Confident Multiple Choice Learning

Kimin Lee[1], Changho Hwang[1], KyoungSoo Park[1], Jinwoo Shin[1]

[1]Korea Advanced Institute of Science and Technology (KAIST)

ICML | Sydney

34-th International Conference on Machine Learning

Ensemble Methods

Ensemble method has been successfully applied to many applications:

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

WMT 2016

Rank	Submitter	System	BLEU
1	University of Edinburgh	Uedin-nmt-ensemble	34.8
2	Salesforce metamind	Metamind-ensemble	32.8
3	University of Edinburgh	Uedin-nmt-single	32.2

Independent ensemble (IE) [Ciregan et al., 2012] trains models independently:

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

- It generally improves the performance by reducing the variance.

Multiple choice learning (MCL) [Guzman et al., 2012] makes models specialized for subset:

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

- It can produce diverse and plausible outputs.

Our Contribution and Key Ideas

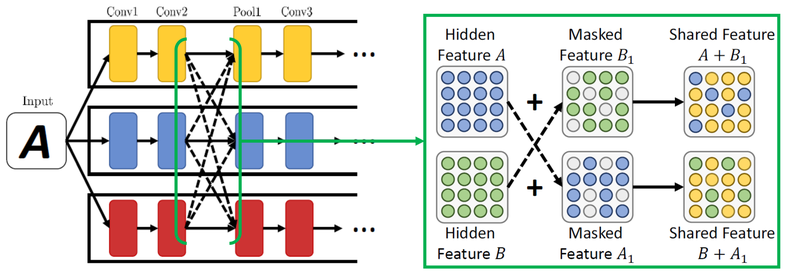
We propose a **confident multiple choice learning (CMCL).**

- Confident oracle loss: integer programming variant of $L_O(\mathcal{D})$.

$$L_C(\mathcal{D}) = \min_{v_i^m} \sum_{i=1}^{N}\sum_{m=1}^{M} \Big(v_i^m \ell(y_i, P_{\theta_m}(y \mid \mathbf{x}_i)) + \beta(1-v_i^m) D_{KL}(\mathcal{U}(y) \parallel P_{\theta_m}(y \mid \mathbf{x}_i)) \Big) \quad (1a)$$

$$\text{subject to } \sum_{m=1}^{M} v_i^m = 1,\ \forall i.\ v_i^m \in \{0,1\}, \quad \forall i, m. \quad (1b)$$

- Feature sharing: stochastically shares the features from models.



- Random labeling: noisy unbiased estimator of gradients.

$$\nabla_\theta D_{KL}(\mathcal{U}(y) \parallel P_\theta(y \mid \mathbf{x})) \backsimeq -\frac{1}{S}\sum_{s} \nabla_\theta \log P_\theta(y^s \mid \mathbf{x}).$$

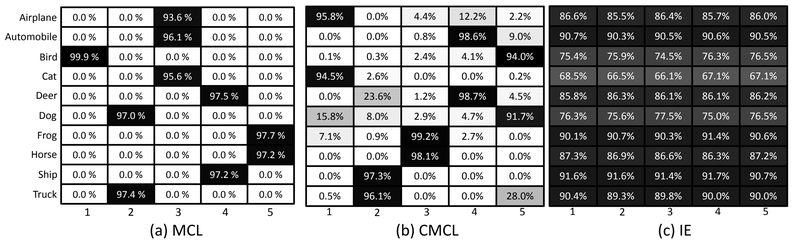
Training Algorithm and Effects of CMCL

Algorithm for optimizing the confident oracle loss:

1. Sample random batch $\mathcal{B} \subset \mathcal{D}$.
2. Compute the loss of each model per each batch.
3. Most accurate model trains the task-specific loss.
4. Other models minimize the KL divergence from predictive distribution to uniform one.
5. Repeat steps $1 \sim 4$ until convergence.

