Diverse Ensemble Evolution: Curriculum based Data-Model Marriage

Tianyi Zhou, Shengjie Wang, Jeff 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 complementary to each other.
- **Basic Idea:** teacher adaptively assigns training samples to multiple learners for each epoch of their training processes.
- **Previous works on ensemble training:** only encourage diversity before training begins, not adaptive, always train each model on the whole training set.
- **Our Method (DivE2):** can evolve diverse and complementary expertise on different models faster, achieve better performance with less training time.
- **Applications:** distributed/federated machine learning; edge computing.

Combinatorial-Continuous Hybrid Optimization

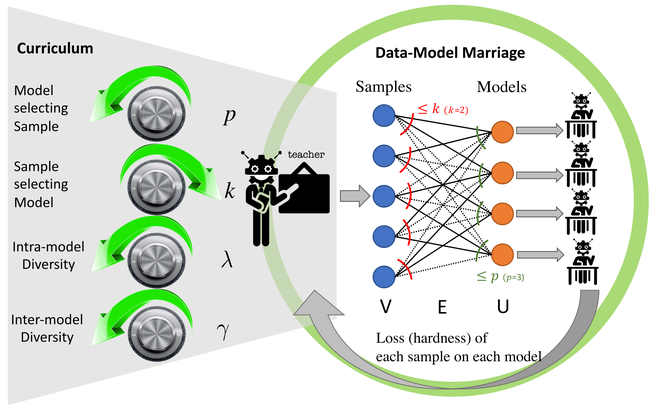
$$\max_{W} \max_{A\subseteq E, A\in\mathcal{I}_v\cap\mathcal{I}_u} G(A,W) \triangleq \sum_{(v_j,u_i)\in A} \underbrace{(\beta-\ell(v_j;w_i))}_{\text{Reward}} + \underbrace{\gamma F_{inter}(A)}_{\text{Inter-model diversity}} + \underbrace{\lambda F_{intra}(A)}_{\text{Intra-model diversity}}$$

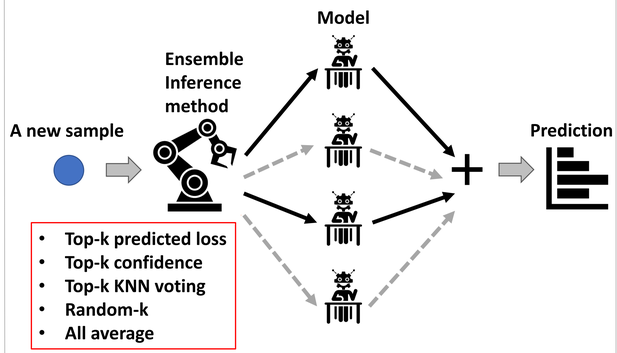
$\mathcal{I}_u = \{A | A\subseteq E, |A\cap\delta(u)| \le p, \forall u\in U\}$ Model Selecting (at most p) Sample

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

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

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





- **Curriculum for Data-Model Marriage**: gradually tune four knobs controlling the data assignment process (formulated as a bipartite graph matching), which includes two matroid constraints and two diversity terms.
- **Model selecting Sample & Sample Selecting Model:** in earlier stages, every model selects p samples 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 expand their expertise regions and gradually turn focus on data they are good at.

Algorithm

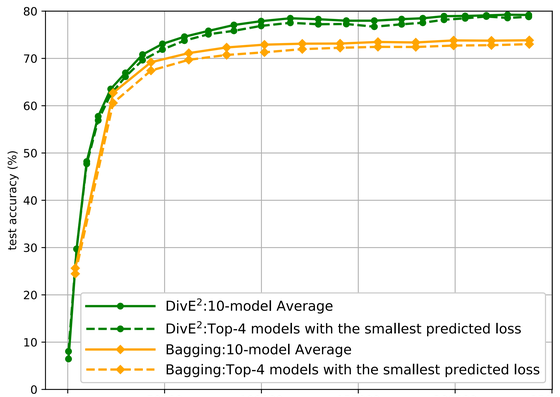
Algorithm 1 SELECTLEARN$(k, p, \lambda, \gamma, \{w_i^0\}_{i=1}^m)$

