Wisdom of the ensemble: Improving consistency of deep learning models

Lijing Wang[1] , Dipanjan Ghosh[2], Maria Teresa Gonzalez Diaz[2], Ahmed Farahat[2], Mahbubul Alam[2], Chetan Gupta[2], Jiangzhuo Chen[1], Madhav Marathe[1]

[1] University of Virginia, [2] Industrial AI Lab, Hitachi America, Ltd.

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Trust is often a function of constant behavior. From an AI model perspective it means given the same input the user would expect the same output, especially for correct outputs, or in other words consistently correct outputs. This paper studies a model behavior in the context of periodic retraining of deployed models where the outputs from successive generations of the models might not agree on the correct labels assigned to the same input. We formally define ***consistency*** and ***correct-consistency*** of a learning model. We prove that consistency and correct-consistency of an ensemble learner is not less than the average consistency and correct-consistency of individual learners and correct-consistency can be improved with a probability by combining learners with accuracy not less than the average accuracy of ensemble component learners. To validate the theory using three datasets and two state-of- the-art deep learning classifiers we also propose an efficient dynamic snapshot ensemble method and demonstrate its value.

Definition

Consistency is defined as the ability of a model to reproduce an output for the same input across model generations.

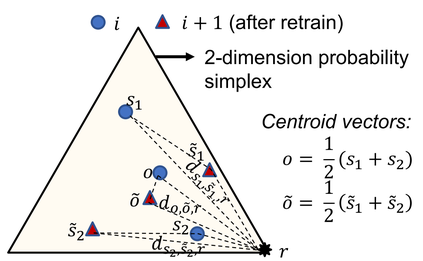
Correct-consistency is defined as the ability of a model to reproduce a correct output for the same input across model generations.

Model$_{(\text{Generation } i)}$	Model$_{(\text{Generation } i+1)}$	Definition and Impact on Users
Correct	Correct	Consistency, Correct-Consistency (High Impact)
Incorrect	Correct	Accuracy, Inconsistent (High Impact)
Incorrect	Incorrect	Consistency (Less Impact)
Correct	Incorrect	Inconsistency (High Impact)

Theory Background

Assumption: A supervised classification problem has p class labels.

Example p = 3



Centroid vectors:

$$o = \frac{1}{2}(s_1 + s_2)$$

$$\tilde{o} = \frac{1}{2}(\tilde{s}_1 + \tilde{s}_2)$$

r : ground truth vector
$o, \tilde{o}$: prediction vectors from two generations of the ensemble
$s_i, \tilde{s}_i$: prediction vectors from two generations of a component
$d_{o,\tilde{o}}$: Euclidean distance of two vectors (consistency)
$d_{o,\tilde{o},r} = d_{o,\tilde{o}} + d_{o,r} + d_{\tilde{o},r}$ (correct-consistency)

Why ensemble?

Based on Minkowski's inequality for sums, we prove that:

Consistency of ensemble

$$d_{o,\tilde{o}} \le \frac{1}{2}(d_{s_1,\tilde{s}_1} + d_{s_2,\tilde{s}_2})$$

Average consistency of components

Correct-consistency of ensemble

$$d_{o,\tilde{o},r} \le \frac{1}{2}(d_{s_1,\tilde{s}_1,r} + d_{s_2,\tilde{s}_2,r})$$

Average correct-consistency of components

The inequality can be generalized to $p \ge 1$ class labels ($p = 1$ is regression problem) and Minkowski distance with order $q > 1$.

Theorem 1 For I_t, the distance … **Theorem 2** For I_t, let …
Theorem 3 For I_t, the sum … **Theorem 4** For I_t, let …
Theorem 5 For I, the aggregate correct-consistency … **Corollary 5.1** For I, let …

Theoretical Findings

- *According to Theorem 1 and 2:* The consistency of an ensemble learner is not less than the average consistency of individual learners.
- *According to Theorem 3 and 4:* The correct-consistency of an ensemble learner is not less than the average correct-consistency of individual learners.
- *According to Theorem 5 and Corollary 5.1:* A better aggregate correct-consistency performance of an ensemble can be achieved by combining components with accuracy that is higher than the average accuracy of the ensemble members with a quantifiable probability.

