Improving Confidence Estimates for Unfamiliar Examples
Supplemental Material

Zhizhong Li, Derek Hoiem
Department of Computer Science,
University of Illinois Urbana Champaign
{zli115,dhoiem}@illinois.edu

In Sec. 1, we compare the entropy and cross-entropy (NLL) of three approaches to analyze overconfidence.

In Sec. 2, we show experimental results on a simple dataset to illustrate why ensembles perform well for unfamiliar samples and how use of unsupervised samples by G-distill can lead it to mimic the performance of the ensemble (at least in the ideal case where unsupervised samples cover a superset of the unfamiliar samples).

In Sec. 3, we show the complete table of results, mainly to simplify comparisons by any later works. Note that in the supplemental material methods without "T-scaling" in the name do not use calibration. In the main table of the paper, for brevity, only results with calibration are shown except where noted. So "Ensemble" in the main paper is "Ensemble of T-scaled models" here.

In Sec. 4, we show results on one of the tasks with DenseNet-161, supporting the same conclusions as we found based on experiments with ResNet-18. We leave a more complete exploration of depth and architecture of network to future work.

1. Entropy vs. NLL/Cross-Entropy

We thank one of the reviewers for suggesting analysis of prediction entropy, which we include in Table 3. Prediction entropy measures the uncertainty of classification and is maximized if the classifier outputs uniform probabilities for each class. NLL, equivalent to cross-entropy when using hard labels, measures the uncertainty in the correct label. When entropy is lower than cross-entropy (i.e. more confident than confidently correct), the classifier is overconfident.

	NLL		Entropy		NLL-Ent.	
Gender	**fam.**	**unf.**	**fam.**	**novel**	**fam.**	**unf.**
Single	0.083	0.542	0.036	0.089	0.047	0.453
Sin. T-scale	0.073	0.400	0.069	0.139	0.005	0.261
Ens. T-scale	0.063	0.363	0.079	0.158	-0.016	0.205
Cat vs. Dog						
Single	0.053	0.423	0.014	0.043	0.039	0.380
Sin. T-scale	0.041	0.295	0.033	0.090	0.007	0.206
Ens. T-scale	0.032	0.229	0.039	0.113	-0.007	0.116
Animals						
Single	0.326	1.128	0.146	0.263	0.180	0.864
Sin. T-scale	0.284	0.866	0.282	0.451	0.002	0.415
Ens. T-scale	0.254	0.772	0.311	0.504	-0.057	0.269

Table 3. When prediction entropy is lower than NLL (cross-entropy), the classifier is overconfident, i.e. more confident than confidently correct. We see, e.g., that single-model calibration eliminates overconfidence for familiar examples, but the calibrated ensemble achieves much further reduction in overconfidence for unfamiliar examples by increasing uncertainty and improving accuracy (lower NLL and label error).

2. Toy Experiment

Figure 6 shows results of single models, ensembles, and distillation models on simple datasets with two dimensional features. We take 1200 samples for both train and validation. The test set is densely sampled. For these experimetns, we use a 3-hidden-layer network, both layers with 1024 hidden units and Glorot initialization similar to popular deep networks, to avoid bad local minima when layer widths are too small [1]. Batchnorm [2] and then dropout [3] are applied after ReLU. The same hyperparameter tuning, initialization, and training procedures are used as described in the main paper.

3. Complete Results Table

Table 4 shows the complete table of absolute errors for all methods tested across all datasets. In the main paper, a subset of methods is shown due to space constraints (and to save the reader from being overwhelmed), with performance relative to baseline (single model) shown. This table is provided for completeness and to facilitate comparison by other methods.

