Combining Ensembles and Data Augmentations can Harm your Calibration

Yeming Wen*, Ghassen Jerfel*, Rafael Muller, Michael W. Dusenberry
Jasper Snoek, Balaji Lakshminarayanan & Dustin Tran
presented by Yeming Wen (work done as an intern at Google brain)

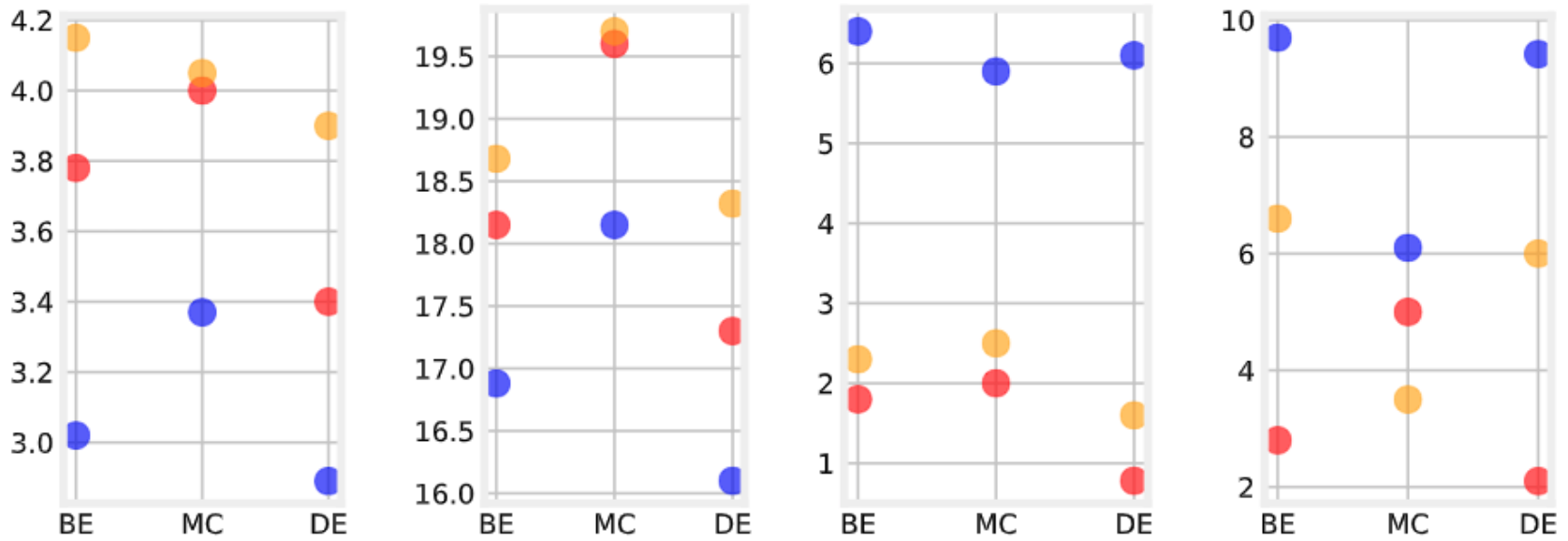
April 2, 2021

Backgroud

- Ensembles are great to improve model accuracy and calibration.
  1. Deep Ensembles
  2. MC-dropout
  3. BatchEnsemble (*Wen et al., 2020*)
- Data augmentation is an orthogonal approach to improve model accuracy and calibration.
  1. Random Corp and Random Flip
  2. AugMix (*Hendrycks et al., 2020*)
  3. Mixup (*Zhang et al., 2018*)
- Naturally, we expect ensembles + data augmentation leads to even better calibration.
- This is not true as our experiments demonstrated.

Motivating examples

- We reported the model accuracy and calibration when various ensemble methods combined with Mixup.



