Supplementary Material : Knowledge Diversification in Ensembles of Identical Neural Networks

BMVC 2022 Submission # 0798

1 Additional Results

Table 1: Results of experiments on a single layer ConvNet with M filters trained on MNIST. The middle column indicates accuracy results obtained with a single network of M filters. The right column indicates the accuracy results obtained with two networks with M/2 filters each, trained using feature difference loss across the two networks. Results indicate that two networks trained with FDL learns better representations, achieves higher accuracy and thus makes better use of the model capacity.

Filters	1x Network (%)	2x Half-Networks (%)
(M)	(M filters)	(M/2 filters each) + FDL
1	87.87	-
2	91.96	**92.7**
4	**94.99**	94.55
8	97.06	**97.52**
16	98	**98.04**
32	98.23	**98.37**
64	98.27	**98.49**
128	98.31	**98.55**
256	98.32	**98.54**

Table 2: FDL Ensemble of multiple neural networks for the CIFAR-10 dataset. We report accuracy metrics of each base model along with the ensemble. Best individual and ensemble accuracies are marked in bold.

Model	CIFAR-10				FDL
	Network 1	Network 2	Network 3	Network 4	Ensemble
VGG-16 (x1)	93.66				93.66
VGG-16 (x2)	93.57	93.96			94.93
VGG-16 (x3)	93.53	93.93	93.60		95.14
VGG-16 (x4)	93.82	93.79	**93.97**	93.58	**95.22**
ResNet-20 (x1)	92.20				92.20
ResNet-20 (x2)	91.79	92.02			93.56
ResNet-20 (x3)	92.36	92.51	92.28		94.29
ResNet-20 (x4)	92.18	91.98	92.10	**92.70**	**94.44**
ResNet-32 (x1)	93.21				93.21
ResNet-32 (x2)	92.37	92.10			93.81
ResNet-32 (x3)	93.33	**93.48**	93.11		94.78
ResNet-32 (x4)	92.69	92.91	93.03	93.02	**94.88**

Table 3: FDL Ensemble of multiple neural networks for the CIFAR-100 dataset. We report accuracy metrics of each base model along with the ensemble. Best individual and ensemble accuracies are marked in bold.

Model	CIFAR-100				FDL
	Network 1	Network 2	Network 3	Network 4	Ensemble
VGG-16 (x1)	74.61				74.61
VGG-16 (x2)	74.27	74.44			77.02
VGG-16 (x3)	74.03	74.42	73.72		77.66
VGG-16 (x4)	**74.82**	73.88	73.73	74.26	**78.34**
ResNet-20 (x1)	67.82				67.82
ResNet-20 (x2)	67.63	68.26			71.48
ResNet-20 (x3)	67.57	67.57	**68.50**		73.43
ResNet-20 (x4)	67.11	67.76	67.48	67.71	**73.71**
ResNet-32 (x1)	69.42				69.42
ResNet-32 (x2)	69.25	68.80			73.46
ResNet-32 (x3)	69.40	**69.70**	68.77		75.38
ResNet-32 (x4)	69.20	69.60	69.45	69.10	**76.30**

Algorithm 1: Training routine for two identical base networks in an FDL ensemble.

Given: Identical networks N_1 and N_2, Ensemble Head network N_E
Data: **I**=Image, **y**=Label.

Pretrain N_1 for one epoch. `/*Phase 0*/`
Pretrain N_2 for one epoch.

while $iter \leq iter_{max}$ **do**

$\hat{\mathbf{y}}_1 \leftarrow N_1(I)$ `/*Phase 1*/`
$L_1 \leftarrow L_X(\hat{\mathbf{y}}_1, \mathbf{y})$
$BackpropagateAndUpdate_{N_1}(L_1)$
$\hat{\mathbf{y}}_2 \leftarrow N_2(I)$
$L_2 \leftarrow L_X(\hat{\mathbf{y}}_2, \mathbf{y})$
$BackpropagateAndUpdate_{N_2}(L_2)$

$L_1 \leftarrow L_X(N_1(I), \mathbf{y})$ `/*Phase 2*/`
$L_2 \leftarrow L_X(N_2(I), \mathbf{y})$
$S^{N_1,N_2} = (L_1 - L_2)^2$
$BackpropagateAndUpdate_{N_1,N_2}(S)$

$\hat{\mathbf{y}}_1 \leftarrow N_1(I)$ `/*Phase 3*/`
$\hat{\mathbf{y}}_2 \leftarrow N_2(I)$
Collect all feature tensors into buckets F_{N_1} and F_{N_2}.
Compute feature difference loss, L_{FDL} from Eq. **??**.
$BackpropagateAndUpdate_{N_1,N_2}(-L_{FDL})$

$\hat{\mathbf{y}}_c \leftarrow [N_1(I)^T, N_2(I)^T]^T$ `/*Phase 4*/`
$\hat{\mathbf{y}}_E \leftarrow N_E(\hat{\mathbf{y}}_c)$
$L_E \leftarrow L_X(\hat{\mathbf{y}}_E, \mathbf{y})$
$BackpropagateAndUpdate_{N_E}(L_E)$

$iter{+}{+}$

end

2 Hyperparameters

Table 4: Hyperparameters for all experiments carried out.

Model	Dataset	N		Network 1 Optimizer		Network 2 Optimizer		Network 3 Optimizer		Network 4 Optimizer		Similarity Optimizer		FDL Optimizer		Ensemble Head Optimizer	
			B	lr	m	lr	m	lr	m	lr	m	lr	m	lr	m	lr	m
VGG-16	CIFAR-10	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-5	0.85	1e-1	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-4	0.85	1e-1	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-4	0.85	1e-2	0.9
	CIFAR-100	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-5	0.85	1e-1	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-5	0.85	1e-1	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-4	0.85	1e-2	0.9
ResNet-20	CIFAR-10	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-4	0.85	1e-2	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-4	0.85	1e-2	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-4	0.85	1e-2	0.9
	CIFAR-100	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-5	0.85	1e-2	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-5	0.85	1e-2	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-5	0.85	1e-2	0.9
ResNet-32	CIFAR-10	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-4	0.85	1e-2	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-4	0.85	1e-2	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-4	0.85	1e-2	0.9
	CIFAR-100	x1	128	1e-2	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	128	1e-2	0.9	1e-2	0.9	-	-	-	-	1e-2	0.9	1e-5	0.85	1e-2	0.9
		x3	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	-	-	1e-2	0.9	1e-5	0.85	1e-2	0.9
		x4	128	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-2	0.9	1e-5	0.85	1e-2	0.9
ResNet-18	ImageNet	x1	256	1e-1	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	256	1e-1	0.9	1e-1	0.9	-	-	-	-	1e-5	0.9	2e-5	0.9	1e-4	0.9
ResNet-50	ImageNet	x1	256	1e-1	0.9	-	-	-	-	-	-	-	-	-	-	-	-
		x2	256	1e-1	0.9	1e-1	0.9	-	-	-	-	1e-5	0.9	1e-1	0.9	1e-4	0.9

N = Number of networks in the ensemble. B = Batch Size, lr = Learning Rate, m = Momentum

Other hyperparameters that are common across all experiments are the following:

Weight Decay = $5e-4$

Nesterov = $True$