Generalized Depthwise-Separable Convolutions for Adversarially Robust and Efficient Neural Networks

NEURAL INFORMATION PROCESSING SYSTEMS

Spotlight

Hassan Dbouk & Naresh Shanbhag - *University of Illinois at Urbana-Champaign*

{hdbouk2,shanbhag}@illinois.edu

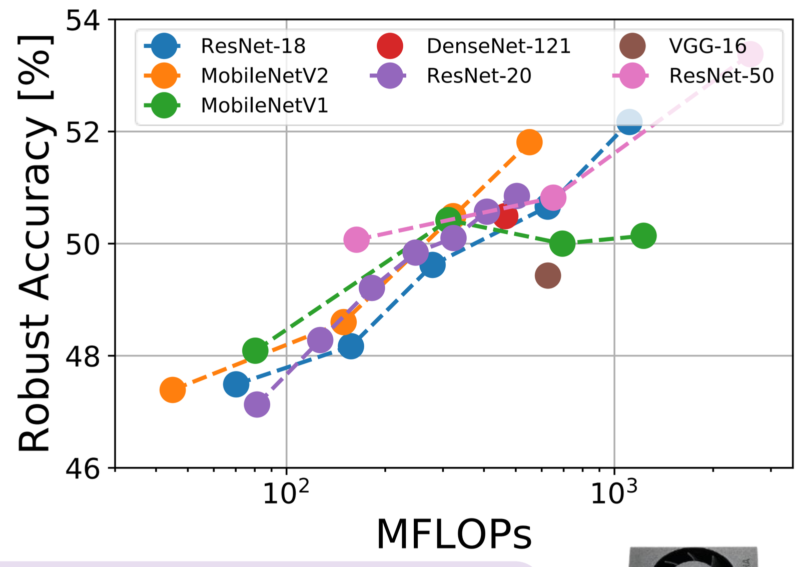
Motivation

deep nets are vulnerable

decision: 'panda'

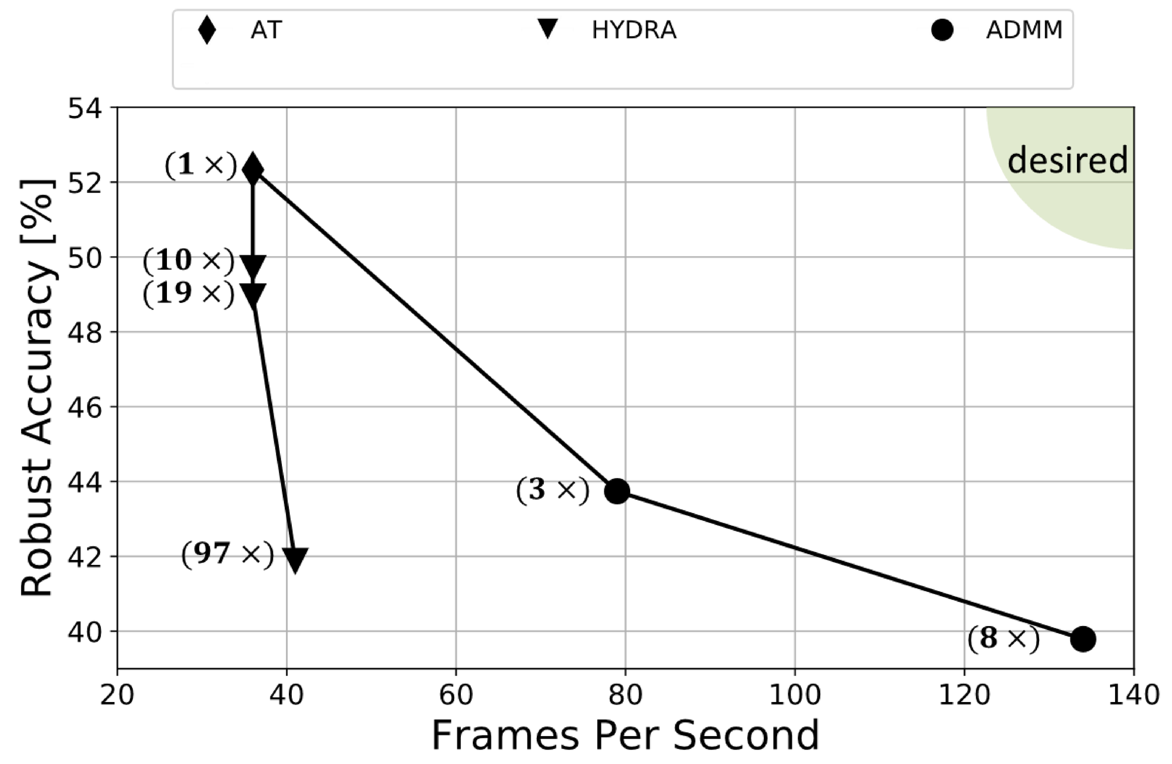
decision: 'gibbon'

