When Does Label Smoothing help?

Rafael Muller, Simon Kornblith, Geoffrey Hinton
Google Brain
Toronto

NIPS 2019 (Oral)

Presented by Eungyeup Kim

Vision Seminar
03 DEC 2019

Motivation

Despite its widespread use, label smoothing is still poorly understood

- Since Szegedy et al.(2016), label smoothing has improved the accuracy of deep learning models across a range of tasks including image classification, speech recognition and machine translation

- It also helps improve the model calibration

- Still not much is known about why and when label smoothing should work

DATA SET	ARCHITECTURE	METRIC	VALUE W/O LS	VALUE W/ LS
IMAGENET	INCEPTION-V2 [6]	TOP-1 ERROR	23.1	**22.8**
		TOP-5 ERROR	6.3	**6.1**
EN-DE	TRANSFORMER [11]	BLEU	25.3	**25.8**
		PERPLEXITY	**4.67**	4.92
WSJ	BILSTM+ATT.[10]	WER	8.9	7.0/**6.7**

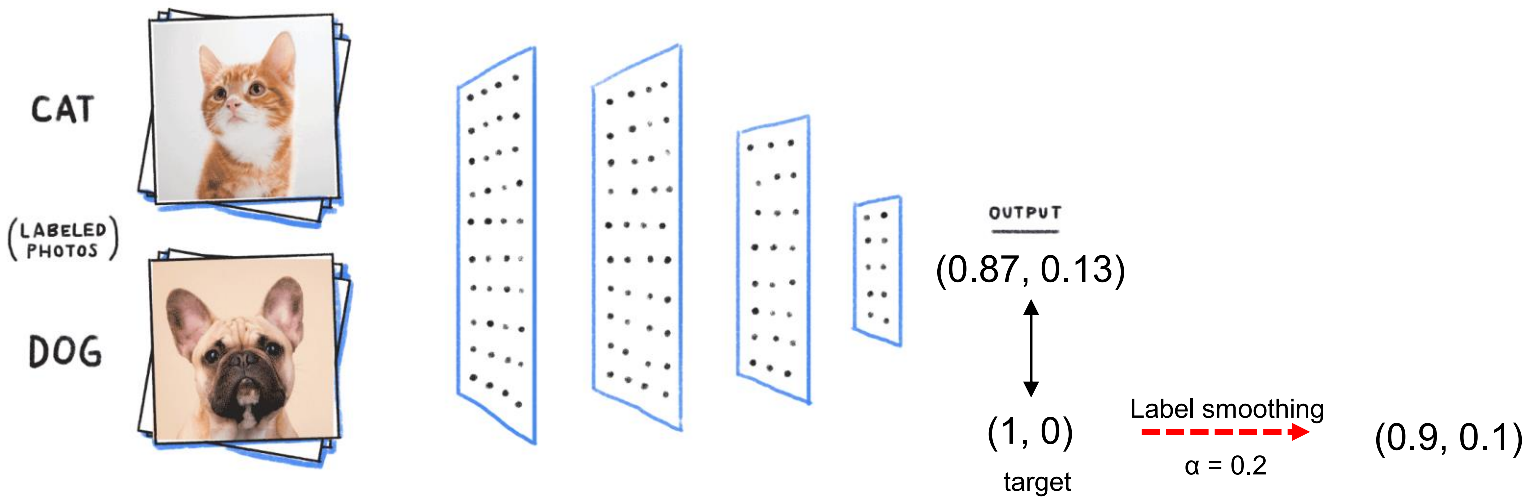
Introduction

Label Smoothing

- Let the true targets y_k where y_k is "1" for the correct class and "0" for the rest
- Networks' outputs $p_k = \frac{e^{x^T w_k}}{\sum_{l=1}^{L} e^{x^T w_l}}$, where p_k is the likelihood the model assigns to the k^{th} class, w_k represents the weights and biases of the last layer, x is the vector containing the activations of the penultimate layer of a neural network.

