SuperLoss: A Generic Loss for Robust Curriculum Learning

Satyam Kumar Gupta
B20ME068

---

1. Introduction

The basic idea behind Curriculum Learning (CL) is that the model should be trained first with easy samples, then medium samples and at last with hard samples. The motivation for this algorithm comes from how humans learn, where they start with simple concepts and on top of that they start building tougher concepts. The key idea for this algorithm is to somehow estimate the weights of the samples during training based on the observation that easy and hard samples behave differently and can therefore be separated. In this paper, we propose instead a simple yet generic approach to dynamic curriculum learning. It is inspired by recent confidence-aware loss functions that yield the capability to jointly learn network parameters and sample weights, or confidences, in a unified framework. In this paper we have talked about SuperLoss and its performance over different applications. SuperLoss simply plugged on top of original task loss during training and monitors the loss of each sample and determines the sample contribution dynamically by applying the core principle of curriculum learning.

2. SuperLoss

2.1 Confidence-aware loss

Consider a dataset $\{(x_i , y_i)\}^N_{i=1}$, where sample x_i has label y_i , and let $f(\cdot)$ be a trainable predictor to be optimised using empirical risk minimization. In contrast to traditional loss functions of the form $L(f(x_i), y_i)$, a confidence-aware loss function $L(f(x_i), y_i , \sigma_i)$ takes an additional learnable parameter $\sigma_i \geq 0$ as input. Such a parameter is associated with each sample x_i and represents the confidence or reliability of the corresponding prediction $f(x_i)$. The goal of a confidence-aware loss is to handle difficult samples without resorting to heuristics such as using robust versions of the loss. Regardless of the manner the confidence intervenes in the loss formula, a key property that they noticeably share is that the gradient of the loss w.r.t. the network parameters monotonously increases with the confidence, all other parameters staying fixed. This property makes confidence-aware losses particularly well suited to dynamic CL, as it allows to learn the confidence, i.e. weight, of each sample automatically through back-propagation and without further modification of the learning procedure.

2.2 Optimal confidence and SuperLoss

Instead of waiting for the confidence parameters to converge, we therefore propose to directly use their converged value at the limit, which only depends on the input loss L_i :

$$\sigma^*_\lambda (l_i) = \arg\min_{\sigma_i} L_\lambda (l_i , \sigma_i)$$

As a consequence, the confidence parameters do not need to be learned and are up-to-date with the sample status. The new loss function that we obtain takes a single parameter as input and can therefore simply be appended on top of any given task loss, hence its name of SuperLoss (SL)

$$SL_\lambda (l_i) = L_\lambda (l_i , \sigma^*_\lambda (l_i)) = \min_{\sigma_i} L_\lambda (l_i , \sigma_i)$$

3. Applications and Experimental results

3.1 Regression

We first evaluate our SuperLoss on digit regression on MNIST and human age regression on UTKFace, with both a robust loss (smooth-L1) and a non-robust one (L2), and with different noise levels. Models trained using the SuperLoss consistently outperform the baseline by a significant margin, regardless of the noise level, this is particularly true when the network is trained with a non-robust loss (L2), suggesting that the SuperLoss makes a non-robust loss more robust. Even when the baseline is trained using a robust loss (smooth-L1), the SuperLoss still significantly reduces the error.

3.2 Classification

We evaluate our SuperLoss for the image classification task on CIFAR-10, CIFAR-100 and WebVision. CIFAR-10 and CIFAR-100 consist of 50K training and 10K test images belonging to C = 10 and C = 100 classes respectively, and WebVision has 2.4 million images with C = 1000. We straightforwardly plug the Cross-Entropy loss into the SuperLoss.

3.3 Object detection

We perform experiments for the object detection task on Pascal VOC and its noisy version from where symmetric label noise is applied to 20%, 40% or 60% of the instances. We use two object detection frameworks from detectron: Faster R-CNN and RetinaNet. While the baseline and the SuperLoss are on par on clean data, the SuperLoss again significantly outperforms the baseline in the presence of noise.

3.4 Image retrieval

We evaluate the SuperLoss on the image retrieval task using the Revisited Oxford and Paris benchmark. To train our method, we use the large-scale Landmarks dataset that is composed of about 200K images. As retrieval model, we use ResNet-50 with Generalized-Mean (GeM) pooling and a contrastive loss. On clean data, the SuperLoss has minor impact. However, it enables an impressive performance boost on noisy data (Landmarks-full), overall outperforming the baseline trained using clean data. This result shows

that our SuperLoss makes it possible to train a model from a large automatically-collected dataset with a better performance than from a manually labelled subset.

4. Conclusion

SuperLoss, a straightforward task-agnostic loss function that may be added on top of any loss during training, has been introduced in the study. The ideal confidence can be stated as a closed-form solution, making it a confidence-aware loss function. The use of the SuperLoss results in implicit curriculum learning, which produces intrinsic noise robustness features, according to our results on a range of tasks.