Confident Multiple Choice Learning

Kimin Lee, Changho Hwang, KyoungSoo Park, Jinwoo Shin

Korea Advanced Institute of Science and Technology (KAIST)

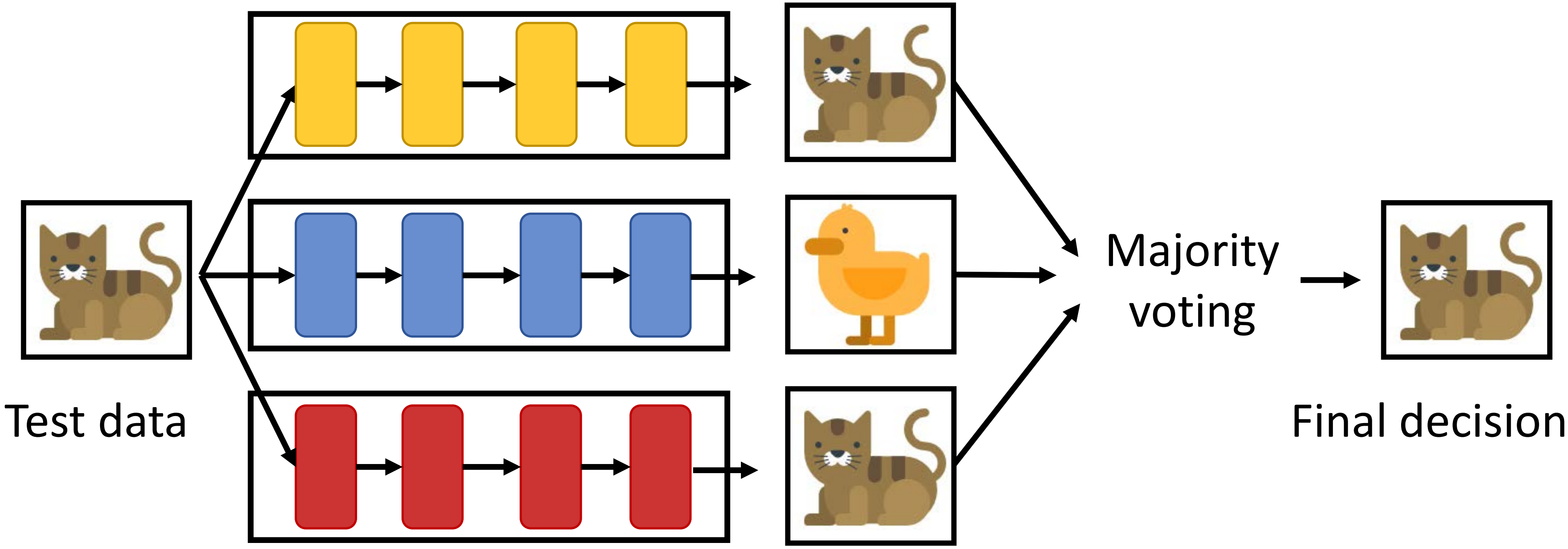
ICML 2017, Sydney

Outline

- Summary
  - Motivation
  - Main contribution
- Confident multiple choice learning (CMCL)
  - Preliminaries
  - Confident oracle loss
  - Feature sharing
  - Random labeling
- Experimental results
  - Image classification
  - Foreground-background segmentation

What is Ensemble Learning ?

- Train multiple models to try and solve the same problem
- Combine the outputs of them to obtain the final decision



- Bagging [Breiman’ 96], boosting [Freund’ 99] and mixture of experts [Jacobs’ 91]

[Freund’ 99] Freund, Yoav, Schapire, Robert, and Abe, N. A short introduction to boosting. Journal-Japanese Society For Arti- ficial Intelligence, 14(771-780):1612, 1999.
[Breiman’ 96] Breiman, Leo. Bagging predictors. Machine learning, 24 (2):123–140, 1996.
[Jacobs’ 91] Jacobs, Robert A, Jordan, Michael I, Nowlan, Steven J, and Hinton, Geoffrey E. Adaptive mixtures of local experts. Neural computation, 1991.

Successes of Ensemble Methods

- Ensemble methods have been successfully applied to enhancing performance in many machine learning applications

ImageNet 2017 – Object Detection

Rank	Team name	Entry description	Mean AP	# of categories won
1	BDAT	Submission4	0.732227	65
2	NUS-Qihoo_DPNs (DET)	Ensemble of DPN models	0.656932	9
3	**KAISTNIA_ETRI**	Ensemble Model 5	0.61022	1

* Table is from http://image-net.org/challenges/LSVRC/2017/results

WMT 2016 competition results

Rank	System	Submitter	BLEU	System Notes
1	Uedin-nmt-ensemble	University of Edinburgh	34.8	~. Ensemble of 4, reranked with right-to-~
2	Metamind-ensemble	Salesforce metamind	32.8	~. Ensemble of 3 checkpoints ~
3	Uedin-nmt-single	University of Edinburgh	32.2	~. Single model

* Table is from http://cs224d.stanford.edu/lectures/CS224d-Lecture9.pdf

⇨ High-performance teams employ ensemble methods !

Main Contribution

Problem

- Simple ensemble methods have been of typical choice for most applications involving deep neural networks
  - Relatively slow progress on developing more advanced ensembles specialized for deep neural networks

Main contributions

- We propose a new ensemble method specialized for deep neural networks based on advanced collaboration of ensemble members
- For Image classification,
  - Our method using residual networks provides 14.05% and 6.60% relative reductions in their errors from standard ensemble method on CIFAR and SVHN datasets, respectively
- For foreground-background segmentation,
  - Our method using fully convolutional networks achieve up to 6.77% relative reduction in their errors from standard ensemble method on iCoseg dataset

Outline

- Summary
  - Motivation
  - Main contribution

- Confident multiple choice learning (CMCL)
  - Preliminaries
  - Confident oracle loss
  - Feature sharing
  - Random labeling

- Experimental results
  - Image classification
  - Foreground-background segmentation

Ensemble Methods for Deep Neural Networks

- Independent Ensemble (IE) [Ciregan' 12]
  - Independently train each model with random initialization

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

Var	Definition
$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$	training data
$(f_1, \ldots, f_M)$	M models
$\ell\left(y_i, f\left(\mathbf{x}\right)\right)$	task-specific loss

