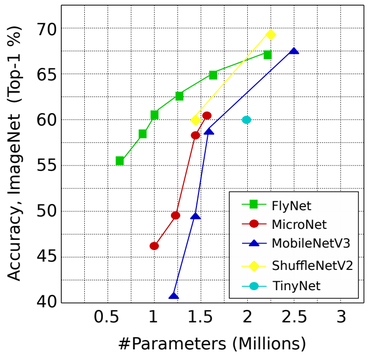
FlyNet: Max it, Excite it, Quantize it

Luis Guerra[1], Tom Drummond[2]

[1]Monash University, [2]University of Melbourne

Networks for TinyML



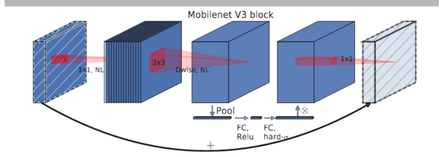
Comparison of FlyNet-h3 (3 MDW heads) architecture with different width multipliers against state-of-the-art models. FlyNet performs favorably in terms of parameter count against both popular and recently published works. Particularly, our method excels as the model shrinks, towards the extremely low parameter regime.

Why trading parameters with FLOPS?

Operation	Energy (pJ)
8 bit int ADD	0.03
16 bit int ADD	0.05
32 bit int ADD	0.1
16 bit float ADD	0.4
32 bit float ADD	0.9
8 bit MULT	0.2
32 bit MULT	3.1
16 bit float MULT	1.1
32 bit float MULT	3.7
32 bit SRAM READ	5.0
32 bit DRAM READ	640

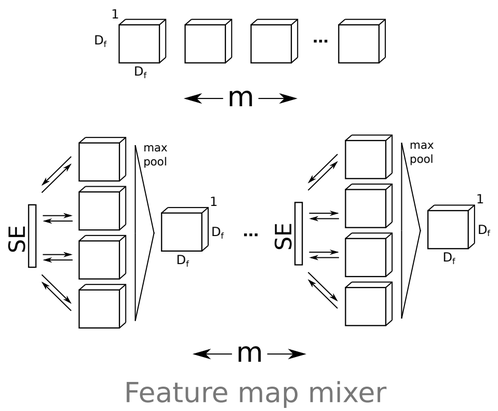
Energy Consumption of different processor operations

MobileNetV3 backbone



FlyNet

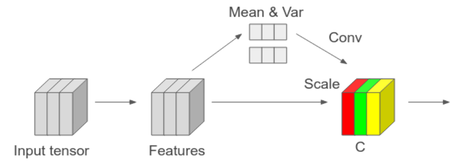
Multihead-Depthwise Convolution



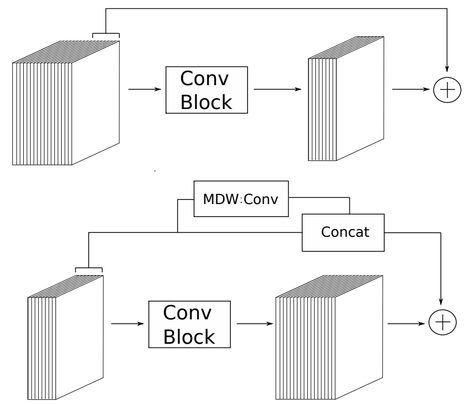
Model	Activation	Reduction	Top-1 Acc	Top-5 Acc
FlyNet-h3 0.5x	ECA+Sigmoid	Sum+Relu	57.6	80.5
FlyNet-h3 0.5x	ECA+Sigmoid	Funnel Max-out	58.8	80.9
FlyNet-h3 0.5x	ECA+Sigmoid	Max-out	58.9	81.0
FlyNet-h3 0.5x	ECA+Softmax(T=3)	Max-out	58.9	81.0
FlyNet-h3 0.5x	ECA+Softmax(T=10→ 1)	Max-out	**59.1**	**81.3**

Model	Heads	Top-1 Acc	Top-5 Acc	Params	MAdds
FlyNet 0.5x	1	56.7	79.6	0.80M	24M
FlyNet 0.5x	2	58.2	81.0	0.83M	29M
FlyNet 0.5x	3	59.1	81.4	0.86M	34M
FlyNet 0.5x	4	59.8	81.9	0.89M	39M
FlyNet 0.5x	5	**59.9**	**82.0**	0.92M	44M

