Interpolation Consistency Training forSemi-Supervised Learning (2019)

Vikas Verma et al.

Notes

1 Introduction

In this paper the authors introducte ICT (Interpolation Consistency Training) which is a simple algorithm to be used in a semi supervised learning (SSL) setting where we have limited number of labeled data, and leverages the big number of unsupervised data to learn additionnal structure about the input distribution, ICT tries to move the decision boundary to low density regions based on the low density seperation assumption (will have better generalization if the models boundaries are in low density regions), given that a boundary in a high density region will cut a cluster into two different classes.

Some SSL methods tries to push the boundaries to low density regions based on consistency regularization techniques, by enouraging invariant predictions $f(x) = f(x + \epsilon)$ for small perturbations ϵ) to the input x.

2 ICT

ICT applies mixup operation to the unlabeled data, the labels to be mixed in this case are the model's predictions, but these prediction are not comming from the model f_θ we're training, but a EMI-model $f_{\theta'}$ (exponentially moving average) of the original model's weights :

$$f_\theta\left(\text{Mix}_\lambda\left(u_j, u_k\right)\right) \approx \text{Mix}_\lambda\left(f_{\theta'}\left(u_j\right), f_{\theta'}\left(u_k\right)\right)$$

$$\text{Where } \text{Mix}_\lambda(a, b) = \lambda \cdot a + (1 - \lambda) \cdot b$$

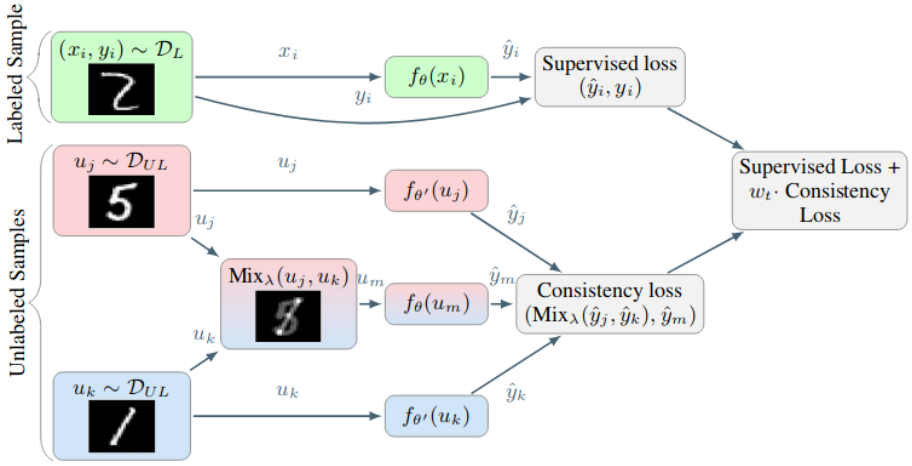

Figure 2: Interpolation Consistency Training (ICT) learns a student network f_θ in a semi-supervised manner. To this end, ICT uses a mean-teacher $f_{\theta'}$, where the teacher parameters θ' are an exponential moving average of the student parameters θ. During training, the student parameters θ are updated to encourage consistent predictions $f_\theta(\text{Mix}_\lambda(u_j, u_k)) \approx \text{Mix}_\lambda(f_{\theta'}(u_j), f_{\theta'}(u_k))$, and correct predictions for labeled examples x_i.

The mixup is only applied to unlabeled examples, for labeled examples we calculate the regular corss entropy loss between the predictions and the labels, and for the unlabeled examples we compare the mixed labels ($\hat{y}_j$ and $\hat{y}_k$) and the models prediction of the mixed inputs (u_j and u_k) using KL-divergence.

More formally, given a labeled example $(x_i, y_i) \sim \mathcal{D}_L$ from the joint distribution and an unlabeled example $u_j, u_k \sim \mathcal{D}_{UL}$ from the marginal disribution $P(X) = \frac{P(X,Y)}{P(Y|X)}$, the goal is to the train a model f_θ to be able to predict Y from X, where at each iteartion t we update the parameters θ to minimise:

$$L = L_S + w(t) \cdot L_{US}$$

Where L_S is the cross entropy and L_{US} is the consistency regularization term (KL-divergence or MSE between the mixup labels and the prediction). In sum the ICT term can be written as:

$$\mathcal{L}_{US} = \mathop{\mathbb{E}}_{u_j,u_k \sim P(X)} \mathop{\mathbb{E}}_{\lambda \sim \mathrm{Beta}(\alpha,\alpha)} \ell\left(f_\theta\left(\mathrm{Mix}_\lambda\left(u_j, u_k\right)\right), \mathrm{Mix}_\lambda\left(f_{\theta'}\left(u_j\right), f_{\theta'}\left(u_k\right)\right)\right)$$

