Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles

Stefan Lee
Virginia Tech
steflee@vt.edu

Senthil Purushwalkam
Carnegie Mellon University
spurushw@andrew.cmu.edu

Michael Cogswell
Virginia Tech
cogswell@vt.edu

Viresh Ranjan
Virginia Tech
rviresh@vt.edu

David Crandall
Indiana University
djcran@indiana.edu

Dhruv Batra
Virginia Tech
dbatra@vt.edu

Abstract

Many practical perception systems exist within larger processes that include interactions with users or additional components capable of evaluating the quality of predicted solutions. In these contexts, it is beneficial to provide these oracle mechanisms with multiple highly likely hypotheses rather than a single prediction. In this work, we pose the task of producing multiple outputs as a learning problem over an ensemble of deep networks – introducing a novel stochastic gradient descent based approach to minimize the loss with respect to an oracle. Our method is simple to implement, agnostic to both architecture and loss function, and parameter-free. Our approach achieves lower oracle error compared to existing methods on a wide range of tasks and deep architectures. We also show qualitatively that the diverse solutions produced often provide interpretable representations of task ambiguity.

1 Introduction

Perception problems rarely exist in a vacuum. Typically, problems in Computer Vision, Natural Language Processing, and other AI subfields are embedded in larger applications and contexts. For instance, the task of recognizing and segmenting objects in an image (semantic segmentation [6]) might be embedded in an autonomous vehicle [7], while the task of describing an image with a sentence (image captioning [18]) might be part of a system to assist visually-impaired users [22, 29].

In these scenarios, the goal of perception is often not to generate a single output but *a set of plausible hypotheses* for a 'downstream' process, such as a verification component or a human operator. These downstream mechanisms may be abstracted as *oracles* that have the capability to pick the correct solution from this set. Such a learning setting is called Multiple Choice Learning (MCL) [8], where the goal for the learner is to minimize oracle loss achieved by a set of M solutions. More formally, given a dataset of input-output pairs $\{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, the goal of classical supervised learning is to search for a mapping $F : \mathcal{X} \to \mathcal{Y}$ that minimizes a task-dependent loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}^+$ capturing the error between the actual labeling y_i and predicted labeling $\hat{y}_i$. In this setting, the learned function f makes a single prediction for each input and pays a penalty for that prediction. In contrast, Multiple Choice Learning seeks to learn a mapping $g : \mathcal{X} \to \mathcal{Y}^M$ that produces M solutions $\hat{Y}_i = (\hat{y}_i^1, \ldots, \hat{y}_i^M)$ such that oracle loss $\min_m \ell\left(y_i, \hat{y}_i^m\right)$ is minimized.

In this work, we fix the form of this mapping g to be the union of outputs from an ensemble of predictors such that $g(x) = \{f_1(x), f_2(x), \ldots, f_M(x)\}$, and address the task of training ensemble members $f_1, \ldots, f_M$ such that g minimizes oracle loss. Under our formulation, different ensemble members are free to specialize on subsets of the data distribution, so that collectively they produce a set of outputs which covers the space of high probability predictions well.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

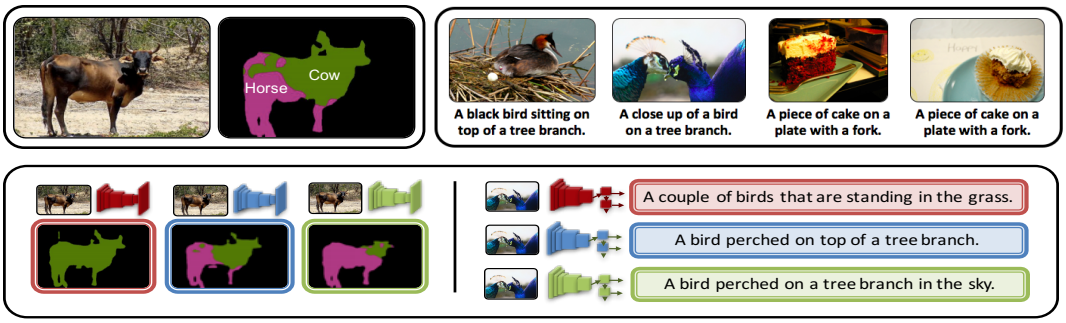


Figure 1: Single-prediction based models often produce solutions with low *expected loss* in the face of ambiguity; however, these solutions are often unrealistic or do not reflect the image content well (row 1). Instead, we train ensembles under a unified loss which allows each member to produce different outputs reflecting multi-modal beliefs (row 2). We evaluate our method on image classification, segmentation, and captioning tasks.

