Abstract

- This study extends the comparative analysis of various neural network pruning techniques—specifically SNIP [2], GraSP [4], SynFlow [3], random pruning, and magnitude based pruning—by integrating modern evaluation metrics and introducing an iterative pruning method, Iterative Magnitude Pruning (IMP) [1].
- Our objectives are to enhance the understanding and efficiency of these techniques for more effective neural network model development.
- We implemented IMP to evaluate its theoretical effectiveness and incorporated additional metrics such as CPU memory usage, GPU allocation, and cache memory tracking.
- Our comparative analysis across different compression levels reveals that iterative pruning methods like IMP tend to outperform one-shot approaches. Furthermore, initial results suggest that each one-shot pruning method presents distinct advantages and limitations. This comprehensive assessment aids in identifying optimal pruning strategies for various network architectures and applications.

Iterative Magnitude Pruning

- Iterative Magnitude Pruning (IMP) [1] is a neural network pruning technique that employs an iterative process to re move weights based on their magnitudes. It seeks to identify a sparse but capable subnetwork that, when trained from the beginning, could match or surpass the performance of the unpruned network. IMP is inspired by the Lottery Ticket Hypothesis, which suggests that effective subnetworks can exist within randomly initialized networks.
- IMP differs significantly from traditional one-shot pruning methods. While one-shot pruning involves removing a pre determined percentage of weights based solely on a single pass or criterion (such as weight magnitude), IMP applies a more nuanced approach. It uses multiple iterations of pruning followed by training, where each cycle aims to eliminate a fixed percentage of the smallest weights and then retrain the network to regain performance. This methodical reduction

Iterative Magnitude Pruning (Cont'd)

and training process allows IMP to refine the network's structure iteratively, enhancing its ability to maintain or improve performance despite increased sparsity.

of each method in reducing computational complexity. IMP

(Iterative Magnitude Pruning) shows a consistent and signifi-

cant increase in sparsity with more iterations, highlighting its

effectiveness in achieving higher computational savings.

• This cyclic nature of IMP is crucial for its success. It allows the network to adapt gradually to the loss of weight, which can prevent the significant performance degradation often ob served with one-shot pruning methods after aggressive weight removal. By continuously adjusting and retraining, IMP can discover more efficient and robust network configurations capable of achieving similar or even superior performance compared to the original unpruned model.

