Semi-Supervised Learning

Jay Sawant
18D070050

Guide: Prof. Amit Sethi

Indian Institute of Technology, Bombay

December 9, 2021

Table of Contents

What is Semi-Supervised Learning?

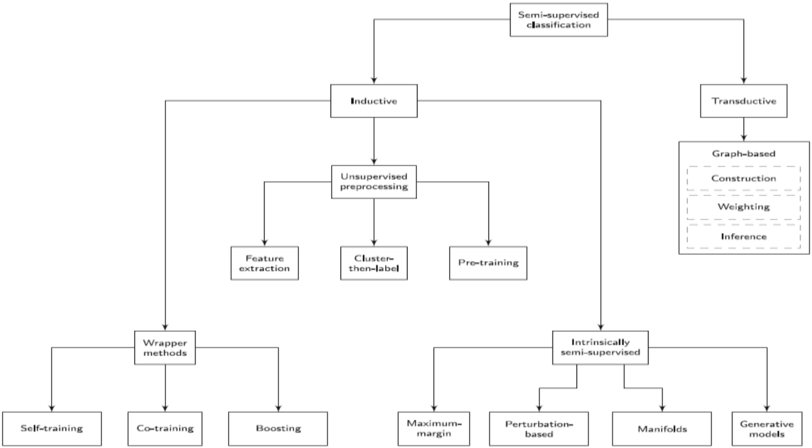
Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training

In order to make any use of unlabeled data, some relationship to the underlying distribution of data must exist. Semi-supervised learning algorithms make use of at least one of the following assumptions:

1. **Smoothness Assumption** - Points close to each other share same label
2. **Low Density Assumption** - The decision boundary of classifier should pass through low density region in input space
3. **Manifold Assumption** - The high dimensional data roughly lie on a low dimensional manifold.

Popular Semi-supervised Learning Methods

Taxonomy of SSL methods:



Reference: https://link.springer.com/content/pdf/10.1007/s10994-019-05855-6.pdf

Popular Semi-Supervised Learning Methods

Perturbation based SSL methods: The smoothness assumption entails that a predictive model should be robust to local perturbations in its input. This means that, when we perturb a data point with a small amount of noise, the predictions for the noisy and the clean inputs should be similar

Popular Examples of Perturbation based SSL methods:

1. Ladder Networks
2. Temporal Ensembling and Pi Model
3. Mean Teacher

Ladder Networks

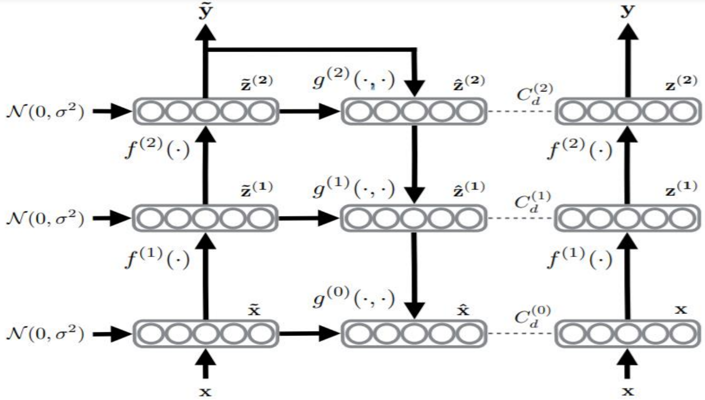


Figure 2: A conceptual illustration of the Ladder network when $L = 2$. The feedforward path ($\mathbf{x} \rightarrow \mathbf{z}^{(1)} \rightarrow \mathbf{z}^{(2)} \rightarrow \mathbf{y}$) shares the mappings $f^{(l)}$ with the corrupted feedforward path, or encoder ($\mathbf{x} \rightarrow \tilde{\mathbf{z}}^{(1)} \rightarrow \tilde{\mathbf{z}}^{(2)} \rightarrow \tilde{\mathbf{y}}$). The decoder ($\tilde{\mathbf{z}}^{(l)} \rightarrow \hat{\mathbf{z}}^{(l)} \rightarrow \hat{\mathbf{x}}$) consists of the denoising functions $g^{(l)}$ and has cost functions $C_d^{(l)}$ on each layer trying to minimize the difference between $\hat{\mathbf{z}}^{(l)}$ and $\mathbf{z}^{(l)}$. The output $\tilde{\mathbf{y}}$ of the encoder can also be trained to match available labels $t(n)$.

