A Tables of p-values for the hypothesis tests

Table 1: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_{t_i}^M] \geq \mathbb{E}[\alpha_{t_{i+1}}^M]$ vs. $H_a : \mathbb{E}[\alpha_{t_i}^M] < \mathbb{E}[\alpha_{t_{i+1}}^M]$ within each architecture M, for $i = 1, 2, 3, 4$ using CIFAR-10 and MP, Bonferroni-corrected by column.

Ratios	ResNet-20	ResNet-32	ResNet-44	ResNet-56	ResNet-110
2 vs 4	<0.001	0.001	0.708	0.036	0.079
4 vs 10	<0.001	<0.001	<0.001	<0.001	<0.001
10 vs 20	<0.001	<0.001	<0.001	<0.001	<0.001
20 vs 50	<0.001	<0.001	<0.001	<0.001	<0.001

Table 2: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_{t_i}^M] \geq \mathbb{E}[\alpha_{t_{i+1}}^M]$ vs. $H_a : \mathbb{E}[\alpha_{t_i}^M] < \mathbb{E}[\alpha_{t_{i+1}}^M]$ within each architecture M, for $i = 1, 2, 3, 4$ using CIFAR-10 and UP, Bonferroni-corrected by column.

Ratios	ResNet-20	ResNet-32	ResNet-44	ResNet-56	ResNet-110
2 vs 4	0.029	0.015	0.002	0.885	0.100
4 vs 10	<0.001	<0.001	<0.001	<0.001	<0.001
10 vs 20	<0.001	<0.001	<0.001	<0.001	<0.001
20 vs 50	<0.001	<0.001	<0.001	<0.001	<0.001

Table 3: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_{t_i}^D] \geq \mathbb{E}[\alpha_{t_{i+1}}^D]$ vs. $H_a : \mathbb{E}[\alpha_{t_i}^D] < \mathbb{E}[\alpha_{t_{i+1}}^D]$ within each dataset D, for $i = 1, 2, 3, 4$ using MP, Bonferroni-corrected by column.

Ratios	MNIST	Fashion	CIFAR-10	CIFAR-100
2 vs 4	0.389	1.000	0.036	<0.001
4 vs 10	1.000	0.700	<0.001	<0.001
10 vs 20	0.071	0.001	<0.001	<0.001
20 vs 50	<0.001	<0.001	<0.001	<0.001

Table 4: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_{t_i}^P] \geq \mathbb{E}[\alpha_{t_{i+1}}^P]$ vs. $H_a : \mathbb{E}[\alpha_{t_i}^P] < \mathbb{E}[\alpha_{t_{i+1}}^P]$ within each pruning algorithm P, for $i = 1, 2, 3, 4$ on CIFAR-10 at ResNet-56, Bonferroni-corrected by column.

Ratios	MP	GP	UP	RP
2 vs 4	0.036	<0.001	0.885	0.034
4 vs 10	<0.001	<0.001	<0.001	1.000
10 vs 20	<0.001	<0.001	<0.001	1.000
20 vs 50	<0.001	<0.001	<0.001	1.000

Table 5: p-values for independent-samples t-tests of $H_0 : \mathbb{E}[\alpha_t^{M_i}] \leq E[\alpha_t^{M_{i+1}}]$ vs. $H_a : \mathbb{E}[\alpha_t^{M_i}] > E[\alpha_t^{M_{i+1}}]$ within each ratio t, for $i = 1, 2, 3, 4$ using CIFAR-10 and MP, Bonferroni-corrected by column.

ResNet Sizes	$t = 2$	$t = 4$	$t = 10$	$t = 20$	$t = 50$
20 vs 32	1.000	0.662	0.021	0.002	<0.001
32 vs 44	1.000	0.291	0.125	<0.001	<0.001
44 vs 56	0.014	0.147	0.002	<0.001	<0.001
56 vs 110	1.000	1.000	0.410	0.121	0.608

Table 6: p-values for independent-samples t-tests of $H_0 : \mathbb{E}[\alpha_t^{M_i}] \leq E[\alpha_t^{M_{i+1}}]$ vs. $H_a : \mathbb{E}[\alpha_t^{M_i}] > E[\alpha_t^{M_{i+1}}]$ within each ratio t, for $i = 1, 2, 3, 4$ using CIFAR-10 and UP, Bonferroni-corrected by column.