Compression	n Rand	Mag	SNIP	GraSP	SynFlow	Compression	Rand	Mag	SNIP	GraSP	• SynFlow
0.05	88.08	88.69	88.20	79.03	88.05	0.05	0.653	0.630	0.704	0.598	0.617
0.1	87.51	89.47	88.17	72.34	88.16	0.1	0.705	0.676	0.606	0.621	0.620
0.2	88.16	89.36	87.87	78.87	88.50	0.2	0.631	0.637	0.600	0.618	0.682
0.5	86.80	89.95	88.55	80.86	87.14	0.5	0.693	0.639	0.602	0.618	0.603
1	10.00	88.88	87.77	81.46	87.83	1	0.606	0.604	0.606	0.613	0.610
2	10.00	42.61	81.55	82.72	10.00	2	0.621	0.602	0.668	0.627	0.618
Compression	n IMP (2 Iters)	IMP (3	Iters)	IMP (4 Iters)	Compression	IMP (2	2 Iters)	IMP (3	Iters)	IMP (4 Iters
0.05	88	8.56	88.6	55	89.81	0.05	0.7	43	0.70)0	0.689
0.1	89	0.15	88.6	55	89.06	0.1	0.6	46	0.63	34	0.653
0.2	88	3.84	89.2	20	89.69	0.2	0.6	38	0.65	55	0.666
0.5	89	0.14	89.3	37	88.94	0.5	0.7	40	0.67	72	0.641
1	89	89.52		57	89.80	1	0.6	49	0.69	98	0.641
2	81	.60	75.37		19.34	2	0.638		0.653		0.667
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability	present th FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta	cross var ng metho ve Magni consistent mpressio in accura	ious com ds (Rand tude Prun ly outper on levels, acy while	pression l, Mag, S ning (IM forms ot demonst reducing	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning me GraSP, SynFlo tently shows op iterations, which work operation	oss vary ethods, in w) and it ptimized ch highlins post-p	ving con ncluding erative (inference ghts its runing.	npression g one-sho IMP) app ce times, efficiency	n ratios ot (Rand, proaches. particula y in strea	using differ- Mag, SNIP, IMP consis- rly at higher mlining net-
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability	present th FAR-10 a shot prunin ith Iterativ ons. IMP c thigher co to mainta	cross var ng metho ve Magni consistent mpressio in accura	ious com ds (Rand tude Prun ly outper n levels, acy while	pression l, Mag, S ning (IM forms ot demonst reducing	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning me GraSP, SynFlo tently shows op iterations, white work operation	ross vary ethods, in w) and it ptimized ch highli ns post-p	ving con ncluding erative (inference ghts its runing.	npressior g one-sho IMP) app ce times, efficiency	n ratios ot (Rand, oroaches, particula y in strea	using differ- Mag, SNIP, IMP consis- rly at higher mlining net-
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability	FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand	ross var ng metho ve Magni consistent mpressio in accura Mag	ious com ds (Rand tude Prut ly outper n levels, acy while SNIP	pression l, Mag, S ning (IM forms ot demonst reducing GraSI	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning me GraSP, SynFlo tently shows op iterations, white work operation	ross vary ethods, in w) and it ptimized ch highli ns post-p Rand	ving com ncluding erative (inference ghts its runing. Mag	npression g one-sho IMP) app ce times, efficiency SNIP	n ratios ot (Rand, oroaches. particula y in strea GraSP	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- SynFlow
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05	Fresent th FAR-10 a shot prunin ith Iterativo ons. IMP c higher co to mainta Rand 0.8916	cross var ng metho ve Magni consistent mpressio in accura Mag 0.9477	ious com ds (Rand tude Prut ly outper on levels, icy while SNIP 0.9377	pression I, Mag, S ning (IM forms ot demonst reducing GraSI 0.820 0.720	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning ma GraSP, SynFlo tently shows op iterations, white work operation	ross vary ethods, ir w) and it ptimized ch highli ns post-p Rand 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8	npression g one-sho IMP) app ce times, efficiency SNIP 16.0	ratios ot (Rand, proaches, particula y in strea GraSP 16.1	using differ- Mag, SNIP, IMP consis- rrly at higher mlining net- SynFlow 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1	Fresent th FAR-10 a shot prunin ith Iterativo ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310	Mag 0.9477 0.8991	ious com ds (Rand tude Pru ly outper n levels, acy while SNIP 0.9377 0.9268 0.7812	GraSI 0.820 0.729 0.571	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning ma GraSP, SynFlo tently shows op iterations, white work operation Compression 0.05 0.1	ross vary ethods, ir w) and it ptimized ch highli ns post-p Rand 10.0 10.1	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0	ratios ot (Rand, proaches, particula y in strea GraSP 16.1 13.0	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- SynFlow 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5	FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0 3165	ross var ng metho ve Magni consistent ompressio in accura Mag 0.9477 0.8991 0.8151 0.5634	ious com ds (Rand tude Prui ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622	GraSI 0.820 0.729 0.571 0.378	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size.	