Diverse solution sets are especially useful for structured prediction problems with multiple reasonable interpretations, only one of which is correct. Situations that often arise in practical systems include:

– **Implicit class confusion.** The label space of many classification problems is often an arbitrary quantization of a continuous space. For example, a vision system may be expected to classify between tables and desks, despite many real-world objects arguably belonging to both classes. By making multiple predictions, this implicit confusion can be viewed explicitly in system outputs.
– **Ambiguous evidence.** Often there is simply not enough information to make a definitive prediction. For example, even a human expert may not be able to identify a fine-grained class (e.g., particular breed of dog) given an occluded or distant view, but they likely can produce a small set of reasonable guesses. In such cases, the task of producing a diverse set of possibilities is more clearly defined than producing one correct answer.
– **Bias towards the mode.** Many models have a tendency to exhibit mode-seeking behaviors as a way to reduce expected loss over a dataset (e.g., a conversation model frequently producing 'I don't know'). By making multiple predictions, a system can improve coverage of lower density areas of the solution space, without sacrificing performance on the majority of examples.

In other words, by optimizing for the oracle loss, a multiple-prediction learner can respond to ambiguity much like a human does, by making multiple guesses that capture multi-modal beliefs. In contrast, a single-prediction learner is forced to produce a solution with low expected loss in the face of ambiguity. Figure 1 illustrates how this can produce solutions that are not useful in practice. In semantic segmentation, for example, this problem often causes objects to be predicted as a mixture of multiple classes (like the horse-cow shown in the figure). In image captioning, minimizing expected loss encourages generic sentences that are 'safe' with respect to expected error but not very informative. For example, Figure 1 shows two pairs of images each having different image content but very similar, generic captions – the model knows it is safe to assume that birds are on branches and that cakes are eaten with forks.

In this paper, we generalize the Multiple Choice Learning paradigm [8, 9] to jointly learn ensembles of deep networks that *minimize the oracle loss directly*. We are the first to adapt these ideas to deep networks and we present a novel training algorithm that avoids costly retraining [8] and learning difficulty [5] of past methods. Our primary technical contribution is the formulation of a stochastic block gradient descent optimization approach well-suited to minimizing the oracle loss in ensembles of deep networks, which we call **Stochastic Multiple Choice Learning (sMCL)**. Our formulation is applicable to any model trained with stochastic gradient descent, is agnostic to the form of the task dependent loss, is parameter-free, and is time efficient, training all ensemble members concurrently.

We demonstrate the broad applicability and efficacy of sMCL for training diverse deep ensembles with *interpretable emergent expertise* on a wide range of problem domains and network architectures, including Convolutional Neural Network (CNN) [1] ensembles for image classification [17], Fully-Convolutional Network (FCN) [20] ensembles for semantic segmentation [6], and combined CNN and Recurrent Neural Network (RNN) ensembles [14] for image captioning [18]. We provide detailed analysis of the training and output behaviors of the resulting ensembles, demonstrating how ensemble member specialization and expertise *emerge* automatically when trained using sMCL. Our method outperforms existing baselines and produces sets of outputs with high oracle performance.

2 Related Work

Ensemble Learning. Much of the existing work on training ensembles focuses on diversity between member models as a means to improve performance by decreasing error correlation. This is often accomplished by resampling existing training data for each member model [27] or by producing artificial data that encourages new models to be decorrelated with the existing ensemble [21]. Other approaches train or combine ensemble members under a joint loss [19, 26]. More recently, work of Hinton et al. [12] and Ahmed et al. [2] explores using 'generalist' network performance statistics to inform the design of ensemble-of-expert architectures for classification. In contrast, sMCL *discovers* specialization as a consequence of minimizing oracle loss. Importantly, most existing methods do not generalize to structured output labels, while sMCL seamlessly adapts, discovering different task-dependent specializations automatically.

Generating Multiple Solutions. There is a large body of work on the topic of extracting multiple diverse solutions from a single model [3, 15, 16, 23, 24]; however, these approaches are designed for probabilistic structured-output models and are not directly applicable to general deep architectures.

