

## A UNSTRUCTURED PRUNING NATURALLY REDUCES THE ENTROPY

Let us assume the input  $x$  for a given neuron is a sequence of random variables  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ . Similarly, we can assume the  $N$  parameters populating such neuron, for a large  $N$  limit, follow as well a Gaussian distribution, and we model it as  $W \sim \mathcal{N}(\mu_W, \sigma_W^2)$ . [These assumptions are empirically validated in Appendix B](#). Let us assume we apply a magnitude-based pruning mask to the neuron’s parameters, where we apply some threshold  $t$ . As such, we obtain a modified distribution for the layer’s parameters:

$$f_{\widehat{W}}(w, t) = \begin{cases} \frac{1}{\sigma_W \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{w - \mu_W}{\sigma_W} \right)^2 \right] & |w| > t \\ \zeta(t) \delta(w) & |w| \leq t, \end{cases} \quad (11)$$

where

$$\zeta(t) = \frac{1}{2} \left[ \operatorname{erf} \left( \frac{t - \mu_W}{\sigma_W \sqrt{2}} \right) - \operatorname{erf} \left( \frac{-t - \mu_W}{\sigma_W \sqrt{2}} \right) \right] \quad (12)$$

is the fraction of parameters pruned, or *pruning rate*,  $\delta$  is the Dirac delta and  $\operatorname{erf}$  is the error function. Fig. 2a displays an example of distribution when applying magnitude pruning having threshold  $t$  against the original distribution. Under the assumption of independent-centered distributions having a unitary variance, we can obtain the distribution for the pre-activation  $z$  (resulting from the product of the weights and the input, modeled through the random variable  $Z$ ), according to the result obtained by Craig (1936); Seijas-Macías and Oliveira (2012), follows

$$f_Z(z, t) = \frac{1}{\pi} K_0 \left( \left| \frac{1}{q(t)} \cdot z \right| \right), \quad (13)$$

where

$$q(t) = 1 - \operatorname{erf} \left( \frac{t}{\sqrt{2}} \right) \quad (14)$$

and  $K_n$  is the  $n$ -th order modified Bessel function of the second kind. We can observe, from Fig. 2b, how  $f_Z$  is affected by increasing the thresholding  $t$ . Now, let us assume the activation function of such a neuron is a rectifier function, and we are interested in observing what is the probability of the post-activation output being in the linear region: we are interested in measuring

$$p[Z > \epsilon] = \frac{1}{\pi} \int_{\epsilon}^{+\infty} K_0 \left( \left| \frac{1}{q(t)} \cdot z \right| \right) dz = \frac{1}{2} \left[ 1 - I \left( \frac{\epsilon}{q(t)} \right) \right], \quad (15)$$

where

$$I(x) = x[L_{-1}(x)K_0(x) + L_0(x)K_1(x)], \quad (16)$$

$L_n$  is the  $n$ -th order modified Struve function, and  $\epsilon$  is a positive small value. From this, we can easily obtain the complementary probability  $p[Z \leq \epsilon]$  and for instance calculate the entropy between the two States.

Fig. 2c displays the entropy as a function of the thresholding parameter  $t$ : as we observe, the entropy decreases given that the threshold increases: through unstructured pruning, the neuron’s output entropy is naturally minimized when employing rectified activations, even in the oversimplified case here treated.

## B PRACTICAL SANITY CHECKS ON GAUSSIAN ASSUMPTIONS

Our derivation is based on the assumption that the weights of a layer and the inputs for its corresponding rectifier activation are Gaussian distributed. We validate here this assumption empirically. Fig 4, 6, 8, 10, 12 show the weights distribution for each rectifier activated layer for all the models employed in our experiments. Moreover, Fig 5, 7, 9, 11, 13 show the input distribution for each rectifier layer for each model used in our experiments. Inputs from CIFAR-10 are used for ResNet-18, Swin-T, and MobileNet-V2 models, while inputs from QNLI are employed for BERT and RoBERTa models. It appears that the weights and the inputs of the corresponding layers are following a Gaussian distribution.