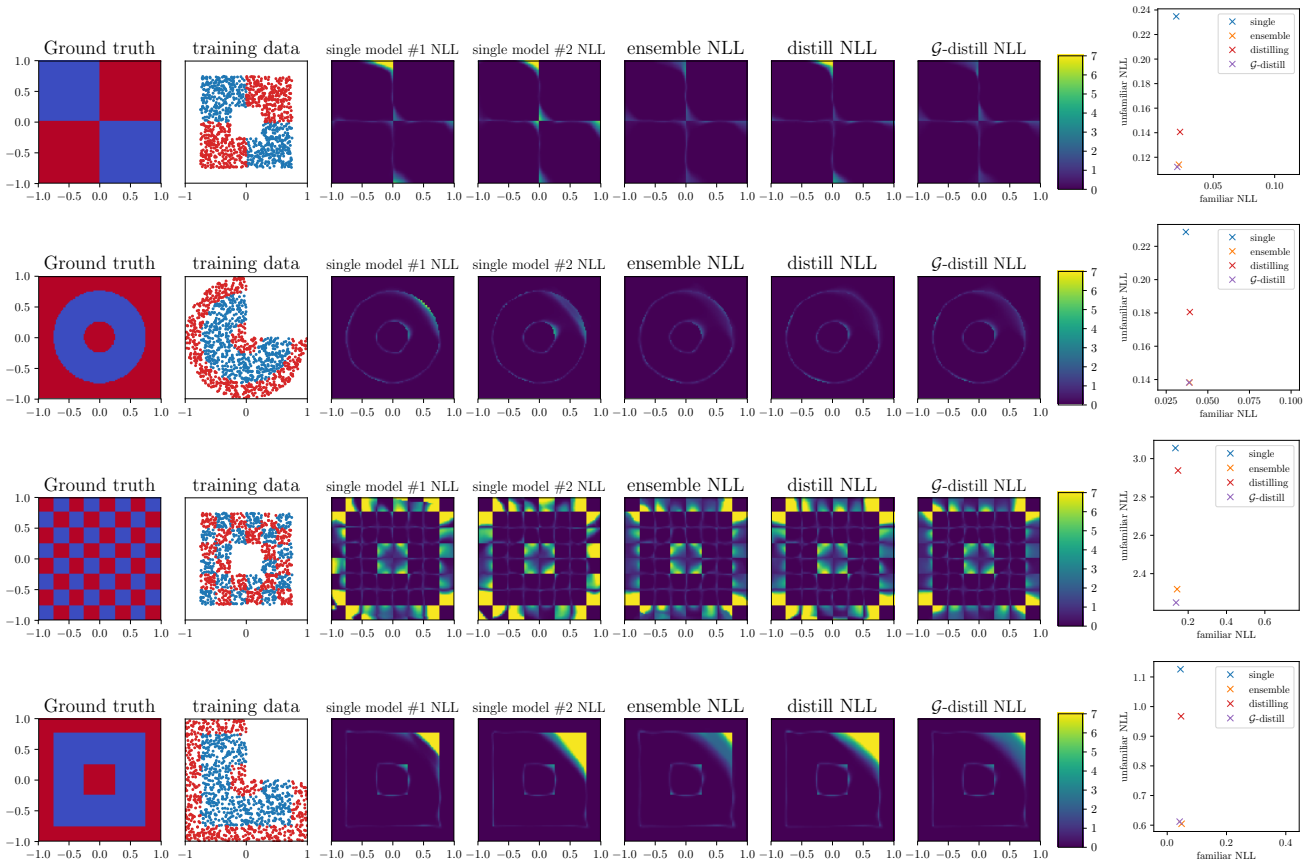

Figure 6. **Illustration on toy datasets.** Ground truth for each class are shown in red and blue. "Familiar" data is sampled from a portion of the 2D feature space. Negative log likelihood (NLL) errors are shown for two single models, ensembles, and two distillation methods. On right, average NLL for familiar and unfamiliar samples are shown. Different single models can make mistakes in different areas, while ensembles average out these differences. Distillation, when based only on familiar samples, fails to mimic the ensemble's behavior in the unfamiliar areas. G-distillation, which incorporates unsupervised unfamiliar samples, performs similarly to the ensemble but does not require multiple models at test time. In experiments on real data, however, (see main paper), G-distill underperforms the ensemble, likely because it is not possible to densely sample the unfamiliar space in practice. Figure best read in color.

4. Results on DenseNet-161

We also ran with all models on the Gender task using the DenseNet-161 architecture, as shown in Table 5. In this case Dropout was used for all layers of the network for "Bayesian". Ensemble of T-scaled Networks is still the clear leader for this architecture.