Most related to our approach is the work of Guzman-Rivera et al. [8, 9] which explicitly minimizes oracle loss over the outputs of an ensemble, formalizing this setting as the Multiple Choice Learning (MCL) paradigm. They introduce a general alternating block coordinate descent training approach which requires retraining models multiple times. More recently, Dey et al. [5] reformulated this problem as a submodular optimization task in which ensemble members are learned sequentially in a boosting-like manner to maximize marginal gain in oracle performance. Both these methods require either costly retraining or sequential training, making them poorly suited to modern deep architectures that can take weeks to train. To address this serious shortcoming and to provide the first practical algorithm for training diverse deep ensembles, we introduce a stochastic gradient descent (SGD) based algorithm to train ensemble members concurrently.

3 Multiple-Choice Learning as Stochastic Block Gradient Descent

We consider the task of training an ensemble of differentiable learners that together produce a set of solutions with minimal loss with respect to an oracle that selects only the lowest-error prediction.

Notation. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. Given a training set of input-output pairs $D = \{(x_i, y_i) \,|\, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, our goal is to learn a function $g : \mathcal{X} \to \mathcal{Y}^M$ which maps each input to M outputs. We fix the form of g to be an ensemble of M learners f such that $g(x) = (f_1(x), \ldots, f_M(x))$. For some task-dependent loss $\ell(y, \hat{y})$, which measures the error between true and predicted outputs y and $\hat{y}$, we define the oracle loss of g over the dataset D as

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

Minimizing Oracle Loss with Multiple Choice Learning. In order to directly minimize the oracle loss for an ensemble of learners, Guzman-Rivera et al. [8] present an objective which forms a (potentially tight) upper-bound. This objective replaces the $\min$ in the oracle loss with indicator variables $(p_{i,m})_{m=1}^{M}$ where $p_{i,m}$ is 1 if predictor m has the lowest error on example i,

$$\begin{aligned} \underset{f_m, p_{m,i}}{\operatorname{argmin}} \quad & \sum_{i=1}^{n} \sum_{m=1}^{M} p_{i,m}\, \ell\left(y_i, f_m(x_i)\right) \\ s.t. \quad & \sum^{M} p_{i,m} = 1, \quad p_{i,m} \in \{0, 1\}. \end{aligned} \tag{1}$$

The resulting minimization is a constrained joint optimization over ensemble parameters and data-point assignments. The authors propose an alternating block algorithm, shown in Algorithm 1, to approximately minimize this objective. Similar to K-Means or 'hard-EM,' this approach alternates between assigning examples to their min-loss predictors and training models to convergence on the partition of examples assigned to them. Note that this approach is not feasible with training deep networks, since modern architectures [11] can take weeks or months to train a single model once.

Stochastic Multiple Choice Learning. To overcome this shortcoming, we propose a stochastic algorithm for differentiable learners which interleaves the assignment step with batch updates in

Input: Dataset $D = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, loss $\ell(y, \hat{y})$, and randomly initialized differentiable learners $f_1, \ldots, f_M$ parameterized by $\theta_1, \ldots, \theta_M$

Output: Ensemble of M trained predictors $f_1, \cdots f_M$

for $t \leftarrow 1$ *to convergence* **do**

Algorithm 1: MCL

for $m \leftarrow 1$ *to* M **do**
  `//Train models to convergence`
  $f_m \leftarrow \text{argmin}_{f_m} \sum_{(x_i, y_i) \in P_m} \ell(y_i, f_m(x_i))$
for $m \leftarrow 1$ *to* M **do**
  `//Reassign examples to min-loss predictor`
  $P_m \leftarrow \{(x_i, y_i) | f_m = \text{argmin}_{f_1, \ldots, f_M} \ell(x_i, f_j(x_i))\}$

Algorithm 2: sMCL

Sample batch $B \subset D$
for $m \leftarrow 1$ *to* M **do**
  `//Compute outputs for batch`
  `//for each model (forward-prop)`
  $y_{m,1}, \ldots, y_{m,B} \leftarrow f_m(B)$
for $i \leftarrow 1$ *to* $|B|$ **do**
  `//Select lowest error model per example`
  $m^* \leftarrow \text{argmin}_{j \in [1, \ldots, M]} \ell(y_i, y_{m,i})$
  `//and update only that model (back-prop)`
  $\theta_{m^*} = \theta_{m^*} - \lambda \partial \ell_{m^*,i} / \partial \theta_{m^*}$

return $f_1, \ldots f_M$

Figure 2: The MCL approach of [8] (Alg. 1) requires costly retraining while our sMCL method (Alg. 2) works within standard SGD solvers, training all ensemble members under a joint loss.

