Soft Augmentation for Image Classification

Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, Deva Ramanan
Argo AI
youngleoel@gmail.com, shenyan@google.com, leal.taixe@tum.de, hays@gatech.edu, deva@cs.cmu.edu

Abstract

Modern neural networks are over-parametrized and thus rely on strong regularization such as data augmentation and weight decay to reduce overfitting and improve generalization. The dominant form of data augmentation applies invariant transforms, where the learning target of a sample is invariant to the transform applied to that sample. We draw inspiration from human visual classification studies and propose generalizing augmentation with invariant transforms to soft augmentation where the learning target softens non-linearly as a function of the degree of the transform applied to the sample: e.g., more aggressive image crop augmentations produce less confident learning targets. We demonstrate that soft targets allow for more aggressive data augmentation, offer more robust performance boosts, work with other augmentation policies, and interestingly, produce better calibrated models (since they are trained to be less confident on aggressively cropped/occluded examples). Combined with existing aggressive augmentation strategies, soft targets 1) double the top-1 accuracy boost across Cifar-10, Cifar-100, ImageNet-1K, and ImageNet-V2, 2) improve model occlusion performance by up to $4 \times$, and 3) half the expected calibration error (ECE). Finally, we show that soft augmentation generalizes to self-supervised classification tasks. Code available at https://github.com/youngleox/soft_augmentation

1. Introduction

Deep neural networks have enjoyed great success in the past decade in domains such as visual understanding [42], natural language processing [5], and protein structure prediction [41]. However, modern deep learning models are often over-parameterized and prone to overfitting. In addition to designing models with better inductive biases, strong regularization techniques such as weight decay and data augmentation are often necessary for neural networks to achieve ideal performance. Data augmentation is often a computationally cheap and effective way to regularize models and mitigate overfitting. The dominant form of data augmentation modifies training samples with invariant transforms – transformations of the data where it is assumed that the identity of the sample is invariant to the transforms.

Indeed, the notion of visual invariance is supported by evidence found from biological visual systems [54]. The robustness of human visual recognition has long been documented and inspired many learning methods including data augmentation and architectural improvement [19, 47]. This paper focuses on the counterpart of human visual robustness, namely how our vision fails. Instead of maintaining perfect invariance, human visual confidence degrades non-linearly as a function of the degree of transforms such as occlusion, likely as a result of information loss [44]. We propose modeling the transform-induced information loss for learned image classifiers and summarize the contributions as follows:

- We propose Soft Augmentation as a generalization of data augmentation with invariant transforms. With Soft Augmentation, the learning target of a transformed training sample softens. We empirically compare several softening strategies and prescribe a robust non-linear softening formula.
- With a frozen softening strategy, we show that replacing standard crop augmentation with soft crop augmentation allows for more aggressive augmentation, and doubles the top-1 accuracy boost of RandAugment [8] across Cifar-10, Cifar-100, ImageNet-1K, and ImageNet-V2.
- Soft Augmentation improves model occlusion robustness by achieving up to more than $4 \times$ Top-1 accuracy boost on heavily occluded images.
- Combined with TrivialAugment [37], Soft Augmentation further reduces top-1 error and improves model calibration by reducing expected calibration error by more than half, outperforming 5-ensemble methods [25].
- In addition to supervised image classification models, Soft Augmentation also boosts the performance of self-supervised models, demonstrating its generalizability.
2. Related Work

2.1. Neural Networks for Vision

Since the seminal work from Krizhevsky et al. [24], neural networks have been the dominant class of high performing visual classifiers. Convolutional Neural Networks (CNNs) are a popular family of high performing neural models which borrow a simple idea of spatially local computations from biological vision [12, 18, 26]. With the help of architectural improvements [15], auxiliary loss [42], and improved computational power [13], deeper, larger, and more efficient neural nets have been developed in the past decade.

2.2. Data Augmentation

Data augmentation has been an essential regularizer for high performing neural networks in many domains including visual recognition. While many other regularization techniques such as weight decay [32] and batch normalization [4] are shown to be optional, we are aware of no competitive vision models that omit data augmentation.

Accompanying the influential AlexNet model, Krizhevsky et al. [24] proposed horizontal flipping and random cropping transforms which became the backbone of image data augmentation. Since the repertoire of invariant transformations has grown significantly in the past decade [42], choosing which subset to use and then finding the optimal hyperparameters for each transform has become computationally burdensome. This sparked a line of research [7, 28] which investigates optimal policies for data augmentation such as RandAugment [8] and TrivialAugment [37].