(a) CIFAR10-Error (%) (b) CIFAR100-Error (%) (c) CIFAR10-ECE (%) (d) CIFAR100-ECE (%)

Figure 1: WideResNet 28-10 on CIFAR-10/CIFAR-100. Red: Ensembles without Mixup; Blue: Ensembles with Mixup; Orange: Individual models in ensembles without Mixup. **(a) & (b)**: Applying Mixup to different ensemble methods leads to consistent improvement on test accuracy. **(c) & (d)**: Applying Mixup to different ensemble methods harms calibration. Averaged over 5 random seeds.

Reliability

- We plot a variant of reliability diagrams (DeGroot and Fienberg, 1983) on BatchEnsemble.
- Ensembles (underconfidence) + Mixup (underconfidence) leads to too much underconfidence.

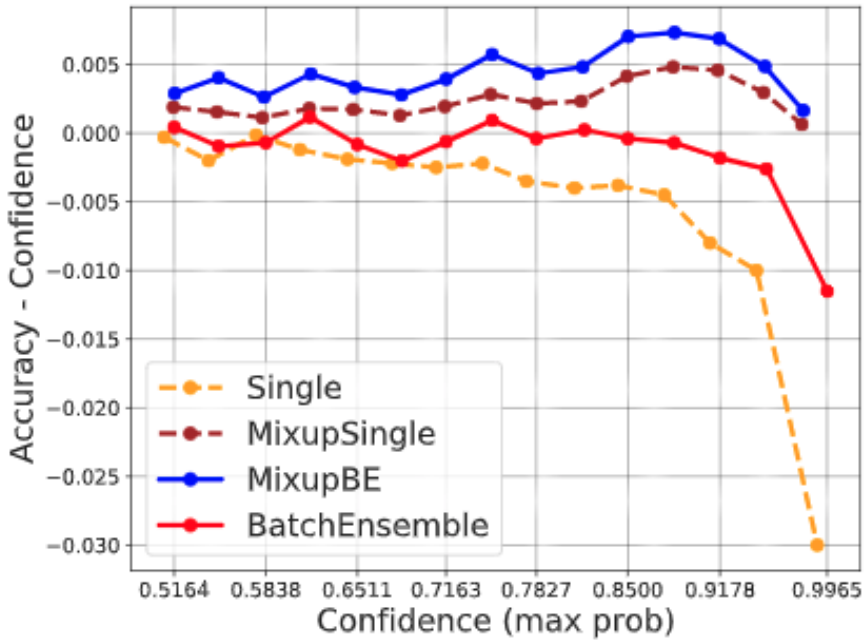


Figure 2: Reliability diagrams on CIFAR-100 with a WideResNet 28-10.

