Attention Based Pruning for Shift Networks

Ghouthi BOUKLI HACENE, Carlos Lassance, Vincent Gripon,
Matthieu Courbariaux and Yoshua Bengio

Mila

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

January 13th, 2021

Pruning

Layer 1

w_1^1	w_2^1	w_3^1	w_4^1	w_5^1
w_6^1	w_7^1	w_8^1	w_9^1	w_{10}^1
w_{11}^1	w_{12}^1	w_{13}^1	w_{14}^1	w_{15}^1
w_{16}^1	w_{17}^1	w_{18}^1	w_{19}^1	w_{20}^1
w_{21}^1	w_{22}^1	w_{23}^1	w_{24}^1	w_{25}^1

$\longrightarrow$

Layer 2

w_1^2	w_2^2	w_3^2	w_4^2	w_5^2
w_6^2	w_7^2	w_8^2	w_9^2	w_{10}^2
w_{11}^2	w_{12}^2	w_{13}^2	w_{14}^2	w_{15}^2
w_{16}^2	w_{17}^2	w_{18}^2	w_{19}^2	w_{20}^2
w_{21}^2	w_{22}^2	w_{23}^2	w_{24}^2	w_{25}^2

Baseline

Pruning

Layer 1

0	w^1_2	0	w^1_4	w^1_5
0	0	w^1_8	w^1_9	w^1_{10}
w^1_{11}	0	w^1_{13}	w^1_{14}	0
w^1_{16}	w^1_{17}	0	w^1_{19}	0
w^1_{21}	w^1_{22}	0	w^1_{24}	w^1_{25}

$\longrightarrow$

Layer 2

0	0	0	0	0
w^2_6	w^2_7	w^2_8	w^2_9	w^2_{10}
w^2_{11}	w^2_{12}	w^2_{13}	w^2_{14}	w^2_{15}
0	0	0	0	0
w^2_{21}	w^2_{22}	w^2_{23}	w^2_{24}	w^2_{25}

Non structured

Structured

Pruning

Pruning

- Evaluate the importance of neurons and eliminate the least important ones to reduce neural network size.
- Non structured pruning: eliminate neurons independently, only exploitable for very large levels of sparsity.
- Structured pruning: eliminate kernels, filters or even layers, exploitable for even low levels of sparsity.

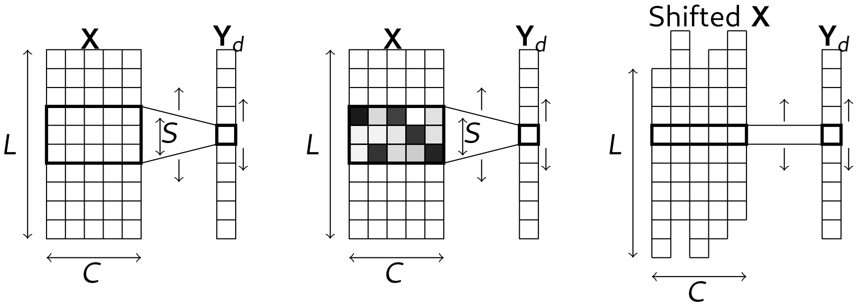
Pruning

- Evaluate the importance of neurons and eliminate the least important ones to reduce neural network size.
- Non structured pruning: eliminate neurons independently, only exploitable for very large levels of sparsity.
- Structured pruning: eliminate kernels, filters or even layers, exploitable for even low levels of sparsity.

Pruning

- Evaluate the importance of neurons and eliminate the least important ones to reduce neural network size.
- Non structured pruning: eliminate neurons independently, only exploitable for very large levels of sparsity.
- Structured pruning: eliminate kernels, filters or even layers, exploitable for even low levels of sparsity.

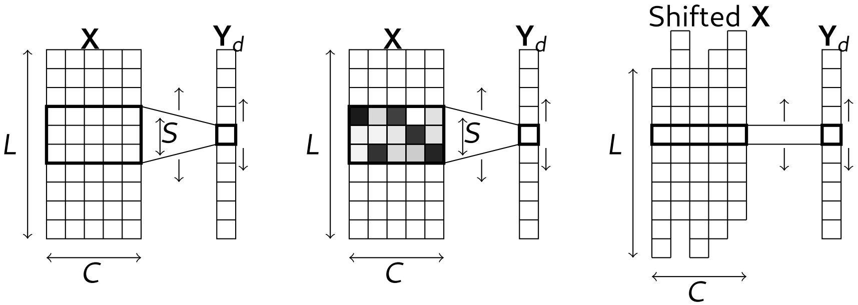
Structured pruning and shift layers



Shift Attention Layer (SAL)

- Simplified operations,
- Reduced number of parameters,
- Fully exploitable technique.