2.3. Learning from Soft Targets

While minimizing the cross entropy loss between model logits and hard one-hot targets remains the go-to recipe for supervised classification training, learning with soft targets has emerged in many lines of research. Label Smoothing [36, 43] is a straightforward method which applies a fixed smoothing (softening) factor $\alpha$ to the hard one-hot classification target. The motivation is that label smoothing prevents the model from becoming over-confident. Müller et al. [36] shows that label smoothing is related to knowledge distillation [17], where a student model learns the soft distribution of a (typically) larger teacher model.

A related line of research [49, 53] focuses on regularizing how a model interpolates between samples by linearly mixing two or more samples and linearly softening the resulting learning targets. Mixing can be in the form of per-pixel blending [53] or patch-level recombination [49].

2.4. Robustness of Human Vision

Human visual classification is known to be robust against perturbations such as occlusion. In computer vision research, the robustness of human vision is often regarded as the gold standard for designing computer vision models [34, 54]. These findings indeed inspire development of robust vision models, such as compositional, recurrent, and occlusion aware models [22, 46, 47]. In addition to specialty models, much of the idea of using invariant transforms to
augment training samples come from the intuition and observation that human vision are robust against these transforms such as object translation, scaling, occlusion, photometric distortions, etc.

Recent studies such as Tang et al. [44] indeed confirm the robustness of human visual recognition against mild to moderate perturbations. In a 5-class visual classification task, human subjects maintain high accuracy when up to approximately half of an object is occluded. However, the more interesting observation is that human performance starts to degenerate rapidly as occlusion increases and falls to chance level when the object is fully occluded (see Figure 2 right $k = 2, 3, 4$ for qualitative curves).

3. Soft Augmentation

In a typical supervised image classification setting, each training image $x_i$ has a ground truth learning target $y_i$ associated to it thus forming tuples:

$$\{(x_i, y_i)\}, \tag{1}$$

where $x_i \in \mathbb{R}^{C \times W \times H}$ denotes the image and $y_i \in [0, 1]^N$ denotes a $N$-dimensional one-hot vector representing the target label (Figure 2 left, “Hard Target”). As modern neural models have the capacity to memorize even large datasets [1], data augmentation mitigates the issue by hallucinating data points through transformations of existing training samples.

(Hard) data augmentation relies on the key underlying assumption that the augmented variant of $x_i$ should maintain the original target label $y_i$:

$$\langle x_i, y_i \rangle \Rightarrow \langle t_{\phi \sim S}(x_i), y_i \rangle, \tag{2}$$

where $t_{\phi \sim S}(x_i)$ denotes the image transform applied to sample $x_i$, $\phi$ is a random sample from the fixed transform range $S$. Examples of image transforms include translation, rotation, crop, noise injection, etc. As shown by Tang et al. [44], transforms of $x_j$ such as occlusion are approximately perceptually invariant only when $\phi$ is mild. Hence $S$ often has to be carefully tuned in practice, since naively increasing it can lead to degraded performance (Figure 1). In the extreme case of 100% occlusion, total information loss occurs, making it detrimental for learning.

Label Smoothing applies a smoothing function $g$ to the target label $y_i$, parameterized by a handcrafted, fixed smoothing factor $\alpha$. Specifically, label smoothing replaces the indicator value ‘1’ (for the ground-truth class label) with $p = 1 - \alpha$, distributing the remaining $\alpha$ probability mass across all other class labels (Figure 2 left, “Soft Target”). One can interpret label smoothing as accounting for the average loss of information resulting from averaging over transforms from the range $S$. From this perspective, the smoothing factor $\alpha$ can be written as a function of the fixed transform range $S$:

$$\langle x_i, y_i \rangle \Rightarrow \langle t_{\phi \sim S}(x_i), g_\alpha(S)(y_i) \rangle, \tag{3}$$

Soft Augmentation, our proposed approach, can now be described succinctly as follows: replace the fixed smoothing factor $\alpha(S)$ with an adaptive smoothing factor $\alpha(\phi)$, that depends on the degree of the specific sampled augmentation $\phi$ applied to $x_i$:

$$\langle x_i, y_i \rangle \Rightarrow \langle t_{\phi \sim S}(x_i), g_\alpha(\phi)(y_i) \rangle, \tag{4}$$

Crucially, conditioning on the information loss from a particular $\phi$ allows one to define far larger augmentation ranges $S$. We will show that such a strategy consistently produces robust performance improvements with little tuning across a variety of datasets, models, and augmentation strategies.