ResNet Sizes	$t = 2$	$t = 4$	$t = 10$	$t = 20$	$t = 50$
20 vs 32	1.000	1.000	0.005	<0.001	<0.001
32 vs 44	0.063	0.057	0.085	0.005	<0.001
44 vs 56	1.000	0.484	0.005	0.004	<0.001
56 vs 110	0.007	0.275	0.153	<0.001	<0.001

Table 7: p-values for independent-samples t-tests of $H_0 : \mathbb{E}[\alpha_t^{D_i}] \geq E[\alpha_t^{D_j}]$ vs. $H_a : \mathbb{E}[\alpha_t^{D_i}] < E[\alpha_t^{D_j}]$ within each ratio t, for three dataset pairs (D_i, D_j) using MP, Bonferroni-corrected by column.

Datasets	$t = 2$	$t = 4$	$t = 10$	$t = 20$	$t = 50$
MNIST vs CIFAR-10	<0.001	<0.001	<0.001	<0.001	1.000
Fashion vs CIFAR-10	1.000	1.000	<0.001	<0.001	1.000
CIFAR-10 vs CIFAR-100	0.380	0.500	0.059	0.001	0.025

Table 8: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_t^{P_i}] = E[\alpha_t^{P_j}]$ vs. $H_a : \mathbb{E}[\alpha_t^{P_i}] \neq E[\alpha_t^{P_j}]$ within each ratio t, for all algorithm pairs (P_i, P_j) on CIFAR-10 and ResNet-56, Bonferroni-corrected by column.

Methods	$t = 2$	$t = 4$	$t = 10$	$t = 20$	$t = 50$
MP vs GP	1.000	0.286	0.003	<0.001	0.037
MP vs UP	1.000	0.664	0.004	0.004	<0.001
GP vs UP	1.000	0.003	<0.001	<0.001	<0.001
MP vs RP	<0.001	<0.001	0.001	0.001	0.003
GP vs RP	<0.001	<0.001	<0.001	0.001	0.002
UP vs RP	<0.001	<0.001	0.001	0.002	0.008

Table 9: p-values for paired-samples t-tests of $H_0 : \mathbb{E}[\alpha_{t,P_i}^M] = E[\alpha_{t,P_j}^M]$ vs. $H_a : \mathbb{E}[\alpha_{t,P_i}^M] \neq E[\alpha_{t,P_j}^M]$ within each rate t and architecture M for CIFAR-10, always comparing only algorithm pairs $(P_i, P_j) = (\text{MP}, \text{UP})$. Unlike the other tables, these p-values are Bonferroni-corrected for all 25 comparisons at once.

Rate	ResNet-20	ResNet-32	ResNet-44	ResNet-56	ResNet-110
2	1.000	1.000	0.001	1.000	0.011
4	0.097	1.000	0.526	1.000	0.015
10	0.125	0.098	0.038	0.010	0.005
20	<0.001	<0.001	0.001	0.015	<0.001
50	0.003	0.051	0.001	<0.001	<0.001

