1 Supplementary Materials

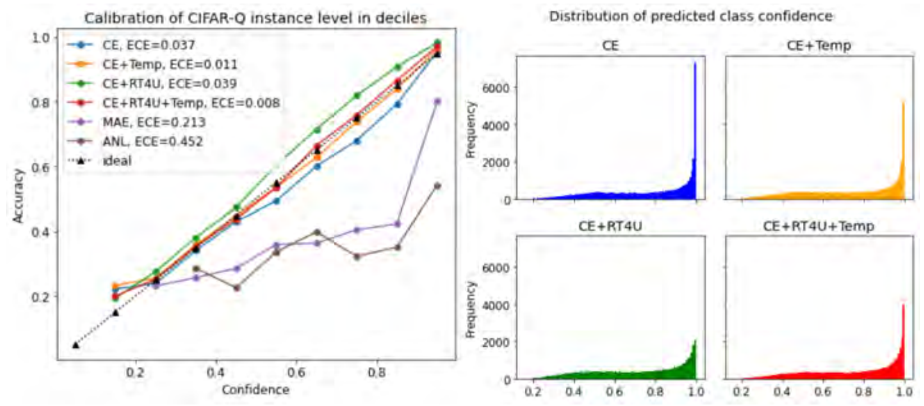

Fig. 1. Left: Calibration performance of cross-entropy variants with RT4U and temperature scaling, evaluated at the instance-level on the CIFAR-Q dataset. Right: Histograms of network confidence for the predicted class.

Table 1. Experimental settings. Values in parentheses indicate experiments with more than 1 value were conducted with the bolded value being chosen based on the validation set performance.

Dataset	CIFAR-Q	TMED-2	AS Private	
Architecture	ResNet-18	ResNet-18	R(2+1)D-18	ProtoASNet
Learning rate	1e-4	(1e-4, **7e-4**)	1e-4	(**1e-4**, 5e-4)
Batch size	256	128	32	32
# of classes	10	3	3	3
# of epochs	10	(**15**, 30)	30	100
Conformal α	0.05	0.1	0.1	0.1

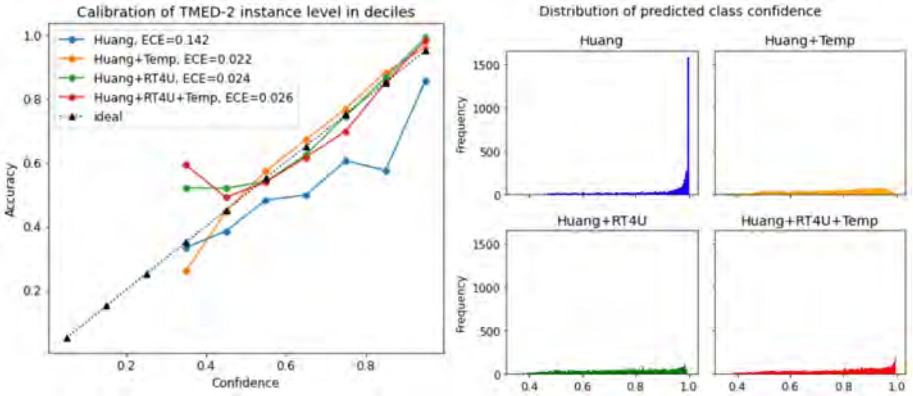

Fig. 2. Left: Calibration performance of cross-entropy variants with RT4U and temperature scaling, evaluated at the instance-level on the TMED-2 dataset. Right: Histograms of network confidence for the predicted class.

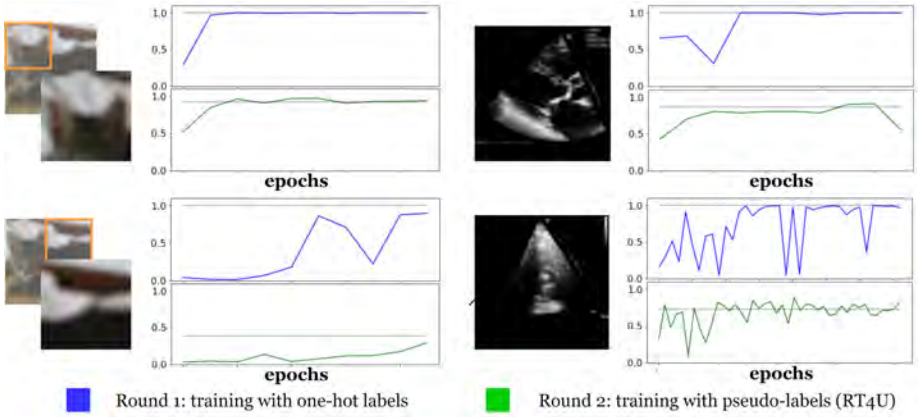

Fig. 3. More examples of the evolution of model predictions during one-hot training (blue) and R4TU training (green). Series show the confidence ($\in [0, 1]$) for the GT class as a function of epoch. One-hot training leads to overfitting in images absent of the characteristic features of their class. RT4U sets a new, non-one-hot target (illustrated by dotted line) for the network to follow.