- Cross-entropy loss that we minimize is $H(y, p) = \sum_{k=1}^{K} - y_k \log(p_k)$
- Modified target after label smoothing with parameter α is y_k^{LS},

$$y_k^{LS} = y_k(1 - \alpha) + \alpha / K$$



Analysis

Penultimate layer representations

$$p_k = \frac{e^{x^T w_k}}{\sum_{l=1}^{L} e^{x^T w_l}} \propto e^{x^T w_k}$$

- The logit $x^T w_k$ can be thought of as a measure of the squared Euclidean distance between x^T and w_k, as $\|x - w_k\|^2 = x^T \mathrm{x} - 2\mathrm{x}^\mathrm{T} w_k + w_k^T w_k$
- Training with Cross-entropy loss

$$H(y,p) = \sum_{k=1}^{K} -y_k \log(p_k) = -\, y_1 \log(p_1) - y_2 \log(p_2) - \cdots - y_k \log(p_k) - \cdots$$

enforces the penultimate layer representation x to be close to the template of the class w

- **Hard Target** encourages the activation of the penultimate layer to be close to the template of the correct class w_k **with no constraints,** resulting in the correct logit being much larger than any other logits of incorrect classes

- **Soft Target** encourages the activation of the penultimate layer to be close to the template of the correct class w_k **while encouraging to be close to the template of the incorrect classes $w_{/k}$ in some degree,** resulting in the differences between the logit of correct class and incorrect to be a constant dependent on α

Analysis

Penultimate layer representations

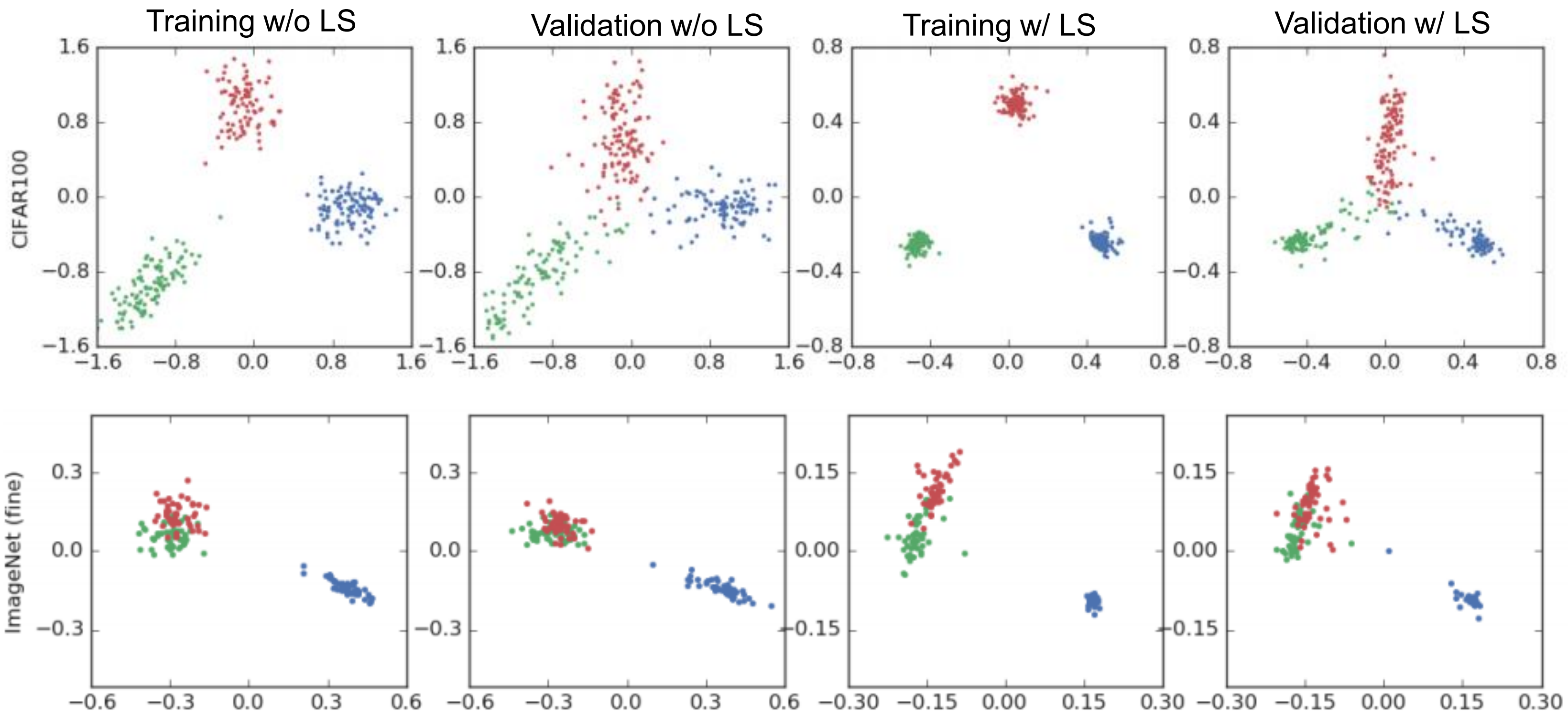
- Without label smoothing, the larger the logit for correct class $x^T w_k$ is, the lower the loss we minimize
  => This causes an overconfidence of our model, which results in overfitting