Extensions to Soft Augmentation may be proposed by also considering loss reweighting [40, 48], which is an alternative approach for softening the impact of an augmented example by down-weighting its contribution to the loss. To formalize this, let us write the training samples of a supervised dataset as triples including a weight factor $w_i$ (that is typically initialized to all ‘1’s). One can then re-purpose our smoothing function $g$ to modify the weight instead of (or in addition to) the target label (Figure 2 left):

$$\langle x_i, y_i, w_i \rangle \Rightarrow \langle t_{\phi \sim S}(x_i), y_i, g_\alpha(\phi)(w_i) \rangle, \tag{5}$$

Soft Augmentation (Target & Weight) \hspace{1cm} \tag{6}

Finally, one may wish to soften targets by exploiting class-specific confusions when applying $\alpha(\phi)$; the smoothed target label of a highly-occluded truck example could place more probability mass on other vehicle classes, as opposed to distributing the remaining probability equally across all other classes. Such extensions are discussed in Section 5.

4. Experiments

4.1. Soft Augmentation with Crop

As a concrete example of the proposed Soft Augmentation, we consider the crop transform $t_{(tx, ty, w, h)}(x)$. In the case of $32 \times 32$ pixel Cifar images [23], the cropped images typically have a constant size $w = h = 32$, and $t(x)$ is fully parameterized by $tx$ and $ty$, which are translational offsets between the cropped and the original image. As shown in
Figure 1 (left), the standard hard crop augmentation for the Cifar-10/100 classification tasks draws $tx, ty$ independently from a uniform distribution of a modest range $U(-4, 4)$. Under this distribution, the minimal visibility of an image is 77% and ResNet-18 models trained on the Cifar-100 task achieve mean top-1 validation error of 20.80% across three independent runs (Figure 1 left). Naively applying aggressive hard crop augmentation drawn from a more aggressive range $U(-16, 16)$ increases top-1 error by 2.19% (Figure 1 middle). We make two changes to the standard crop augmentation.

We first propose drawing $tx, ty$ independently from a scaled normal distribution $S* \sim N(0, \sigma L)$ (with clipping such that $|tx| < L, |ty| < L$), where $L$ is the length of the longer edge of the image ($L = 32$ for Cifar). The distribution has zero mean and $\sigma$ controls the relative spread of the distribution hence the mean occlusion level. Following the $3\sigma$ rule of normal distribution, an intuitive tuning-free choice is to set $\sigma \approx 0.3$, where $\approx 99\%$ of cropped samples have visibility $\geq 0$. Figure 3 (left, $\alpha = 0$) shows that changing the distribution alone without target softening provides a moderate $\sim 0.4\%$ performance boost across crop strength $\sigma$.

Directly borrowing the findings from human vision research [44], one can define an adaptive softening $\alpha(tx, ty, k)$ that softens the ground truth learning target. Similar to Label Smoothing [43], a hard target can be softened to confidence $p \in [0, 1]$. Instead of a fixed $\alpha$, consider a family of power functions that produces target hardness $p$ given crop parameters $tx, ty$ and curve shape $k$:

$$p = 1 - \alpha(tx, ty, k) = 1 - (1 - p_{min})(1 - v_{(tx,ty)})^k,$$

where $v_{(tx,ty)} \in [0, 1]$ is the image visibility which is a function of $tx$ and $ty$. The power function family is a simple one-parameter formulation that allows us to test both linear ($k = 1$) and non-linear ($k \neq 1$) softening: higher $k$ provides flatter plateaus in high visibility regime (see Figure 2 right). As seen in Label Smoothing, $p$ can be interpreted as ground
true class probability of the one-hot learning target. $p_{\text{min}}$ is the chance probability depending on the task prior. For example, for Cifar-100, $p_{\text{min}} = \frac{1}{100} = 0.01$.

Equation 7 has three assumptions: 1) the information loss is a function of image visibility and all information is lost only when the image is fully occluded, 2) the original label of a training image has a confidence of $100\%$, which suggests that there is no uncertainty in the class of the label, and 3) the information loss of all images can be approximated by a single confidence-visibility curve. While the first assumption is supported by observations of human visual classification research [44], and empirical evidence in Sections 4.2 and 4.3 suggests that the second and the third assumptions approximately hold, the limitations to these assumptions will be discussed in Section 5.