Algorithm: Dynamic snapshot ensemble

Snapshot ensemble

Save multiple single learners during one training process.

- *DynEns-cyc*: cyclic cosine learning rate schedule + cyclic snapshot
- *DynEns-step*: step-wise decay learning rate schedule + top-N snapshot

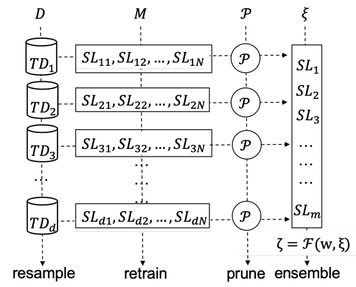
Dynamic pruning

Pruning criteria $\mathcal{P}$ For a single snapshot learning on a resampled training dataset, N single learners are snapshotted, $\xi_i = \{SL_1, \dots, SL_N\}$, with their validation accuracy, $w_i = \{W_{i1}, \dots, W_{iN}\}$. The pruning criteria $\mathcal{P}$ is defined as: SL_{ij} is included in the final ensemble ζ if W_{ij} is larger or equal to a threshold τ:

$$\tau = (1-\beta) * \max(w_i) + \beta * \min(w_i)$$

Based on the theorems, $\tau = \frac{1}{N}\sum w_i$ i.e. $\beta = \frac{\max(w_i) - \frac{1}{N}\sum w_i}{\max(w_i) - \min(w_i)}$, selects SL_{ij} that can lead to better correct-consistency of the ensemble than the correct-consistency of ξ_i, resulting into an ideal β for ξ_i obtained empirically.

Dynamic snapshot ensemble: Advantages



Algorithm Advantages

- *Data diversity*: Random shuffle of training and validation datasets.
- *Parameter diversity*: Random initialization of model parameters.
- *Metrics vs Computational Cost*: Better accuracy, consistency, correct-consistency without compromising computational cost.

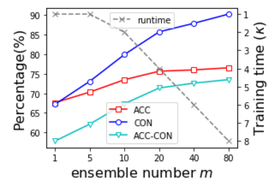
Empirical validation

Compared to SingleBase: All ensemble methods achieve improvements:

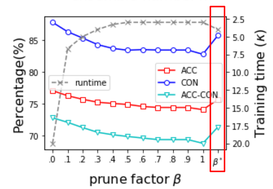
- ACC: 1.8%-3.3%, 5.4%-8.3%, and 0.7%-2%
- CON: 4.5%-8.2%, 9.8%-19.3%, and 3.4%-4.5%
- ACC-CON: 3.7%- 6.5%, 8.3%-14.1%, and 1.4%-2.3%

Compared between ensemble methods: DynSnap-cyc/DynSnap-step achieve comparable ACC, CON, ACC-CON performance with ExtBagging (best ACC) but with much smaller training time, and better performance than MCDropout and Snapshot (least training cost).

Sensitivity analysis: (CIFAR100 + ResNet56 + DynSnap-cyc + AVG)



Conform with theoretical findings that better consistency/correct-consistency can (but not guarantee) be achieved by combining more components.



The proposed pruning algorithm can achieve a good trade-off between model performance and training cost empirically.

‣ Single learner: SingleBase ‣ Ensemble learner: ExtBagging, MCDropout, Snapshot, DynSnap-cyc (ours), DynSnap-step (ours) ‣ Accuracy (ACC); Consistency (CON); Correct-Consistency (ACC-CON) ‣ Majority Voting (MV); Weighted Majority Voting (WMV); Averaging (AVG); Weighted Averaging (WAVG)

References

M.I. Voitsekhovskii (2001) [1994], "Minkowski inequality", Encyclopedia of Mathematics, EMS Press

Huang, G. et al. (2017). "Snapshot ensembles: Train 1, get m for free", arXiv preprint arXiv:1704.00109.