B Additional scatterplots

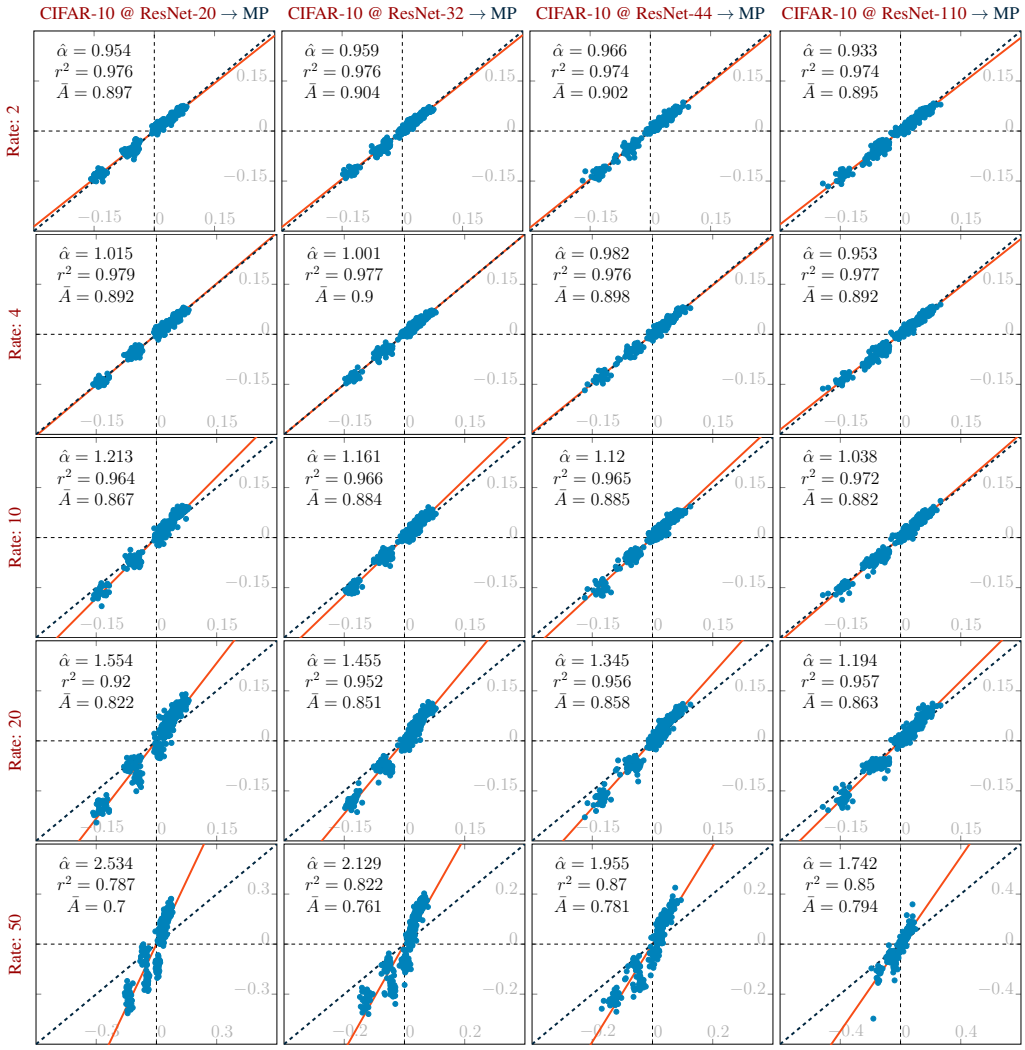

Figure 6: Scatterplot matrix of $\bar{B}^c(m)$ (x-axis) vs $\bar{B}^c_t(m)$ (y-axis), at several values of t (rows) and M (columns) for $P =$ MP. Each scatterplot point corresponds to one c for one m. See Section 6.2.