  - IE generally improves the performance by reducing the variance
  - However, IE does not produce diverse solution well

[Ciregan' 12] Ciregan, D., Meier, U. and Schmidhuber, J. Multi-column deep neural networks for image classification. In *CVPR, 2012.*

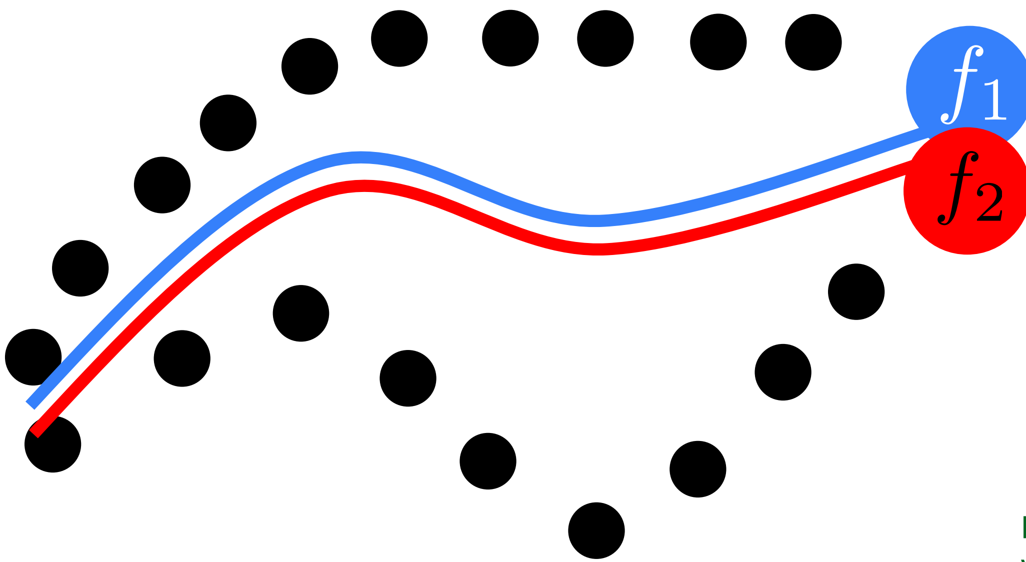
Ensemble Methods for Deep Neural Networks

- Independent Ensemble (IE) [Ciregan' 12]
  - Independently train each model with random initialization

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

Var	Definition
$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$	training data
$(f_1, \ldots, f_M)$	M models
$\ell(y_i, f(\mathbf{x}))$	task-specific loss

  - Toy example: regression using 2 models with mean squared error



Moving predictions "towards the mean"

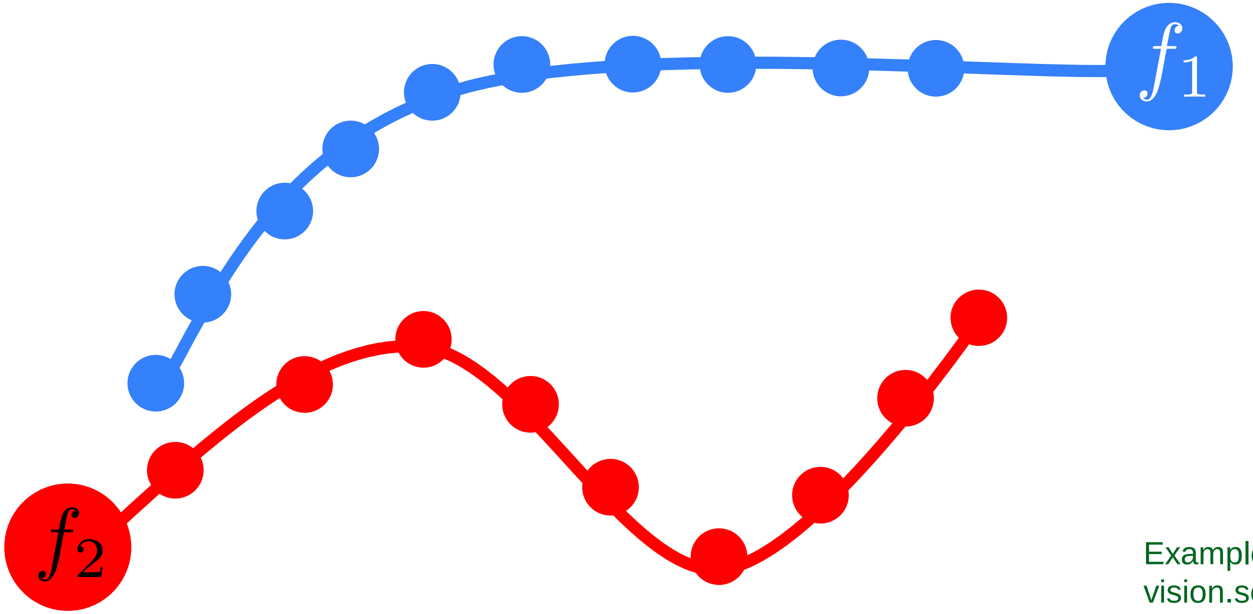
Example and figure are from vision.soic.indiana.edu/pres/diversity2016cvpr-slides.pptx

Ensemble Methods for Deep Neural Networks

- Multiple choice learning (MCL) [Guzman' 12]
  - Making each model specialized for certain subset of data