- With label smoothing, the logit $x^T w_k$ with overly large magnitude is being penalized to maintain to be equally distant to the incorrect logits
  => This regularizes the models' confidence not to be overfitted, showing better results in test time

Analysis

Visualization

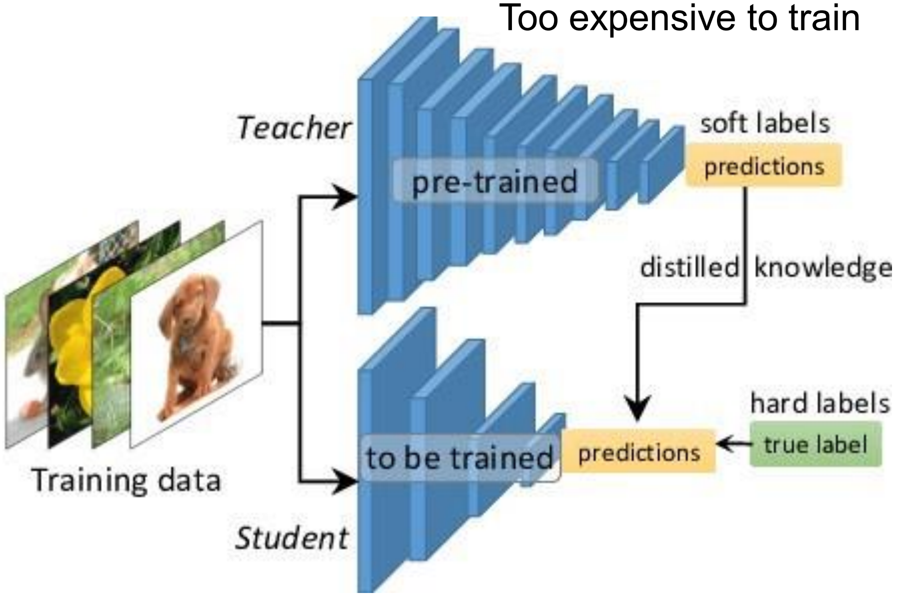
Setup
(1) Pick three classes
(2) Find an orthonormal basis of the plane crossing the templates of these three classes
(3) Project the penultimate layer activations of examples onto this plane



Analysis

Knowledge Distillation

Too expensive to train



$$minimize\ (1-\beta)H(y,p) + \beta H(p^t(T), p(T)),$$

- $p^t(T)$ *and* $p(T)$ are the outputs of the student and teacher after temperature scaling with temperature T, respectively

- Temperature exaggerates the differences between the probabilities of incorrect answers

=> Both label smoothing and knowledge distillation involve fitting a model using soft targets

