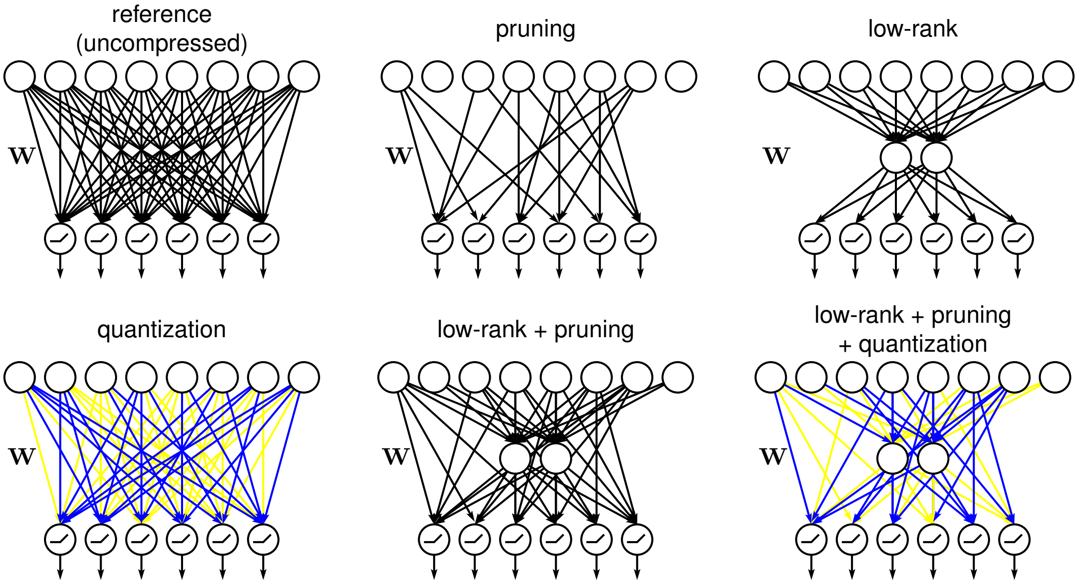
A flexible, extensible software framework for model compression based on the LC algorithm

Yerlan Idelbayev and **Miguel Á. Carreira-Perpiñán**
Department of CSE, University of California, Merced
`http://eecs.ucmerced.edu`

The code is available at:
`https://github.com/UCMerced-ML/LC-model-compression`

The fundamental problem of model compression: what to choose?



Images are from the slides of Miguel Á. Carreira-Perpiñán

Challenges

In principle, we want to explore all possible combinations, and select the best. But:

- Many compression schemes $\Longrightarrow$ many algorithms
- How to maintain a library of many compressions?
- How to make it user friendly?
  - many algorithms $\Longrightarrow$ many failure points
- How to make it extensible and easily maintainable?

We propose a software based on the Learning-Compression (LC) algorithm:

- single algorithm—many compressions
- extensible, modular, and fast
- impressive compression results
- open source: `BSD 3-clause` license

The LC algorithm: formulation

Given a network with weights $\mathbf{w}$ and loss L:

uncompressed weights

low-dim. params

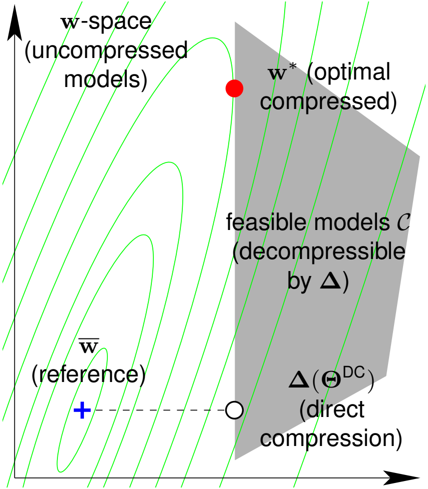
$$\min_{\mathbf{w},\boldsymbol{\Theta}} L(\mathbf{w}) \quad \text{s.t.} \quad \mathbf{w} = \boldsymbol{\Delta}(\boldsymbol{\Theta}) \qquad (1)$$

task loss

decompression mapping
$\boldsymbol{\Delta}\colon \boldsymbol{\Theta} \to \mathbf{w} \in \mathbb{R}^P$