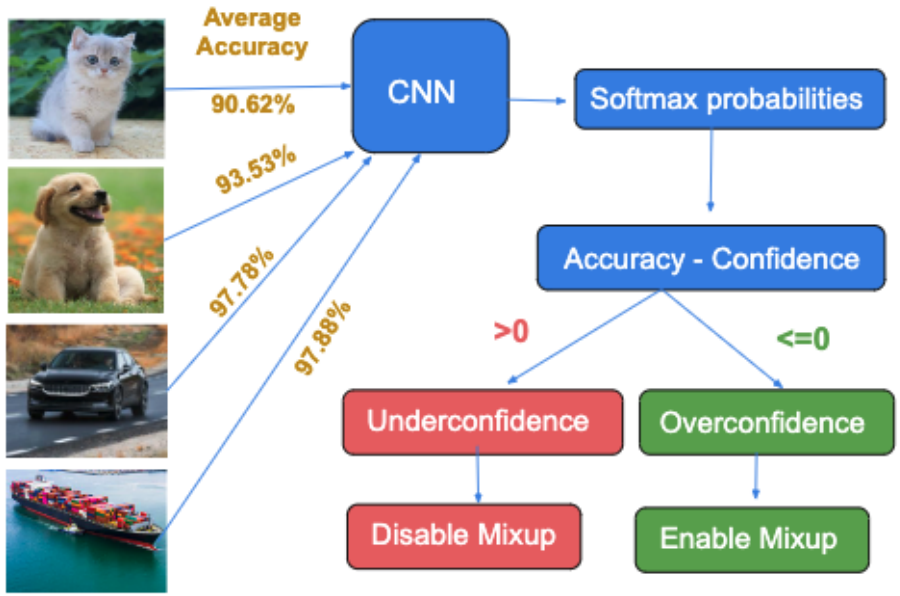
Correction

- We cannot change the underconfidence of ensembles.
- Mixup mixes all training examples.
- Ideally, we only want to mix overconfident examples.
- In this paper, we proposed to apply mixup on classes which are overconfident in the validation set.
- Denote the accuracy and confidence of class i as $\mathrm{Acc}(C_i)$ and $\mathrm{Conf}(C_i)$. We adjust Mixup's λ by the sign of $\mathrm{Acc}(C_i) - \mathrm{Conf}(C_i)$,

$$\lambda_i = \begin{cases} 0 & \mathrm{Acc}(C_i) > \mathrm{Conf}(C_i) \\ \lambda & \mathrm{Acc}(C_i) \leq \mathrm{Conf}(C_i). \end{cases} \tag{1}$$

Confidence Adjusted Mixup (CAMixup)

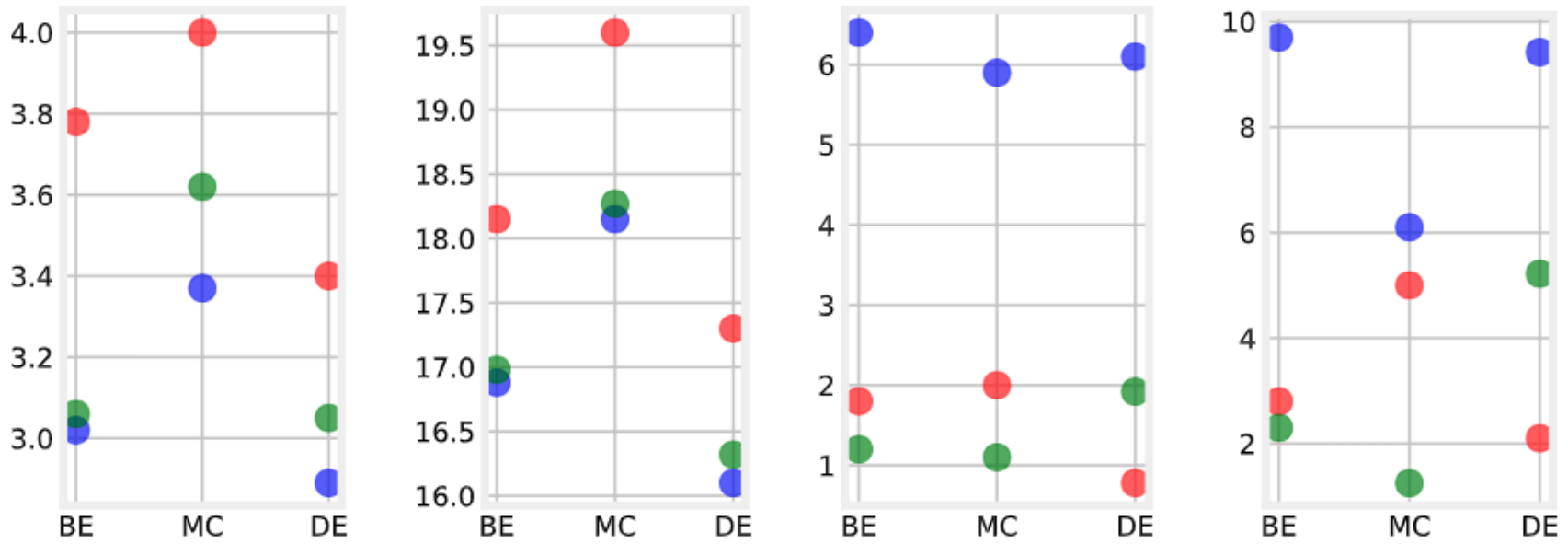
- We demonstrated class based confidence. One can also use forgetting counts (*Toneva et al., 2018*) during the training.