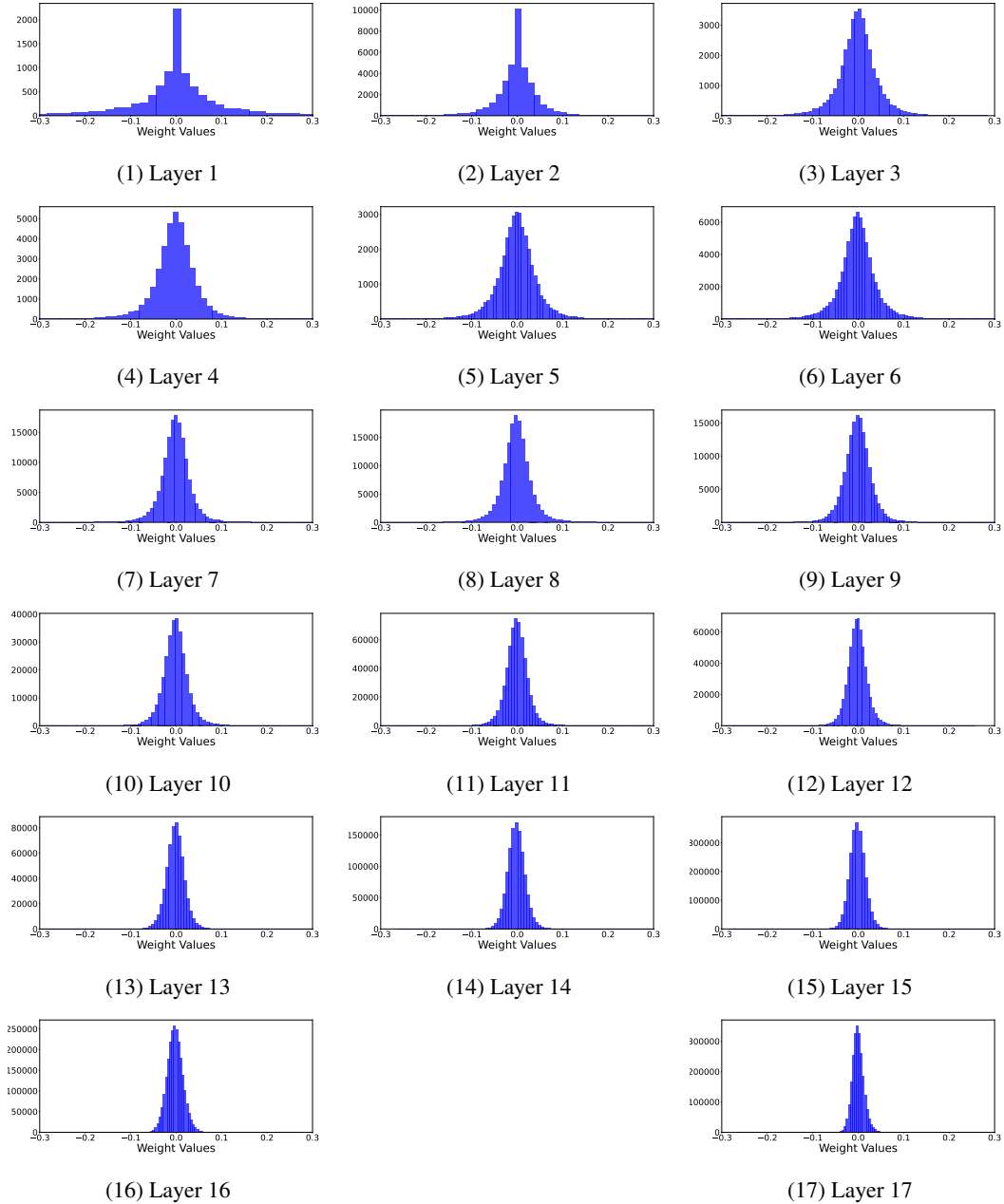


Figure 4: Weights distribution of each layer in ResNet-18

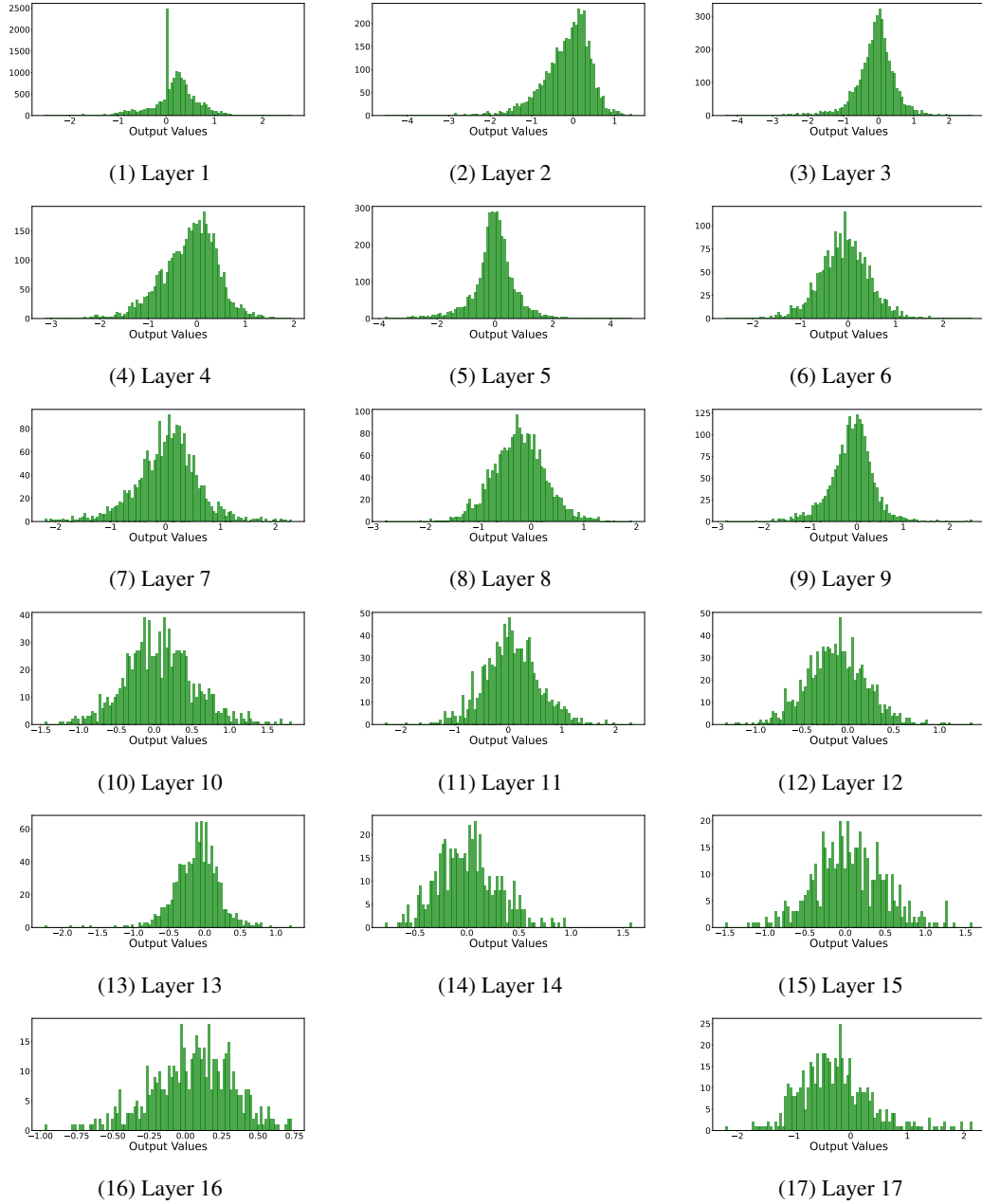


Figure 5: Input distribution of each layer in ResNet-18

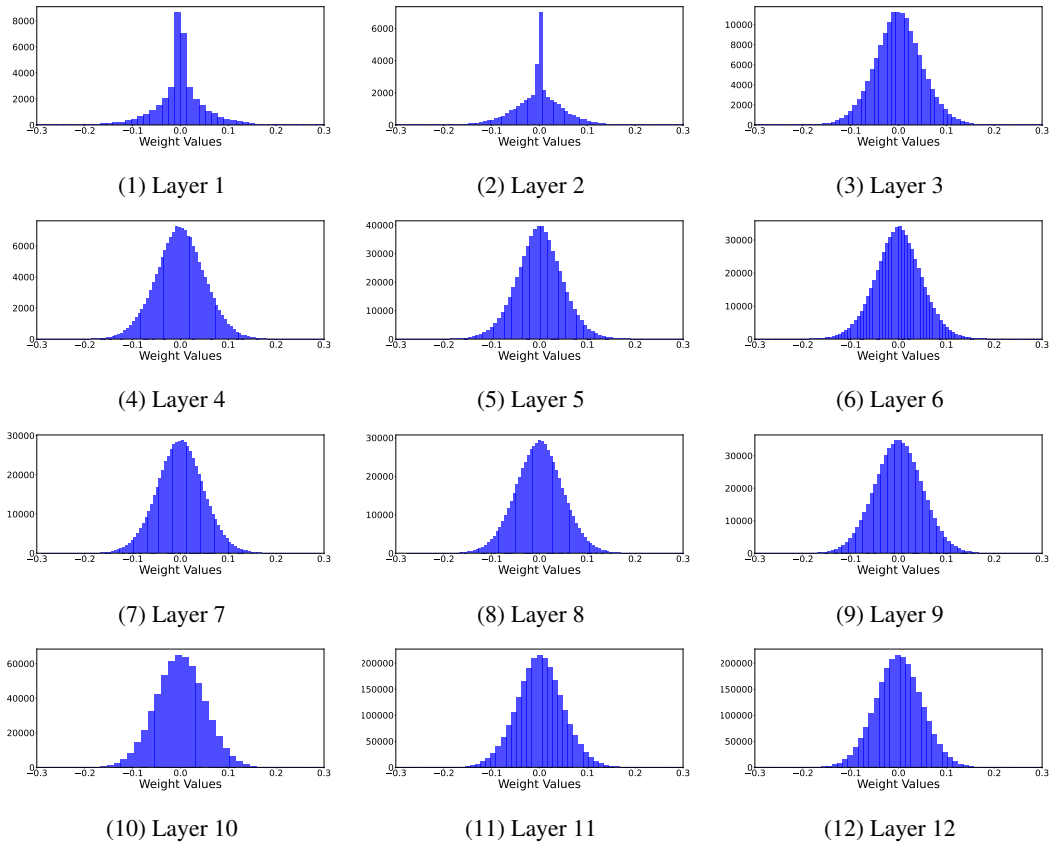


Figure 6: Weights distribution of each layer in Swin-T

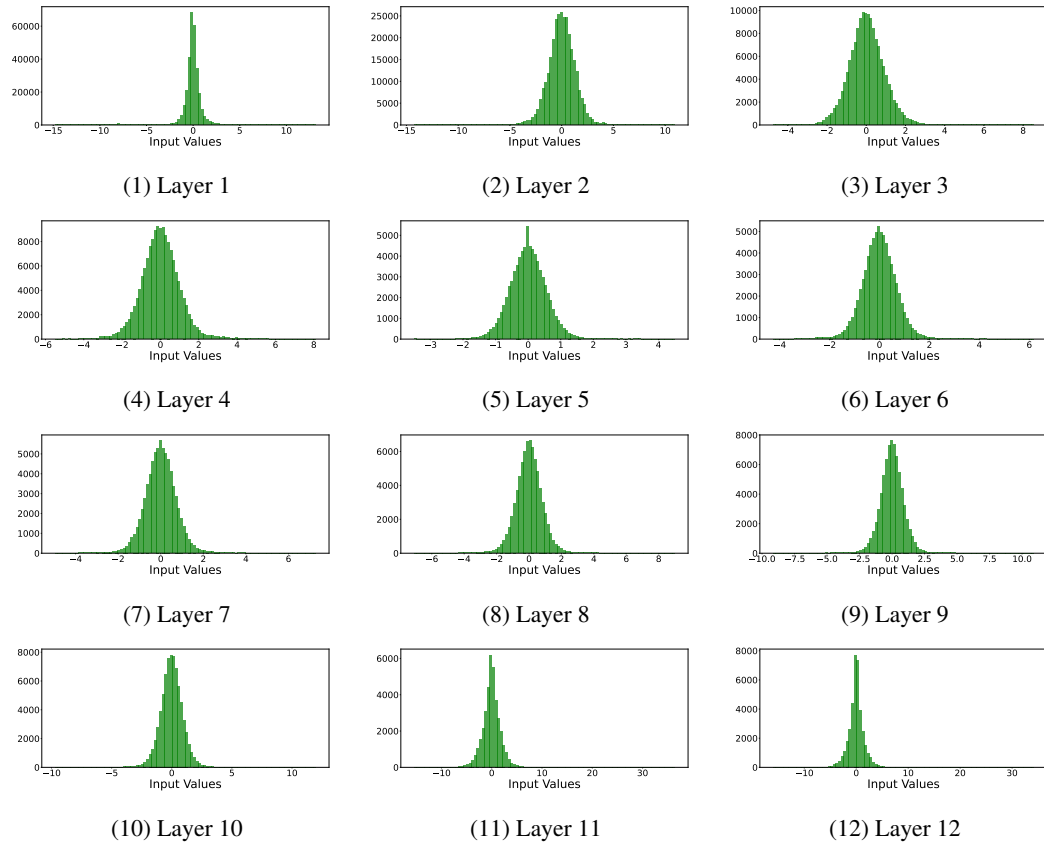


Figure 7: Input distribution of each layer in Swin-T

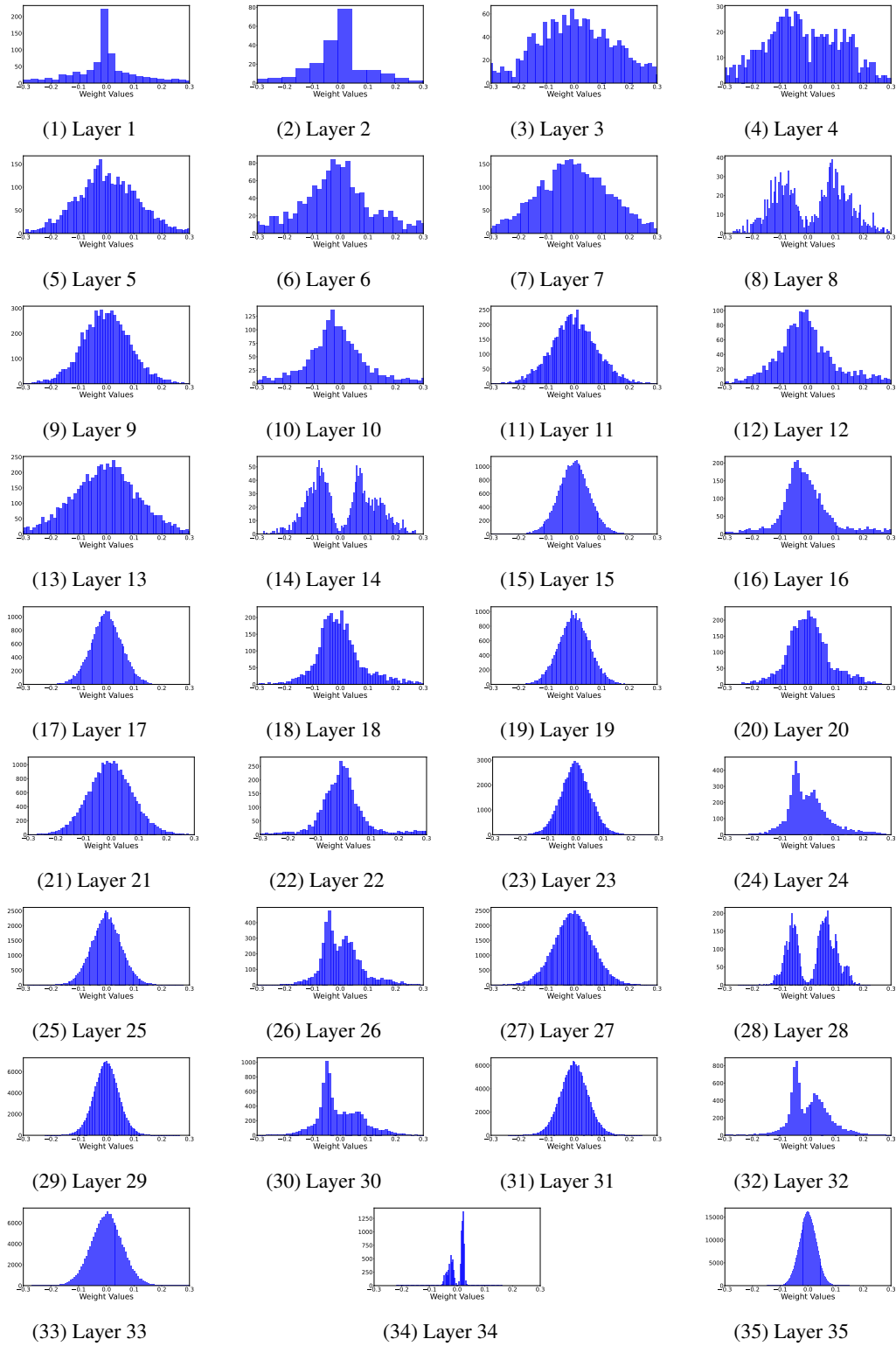


Figure 8: Weights distribution of each layer in MobileNet

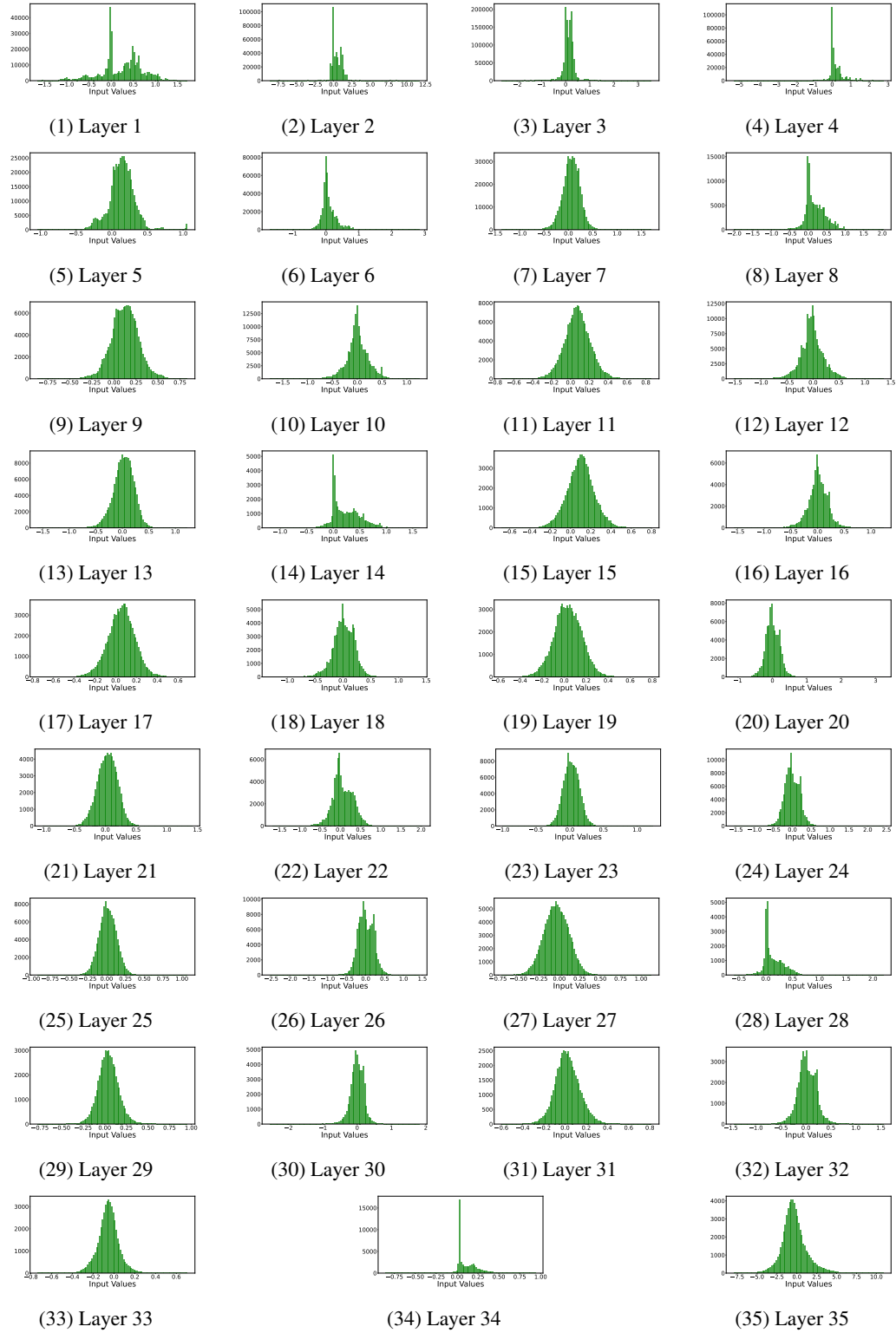


Figure 9: Input distribution of each layer in MobileNet

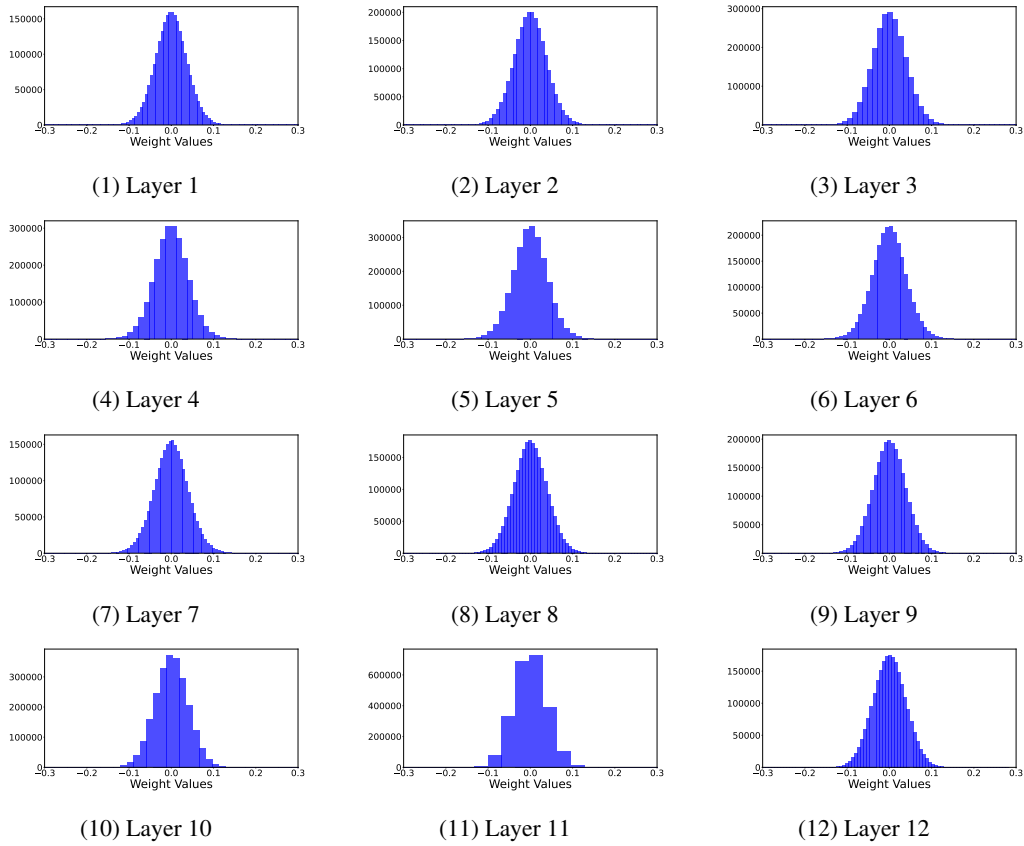


Figure 10: Weights distribution of each layer in BERT



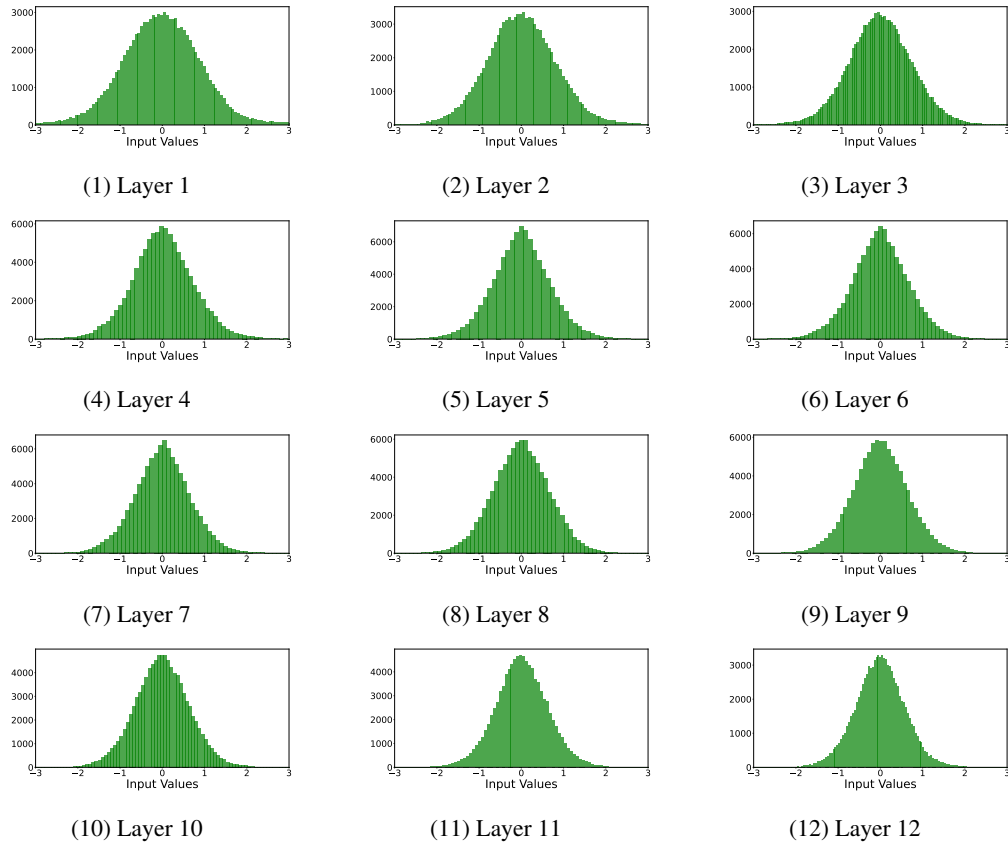


Figure 11: Input distribution of each layer in BERT

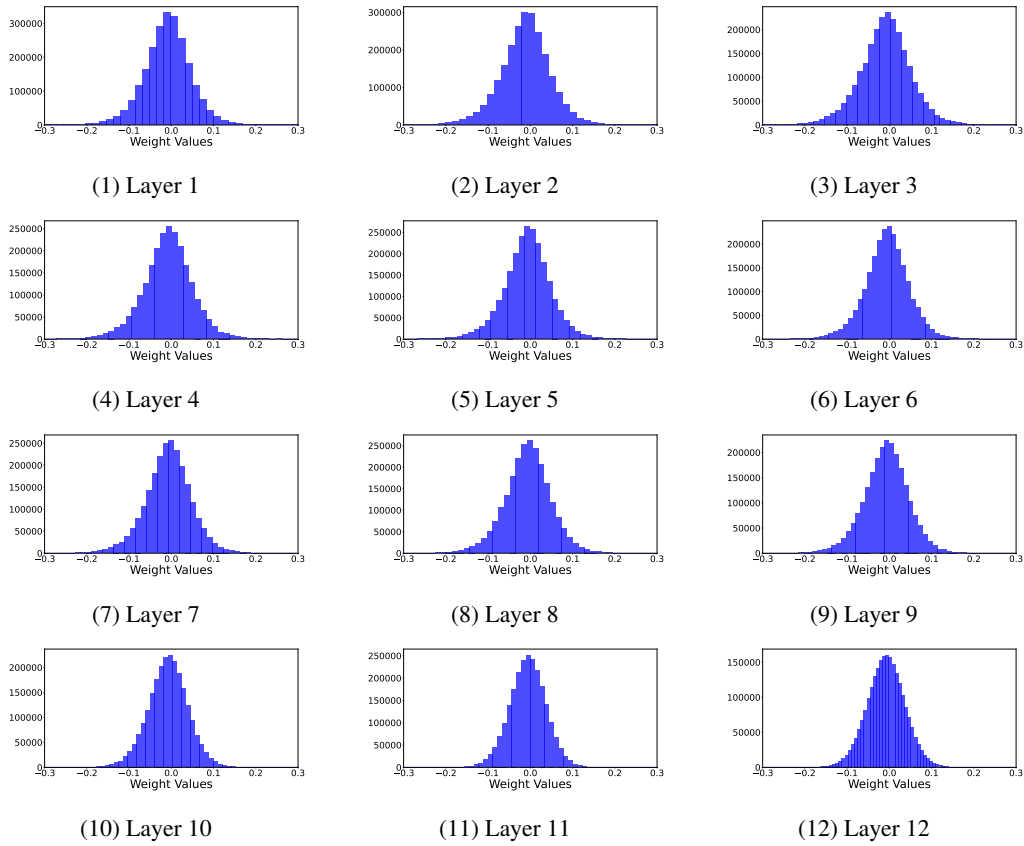


Figure 12: Weights distribution of each layer in RoBERTa

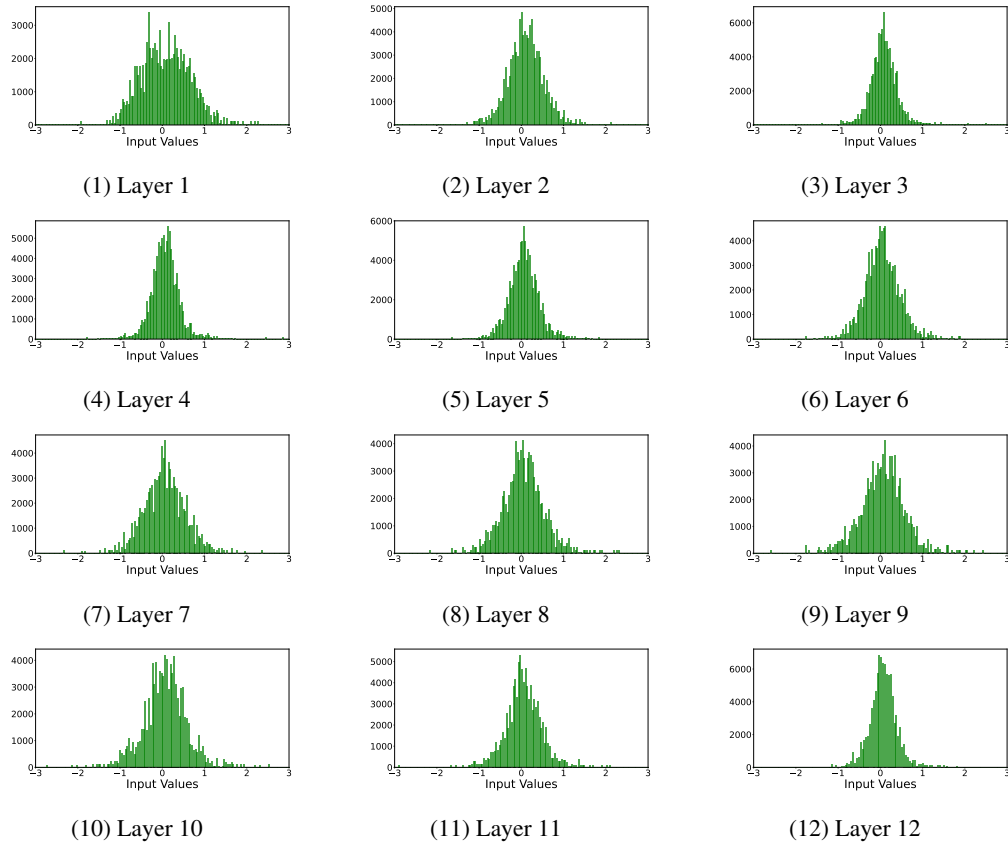


Figure 13: Input distribution of each layer in RoBERTa

**Algorithm 2** Function `Weights_to_prune`


---

```

1: function WEIGHTS_TO_PRUNE( $L, \hat{\mathcal{H}}, \|\mathbf{w}\|_0^{\text{PRUNED}}, \mathcal{D}$ )
2:   for  $l \in L$  do
3:      $\mathcal{I}_l \leftarrow \frac{1}{N_l} \sum_{i=1}^{N_l} \hat{\mathcal{H}}_{l,i} \cdot \frac{1}{\|\mathbf{w}_{l,i}\|_0} |\mathbf{w}_{l,i}|$ 
4:      $\mathcal{R}_l \leftarrow \begin{cases} \frac{\sum_{j \in L} \mathcal{I}_j}{\mathcal{I}_l} & \text{if } \mathcal{I}_l \neq 0 \\ 0 & \text{otherwise.} \end{cases}$ 
5:      $\|\mathbf{w}_L\|_0^{\text{pruned}} \leftarrow \|\mathbf{w}\|_0^{\text{pruned}} \cdot \frac{\exp[\mathcal{R}_l]}{\sum_j \exp[\mathcal{R}(j)]}$ 
6:   end for
7:   return  $\|\mathbf{w}_L\|_0^{\text{pruned}}$ 
8: end function

```

---

**C DETAILS ON THE LEARNING STRATEGIES EMPLOYED**

The implementation details used in this paper are presented here.

Alg. 2 is the function we use to calculate the number of parameters to be removed from each layer.

