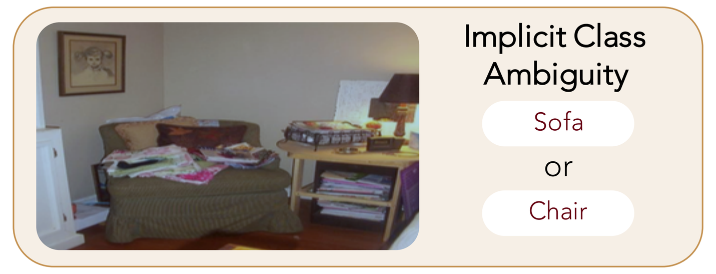
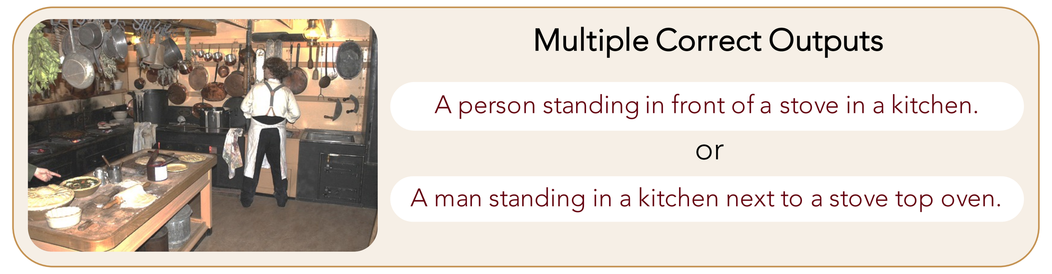
Stochastic Multiple Choice Learning For Training Diverse Deep Ensembles

Stefan Lee[1], Senthil Purushwalkam[3], Michael Cogswell[1], Viresh Ranjan[1], David Crandall[4], Dhruv Batra[2]
Virginia Tech[1] Georgia Tech[2] Carnegie Mellon University[3] Indiana University[4]

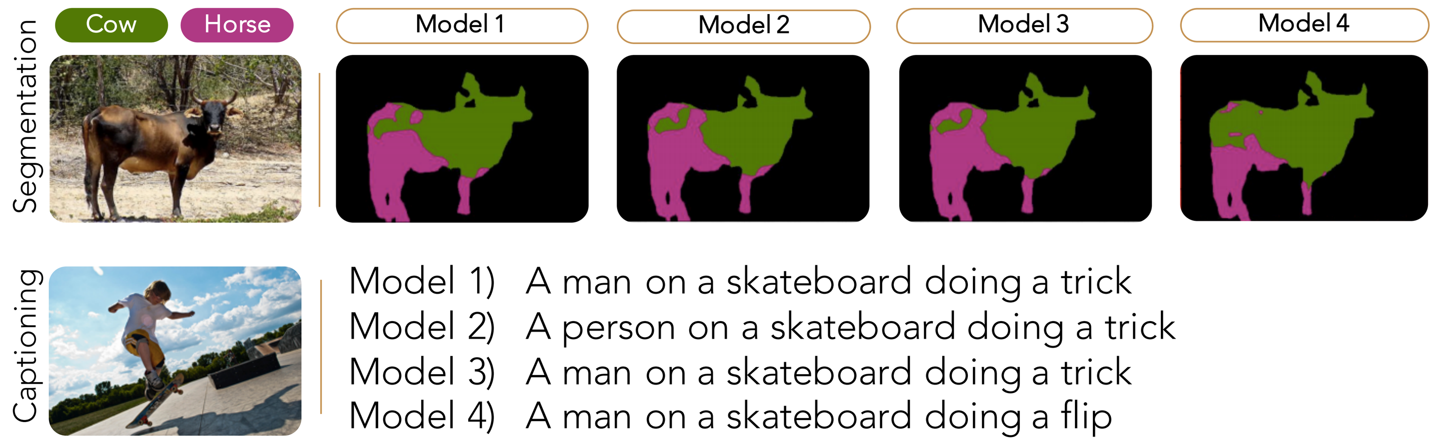
1 Overview: The Need for Diversity

Many interesting inference problems have some degree of ambiguity, often as an implicit property of an uncertain world.





In the face of ambiguity, humans can give multiple likely answers to articulate multimodal beliefs. One natural method to generate multiple outputs is to train an ensemble of models; however, we find independently trained networks typically produce similar outputs.