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

  - Toy example: regression using 2 models with mean squared error



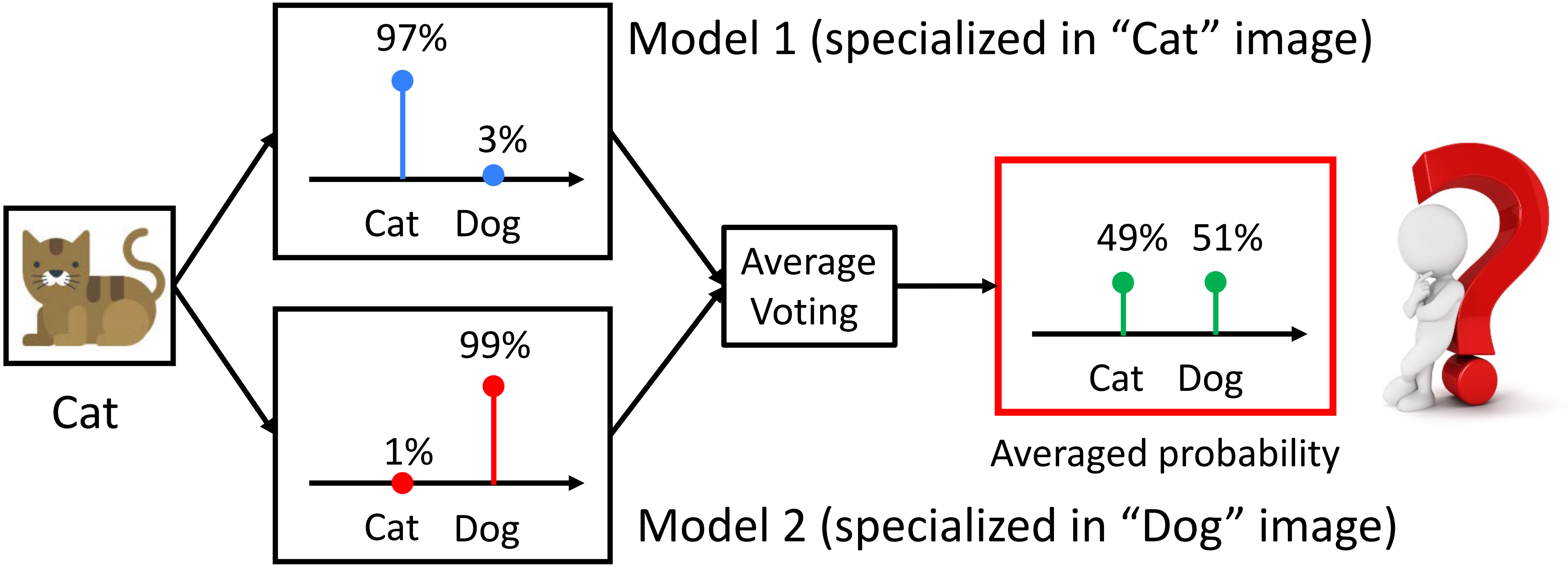
Example and figure are from vision.soic.indiana.edu/pres/diversity2016cvpr-slides.pptx

Ensemble Methods for Deep Neural Networks

- Multiple choice learning (MCL) [Guzman' 12]
  - Making each model specialized for certain subset of data

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

  - Overconfidence issues of MCL



Confident Multiple Choice Learning (CMCL)

- Making the specialized models with confident predictions
- Main components of our contributions

New loss: confident oracle loss

New architecture: feature sharing

New training method: random labeling

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

Ensemble Method	Feature Sharing	Stochastic Labeling	Top-1 Error Rate
IE	-	-	15.34%
MCL	-	-	60.40%
CMCL	-	-	15.65%
	✓	-	14.83%
	✓	✓	**14.78%**

Confident Oracle Loss

- Confident oracle loss

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

$$\text{subject to} \quad \sum_{m=1}^{M} v_i^m = 1, \quad \forall i, \tag{1b}$$

$$v_i^m \in \{0,1\}, \quad \forall i, m \tag{1c}$$

- Generating confident predictions by minimizing the KL divergence

D_{KL}: the KullbackLeibler (KL) divergence
$\mathcal{U}\left(y\right)$: the uniform distribution
v_i^m: a flag variable to decide the assignment of $\mathbf{x}_i$ to the m-th model
β: a penalty parameter
θ_m : model parameters
$P_{\theta_m}\left(y \mid \mathbf{x}\right)$: Predictive distribution of m-th model

Confident Oracle Loss

- Confident oracle loss

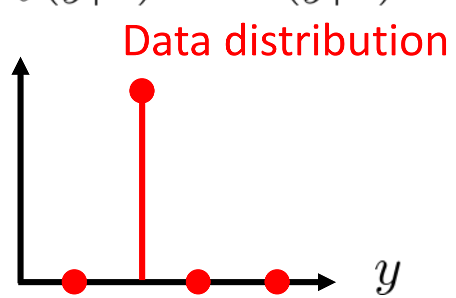
$$L_C(\mathcal{D}) = \min_{v_i^m} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\boxed{v_i^m \ell\left(y_i, P_{\theta_m}\left(y_i \mid \mathbf{x}_i\right)\right)} + \beta\left(1 - v_i^m\right) D_{KL}\left(\mathcal{U}\left(y\right) \parallel P_{\theta_m}\left(y \mid \mathbf{x}_i\right)\right) \right) \quad (1a)$$

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

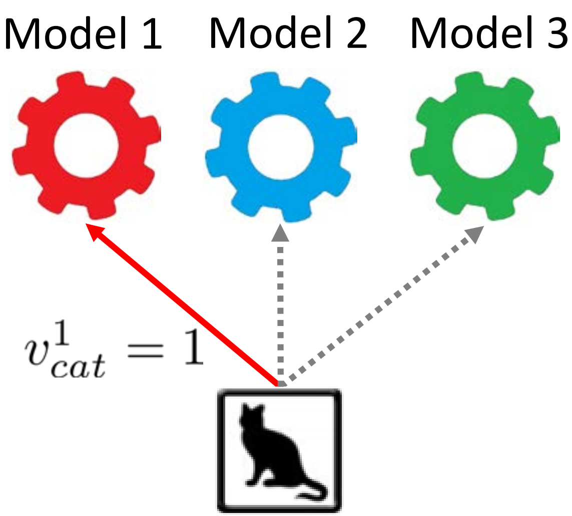
$$v_i^m \in \{0, 1\}, \quad \forall i, m \quad (1c)$$