Reference: https://arxiv.org/pdf/1507.02672v2.pdf

Temporal Ensembling and Pi Model

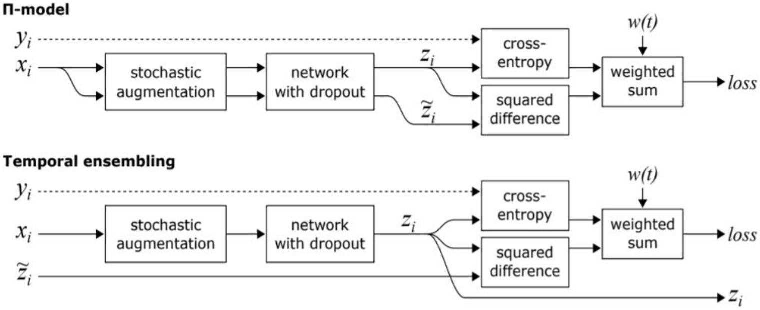
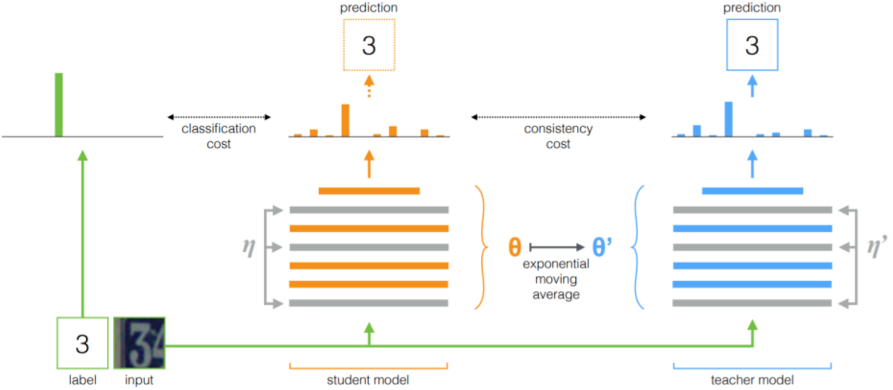


Figure 1: Structure of the training pass in our methods. Top: Π-model. Bottom: temporal ensembling. Labels y_i are available only for the labeled inputs, and the associated cross-entropy loss component is evaluated only for those.

Reference: https://arxiv.org/pdf/1610.02242v3.pdf

Mean Teacher



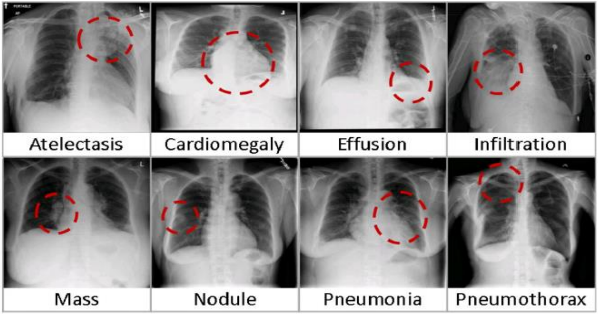
Reference: https://arxiv.org/pdf/1703.01780v6.pdf

Mean Teacher implementation on NIH ChestXray Data

NIH ChestXray Dataset:

NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories:

(1, Atelectasis; 2, Cardiomegaly; 3, Effusion; 4, Infiltration; 5, Mass; 6, Nodule; 7, Pneumonia; 8, Pneumothorax; 9, Consolidation; 10, Edema; 11, Emphysema; 12, Fibrosis; 13, Pleural_Thickening; 14 Hernia)



Reference: https://www.kaggle.com/nih-chest-xrays/data

Mean Teacher implementation on NIH ChestXray Data

Dataset overview:

1. A total of 112,120 frontal-view X-ray images of shape 1024x1024
2. Data Split → Training = 78468, Validation = 11219, Test = 22433
   Training Data → 7000 labelled, 71468 unlabelled
1. Each label is a multi-class label i.e. each X-ray image can have multiple diseases.