Like in He et al. (2022) setup, for the ResNet-18 network, a modified version of the `torchvision` model is used: the first convolutional layer is set with a filter of size  $3 \times 3$  and the max-pooling layer that follows has been eliminated to adapt ResNet-18 for CIFAR-10.

CIFAR-10 is augmented with per-channel normalization, random horizontal flipping, and random shifting by up to four pixels in any direction. For the datasets of DomainBed, the images are augmented with per-channel normalization, random horizontal flipping, random cropping, and resizing to 224. The brightness, contrast, saturation, and hue are also randomly affected with a factor fixed to 0.4. Tiny ImageNet is augmented with per-channel normalization and random horizontal flipping. ImageNet is augmented with per-channel normalization, random horizontal flipping, random cropping, and resizing to 224. The sequence length of SST-2, QNLI from Williams et al, and RTE is set to 128.

All weights from ReLU-activated layers are set as prunable for ResNet-18. For Swin-T, BERT, and RoBERTa, all weights from GELU-activated layers are prunable. while for MobileNetv2 all weights from ReLU6-activated layers are considered in the pruning. Neither biases nor batch normalization parameters are pruned.

$\zeta$  is a crucial parameter. As Table 8 shows, in our approach, since we utilize iterative pruning, if the sparse ratio is set too low, more iterations will be required to get removable layers (because we remove fewer parameters at a time). On the other hand, if the sparse ratio is set too high, the model’s performance will be destroyed quickly (because essential parameters are removed). Currently, we follow commonly used sparse ratios from unstructured pruning.

The training hyperparameters used in the Image classification experiments are presented in Table 9, hyper-parameters used in the NLP experiments are presented in Table 10. We have performed our experiments on an NVIDIA A4500 GPU. Our code is attached to this supplementary material and will be publicly available upon acceptance of the article.