stochastic gradient descent. Consider the partial derivative of the objective in Eq. 1 with respect to the output of the m^{th} individual learner on example x_i,

$$\frac{\partial \mathcal{L}_O}{\partial f_m(x_i)} = p_{i,m} \frac{\partial \ell(y_i, f_m(x_i))}{\partial f_m(x_i)}. \tag{2}$$

Notice that if f_m is the minimum error predictor for example x_i, then $p_{i,m} = 1$, and the gradient term is the same as if training a single model; otherwise, the gradient is zero. This behavior lends itself to a straightforward optimization strategy for learners trained by SGD based solvers. For each batch, we pass the examples through the learners, calculating losses from each ensemble member for each example. During the backward pass, the gradient of the loss for each example is backpropagated only to the lowest error predictor on that example (with ties broken arbitrarily).

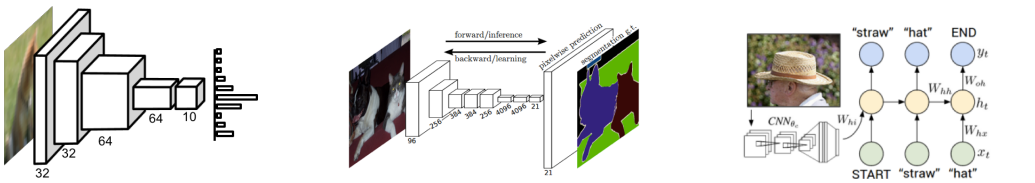
This approach, which we call Stochastic Multiple Choice Learning (sMCL), is shown in Algorithm 2. sMCL is generalizable to any learner trained by stochastic gradient descent and is thus applicable to an extensive range of modern deep networks. Unlike the iterative training schedule of MCL, sMCL ensembles need only be trained to convergence once in parallel. sMCL is also agnostic to the exact form of loss function ℓ such that it can be applied without additional effort on a variety of problems.

4 Experiments

In this section, we present results for sMCL ensembles trained for the tasks and deep architectures shown in Figure 3. These include CNN ensembles for image classification, FCN ensembles for semantic segmentation, and a CNN+RNN ensembles for image caption generation.

Baselines. Many existing general techniques for inducing diversity are not directly applicable to deep networks. We compare our proposed method against:

- **Classical** ensembles in which each model is trained under an independent loss with differing random initializations. We will refer to these as **Indp.** ensembles in figures.
- **MCL [8]** that alternates between training models to convergence on assigned examples and allocating examples to their lowest error model. We repeat this process for 5 meta-iterations and initialize ensembles with (different) random weights. We find MCL performs similarly to sMCL on small classification tasks; however, MCL performance drops substantially on segmentation and captioning tasks. Unlike sMCL which can effectively reassign an example once per epoch, MCL only does this after convergence, limiting its capacity to specialize compared to sMCL. We also note that sMCL is 5x faster than MCL, where the factor 5 is the result of choosing 5 meta-iterations (other applications may require more, further increasing the gap.)
- **Dey et al. [5]** train models sequentially in a boosting-like fashion, each time reweighting examples to maximize marginal increase of the evaluation metric. We find these models saturate quickly as the ensemble size grows. As performance increases, the marginal gain and therefore the weights approach zero. With low weights, the average gradient backpropagated for stochastic learners drops substantially, reducing the rate and effectiveness of learning without careful tuning. To compute



(a) Convolutional classification model of [1] for CIFAR10 [17]

(b) Fully-convolutional segmentation model of Long et al. [20]

(c) CNN+RNN based captioning model of Karpathy et al. [14]

Figure 3: We experiment with three problem domains using the various architectures shown above.

weights, [5] requires an error measure bounded above by 1: accuracy (for classification) and IoU (for segmentation) satisfy this; the CIDEr-D score [28] divided by 10 guarantees this for captioning.