Despite random initializations and batches, we find deep networks converge to very similar solutions.

We propose that one cause for this is that training drives each model to have *low expected loss* across the training set, inhibiting specialization.

2 Stochastic Multiple Choice Learning (sMCL)

To encourage the specialization of ensemble members, we consider a loss with respect to a perfect oracle which picks the most correct solution from the M ensemble outputs,

$$\mathcal{L}_O(D) = \sum_{i=1}^{n} \min_{m \in [M]} \ell\left(y_i, f_m(x_i)\right) \; = \; \sum_{i=1}^{n} \sum_{m=1}^{M} p_{i,m} \, \ell\left(y_i, f_m(x_i)\right)$$

where $p_{i,m}$ is1 if model m has the lowest loss on example i and 0 otherwise. Holding $p_{i,m}$ fixed, the gradient with respect to a single models output $f_m(x_i)$ is

$$\frac{\partial \mathcal{L}_O}{\partial f_m(x_i)} = p_{i,m} \frac{\partial \ell(y_i, f_m(x_i))}{\partial f_m(x_i)}$$

As $p_{i,m}$ is only non-zero for the minimum predictor, this gradient is only zero for all other predictors.

Leads to to a simple training algorithm to minimize the oracle loss in SGD-based learners which we call Stochastic Multiple Choice Learning (sMCL).

sMCL Training Algorithm:

For each example in a batch:

1) Compute the loss of the example for each model in the ensemble.
2) Back-propagate the gradient only to the model with lowest loss.

This **'Winner-Take-Gradient'** training is agnostic to both model architecture and loss.

3 Specialization in Image Classification

To test sMCL in a simple setting, we train ensembles on CIFAR10 using a small CNN model. We find sharp, class-based specializations emerge in sMCL trained ensembles.

	Oracle Accuracy for Ensemble of Size				
	M = 2	3	4	5	6
sMCL	**85.47**	**88.65**	**93.10**	**94.29**	**96.20**
Guzman-Rivera et al. (2012)	84.69	88.44	92.09	94.64	95.53
Dey et al. (2015)	83.30	86.04	87.35	88.25	88.84
Independent Ensemble	83.03	86.58	88.51	90.09	92.33

sMCL results in substantial gains over other methods and between **2-5% accuracy over independent ensembles.**

sMCL Ensemble

	0	1	2	3
airplaine	0.10%	99.60%	0.10%	0.20%
automobile	0.20%	0.00%	99.80%	0.00%
bird	99.50%	0.10%	0.30%	0.10%
cat	0.10%	99.90%	0.00%	0.00%
deer	37.60%	0.00%	62.40%	0.00%
dog	0.10%	0.00%	0.00%	99.90%
frog	99.90%	0.10%	0.00%	0.00%
horse	0.00%	99.90%	0.00%	0.10%
ship	0.00%	0.00%	100.00%	0.00%
truck	0.00%	0.00%	0.20%	99.80%

Independent Ensemble

	0	1	2	3
airplaine	22.60%	33.20%	25.20%	19.00%
automobile	30.30%	20.30%	26.10%	23.30%
bird	19.70%	27.70%	26.30%	26.30%
cat	26.30%	26.40%	24.30%	23.00%
deer	20.00%	23.60%	31.70%	24.70%
dog	29.30%	21.40%	27.90%	21.40%
frog	17.30%	18.30%	32.50%	31.90%
horse	26.30%	26.80%	22.60%	24.30%
ship	25.30%	22.70%	24.40%	27.60%
truck	23.80%	20.60%	27.10%	28.50%