References

[1] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015. 1

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015. 1

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 1

	NLL		Brier		Label Error		ECE		E99	
Gender	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**
Single Model	0.08324	0.54208	0.14663	0.35199	0.02772	**0.14682**	0.01348	0.10902	0.00470	0.06021
Single + T-scaling	0.07348	0.39971	0.14332	**0.33715**	0.02772	**0.14682**	0.00361	0.08737	0.00192	0.02324
Ensemble	**0.06233**	0.45471	**0.13077**	0.34312	**0.02195**	**0.14714**	**0.00272**	0.09341	0.00171	0.03856
Ensemble of T-scaled models	**0.06312**	0.36266	**0.13153**	**0.33246**	**0.02170**	**0.14714**	0.00856	**0.07723**	**0.00131**	0.01003
Distill	0.07661	0.36426	0.14283	0.33963	0.02690	0.15641	0.00797	0.08596	**0.00141**	**0.00244**
Distill + T-scaling	0.07457	0.41629	0.14304	0.34873	0.02690	0.15641	0.00532	0.09974	0.00230	0.01456
G-distill	0.07268	**0.33729**	0.13885	**0.33216**	0.02519	0.15346	0.00928	**0.07535**	**0.00117**	**0.00121**
G-distill + T-scaling	0.06972	0.36859	0.13853	0.33913	0.02519	0.15346	0.00416	0.08600	0.00174	0.00424
Novelty scaling	0.07348	0.39971	0.14332	**0.33715**	0.02772	**0.14682**	0.00361	0.08737	0.00192	0.02324
Bayesian	0.08005	0.56668	0.14216	0.35391	0.02570	**0.14709**	0.01249	0.11104	0.00504	0.06768
Bayesian + T-scaling	0.06955	0.40056	0.13907	0.33765	0.02585	**0.14797**	**0.00315**	0.08884	0.00165	0.02300
Cat vs. Dog										
Single Model	0.05296	0.42285	0.11158	0.29026	0.01555	0.09518	0.00976	0.07777	0.00394	0.05251
Single + T-scaling	0.04059	0.29537	0.10686	0.27653	0.01555	0.09518	0.00356	0.05953	0.00074	0.02212
Ensemble	**0.03271**	0.28633	**0.09343**	0.26247	**0.01180**	**0.08756**	0.00248	0.05493	0.00074	0.02396
Ensemble of T-scaled models	**0.03154**	**0.22931**	**0.09252**	**0.25532**	**0.01215**	**0.08756**	**0.00202**	**0.04222**	**0.00040**	**0.01313**
Distill	0.05975	0.33184	0.12161	0.28798	0.01836	0.09937	0.00438	0.05785	0.00193	0.03610
Distill + T-scaling	0.06411	0.41595	0.12309	0.29453	0.01836	0.09937	0.00860	0.07354	0.00490	0.05314
G-distill	0.06232	0.30747	0.12732	0.28745	0.02065	0.10254	0.00577	0.05390	0.00123	0.02163
G-distill + T-scaling	0.06509	0.37850	0.12892	0.29445	0.02065	0.10254	0.00857	0.07182	0.00314	0.04420
Novelty scaling	0.04024	0.29713	0.10635	0.27432	0.01555	0.09518	**0.00255**	0.05662	0.00081	0.02396
Bayesian	0.05551	0.41758	0.11221	0.28485	0.01541	0.09264	0.00986	0.07454	0.00444	0.05306
Bayesian + T-scaling	0.04381	0.31260	0.10826	0.27497	0.01558	0.09264	0.00563	0.06152	0.00173	0.02825
Animals										
Single Model	0.32575	1.12785	0.19922	0.34062	0.10375	0.29056	0.04807	0.18714	0.01339	0.08701
Single + T-scaling	0.28425	0.86575	0.19386	0.32398	0.10375	0.29056	**0.01208**	0.11751	**0.00219**	0.02567
Ensemble	0.25623	0.92980	**0.18108**	0.32221	0.09437	**0.27563**	0.02236	0.13766	0.00521	0.04509
Ensemble of T-scaled models	**0.25377**	**0.77222**	**0.18149**	**0.31193**	**0.09250**	**0.27438**	0.02408	**0.07979**	**0.00174**	**0.01329**
Distill	0.30180	0.92112	0.19639	0.32732	0.10450	0.29000	**0.01329**	0.12952	0.00650	0.04228
Distill + T-scaling	0.30167	0.86280	0.19657	0.32303	0.10450	0.29000	0.01646	0.10353	0.00481	0.03457
G-distill	0.27929	0.86841	0.18950	0.32109	0.09644	0.28444	0.01489	0.11629	0.00651	0.03399
G-distill + T-scaling	0.28152	0.82950	0.19011	0.31806	0.09644	0.28444	0.02105	0.09629	0.00302	0.03354
Novelty scaling	0.28425	0.86575	0.19386	0.32398	0.10375	0.29056	**0.01208**	0.11751	**0.00219**	0.02567
Bayesian	0.30986	1.12297	0.19370	0.33759	0.09906	0.28694	0.04440	0.18238	0.01439	0.09471
Bayesian + T-scaling	0.27239	0.86154	0.18905	0.32226	0.09863	0.28694	**0.01437**	0.11515	**0.00290**	0.03181
Objects										
Single Model	0.08597	0.12815	0.15392	0.18553	0.19494	0.45523	0.00475	0.01021	0.00222	0.00546
Single + T-scaling	0.08589	0.12780	0.15388	0.18550	0.19494	0.45523	0.00466	0.01003	0.00207	0.00519
Ensemble	**0.08222**	0.12292	**0.15063**	0.18207	**0.18298**	**0.44095**	**0.00435**	0.00950	0.00179	0.00459
Ensemble of T-scaled models	**0.08227**	0.12274	**0.15063**	0.18207	**0.18299**	**0.44095**	0.00459	0.00953	0.00171	0.00437
Distill	0.08658	**0.12165**	0.15437	**0.18166**	0.19308	0.45322	0.00624	**0.00918**	**0.00122**	0.00325
Distill + T-scaling	0.08583	0.12218	0.15421	**0.18160**	0.19308	0.45322	0.00450	**0.00903**	0.00191	0.00456
G-distill	0.08736	0.12196	0.15527	0.18188	0.19822	0.45861	0.00670	0.00951	**0.00119**	**0.00315**
G-distill + T-scaling	0.08661	0.12229	0.15511	**0.18180**	0.19822	0.45861	0.00485	**0.00905**	0.00187	0.00441
Novelty scaling	0.08582	0.12809	0.15385	0.18561	0.19452	0.45573	0.00460	0.01007	0.00205	0.00516
Bayesian	0.08597	0.12884	0.15359	0.18577	0.19440	0.45674	0.00474	0.01046	0.00254	0.00581
Bayesian + T-scaling	0.08580	0.12789	0.15356	0.18569	0.19444	0.45686	0.00460	0.01008	0.00211	0.00504

