Note: ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation

Sinkoo

April 17, 2022

1 Introduction

Authors found injecting strong data augmentation(SDA) on unlabeled images extremely beneficial to decouple their predictions as well as alleviate overfittingon noisy pseudo labels. And they uses it on basic self-training framework and named as ST as a strong baseline. And they think that utilizing all unlabeled images at the same time degrades the models performance, to deal with that authors further proposed ST++ that automatically selects and prioritizes more reliable images in the re-training phase to produce higher-quality artificial labels on the remaining less reliable images. Both ST and ST++ outperform the current methods.

2 Method

2.1 ST: Inject SDA on Unlabeled Images

To deal with overfitting noisy labels and prediction coupling between student and teacher, authors propose SDA(including colorjitter, grayscale, blur and Cutout) on unlabeled images.Oversampling D^l to around the size of $\hat{D}^u$ And the loss is: $A^u_{ST} = H(T(x), S(A^s(A^w(x))))$ (why there still remains the WDA? : For labeled samples)

Algorithm 1: ST Pseudocode

Input: Labeled training set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^M$,
Unlabeled training set $\mathcal{D}^u = \{u_i\}_{i=1}^N$,
Weak/strong augmentations $\mathcal{A}^w/\mathcal{A}^s$,
Teacher/student model T/S

Output: Fully trained student model S

Train T on $\mathcal{D}^l$ with cross-entropy loss $\mathcal{L}_{ce}$
Obtain pseudo labeled $\hat{\mathcal{D}}^u = \{(u_i, T(u_i))\}_{i=1}^N$
Over-sample $\mathcal{D}^l$ to around the size of $\hat{\mathcal{D}}^u$
for *minibatch* $\{(x_k, y_k)\}_{k=1}^B \subset (\mathcal{D}^l \cup \hat{\mathcal{D}}^u)$ **do**
  for $k \in \{1, \ldots, B\}$ **do**
    if $x_k \in \mathcal{D}^u$ **then**
      $x_k, y_k \leftarrow \mathcal{A}^s(\mathcal{A}^w((x_k, y_k))$
    else
      $x_k, y_k \leftarrow \mathcal{A}^w(x_k, y_k)$
    $\hat{y}_k = S(x_k)$
  Update S to minimize $\mathcal{L}_{ce}$ of $\{(\hat{y}_k, y_k)\}_{k=1}^B$
return S

3 ST++: Select and Prioritize Reliable Images

Using

$$s_i = \sum_{j=1}^{K-1} meanIOU(M_{ij}, M_{iK})$$

to measure the reliability.Choose the R images largest stability score for the first re-training and in the second re-training relabel the rest images.

Algorithm 2: ST++ Pseudocode

Input: Same as Algorithm 1

Output: Same as Algorithm 1

Train T on $\mathcal{D}^l$ and save K checkpoints $\{T_j\}_{j=1}^K$

for $u_i \in \mathcal{D}^u$ **do**

  for $T_j \in \{T_j\}_{j=1}^K$ **do**

    Pseudo mask $M_{ij} = T_j(u_i)$

  Compute s_i with Equation 4 and $\{M_{ij}\}_{j=1}^K$

Select R highest scored images to compose $\mathcal{D}^{u_1}$

$\mathcal{D}^{u_2} = \mathcal{D}^u - \mathcal{D}^{u_1}$

$\mathcal{D}^{u_1} = \{(u_k, T(u_k))\}_{u_k \in \mathcal{D}^{u_1}}$

Train S on $(\mathcal{D}^l \cup \mathcal{D}^{u_1})$ with ST re-training

$\mathcal{D}^{u_2} = \{(u_k, S(u_k))\}_{u_k \in \mathcal{D}^{u_2}}$

Re-initialize S

Train S on $(\mathcal{D}^l \cup \mathcal{D}^{u_1} \cup \mathcal{D}^{u_2})$ with ST re-training

return S

4 Results

ST and ST++ outperform most of current method and as number of labeled images increasing, the gender between proposed methods and other method becoming even larger.
On Cityscape, ST and ST++ with ResNet-50 already surpasses other methods with ResNet-101.

5 Ablation Study

1, Different data augmentation takes different performance improvement: Colorjitter works best when only one method is implemented and Full SDA works the best.
2, Apply SDA on unlabeled data works even better than adopt on both labeled data and unlabeled data. 3, Image-level selection brings consistent improvements over ST framework and is superior to the pixel-level counterpart. 4, The

performance can be further boosted with iterate training.

6 Summary

This paper proposed ST and ST++(Adopted prioritize reliability on pseudo labels) and outperform previous method by a large margin. ST++ is really simple: it doesn't use iterate training(actually iterate training helps). And this paper shows that SDA on unlabeled data can solve: 1, similar prediction from student and teacher; 2, overfitting on noisy.