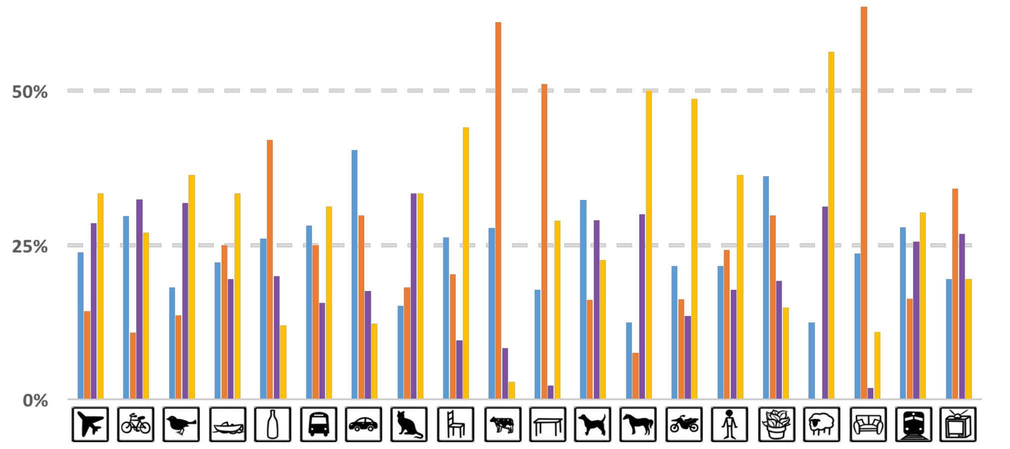
Percentage of each class assigned to each model at test time for sMCL and classical ensembles. The sMCL models becomes specialist on subsets of the classes.

4 Specialization in Semantic Segmentation

We train ensembles for semantic segmentation on PASCAL VOC 2011 using the fully-convolutional CNN architecture of Long et al. (2015).

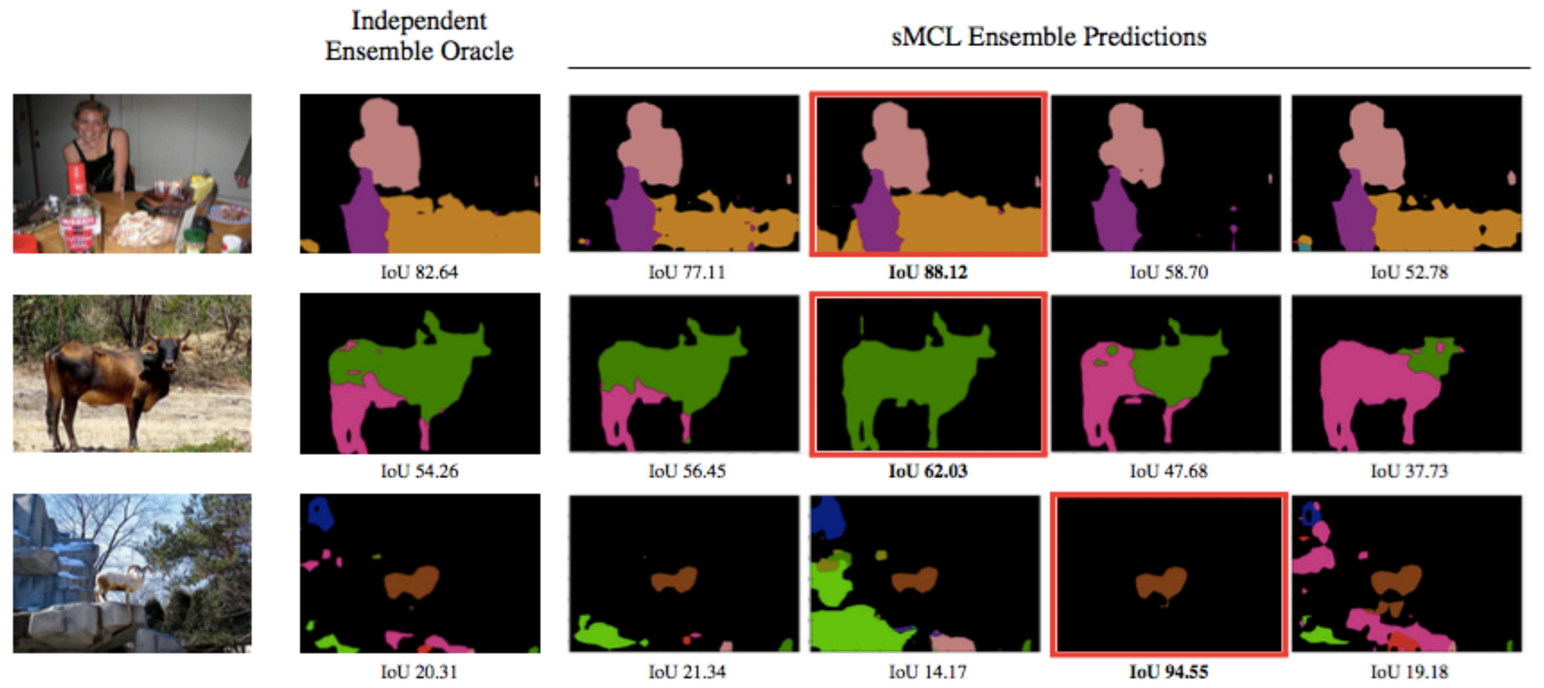
	Oracle Mean IoU for Ensemble of Size				
	M = 2	3	4	5	6
sMCL	**65.75**	**68.14**	**69.09**	**70.49**	**71.58**
Guzman-Rivera et al. (2012)	63.39	67.53	65.94	66.43	67.10
Dey et al. (2015)	63.63	63.70	63.71	64.73	63.75
Independent Ensemble	64.05	64.78	65.10	65.97	66.60