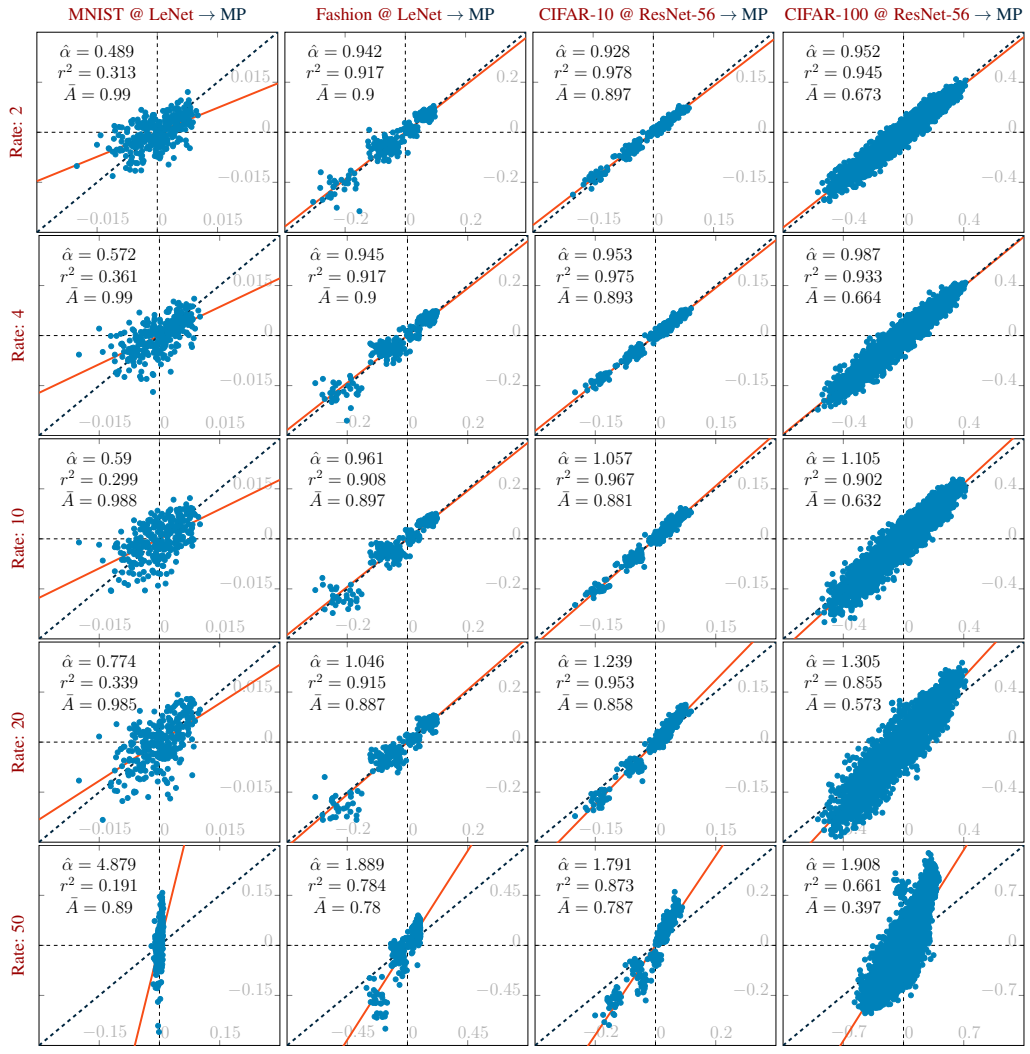

Figure 7: Scatterplot matrix of $\bar{B}^c(m)$ (x-axis) vs $\bar{B}^c_t(m)$ (y-axis), at several values of t (rows) and D (columns). Each scatterplot point corresponds to one c for one m. See Section 6.2.