4.2. How to Soften Targets

As prescribed by Equations 4, 5, and 6, three versions of Soft Augmentation are compared with Label Smoothing across a range of crop strength $\sigma$. The popular ResNet-18 models [16] are trained on the 100-class classification Cifar-100 training set. Top-1 error reductions on the validation set are reported (details in Appendix B). Consistent with prior studies, label smoothing can boost model performance by $\sim 1.3\%$ when the smoothing factor $\alpha$ is properly tuned (Figure 3 left).

Combining both target and weight softening (Equation 6) with $k = 2$ and $\sigma \approx 0.3$ boosts model performance by $2.5\%$ (Figure 3 right). Note that $k = 2$ qualitatively resembles the shape of the curve of human visual confidence degradation under occlusion reported by Tang et al. [44]. Interestingly, the optimal $\sigma \approx 0.3$ fits the intuitive 3-σ rule. In the next section we freeze $k = 2$ and $\sigma = 0.3$ and show robust improvements that generalize to Cifar-10 [23], ImageNet-1K [9], and ImageNet-V2 tasks [39].

4.3. Supervised Classification

4.3.1 Comparison with Related Methods

As mentioned in Section 2, many approaches similar to soft augmentation have demonstrated empirical performance gains, including additional data augmentation transforms [10], learning augmentation policies [8], softening learning targets [43], and modifying loss functions [29]. However, as training recipes continued to evolve over the past decade, baseline model performance has improved accordingly. As seen in Table 1 (Baseline), our baseline ResNet-18 models with a 500-epoch schedule and cosine learning rate decay [33] not only outperform many recent baseline models of the same architecture, but also beat various published results of Mixup and Cutout. To ensure fair comparisons, we reproduce 6 popular methods: Mixup, Cutout, Label Smoothing, Online Label Smoothing, Focal Loss, and RandAugment, and report the Top-1 Error on Cifar-100 in Table 1. Additional comparisons with the self-reported results are available in Appendix Table 5.

Table 1 shows that Soft Augmentation outperforms all other single methods. It is worth noting that although focal loss [29] proposed for detection tasks, it can be tuned to slightly improve classification model performance.

4.3.2 Soft Augmentation Compliments RandAugment

This section investigates the robustness of Soft Augmentation across models and tasks, and how well it compares or complements more sophisticated augmentation policies such as RandAugment [8]. The ImageNet-1K dataset [9] has larger and variable-sized images compared to the Cifar [23] datasets. In contrast with the fixed-sized crop augmentation for Cifar, a crop-and-resize augmentation $t_{(tx, ty, w, h)}(x)$ with random location $tx, ty$ and random size $w, h$ is standard for ImageNet training recipes [7,8,42]. The resizing step is necessary to produce fixed-sized training images (e.g. $224 \times 224$). We follow the same $\sigma = 0.3$ principle for drawing $tx, ty$ and $w, h$ (details in Appendix B).