sMCL consistently outperforms baseline methods at oracle mean Intersection over Union (IoU). **Over 5 mean IoU over independent ensembles.**



Percentage of each class assigned to each model at test time (one color per model). Some class based specialization between models has emerged.

Samples images and segmentations from an sMCL ensemble and the top output of a classical ensemble. Minimum loss outputs are outlined in red. Notice that sMCL ensembles vary in the shape, class, and frequency of predicted segments.



5 Specialization in Image Captioning

We evaluate on the MSCOCO image captioning task, training ensembles of the CNN+LSTM model of Karpathy et al. (2015, with and without CNN fine-tuning.

Input	Independently Trained Networks	sMCL Ensemble
	A man riding a wave on top of a surfboard. A man riding a wave on top of a surfboard. A man riding a wave on top of a surfboard. A man riding a wave on top of a surfboard.	A man riding a wave on top of a surfboard. A person on a surfboard in the water. A surfer is riding a wave in the ocean. A surfer riding a wave in the ocean.
	A group of people standing on a sidewalk. A man is standing in the middle of the street. A group of people standing around a fire hydrant. A group of people standing around a fire hydrant	A man is walking down the street with an umbrell. A group of people sitting at a table with umbrellas. A group of people standing around a large plane. A group of people standing in front of a building
	A kitchen with a stove and a microwave. A white refrigerator freezer sitting inside of a kitchen. A white refrigerator sitting next to a window. A white refrigerator freezer sitting in a kitchen	A cat sitting on a chair in a living room. A kitchen with a stove and a sink. A cat is sitting on top of a refrigerator. A cat sitting on top of a wooden table
	A bird is sitting on a tree branch. A bird is perched on a branch in a tree. A bird is perched on a branch in a tree. A bird is sitting on a tree branch	A small bird perched on top of a tree branch. A couple of birds that are standing in the grass. A bird perched on top of a branch. A bird perched on a tree branch in the sky

Captions generated by classical ensembles tend to be only slightly different for a given image (row 1) and often produce outputs that are poor fits to individual images (row 4). sMCL ensembles are capable of specialization and their outputs are much more diverse and capture individual image characteristics well.

	Oracle CIDEr-D for Ensemble of Size				# Unique n-grams (M=5)				Avg. Length
	M = 2	3	4	5	n = 1	2	3	4	
sMCL	**0.822**	**0.862**	**0.911**	**0.922**	**713**	**2902**	**6464**	**15427**	10.21
Guzman-Rivera et al. (2012)	0.752	0.810	0.823	0.852	384	1565	3586	9551	9.87
Dey et al. (2015)	0.798	0.850	0.887	0.910	584	2266	4969	12208	10.26
Independent Ensemble	0.757	0.784	0.809	0.831	540	2003	4312	10297	10.24
sMCL (fine-tuned CNN)	**1.064**	**1.130**	**1.179**	**1.184**	**1135**	**6028**	**15184**	**35518**	10.43
Independent (fine-tuned CNN)	1.001	1.050	1.073	1.095	921	4335	10534	23811	10.33

sMCL trained ensembles consistently outperform other techniques and independent ensembles on oracle metrics and produce significantly more unique n-grams at similar sentence length

6 Conclusion

For many complex inference tasks, there is implicit ambiguity and/or multiple correct possible outputs. By directly optimizing for the oracle loss, our sMCL allows an ensemble to specialize in response to ambiguity and multimodal outputs distributions.

sMCL is **effective**, **easy to implement**, and **model and loss agnostic.**

PAPER

SLIDES