Diverse Ensemble Evolution: Curriculum based Data-Model Marriage

Tianyi Zhou, Shengjie Wang, Jeff A. Bilmes, University *of* Washington, Seattle

- **Background:** An ensemble of light-weight models can outperform a large model, and can be trained at multiple edge devices with limited computational power.
- **Problem:** How to train an ensemble of models jointly, so they all have different and diverse expertise, but are mutually complementary?
- **Basic Idea:** A teacher adaptively and jointly assigns subsets of training samples to multiple learners at each epoch of their training processes.
- **Previous works on ensemble training:** only encourages diversity before training begins, is not adaptive, and always trains each model on the whole training set.
- **Our Method (DivE²):** can evolve diverse and complementary expertise on different models faster, achieve better performance using less training time.
- **Applications:** distributed/federated machine learning; edge computing.

Combinatorial-Continuous Hybrid Optimization

Reward | Inter-model diversity | Intra-model diversity

$$\max_{W} \max_{A\subseteq E, A\in\mathcal{I}_v\cap\mathcal{I}_u} G(A,W) \triangleq \sum_{(v_j,u_i)\in A} (\beta - \ell(v_j;w_i)) + \gamma F_{inter}(A) + \lambda F_{intra}(A)$$

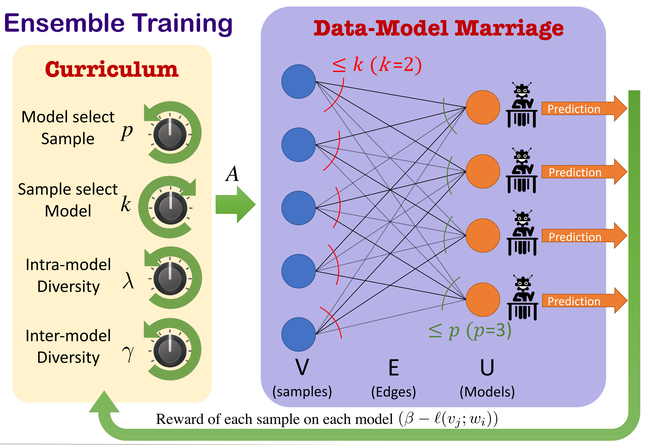
$\mathcal{I}_u = \{A | A\subseteq E, |A\cap\delta(u)| \le p, \forall u\in U\}$ Model Selecting (at most p) Samples

$\mathcal{I}_v = \{A | A\subseteq E, |A\cap\delta(v)| \le k, \forall v\in V\}$ Sample Selecting (at most k) Models

$F_{inter}(A) \triangleq \sum_{i,j\in[m],i<j} F(\delta(u_i)\cup\delta(u_j)\cap A)$ Inter-model Diversity (Submodular)

$F_{intra}(A) = \sum_{i\in[m]} F'(\delta(u_i)\cap A)$ Intra-model Diversity (Submodular)

Ensemble Training

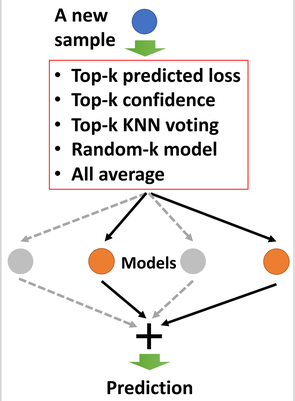


Algorithm 1 SELECTLEARN$(k, p, \lambda, \gamma, \{w_i^0\}_{i=1}^m)$

Input: $\{v_j\}_{j=1}^n$, $\{l(\cdot;w_i)\}_{i=1}^m$, $\pi(\cdot;\eta)$
Output: $\{w_i\}_{i=1}^m$
Initialize: $w_i \leftarrow w_i^0\ \forall i\in[m], t=0$
while not "converged" **do**
  $W \leftarrow \{w_i^t\}_{i=1}^m$, define $G(\cdot,W)$ by W;
  $A \leftarrow$ SUBMODULARMAX$(G(\cdot,W),k,p)$;
  if $G(\hat{A},W) > G(A,W)$ **then**
    $A \leftarrow \hat{A}$;
  end if
  for $i\in\{1,\cdots,m\}$ **do**
    $-\nabla\hat{H}(w_i^t) \leftarrow \frac{\partial}{\partial w_i^t}\sum_{v_j\in V(A\cap\delta(u_i))}\ell(v_j;w_i^t)$;
    $w_i^{t+1} \leftarrow w_i^t + \pi\left(\{w_i^\tau, -\nabla\hat{H}(w_i^\tau)\}_{\tau\in[1,t]};\eta^t\right)$;
  end for
  $t \leftarrow t+1$;
end while

