When Does Label Smoothing Help?

1 Introduction

- In these early days of neural network research, mor exotic objectives could outperform the standard cross entropy loss.
- Label smoothing has been used successfully to improve the accuracy of deep learning models
- paper's contributions
  - label smoothing calibrates learned models, so the confidences of their predictions are more aligned with the accuracies of their predictions.
  - label smoothing impairs distillation. It is from loss of information in the logits

1.1 Preliminaries

- p_k : the likelihood the model assigns to the k-th class $\left(p_k = \frac{\exp(x^T w_k)}{\sum_{l=1}^{L} \exp(x^T w_l)}\right)$
- $\boldsymbol{w}_k$: the weights and biases of the last layer
- $\boldsymbol{x}$: the vector containing the activations of the penultimate layer (마지막 layer 이전의 layer를 의미) concatenated with "1" to account for the bias
- With hard target, minimize the expected value of the cross-entropy $H(\boldsymbol{y}, \boldsymbol{p}) = \sum_{k=1}^{K} -y_k \log(p_k)$, where y_k is "1" for the correct class and "0" else class.
- With label smoothing of parameter α, use the modified targets $y_k^{LS} = y_k(1-\alpha) + \alpha/\text{K}$.
  - 예시) Hard Target : [0,1,0,0] vs Label smoothing with $\alpha = 0.1$: [0.025,0.925,0.025,0.025]

2 Penultimate layer representations

- logit$\boldsymbol{x^T w}$ $\left(= \log\left(\frac{x^T w_k}{1-x^T w_k}\right)\right)$ of the k-th class can be thought of as a measure of the squared Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{w}_k$.
  - $||\boldsymbol{x} - \boldsymbol{w}_k||^2 = \boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{w}_k + \boldsymbol{w}_k^T\boldsymbol{w}_k$
  - $\boldsymbol{x}^T\boldsymbol{x}$ is factored out in calculating softmax outputs
  - $\boldsymbol{w}_k^T\boldsymbol{w}_k$ is usually constant across classes
  - 위의 Euclidean distance 수식으로 이해해도 좋고 바로 inner product로 이해해도 될 것 같다.
- *label smoothing encourages the activations of the penultimate layer to be close to the template of the correct class and equally distant to the templates of the incorrect classes.*
  - *activations of the penultimate layer :* $\boldsymbol{x}$
  - *template :* $\boldsymbol{w}$
- To obseve the property of label smoothing : 4 case visualizing

  (1) Pick 3 classes

  (2) Orthonormal basis of the plane crossing the templates of these 3 classes

  (3) Project the penultimate layer activations of examples from these three classes