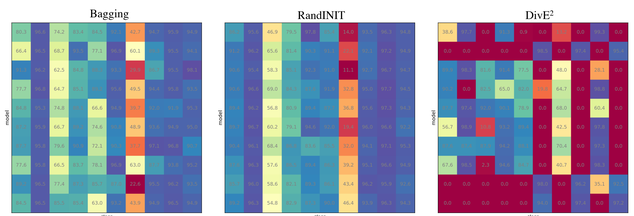
Input: $\{v_j\}_{j=1}^n, \{l(\cdot;w_i)\}_{i=1}^m, \pi(\cdot;\eta)$
Output: $\{w_i\}_{i=1}^m$
Initialize: $w_i \leftarrow w_i^0 \ \ \forall i\in[m], t=0$
while *not "converged"* **do**
  $W \leftarrow \{w_i^t\}_{i=1}^m$, define $G(\cdot,W)$ by W;
  $\hat{A} \leftarrow$ SUBMODULARMAX$(G(\cdot,W), k, p)$;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

Algorithm 2 Diverse Ensemble Evolution (DIVE^2)

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

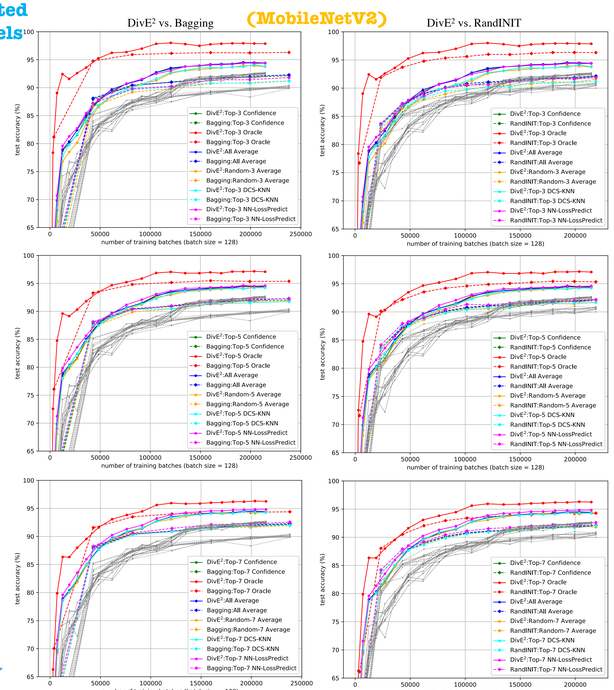
Experiments



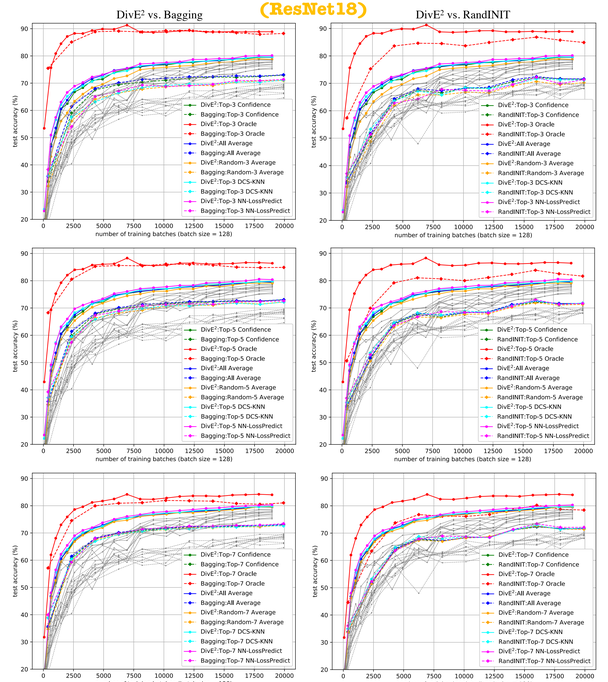


Increase Selected Models

CIFAR10 (MobileNetV2)



CIFAR100 (ResNet18)



STL10 (6-layer CNN)