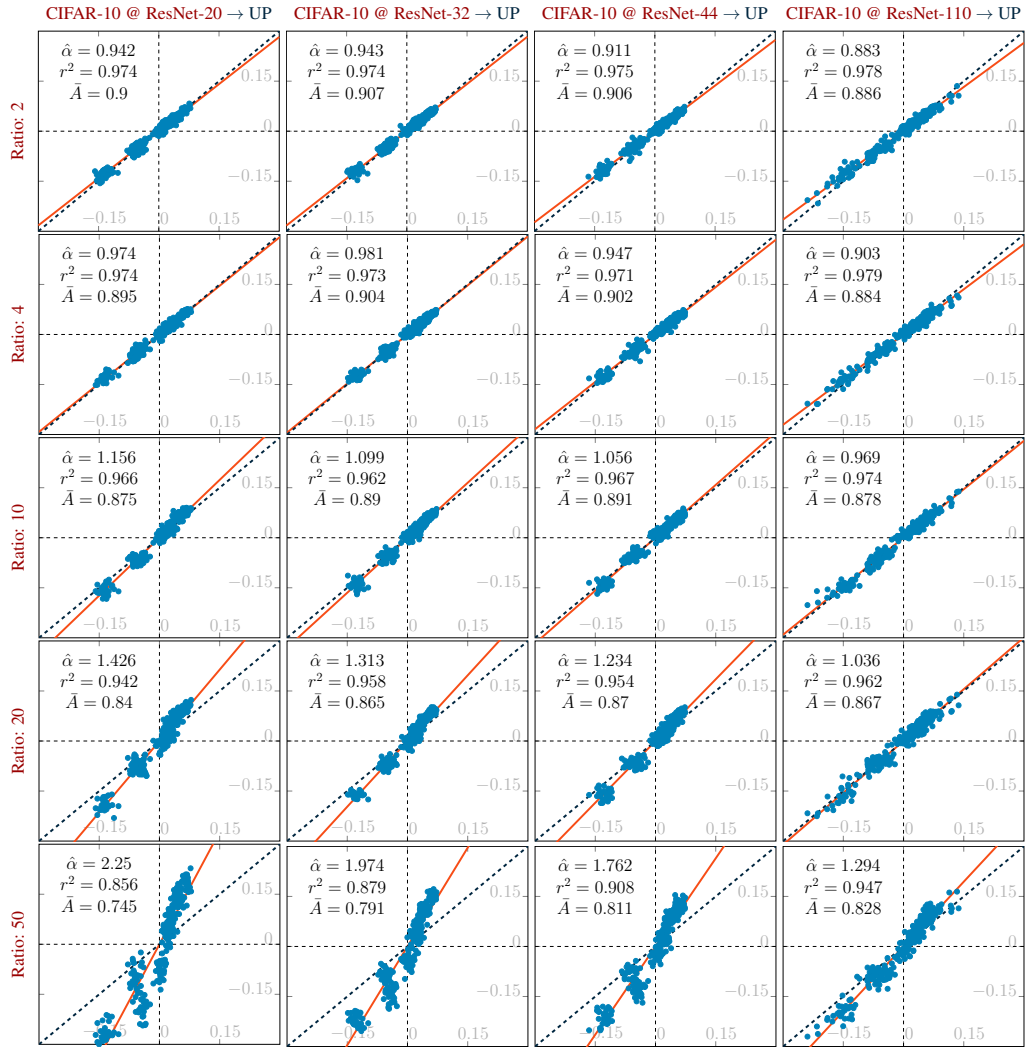

Figure 8: Scatterplot matrix of $\bar{B}^c(m)$ (x-axis) vs $\bar{B}^c_t(m)$ (y-axis), at several values of t (rows) and M (columns) for $P =$ UP. Each scatterplot point corresponds to one c for one m. See Section 6.2.

C Confidence intervals for each boxplot

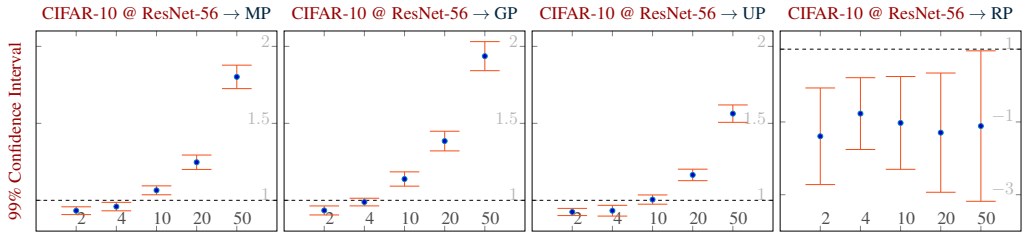

Figure 9: 99% confidence intervals for $\alpha_{t,P}^{D,M}$ at each t within each P associated with Figure 1.

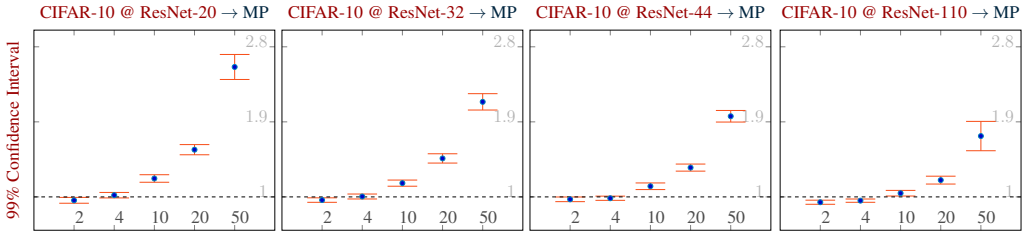

Figure 10: 99% confidence intervals for $\alpha_{t,P}^{D,M}$ at each t within each M for P = MP associated with Figure 2.

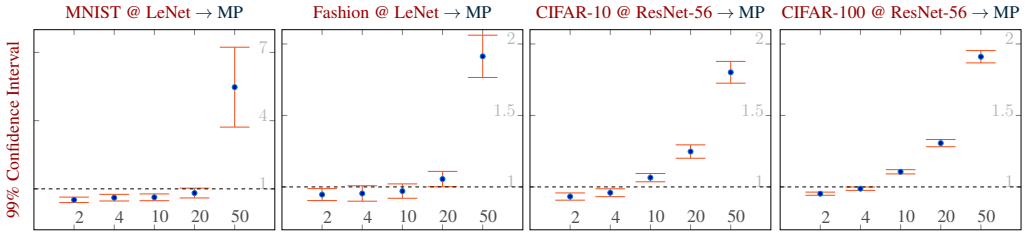

Figure 11: 99% confidence intervals for $\alpha_{t,P}^{D,M}$ at each t within each D associated with Figure 3.

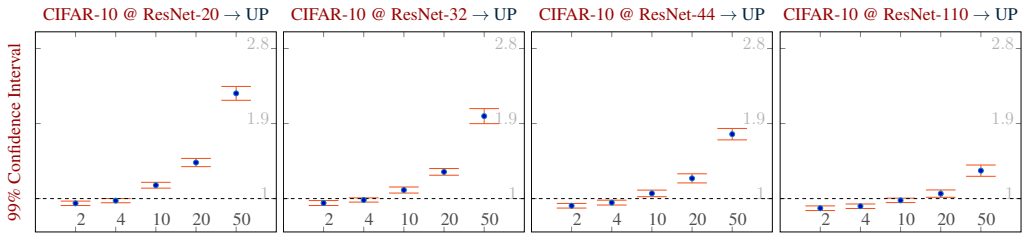

Figure 12: 99% confidence intervals for $\alpha_{t,P}^{D,M}$ at each t within each M for P=UP associated with Figure 4.