Table 8: Iterations (Iter.), Test performance (top-1), and the number of removed layers (Rem.) for MobileNet-V2 trained on CIFAR-10 and pruned by NEPENTHE with different  $\zeta$

(a) $\zeta = 0.05$			(b) $\zeta = 0.1$			(c) $\zeta = 0.25$		
Iter.	top-1	Rem.	Iter.	top-1	Rem.	Iter.	top-1	Rem.
1	93.42	1/35	1	93.55	1/35	1	94.12	1/35
3	93.56	2/35	3	93.14	2/35	3	10	18/35
7	93.36	5/35	7	93.26	7/35	7	-	-
9	93.54	7/35	9	92.78	9/35	9	-	-

Table 9: Table of the different employed learning strategies.

Model	Dataset	Epochs	Batch	Opt.	Mom.	LR	Milestones	Drop Factor	Weight Decay
ResNet-18	CIFAR-10	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
ResNet-50	CIFAR-10	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
ResNet-152	CIFAR-10	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
Swin-T	CIFAR-10	160	128	SGD	0.9	0.001	[80, 120]	0.1	1e-4
MobileNetv2	CIFAR-10	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
MobileNetV2-0.75	CIFAR-10	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
ResNet-18	PACS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
Swin-T	PACS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
MobileNetv2	PACS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
ResNet-18	VLCS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
Swin-T	VLCS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
MobileNetv2	VLCS	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
ResNet-18	SVIRO	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
Swin-T	SVIRO	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
MobileNetv2	SVIRO	30	16	SGD	0.9	0.001	[24]	0.1	5e-4
ResNet-18	Tiny ImageNet	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
Swin-T	Tiny ImageNet	160	128	SGD	0.9	0.001	[80, 120]	0.1	1e-4
MobileNetv2	Tiny ImageNet	160	128	SGD	0.9	0.1	[80, 120]	0.1	1e-4
ResNet-18	ImageNet	90	128	SGD	0.9	0.1	[30, 90]	0.1	1e-4

Table 10: Table of the different employed learning strategies.