CIFAR-10 acr ent pruning me GraSP, SynFlor tently shows op iterations, white work operation Compression 0.05 0.1 0.2	ross vary ethods, ir w) and it ptimized ch highli as post-p Rand 10.0 10.1 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0 9.8	GraSP 16.1 13.0 13.0	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- SynFlow 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1	FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009	Mag 0.9477 0.8991 0.5634 0.2283	ious com ds (Rand tude Prun ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979	GraSI 0.820 0.729 0.371 0.378 0.175	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571	CIFAR-10 acr ent pruning me GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5	ross vary ethods, in w) and it ptimized ch highlins post-p Rand 10.0 10.1 10.0 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.1	sNIP 16.0 13.0 9.8 9.8	GraSP 16.1 13.0 9.8	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- SynFlow 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2	FAR-10 a shot prunin ith Iterativ ons. IMP c thigher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108	Mag 0.9477 0.8991 0.5634 0.0322	ious com ds (Rand tude Prui ly outper in levels, icy while SNIP 0.9377 0.9268 0.7812 0.4622 0.4622 0.1979 0.0411	GraSI 0.820 0.729 0.571 0.378 0.058	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 5 0.1845	CIFAR-10 acr ent pruning me GraSP, SynFlov tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1	Rand Rand 10.0 10.0 10.0 10.0 10.0 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.1 9.8	sNIP 16.0 13.0 9.8 9.8 9.8	GraSP 16.1 13.0 9.8 9.5	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c thigher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108	Mag 0.9477 0.8991 0.8151 0.5634 0.0322	ious com ds (Rand tude Prun ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411	GraSI 0.820 0.729 0.371 0.378 0.175 0.058	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 5 0.1845	CIFAR-10 acr ent pruning ma GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2	Rand Rand 10.0 10.0 10.0 10.0 10.0 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8	sNIP 16.0 13.0 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2	Mag 0.9477 0.8991 0.5634 0.2283 0.0322	ious com ds (Rand tude Prui ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.4622 0.1979 0.0411 IMP (3	GraSI 0.820 0.729 0.571 0.378 0.175 0.058	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 6 0.1845 IMP (4 Iters)	CIFAR-10 acr ent pruning ma GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression	Rand 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 9.8	sNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 14.1 15.0 15.0 15.0 15.0 15.0 15.0 15.0 15	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c ithigher co it to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2 0.92	Mag 0.9477 0.8991 0.8151 0.5634 0.2283 0.0322 2 Iters) 290	ious com ds (Rand tude Prui ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411 IMP (3 0.922	GraSI 0.820 0.729 0.571 0.378 0.175 0.058 Iters)	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 5 0.1845 IMP (4 Iters) 0.9229 0.9229	CIFAR-10 acr ent pruning ma GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression	Rand Rand 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 2 Iters)	SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 1MP (3	GraSP 16.1 13.0 9.5 9.5 Iters)	using differ- Mag, SNIP, IMP consis- rly at higher unlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.05 0.1 0.05 0.1 0.05	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c ithigher coordinate ithigher coordithigher coordin	Mag 0.9477 0.8991 0.8151 0.5634 0.2283 0.0322 2 Iters) 290 655	ious com ds (Rand tude Pru- ly outper on levels, icy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411 IMP (3 0.925 0.86	GraSI 0.820 0.729 0.571 0.378 0.175 0.058	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 5 0.1845 IMP (4 Iters) 0.9229 0.8560 0.7414	CIFAR-10 acr ent pruning mo GraSP, SynFlov tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression 0.05	ross vary ethods, in w) and it ptimized