1. *If $k < {}^{mp+p}/_{n+(p-1)}$, $\mathcal{I}_v$ saturates, i.e., $|\hat{A}\cap\delta(v)| = k, \forall v\in V$, and $|\hat{A}| = nk$;*
2. *If $k > {}^{mp-p}/_{n-(p-1)}$, $\mathcal{I}_u$ saturates, i.e., $|\hat{A}\cap\delta(u)| = p, \forall u\in U$, and $|\hat{A}| = mp$;*
3. *If ${}^{mp+p}/_{n+(p-1)} \le k \le {}^{mp-p}/_{n-(p-1)}$, $|\hat{A}| \ge \min\{(k-1)+(m-k+1)p, (p-1)+(n-p+1)k\}$.*

Ensemble Inference



- **Curriculum for Data-Model Marriage**: gradually tune four knobs controlling the data assignment process (formulated as a generalized bipartite matching), which involves submodular maximization (two diversity terms) subject to two matroid independence constraints.
- **Model selecting Sample & Sample Selecting Model:** in earlier stages (epochs) of training, every model selects p samples that are easier to learn; in later stages, every sample selects k models which provide better prediction; use them together to avoid imbalance.
- **Inter-model & Intra-model Diversity**: intra-model diversity selects diverse training samples for each model; inter-model diversity encourages less overlap between training samples assigned to different models; we start from large diversity and gradually reduce it, so the learners improve their expertise and specialization by gradually focusing on the individualized data subset they are each good at.

Algorithm

Algorithm 2 Diverse Ensemble Evolution (DIVE²)

Input: $\{(x_j,y_j)\}_{j=1}^n, \{w_i^0\}_{i=1}^m, \pi(\cdot;\eta), \mu, \Delta_k, \Delta_p, T$
Output: $\{w_i^t\}_{i=1}^m$
Initialize: $k \le m, p \ge 1$ s.t. $mp \le nk$, $\lambda\in[0,1], \gamma\in[0,1]$
for $t\in\{1,\cdots,T\}$ **do**
  $\{w_i^t\}_{i=1}^m \leftarrow$ SELECTLEARN$(k,p,\lambda,\gamma,\{w_i^{t-1}\}_{i=1}^m)$;
  $\lambda \leftarrow (1-\mu)\cdot\lambda, \gamma \leftarrow (1-\mu)\cdot\gamma$;
  $k \leftarrow \max\{\lceil k-\Delta_k\rceil, 1\}, p \leftarrow \min\{\lfloor p+\Delta_p\rfloor, n\}$;
end for

The hybrid optimization equals to maximization of a piecewise function $H(W) \triangleq \max_{A\subseteq E, A\in\mathcal{I}_v\cap\mathcal{I}_u} G(A,W)$, with each piece defined by a fixed A achieving the maximum of $G(A,W)$ in a local region of W.

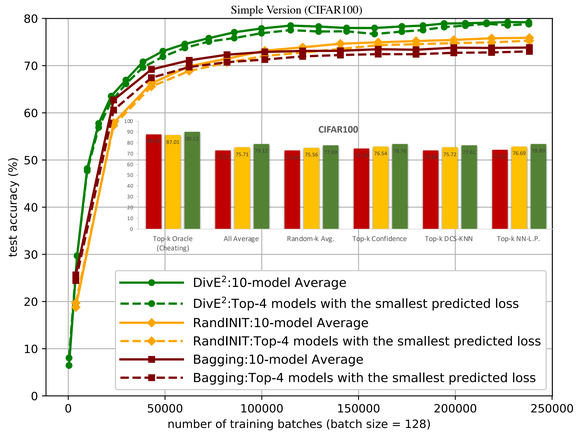
Proposition 1. *Algorithm 2:*

1. *generates a monotonically non-decreasing sequence of objective values $G(A;W)$ (assuming $\pi(\cdot;\eta)$ does the same);*
2. *converges to a stationary point on $\hat{H}(W)$; and (3) for any loss $\ell(u,w)$ that is β-strongly convex w.r.t. w, if* SUBMODULARMAX *has approximation factor α, it converges to a local optimum $\hat{W}\in\text{argmax}_{W\in\mathcal{K}}\hat{H}(W)$ (i.e., $\hat{W}$ is optimal in an local area $\mathcal{K}$) such that for any local optimum $W_{loc}^*\in\mathcal{K}$ on the true objective $H(W)$, we have*