Table 4. Errors of all tested methods across all datasets. Bold numbers are best or not significantly different than the best.

	NLL		Brier		Label Error		ECE		E99	
Gender	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**	**familiar**	**unfam.**
Single Model	0.0769	0.5608	0.1332	0.3499	0.0219	0.1430	0.0132	0.1139	0.0063	0.0658
Single + T-scaling	0.0611	0.3553	0.1291	0.3262	0.0219	0.1430	**0.0024**	0.0850	0.0015	0.0131
Ensemble	**0.0513**	0.4103	**0.1163**	0.3281	**0.0180**	**0.1342**	**0.0031**	0.0861	0.0022	0.0348
Ensemble of T-scaled models	**0.0507**	**0.2995**	**0.1165**	**0.3116**	**0.0185**	**0.1338**	0.0070	**0.0653**	**0.0003**	**0.0023**
Distill	0.0706	0.4117	0.1321	0.3347	0.0211	**0.1352**	0.0081	0.0951	0.0039	0.0245
Distill + T-scaling	0.0694	0.3947	0.1317	0.3326	0.0211	**0.1352**	0.0067	0.0920	0.0034	0.0186
G-distill	0.0645	0.3559	0.1265	0.3250	0.0196	0.1391	0.0049	0.0815	0.0030	0.0138
G-distill + T-scaling	0.0641	0.3477	0.1263	0.3235	0.0196	0.1391	0.0041	0.0791	0.0026	0.0118
Novelty scaling	0.0611	0.3553	0.1291	0.3262	0.0219	0.1430	**0.0024**	0.0850	0.0015	0.0131
Bayesian	0.0795	0.5930	0.1341	0.3512	0.0218	0.1416	0.0139	0.1155	0.0070	0.0738
Bayesian + T-scaling	0.0617	0.3934	0.1295	0.3324	0.0217	0.1412	0.0050	0.0932	0.0018	0.0206

Table 5. Performance for DenseNet-161 classifier. Best and within significance range of best is in bold.