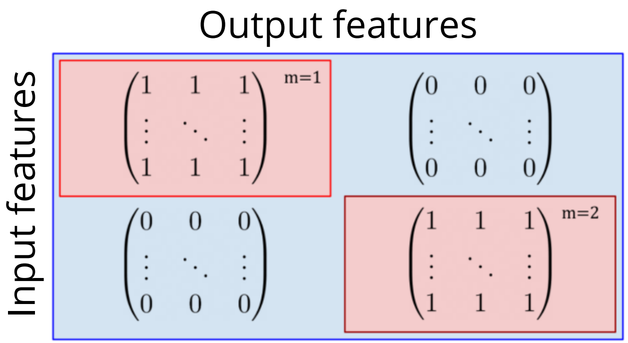
The output channel c is produced by a specific group, identified by the integer $g = \left\lfloor \frac{\text{groups} \times c}{C_{out}} \right\rfloor$, which only uses $\frac{1}{\text{groups}}$ of the input channels:

$$\text{out}(c,:,:) = \sum_{k=0}^{\frac{C_{in}}{\text{groups}}-1} \text{weight}_g\,(c,k,:,:) \star \text{input}\left(k + g \times \frac{C_{in}}{\text{groups}},:,:\right)$$

C_{in} : number of input channels.

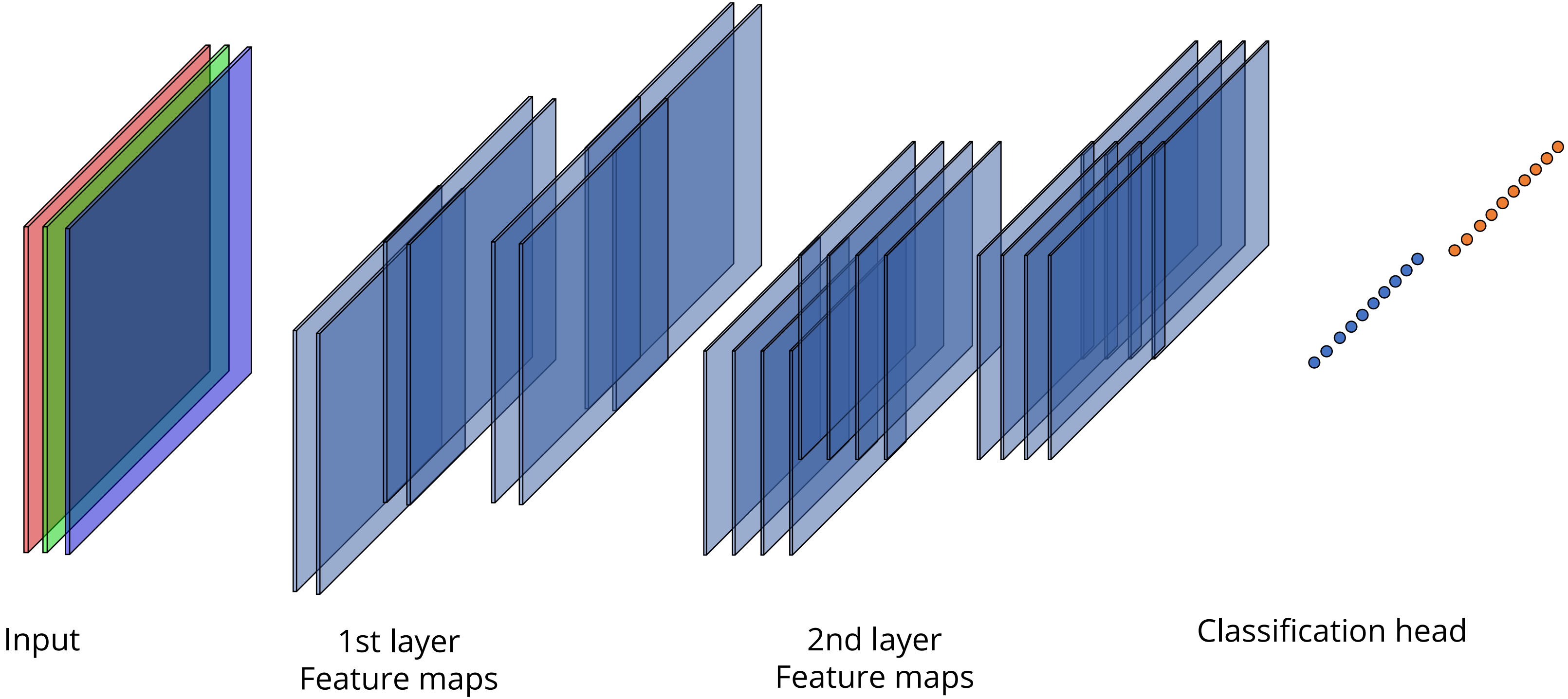
C_{out}: number of output channels.