Algorithm 1 The Interpolation Consistency Training (ICT) Algorithm

Require: $f_\theta(x)$: neural network with trainable parameters θ
Require: $f_{\theta'}(x)$ mean teacher with θ' equal to moving average of θ
Require: $\mathcal{D}_L(x, y)$: collection of the labeled samples
Require: $\mathcal{D}_{UL}(x)$: collection of the unlabeled samples
Require: α: rate of moving average
Require: $w(t)$: ramp function for increasing the importance of consistency regularization
Require: T: total number of iterations
Require: Q: random distribution on [0,1]
Require: $\mathrm{Mix}_\lambda(a, b) = \lambda a + (1 - \lambda)b$.
  for $t = 1, \ldots, T$ **do**
    Sample $\{(x_i, y_i)\}_{i=1}^B \sim \mathcal{D}_L(x, y)$ ▷ Sample labeled minibatch
    $L_S = \mathrm{CrossEntropy}(\{(f_\theta(x_i), y_i)\}_{i=1}^B)$ ▷ Supervised loss (cross-entropy)
    Sample $\{u_j\}_{j=1}^U, \{u_k\}_{k=1}^U \sim \mathcal{D}_{UL}(x)$ ▷ Sample two unlabeled examples
    $\{\hat{y}_j\}_{j=1}^U = \{f_{\theta'}(u_j)\}_{j=1}^U, \{\hat{y}_k\}_{k=1}^U = \{f_{\theta'}(u_k)\}_{k=1}^U$ ▷ Compute fake labels
    Sample $\lambda \sim Q$ ▷ sample an interpolation coefficient
    $(u_m = \mathrm{Mix}_\lambda(u_j, u_k), \hat{y}_m = \mathrm{Mix}_\lambda(\hat{y}_j, \hat{y}_k))$ ▷ Compute interpolation
    $L_{US} = \mathrm{ConsistencyLoss}(\{(f_\theta(u_m), \hat{y}_m)\}_{m=1}^U)$ ▷ e.g., mean squared error
    $L = L_S + w(t) \cdot L_{US}$ ▷ Total Loss
    $g_\theta \leftarrow \nabla_\theta L$ ▷ Compute Gradients
    $\theta' = \alpha\theta' + (1 - \alpha)\theta$ ▷ Update moving average of parameters
    $\theta \leftarrow \mathrm{Step}(\theta, g_\theta)$ ▷ e.g. SGD, Adam
  end for
  return θ

3 Experiments

Table 1: Error rates (%) on CIFAR-10 using CNN-13 architecture. We ran three trials for ICT.

Model	1000 labeled 50000 unlabeled	2000 labeled 50000 unlabeled	4000 labeled 50000 unlabeled
Supervised	39.95 ± 0.75	31.16 ± 0.66	21.75 ± 0.46
Supervised (Mixup)	36.48 ± 0.15	26.24 ± 0.46	19.67 ± 0.16
Supervised (Manifold Mixup)	34.58 ± 0.37	25.12 ± 0.52	18.59 ± 0.18
Π model (Laine & Aila, 2016)	31.65 ± 1.20	17.57 ± 0.44	12.36 ± 0.31
TempEns (Laine & Aila, 2016)	23.31 ± 1.01	15.64 ± 0.39	12.16 ± 0.24
MT (Tarvainen & Valpola, 2017)	21.55 ± 1.48	15.73 ± 0.31	12.31 ± 0.28
VAT (Miyato et al., 2018)	–	–	$11.36 \pm$ NA
VAT+Ent (Miyato et al., 2018)	–	–	$10.55 \pm$ NA
VAdD (Park et al., 2018)	–	–	11.32 ± 0.11
SNTG (Luo et al., 2018)	18.41 ± 0.52	13.64 ± 0.32	10.93 ± 0.14
MT+ Fast SWA (Athiwaratkun et al., 2019)	$15.58 \pm$ NA	$11.02 \pm$ NA	$9.05 \pm$ NA
ICT	$\mathbf{15.48 \pm 0.78}$	$\mathbf{9.26 \pm 0.09}$	$\mathbf{7.29 \pm 0.02}$

Table 2: Error rates (%) on SVHN using CNN-13 architecture. We ran three trials for ICT.

Model	250 labeled 73257 unlabeled	500 labeled 73257 unlabeled	1000 labeled 73257 unlabeled
Supervised	40.62 ± 0.95	22.93 ± 0.67	15.54 ± 0.61
Supervised (Mixup)	33.73 ± 1.79	21.08 ± 0.61	13.70 ± 0.47
Supervised (Manifold Mixup)	31.75 ± 1.39	20.57 ± 0.63	13.07 ± 0.53
Π model (Laine & Aila, 2016)	9.93 ± 1.15	6.65 ± 0.53	4.82 ± 0.17
TempEns (Laine & Aila, 2016)	12.62 ± 2.91	5.12 ± 0.13	4.42 ± 0.16
MT (Tarvainen & Valpola, 2017)	4.35 ± 0.50	4.18 ± 0.27	3.95 ± 0.19
VAT (Miyato et al., 2018)	–	–	$5.42 \pm$ NA
VAT+Ent (Miyato et al., 2018)	–	–	$3.86 \pm$ NA
VAdD (Park et al., 2018)	–	–	4.16 ± 0.08
SNTG (Luo et al., 2018)	$\mathbf{4.29 \pm 0.23}$	$\mathbf{3.99 \pm 0.24}$	$\mathbf{3.86 \pm 0.27}$
ICT	4.78 ± 0.68	4.23 ± 0.15	3.89 ± 0.04