- Interpretation

$$P_\theta(y|\mathbf{x}) \rightarrow P(y|\mathbf{x})$$



$[\text{Target data } (v_i = 1)]$



Confident Oracle Loss

- Confident oracle loss

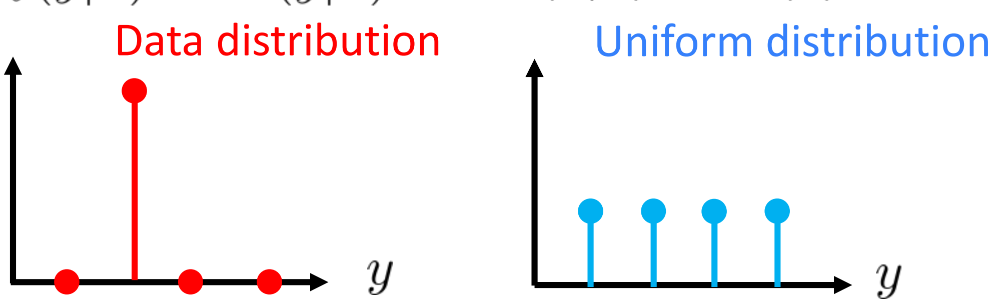
$$L_C(\mathcal{D}) = \min_{v_i^m} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(v_i^m \ell\left(y_i, P_{\theta_m}\left(y_i \mid \mathbf{x}_i\right)\right) \boxed{+ \beta\left(1 - v_i^m\right) D_{KL}\left(\mathcal{U}\left(y\right) \parallel P_{\theta_m}\left(y \mid \mathbf{x}_i\right)\right)} \right) \tag{1a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} v_i^m = 1, \quad \forall i, \tag{1b}$$

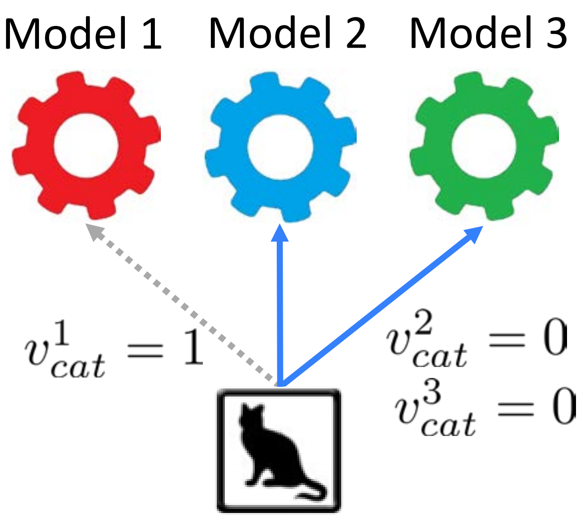
$$v_i^m \in \{0, 1\}, \quad \forall i, m \tag{1c}$$

- Interpretation

$P_\theta(y|\mathbf{x}) \to P(y|\mathbf{x})$ $\quad$ $P_\theta(y|\mathbf{x}) \to \mathcal{U}(y)$