groups: number of groups.



Grouped Linear mask
$(\text{groups}=2)$

Seamlessly training ensembles



Modulating model capacity

Number of parameters of a convolution

$$\underbrace{C_{out}}^{\text{Output channels}} \times \underbrace{C_{in}}_{\text{Input channels}} \times \overbrace{k_{height} \times k_{width}}^{\text{Kernel size}}$$

$$\frac{(\alpha \times C_{out} \times C_{in} \times k_{height} \times k_{width}}{M}$$

Grouped convolution / Packed-Ensembles layer with α

Two specific cases:

$\alpha = \sqrt{M}$ ➡ Single Network

$\alpha = M$ ➡ Deep Ensembles w/ M subnetworks

M: number of subnetworks

α : width factor

γ : number of subgroups

Packed-Ensembles

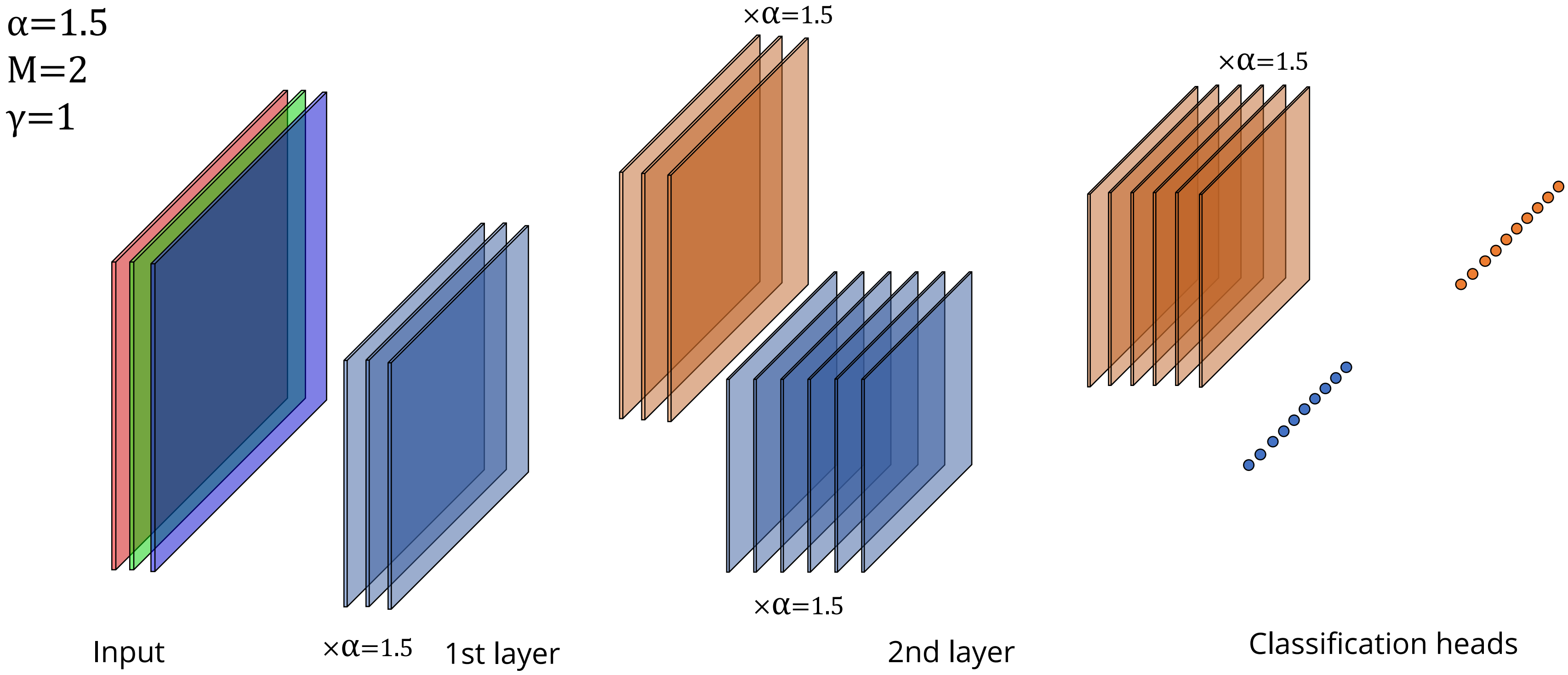


Figure 3 – Animation of Packed-Ensembles

Experiments - CIFAR

First
Second

	Method	Acc ↑	NLL ↓	ECE ↓	AUPR ↑	AUC ↑	FPR95 ↓	Params (10^6) ↓	Mult-Adds ↓
		Classification		Calibration	OOD Detection			Complexity	
CIFAR-10	Single Model	95.1	0.211	0.031	95.2	91.9	23.6	23.52	1.30
	BatchEnsemble	93.9	0.255	0.033	94.7	91.3	20.1	23.63	5.19
	MIMO ($\rho = 1$)	95.4	0.197	0.030	95.1	90.8	26.0	23.59	1.30
	Masksembles	95.3	0.175	0.019	95.7	92.2	22.1	23.81	5.19
	Packed-Ensembles	95.9	0.137	**0.008**	**97.3**	**95.2**	**14.4**	**14.55**	**1.00**
	Deep Ensembles	**96.0**	**0.136**	**0.008**	97.0	94.7	15.5	94.08	5.19
CIFAR-100	Single Model	78.3	0.905	0.089	87.4	77.9	57.6	23.70	1.30
	BatchEnsemble	66.6	1.788	0.182	85.2	74.6	60.6	23.81	5.19
	MIMO ($\rho = 1$)	79.0	0.876	0.079	87.5	76.9	64.7	24.33	1.30
	Masksembles	78.5	0.832	0.046	**90.3**	**81.9**	**52.3**	23.81	5.19
	Packed-Ensembles	**81.2**	**0.703**	**0.020**	90.0	81.7	56.5	**15.55**	**1.00**
	Deep Ensembles	80.9	0.713	0.026	89.2	80.8	52.5	94.82	5.19

Table 1 - Performance of various ensembles methods on CIFAR - M=4, $\alpha=\gamma=2$ – ResNet-50