The compression details are abstracted in $\boldsymbol{\Delta}(\boldsymbol{\Theta})$:

- e.g., low-rank: $\boldsymbol{\Delta}(\boldsymbol{\Theta}) = \mathbf{U}\mathbf{V}^T$ where $\boldsymbol{\Theta} = \{\mathbf{U}, \mathbf{V}\}$



feasible set $\mathcal{C} = \{\mathbf{w} \in \mathbb{R}^P\colon\ \mathbf{w} = \boldsymbol{\Delta}(\boldsymbol{\Theta})$ for $\boldsymbol{\Theta} \in \mathbb{R}^Q\}$
figure from the slides of Miguel Á. Carreira-Perpiñán

The LC algorithm (cont.)

The problem (1) can be solved by alternation of these two steps (while driving $\mu \to \infty$), which form the basis of our software:

- Learning (L) step:

$$\min_{\mathbf{w}} L(\mathbf{w}) + \frac{\mu}{2} \|\mathbf{w} - \boldsymbol{\Delta}(\boldsymbol{\Theta})\|^2$$

  - This is a regular training of the model, but with a quadratic regularization term
  - When you train a network, you already have the L step.

- Compression (C) step:

$$\min_{\boldsymbol{\Theta}} \|\mathbf{w} - \boldsymbol{\Delta}(\boldsymbol{\Theta})\|^2$$

  - Independent of the loss, neural network structure, and the dataset.
  - We provide a library of different C steps for many different compressions.

The library of implemented compressions

Type	Forms
Quantization	Adaptive Quantization into $\{c_1, c_2, \dots c_K\}$ Binarization into $\{-1, 1\}$ and $\{-c, c\}$ Ternarization into $\{-c, 0, c\}$
Pruning	ℓ_0-constraint (s.t., $\|\mathbf{w}\|_0 \leq \kappa$) ℓ_1-constraint (s.t., $\|\mathbf{w}\|_0 \leq \kappa$) ℓ_0-penalty ($\alpha\|\mathbf{w}\|_0$) ℓ_1-penalty ($\alpha\|\mathbf{w}\|_1$)
Low-rank	Low-rank compression to a given rank Low-rank with *automatic* rank selection for FLOPs reduction Low-rank with *automatic* rank selection for storage compression
Additive Combinations	Quantization + Pruning Quantization + Low-rank Pruning + Low-rank Quantization + Pruning + Low-rank

Easy exploration of compressions

Having an L-step implementation (you only need one), definition of compression is very simple:

quantize each layer with separate codebooks

```
compression_tasks = {
  Param(l1.weight): (AsVector, AdaptiveQuantization(k=2)),
  Param(l2.weight): (AsVector, AdaptiveQuantization(k=2)),
  Param(l3.weight): (AsVector, AdaptiveQuantization(k=2))
}
```

---

prune all but 5%

```
compression_tasks = {
  Param([l1.weight, l2.weight, l3.weights]):
    (AsVector, ConstraintL0Pruning(kappa=13310)) # 13310 = 5%
}
```

---

prune first layer, low-rank to second, quantize third

```
compression_tasks = {
  Param(l1.weight): (AsVector, ConstraintL0Pruning(kappa=5000)),
  Param(l2.weight): (AsIs,     LowRank(target_rank=10))
  Param(l3.weight): (AsVector, AdaptiveQuantization(k=2))
}
```

Example: Low-rank AlexNet models with automatic rank selection

Our framework achieves competitive results in many compression schemes.
For example, using our code for rank-selection, we can achieve considerable speed-up on AlexNet:

	MFLOPs	top-1	top-5	ρ_{FLOPs}
Caffe-AlexNet [1]	724	42.70	**19.80**	1.00
our, scheme 2, (L_1)	238	41.81	19.40	3.01
our, scheme 2, (L_2)	190	42.07	19.54	3.81
our, scheme 2, (L_3)	**151**	**42.69**	19.83	**4.79**
Kim et al. [2], Tucker	272	n/a	21.67	2.66
Tai et al. [3], scheme 2	185	n/a	20.34	3.90
Wen et al. [4], scheme 1	269	n/a	20.14	2.69
Kim et al. [5], scheme 2	272	43.40	20.10	2.66
Yu et al. [6], filter prun.	232	44.13	n/a	3.12
Li et al. [7], filter prun.	334	43.17	n/a	2.16
Ding et al. [8], filter prun.	492	43.83	20.47	1.47

ρ_{FLOPs} — reduction in FLOPs. see Idelbayev and Carreira-Perpiñán [9] for full details

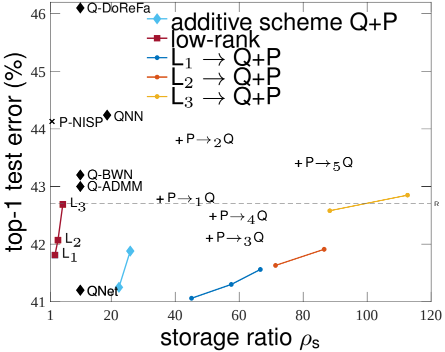
Not only theoretical reduction!

Model	GPU of Jetson Nano time, ms	speed-up
AlexNet	23.36	1.00
L_1	11.59	2.01
L_2	8.88	2.63
L_3	7.11	3.29

Example: Additive compressions to achieve smallest AlexNet-s

The codebase allows easy exploration of new compression mechanisms. For example, we can further compress low-rank AlexNet models to target storage:

	Model	top-1	size, MB	MFLOPs
	Caffe-AlexNet Jia et al. [1]	42.70	243.5	724
our	$L_1 \to$ Q (1-bit) + P (0.25M)	**41.56**	3.7	238
our	$L_2 \to$ Q (1-bit) + P (0.25M)	41.91	2.8	190
our	$L_3 \to$ Q (1-bit) + P (0.25M)	42.85	**2.2**	**151**
	AlexNet-QNN of Wu et al. [10]	44.24	13.0	175
	$P{\to_1}Q$ of Han et al. [11]	42.78	6.9	724
	$P{\to_2}Q$ of Choi et al. [12]	43.80	5.9	724
	$P{\to_3}Q$ of Tung and Mori [13]	42.10	4.8	724
	$P{\to_4}Q$ of Yang et al. [14]	42.48	4.7	724
	$P{\to_5}Q$ of Yang et al. [14]	43.40	3.1	724
	filter pruning of Li et al. [7]	43.17	232.0	334



Source code and library features

Our code is written in `Python` using `PyTorch`, and open source under `BSD 3-clause` license:

`https://github.com/UCMerced-ML/LC-model-compression`

Using the provided code, you will be able to:

- replicate all reported experiments
- compress your own models with many available compression schemes

Our library is:

- modular and easily extensible
- only requires the L-step implementation: the regular learning of the model (using SGD)
- based on solid optimization principles
- single algorithm—many compressions
- time proven (development since 2017), with many publications [9, 15, 16, 17, 18, 19, 20, 21]

References

[1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," Jun. 20 2014, arXiv:1408.5093 [cs.CV].

[2] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," in *Proc. of the 4th Int. Conf. Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, May 2–4 2016.

[3] C. Tai, T. Xiao, Y. Zhang, X. Wang, and W. E, "Convolutional neural networks with low-rank regularization," in *Proc. of the 4th Int. Conf. Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, May 2–4 2016.