Structured pruning and shift layers



Shift Attention Layer (SAL)

- Simplified operations,
- Reduced number of parameters,
- Fully exploitable technique.

Attention Shift Layer

$$\begin{array}{|c|c|c|}\hline w_1 & w_2 & w_3 \\ \hline w_4 & w_5 & w_6 \\ \hline w_7 & w_8 & w_9 \\ \hline \end{array} \times \begin{array}{|c|c|c|}\hline a_1 & a_2 & a_3 \\ \hline a_4 & a_5 & a_6 \\ \hline a_7 & a_8 & a_9 \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline \tilde{w}_1 & \tilde{w}_2 & \tilde{w}_3 \\ \hline \tilde{w}_4 & \tilde{w}_5 & \tilde{w}_6 \\ \hline \tilde{w}_7 & \tilde{w}_8 & \tilde{w}_9 \\ \hline \end{array}$$

Figure: Using attention tensor A to select which weight should be kept

Conv Operation

$$
\begin{array}{|c|c|c|c|c|c|}
\hline
8 & 2 & 4 & 6 & 1 & 1 \\ \hline
3 & 2 & 0 & 9 & 7 & 4 \\ \hline
0 & 6 & 4 & 0 & 0 & 3 \\ \hline
5 & 1 & 2 & 0 & 3 & 4 \\ \hline
1 & 3 & 4 & 0 & 2 & 1 \\ \hline
0 & 2 & 3 & 1 & 2 & 2 \\ \hline
\end{array}
\ast
\begin{array}{|c|c|c|}
\hline
2 & 0 & 0 \\ \hline
0 & 0 & 0 \\ \hline
0 & 0 & 0 \\ \hline
\end{array}
=
\begin{array}{|c|c|c|c|}
\hline
16 & 4 & 8 & 12 \\ \hline
6 & 4 & 0 & 18 \\ \hline
0 & 12 & 8 & 0 \\ \hline
10 & 2 & 4 & 0 \\ \hline
\end{array}
$$

Shift Operation

$$
\begin{array}{|c|c|c|c|c|c|}
\hline
8 & 2 & 4 & 6 & 1 & 1 \\ \hline
3 & 2 & 0 & 9 & 7 & 4 \\ \hline
0 & 6 & 4 & 0 & 0 & 3 \\ \hline
5 & 1 & 2 & 0 & 3 & 4 \\ \hline
1 & 3 & 4 & 0 & 2 & 1 \\ \hline
0 & 2 & 3 & 1 & 2 & 2 \\ \hline
\end{array}
\times \boxed{2} =
\begin{array}{|c|c|c|c|}
\hline
16 & 4 & 8 & 12 \\ \hline
6 & 4 & 0 & 18 \\ \hline
0 & 12 & 8 & 0 \\ \hline
10 & 2 & 4 & 0 \\ \hline
\end{array}
$$

Experimental Results

Table: Comparison of accuracy, number of parameters and FLOPs between a standard CNN, SAL and vanilla Shiftnet on ImageNet ILSVRC 2012.

		Top-1	Params	FLOPs
Large budget	ResNet-w24 (CLs)	63.47%	**3.2**M	664M
	ShiftNet-A	70.1%	4.1M	1.4G
	ResNet-w64 + SAL	**71**%	3.3M	**538**M
Small budget	ResNet-w16 (CLs)	56.6%	1.4M	295M
	ShiftNet-B	61.2%	1.1M	371M
	ResNet-w32 + SAL	**62.7**%	**0.97**M	136M
Mobile Architecture	MobileNetV2	56.6%	1.76M	57M
	ShuffleNetV2	60.7%	1.3M	**41**M

Conclusion and Future Work

Conclusion

- We introduced novel attention-based pruning method.
- The pruning method aims at replacing convolutional layers by shift layers.
- We showed SAL outperformed other existing methods.

Future Work

- Extend SAL to all kernel shapes, and to other domains than classification.
- Work on reducing complexity of the training process.

Conclusion and Future Work

Conclusion

- We introduced novel attention-based pruning method.
- The pruning method aims at replacing convolutional layers by shift layers.
- We showed SAL outperformed other existing methods.

Future Work

- Extend SAL to all kernel shapes, and to other domains than classification.
- Work on reducing complexity of the training process.

Thank you

Thank you for watching this presentation. I will be glad to answer any questions you have via e-mail ghouthi.bouklihacene@imt-atlantique.fr.

References

Wu, Bichen, et al. "Shift: A zero flop, zero parameter alternative to spatial convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Jeon, Yunho, and Junmo Kim. "Constructing fast network through deconstruction of convolution." Advances in Neural Information Processing Systems. 2018.