$[\text{Target data } (v_i = 1)]$ $[\text{Non-target data } (v_i = 0)]$



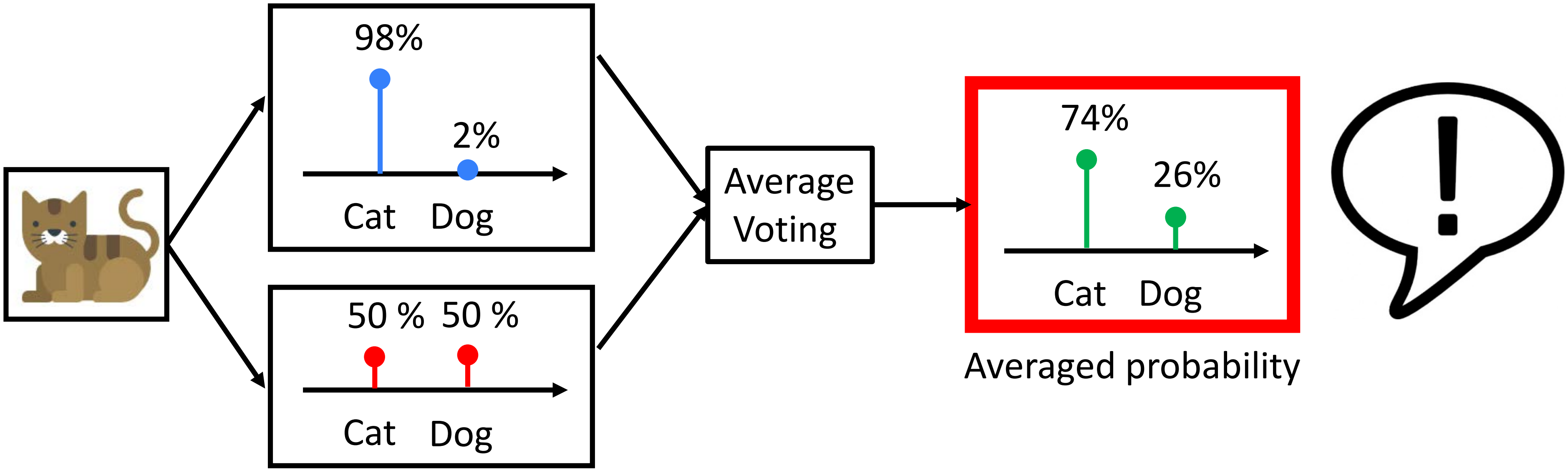
Confident Oracle Loss

- Confident oracle loss

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

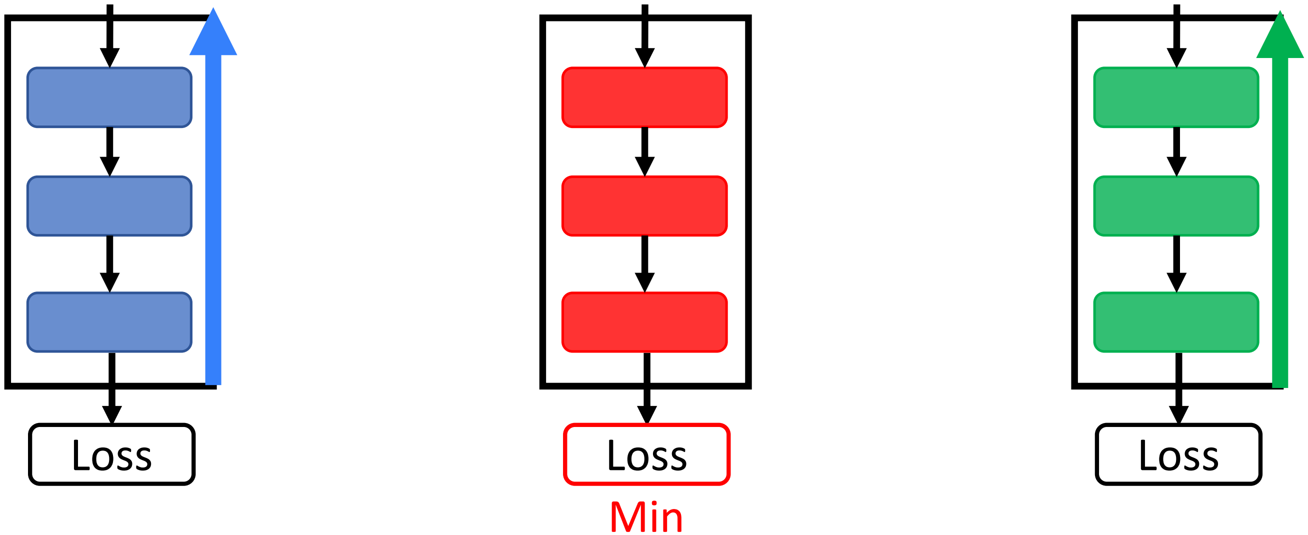
$$\text{subject to} \quad \sum_{m=1}^{M} v_i^m = 1, \quad \forall i, \tag{1b}$$

$$v_i^m \in \{0,1\}, \quad \forall i, m \tag{1c}$$



Algorithm Description

- Stochastic alternating procedure based on [Lee' 16]
  - Assumption: models are trained by stochastic gradient
  - For each batch
    - Compute the confident oracle loss of each model
    - Most accurate model trains the task-specific loss
    - Other models minimize the KL divergence loss
  - Repeat until convergence



[Lee' 16] Lee, S., Prakash, S.P.S., Cogswell, M., Ranjan, V., Crandall, D. and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. In NIPS, 2016.

Effect of Confident Oracle Loss

- Experiments on CIFAR-10 using 5 CNNs (2 Conv + 2 FC)
  - Class-wise test set accuracy

	1	2	3	4	5
Airplane	0.0 %	0.0 %	93.6 %	0.0 %	0.0 %
Automobile	0.0 %	0.0 %	96.1 %	0.0 %	0.0 %
d	99.9 %	0.0 %	0.0 %	0.0 %	0.0 %
Cat	0.0 %	0.0 %	95.6 %	0.0 %	0.0 %
Deer	0.0 %	0.0 %	0.0 %	97.5 %	0.0 %
Dog	0.0 %	97.0 %	0.0 %	0.0 %	0.0 %
Frog	0.0 %	0.0 %	0.0 %	0.0 %	97.7 %
Horse	0.0 %	0.0 %	0.0 %	0.0 %	97.2 %
Ship	0.0 %	0.0 %	0.0 %	97.2 %	0.0 %
Truck	0.0 %	97.4 %	0.0 %	0.0 %	0.0 %

(a) Multiple choice learning (MCL)

	1	2	3	4	5
Airplane	95.8%		4.4%	12.2%	2.2%
Automobile	0.0%	0.0%	0.8%	98.6%	9.0%
Bird	0.1%	0.3%	2.4%	4.1%	94.0%
Cat	94.5%		0.0%	0.0%	0.2%
Deer	0.0%	23.6%	1.2%	98.7%	4.5%
Dog	15.8%	8.0%	2.9%	4.7%	91.7%
Frog	7.1%	0.9%	99.2%	2.7%	0.0%
Horse	0.0%	0.0%	98.1%	0.0%	0.0%
Ship	0.0%	97.3%	0.0%	0.0%	0.0%
Truck	0.5%	96.1%	0.0%	0.0%	28.0%

(b) Confident MCL (CMCL)

	1	2	3	4	5
Airplane	86.6%	85.5%	86.4%	85.7%	86.0%
Automobile	90.7%	90.3%	90.5%	90.6%	90.5%
Bird	75.4%	75.9%	74.5%	76.3%	76.5%
Cat	68.5%	66.5%	66.1%	67.1%	67.1%
Deer	85.8%	86.3%	86.1%	86.1%	86.2%
Dog	76.3%	75.6%	77.5%	75.0%	76.5%
Frog	90.1%	90.7%	90.3%	91.4%	90.6%
Horse	87.3%	86.9%	86.6%	86.3%	87.2%
Ship	91.6%	91.6%	91.4%	91.7%	90.7%
Truck	90.4%	89.3%	89.8%	90.0%	90.0%

(c) Independent ensemble (IE)

⇨ Both MCL and CMCL make each model specialized for certain classes, while IE does not

Effect of Confident Oracle Loss

- Experiments on CIFAR-10 using 5 CNNs (2 Conv + 2 FC)
  - Class-wise test set accuracy

	1	2	3	4	5
Airplane	0.0 %	0.0 %	93.6 %	0.0 %	0.0 %
Automobile	0.0 %	0.0 %	96.1 %	0.0 %	0.0 %
Bird	99.9 %	0.0 %	0.0 %	0.0 %	0.0 %
Cat	0.0 %	0.0 %	95.6 %	0.0 %	0.0 %
Deer	0.0 %	0.0 %	[illegible]	97.5 %	0.0 %
Dog	0.0 %	97.0 %	[illegible]	0.0 %	0.0 %
Frog	0.0 %	0.0 %	[illegible]	0.0 %	97.7 %
Horse	0.0 %	0.0 %	0.0 %	0.0 %	97.2 %
Ship	0.0 %	0.0 %	0.0 %	97.2 %	0.0 %
Truck	0.0 %	97.4 %	0.0 %	0.0 %	0.0 %

(a) Multiple choice learning (MCL)

	1	2	3	4	5
Airplane	95.8%	0.0%	4.4%	12.2%	2.2%
Automobile	0.0%	0.0%	0.8%	98.6%	9.0%
Bird	0.1%	0.3%	2.4%	4.1%	94.0%
Cat	94.5%	2.6%	0.0%	0.0%	0.2%
Deer	[illegible]	23.6%	1.2%	98.7%	4.5%
Dog	[illegible]	8.0%	2.9%	4.7%	91.7%
Frog	7.1%	0.9%	99.2%	2.7%	0.0%
Horse	0.0%	0.0%	98.1%	0.0%	0.0%
Ship	0.0%	97.3%	0.0%	0.0%	0.0%
Truck	0.5%	96.1%	0.0%	0.0%	28.0%