Comparing single methods, Soft Augmentation with crop only consistently outperforms the more sophisticated RandAugment with 14 transforms (Table 2). The small ResNet-18 models trained with SA on Cifar-10/100 even outperforms much larger baseline ResNet-50 [39] and WideResNet-28 [50] models.
Table 2. Soft Augmentation (SA) with a fixed softening curve of \(k = 2\) doubles the top-1 error reduction of RandAugment (RA) across datasets and models. Note that the ResNet-18 models trained with SA on Cifar-10/100 even outperform larger baseline ResNet-50 and WideResNet-28 models. All results are mean ± standard error of top-1 validation error in percentage. Best results are shown in bold, runners-up are underlined, and results in parentheses indicate improvement over baseline. Statistics are computed from three runs.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Baseline</th>
<th>SA</th>
<th>RA</th>
<th>SA+RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cifar100</td>
<td>EfficientNet-B0</td>
<td>47.00±0.55</td>
<td>42.13±0.42 (−7.57)</td>
<td>46.68±0.52 (−3.02)</td>
<td>38.72±0.71 (−10.98)</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>20.80±0.11</td>
<td>18.31±0.17 (−2.49)</td>
<td>19.09±0.11 (−1.70)</td>
<td>18.10±0.20 (−2.70)</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>20.18±0.30</td>
<td>18.06±0.24 (−2.12)</td>
<td>22.57±0.09 (−9.16)</td>
<td>16.72±0.06 (−3.46)</td>
</tr>
<tr>
<td></td>
<td>WideResNet-28</td>
<td>18.60±0.19</td>
<td>16.47±0.18 (−2.13)</td>
<td>17.65±0.14 (−0.95)</td>
<td>15.37±0.17 (−3.23)</td>
</tr>
<tr>
<td></td>
<td>PyramidNet + ShakeDrop</td>
<td>15.77±0.17</td>
<td>14.03±0.05 (−1.75)</td>
<td>14.10±0.28 (−1.76)</td>
<td>12.78±0.16 (−2.99)</td>
</tr>
<tr>
<td>Cifar10</td>
<td>EfficientNet-B0</td>
<td>17.73±0.69</td>
<td>12.21±0.22 (−5.52)</td>
<td>14.54±0.47 (−3.19)</td>
<td>11.67±0.26 (−6.06)</td>
</tr>
<tr>
<td></td>
<td>ResNet-18</td>
<td>4.38±0.05</td>
<td>3.54±0.08 (−0.87)</td>
<td>3.89±0.06 (−0.49)</td>
<td>3.27±0.06 (−1.11)</td>
</tr>
<tr>
<td></td>
<td>ResNet-50</td>
<td>4.34±0.14</td>
<td>3.67±0.06 (−0.67)</td>
<td>3.91±0.14 (−0.43)</td>
<td>3.01±0.02 (−1.33)</td>
</tr>
<tr>
<td></td>
<td>WideResNet-28</td>
<td>3.67±0.08</td>
<td>2.85±0.02 (−0.82)</td>
<td>3.26±0.04 (−0.41)</td>
<td>2.45±0.04 (−1.20)</td>
</tr>
<tr>
<td></td>
<td>PyramidNet + ShakeDrop</td>
<td>2.86±0.03</td>
<td>2.26±0.02 (−0.60)</td>
<td>2.32±0.08 (−0.54)</td>
<td>2.02±0.01 (−0.84)</td>
</tr>
<tr>
<td>ImageNet-1K</td>
<td>ResNet-50</td>
<td>22.62±0.01</td>
<td>21.66±0.02 (−0.96)</td>
<td>22.02±0.02 (−0.60)</td>
<td>21.27±0.06 (−1.35)</td>
</tr>
<tr>
<td></td>
<td>ResNet-100</td>
<td>20.91±0.04</td>
<td>20.63±0.00 (−0.28)</td>
<td>20.39±0.07 (−0.52)</td>
<td>19.86±0.00 (−1.05)</td>
</tr>
<tr>
<td>ImageNet-V2</td>
<td>ResNet-50</td>
<td>34.97±0.03</td>
<td>33.32±0.10 (−1.65)</td>
<td>34.16±0.21 (−0.81)</td>
<td>32.38±0.16 (−2.59)</td>
</tr>
<tr>
<td></td>
<td>ResNet-101</td>
<td>32.68±0.04</td>
<td>31.81±0.16 (−0.87)</td>
<td>32.08±0.19 (−0.60)</td>
<td>31.26±0.12 (−1.42)</td>
</tr>
</tbody>
</table>

Because RandAugment is a searched policy that is originally prescribed to be applied in addition to the standard crop augmentation [8], one can easily replace the standard crop with soft crop and combine Soft Augmentation and RandAugment. As shown in Table 2, Soft Augmentation complements RandAugment by doubling its top-1 error reduction across tasks and models.

Note that for the small ResNet-18 model trained on Cifar-100, the fixed RandAugment method slightly degrades its performance. Consistent with observations from Cubuk et al. [8], the optimal hyperparameters for RandAugment depend on the combination of model capacity and task complexity. Despite the loss of performance of applying RandAugment alone, adding Soft Augmentation reverses the effect and boosts performance by 2.7%.

For the preceding experiments, a fixed \(k = 2\) is used for Soft Augmentation and the official PyTorch RandomAugment [38] is implemented to ensure a fair comparison and to evaluate robustness. It is possible to fine-tune the hyperparameters for each model and task to achieve better empirical performance.