Experiments - ImageNet

First
Second

Method	Net	Acc	ECE	Texture Dataset			ImageNet - O			ImageNet - R		
				AUPR - T	**AUC - T**	**FPR95 - T**	**AUPR - IO**	**AUC - IO**	**FPR95 - IO**	**rAcc**	**rNLL**	**rECE**
Single Model	R50	77.8	0.121	18.0	80.9	68.6	3.6	50.8	90.8	23.5	5.187	0.082
BatchEnsemble	R50	75.9	0.035	20.2	81.6	66.5	4.0	55.2	82.3	21.0	6.148	0.165
MIMO ($\rho = 1$)	R50	77.6	0.147	18.4	81.6	66.8	3.7	52.2	90.6	23.4	5.115	0.059
Masksembles	R50	73.6	0.209	13.6	79.7	68.3	3.3	47.7	87.7	21.2	5.139	0.011
Packed-Ensembles $\alpha = 3$	R50	77.9	0.180	**35.1**	**88.2**	**43.7**	**9.9**	**68.4**	**80.9**	23.8	4.978	0.022
Deep Ensembles	R50	**79.2**	0.233	19.6	83.4	62.1	3.7	52.5	85.5	**24.9**	**4.879**	**0.018**
Single Model	R50×4	80.2	0.022	20.5	82.6	63.9	4.9	60.2	87.4	26.0	5.190	0.1721
BatchEnsemble	R50×4	77.7	0.024	23.8	82.8	63.8	4.4	58.4	80.5	23.4	6.079	0.203
MIMO ($\rho = 1$)	R50×4	80.3	0.015	19.3	82.5	66.1	4.9	60.7	86.4	25.8	5.278	0.189
Masksembles	R50×4	79.8	0.137	21.5	83.3	63.5	4.4	58.4	80.5	23.4	6.079	0.207
Packed-Ensembles $\alpha = 2$	R50×4	81.3	0.103	**34.6**	**88.1**	**50.3**	**9.6**	**69.9**	**79.2**	26.6	4.848	**0.075**
Deep Ensembles	R50×4	**82.1**	0.053	23.0	85.6	58.1	5.0	62.7	81.9	**28.2**	**4.789**	0.105

Table 2 - Performance of various ensembles methods on ImageNet - $M=4$, $\gamma=1$

ImageNet-O: Hendrycks et al. 2021a
ImageNet-R : Hendrycks et al. 2021b
Texture: Wang et al. 2022

Diversity

Diversity in Ensembles is essential (Fort et al.). But where does it come from?

Stochasticity			ResNet-50				
ND	DI	DB	Acc (↑)	ECE (↓)	AUPR (↑)	IDMI	OODMI
-	-	-	77.63±0.23	0.0825±0.0018	89.19±0.65	0±0	0±0
✓	-	-	80.94±0.10	0.0179±0.0010	90.23±0.62	0.1513	0.4022
-	✓	-	81.01±0.06	0.0202±0.0011	91.10±0.39	0.1524	0.4088
-	-	✓	80.87±0.10	0.0178±0.0010	90.80±0.30	0.1505	0.4115
✓	✓	-	81.16±0.10	0.0210±0.0008	91.69±0.56	0.1584	0.4135
✓	✓	✓	81.08±0.08	0.0198±0.0013	90.68±0.25	0.1534	0.4031