[4] W. Wen, C. Xu, C. Wu, Y. Wang, Y. Chen, and H. Li, "Coordinating filters for faster deep neural networks," in *Proc. 17th Int. Conf. Computer Vision (ICCV'17)*, Venice, Italy, Dec. 11–18 2017.

[5] H. Kim, M. U. K. Khan, and C.-M. Kyung, "Efficient neural network compression," in *Proc. of the 2019 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'19)*, Long Beach, CA, Jun. 16–20 2019, pp. 12 569–12 577.

[6] R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V. I. Morariu, X. Han, M. Gao, C.-Y. Lin, and L. S. Davis, "NISP: Pruning networks using neuron importance score propagation," in *Proc. of the 2018 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'18)*, Salt Lake City, UT, Jun. 18–22 2018, pp. 9194–9203.

[7] J. Li, Q. Qi, J. Wang, C. Ge, Y. Li, Z. Yue, and H. Sun, "OICSR: Out-in-channel sparsity regularization for compact deep neural networks," in *Proc. of the 2019 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'19)*, Long Beach, CA, Jun. 16–20 2019, pp. 7046–7055.

[8] X. Ding, G. Ding, Y. Guo, J. Han, and C. Yan, "Approximated oracle filter pruning for destructive CNN width optimization," in *Proc. of the 36th Int. Conf. Machine Learning (ICML 2019)*, K. Chaudhuri and R. Salakhutdinov, Eds., Long Beach, CA, Jun. 9–15 2019, pp. 1607–1616.

[9] Y. Idelbayev and M. Á. Carreira-Perpiñán, "Low-rank compression of neural nets: Learning the rank of each layer," in *Proc. of the 2020 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'20)*, Seattle, WA, Jun. 14–19 2020, pp. 8046–8056.

[10] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. of the 2016 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'16)*, Las Vegas, NV, Jun. 26 – Jul. 1 2016, pp. 4020–4028.

[11] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. of the 4th Int. Conf. Learning Representations (ICLR 2016)*, San Juan, Puerto Rico, May 2–4 2016.

[12] Y. Choi, M. El-Khamy, and J. Lee, "Towards the limit of network quantization," in *Proc. of the 5th Int. Conf. Learning Representations (ICLR 2017)*, Toulon, France, Apr. 24–26 2017.

[13] F. Tung and G. Mori, "CLIP-Q: Deep network compression learning by in-parallel pruning-quantization," in *Proc. of the 2018 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'18)*, Salt Lake City, UT, Jun. 18–22 2018, pp. 7873–7882.

[14] H. Yang, S. Gui, Y. Zhu, and J. Liu, "Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach," in *Proc. of the 2020 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'20)*, Seattle, WA, Jun. 14–19 2020, pp. 2175–2185.

[15] M. Á. Carreira-Perpiñán, "Model compression as constrained optimization, with application to neural nets. Part I: General framework," Jul. 5 2017, arXiv:1707.01209.

[16] M. Á. Carreira-Perpiñán and Y. Idelbayev, "Model compression as constrained optimization, with application to neural nets. Part II: Quantization," Jul. 13 2017, arXiv:1707.04319.

[17] ——, ""Learning-compression" algorithms for neural net pruning," in *Proc. of the 2018 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR'18)*, Salt Lake City, UT, Jun. 18–22 2018, pp. 8532–8541.

[18] Y. Idelbayev and M. Á. Carreira-Perpiñán, "A flexible, extensible software framework for model compression based on the LC algorithm," May 15 2020, arXiv:2005.07786.

[19] ——, "More general and effective model compression via an additive combination of compressions," 2020, submitted.

[20] ——, "Neural network compression via additive combination of reshaped, low-rank matrices," in *Proc. Data Compression Conference (DCC 2021)*.

[21] ——, "Optimal selection of matrix shape and decomposition scheme for neural network compression," in *Proc. of the IEEE Int. Conf. Acoustics, Speech and Sig. Proc. (ICASSP'21)*, Toronto, Canada, Mar. 21–25 2021.