4.3.3 Occlusion Robustness

As discussed in Section 2, occlusion robustness in both human vision [34,44,54] and computer vision [22,46,47] have been an important property for real world applications of vision models as objects. To assess the effect of soft augmentation on occlusion robustness of computer vision models, ResNet-50 models are tested with occluded ImageNet validation images (Figure 4 and Appendix Figure 7). 224 × 224 validation images of ImageNet are occluded with randomly placed square patches that cover \(\lambda\) of the image area. \(\lambda\) is set to \(\{0\%, 20\%, 40\%, 60\%, 80\%\}\) to create a range of occlusion levels.

As shown in Figure 5, both RandAugment (RA) and Soft Augmentation (SA) improve occlusion robustness independently across occlusion levels. Combining RA with SA reduces Top-1 error by up to 17%. At 80% occlusion level, SA+RA achieves more than 4 times accuracy improvement over the baseline (18.98% vs 3.42%).

4.3.4 Confidence Calibration

In addition to top-1 errors, reliability is yet another important aspect of model performance. It measures how close a model’s predicted probability (confidence) tracks the true correctness likelihood (accuracy). Expected Calibration Error (ECE) is a popular metric [14,25,35] to measure confidence calibration by dividing model predictions into \(M\) confidence bins \((B_m)\) and compute a weighted average error between accuracy and confidence:

\[
ECE = \frac{1}{n} \sum_{m=1}^{M} |B_m| \left(\frac{acc(B_m) - conf(B_m)}{n}\right),
\]

where \(n\) is the number of samples, \(acc(B_m)\) denotes the accuracy of bin \(m\), and \(conf(B_m)\) denotes mean model confidence of bin \(m\). Consistent with Guo et al. [14], we set \(M = 10\) and compute ECE for Cifar-10 and Cifar-100 tasks.

As shown in Table 3, many methods [25,30,35,45] have been proposed to improve confidence calibration, sometimes at the cost of drastically increased computational overhead [25], or degraded raw performance [30]. We show in Table 3 (and Appendix Table 7) that it is possible to further reduce model top-1 error and expected calibration error simultaneously.
Figure 4. Examples of occluded ImageNet validation images and model predictions of ResNet-50. 224 × 224 validation images of ImageNet are occluded with randomly placed square patches that cover λ of the image area. λ is set to {0%, 20%, 40%, 60%, 80%} to create a range of occlusion levels.

Figure 5. Soft Augmentation improves occlusion robustness of ResNet-50 on ImageNet. Both RandAugment (RA) and Soft Augmentation (SA) improve occlusion robustness independently. Combining RA with SA reduces Top-1 error by up to 17%. At 80% occlusion level, compared with baseline accuracy (3.42%), SA+RA achieves more than 4× accuracy (18.98%).

Compared to previous single-model methods, our strong baseline WideResNet-28 models achieve lower top-1 error at the cost of higher ECE. Combining Soft Augmentation with more recently developed augmentation policies such as TrivialAugment [37] (SA+TA) reduces top-1 error by 4.36% and reduces ECE by more than half on Cifar-100, outperforming the 4× more computationally expensive 5-ensemble model [25]. To the best of our knowledge, this is state of the art ECE performance for WideResNet-28 on Cifar without posthoc calibration.

4.4. Soft Augmentation Boosts Self-Supervised Learning

In contrast with supervised classification tasks where the learning target \( y_t \) is usually a one-hot vector, many self-supervised methods such as SimSiam [6] and Barlow Twins [51] learn visual feature representations without class labels by encouraging augmentation invariant feature representations. This section investigates whether Soft Augmentation generalizes to learning settings where no one-hot style labeling is provided.

In a typical setting, two random crops of the same image are fed into a pair of identical twin networks (e.g., ResNet-18) with shared weights and architecture. The learning target can be the maximization of similarity between the feature representations of the two crops [6], or minimization of redundancy [51]. By default, all randomly cropped pairs have equal weights. We propose and test two alternative hypotheses for weight softening with SimSiam. To accommodate self-supervised learning, Equation 7 is modified by replacing visibility \( v_{tx,ty} \) with intersection over union \( IoU \) of two crops of an image:

\[
p = 1 - \alpha(\phi_1, \phi_2, k) = 1 - (1 - p_{\text{min}})(1 - IoU_{\phi_1, \phi_2})^k,
\]

