

Ensemble Methods

ImageNet 2017 – Object detection

Rank	Team name	Entry description	Mean AP
1	BDAT	Submission4	0.73222
2	NUS-Qihoo_DPNs (DET)	Ensemble of DPN models	0.65693
3	KAISTNIA_ETRI	Ensemble Model 5	0.61022

Rank	Submitter	System
1	University of Edinburgh	Uedin-
2	Salesforce metamind	Metarr
3	University of Edinburgh	Uedin-

$$L_E(\mathcal{D}) = \sum_{i=1}^{N} \sum_{m \in [M]} \ell(y_i, f_m(\mathbf{x}_i)).$$

$$L_O(\mathcal{D}) = \sum_{i=1}^N \min_{m \in [M]} \ell(y_i, f_m(\mathbf{x}_i)).$$

Algorithm for optimizing the confident oracle loss: Ensemble method has been successfully applied to many applications: WMT 2016 1. Sample random batch $\mathcal{B} \subset \mathcal{D}$. BLEU 2. Compute the loss of each model per each batch. 34.8 nmt-ensemble 3. Most accurate model trains the task-specific loss. 32.2 nmt-single Independent ensemble (IE) [Ciregan et al., 2012] trains models indetion to uniform one. pendently: 5. Repeat steps $1 \sim 4$ until convergence. Classification on CIFAR-10 using 5 CNNs (2 Conv + 2 FC): Class-wise test set accuracy • It generally improves the performance by reducing the variance. Multiple choice learning (MCL) [Guzman et al., 2012] makes models specialized for subset: • It can produce diverse and plausible outputs. • e_{ij} = test set accuracy of *j*-th model on class *i* data. • MCL and CMCL make each model specialized for certain classes, Our Contribution and Key Ideas while IE does not. We propose a **confident multiple choice learning (CMCL)**. 2 Histogram of the predictive entropy • Confident oracle loss: integer programming variant of $L_O(\mathcal{D})$. CIFAR-10 (non-specialized) CIFAR-10 (non-specialized) SVHN (unseen) 🔶 SVHN (unseen) 0.4 ś 0.4 – **c** w 0.4 $+\beta\left(1-v_{i}^{m}\right)D_{KL}\left(\mathcal{U}\left(y\right)\parallel P_{\theta_{m}}\left(y\mid\mathbf{x}_{i}\right)\right)$ (1a) 0.5 1.0 1.5 2.0 2.5 0 0.5 1.0 1.5 2.0 2.5 Entropy Entropy (c) IE with AT (a) MCL (b) CMCL subject to $\sum v_i^m = 1, \ \forall i. \ v_i^m \in \{0, 1\}, \quad \forall i, m.$ (1b) • For non-specialized data (i.e., accuracy < 80%) and unseen dataset (i.e., SVHN), ensemble members of CMCL are not over-• Feature sharing: stochastically shares the features from models. confident. Pool1 Hidden Shared Feature Masked 3 Contribution by each technique Feature A Feature B $A + B_{1}$ $\left(\bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \right)$ Input |0000|Α 0000L 10000 0000 Shared Feature Hidden Masked $B + A_1$ Feature A Feature B • Random labeling: noisy unbiased estimator of gradients. $\nabla_{\theta} D_{KL} \left(\mathcal{U} \left(y \right) \parallel P_{\theta} \left(y \mid \mathbf{x} \right) \right) \simeq -\frac{1}{S} \sum \nabla_{\theta} \log P_{\theta} \left(y^s \mid \mathbf{x} \right).$ mance of CMCL.

$$L_{C}(\mathcal{D}) = \min_{v_{i}^{m}} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(v_{i}^{m} \ell\left(y_{i}, P_{\theta_{m}}\left(y \mid \mathbf{x}_{i}\right)\right) \right)$$



Confident Multiple Choice Learning

Kimin Lee¹, Changho Hwang¹, KyoungSoo Park¹, Jinwoo Shin¹

¹Korea Advanced Institute of Science and Technology (KAIST)