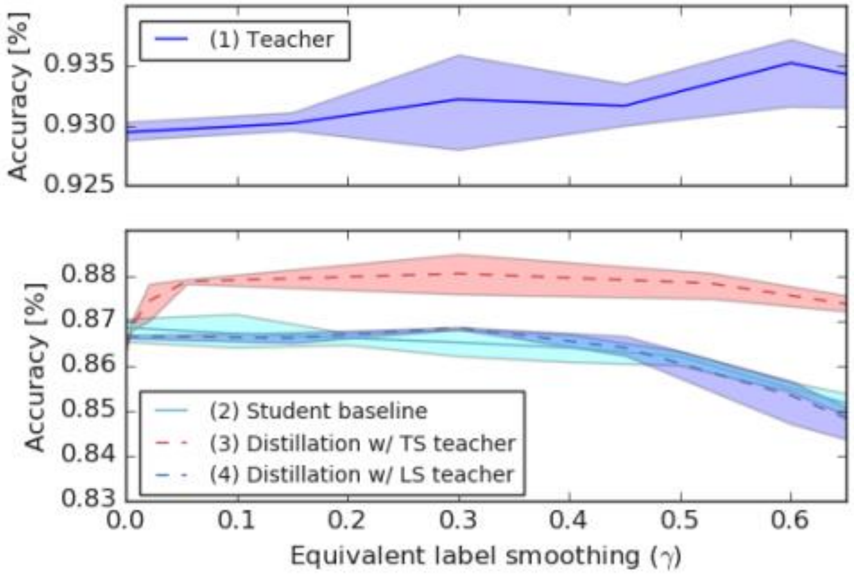
It has been observed that use of label smoothing to train a teacher network degrades the ability to distill the teacher's knowledge into a student network…

Analysis

Knowledge Distillation and Label Smoothing

- Teacher model: ResNet-56
- Student model: AlexNet

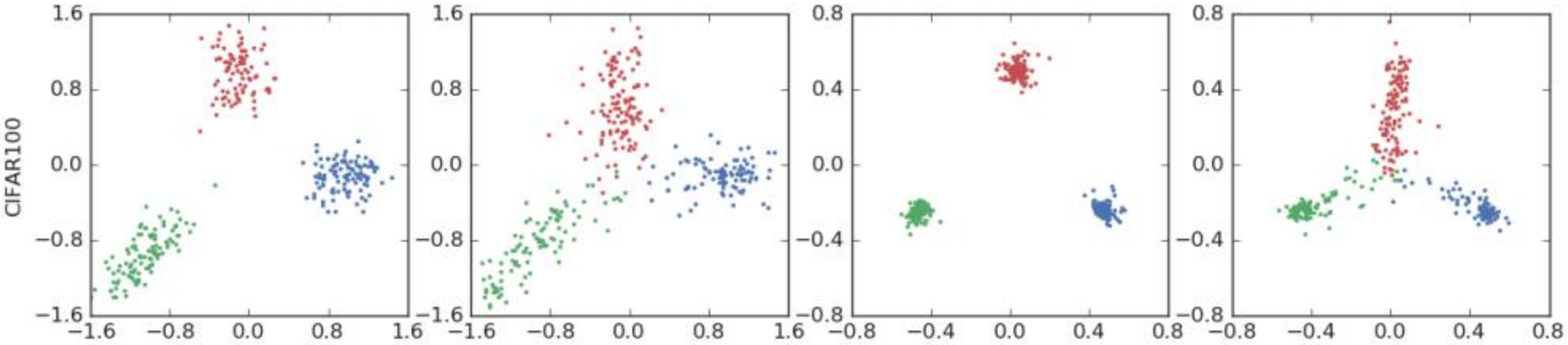
1. The teacher's accuracy as a function of the label smoothing factor
2. The student's baseline accuracy as a function of the label smoothing factor without distillation
3. The student's accuracy after distillation with temperature scaling to control the smoothness of the teacher's provided targets (teacher trained with hard targets)
4. The student's accuracy after distillation with fixed temperature (T=1.0 and teach trained with label smoothing to control the smoothness of the teacher's provided targets)



⇒ Label smoothing of teacher model degrades the accuracy, as the relative information between logits is **erased** when the teacher is trained with label smoothing

Analysis

Knowledge Distillation and Label Smoothing



- Different examples from the same class can have very different similarities to other classes

- Every example from the same class has very similar proximities to examples of the other classes

⇒ Achieves the better accuracy, but loses the information to distill to student…

Thank you