

SUPPLEMENTARY MATERIAL FOR WHAT ARE EFFECTIVE LABELS FOR AUGMENTED DATA? IMPROVING ROBUSTNESS WITH AUTO LABEL

Anonymous authors

Paper under double-blind review

A THE PROPOSED AUTO LABEL ALGORITHM

In Algorithm 1 we describe how AutoLabel works.

Algorithm 1 Pseudocode of AutoLabel

- 1: **Input:** A training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1, \dots, m}$, a validation dataset \mathcal{D}_V drawn i.i.d. from the same distribution, an augmentation method `Aug`. Number of classes K , number of training epochs T , number of distance buckets N and the hyperparameter α .
 - 2: We perform `Aug` to obtain the augmented training data $\text{Aug}(x, s)$, where the transformation distance s is determined by the hyperparameters in the `Aug`. We discretize the transformation distance s into N buckets $\{S_1, \dots, S_N\}$, where each S_n is a range.
 - 3: For each distance bucket S_n , we initialize $\tilde{y}^0(S_n)$ as the one-hot label.
 - 4: **for** $t = 0$ **to** $T - 1$ **do**
 - 5: Minimize cross-entropy loss over the augmented training data with smoothed labels $\tilde{y}^t(S_n)$.
 - 6: **for** $n = 1$ **to** N **do**
 - 7: Generate an augmented validation set:
 $\mathcal{Q}(S_n) = \{(\text{Aug}(x_i, s), y_i) | (x_i, y_i) \in \mathcal{D}_V, s \sim \mathcal{U}(S_n)\}$.
 - 8: Update the label for the true class $\tilde{y}_{k=y}^{t+1}(S_n)$: \triangleright according to Eqn (1)
 $\tilde{y}_{k=y}^{t+1}(S_n) = \tilde{y}_{k=y}^t(S_n) - \alpha \cdot \text{ECE}^t(\mathcal{Q}(S_n)) \cdot \text{sign}(\text{Conf}^t(\mathcal{Q}(S_n)) - \text{Acc}^t(\mathcal{Q}(S_n)))$
 - 9: Clip $\tilde{y}_{k=y}^{t+1}(S_n)$ to be within $[\text{Acc}^t(\mathcal{Q}(S_n)), 1]$
 - 10: Update the label for other classes $\tilde{y}_{k \neq y}^{t+1}(S_n)$: \triangleright according to Eqn (2)
 $\tilde{y}_{k \neq y}^{t+1}(S_n) = (1 - \tilde{y}_{k=y}^{t+1}(S_n)) \cdot \frac{1}{K-1}$ \triangleright according to Eqn (2)
 - 11: **end for**
 - 12: **end for**
-

B ABLATION STUDY

In this section, we compare AutoLabel with:

- Label Smoothing (LS) (Szegedy et al., 2016): which softs labels by sweeping a hyperparameter ϵ which controls the smoothing degree in a range to find the best hyperparameter ρ ,
- Temperature Scaling (TS) (Guo et al., 2017): a post-hoc calibration method which divides the predicted logits by a temperature,
- multiple ϵ for adversarial training: which constructs adversarial examples that are bounded by randomly samples $\epsilon \sim \mathcal{U}(0, \epsilon_{max})$ as AutoLabel but assigning one-labels to the adversarial examples as standard adversarial training (Madry et al., 2017).

We apply these methods to the three data augmentation technique: AugMix (Hendrycks et al., 2020), mixup (Zhang et al., 2018) and adversarial training (Madry et al., 2017). For adversarial training based methods, we use $\epsilon = 0.01$ for “Adv. Train” and “Adv. Train + LS”, and $\epsilon_{max} = 0.01$ for “multiple ϵ ” as well as AutoLabel. The accuracy and calibration performance on clean and corrupted datasets: CIFAR10 and CIFAR100 are presented in Table 1 and Table 2.

Table 1: Ablation study of `AutoLabel` on improving model’s calibration performance. We report Accuracy and ECE on the CIFAR100 dataset and cAcc and cECE on the CIFAR100-C dataset. All numbers are in % and the best result are highlighted in **bold**.