MI: Mutual Information

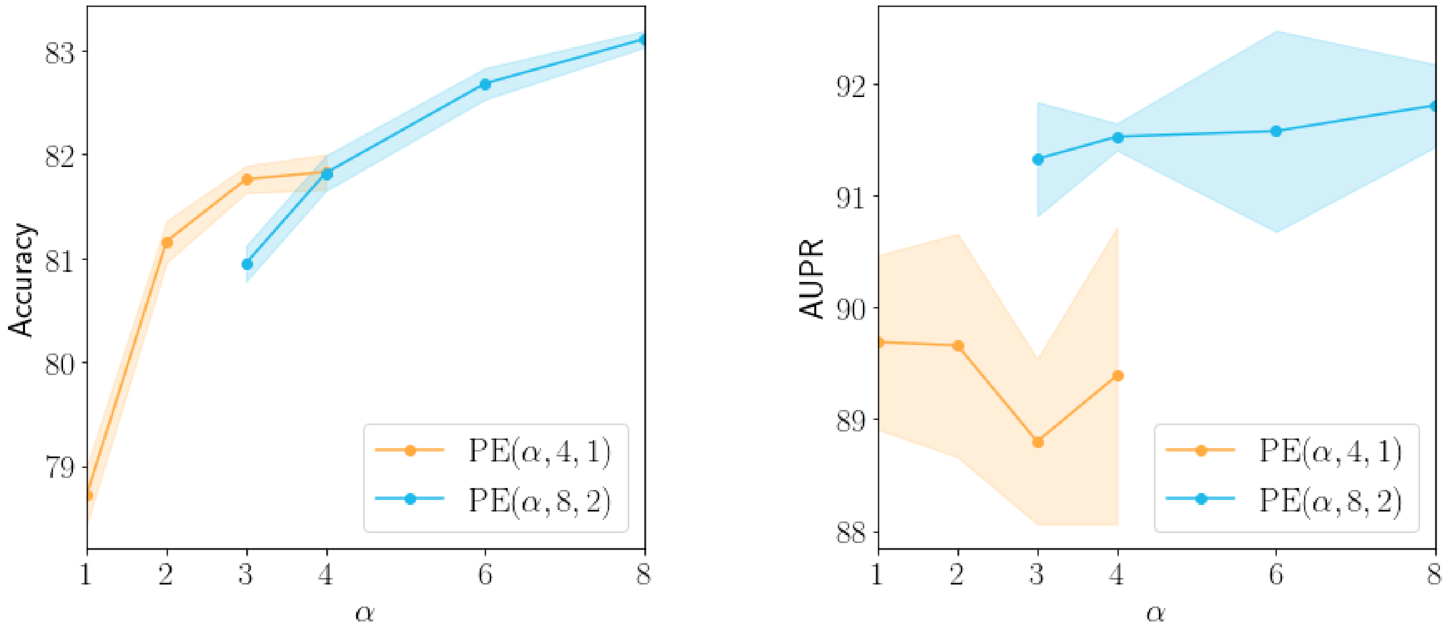
Stochasticity in:
Packed-Ensembles
Deep Ensembles

Table 3 - Performance with respect to stochasticity sources CIFAR-100

ND: **N**on-**D**eterministic backward propagation