(b) Confident MCL (CMCL)

	1	2	3	4	5
Airplane	86.6%	85.5%	86.4%	85.7%	86.0%
Automobile	90.7%	90.3%	90.5%	90.6%	90.5%
Bird	75.4%	75.9%	74.5%	76.3%	76.5%
Cat	68.5%	66.5%	66.1%	67.1%	67.1%
Deer	85.8%	86.3%	86.1%	86.1%	86.2%
Dog	76.3%	75.6%	77.5%	75.0%	76.5%
Frog	90.1%	90.7%	90.3%	91.4%	90.6%
Horse	87.3%	86.9%	86.6%	86.3%	87.2%
Ship	91.6%	91.6%	91.4%	91.7%	90.7%
Truck	90.4%	89.3%	89.8%	90.0%	90.0%

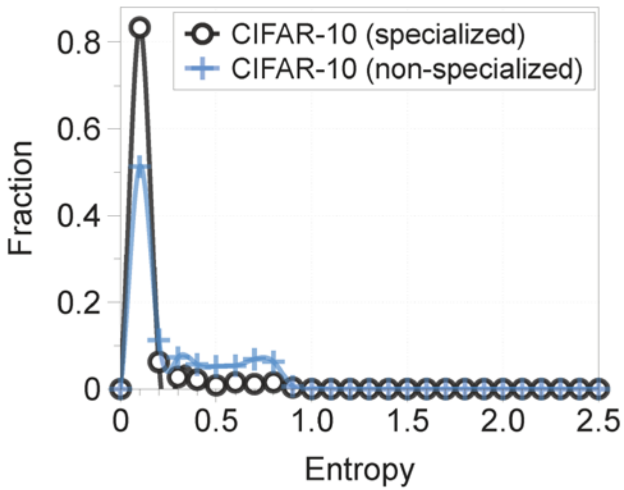
(c) Independent ensemble (IE)

⇨ Both MCL and CMCL make each model specialized for certain classes, while IE does not

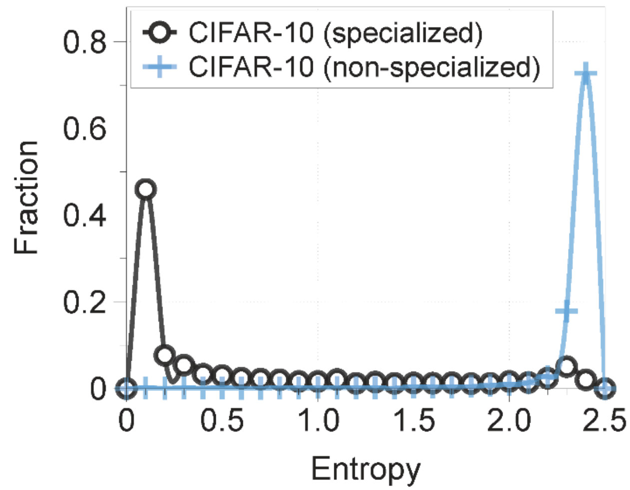
⇨ For specialized data, model trained by CMCL and MCL outperforms model trained by IE

Effect of Confident Oracle Loss

- Experiments on CIFAR-10 using 5 CNNs (2 Conv + 2 FC)
  - Histogram of the predictive entropy of model trained by each method



(a) MCL

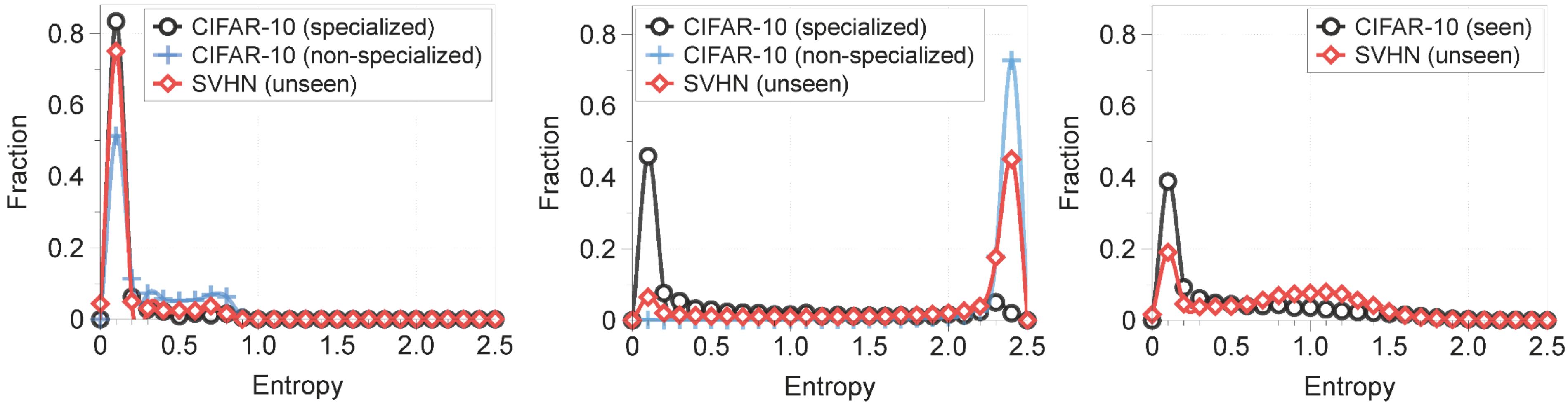


(b) CMCL

⇨ For non-specialized data (i.e., accuracy < 80%), ensemble members of CMCL are not overconfident compared to MCL

Effect of Confident Oracle Loss

- Experiments on CIFAR-10 using 5 CNNs (2 Conv + 2 FC)
  - Histogram of the predictive entropy of model trained by each method



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

⇨ For non-specialized data (i.e., accuracy < 80%), ensemble members of CMCL are not overconfident compared to MCL

⇨ For unseen dataset (SVHN), ensemble members of CMCL are not overconfident while models trained by MCL and IE are overconfident