Method	Acc	cAcc	Method	Acc	cAcc	Method	Acc	cAcc
AugMix	80.6	63.9	mixup	80.8	55.6	Adv. Train	71.5	58.1
+ LS	80.7	64.7	+ LS	-	-	+ LS	71.9	58.1
+ TS	80.6	63.9	+ TS	80.8	55.6	+ multiple ϵ	74.6	12.9
+ <code>AutoLabel</code>	81.6	65.0	+ <code>AutoLabel</code>	81.2	56.9	+ <code>AutoLabel</code>	75.3	60.2
Method	ECE	cECE	Method	ECE	cECE	Method	ECE	cECE
AugMix	5.1	11.8	mixup	1.8	11.2	Adv. Train	8.0	13.5
+ LS	2.5	6.8	+ LS	-	-	+ LS	6.4	6.5
+ TS	2.5	7.1	+ TS	3.1	13.9	+ multiple ϵ	7.0	12.9
+ <code>AutoLabel</code>	1.8	4.3	+ <code>AutoLabel</code>	1.2	10.0	+ <code>AutoLabel</code>	4.2	6.9

Table 2: Ablation study of `AutoLabel` on improving model’s calibration performance. We report Accuracy and ECE on the CIFAR10 dataset and cAcc and cECE on the CIFAR10-C dataset. All numbers are in % and the best result are highlighted in **bold**.

Method	Acc	cAcc	Method	Acc	cAcc	Method	Acc	cAcc
AugMix	96.9	87.7	mixup	96.2	81.1	Adv. Train	93.6	83.9
+ LS	96.8	88.2	+ LS	-	-	+ LS	93.1	83.6
+ TS	96.9	87.7	+ TS	96.2	81.1	+ multiple ϵ	94.3	84.5
+ <code>AutoLabel</code>	96.9	88.2	+ <code>AutoLabel</code>	96.7	81.3	+ <code>AutoLabel</code>	94.6	83.6
Method	ECE	cECE	Method	ECE	cECE	Method	ECE	cECE
AugMix	1.0	4.1	mixup	0.8	8.8	Adv. Train	3.7	10.5
+ LS	0.9	3.1	+ LS	-	-	+ LS	6.5	7.0
+ TS	0.6	2.9	+ TS	0.5	9.4	+ multiple ϵ	3.4	10.0
+ <code>AutoLabel</code>	0.9	2.7	+ <code>AutoLabel</code>	0.6	8.5	+ <code>AutoLabel</code>	2.0	6.5

C IMPLEMENTATION DETAILS

We train the vanilla models on CIFAR10, CIFAR100 and ImageNet using the open-sourced code for uncertainty baselines at <https://github.com/google/uncertainty-baselines/tree/master/baselines>.

C.1 AUGMIX

For AugMix (Hendrycks et al., 2020), the max depth of the augmentation chain is $d_{max} = 3$ for three datasets following the original work.

When applying label smoothing to AugMix, we sweep the hyperparameter ρ which decides the smoothing degree in a range $[0, 0.1]$ with a step size 0.01 and find the best $\rho = 0.01$ for CIFAR100 $\rho = 0.02$ for CIFAR100 and $\rho = 0.01$ for ImageNet.

When applying `AutoLabel` to AugMix, we set the number of distance buckets to be $d_{max} \cdot N = 3 \cdot 5 = 15$ for three datasets. The hyperparameter α in Eqn (1) is sweep in a set and we choose the best $\alpha = 0.01$ for CIFAR10 and $\alpha = 0.02$ for CIFAR100 and ImageNet.

C.2 MIXUP

When applying `AutoLabel` to mixup, we set the number of distance buckets to be $N = 5$ for three datasets. The hyperparameter α in Eqn (1) is sweep in a set and we choose the best $\alpha = 0.005$ for CIFAR10 and $\alpha = 0.008$ for CIFAR100.

C.3 ADVERSARIAL TRAINING

To improve adversarial training to be beneficial to calibration, we train all the models with ℓ_∞ norm based PGD attacks bounded by $\epsilon_{max} = 0.01$. We construct PGD attacks with 10 iterations and the step size is set to be $\epsilon/4$. When we apply label smoothing to adversarial training, we sweep the hyperparameter $\rho \in \{0.1, 0.2, 0.3\}$ and use the best $\rho = 0.1$ for both CIFAR10 and CIFAR100. When applying AutoLabel to adversarial training, we set the number of distance buckets to be $N = 10$ and the hyperparameter $\alpha = 0.5$ for CIFAR10 and $\alpha = 0.01$ for CIFAR100.

To show the adversarial robustness of all the models involving adversarial training, we train each model with $\epsilon_{max} \in \{0.01, 0.02, 0.03, 0.05, 0.1\}$ and then report the best model with the strongest adversarial robustness. In Figure 3 of the main text, we report the best performance of each model, where $\epsilon_{max} = 0.03$ for AT on both CIFAR10 and CIFAR100, $\epsilon_{max} = 0.1$ for CCAT on both CIFAR10 and CIFAR100. For AutoLabel, we set the number of distance buckets to be $N = 10$ and the hyperparameter $\alpha = 0.05$, $\epsilon_{max} = 0.01$ for CIFAR10 and $\alpha = 1$, $\epsilon_{max} = 0.1$ for CIFAR100.

REFERENCES

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.
- Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2017.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.
- Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *International Conference on Learning Representation*, 2018.