Note: Self-Distillation from the Last Mini-Batch for Consistency Regularization.[1]

sgc

May 1, 2022

1 Introduction

Conventional KD requires a complex teacher resulting its low-efficiency. And recent self-KD need extra network or hard to parallelize. This paper proposed Self-Distillation from Last Mini-Batch (DLB). Rather than storing network configuration, this method only stores the soft labels.
Contributions:
1, With no network architecture modifications, DLB method requires very few additional computation costs as well as the run-time memory to implement.

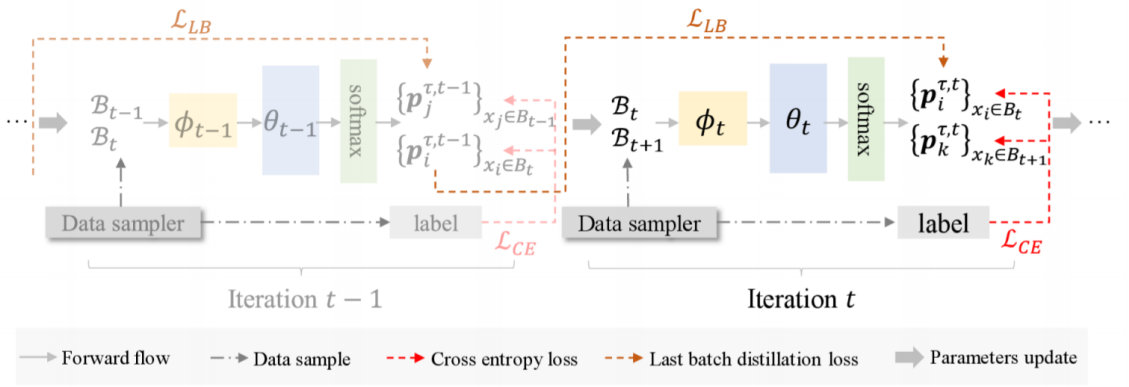

Figure 1. The overall architecture of our DLB. We write $\mathcal{B}_t$, ϕ_t, θ_t for a mini-batch of data samples, random augmentation and the trainable parameters indexed in the t^{th} iteration.

2 Method

2.1 Self-Distillation from Last Batch

The total loss consists of two parts:

$$L = L_{CE} + \alpha L_{LB}$$

Last-batch consistency regularization loss:

$$\mathcal{L}_{LB} = \frac{1}{n} \sum_{i=1}^{n} \tau^2 \cdot D_{KL}\left(\mathbf{p}_i^{\tau,t-1} \| \mathbf{p}_i^{\tau,t}\right).$$

where $p_i^{\tau,t-1}$ is the prediction of $t-1$ iteration, and $p_i^{\tau,t}$ is t's.
Both the predictions from B_{t-1} and B_t in $t-1^{th}$ iteration update the L_{CE}.

Algorithm 1 Pseudo code for DLB.

Input: balancing coefficient α
Input: distillation temperature τ
Require: data_loader samples batches as in Figure 1

```
last_logits = None # initialization
for (x,gt_labels) in data_loader do
    n̂ = gt_labels.size(0) # batch size
    logits = model.forward(x)
    loss = CELoss(logits, gt_labels)
    if last_logits != None then
        soft_targets = Softmax(last_logits/τ)
        pred = Softmax(logits[:n̂//2]/τ)
        loss += α*LBLoss(pred, soft_targets)*τ^2
    end if
    loss.backward() # update parameters
    last_logits = logits[n̂//2:].detach() # no gradient
end for
```

3 Results

DLB consistently improved the performance on various backbones, outperformed the SOTA. DLB and augmentation regularization work orthogonally.

4 Summary

This paper proposed DLB which does not require a complex teacher or ensemble peer students. It boosts the robustness to label noise. It uses half samples from the last iteration to keep consistency of samples and thus make it robust to noise.

References

[1] Y. Shen, L. Xu, Y. Yang, Y. Li, and Y. Guo, "Self-distillation from the last mini-batch for consistency regularization," *arXiv preprint arXiv:2203.16172*, 2022. (document)