Training Algorithm and Effects of CMCL

- 4. Other models minimize the KL divergence from predictive distribu-

	UIU.	55 VI			301	a	Juan	acy								
Airplane	0.0 %	0.0 %	93.6 %	0.0 %	0.0 %		95.8%	0.0%	4.4%	12.2%	2.2%	86.6%	85.5%	86.4%	85.7%	86.0%
itomobile	0.0 %	0.0 %	96.1 %	0.0 %	0.0 %	1 [0.0%	0.0%	0.8%	98.6%	9.0%	90.7%	90.3%	90.5%	90.6%	90.5%
Bird	99.9 %	0.0 %	0.0 %	0.0 %	0.0 %		0.1%	0.3%	2.4%	4.1%	94.0%	75.4%	75.9%	74.5%	76.3%	76.5%
Cat	0.0 %	0.0 %	95.6 %	0.0 %	0.0 %		94.5%	2.6%	0.0%	0.0%	0.2%	68.5%	66.5%	66.1%	67.1%	67.1%
Deer	0.0 %	0.0 %	0.0 %	97.5 %	0.0 %		0.0%	23.6%	1.2%	98.7%	4.5%	85.8%	86.3%	86.1%	86.1%	86.2%
Dog	0.0 %	97.0 %	0.0 %	0.0 %	0.0 %		15.8%	8.0%	2.9%	4.7%	91.7%	76.3%	75.6%	77.5%	75.0%	76.5%
Frog	0.0 %	0.0 %	0.0 %	0.0 %	97.7 %		7.1%	0.9%	99.2%	2.7%	0.0%	90.1%	90.7%	90.3%	91.4%	90.6%
Horse	0.0 %	0.0 %	0.0 %	0.0 %	97.2 %		0.0%	0.0%	98.1%	0.0%	0.0%	87.3%	86.9%	86.6%	86.3%	87.2%
Ship	0.0 %	0.0 %	0.0 %	97.2 %	0.0 %		0.0%	97.3%	0.0%	0.0%	0.0%	91.6%	91.6%	91.4%	91.7%	90.7%
Truck	0.0 %	97.4 %	0.0 %	0.0 %	0.0 %] [0.5%	96.1%	0.0%	0.0%	28.0%	90.4%	89.3%	89.8%	90.0%	90.0%
	1	2	3	4	5		1	2	3	4	5	1	2	3	4	5
(a) MCL						(b) CMC	CL				(c) IE				



Ensemble	Feature	Stochastic	Oracle	Top-1
Method	Sharing	Labeling	Error Rate	Error Rate
IE	-	-	10.65%	15.34%
MCL	-	-	4.40%	60.40%
	-	-	4.49%	15.65%
CMCL	\checkmark	-	5.12%	14.83%
	\checkmark	\checkmark	3.32%	14.78%

• Both feature sharing and stochastic labeling improve the perfor-

ICML Sydney 34-th International Conference on Machine Learning

Experiments

Image classification • Ensemble of small-scale CNN models.

	т.	Ensemble	Size $M = 5$	Ensemble Size $M = 10$		
Ensemble Method	K	Oracle Error Rate	Top-1 Error Rate	Oracle Error Rate	Top-1 Error Rate	
IE	IE -		15.34%	9.26%	15.34%	
	1	4.40%	60.40%	0.00%	76.88%	
МСІ	2	3.75%	20.66%	1.46%	49.31%	
MCL	3	4.73%	16.24%	1.52%	22.63%	
	4	5.83%	15.65%	1.82%	17.61%	
	1	3.32%	14.78%	1.96%	14.28%	
CMCI	2	3.69%	14.25% (-7.11%)	1.22%	13.95%	
CIVICL	3	4.38%	14.38%	1.53%	14.00%	
	4	5.82%	14.49%	1.73%	13.94% (-9.13%)	

• Ensemble of 5 large-scale CNN models.

Madal Nama	Ensemble	CIFAF	R-10	SVHN			
wodel mame	Method	Oracle Error Rate	Top-1 Error Rate	Oracle Error Rate	Top-1 Error Rate		
	- (single)	10.65%	10.65%	5.22%	5.22%		
VCCNat 17	IE	3.27%	8.21%	1.99%	4.10%		
VGGINEL-1/	MCL	2.52%	45.58%	1.45%	45.30%		
	CMCL	2.95%	7.83% (-4.63%)	1.65%	3.92% (-4.39%)		
	- (single)	10.15%	10.15%	4.59%	4.59%		
CooglaNat 19	IE	3.37%	7.97%	1.78%	3.60%		
GoogLeinet-18	MCL	2.41%	52.03%	1.39%	37.92%		
	CMCL	2.78%	7.51% (-5.77%)	1.36%	3.44% (-4.44%)		
	- (single)	14.03%	14.03%	5.31%	5.31%		
DasNat 20	IE	3.83%	10.18%	1.82%	3.94%		
Keshet-20	MCL	2.47%	53.37%	1.29%	40.91%		
	CMCL	2.79%	8.75% (-14.05%)	1.42%	3.68% (-6.60%)		

• Figure 1(a) compares the effects of feature sharing.

Foreground-background segmentation



ensemble methods.



Figure 1: (a) Ensemble of *M* ResNets with 20 layers, (b) Top-1 error and (c) oracle error on iCoseg.

• Pixel-level classification problem with 2 classes, i.e., 0 (background) or 1 (foreground) using fully convolutional networks. • Foreground-background segmentation for a few samples.

• Figure 1(b) and 4(c) show both top-1 and oracle error rates for all