D Mean accuracy before training

Table 10: Mean accuracy before pruning of the models used for each set of experiments.

Model	Dataset	Pruning Algorithm	Accuracy
LeNet	MNIST	MP	0.989
LeNet	Fashion	MP	0.899
ResNet-20	CIFAR-10	MP	0.896
ResNet-20	CIFAR-10	UP	0.896
ResNet-32	CIFAR-10	MP	0.903
ResNet-32	CIFAR-10	UP	0.903
ResNet-44	CIFAR-10	MP	0.900
ResNet-44	CIFAR-10	UP	0.901
ResNet-56	CIFAR-10	MP	0.893
ResNet-56	CIFAR-10	GP	0.896
ResNet-56	CIFAR-10	UP	0.893
ResNet-56	CIFAR-10	RP	0.893
ResNet-56	CIFAR-100	MP	0.671
ResNet-56	CIFAR-100	UP	0.670
ResNet-110	CIFAR-10	MP	0.889
ResNet-110	CIFAR-10	UP	0.876

E Tradeoff between recall distortion and accuracy

The supplemental results in this section illustrate the tradeoff between recall distortion and accuracy, which are illustrated in Figure 13. We refine our study in lower pruning ratios by evaluating 30 models at ratios 2, 4, 6, 8, and 10. In those plots, the blue curve associated with the left y-axis represents the mean accuracy at each compression ratio, with the initial observation at compression ratio 1 corresponding to the model accuracy before pruning. The orange curve associated with the right y-axis represents the mean intensification at each compression ratio. We have aligned both y-axes so that the center of the left axis represents the mean accuracy before pruning and the right axis represents an intensification ratio of 1, and then a dashed horizontal line is drawn at the center of the plot. Whenever we see the blue plot above that line and the orange plot below that line, which is typical for the lower pruning ratios, we are observing accuracy going up while intensification is going down.

We observe some agreement between recall distortion and accuracy for the pruning ratios at which both are more beneficial. More specifically, we observe model accuracy improving at the same time that intensification is reduced if a small pruning ratios are used. These plots suggest that intensification could be another axis along which pruning methods should be evaluated. When the pruning ratio increases, model accuracy and intensification no longer move in the same direction. Whereas heavy pruning makes the accuracy worse and intensification stronger, lighter pruning makes the accuracy better while not pushing intensification above 1.

We believe that these results are actionable to the extent that they encourage the use of a moderate amount of pruning both for the sake of improving generalization as well as to reduce the performance balance across classes. Moreover, these results provide a qualitative understanding of how to adjust for different cases. Namely, if we work with a comparatively more complex task we should compensate with a larger model or a lower pruning ratio if we would like to obtain a pruned model with similar improvements for both metrics.

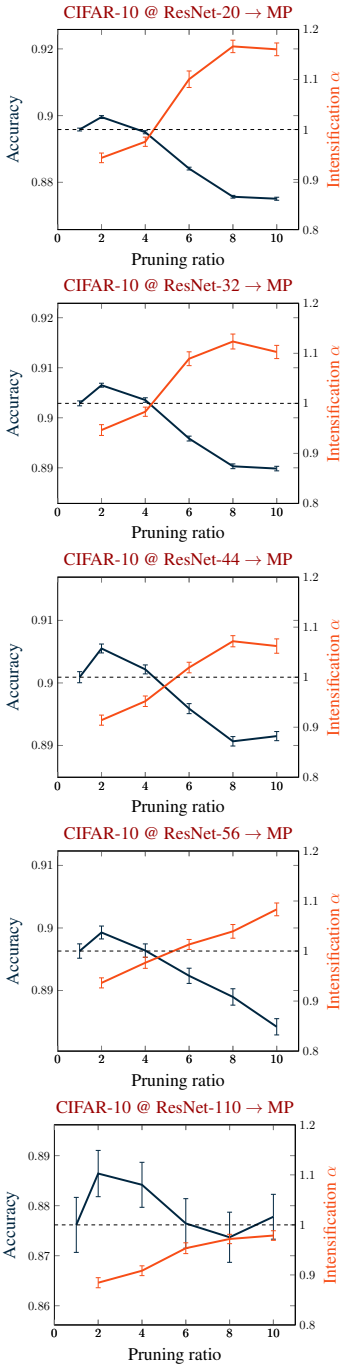

Figure 13: Comparison between mean accuracy and mean intensification along with their standard errors for 30 models trained on CIFAR-10 and pruned with pruning ratios 2, 4, 6, 8, and 10.