Model	Dataset	Epochs	Batch	Opt.	LR	$\beta_1$	$\beta_2$	$\epsilon$
BERT	QNLI	3	32	AdamW	2e-5	0.9	0.999	1e-8
RoBERTa	QNLI	3	32	AdamW	2e-5	0.9	0.999	1e-8
BERT	RTE	3	32	AdamW	2e-5	0.9	0.999	1e-8
RoBERTa	RTE	3	32	AdamW	2e-5	0.9	0.999	1e-8
BERT	SST-2	3	32	AdamW	2e-5	0.9	0.999	1e-8
RoBERTa	SST-2	3	32	AdamW	2e-5	0.9	0.999	1e-8

## D MORE DETAILED RESULTS

### D.1 LAYER STATES FOR MODELS TRAINED ON CIFAR-10

Fig. 14 shows the layers’ states for ResNet-18 trained on CIFAR-10 with different methods.

### D.2 EXPERIMENTS ON MORE ARCHITECTURES

Table 11 presents the results for ResNet-50, ResNet-152, and MobileNetV2-0.75 models trained on Cifar-10 dataset. Table 12 presents the results for ResNet-18 model trained on ImageNet dataset.

Table 11: est performance (top-1) and the number of removed layers (Rem.) for ResNet-50, ResNet-152, and MobileNetV2-0.75 models trained on CIFAR-10.

Dataset	Approach	ResNet-50		ResNet-152		MobileNetV2-0.75	
		Top-1	Rem.	Top-1	Rem.	Top-1	Rem.
CIFAR-10	Dense	87.37	0/49	85.61	0/151	85.17	0/35
	NEPENTHE	89.06	16/49	89.20	82/151	87.06	12/35

### D.3 DETAILED EXPERIMENTAL RESULTS

Tables 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 29, 29, 30, 31, 32, 33, present the entropy trends in the six layers exhibiting the lowest entropy and the training time, for all the unstructured pruning and NEPENTHE setups. This illustrates that while unstructured pruning inherently reduces the entropy of certain layers, as detailed in Section 4.2, it lacks the capability to entirely eliminate any specific layer. In contrast, our methodology, NEPENTHE, aims to push all the encoded information from layers with low entropy to those with already high entropy. This strategy enables the removal of zero-entropy layers.

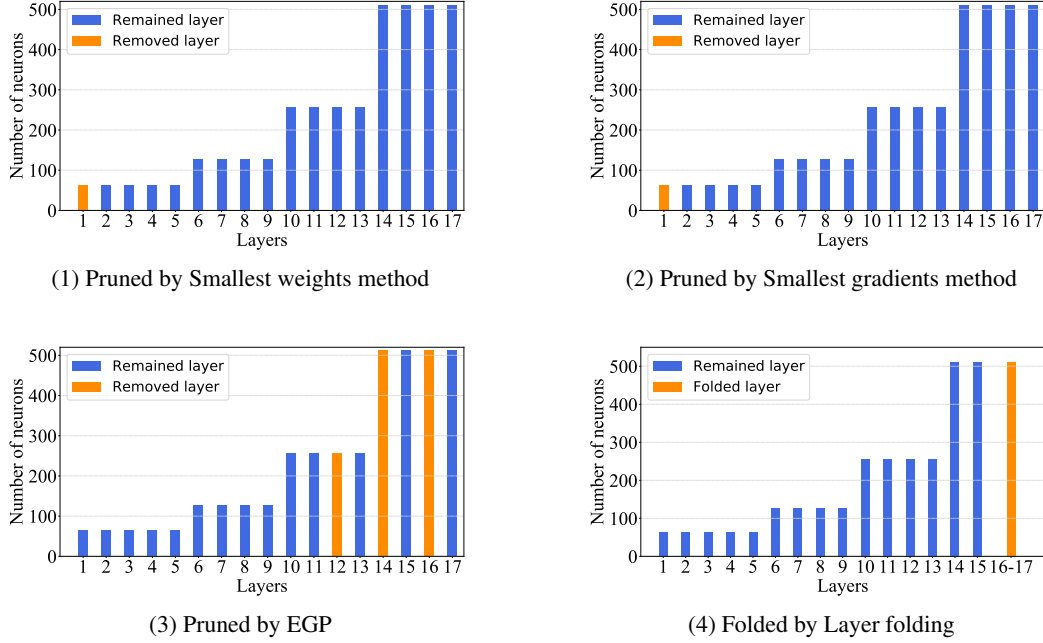


Figure 14: Layer states for ResNet-18 trained on CIFAR-10 with different methods.

Table 12: Test performance (top-1) and the number of removed layers (Rem.) for Resnet-18 trained on Imagenet.

Dataset	Approach	ResNet-18	
		top-1	Rem.
ImageNet	Dense model	68.20	0/17
	IMP (low prune)	68.38	0/17
	IMP (mid prune)	67.88	0/17
	IMP (high prune)	66.63	0/17
	NEPENTHE (low prune)	66.17	0/17
	NEPENTHE (mid prune)	62.74	1/17
	NEPENTHE (high prune)	62.15	3/17

Table 13: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for ResNet-18 on CIFAR-10.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.647	0.680	0.728	0.785	0.791	0.797	91.66	0/17	0h48
IMP(iter #1)	0.585	0.650	0.699	0.725	0.767	0.778	92.29	0/17	1h36
IMP(iter #2)	0.506	0.580	0.647	0.654	0.700	0.722	92.25	0/17	2h24
IMP(iter #3)	0.256	0.623	0.658	0.672	0.682	0.737	92.46	0/17	3h12
IMP(iter #4)	0.192	0.660	0.667	0.676	0.698	0.763	92.27	0/17	4h00
IMP(iter #5)	0.136	0.589	0.648	0.727	0.728	0.791	92.44	0/17	4h48
IMP(iter #6)	0.093	0.447	0.640	0.650	0.764	0.765	91.89	0/17	5h36
IMP(iter #7)	0.055	0.335	0.487	0.592	0.640	0.775	91.66	0/17	6h24
NEPENTHE(iter #1)	0	0.168	0.581	0.654	0.681	0.714	92.25	1/17	1h37
NEPENTHE(iter #2)	0	0.076	0.615	0.619	0.633	0.644	92.60	1/17	2h26
NEPENTHE(iter #3)	0	0	0	0.121	0.139	0.642	92.55	3/17	3h15
NEPENTHE(iter #4)	0	0	0	0.003	0.242	0.320	91.93	3/17	4h04
NEPENTHE(iter #5)	0	0	0	0	0	0.114	89.30	5/17	4h53
NEPENTHE(iter #6)	0	0	0	0	0	0.019	89.43	5/17	5h42
NEPENTHE(iter #7)	0	0	0	0	0	0	83.42	6/17	6h31

Table 14: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for Swin-T on CIFAR-10.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.03	0.054	0.382	0.394	0.394	0.44	91.54	0/12	1h53
IMP(iter #1)	0.03	0.055	0.383	0.397	0.398	0.444	91.73	0/12	3h46
IMP(iter #2)	0.031	0.057	0.382	0.392	0.399	0.443	91.80	0/12	5h39
IMP(iter #3)	0.034	0.063	0.375	0.383	0.393	0.438	91.43	0/12	7h32
IMP(iter #4)	0.036	0.072	0.365	0.379	0.382	0.426	91.46	0/12	9h25
IMP(iter #5)	0.041	0.080	0.349	0.361	0.369	0.409	91.27	0/12	11h18
IMP(iter #6)	0.048	0.096	0.334	0.343	0.350	0.386	91.05	0/12	13h11
IMP(iter #7)	0.055	0.113	0.31	0.325	0.327	0.355	90.53	0/12	15h04
NEPENTHE(iter #1)	0.001	0.215	0.385	0.397	0.407	0.443	91.77	0/12	3h48
NEPENTHE(iter #2)	0.001	0.219	0.387	0.399	0.409	0.445	92.06	0/12	5h45
NEPENTHE(iter #3)	0.001	0.254	0.380	0.395	0.405	0.440	92.08	0/12	7h38
NEPENTHE(iter #4)	0.001	0.001	0.377	0.388	0.404	0.433	92.31	0/12	9h33
NEPENTHE(iter #5)	0	0	0.363	0.373	0.406	0.423	92.29	2/12	11h28
NEPENTHE(iter #6)	0	0	0.344	0.359	0.407	0.412	92.21	2/12	13h23
NEPENTHE(iter #7)	0	0	0.287	0.317	0.405	0.405	92.11	2/12	15h18

Table 15: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for MobileNetv2 on CIFAR-10.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.386	0.474	0.486	0.504	0.528	0.544	93.68	0/35	0h43
IMP(iter #1)	0.186	0.206	0.233	0.241	0.249	0.260	94.07	0/35	1h26
IMP(iter #2)	0.116	0.117	0.153	0.167	0.167	0.168	93.67	0/35	2h09
IMP(iter #3)	0.082	0.084	0.111	0.119	0.119	0.125	93.83	0/35	2h52
IMP(iter #4)	0.063	0.066	0.090	0.094	0.095	0.101	93.32	0/35	3h35
IMP(iter #5)	0.056	0.057	0.074	0.075	0.086	0.088	93.26	0/35	4h18
IMP(iter #6)	0.050	0.050	0.065	0.067	0.071	0.076	93.47	0/35	5h01
IMP(iter #7)	0.046	0.047	0.059	0.060	0.061	0.064	93.50	0/35	5h44
NEPENTHE(iter #1)	0	0.198	0.229	0.232	0.244	0.248	93.55	1/35	1h27
NEPENTHE(iter #2)	0	0.109	0.127	0.138	0.139	0.149	93.42	1/35	2h11
NEPENTHE(iter #3)	0	0.082	0.085	0.103	0.106	0.107	93.14	1/35	2h55
NEPENTHE(iter #4)	0	0	0.063	0.065	0.074	0.076	93.25	2/35	3h39
NEPENTHE(iter #5)	0	0	0.064	0.067	0.049	0.051	93.37	4/35	4h23
NEPENTHE(iter #6)	0	0	0	0	0	0.042	93.15	5/35	5h07
NEPENTHE(iter #7)	0	0	0	0	0	0	93.26	7/35	5h51