Outline

- Summary
  - Motivation
  - Main contribution

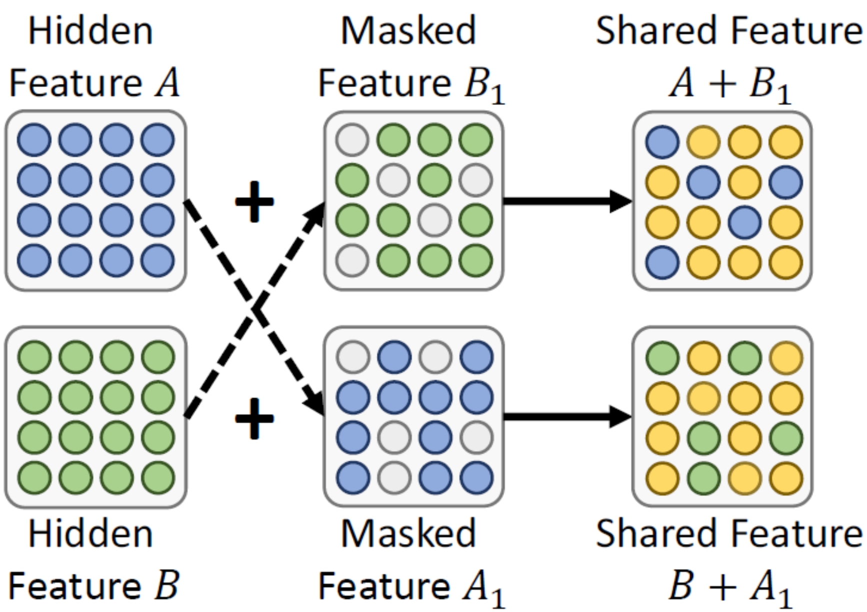
- Confident multiple choice learning (CMCL)
  - Preliminaries
  - Confident oracle loss
  - Feature sharing
  - Random labeling

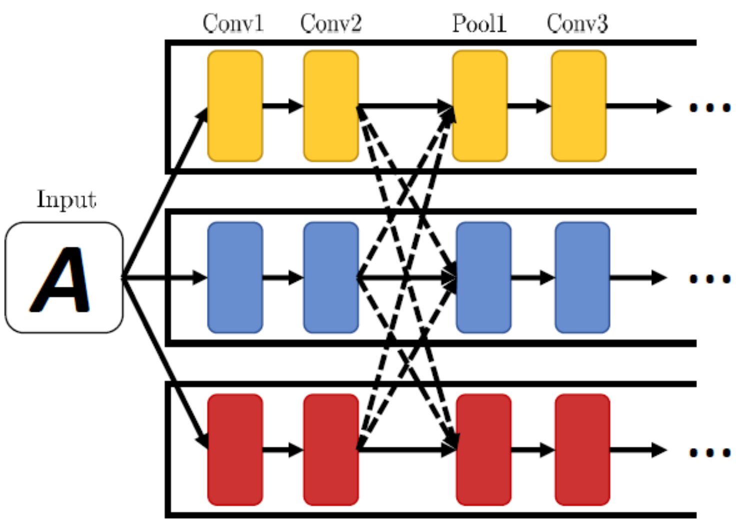
- Experimental results
  - Image classification
  - Foreground-background segmentation

Regularization Techniques for CMCL

- Feature sharing
  - Motivation: extracting general features from data
  - Stochastically shares the features from ensemble members

$$\mathbf{h}_m^\ell(\mathbf{x}) = \phi\left(\mathbf{W}_m^\ell\left(\mathbf{h}_m^{\ell-1}(\mathbf{x}) + \sum_{n\neq m}\boldsymbol{\sigma}_{nm}^\ell \star \mathbf{h}_n^{\ell-1}(\mathbf{x})\right)\right)$$





Sharing lower layer (before 1st pooling layer)

Regularization Techniques for CMCL

- Random labeling
  - Motivation: efficiency in computation and regularization effect
  - By definition,

$$\nabla_\theta D_{KL}\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y \mid \mathbf{x}\right)\right) = -\mathbb{E}_{\mathcal{U}(y)}[\nabla_\theta \log P_\theta\left(y \mid \mathbf{x}\right)].$$

  - Noisy unbiased estimator with Monte Carlo samples

$$\boxed{\underbrace{\nabla_\theta D_{KL}\left(\mathcal{U}\left(y\right) \parallel P_\theta\left(y \mid \mathbf{x}\right)\right)}_{\text{Exact gradient}} \backsimeq -\frac{1}{S}\sum_s \underbrace{\nabla_\theta \log P_\theta\left(y^s \mid \mathbf{x}\right)}_{\text{Gradient of cross entropy}},\ y^s \sim \mathcal{U}\left(y\right)}$$

    - Training using random labels ! ($S = 1$)

Outline

- Summary
  - Motivation
  - Main contribution

- Confident multiple choice learning (CMCL)
  - Preliminaries
  - Confident oracle loss
  - Feature sharing
  - Random labeling

- Experimental results
  - Image classification
  - Foreground-background segmentation

Experimental Results: Image Classification

- Classification test set error rates on CIFAR-10 and SVHN

CIFAR-10 [Krizhevsky' 09]

- 32 × 32 RGB
- 10 classes
- 50,000 training set
- 10,000 test set

SVHN [Netzer' 11]

- 32 × 32 RGB
- 10 classes
- 73,257 training set
- 26,032 test set

- Top-1 error
  - Select the class from averaged probability
- Oracle error
  - Measuring whether none of the members predict the correct class
- We use both feature sharing and random labeling for all experiments

Experimental Results: Image Classification

- Ensemble of small-scale CNNs (2 Conv + 2 FC)

Ensemble Method	K	Ensemble Size $M = 5$ Oracle Error Rate	Ensemble Size $M = 5$ Top-1 Error Rate	Ensemble Size $M = 10$ Oracle Error Rate	Ensemble Size $M = 10$ 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%)**

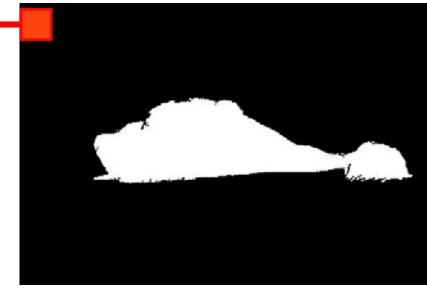
- Ensemble of 5 large-scale CNNs

Model Name	Ensemble Method	CIFAR-10 Oracle Error Rate	CIFAR-10 Top-1 Error Rate	SVHN Oracle Error Rate	SVHN 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%)**