F Variance of recall balance

Table 11 reports mean recall variance on each model before and after pruning for each type of model and pruning ratio used in our study. All the cases in which the recall variance after pruning is greater than before pruning are in bold. That probably confirms the intuition of one of the anonymous reviewers that recall variance increases with pruning. We would emphasize, however, that variance alone is not a clear indicator if recall differences increase or decrease.

Table 11: Mean recall variance for each model and pruning ratio considered.

Model	Before	Ratio 2	Ratio 4	Ratio 10	Ratio 20	Ratio 50
CIFAR-10 @ ResNet-20 → MP	0.0036	**0.0036**	**0.0039**	**0.0057**	**0.0097**	**0.0318**
CIFAR-10 @ ResNet-32 → MP	0.0032	0.0031	**0.0033**	**0.0045**	**0.0074**	**0.0189**
CIFAR-10 @ ResNet-44 → MP	0.0038	0.0035	0.0037	**0.0048**	**0.0068**	**0.0153**
CIFAR-10 @ ResNet-110 → MP	0.0039	0.0036	0.0037	**0.0045**	**0.0061**	**0.0135**
CIFAR-100 @ ResNet-56 → MP	0.0004	0.0004	0.0004	**0.0005**	**0.0009**	**0.0021**
Fashion @ LeNet → MP	0.0099	0.0098	0.0088	0.0088	**0.0115**	**0.0416**
MNIST @ LeNet → MP	0.0038	**0.0038**	**0.0038**	**0.0038**	**0.0039**	**0.0111**
CIFAR-10 @ ResNet-56 → MP	0.0038	0.0034	0.0036	**0.0045**	**0.0062**	**0.0142**
CIFAR-10 @ ResNet-56 → GP	0.0039	0.0033	0.0037	**0.0048**	**0.0072**	**0.0149**
CIFAR-10 @ ResNet-56 → UP	0.004	0.0034	0.0033	**0.004**	**0.0056**	**0.0104**
CIFAR-10 @ ResNet-56 → RP	0.0038	**0.1504**	**0.132**	**0.1413**	**0.1487**	**0.227**

G On the use of α instead of I to measure intensification

One concern with using α is that we do not attribute the same weight to large variations around the origin. However, we note that across all scatterplots in the experimental data that we collected, we did not observe any behavior around the origin that would differ substantially from the linear trend. Hence, the slope (as measured by α) seems to be an appropriate summary of trends seen in our plots. On the other hand, points near the origin correspond to intensification ratios whose denominators are near zero, and hence on the scale of y/x ratios they are often volatile outliers—even though on our scatterplots their (x,y) pairs are not outliers. Since the equally-weighted mean of intensification ratios is not robust to outliers, it is not an appropriate summary of trends seen in our data.

To illustrate that, consider the results for CIFAR-10 @ ResNet-32 with MP at rate 20 in Figure 6. We picked this plot because it has many points concentrated around the origin while presenting a clearly linear behavior. The value of α reported in the plot (1.455) corresponds to a linear regression using data from all models and classes. If we calculate for each of the 30 models separately, we obtain 1.5 ± 0.1 with a minimum of 1.3 and a maximum of 1.7. In turn, if we calculate the mean of the intensifications averaging all classes for each model, we obtain 1.4 ± 1.4 with a minimum of -5.0 and a maximum of 4.5.

For the model that yields the minimum of -5.0, there is a single outlier intensification ratio, corresponding to the pair of (x,y) values (0.0004, -0.0285). For the class associated with these values, the recall before pruning is 90.3% and the model accuracy is 90.26%, hence implying that the model overperforms for this class. Since the test set has 1,000 samples for each class, it would take only one more sample being incorrect for the model to underperform for this class. However, this class alone contributes with an intensification of -64, which would have been positive but similarly large in absolute value if one more test sample were incorrect before pruning.

For the model that yields the maximum of 4.5, there is a single outlier intensification ratio, corresponding to the pair of (x,y) values (-0.0003, -0.0099). The recall for this class before pruning is again 90.3% and the model accuracy is 90.33%, hence implying that the model slightly underperforms for this class. Although the intensification in this case is 30, we note that the normalized recall balance after pruning remains the smallest across all classes. Furthermore, it would take only one more sample being incorrect for the model to overperform for this class, in which case the intensification would be negative but again similarly large in absolute value.