Table 16: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for ResNet-18 on Tiny ImageNet.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.471	0.500	0.592	0.625	0.627	0.783	41.44	0/17	2h15
IMP(iter #1)	0.470	0.540	0.621	0.662	0.666	0.780	42.24	0/17	4h30
IMP(iter #2)	0.461	0.621	0.637	0.697	0.726	0.781	42.12	0/17	6h45
IMP(iter #3)	0.487	0.643	0.735	0.736	0.776	0.779	42.10	0/17	9h00
IMP(iter #4)	0.488	0.643	0.760	0.783	0.831	0.831	41.18	0/17	11h15
IMP(iter #5)	0.482	0.605	0.727	0.839	0.845	0.872	39.92	0/17	13h30
IMP(iter #6)	0.469	0.585	0.690	0.814	0.834	0.834	37.16	0/17	15h45
IMP(iter #7)	0.464	0.544	0.641	0.661	0.725	0.741	39.14	0/17	18h00
NEPENTHE(iter #1)	0	0	0.063	0.559	0.633	0.699	41.42	2/17	4h33
NEPENTHE(iter #2)	0	0	0	0	0	0.129	39.56	5/17	6h51
NEPENTHE(iter #3)	0	0	0	0	0	0.169	40.00	5/17	9h09
NEPENTHE(iter #4)	0	0	0	0	0	0.109	39.40	5/17	11h27
NEPENTHE(iter #5)	0	0	0	0	0	0.107	38.58	5/17	13h45
NEPENTHE(iter #6)	0	0	0	0	0	0.125	37.34	5/17	16h03
NEPENTHE(iter #7)	0	0	0	0	0	0.138	35.80	5/17	18h21



Table 17: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for Swin-T on Tiny ImageNet.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.067	0.133	0.38	0.388	0.395	0.411	75.60	0/12	5h41
IMP(iter #1)	0.069	0.131	0.370	0.373	0.384	0.399	75.86	0/12	11h22
IMP(iter #2)	0.073	0.133	0.355	0.356	0.367	0.380	75.26	0/12	17h03
IMP(iter #3)	0.080	0.143	0.335	0.336	0.346	0.357	73.60	0/12	22h44
IMP(iter #4)	0.089	0.156	0.313	0.314	0.319	0.330	72.32	0/12	28h25
IMP(iter #5)	0.096	0.169	0.291	0.291	0.295	0.309	70.90	0/12	34h06
IMP(iter #6)	0.102	0.184	0.268	0.269	0.275	0.294	69.80	0/12	39h47
IMP(iter #7)	0.104	0.193	0.249	0.255	0.266	0.289	67.56	0/12	45h28
NEPENTHE(iter #1)	0	0.139	0.370	0.377	0.392	0.394	72.58	1/12	11h27
NEPENTHE(iter #2)	0	0.143	0.150	0.183	0.195	0.381	71.02	1/12	17h13
NEPENTHE(iter #3)	0	0.143	0.158	0.183	0.192	0.269	70.76	1/12	22h59
NEPENTHE(iter #4)	0	0.133	0.137	0.165	0.178	0.187	70.12	1/12	28h45
NEPENTHE(iter #5)	0	0.128	0.132	0.172	0.173	0.180	69.68	1/12	34h36
NEPENTHE(iter #6)	0	0.124	0.129	0.164	0.174	0.176	70.06	1/12	39h17
NEPENTHE(iter #7)	0	0.123	0.128	0.160	0.170	0.180	69.42	1/12	46h03

Table 18: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for MobileNetv2 on Tiny ImageNet.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.076	0.112	0.131	0.133	0.153	0.154	45.860	0/35	1h47
IMP(iter #1)	0.036	0.039	0.039	0.045	0.048	0.054	45.84	0/35	3h34
IMP(iter #2)	0.025	0.026	0.031	0.035	0.037	0.037	47.10	0/35	5h21
IMP(iter #3)	0.018	0.021	0.030	0.030	0.031	0.031	47.740	0/35	7h08
IMP(iter #4)	0.016	0.017	0.028	0.028	0.030	0.030	46.800	0/35	8h55
IMP(iter #5)	0.013	0.016	0.025	0.025	0.028	0.029	47.560	0/35	10h42
IMP(iter #6)	0.008	0.011	0.022	0.023	0.028	0.028	47.580	0/35	12h29
IMP(iter #7)	0.007	0.01	0.022	0.023	0.028	0.029	47.440	0/35	14h16
NEPENTHE(iter #1)	0.001	0.004	0.095	0.096	0.098	0.110	46.70	0/35	3h37
NEPENTHE(iter #2)	0	0.003	0.058	0.064	0.071	0.073	47.22	1/35	5h27
NEPENTHE(iter #3)	0	0	0	0	0	0	47.26	6/35	7h17
NEPENTHE(iter #4)	0	0	0	0	0	0	47.82	9/35	9h07
NEPENTHE(iter #5)	0	0	0	0	0	0	47.92	12/35	10h57
NEPENTHE(iter #6)	0	0	0	0	0	0	0.50	34/35	12h37

Table 19: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for ResNet-18 on PACS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.332	0.439	0.602	0.667	0.686	0.687	94.70	0/17	0h46
IMP(iter #1)	0.331	0.423	0.608	0.669	0.688	0.688	95.40	0/17	1h31
IMP(iter #2)	0.319	0.429	0.602	0.668	0.670	0.683	95.30	0/17	2h17
IMP(iter #3)	0.324	0.419	0.607	0.631	0.682	0.682	94.60	0/17	3h03
IMP(iter #4)	0.318	0.441	0.613	0.613	0.661	0.688	95.10	0/17	3h49
IMP(iter #5)	0.300	0.452	0.587	0.621	0.636	0.694	94.00	0/17	4h35
IMP(iter #6)	0.285	0.458	0.533	0.643	0.647	0.694	92.30	0/17	5h21
IMP(iter #7)	0.280	0.418	0.479	0.584	0.646	0.657	90.80	0/17	6h07
NEPENTHE(iter #1)	0.129	0.430	0.482	0.634	0.668	0.669	94.20	0/17	1h32
NEPENTHE(iter #2)	0	0.041	0.091	0.482	0.596	0.596	92.40	1/17	2h19
NEPENTHE(iter #3)	0	0.030	0.066	0.527	0.559	0.599	93.00	1/17	3h06
NEPENTHE(iter #4)	0	0	0.033	0.067	0.422	0.565	90.40	2/17	3h53
NEPENTHE(iter #5)	0	0	0.032	0.061	0.084	0.217	89.50	2/17	4h40
NEPENTHE(iter #6)	0	0	0	0.001	0.002	0.028	90.10	3/17	5h27
NEPENTHE(iter #7)	0	0	0	0.002	0.002	0.040	86.30	3/17	6h14

Table 20: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for Swin-T on PACS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.057	0.184	0.344	0.362	0.380	0.381	97.10	0/12	0h57
IMP(iter #1)	0.060	0.204	0.349	0.362	0.382	0.389	96.60	0/12	1h54
IMP(iter #2)	0.067	0.214	0.357	0.370	0.383	0.388	96.20	0/12	2h51
IMP(iter #3)	0.081	0.245	0.362	0.366	0.375	0.378	96.00	0/12	3h48
IMP(iter #4)	0.095	0.269	0.352	0.353	0.355	0.372	96.60	0/12	4h45
IMP(iter #5)	0.110	0.303	0.314	0.335	0.339	0.341	95.00	0/12	5h42
IMP(iter #6)	0.113	0.278	0.306	0.318	0.321	0.329	94.60	0/12	6h39
IMP(iter #7)	0.101	0.232	0.269	0.284	0.293	0.298	93.90	0/12	7h36
NEPENTHE(iter #1)	0.086	0.240	0.344	0.376	0.376	0.403	96.90	0/12	1h55
NEPENTHE(iter #2)	0.001	0.369	0.372	0.383	0.398	0.416	96.50	0/12	2h53
NEPENTHE(iter #3)	0.001	0.368	0.383	0.385	0.390	0.406	95.90	0/12	3h51
NEPENTHE(iter #4)	0.001	0.360	0.369	0.369	0.392	0.394	96.30	0/12	4h49
NEPENTHE(iter #5)	0	0	0.001	0.335	0.359	0.366	95.10	2/12	5h47
NEPENTHE(iter #6)	0	0	0.107	0.298	0.348	0.349	94.60	2/12	6h45
NEPENTHE(iter #7)	0	0	0.001	0.001	0.161	0.232	93.30	2/12	7h43

Table 21: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for MobileNetv2 on PACS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.207	0.291	0.324	0.372	0.463	0.471	93.20	0/35	0h34
IMP(iter #1)	0.218	0.263	0.284	0.390	0.457	0.467	95.70	0/35	1h09
IMP(iter #2)	0.216	0.235	0.257	0.373	0.453	0.462	95.50	0/35	1h44
IMP(iter #3)	0.224	0.244	0.252	0.403	0.464	0.478	95.40	0/35	2h19
IMP(iter #4)	0.222	0.229	0.241	0.386	0.459	0.476	95.70	0/35	2h54
IMP(iter #5)	0.212	0.223	0.233	0.397	0.464	0.47	95.60	0/35	3h29
IMP(iter #6)	0.196	0.212	0.237	0.405	0.470	0.485	96.20	0/35	4h04
IMP(iter #7)	0.170	0.207	0.234	0.412	0.468	0.472	95.40	0/35	4h39
NEPENTHE(iter #1)	0.119	0.139	0.151	0.192	0.200	0.225	93.30	0/35	1h10
NEPENTHE(iter #2)	0.093	0.128	0.129	0.130	0.135	0.165	93.20	0/35	1h46
NEPENTHE(iter #3)	0.077	0.093	0.112	0.125	0.140	0.141	92.50	0/35	2h22
NEPENTHE(iter #4)	0	0.076	0.083	0.105	0.106	0.116	92.20	1/35	2h58
NEPENTHE(iter #5)	0	0.054	0.068	0.096	0.097	0.115	89.70	1/35	3h34
NEPENTHE(iter #6)	0	0.014	0.016	0.036	0.050	0.051	89.00	1/35	4h10
NEPENTHE(iter #7)	0	0.004	0.008	0.023	0.027	0.034	88.70	1/35	4h46