Oracle Evaluation. We present results as oracle versions of the task-dependent performance metrics. These oracle metrics report the highest score over all outputs for a given input. For example, in classification tasks, oracle accuracy is exactly the top-k criteria of ImageNet [25], *i.e.* whether at least one of the outputs is the correct label. Likewise, the oracle intersection over union (IoU) is the highest IoU between the ground truth segmentation and any one of the outputs. Oracle metrics allow the evaluation of multiple-prediction systems separately from downstream re-ranking or selection systems, and have been extensively used in previous work [3, 5, 8, 9, 15, 16, 23, 24].

Our experiments convincingly demonstrate the broad applicability and efficacy of sMCL for training diverse deep ensembles. In all three experiments, sMCL significantly outperforms classical ensembles, Dey et al. [5] (typical improvements of 6-10%), and MCL (while providing a 5x speedup over MCL). Our analysis shows that the exact same algorithm (sMCL) leads to the automatic emergence of different interpretable notions of specializations among ensemble members.

4.1 Image Classification

Model. We begin our experiments with sMCL on the CIFAR10 [17] dataset using the small convolutional neural network *"CIFAR10-Quick"* provided with the Caffe deep learning framework [13]. CIFAR10 is a ten way classification task with small 32×32 images. For these experiments, the reference model is trained using a batch size of 350 for 5,000 iterations with a momentum of 0.9, weight decay of 0.004, and an initial learning rate of 0.001 which drops to 0.0001 after 4000 iterations.

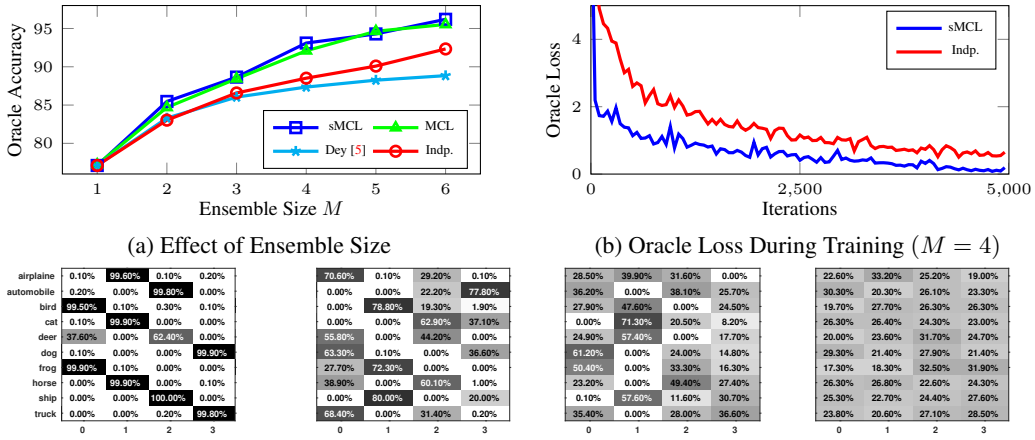
Results. Oracle accuracy for sMCL and baseline ensembles of size 1 to 6 are shown in Figure 4a. The sMCL trained ensembles result in higher oracle accuracy than the baseline methods, and are comparable to MCL while being 5x faster. The method of Dey et al. [5] performs worse than independent ensembles as ensemble size grows. Figure 4b shows the oracle loss during training for sMCL and regular ensembles. The sMCL trained models optimize for the oracle cross-entropy loss directly, not only arriving at lower loss solutions but also reducing error more quickly.

Interpretable Expertise: sMCL Induces Label-Space Clustering. Figure 4c shows the class-wise distribution of the assignment of test datapoints to the oracle or 'winning' predictor for an $M = 4$ sMCL ensemble. The level of class division is *striking* – most predictors *become specialists for certain classes*. Note that these divisions emerge from training under the oracle loss and are not hand-designed or pre-initialized in any way. In contrast, Figure 4f show that the oracle assignments for a standard ensemble are nearly uniform. To explore the space between these two extremes, we loosen the constraints of Eq. 1 such that the lowest k error predictors are penalized. By varying k between 1 and the number of ensemble members M, the models transition from minimizing oracle loss at $k = 1$ to a traditional ensemble at $k = M$. Figures 4d and 4e show these results. We find a direct correlation between the degree of specialization and oracle accuracy, with $k = 1$ netting highest oracle accuracy.

4.2 Semantic Segmentation

We now present our results for the semantic segmentation task on the Pascal VOC dataset [6].