Variance aware ECA-Net

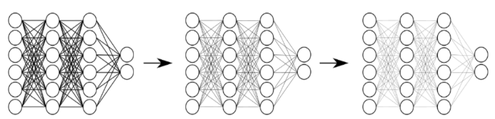


Dense Light Residuals



Backbone	Residual	Top-1 Acc	Top-5 Acc	Params	Madds
FlyNet-h3 0.4x	No Residual	54.5	78.0	0.657M	26.5M
FlyNet-h3 0.4x	Default in MobileNetV3	55.5	78.8	0.657M	26.5M
FlyNet-h3 0.4x	3x3 Conv	56.0	79.3	0.673M	30.9M
FlyNet-h3 0.4x	3x3 DW - PW	55.6	78.9	0.659M	27.5M
FlyNet-h3 0.4x	3x3 MDW - Concat	55.9	79.2	0.658M	27.0M

Quantization as regularization



Model	q-levels	Top-1 Acc	Top-5 Acc
FlyNet-h3 1.0x	2^{32}	67.2	87.5
FlyNet-h3 1.0x	16	67.3	87.7
FlyNet-h3 1.0x	8	**67.4**	**87.8**
FlyNet-h3 1.0x	6	67.1	87.4
FlyNet-h3 1.0x	4	66.1	87.1

Experiments

COCO Object Detection

Backbone	DET Framework	Params	MAdds	mAP
FlyNet-h3 0.4x	RCNN	0.14M	26M	21.9
FlyNet-h3 0.5x		0.19M	34M	23.2
MicroNet-M2		0.58M	12M	22.7
FlyNet-h3 0.6x		0.24M	46M	24.4
MobileNetV3 1.0x		0.89M	56M	25.9
MicroNet-M3		0.69M	21M	26.2
FlyNet-h3 0.8x		0.41M	86M	**27.0**
FlyNet-h3 0.4x	RetinaNet	0.14M	26M	22.9
MicroNet-M2		0.58M	12M	22.6
FlyNet-h3 0.5x		0.19M	34M	23.7
MobileNetV3 1.0x		0.89M	56M	24.0
FlyNet-h3 0.6x		0.24M	46M	24.6
MicroNet-M3		0.69M	21M	25.4
FlyNet-h3 0.8x		0.41M	86M	**27.2**

ImageNet Classification

Model	Top-1 Acc	Top-5 Acc	Params	MAdds
MobileNetV3 0.15x[37]	33.7	57.2	1.0M	4M
MicroNet-M0 [37]	46.6	70.6	1.0M	4M
FlyNet-h3 0.4x	**55.9**	**79.2**	0.65M	26M
MobileNetV3 0.2x[37]	41.1	65.2	1.2M	6M
MicroNet-M1#	49.4	72.9	1.2M	5M
MicroNet-M1	51.4	74.5	1.8M	6M
EfficientNet-B [56]	56.7	79.8	1.3M	24M
FlyNet-h3 0.5x	**59.1**	**81.3**	0.86M	34M
MobileNetV3 0.35x+BFT [59]	55.2	-	1.4M	15M
MobileNetV3 0.5x [24]	58.0	-	1.6M	21M
MicroNet-M2#	58.2	80.1	1.4M	11M
MicroNet-M2	59.4	80.9	2.4M	12M
TinyNet-E [19]	59.9	81.8	2.0M	24M
ShuffleNetV2 0.5x [42]	60.3	-	1.4M	41M
FlyNet-h3 0.6x	**61.5**	**83.4**	1M	46M
ShuffleNetV2 0.5x+BFT [59]	61.3	-	1.4M	41M
MicroNet-M3#	61.3	82.9	1.6M	20M
FlyNet-h3 0.7x	**63.3**	**84.5**	1.3M	54M
MicroNet-M3	62.5	83.1	2.6M	21M
Mobile-Former-26M [5]	64.0	-	3.2M	26M
EtinyNet [66]	65.5	86.2	0.98M	117M
FlyNet-h3 0.8x	**65.7**	**86.2**	1.6M	66M

We developed a series of simple but effective architectural modifications that can be integrated into any neural architecture to provide accuracy boosts at very little overhead. Particularly, our contributions are aimed at compressed networks in the extremely low parameter regime (sub 1M).