Table 22: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for ResNet-18 on VLCS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.382	0.457	0.647	0.676	0.681	0.698	80.89	0/17	3h49
IMP(iter #1)	0.387	0.471	0.647	0.679	0.681	0.703	82.76	0/17	8h37
IMP(iter #2)	0.392	0.476	0.644	0.654	0.69	0.703	82.01	0/17	13h06
IMP(iter #3)	0.378	0.474	0.620	0.658	0.707	0.707	82.01	0/17	17h15
IMP(iter #4)	0.391	0.491	0.595	0.672	0.711	0.726	80.15	0/17	22h24
IMP(iter #5)	0.372	0.479	0.571	0.665	0.716	0.739	79.31	0/17	27h13
IMP(iter #6)	0.383	0.519	0.531	0.699	0.721	0.750	78.84	0/17	31h22
IMP(iter #7)	0.357	0.409	0.502	0.64	0.707	0.712	74.09	0/17	35h11
NEPENTHE(iter #1)	0.001	0.453	0.497	0.651	0.676	0.680	78.99	0/17	8h44
NEPENTHE(iter #2)	0	0	0.001	0.508	0.516	0.619	78.38	2/17	13h20
NEPENTHE(iter #3)	0	0	0	0	0.518	0.553	76.98	4/17	17h36
NEPENTHE(iter #4)	0	0	0	0	0.516	0.574	78.66	4/17	22h52
NEPENTHE(iter #5)	0	0	0	0	0	0	76.05	6/17	27h48
NEPENTHE(iter #6)	0	0	0	0	0	0	74.28	6/17	32h04
NEPENTHE(iter #7)	0	0	0	0	0	0	74.37	6/17	36h01

Table 23: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for Swin-T on VLCS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.070	0.175	0.373	0.385	0.414	0.427	86.58	0/12	2h22
IMP(iter #1)	0.078	0.179	0.391	0.402	0.421	0.433	84.72	0/12	4h45
IMP(iter #2)	0.093	0.195	0.387	0.403	0.416	0.419	85.65	0/12	8h08
IMP(iter #3)	0.102	0.224	0.388	0.411	0.424	0.424	84.34	0/12	11h31
IMP(iter #4)	0.122	0.236	0.395	0.402	0.415	0.418	84.06	0/12	14h54
IMP(iter #5)	0.140	0.261	0.369	0.394	0.404	0.412	82.01	0/12	18h56
IMP(iter #6)	0.141	0.292	0.331	0.387	0.390	0.393	81.36	0/12	22h38
IMP(iter #7)	0.139	0.277	0.304	0.370	0.373	0.374	80.06	0/12	26h20
NEPENTHE(iter #1)	0.121	0.187	0.400	0.402	0.430	0.437	85.46	0/12	4h57
NEPENTHE(iter #2)	0.132	0.225	0.403	0.406	0.428	0.438	85.09	0/12	8h26
NEPENTHE(iter #3)	0	0.411	0.413	0.432	0.437	0.457	85.27	1/12	11h55
NEPENTHE(iter #4)	0	0.409	0.420	0.441	0.446	0.463	83.88	1/12	15h18
NEPENTHE(iter #5)	0	0.406	0.409	0.428	0.434	0.469	81.73	1/12	19h28
NEPENTHE(iter #6)	0	0.318	0.383	0.398	0.413	0.469	81.55	1/12	23h20
NEPENTHE(iter #7)	0	0.001	0.304	0.369	0.374	0.475	79.22	1/12	27h08

Table 24: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for MobileNetv2 on VLCS.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.257	0.353	0.46	0.513	0.514	0.572	81.83	0/35	3h00
IMP(iter #1)	0.270	0.292	0.490	0.503	0.533	0.573	80.80	0/35	6h01
IMP(iter #2)	0.261	0.263	0.507	0.524	0.528	0.562	81.64	0/35	9h01
IMP(iter #3)	0.269	0.271	0.496	0.501	0.536	0.573	80.43	0/35	12h02
IMP(iter #4)	0.258	0.258	0.491	0.521	0.550	0.579	79.59	0/35	15h03
IMP(iter #5)	0.264	0.271	0.474	0.540	0.545	0.589	79.96	0/35	18h03
IMP(iter #6)	0.268	0.277	0.468	0.547	0.549	0.585	80.80	0/35	21h04
IMP(iter #7)	0.273	0.279	0.470	0.524	0.554	0.592	80.43	0/35	24h05
NEPENTHE(iter #1)	0.184	0.249	0.342	0.505	0.534	0.581	81.08	0/35	6h04
NEPENTHE(iter #2)	0.001	0.077	0.251	0.345	0.417	0.500	80.52	0/35	9h07
NEPENTHE(iter #3)	0	0.002	0.005	0.260	0.354	0.488	78.84	1/35	12h11
NEPENTHE(iter #4)	0	0.001	0.040	0.261	0.363	0.527	77.91	1/35	15h15
NEPENTHE(iter #5)	0	0	0.001	0.274	0.366	0.523	80.06	2/35	18h18
NEPENTHE(iter #6)	0	0	0.001	0.26	0.351	0.485	79.31	2/35	21h22

Table 25: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for ResNet-18 on SVIRO.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.009	0.446	0.631	0.685	0.693	0.707	99.93	0/17	7h03
IMP(iter #1)	0.009	0.456	0.629	0.679	0.684	0.705	99.98	0/17	14h06
IMP(iter #2)	0.008	0.446	0.624	0.646	0.68	0.700	99.98	0/17	21h09
IMP(iter #3)	0.009	0.454	0.638	0.655	0.676	0.689	100	0/17	28h12
IMP(iter #4)	0.007	0.478	0.641	0.658	0.677	0.687	99.95	0/17	35h15
IMP(iter #5)	0.010	0.507	0.658	0.659	0.698	0.702	99.96	0/17	42h18
IMP(iter #6)	0.006	0.530	0.577	0.676	0.688	0.691	100	0/17	49h21
IMP(iter #7)	0.003	0.488	0.495	0.518	0.629	0.657	99.95	0/17	56h26
NEPENTHE(iter #1)	0.001	0.512	0.551	0.649	0.686	0.702	99.98	0/17	14h19
NEPENTHE(iter #2)	0	0	0.001	0.031	0.453	0.460	99.93	2/17	21h35
NEPENTHE(iter #3)	0	0	0.012	0.397	0.440	0.598	99.91	2/17	28h51
NEPENTHE(iter #4)	0	0	0	0.006	0.013	0.371	99.86	3/17	36h07
NEPENTHE(iter #5)	0	0	0	0.001	0.005	0.054	99.84	3/17	43h23
NEPENTHE(iter #6)	0	0	0	0	0	0	99.61	8/17	50h44
NEPENTHE(iter #7)	0	0	0	0	0	0	98.75	8/17	57h57

Table 26: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for Swin-T on SVIRO.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.061	0.205	0.280	0.290	0.321	0.325	99.95	0/12	5h38
IMP(iter #1)	0.046	0.232	0.284	0.285	0.291	0.303	99.84	0/12	12h17
IMP(iter #2)	0.026	0.125	0.273	0.280	0.283	0.289	99.77	0/12	18h56
IMP(iter #3)	0.022	0.062	0.216	0.233	0.238	0.275	99.84	0/12	25h55
IMP(iter #4)	0.027	0.071	0.163	0.183	0.187	0.187	99.68	0/12	32h14
IMP(iter #5)	0.034	0.095	0.101	0.115	0.143	0.149	99.68	0/12	39h33
IMP(iter #6)	0.036	0.047	0.090	0.125	0.127	0.129	99.79	0/12	46h12
IMP(iter #7)	0.026	0.041	0.074	0.124	0.127	0.137	99.75	0/12	52h01
NEPENTHE(iter #1)	0.001	0.269	0.321	0.326	0.343	0.348	99.93	0/12	12h28
NEPENTHE(iter #2)	0	0.338	0.347	0.357	0.362	0.367	99.82	1/12	19h18
NEPENTHE(iter #3)	0	0.156	0.282	0.309	0.376	0.381	99.79	1/12	26h28
NEPENTHE(iter #4)	0	0.001	0.092	0.235	0.267	0.356	99.68	1/12	32h58
NEPENTHE(iter #5)	0	0	0.001	0.001	0.001	0.339	99.77	2/12	40h28
NEPENTHE(iter #6)	0	0	0	0	0	0.162	99.75	5/12	47h18
NEPENTHE(iter #7)	0	0	0	0	0	0.001	99.70	5/12	53h18

Table 27: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for MobileNetv2 on SVIRO.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.187	0.241	0.337	0.341	0.484	0.508	99.98	0/35	5h35
IMP(iter #1)	0.165	0.218	0.248	0.366	0.478	0.515	99.98	0/35	12h11
IMP(iter #2)	0.152	0.191	0.232	0.402	0.471	0.560	100	0/35	18h47
IMP(iter #3)	0.161	0.180	0.290	0.411	0.483	0.521	99.98	0/35	24h23
IMP(iter #4)	0.169	0.215	0.296	0.419	0.430	0.483	99.96	0/35	30h19
IMP(iter #5)	0.162	0.173	0.306	0.372	0.417	0.435	99.93	0/35	36h35
IMP(iter #6)	0.155	0.196	0.333	0.337	0.362	0.417	99.93	0/35	42h31
IMP(iter #7)	0.146	0.185	0.291	0.317	0.322	0.381	99.95	0/35	48h47
NEPENTHE(iter #1)	0.001	0.168	0.220	0.385	0.459	0.508	100	0/35	12h22
NEPENTHE(iter #2)	0.001	0.001	0.163	0.196	0.321	0.367	99.98	0/35	19h09
NEPENTHE(iter #3)	0	0	0.002	0.059	0.218	0.375	99.98	2/35	24h56
NEPENTHE(iter #4)	0	0	0.004	0.020	0.268	0.392	99.95	2/35	30h54
NEPENTHE(iter #5)	0	0	0.001	0.045	0.262	0.388	99.97	2/35	37h19
NEPENTHE(iter #6)	0	0	0	0	0.020	0.147	35.55	4/35	49h53