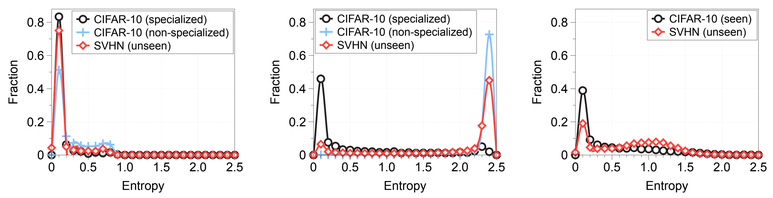
Classification on CIFAR-10 using 5 CNNs (2 Conv + 2 FC):

1 **Class-wise test set accuracy**



- e_{ij} = test set accuracy of j-th model on class i data.
- MCL and CMCL make each model specialized for certain classes, while IE does not.

2 **Histogram of the predictive entropy**



(a) MCL (b) CMCL (c) IE with AT

- For non-specialized data (i.e., accuracy < 80%) and unseen dataset (i.e., SVHN), ensemble members of CMCL are not over-confident.

3 **Contribution by each technique**

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

- Both feature sharing and stochastic labeling improve the performance of CMCL.

Experiments

Image classification

- Ensemble of small-scale CNN models.

Ensemble Method	K	Ensemble Size $M=5$		Ensemble Size $M=10$	
		Oracle Error Rate	Top-1 Error Rate	Oracle Error Rate	Top-1 Error Rate
IE	-	10.65%	15.34%	9.26%	15.34%
MCL	1	4.40%	60.40%	**0.00%**	76.88%
	2	3.75%	20.66%	1.46%	49.31%
	3	4.73%	16.24%	1.52%	22.63%
	4	5.83%	15.65%	1.82%	17.61%
CMCL	1	**3.32%**	14.78%	1.96%	14.28%
	2	3.69%	**14.25% (-7.11%)**	1.22%	13.95%
	3	4.38%	14.38%	1.53%	14.00%
	4	5.82%	14.49%	1.73%	**13.94% (-9.13%)**

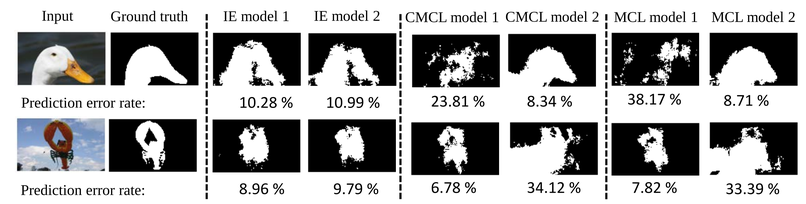
- Ensemble of 5 large-scale CNN models.

Model Name	Ensemble Method	CIFAR-10		SVHN	
		Oracle Error Rate	Top-1 Error Rate	Oracle Error Rate	Top-1 Error Rate
VGGNet-17	- (single)	10.65%	10.65%	5.22%	5.22%
	IE	3.27%	8.21%	1.99%	4.10%
	MCL	**2.52%**	45.58%	**1.45%**	45.30%
	CMCL	2.95%	**7.83% (-4.63%)**	1.65%	**3.92% (-4.39%)**
GoogLeNet-18	- (single)	10.15%	10.15%	4.59%	4.59%
	IE	3.37%	7.97%	1.78%	3.60%
	MCL	**2.41%**	52.03%	1.39%	37.92%
	CMCL	2.78%	**7.51% (-5.77%)**	**1.36%**	**3.44% (-4.44%)**
ResNet-20	- (single)	14.03%	14.03%	5.31%	5.31%
	IE	3.83%	10.18%	1.82%	3.94%
	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

- 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 ensemble methods.

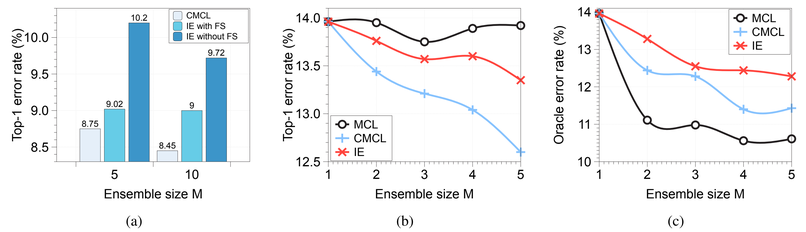


Figure 1: (a) Ensemble of M ResNets with 20 layers, (b) Top-1 error and (c) oracle error on iCoseg.