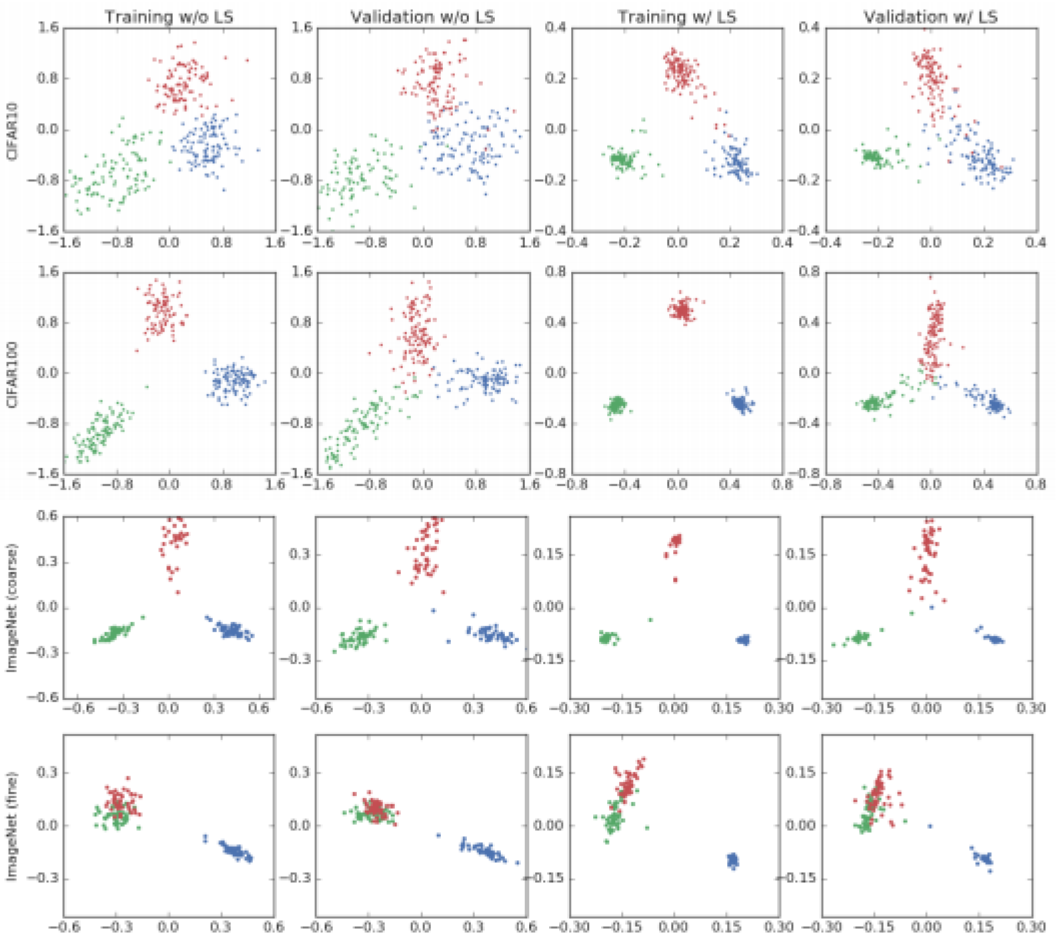


Figure 1: Visualization of penultimate layer's activations of: AlexNet/CIFAR-10 (first row), CIFAR-100/ResNet-56 (second row) and ImageNet/Inception-v4 with three semantically different classes (third row) and two semantically similar classes plus a third one (fourth row).

- result
  - With LS, clusters are much tighter. Because LS encourages that each example in training set to be equidistant from all the other class'sb templates.
  - And scale is wider in w/o LS. (overconfident)
  - (forth row) Semantically similar classes are more seperated well in LS case.

3 Implicit model calibration

- Temperature scaling
  - NN이 예측 과정에서 과신(over-confident)하는 경향이 있어, 이를 완화해 일반화 성능을 높이는 기법
  - multiplying the logits by a scalar before applying the softmax operator.

$$P_k = \frac{\exp(x^T w_k/T)}{\sum_l \exp(x^T w_l/T)}$$

  - logit에 T 만큼 나누어주어서 softmax의 분포가 uniform하게 만들어서 calibration하는 것이다.

- temperature scaling없이도 LS가 ECE(estimated expected calibration error)를 줄이고 network를 calibration한다는 것을 보이려고 한다.

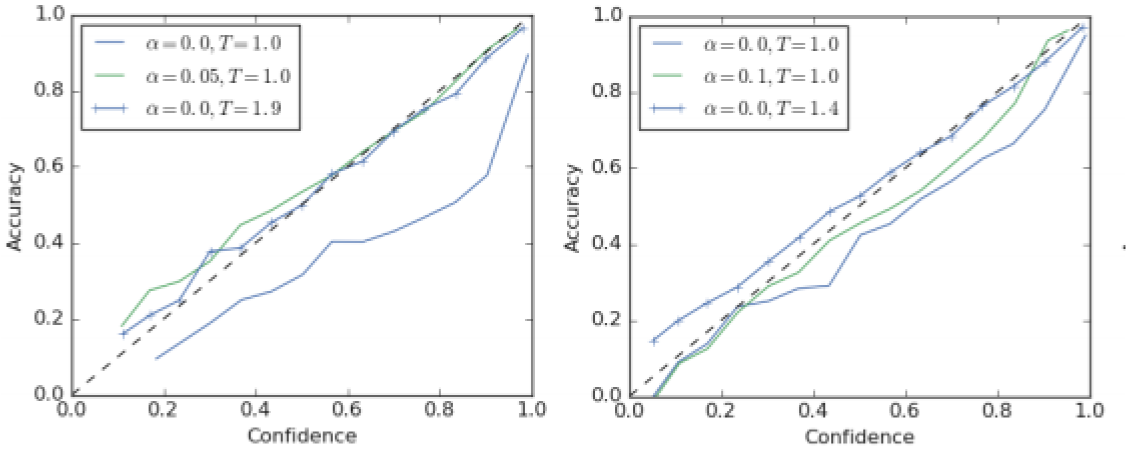
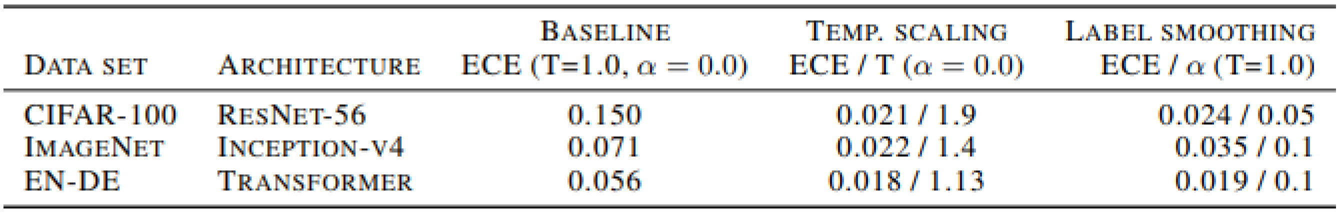


Figure 2: Reliability diagram of ResNet-56/CIFAR-100 (left) and Inception-v4/ImageNet (right).

Table 3: Expected calibration error (ECE) on different architectures/datasets.