DI: **D**ifferent **I**nitialization
DB: **D**ifferent **B**atch composition & order

➡ Obvious sources of diversity seem equivalent

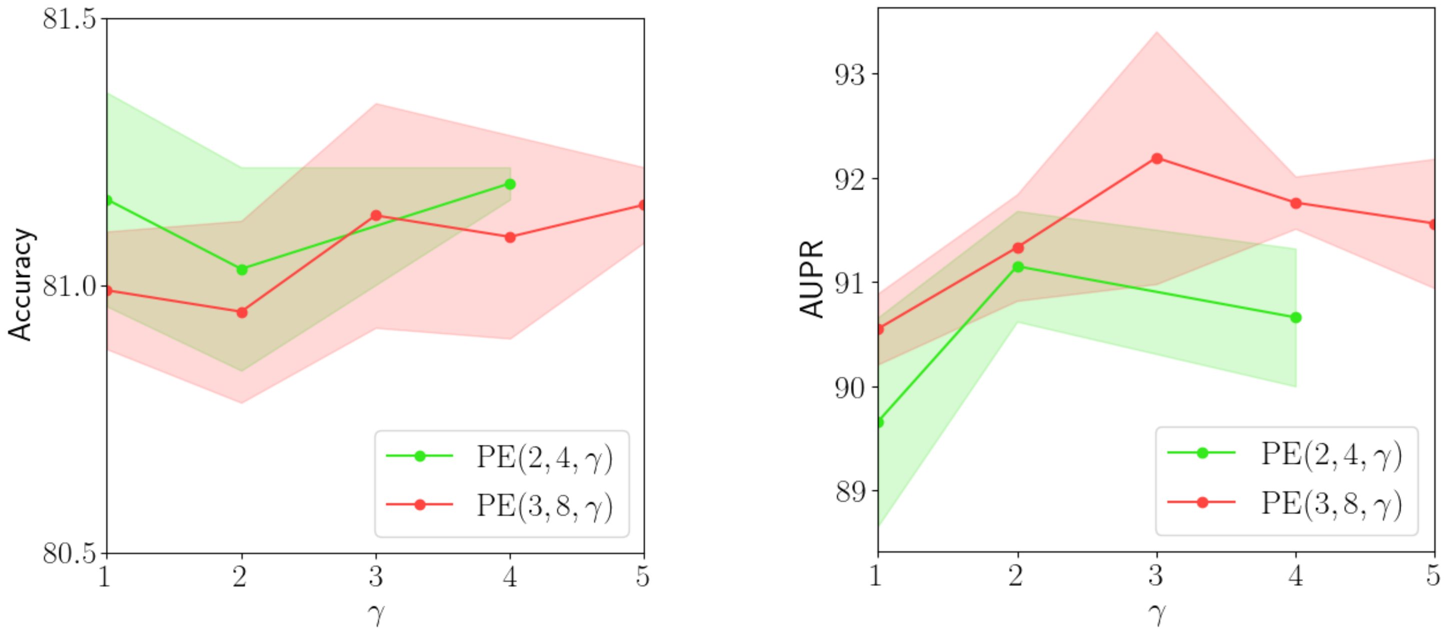
Sensitivity Analysis on α – capacity modulator