Model. We use the fully convolutional network (FCN) architecture presented by Long et al. [20] as our base model. Like [20], we train on the Pascal VOC 2011 training set augmented with extra segmentations provided in [10] and we test on a subset of the VOC 2011 validation set. We initialize



(a) Effect of Ensemble Size (b) Oracle Loss During Training ($M = 4$)

(c) k=1 (d) k=2 (e) k=3 (f) k=M=4

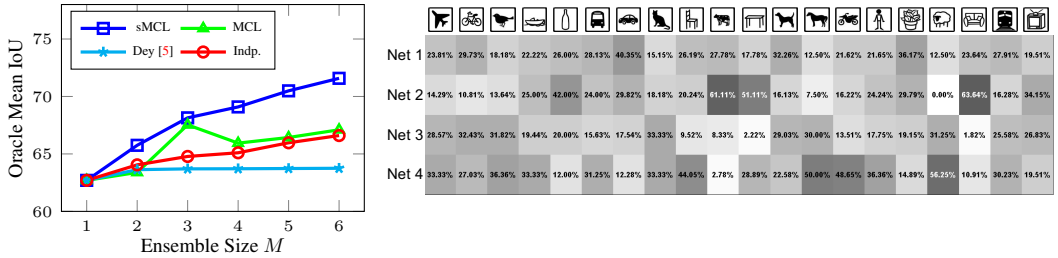
Figure 4: sMCL trained ensembles produce higher oracle accuracies than baselines (a) by directly optimizing the oracle loss (b). By varying the number of predictors k each example can be assigned to, we can interpolate between sMCL and standard ensembles, and (c-f) show the percentage of test examples of each class assigned to each ensemble member by the oracle for various k. These divisions are not preselected and show how specialization is an emergent property of sMCL training.

our sMCL models from a standard ensemble trained for 50 epochs at a learning rate of 10^{-3}. The sMCL ensemble is then fine-tuned for another 15 epochs at a reduced learning rate of 10^{-5}.

Results. Figure 5a shows oracle accuracy (class-averaged IoU) for all methods with ensemble sizes ranging from 1 to 6. Again, sMCL significantly outperforms all baselines (~7% relative improvement over classical ensembles). In this more complex setting, we see the method of Dey et al. [5] saturates more quickly – resulting in performance worse than classical ensembles as ensemble size grows. Though we expect MCL to achieve similar results as sMCL, retraining the MCL ensembles a sufficient number of times proved infeasible so results after five meta-iterations are shown.

Interpretable Expertise: sMCL as Segmentation Specialists. In Figure 5b, we analyze the class distribution of the predictions using an sMCL ensemble with 4 members. For each test sample, the oracle picks the prediction which corresponds to the ensemble member with the highest accuracy for that sample. We find the specialization with respect to classes is much less evident than in the classification experiments. As segmentation presents challenges other than simply selecting the correct class, specialization can occur in terms of shape and frequency of predicted segments in addition to class divisions; however, we do still see some class biases – network 2 captures cows, tables, and sofas well and network 4 has become an expert on sheep and horses.

Figure 6 shows qualitative results from a four member sMCL ensemble. We can clearly observe the diversity in the segmentations predicted by different members. In the first row, we see the majority of the ensemble members produce dining tables of various completeness in response to the visual uncertainty caused by the clutter. Networks 2 and 3 capture this ambiguity well, producing segmentations with the dining table completely present or absent. Row 2 demonstrates the capacity of sMCL ensembles to provide multiple high quality solutions. The models are confused whether the



(a) Effect of Ensemble Size (b) Oracle Assignment Distributions by Class

Figure 5: a) sMCL trained ensembles consistently result in improved oracle mean IoU over baselines on PASCAL VOC 2011. b) Distribution of examples from each category assigned by the oracle for an sMCL ensemble.