Table 28: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for BERT on QNLI.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.173	0.191	0.212	0.219	0.223	0.227	90.48	0/12	0h34
IMP(iter #1)	0.179	0.198	0.224	0.229	0.238	0.248	89.60	0/12	1h11
IMP(iter #2)	0.202	0.209	0.213	0.220	0.224	0.258	89.91	0/12	1h47
IMP(iter #3)	0.215	0.217	0.234	0.243	0.250	0.269	89.80	0/12	2h23
IMP(iter #4)	0.231	0.235	0.251	0.252	0.269	0.300	89.75	0/12	2h59
IMP(iter #5)	0.238	0.241	0.248	0.256	0.284	0.306	89.38	0/12	3h35
IMP(iter #6)	0.265	0.268	0.269	0.280	0.294	0.335	89.15	0/12	4h11
IMP(iter #7)	0.273	0.274	0.280	0.283	0.291	0.356	88.41	0/12	4h47
NEPENTHE(iter #1)	0	0	0.218	0.234	0.240	0.244	89.58	2/12	1h23
NEPENTHE(iter #2)	0	0	0.215	0.246	0.251	0.264	89.46	2/12	2h11
NEPENTHE(iter #3)	0	0	0	0.227	0.239	0.246	88.89	3/12	2h59
NEPENTHE(iter #4)	0	0	0	0.239	0.248	0.252	88.94	3/12	3h47
NEPENTHE(iter #5)	0	0	0	0.192	0.263	0.279	89.38	3/12	4h35
NEPENTHE(iter #6)	0	0	0	0	0.280	0.299	88.21	3/12	5h23
NEPENTHE(iter #7)	0	0	0	0	0.251	0.314	88.69	4/12	6h11

Table 29: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for RoBERTa on QNLI.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.190	0.201	0.210	0.235	0.275	0.303	92.18	0/12	0h35
IMP(iter #1)	0.198	0.214	0.215	0.260	0.271	0.276	92.07	0/12	1h10
IMP(iter #2)	0.205	0.212	0.255	0.263	0.300	0.325	92.07	0/12	1h45
IMP(iter #3)	0.216	0.223	0.265	0.266	0.325	0.334	91.56	0/12	2h21
IMP(iter #4)	0.227	0.234	0.265	0.278	0.343	0.383	91.65	0/12	2h56
IMP(iter #5)	0.227	0.255	0.268	0.321	0.379	0.380	91.51	0/12	3h31
IMP(iter #6)	0.239	0.259	0.261	0.313	0.368	0.386	90.19	0/12	4h06
IMP(iter #7)	0.248	0.254	0.300	0.346	0.393	0.420	90.08	0/12	4h41
NEPENTHE(iter #1)	0.001	0.001	0.230	0.274	0.426	0.475	88.36	0/12	1h22
NEPENTHE(iter #2)	0	0	0.001	0.028	0.031	0.487	87.41	2/12	2h09
NEPENTHE(iter #3)	0	0	0.001	0.001	0.003	0.322	86.43	2/12	2h56
NEPENTHE(iter #4)	0	0	0	0.001	0.001	0.023	87.26	3/12	3h43
NEPENTHE(iter #5)	0	0	0	0.000	0.001	0.017	86.03	3/12	4h30
NEPENTHE(iter #6)	0	0	0	0.001	0.001	0.006	85.21	3/12	5h17
NEPENTHE(iter #7)	0	0	0	0	0	0	84.37	6/12	6h04

Table 30: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for BERT on RTE.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.211	0.216	0.233	0.242	0.252	0.261	61.01	0/12	0h02
IMP(iter #1)	0.224	0.225	0.237	0.250	0.260	0.276	55.60	0/12	0h03
IMP(iter #2)	0.243	0.244	0.245	0.274	0.288	0.294	62.09	0/12	0h04
IMP(iter #3)	0.258	0.263	0.279	0.286	0.309	0.344	59.21	0/12	0h05
IMP(iter #4)	0.275	0.282	0.321	0.326	0.342	0.386	58.84	0/12	0h06
IMP(iter #5)	0.287	0.293	0.328	0.366	0.369	0.397	58.48	0/12	0h07
IMP(iter #6)	0.303	0.313	0.342	0.388	0.403	0.409	59.21	0/12	0h08
IMP(iter #7)	0.335	0.343	0.370	0.374	0.432	0.451	57.76	0/12	0h09
NEPENTHE(iter #1)	0.204	0.223	0.230	0.233	0.244	0.260	57.04	0/12	0h03
NEPENTHE(iter #2)	0.001	0.153	0.173	0.181	0.264	0.280	52.71	0/12	0h04
NEPENTHE(iter #3)	0	0.001	0.201	0.283	0.298	0.311	54.87	1/12	0h05
NEPENTHE(iter #4)	0	0	0	0.246	0.258	0.266	53.43	3/12	0h06
NEPENTHE(iter #5)	0	0	0	0.007	0.287	0.296	56.32	3/12	0h07
NEPENTHE(iter #6)	0	0	0	0.022	0.296	0.309	54.87	3/12	0h08
NEPENTHE(iter #7)	0	0	0	0.008	0.327	0.339	53.79	3/12	0h09

Table 31: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for RoBERTa on RTE.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.236	0.242	0.263	0.292	0.302	0.303	66,787	0/12	0h02
IMP(iter #1)	0.241	0.243	0.282	0.296	0.314	0.321	72.56	0/12	0h03
IMP(iter #2)	0.253	0.256	0.296	0.341	0.352	0.385	72.92	0/12	0h04
IMP(iter #3)	0.249	0.293	0.304	0.376	0.401	0.409	66.06	0/12	0h05
IMP(iter #4)	0.254	0.291	0.326	0.387	0.434	0.435	67.51	0/12	0h06
IMP(iter #5)	0.262	0.291	0.363	0.418	0.459	0.484	63.54	0/12	0h07
IMP(iter #6)	0.285	0.296	0.385	0.426	0.465	0.497	64.98	0/12	0h08
IMP(iter #7)	0.297	0.305	0.405	0.445	0.492	0.502	60.29	0/12	0h09
NEPENTHE(iter #1)	0.001	0.252	0.263	0.298	0.310	0.325	64.26	0/12	0h03
NEPENTHE(iter #2)	0.001	0.271	0.295	0.322	0.347	0.351	68.23	0/12	0h04
NEPENTHE(iter #3)	0	0.001	0.265	0.295	0.363	0.380	66.06	1/12	0h05
NEPENTHE(iter #4)	0	0	0.326	0.333	0.344	0.349	54.15	2/12	0h06
NEPENTHE(iter #5)	0	0	0.263	0.359	0.392	0.399	59.93	2/12	0h07
NEPENTHE(iter #6)	0	0	0.001	0.001	0.325	0.387	54.15	2/12	0h08
NEPENTHE(iter #7)	0	0	0.001	0.001	0.001	0.390	59.21	2/12	0h09

Table 32: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for BERT on SST-2.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.114	0.122	0.130	0.172	0.178	0.205	92,20	0/12	0h21
IMP(iter #1)	0.119	0.126	0.139	0.190	0.213	0.226	91.63	0/12	0h41
IMP(iter #2)	0.138	0.141	0.161	0.241	0.264	0.267	90.83	0/12	1h01
IMP(iter #3)	0.133	0.144	0.185	0.239	0.280	0.307	90.48	0/12	1h22
IMP(iter #4)	0.142	0.167	0.205	0.251	0.303	0.332	90.37	0/12	1h43
IMP(iter #5)	0.158	0.181	0.219	0.279	0.312	0.376	89.91	0/12	2h04
IMP(iter #6)	0.178	0.215	0.241	0.322	0.361	0.363	88.53	0/12	2h25
IMP(iter #7)	0.187	0.226	0.270	0.365	0.370	0.466	87.96	0/12	2h46
NEPENTHE(iter #1)	0.001	0.132	0.137	0.182	0.224	0.229	91.17	0/12	0h48
NEPENTHE(iter #2)	0.001	0.001	0.191	0.209	0.238	0.246	89.68	0/12	1h16
NEPENTHE(iter #3)	0	0	0	0.193	0.205	0.224	89.00	3/12	1h44
NEPENTHE(iter #4)	0	0	0	0.001	0.204	0.243	88.88	3/12	2h12
NEPENTHE(iter #5)	0	0	0	0.001	0.228	0.254	87.39	3/12	3h01
NEPENTHE(iter #6)	0	0	0	0.001	0.246	0.252	88.76	3/12	3h29
NEPENTHE(iter #7)	0	0	0	0.001	0.245	0.253	87.73	3/12	3h57



Table 33: Test performance (top-1), bottom six layer’s entropies  $\hat{\mathcal{H}}$  and the number of removed layers (Rem.) for RoBERTa on SST-2.

Approach	$\hat{\mathcal{H}}_1$	$\hat{\mathcal{H}}_2$	$\hat{\mathcal{H}}_3$	$\hat{\mathcal{H}}_4$	$\hat{\mathcal{H}}_5$	$\hat{\mathcal{H}}_6$	top-1	Rem.	Time
Dense model	0.131	0.145	0.152	0.187	0.192	0.278	92.66	0/12	0h21
IMP(iter #1)	0.129	0.151	0.153	0.189	0.196	0.276	92.32	0/12	0h41
IMP(iter #2)	0.136	0.155	0.156	0.186	0.226	0.266	92.43	0/12	1h01
IMP(iter #3)	0.159	0.173	0.179	0.195	0.234	0.346	93.23	0/12	1h22
IMP(iter #4)	0.178	0.180	0.215	0.227	0.304	0.374	92.66	0/12	1h43
IMP(iter #5)	0.178	0.199	0.253	0.293	0.319	0.419	92.20	0/12	2h04
IMP(iter #6)	0.177	0.208	0.226	0.280	0.323	0.416	91.17	0/12	2h25
IMP(iter #7)	0.174	0.200	0.220	0.285	0.344	0.428	90.83	0/12	2h46
NEPENTHE(iter #1)	0.001	0.128	0.160	0.161	0.202	0.255	92.32	0/12	0h48
NEPENTHE(iter #2)	0.001	0.001	0.166	0.205	0.216	0.251	92.09	0/12	1h16
NEPENTHE(iter #3)	0	0.001	0.162	0.197	0.283	0.322	90.83	1/12	1h44
NEPENTHE(iter #4)	0	0	0.191	0.198	0.257	0.267	90.60	2/12	2h12
NEPENTHE(iter #5)	0	0	0.191	0.212	0.281	0.345	90.48	2/12	3h01
NEPENTHE(iter #6)	0	0	0.001	0.210	0.293	0.378	90.71	2/12	3h29
NEPENTHE(iter #7)	0	0	0.001	0.215	0.272	0.275	89.68	2/12	3h57