Coffee Talk #2

June 29, 2021

Ozgur Taylan Turan

The Deep Bootstrap Framework:Good Learners Are Good Offline Generalizers[1] Conference paper at 2021 ICLR

[1] P. Nakkiran, B. Neyshabur, and H. Sedghi (2020). “The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers”. In: 1. arXiv: 2010.08127

Why This Paper?

Interesting claims

Generalization $\leftrightarrow$ Optimization

Aim & Problem Setting

Aim

- Create a framework for investigating generalization in interpolating regime TrainError ≈ 0

$$\text{TestError} = \cancel{\text{TrainError}}^{\,0} + \underbrace{(\text{TestError} - \cancel{\text{TrainError}})}_{\text{Generalization gap}}{}^{0}$$

Problem Setting

- Supervised classification:

$\mathcal{D} \sim (\boldsymbol{x}, \boldsymbol{y})$ minimize training error with SGD variant on an architecture $\mathcal{F}$ for t steps with the hope of a classifier f_t with low testing error

Real World vs Ideal World-A

For a fixed $\mathcal{D}$ and $\mathcal{F}$:

Ideal World [$\text{Train}_{\mathcal{D},\mathcal{F}}(\infty, t)$]

- Access to $\mathcal{D}$
- Take t steps on mini-batches sampled from $\mathcal{D}$ to get f_t^{iid}

Real World [$\text{Train}_{\mathcal{D},\mathcal{F}}(n, t)$]

- Access to n samples from $\mathcal{D}$
- Take t steps on mini-batches of n samples to get f_t

$$\text{TestError}(f_t) = \text{TestError}(f_t^{\text{iid}}) + \underbrace{(\text{TestError}(f_t) - \text{TestError}(f_t^{\text{iid}}))}_{\text{Bootstrap error}(\varepsilon)}$$

Real vs Ideal-B

- So keeping everything same,

Ideal World $\rightarrow$ minimize population loss
Real World $\rightarrow$ minimize empirical loss

Claim: $\varepsilon(n, \mathcal{D}, \mathcal{F}, t)$ is small for all *realistic* $(n, \mathcal{D}, \mathcal{F})$ at all t

Experimental Setup

Datasets

- CIFAR-5m: generating 6M synthetic data 5M:train 1M:test

Training

- Train the real world optimizer until $\leq 1\%$ or reach specified epochs

Measure

- Soft-Error = 1 - softmax(correct label)

If we have 10 points for training and train for 2 epochs in Real world then, I have to get 20 unique samples and train for 1 epoch for the Ideal world

Results-A

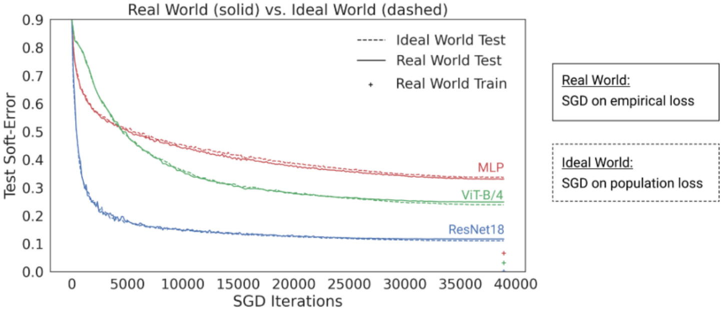
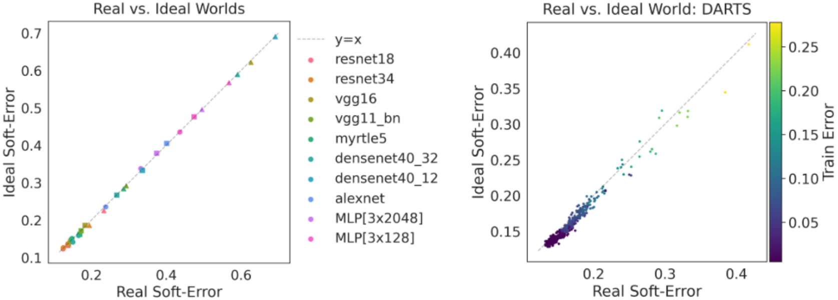


Figure 1: Three architectures trained from scratch on CIFAR-5m, a CIFAR-10-like task. The Real World is trained on 50K samples for 100 epochs, while the Ideal World is trained on 5M samples in 1 pass. The Real World Test remains close to Ideal World Test, despite a large generalization gap.

Claim: Generalization gap MLP $\geq$ CNN because CNN optimize faster in the Ideal world.

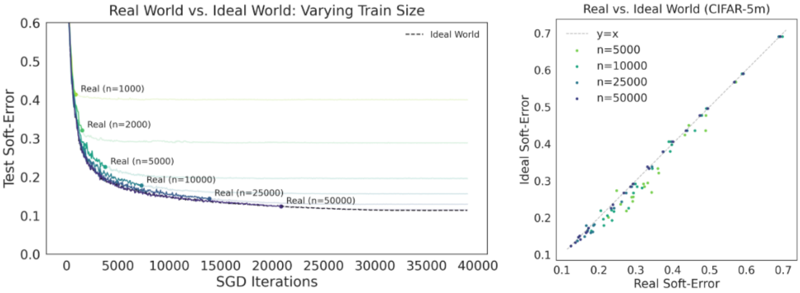
Results-B



(a) Standard architectures. (b) Random DARTS architectures.

Figure 2: **Real vs Ideal World: CIFAR-5m.** SGD with 50K samples. (a): Varying learning-rates 0.1 (●), 0.01 (■), 0.001 (▲). (b): Random architectures from DARTS space (Liu et al., 2019).

Results-C



(a) ResNet-18, varying samples. (b) All architectures, varying samples.

Figure 4: **Effect of Sample Size.**

Conclusions

- To understand the generalization (offline performance) in DL one has to look to population loss (online learning) has to be investigated
- Test performance of modern settings is close between infinite and finite sample sizes $\rightarrow$ quick online learners are well generalizers ?