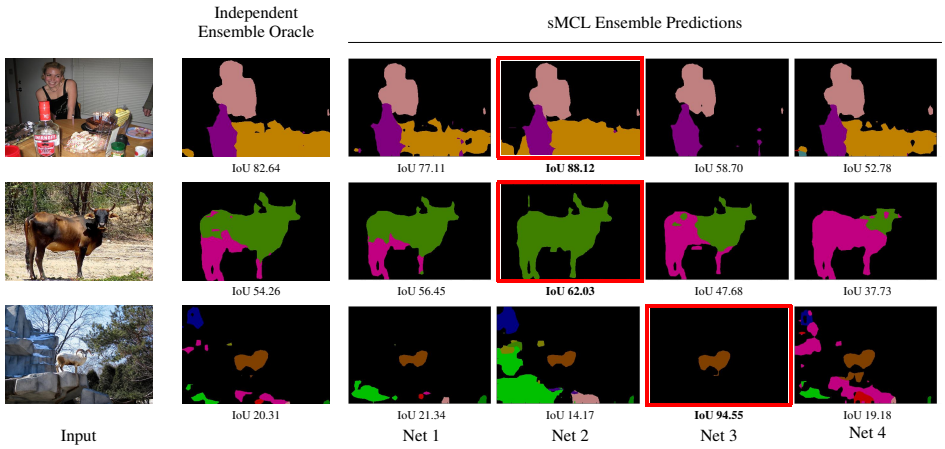


Figure 6: Samples images and corresponding predictions obtained by each member of the sMCL ensemble as well as the top output of a classical ensemble. The output with minimum loss on each example is outlined in red. Notice that sMCL ensembles vary in the shape, class, and frequency of predicted segments.

animal is a horse or a cow – models 1 and 3 produce typical 'safe' responses while models 2 and 4 attempt to give cohesive responses. Finally, row 3 shows how the models can learn biases about the frequency of segments with model 3 presenting only the sheep.

4.3 Image Captioning

In this section, we show that sMCL trained ensembles can produce sets of high quality and diverse sentences, which is essential to improving recall and capturing ambiguities in language and perception.

Model. We adopt the model and training procedure of Karpathy et al. [14], utilizing their publicly available implementation *neuraltalk2*. The model consists of an VGG16 network [4] which encodes the input image as a fixed-length representation for a Long Short-Term Memory (LSTM) language model. We train and test on the MSCOCO dataset [18], using the same splits as [14]. We perform two experimental setups by either freezing or finetuning the CNN. In the first, we freeze the parameters of the CNN and train multiple LSTM models using the CNN as a static feature generator. In the second, we aggregate and back-propagate the gradients from each LSTM model through the CNN in a tree-like model structure. This is largely a construct of memory restrictions as our hardware could not accommodate multiple VGG16 networks. We train each ensemble for 70k iterations with the parameters of the CNN fixed. For the fine-tuning experiments, we perform another 70k iterations of training to fine-tune the CNN. We generate sentences for testing by performing beam search with a beam width of two (following [14]).

Results. Table 1 presents the oracle CIDEr-D [28] scores for all methods on the validation set. We additionally compare with all outputs of a beam search over a single CNN+LSTM model with beam width ranging from 1 to 5. sMCL significantly outperforms the baseline ensemble learning methods (shown in the upper section of the table), increasing both oracle performance and the number of unique n-grams. For $M = 5$, beam search from a single model achieves greater oracle but produces

	Oracle CIDEr-D for Ensemble of Size					# Unique n-Grams (M=5)				
	M = 1	2	3	4	5	n = 1	2	3	4	Avg. Length
sMCL	-	0.822	0.862	0.911	0.922	713	2902	6464	15427	10.21
MCL [8]	-	0.752	0.81	0.823	0.852	384	1565	3586	9551	9.87
Dey [5]	-	0.798	0.850	0.887	0.910	584	2266	4969	12208	10.26
Indp.	0.684	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
Indp. (fine-tuned CNN)	0.912	1.001	1.05	1.073	1.095	921	4335	10534	23811	10.33
Beam Search	0.654	0.754	0.833	0.888	0.943	580	2272	4920	12920	10.62

Table 1: sMCL base methods outperform other ensemble methods a captioning, improve both oracle performance and the number of distinct n-grams. For low M, sMCL also performs better than multiple-output decoders.

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

Figure 7: Comparison of sentences generated by members of a standard independently trained ensemble and an sMCL based ensemble of size four.

significantly fewer unique n-grams. We note that beam search is an inference method and increased beam width could provide similar benefits for sMCL ensembles.

