Explicit and Implicit Entropy Minimization in Proxy-Label-Based Semi-Supervised Learning

Colin Raffel

CVPR Workshop on Learning with Limited and Imperfect Data
Number of labels required to reach 90% accuracy on CIFAR-10
$\mathbb{E}_{x,y} - y \log p_{\theta}(y|x)$
$E_x - \hat{p}_\theta(y|x) \log p_\theta(y|x)$
\( \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \)

“Proxy label”
\[ E_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

“Pseudo-label”
\[ E_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

“Label guess”
Self-training

Unlabeled example

Model

Prediction

Proxy label

\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

\[ \hat{p}_\theta(y|x) = \arg \max_y [p_\theta(y|x)] \]
Self-training with Noisy Student improves ImageNet classification, Xie et al. 2019

Train teacher model with labeled data

Infer pseudo-labels on unlabeled data

Train equal-or-larger student model with combined data and noise injected

Make the student a new teacher

Data augmentation
Dropout
Stochastic depth
Probability of Error of Some Adaptive Pattern-Recognition Machines

H. J. SCUDDER, III, MEMBER, IEEE

We will make the untaught machine from the same basic configuration as the taught machine and use the output of the machine $\hat{\theta}_n$ to iterate the estimate, instead of teaching the machine with the correct observation $\theta_n$ each time.
\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]
\[ \hat{p}_\theta(y|x) = \arg \max_y [p_\theta(y|x)] \]
\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]
\[ \hat{p}_\theta(y|x) = \arg \max_y [p_\theta(y|x)] \]
\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

\[ \hat{p}_\theta(y|x) = \arg\max_y [p_\theta(y|x)] \]

\[ \hat{p}_\theta(y|x) = p_\theta(y|x) \]
\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

\[ \hat{p}_\theta(y|x) = \arg \max_y [p_\theta(y|x)] \]

\[ \hat{p}_\theta(y|x) = p_\theta(y|x) \]

\[ \mathbb{E}_x - p_\theta(y|x) \log p_\theta(y|x) \]
\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

\[ \hat{p}_\theta(y|x) = \arg \max_y [p_\theta(y|x)] \]

\[ \hat{p}_\theta(y|x) = p_\theta(y|x) \]

\[ \mathbb{E}_x - p_\theta(y|x) \log p_\theta(y|x) \]

*Entropy!*
Entropy Minimization

\[ \mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x) \]

\[ \hat{p}_\theta(y|x) = p_\theta(y|x) \]
\[ I(c; \mathbf{x}) = \int \int dc \, d\mathbf{x} \, p(c, \mathbf{x}) \log \frac{p(c, \mathbf{x})}{p(c)p(\mathbf{x})} \]
\[ = \int d\mathbf{x} \, p(\mathbf{x}) \int dc \, p(c|\mathbf{x}) \log \frac{p(c|\mathbf{x})}{p(c)} \]
\[ = \int d\mathbf{x} \, p(\mathbf{x}) \int dc \, p(c|\mathbf{x}) \log \frac{p(c)}{\int d\mathbf{x} \, p(\mathbf{x})p(c|\mathbf{x})} \]

The elements of this expression are separately recognizable:
\( \int d\mathbf{x} \, p(\mathbf{x})(\cdot) \) is equivalent to an average over a training set \( \frac{1}{N_{ts}} \sum_{ts}(\cdot) \);
p(\mathbf{x}|c) is simply the network output \( y_c \);
\( \int dc (\cdot) \) is a sum over the class labels and corresponding network outputs.
Hence:
\[ I(c; \mathbf{x}) = \frac{1}{N_{ts}} \sum_{ts} \sum_{i=1}^{N_c} y_i \log \frac{y_i}{\bar{y}_i} \]
\[ = -\sum_{i=1}^{N_c} \bar{y}_i \log \bar{y}_i + \frac{1}{N_{ts}} \sum_{ts} \sum_{i=1}^{N_c} y_i \log y_i \]
\[ = H(\bar{y}) - H(y) \]
Self-training with Noisy Student improves ImageNet classification, Xie et al. 2019
Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O’Connor, Kevin McGuinness
Insight Centre for Data Analytics, Dublin City University (DCU)
{eric.arazo, diego.ortego}@insight-centre.org

Experiments show that this naive pseudo-labeling is limited by confirmation bias as prediction errors are fit by the network. To deal with this issue, we propose to use mixup augmentation [25] as an effective regularization that helps calibrate deep neural networks [26] and, therefore, alleviates confirmation bias.
\[
\mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x)
\]
\[
\hat{p}_\theta(y|x) = p_\theta(y|x')
\]
Learned feature extraction

Classifier
craffel@gmail.com