A. Experimental Details

Units	Description
Init-conv	3×3 conv, 16
Resunit:1-0	3×3 conv, 64 3×3 conv, 64
(Resunit:1-x)$\times$ 4	3×3 conv, 64 3×3 conv, 64 $\times 4$
(Resunit:2-0)	3×3 conv, 128 3×3 conv, 128
(Resunit:2-x)$\times$ 4	3×3 conv, 128 3×3 conv, 128 $\times 4$
(Resunit:3-0)	3×3 conv, 256 3×3 conv, 256
(Resunit:3-x)$\times$ 4	3×3 conv, 256 3×3 conv, 256 $\times 4$
Average Pool	
Fully Connected - 10 logits	

Table 5. 18 unit Complex Model with 15 ResNet units.

Table 6. Residual Network Model used as the complex model for CIFAR-10 experiments in Section 4.2

Simple Model IDs	Additional Resunits	Rel. Size
SM-3	None	$\approx$ 1/5
SM-5	(Resunit:1-x)$\times$1 (Resunit:2-x)$\times$1	$\approx$ 1/3
SM-7	(Resunit:1-x)$\times$2 (Resunit:2-x)$\times$1 (Resunit:3-x)$\times$1	$\approx$ 1/2

Table 7. Additional Resnet units in the Simple Models apart from the commonly shared ones. The last column shows the approximate size of the simple models relative to the complex neural network model in the previous table.

A.1. Additional Training Details

CIFAR-10 Experiments

Complex Model Training: We trained with an ℓ-2 weight decay rate of 0.0002, sgd optimizer with Nesterov momentum (whose parameter is set to 0.9), 600 epochs and batch size 128. Learning rates are according to the following schedule: 0.1 till $40k$ training steps, 0.01 between $40k$-$60k$ training steps, 0.001 between $60k - 80k$ training steps and 0.0001 for $> 80k$ training steps. This is the standard schedule followed in the code by the Tensorflow authors[2]. We keep the learning rate schedule invariant across all our results.

[2]Code is taken from: https://github.com/tensorflow/models/tree/master/research/resnet.

Simple Models Training:

1. **Standard**: We train a simple model as is on the training set 2.

2. **ConfWeight**: We weight each sample in training set 2 by the confidence score of the last layer of the complex model on the true label. As mentioned before, this is a special case of our method, ProfWeight.

3. **Distilled-temp-t**: We train the simple model using a cross-entropy loss with soft targets. Soft targets are obtained from the softmax ouputs of the last layer of the complex model (or equivalently the last linear probe) rescaled by temperature t as in distillation of (Geoffrey Hinton, 2015). By using cross validation, we pick two temperatures that are competitive on the validation set ($t = 0.5$ and $t = 40.5$) in terms of validation accuracy for the simple models. We cross-validated over temperatures from the set $\{0.5, 3, 10.5, 20.5, 30.5, 40.5, 50\}$.

4. **ProfWeight** ($>= \ell$): Implementation of our ProfWeight algorithm where the weight of every sample in training set 2 is set to the averaged probe confidence scores of the true label of the probes corresponding to units above the ℓ-th unit. We set $\ell = 13, 14$ and 15. The rationale is that unweighted test scores of all the simple models in Table 2 are all below the probe precision of layer 16 on training set 2 but always above the probe precision at layer 12. The unweighted (i.e. Standard model) test accuracies from Table 4 can be checked against the accuracies of different probes on training set 2 given in Table 8 in the supplementary material.

5. **SRatio**: We average confidence scores from $\ell = 13, 14$ and 15 as done above for ProfWeight and divide by the simple models confidence. In each case, we optimize over β which is increased in steps of 0.5 from 1.5 to 10.

A.2. Experimental results for `Distill-proxy 2`

We provide results for the second variant of Distillation `Distill-proxy 2` in Table 9.

Probes	1	2	3	4	5	6	7	8	9
Training Set 2	0.298	0.439	0.4955	0.53855	0.5515	0.5632	0.597	0.6173	0.6418
Probes	10	11	12	13	14	15	16	17	18
Training Set 2	0.66104	0.6788	0.70855	0.7614	0.7963	0.82015	0.8259	0.84214	0.845

Table 8. Probes at various units and their accuracies on the training set 2 for the CIFAR-10 experiment. This is used in the ProfWeight algorithm to choose the unit above which confidence scores needs to be averaged.

Table 9. Below we see the averaged % errors with 95% confidence intervals for Distill-proxy 2 (regression versions of trees and SVM for the simple models that fit soft probabilities from the complex models) on the six real datasets. The results reported using Distill-proxy 1 are in the main paper and are superior to these. Boosted Trees and Random Forest (100 trees) are the complex models (CM), while a single decision tree and linear SVM are the simple models (SM).

Dataset	**Complex Model**	**CM Error**	**Simple Model**	**SM Error**	**Distill-proxy 2 Error (SM)**
Ionosphere	Boosted Trees	8.10 ±0.4	Tree	10.95 ±0.4	10.95 ±0.4
			SVM	12.38 ±0.6	12.17 ±0.3
	Random Forest	6.19 ±0.4	Tree	10.95 ±0.4	10.95 ±0.4
			SVM	12.38 ±0.6	12.38 ±0.6
Ovarian Cancer	Boosted Trees	4.68 ±0.4	Tree	15.62 ±0.8	15.62 ±0.8
			SVM	1.56 ±0.4	1.56 ±0.4
	Random Forest	6.25 ±0.8	Tree	15.62 ±0.8	15.62 ±0.8
			SVM	1.56 ±0.4	1.56 ±0.4
Heart Disease	Boosted Trees	15.55 ±0.6	Tree	23.88 ±0.7	23.69 ±0.2
			SVM	17.22 ±0.2	17.01 ±0.1
	Random Forest	15.88 ±0.6	Tree	23.88 ±0.7	23.88 ±0.7
			SVM	17.22 ±0.2	17.22 ±0.2
Waveform	Boosted Trees	12.96 ±0.1	Tree	25.43 ±0.2	25.26 ±0.1
			SVM	14.70 ±0.2	15.39 ±0.1
	Random Forest	10.90 ±0.1	Tree	25.43 ±0.2	25.43 ±0.2
			SVM	14.70 ±0.2	14.54 ±0.0
Human Activity Recognition	Boosted Trees	6.32 ±0.0	Tree	7.93 ±0.2	7.93 ±0.1
			SVM	14.56 ±0.1	16.04 ±0.2
	Random Forest	2.34 ±0.0	Tree	7.93 ±0.2	7.45 ±0.1
			SVM	14.56 ±0.1	14.23 ±0.3
Musk	Boosted Trees	4.06 ±0.1	Tree	4.49 ±0.1	6.11 ±0.1
			SVM	6.11 ±0.1	6.34 ±0.1
	Random Forest	2.45 ±0.1	Tree	4.49 ±0.1	4.49 ±0.1
			SVM	6.11 ±0.1	6.19 ±0.2