SA#1 (9)

where \( \phi_1 = (tx_1, ty_1, w_1, h_1) \) and \( \phi_2 = (tx_2, ty_2, w_2, h_2) \) are crop parameters for the first and second sample in a pair.
Table 4. Soft Augmentation (SA#1) improves self-supervised learning with SimSiam (ResNet-18) on Cifar-100 by down-weighting sample pairs with small intersection over union (IoU), outperforming the opposite hypothesis (SA#2) of down-weighting pairs with large IoU. For each configuration, we report means and standard errors of 3 runs with best learning rates (LR) found for Cifar-100. The effect of SA#1 (with a fixed $k = 4$) generalizes to Cifar-10 without re-tuning.

$$p = 1 - \alpha(\phi_1, \phi_2, k) = 1 - (1 - p_{\min})(\text{IoU}_{\phi_1, \phi_2})^k.$$  

SA#2 (10)

We first test all three hypotheses (baseline, SA#1, and SA#2) on Cifar-100 with the SimSiam-ResNet-18 models. Table 4 (top) shows that SA#1 outperforms both baseline and SA#2 (details in Appendix B.4). Additional experiments show that models trained with the same SA#1 configuration also generalize to Cifar-10 (Table 4 bottom).

5. Discussion

Other augmentations. While we focus on crop augmentations as an illustrative example, Soft Augmentation can be easily extended to a larger repertoire of transforms such as affine transforms and photometric distortions, as seen in the more sophisticated augmentation policies such as RandAugment. As the formulation of Equation 7 (and Figure 2 right) is directly inspired by the qualitative shape of human vision experiments from Tang et al. [44], optimal softening curves for other transforms may be discovered by similar human experiments. However, results with a second transform in Appendix Table 6 suggest that Equation 7 generalizes to additive noise augmentation as well. A potential challenge is determining the optimal softening strategy when a combination of several transforms are applied to an image since the cost of a naive grid search increases exponentially with the number of hyperparameters. Perhaps reinforcement learning methods as seen in RandAugment can be used to speed up the search.

Other tasks. While we limit the scope of Soft Augmentation to image classification as it is directly inspired by human visual classification research, the idea can be generalized to other types of tasks such as natural language modeling and object detection. Recent studies have shown that detection models benefit from soft learning targets in the final stages [3,27], Soft Augment has the potential to complement these methods by modeling information loss of image transform in the models’ input stage.

Class-dependant augmentations. As pointed out by Balestriero et al. [2], the effects of data augmentation are class-dependent. Thus assumption 3 of Equation 7 does not exactly hold. One can loosen it by adaptively determining the range of transform and softening curve on a per class or per sample basis. As shown in Equation 11,

$$\begin{align*}
(x_i, y_i) &\Rightarrow \left( t_{\phi \sim S(x_i,y_i)}(x_i), g_{\alpha}(\phi, x_i, y_i)(y_i) \right),
\end{align*}$$

(11)

two adaptive improvements can be made: 1) the transformation range $S$ where $\phi$ is drawn from can be made a function of sample $(x_i, y_i)$, 2) the softening factor $\alpha$ can also adapt to $(x_i, y_i)$. Intuitively, the formulation recognizes the heterogeneity of training samples of images at two levels. Firstly, the object of interest can occupy different proportions of an image. For instance, a high-resolution training image with a small object located at the center can allow more aggressive crop transforms without losing its class invariance. Secondly, texture and shape may contribute differently depending on the visual class. A heavily occluded tiger may be recognized solely by its distinctive stripes; in contrast, a minimally visible cloak can be mistaken as almost any clothing.

6. Conclusion

In summary, we draw inspiration from human vision research, specifically how human visual classification performance degrades non-linearly as a function of image occlusion. We propose generalizing data augmentation with invariant transforms to Soft Augmentation where the learning target (e.g. one-hot vector and/or sample weight) softens non-linearly as a function of the degree of the transform applied to the sample.

Using cropping transformations as an example, we empirically show that Soft Augmentation offers robust top-1 error reduction across Cifar-10, Cifar-100, ImageNet-1K, and ImageNet-V2. With a fixed softening curve, Soft Augmentation doubles the top-1 accuracy boost of the popular RandAugment method across models and datasets, and improves performance under occlusion by up to 4×. Combining Soft Augment with the more recently developed TrivialAugment further improves model accuracy and calibration simultaneously, outperforming even compute-heavy 5-ensemble models. Finally, self-supervised learning experiments demonstrate that Soft Augmentation also generalizes beyond the popular supervised one-hot classification setting.
References


[52] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. IEEE Transactions on Image Processing, 30:5984–5996, 2021. 5, 12