Intepretable Expertise: sMCL as N-Gram Specialists. Figure 7 shows example images and generated captions from standard and sMCL ensembles of size four (results from beam search over a single model are similar). It is evident that the independently trained models tend to predict similar sentences independent of initialization, perhaps owing to the highly structured nature of the output space and the mode bias of the underlying language model. On the other hand, the sMCL based ensemble generates diverse sentences which capture ambiguity both in language and perception. The first row shows an extreme case in which all of the members of the standard ensemble predict identical sentences. In contrast, the sMCL ensemble produces sentences that describe the scene with many different structures. In row three, both models are confused about the content of the image, mistaking the pile of suitcases as kitchen appliances. However, the sMCL ensemble widens the scope of some sentences to include the cat clearly depicted in the image. The fourth row is an example of regression towards the mode, with the standard model producing multiple similar sentences describing birds on branches. In the sMCL ensemble, we also see this tendency; however, one model breaks away and captures the true content of the image.

5 Conclusion

To summarize, we propose Stochastic Multiple Choice Learning (sMCL), an SGD-based technique for training diverse deep ensembles that follows a 'winner-take-gradient' training strategy. Our experiments demonstrate the broad applicability and efficacy of sMCL for training diverse deep ensembles. In all experimental settings, sMCL significantly outperforms classical ensembles and other strong baselines including the 5x slower MCL procedure. Our analysis shows that exactly the same algorithm (sMCL) automatically generates specializations among ensemble members along different task-specific dimensions. sMCL is simple to implement, agnostic to both architecture and loss function, parameter free, and simply involves introducing one new sMCL layer into existing ensemble architectures.

Acknowledgments

This work was supported in part by a National Science Foundation CAREER award, an Army Research Office YIP award, ICTAS Junior Faculty award, Office of Naval Research grant N00014-14-1-0679, Google Faculty Research award, AWS in Education Research grant, and NVIDIA GPU donation, all awarded to DB, and by an NSF CAREER award (IIS-1253549), the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract FA8650-12-C-7212, a Google Faculty Research award, and an NVIDIA GPU donation, all awarded to DC. Computing resources used by this work are supported in part by NSF (ACI-0910812 and CNS-0521433), the Lily Endowment, Inc., and the Indiana METACyt Initiative. The U.S. Government is authorized to reproduce and distribute

reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, NSF, or the U.S. Government.

References

[1] CIFAR-10 Quick Network Tutorial. http://caffe.berkeleyvision.org/gathered/examples/cifar10.html, 2016.

[2] K. Ahmed, M. H. Baig, and L. Torresani. Network of experts for large-scale image categorization. In *arXiv preprint arXiv:1604.06119*, 2016.

[3] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[5] D. Dey, V. Ramakrishna, M. Hebert, and J. Andrew Bagnell. Predicting multiple structured visual interpretations. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[8] A. Guzman-Rivera, D. Batra, and P. Kohli. Multiple Choice Learning: Learning to Produce Multiple Structured Outputs. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[9] A. Guzman-Rivera, P. Kohli, D. Batra, and R. Rutenbar. Efficiently enforcing diversity in multi-output structured prediction. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[10] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[12] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NIPS) - Deep Learning Workshop*, 2014.

[13] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[14] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] A. Kirillov, B. Savchynskyy, D. Schlesinger, D. Vetrov, and C. Rother. Inferring m-best diverse solutions in a single one. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[16] A. Kirillov, D. Schlesinger, D. Vetrov, C. Rother, and B. Savchynskyy. M-best-diverse labelings for submodular energies and beyond. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[17] A. Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context, 2014.

[19] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] P. Melville and R. J. Mooney. Creating diversity in ensembles using artificial data. *Information Fusion*, 6(1):99–111, 2005.

[22] Microsoft. Decades of computer vision research, one 'Swiss Army knife'. blogs.microsoft.com /next/2016/03/30/decades-of-computer-vision-research-one-swiss-army-knife/, 2016.

[23] D. Park and D. Ramanan. N-best maximal decoders for part models. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 2627–2634, 2011.

[24] A. Prasad, S. Jegelka, and D. Batra. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[25] O. Russakovsky, J. Deng, J. Krause, A. Berg, and L. Fei-Fei. The ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). http://www.image-net.org/challenges/LSVRC/2012/.

[26] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[27] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, 1996.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[29] WIRED. Facebook's AI Can Caption Photos for the Blind on Its Own. wired.com/2015/10/facebook-artificial-intelligence-describes-photo-captions-for-blind-people/, 2015.