(a) Proposed CAMixup method.

Results

- CAMixup + Ensembles achieve better trade-off between model accuracy and model calibration.

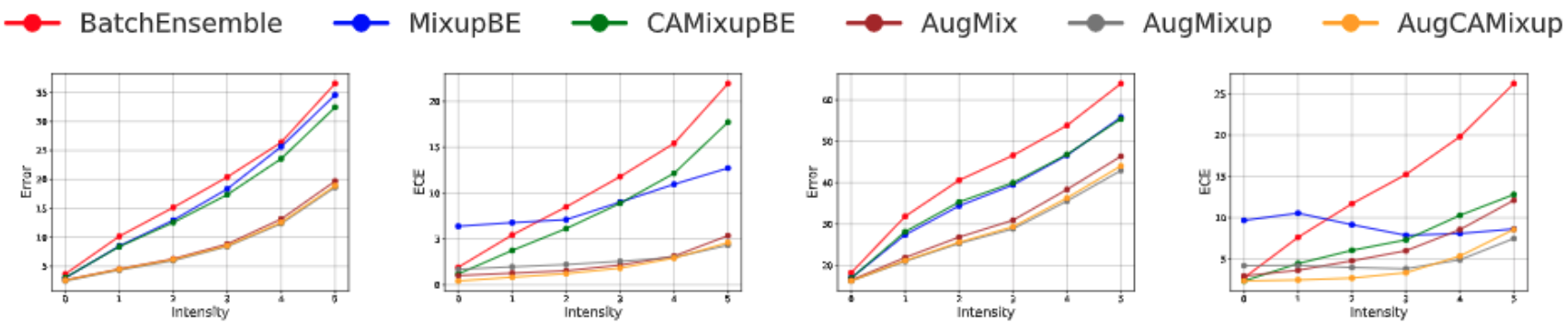


(a) CIFAR10-Error (%) (b) CIFAR100-Error (%) (c) CIFAR10-ECE (%) (d) CIFAR100-ECE (%)

Figure 5: WideResNet 28-10 on CIFAR-10/CIFAR-100. Red: Ensembles without Mixup; Blue: Ensembles with Mixup; Green: Our proposed CAMixup improves both accuracy & ECE of ensembles.

Results

- CAMixup can be combined with AugMix. And this produces the best calibrated model.



(a) CIFAR-10 Error (b) CIFAR-10 ECE (c) CIFAR-100 Error (d) CIFAR-100 ECE

Figure 7: Performance on BatchEnsemble under dataset shift. Mixup and AugMixup improve accuracy and calibration under shift but significantly worsen in-distribution calibration. Our proposed CAMixup and AugCAMixup improve accuracy and calibration.

Method/Metric	CIFAR-10			CIFAR-100		
	Acc(↑)	ECE(↓)	cA/cECE	Acc(↑)	ECE(↓)	cA/cECE
AugMix BE	97.36	1.02%	89.49/2.6%	83.57	2.96%	67.12/7.1%
AugMixup BE	**97.52**	1.71%	**90.05**/2.8%	**83.77**	4.19%	**69.26**/4.8%
AugCAMixup BE	97.47	**0.45%**	89.81/**2.4%**	83.74	**2.35%**	68.71/**4.4%**

Conclusion

- Not all data augmentation methods can be combined with ensembles to improve calibration.
- We proposed CAMixup to fix the compounding underconfidence when combining Mixup and ensembles.
- We showed that combining CAMixup and AugMix generates the best calibrated model.