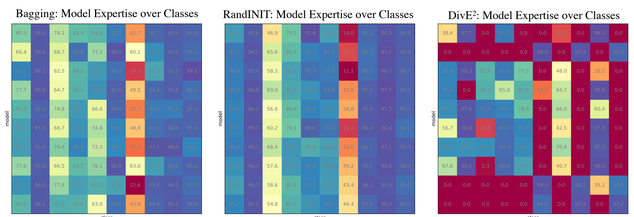
$$\hat{H}(\hat{W}) \ge \alpha H(W_{loc}^*) + \frac{\beta}{2}\cdot\min\{(k-1)+(m-k+1)p, (p-1)+(n-p+1)k\}\cdot\|\hat{W}-W_{loc}^*\|_2^2.$$

where $\alpha = {}^1/_{2+\kappa_G}$ for the greedy algorithm, where κ_G is the curvature of $G(\cdot,W)$. When the weights λ and γ are small, κ_G decreases and $G(\cdot,W)$ becomes more modular.

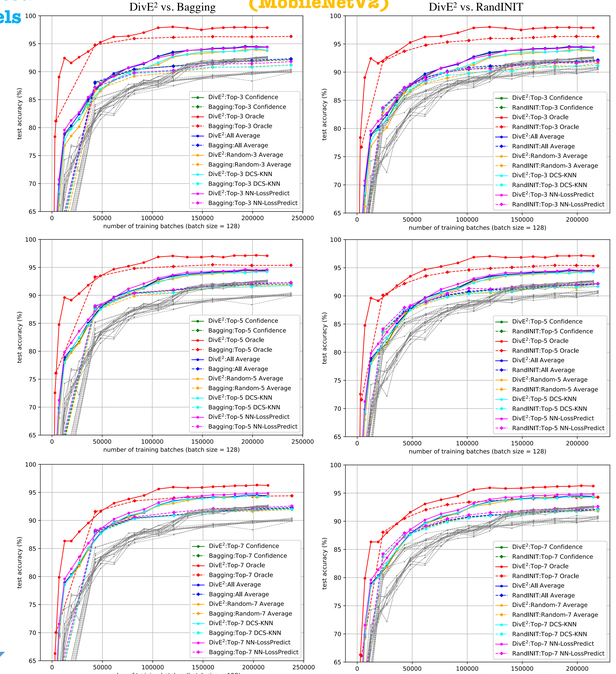
Experiments



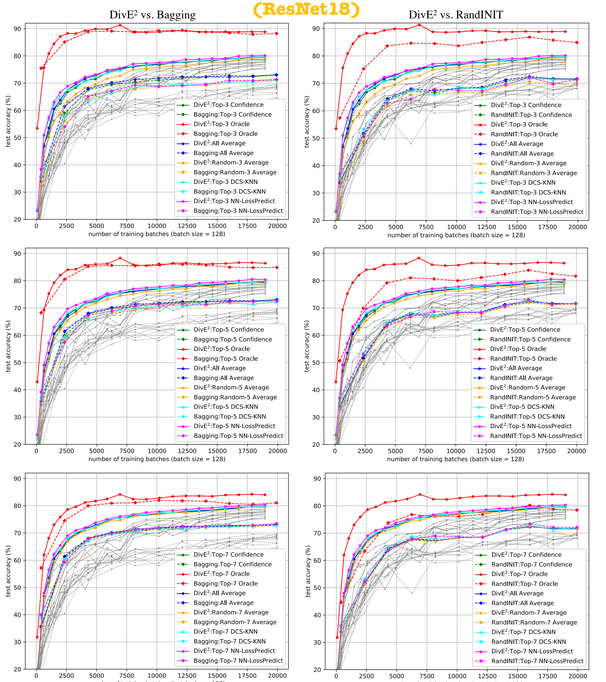
Increase k #Selected Models



CIFAR10 (MobileNetV2)



CIFAR100 (ResNet18)



STL10 (6-layer CNN)