PE(α,M,γ)
M=4, γ=1
M=8, γ=2

Figure 4 - Sensitivity Analysis on α – ResNet-50, CIFAR-100

Sensitivity Analysis on γ – sparsity modulator



PE(α,M,γ)
α=2, M=4
α =3, M=8

Figure 5 - Sensitivity Analysis on γ – ResNet-50, CIFAR-100

Conclusion & Takeaways

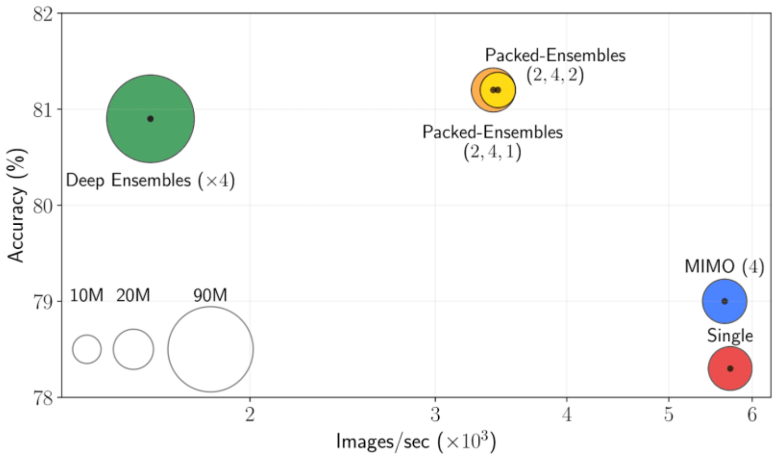
Main takeaway:

Packed-Ensembles: **controlled trade-off** between accuracy & uncertainty vs. model complexity

We propose TorchUncertainty, a new PyTorch library which includes code for PE

Other takeaways:

- Ensembles of small independent neural networks can be as effective as ensembles of large DNNs
- Sources of diversity seem equivalent



TorchUncertainty

torch-uncertainty.github.io

References & QR Codes

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

[2] Wen, et al. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In ICLR, 2019.

[3] Martin Havasi et al. Training independent subnetworks for robust prediction. In ICLR, 2021.

[4] Durasov, et al. Masksembles for uncertainty estimation. In CVPR, 2021.

[5] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In NeurIPS, 2012.

[6] Wang et al. ViM: Out-of-distribution with virtual-logit matching. In CVPR, 2022.

[7] Hendrycks et al. Natural adversarial examples. In CVPR, 2021a.

[8] Hendrycks et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In CVPR, 2021b.

Paper

Poster

TorchUncertainty