In other words, we believe that the value of α represents a more consistent and representative characterization of the intensification effect of pruning on recall balance, since it reflects the consistent

trends across models shown in our scatterplots. It is true that α gives less weight to outlier cases corresponding to classes that had recall very close to the model accuracy before pruning, but we believe this is reasonable because those classes have unstable ratios due to their denominators being near zero. Moreover, since their recalls are so close to the accuracy, it does not seem as appropriate to attribute such changes to intensification.

Another way to think about this is that for the classes in such a situation, the model is only narrowly over- or underperforming, which means that the intensification ratio for the class is not as informative for our purposes. We agree that α is not the only way to aggregate information across classes, but we wish to emphasize that intensification ratios are meant to help us think about questions such as "If a recall-balance is already non-negligible, when does pruning push it even farther away in the same direction?" This is distinct from asking general questions about variability, such as "When does pruning make small recall-balances more variable?"

H Experiments on modern pruning methods

We studied in more detail the effect of the intensification in the recent pruning methods LTH (Lottery Ticket Hypothesis) [17] and CHIP (CHannel Independence-based Pruning) [65] on ResNet-56. In both cases, we observe that the intensification ratio ultimately increases with the pruning ratio and that an intensification above 1 consistently occurs if the pruning ratio exceeds a certain threshold, which depends on the method. In order to reach that threshold, we have experimented with higher pruning ratios than those reported in the papers describing those methods.

We also note that it is not straightforward to adapt multiple methods to successfully operate on exactly the same pruning ratios, since there is a lot of engineering in making sophisticated methods work well. For example, to decide the amount pruned on each iteration with LTH or to decide the amount pruned from each layer with CHIP. For that reason, we emphasize once more our belief that studying classic methods makes it easier to isolate different factors that may influence intensification, as we did in our study. Nevertheless, we appreciate the recommendations by the reviewers to consider other methods, and we believe that the results that we obtained for those methods endorse the main message of our work about how intensification operates at lower and higher pruning ratios.

Table 12 summarizes the outcome of 10 runs of LTH on models trained on the same setting as those of our other experiments using 30,000 steps for training at each level of pruning, with the corresponding pruning ratio next to it. The level 0 (pruning ratio 1) corresponds to the original model without pruning. The LTH paper [17] only goes as far as step 15. We extend the number of steps using the same pruning ratio between steps used up to step 15.

Even if restricted to the first 15 steps, we already observe intensification by step 12. The increase in intensification is consistent since step 6, which is in line with our findings using classic methods.

Table 13 summarize the outcome of 15 runs of CHIP on models trained on the same setting as those of our other experiments, with the corresponding pruning ratio next to it. Pruning ratio 1 corresponds to the original model without pruning. The CHIP paper [65] only goes as far as pruning ratio 3.33. We extend the number of pruning ratios with ratios 8.27 and 19.11 by preserving the proportion of unpruned weights used for pruning ratio 3.33.

If restricted to pruning ratios 1.75 and 3.33, we observe that intensification starts to increase from one pruning ratio to another, which is in line with our findings.

Table 12: Results for accuracy and intensification at each level and corresponding pruning ratio when using the LTH method [17] on 10 ResNet-56 models trained on CIFAR-10.

Level	Pruning ratio	Accuracy	Intensification
0	1.00	0.855	–
1	1.25	0.854	1.010
2	1.56	0.859	0.986
3	1.95	0.859	0.970
4	2.44	0.859	0.974
5	3.05	0.859	0.970
6	3.81	0.859	0.932
7	4.77	0.858	0.957
8	5.96	0.856	0.973
9	7.45	0.853	0.982
10	9.31	0.850	0.980
11	11.64	0.846	0.997
12	14.55	0.841	1.029
13	18.19	0.835	1.055
14	22.74	0.826	1.131
15	28.42	0.811	1.244
16	35.53	0.795	1.293
17	44.41	0.769	1.377
18	55.51	0.739	1.536
19	69.39	0.694	1.767
20	86.74	0.669	1.910

Table 13: Results for accuracy and intensification at each pruning ratio, including extrapolated steps, when using the CHIP method [65] on 15 ResNet-56 models trained on CIFAR-10.

Pruning ratio	Accuracy	Intensification
1.00	0.886	–
1.75	0.934	0.516
3.33	0.920	0.650
8.27	0.882	0.985
19.11	0.824	1.494