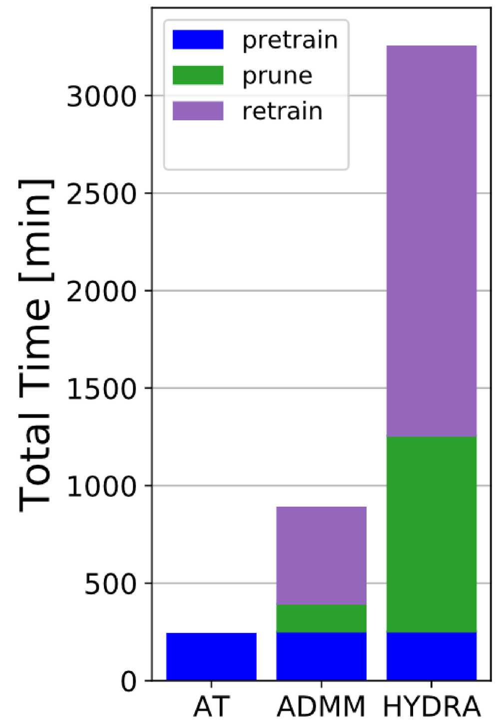
deep nets are expensive



design **robust** and **accurate** deep nets that achieve **high FPS** when mapped onto edge hardware

Limitations of Existing Techniques

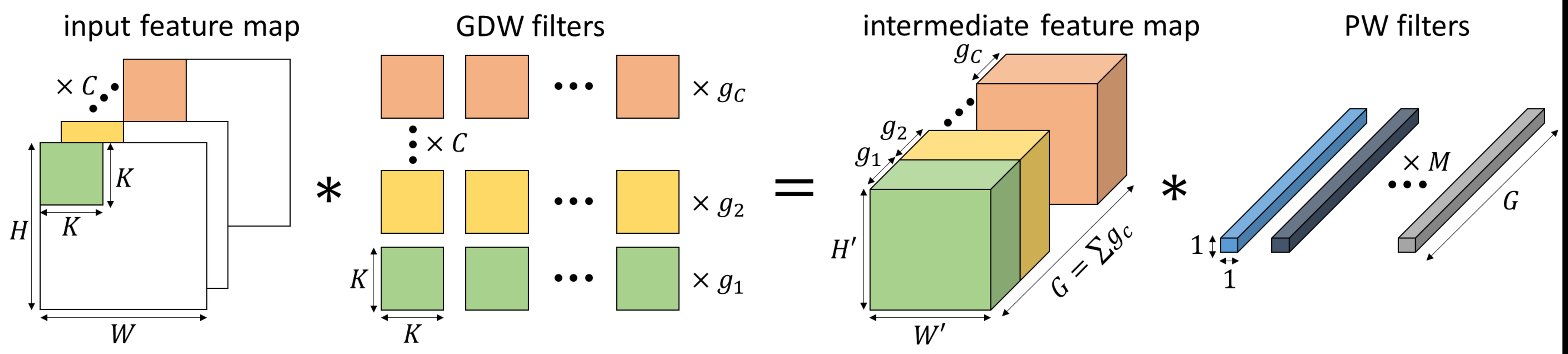




- reductions often **don't** translate to hardware
- makes AT **more expensive**
- ad hoc in nature, **no theoretical** basis behind them

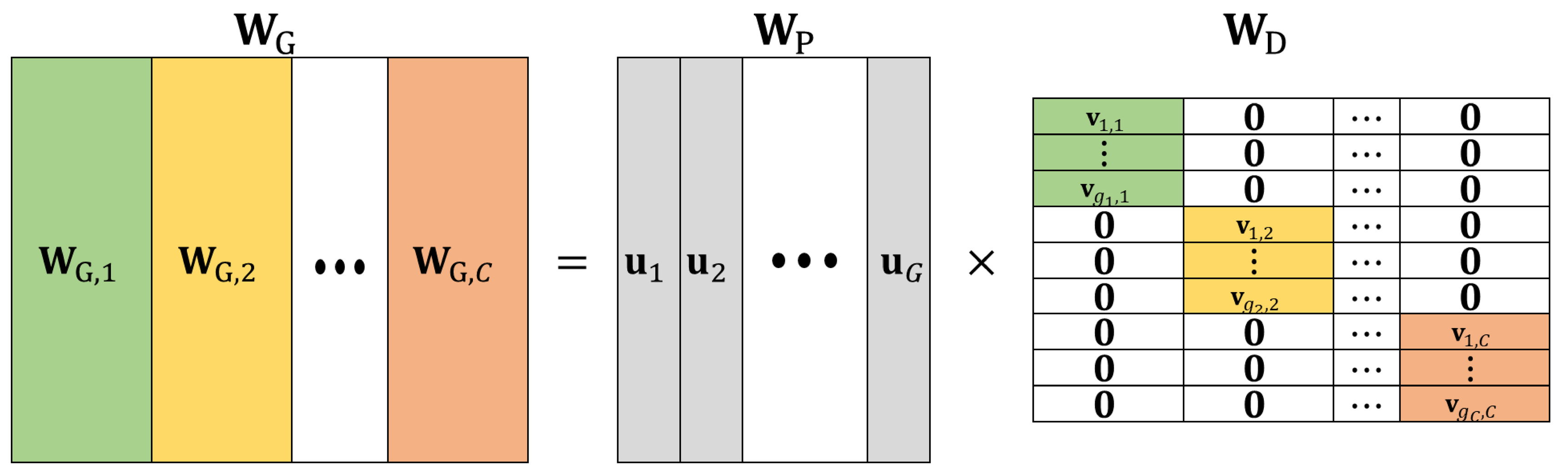
Generalized Depthwise-Separable Convolutions