Dataset pre-processing:

1. Each batch from the training set is of size 16, with 4 labelled examples in it.
2. Transformations like changing the brightness, contrast, RandomAffine is applied on the datapoints.

Mean Teacher implementation on NIH ChestXray Data

Base Model used for student and teacher model:
Pre-trained DenseNet121 with a output of sigmoid classifier of 14 labels.

Losses:
Supervised loss = Classification cost (Binary Cross Entropy)
Unsupervised loss = consistency cost between student and teacher model (MSE loss)

Mean Teacher implementation on NIH ChestXray Data

Algorithm:

1. For a batch in training set, do
2. student_logits = student_model(input1)
   teacher_logits = teacher_model(input2)
   …input1 and input2 are same inputs with random transformations
1. consistency cost = ||stu_logits - tea_logits||^2
   classification cost = ||true_label - stu_logits||^2
   Total cost = classification cost + λ*consistency loss
   … λ = consistency weight
1. Calculate gradients and update parameters of the student model using SGD optimizer
2. Update teacher model weights as $\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$
3. Repeat 1 to 5 for num_epochs

Mean Teacher implementation on NIH ChestXray Data

Metric used for evaluating performance of the model:

AUC-ROC curve → area under the curve of TPR vs FPR

The area under the ROC curve (AUC) results were considered:

1. excellent for AUC values between 0.9-1
2. good for AUC values between 0.8-0.9
3. fair for AUC values between **0.7-0.8**
4. poor for AUC values between 0.6-0.7
5. failed for AUC values between 0.5-0.6

Mean Teacher implementation on NIH ChestXray Data

Results:

```
EPOCH: 0
Total loss 1030.8222995717078
class loss 1030.6795802991837
consistency loss 0.14271906193789619
Total val loss 116.9856825787574
class val loss 116.9851873870939
consistency val loss 0.0004952538775668813
AUC score: 0.7131397116040785
```

```
EPOCH: 15
Total loss 751.6875173393637
class loss 703.1530292080715
consistency loss 48.534487834433094
Total val loss 72.88555252365768
class val loss 115.39641387760639
consistency val loss 0.0008830433657749381
AUC score: 0.7435411251388654
```

```
EPOCH: 5
Total loss 863.3749642847106
class loss 831.5139921056107
consistency loss 31.860972347902134
Total val loss 111.87360934261233
class val loss 111.87314185500145
consistency val loss 0.00046768243399775145
AUC score: 0.7636333017368004
```

```
EPOCH: 30
Total loss 596.3262274060398
class loss 509.4171453532763
consistency loss 86.90908176451921
Total val loss 123.90479649789631
class val loss 123.78775057010353
consistency val loss 0.11704600178563851
AUC score: 0.7176975372780114
```

Mean Teacher implementation on NIH ChestXray Data

Results:

```
EPOCH: 44
Total loss 585.3332506380975
class loss 494.7162082383875
consistency loss 90.61704244487919
Total val loss 125.19013787712902
class val loss 124.96883323788643
consistency val loss 0.2213045152439082
AUC score: 0.7185228960417668
```

```
EPOCH: 53
Total loss 583.6874677641317
class loss 493.72282836632803
consistency loss 89.96463950281031
Total val loss 124.2961710775271
class val loss 124.16575331613421
consistency val loss 0.130417572679562
AUC score: 0.7193798680416584
```

Mean Teacher implementation on NIH ChestXray Data

Future Work:

1. Tuning the hyper-parameters like learning rate, LR scheduler, weight decay in SGD optimizer, consistency weight regulator, etc
2. Adding self-consistency loss i.e. for two different transformations of same input, student model by itself should be self-consistent.