Data set	Architecture	Baseline ECE (T=1.0, $\alpha = 0.0$)	Temp. scaling ECE / T ($\alpha = 0.0$)	Label smoothing ECE / α (T=1.0)
CIFAR-100	ResNet-56	0.150	0.021 / 1.9	0.024 / 0.05
ImageNet	Inception-v4	0.071	0.022 / 1.4	0.035 / 0.1
EN-DE	Transformer	0.056	0.018 / 1.13	0.019 / 0.1

- (image classification, machine translation) Both methods can be used to reduce ECE to a smilar and smaller value than an uncalibrated network trained with hard targets.
- (machine translation) Despite being better calibrated and achieving better BLEU scores, label smoothing results in worse negative log-likelihoods (NLL) than hard targets.

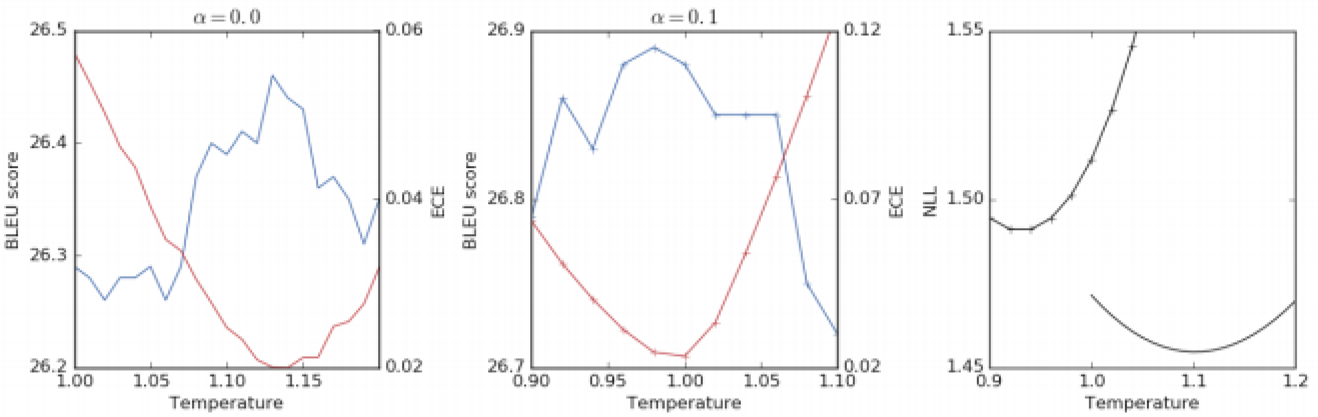


Figure 4: Effect of calibration of Transformer upon BLEU score (blue lines) and NLL (red lines). Curves without markers reflect networks trained without label smoothing while curves with markers represent networks with label smoothing.

- calibration이 항상 만능은 아니다. task와 metric에 따라 잘 사용해야할것 같다.

4 Knowledge distillation

- label smoothing slightly degrades the baseline performance of the student networks.
- why?
  - 오른쪽 그림은 teacher network를 대상으로 hard target과 teacher network의 로짓(logit) 사이의 mutual information을 근사한 결과이다. LS를 적용한 모델의 mutual information 이 적다. LS 를 사용하여 teacher network가 받아들이는 정보가 감소했음을 보여준다.

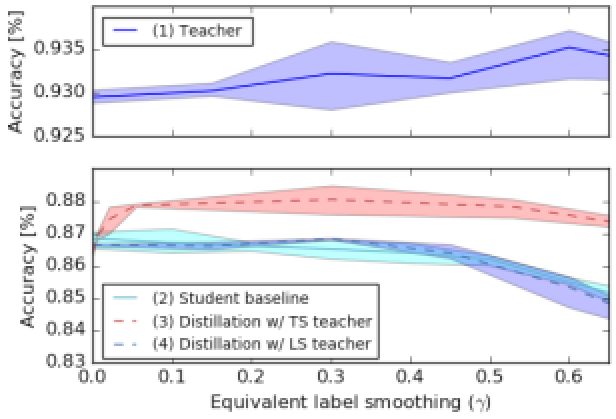


Figure 5: Performance of distillation from ResNet-56 to AlexNet on CIFAR-10.

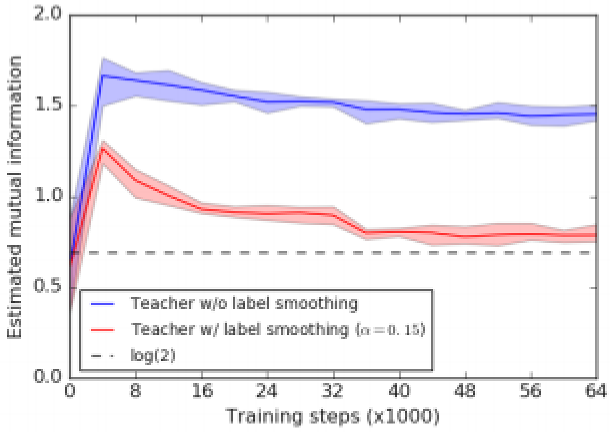


Figure 6: Estimated mutual information evolution during teacher training.

- *a teacher with better accuracy is not necessarily the one that distills better.*

5 Related work

6 Conclusion and future work

References

[1] https://arxiv.org/abs/1906.02629

[2] https://ratsgo.github.io/insight-notes/docs/interpretable/smoothing