Two-stage Convolution



how to choose the g_c's? → optimal approximation algorithm

Structure of Equivalent 2D Convolution



$$\mathbf{W}_{G,c} \in \mathbb{R}^{M\times K^2} \quad \& \quad \mathrm{rank}\left(\mathbf{W}_{G,c}\right) \le \min(g_c, K^2) \quad \forall c \in [C]$$

Main Result: Error-constrained Optimal Approximation

Theorem: Given a (C,K,M) standard 2D convolution with weight matrix $\mathbf{W}$, the $(C,K,\mathbf{g},M)$ GDWS approximation with weight matrix $\widehat{\mathbf{W}}$ that minimizes the complexity $\gamma(\mathbf{g})$ subject to $e\left(\mathbf{W},\widehat{\mathbf{W}},\boldsymbol{\alpha}\right) \le \beta$ (for some $\beta \ge 0$), can be constructed in polynomial time via the LEGO Algorithm.

Experimental Results – CIFAR-10

GDWS vs. HYDRA [NeurIPS'20]

Models	$\mathcal{A}_{nat}$ [%]	$\mathcal{A}_{rob}$ [%]	Size [MB]	FPS
VGG-16	82.72	51.93	58.4	36
+ GDWS ($\beta = 0.5$)	**82.53**	**50.96**	50.6	**102**
VGG-16 ($p = 90\%$)	80.54	49.44	5.9	36
+ GDWS ($\beta = 0.1$)	80.47	49.52	31.5	93
VGG-16 ($p = 95\%$)	78.91	48.74	3.0	36
+ GDWS ($\beta = 0.1$)	78.71	48.53	18.3	106
VGG-16 ($p = 99\%$)	73.16	41.74	0.6	41
+ GDWS ($\beta = 0.02$)	**72.75**	**41.56**	**2.9**	**136**

GDWS vs. ADMM [ICCV'19]

Models	$\mathcal{A}_{nat}$ [%]	$\mathcal{A}_{rob}$ [%]	Size [MB]	FPS
VGG-16	77.45	45.78	56.2	36
+ GDWS ($\beta = 0.5$)	**76.40**	**46.28**	38.8	**119**
VGG-16 ($p = 25\%$)	77.88	43.80	31.6	26
VGG-16 ($p = 50\%$)	75.33	42.93	14.0	113
VGG-16 ($p = 75\%$)	70.39	41.07	3.5	174
ResNet-18	80.65	47.05	42.6	28
+ GDWS ($\beta = 0.75$)	**79.13**	**46.15**	30.4	**105**
ResNet-18 ($p = 25\%$)	81.61	42.67	32.1	31
ResNet-18 ($p = 50\%$)	79.42	42.23	21.7	60
ResNet-18 ($p = 75\%$)	74.62	43.23	11.2	74

- HYDRA: compromises **robustness**, minimal improvements in **FPS**
- ADMM: compromises **robustness**, high **FPS**
- GDWS: preserves **robustness** and boosts **FPS** significantly

GDWS vs. Lightweight Networks

Models	$\mathcal{A}_{nat}$ [%]	$\mathcal{A}_{rob}$ [%]	Size [MB]	FPS
ResNet-18 + GDWS	81.17	50.98	29.1	104
VGG-16 + GDWS	77.17	49.56	28.7	129
MobileNetV1	79.92	49.08	12.3	125
MobileNetV2	79.59	48.55	8.5	70
ResNet-18 (DWS)	80.12	48.52	5.5	120
ResNet-20	74.82	47.00	6.4	125

GDWS vs. RobNet [CVPR'20]

Models	$\mathcal{A}_{nat}$ [%]	$\mathcal{A}_{rob}$ [%]	Size [MB]	FPS
RobNet	82.72	52.23	20.8	5
ResNet-50	84.21	53.05	89.7	16
+ GDWS	83.72	52.94	81.9	37
WRN-28-4	84.00	51.80	22.3	17
+ GDWS	**83.27**	**51.70**	18.9	**65**

- GDWS + standard networks outperform RobNet & MobileNets

Summary

- GDWS: *universal* and *efficient* approximations of 2D convolutions
- dramatically improves **FPS** while preserving **robust accuracy**
- operates on pre-trained models → *no additional training*

Acknowledgement: This work was supported by the Center for Brain-Inspired Computing (C-BRIC) and Artificial Intelligence Hardware (AIHW) program funded by the Semiconductor Research Corporation (SRC) and the Defense Advanced Research Projects Agency (DARPA).