Experimental Results: Image Segmentation

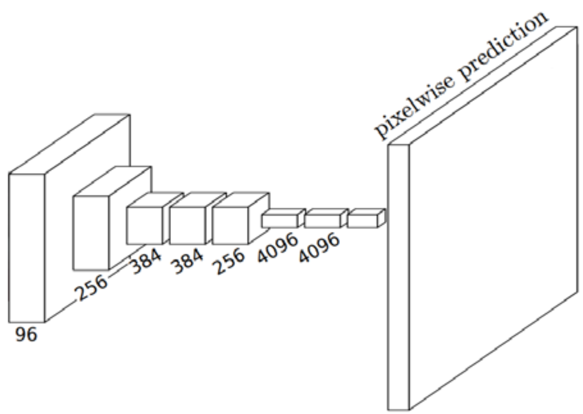
- iCoseg dataset

1(foreground) and 0 (background)



=

Pixel-level classification problem with 2 classes

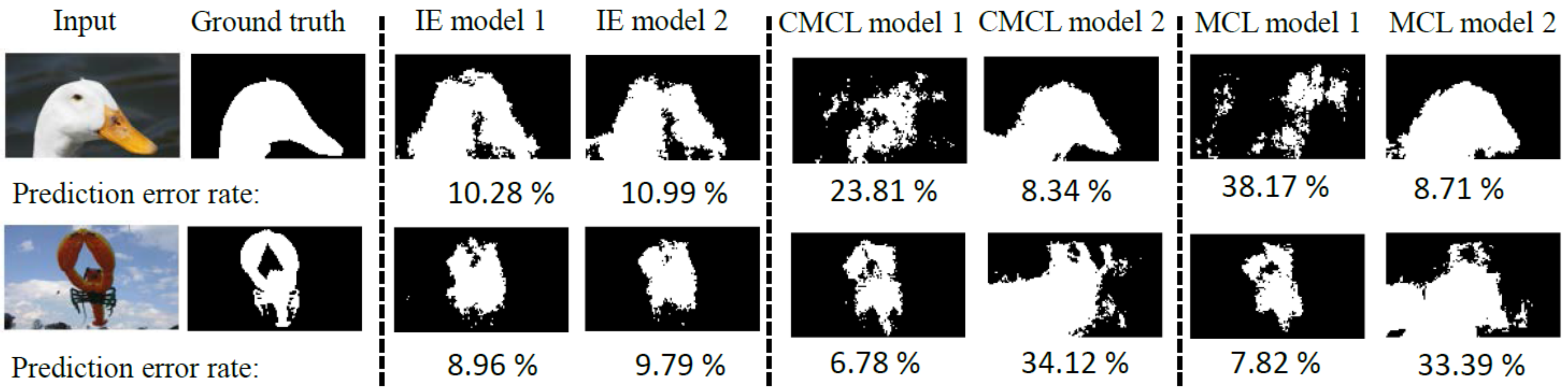


Fully convolutional neural networks (FCNs) [Long' 15]

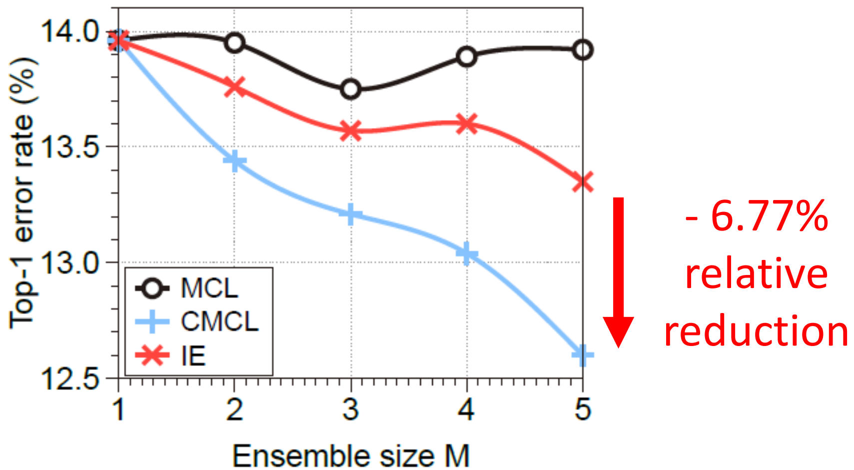
[Long' 15] Long, J., Shelhamer, E. and Darrell, T. Fully convolutional networks for semantic segmentation. In *CVPR, 2015*.

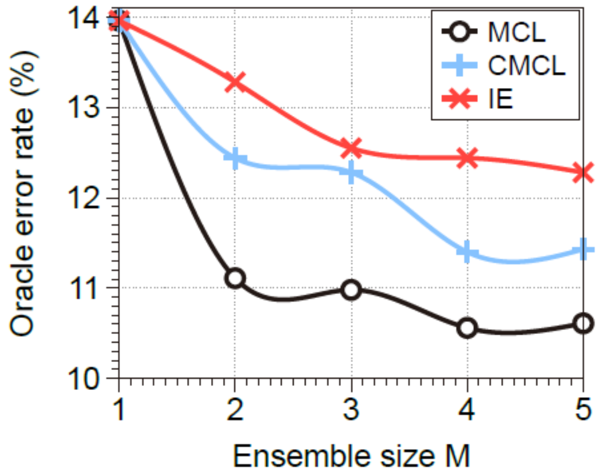
Experimental Results: Image Segmentation

- Prediction results of segmentation for few sample images



  - MCL and CMCL generate high-quality predictions





  - CMCL only outperforms IE in terms of the top-1 error

Conclusion

- We propose a new ensemble method coined CMCL
  - It produces diverse/plausible confident prediction of high quality !

- CMCL outperforms not only the known MCL, but also the traditional independent ensembles in classification and segmentation tasks.

- We believe that our new ensemble approach brings a refreshing angle for developing advanced large-scale deep networks in many related applications

Thank you !