ch highli ns post-p Rand 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 2 Iters) .8	IMP) app ce times, efficiency SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 1ters)	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 0.1 0.2 0.5 0.1 0.2 0.5 0.1 0.2 0.5	Present th FAR-10 a shot prunin ith Iterative ons. IMP c higher coo to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2 0.92 0.88 0.72 0.5	Mag 0.9477 0.8991 0.8151 0.5634 0.2283 0.0322 2 Iters) 290 555 564	ious com ds (Rand tude Pru- ly outper in levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411 IMP (3 0.92: 0.86 0.750 0.51/	pression I, Mag, S ning (IM forms ot demonst reducing GraSI 0.820 0.729 0.571 0.378 0.175 0.058 Iters) 56 12 05 25	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.6423 1 0.6423 1 0.4571 6 0.1845 IMP (4 Iters) 0.9229 0.8560 0.7414 0.5115	CIFAR-10 acr ent pruning ma GraSP, SynFlov tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 2 Compression	Rand 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.1 9.8 9.8 2 Iters) .8	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 1ters) 8	using differ- Mag, SNIP, IMP consis- rly at higher mlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 Compression	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2 0.92 0.86 0.75 0.5 0.2	Mag 0.9477 0.8991 0.8151 0.5634 0.2832 0.0322 2 Iters) 290 555 564 189	ious com ds (Rand tude Prui ly outper in levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.4622 0.4622 0.1079 0.0411 IMP (3 0.925 0.86 0.750 0.512	GraSI 0.820 0.729 0.571 0.378 0.175 0.058 Iters) 56 12 05 25 03	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 6 0.1845 IMP (4 Iters) 0.9229 0.8560 0.7414 0.5115 0.2280	CIFAR-10 acr ent pruning ma GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.05 0.1 0.05 0.1 0.05 0.1 0.2 0.5	ross vary ethods, ir w) and it ptimized ch highlins post-p Rand 10.0 10.1 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 9.8 2 Iters) .8 .8	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 1ters) 8	using differ- Mag, SNIP, IMP consis- rrly at higher mlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 2	present th FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2 0.92 0.86 0.75 0.2	Mag 0.9477 0.8991 0.8151 0.5634 0.2283 0.0322 2 Iters) 290 555 564 189 134 286	ious com ds (Rand tude Prui ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411 IMP (3 0.922 0.86 0.750 0.512 0.219	GraSI 0.820 0.729 0.571 0.378 0.175 0.058 Iters) 56 12 05 25 93 45	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 6 0.1845 IMP (4 Iters) 0.9229 0.8560 0.7414 0.5115 0.2280 0.0239	CIFAR-10 acr ent pruning mo GraSP, SynFlor tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 0.1 0.2 0.5	ross vary ethods, ir w) and it ptimized ch highli ns post-p Rand 10.0 10.1 10.0 10.0 10.0 10.0 10.0 10.	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 2 Iters) .8 .8 .8 .8	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 Iters) 8	using differ- Mag, SNIP, IMP consis- rly at higher unlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5
Table 1: We model on CI paring one-s SynFlow) w and 4 iteration especially at perior ability Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 2	Present th FAR-10 a shot prunin ith Iterativ ons. IMP c higher co to mainta Rand 0.8916 0.7945 0.6310 0.3165 0.1009 0.0108 IMP (2 0.92 0.86 0.72 0.5 0.2 0.02	Mag 0.9477 0.8991 0.8151 0.5634 0.2283 0.0322 2 Iters) 290 655 564 189 134 286	ious com ds (Rand tude Prun ly outper on levels, acy while SNIP 0.9377 0.9268 0.7812 0.4622 0.1979 0.0411 IMP (3 0.92: 0.86 0.750 0.512 0.219	pression I, Mag, S ning (IM forms ot demonst reducing GraSI 0.820 0.729 0.571 0.378 0.175 0.058 Iters) 56 12 05 25 93 45	ratios, com- SNIP, GraSP, P) over 2, 3, her methods, rating its su- g model size. P SynFlow 1 0.9488 5 0.9026 1 0.8217 1 0.6423 1 0.4571 5 0.1845 IMP (4 Iters) 0.9229 0.8560 0.7414 0.5115 0.2280 0.0239	CIFAR-10 acr ent pruning mo GraSP, SynFlot tently shows op iterations, white work operation 0.05 0.1 0.2 0.5 1 2 Compression 0.05 0.1 0.2 0.5 1 2 Compression	Rand 10.0	ving com ncluding erative (inference ghts its runing. Mag 9.8 10.0 10.0 10.0 10.1 9.8 9.8 2 Iters) .8 .8 .8 .8 .4 .4	npression g one-sho IMP) app ce times, efficiency SNIP 16.0 13.0 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8 9.8	GraSP 16.1 13.0 9.5 9.5 1ters) 8 8 4 4	using differ- Mag, SNIP, IMP consis- rly at higher unlining net- 9 SynFlow 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5 9.5

Enhancing Network Pruning Method Evaluation and Integration

Yulai Zhao

Princeton University

VGG16 models at different compression levels. The data demonstrates how iterative pruning (IMP) maintains consistent and lower memory usage across iterations compared to one-shot methods

Main Results (Cont'd)



Figure 1: Comparative visualization of weight distributions across various pruning techniques for a VGG16 model trained on CIFAR-10, targeting 50% compression. The plots demonstrate distinct sparsity patterns: Random (Rand) and Magnitude-based (Mag) pruning show less strategic weight removal than structured approaches like SNIP, GraSP, and SynFlow, which more aggressively prune the later layers. Iterative Magnitude Pruning (IMP) over multiple iterations (2 to 4 Iters) refines sparsity, progressively concentrating weight reductions in the final layers. This suggests an adaptive focus on maintaining early layer density for feature extraction while optimizing later layers for decision-making efficiency.

Main Results (Cont'd)

Rand	Mag	SNIP	GraS	P SynFlow		
0.375	0.376	0.437	0.37	8 0.374		
0.375	0.376	0.437	0.378	8 0.374		
0.375	0.376	0.437	0.37	8 0.374		
0.375	0.376	0.437	0.378	8 0.374		
0.375	0.376	0.437	0.378	8 0.374		
0.375	0.376	0.437	0.378	8 0.374		
IMP (2	l Iters)	IMP (3	Iters)	IMP (4 Iters)		
0.4	94	0.49	5	0.497		
0.4	.99	0.49	9	0.498		
0.4	94	0.49	9	0.494		
0.4	97	0.49	9	0.497		
0.4	97	0.49	9	0.497		
0.4	94	0.49	6	0.498		

Table 5: Maximum GPU memory allocation in GBs for inferencing on the CIFAR-10 test set using VGG16 models pruned via various techniques at different compression ratios. Compared to one-shot pruning methods, IMP demonstrates higher memory efficiency, particularly at finer iterations.

Compression	Rand	Mag	SNIP	GraSP	SynFlow	
0.05	1.124	0.937	1.097	1.267	1.114	
0.1	1.124	0.937	1.097	1.267	1.114	
0.2	1.124	0.937	1.097	1.267	1.114	
0.5	1.124	0.937	1.097	1.267	1.114	
1	1.124	0.937	1.097	1.267	1.114	
2	1.124	0.937	1.097	1.267	1.114	
Compression	IMP (2	l Iters)	IMP (3	Iters)	IMP (4 Iters)	
0.05	1.277		1.277		1.277	
0.1	1.277		1.277		1.277	
0.2	1.277		1.140		1.277	
0.5	1.140		1.277		1.277	
1	1.141		1.141		1.277	
2	1.277		1.277		1.277	

Table 6: Maximum GPU memory cached for inferencing on the CIFAR-10 test set with VGG16 models pruned at various compression ratios.

Conclusion

• Our comparative study underscores the superiority of iterative pruning over traditional one-shot methods.

 IMP retains model accuracy, reduces inference times, and optimizes memory consumption across both CPU and GPU, making it a robust solution for enhancing the operational efficiency of deep neural networks.

• This study highlights the potential of iterative pruning techniques in advancing the SOTA model compression, offering significant benefits for real-world applications where efficiency and performance are critical.

References

[1] FRANKLE, J., DZIUGAITE, G. K., ROY, D., AND CARBIN, M. Linear mode connectivity and the lottery ticket hypothesis. In International Conference on Machine Learning (2020), PMLR, pp. 3259-3269.

[2] LEE, N., AJANTHAN, T., AND TORR, P. H. Snip: Single-shot network pruning based on connection sensitivity. arXiv preprint arXiv:1810.02340 (2018). [3] TANAKA, H., KUNIN, D., YAMINS, D. L., AND GANGULI, S. Pruning neural networks without any data by iteratively conserving synaptic flow. Advances in neural information processing systems 33 (2020), 6377–6389.

[4] WANG, C., ZHANG, G., AND GROSSE, R. Picking winning tickets before training by preserving gradient flow. In International Conference on Learning Representations (2019).