A Model Listing

Model Family	Model	IO Size (kB)		Weights		GPU Execution Latency (ms)				
		Input	Output	Size (MB)	Transfer (ms)	B1	B2	B4	B8	B16
DenseNet [29]	densenet121	602	4	31.8	2.59	3.80	4.52	6.55	10.22	17.91
	densenet161	602	4	114.7	9.33	7.66	10.11	15.13	23.94	40.04
	densenet169	602	4	56.5	4.50	5.18	6.29	8.57	12.82	21.85
	densenet201	602	4	80.0	6.52	6.84	8.45	11.95	18.30	31.03
DLA [63]	dla34	602	4	64.9	5.29	3.06	4.77	7.11	10.66	15.98
GoogLeNet [56]	googlenet	602	4	26.5	2.16	1.54	1.94	2.69	4.19	7.11
Inception v3 [57]	inceptionv3	1073	4	95.3	7.77	4.46	6.85	10.99	16.45	26.17
	xception	602	4	159.3	12.99	4.49	6.64	10.46	18.53	34.55
Mobile Pose [61] + MobileNet [25, 26]	mobile_pose_mobilenet1.0	590	209	20.0	1.63	0.99	1.72	2.99	5.67	10.78
	mobile_pose_mobilenetv3	590	209	19.0	1.55	1.29	1.92	3.13	5.71	11.62
	mobile_pose_resnet18_v1	590	209	51.4	4.19	1.43	2.25	3.52	6.29	11.46
	mobile_pose_resnet50_v1	590	209	102.2	8.31	3.29	5.42	9.00	16.28	29.92
	simple_pose_resnet18_v1b	590	209	61.5	5.00	2.46	3.62	6.67	10.70	18.98
ResNeSt [66]	resnest14	602	4	42.4	3.45	2.70	4.07	6.72	12.61	22.91
	resnest26	602	4	68.2	5.56	4.30	6.07	9.85	18.26	32.52
	resnest50	602	4	109.8	8.93	6.96	9.47	14.27	29.94	56.02
	resnest101	602	4	192.9	15.71	12.31	16.23	25.79	44.65	78.17
ResNet [22]	resnet18_v1	602	4	46.7	3.81	1.27	1.86	2.73	4.06	7.02
	resnet18_v1b	602	4	46.7	3.81	1.25	1.71	2.37	3.93	6.83
	resnet34_v1	602	4	87.2	7.11	2.40	3.39	4.62	7.76	14.40
	resnet34_v1b	602	4	87.2	7.11	2.37	3.37	4.59	7.76	13.32
	resnet50_v1	602	4	102.3	8.33	2.61	3.78	5.61	9.13	15.67
	resnet50_v1b	602	4	102.1	8.33	2.77	3.95	5.88	9.78	16.58
	resnet50_v1c	602	4	102.2	8.31	2.82	4.07	6.11	10.17	17.26
	resnet50_v1d	602	4	102.2	8.31	2.78	4.02	6.01	10.06	17.13
	resnet50_v1s	602	4	102.6	8.35	3.04	4.47	6.99	11.66	20.39
	resnet50_tuned_1.8x	602	4	88.1	7.16	2.24	3.05	4.25	6.65	11.13
	resnet101_v1	602	4	178.3	14.54	5.27	7.62	11.07	18.04	30.30
	resnet101_v1b	602	4	178.0	14.46	5.41	7.80	11.33	18.64	31.18
	resnet101_v1c	602	4	178.1	14.47	5.47	7.91	11.53	19.03	31.98
	resnet101_v1d	602	4	178.1	14.47	5.42	7.87	11.44	18.94	31.84
	resnet101_v1s	602	4	178.5	14.51	5.70	8.35	12.43	20.55	35.10
	resnet101_tuned_1.9x	602	4	136.3	11.08	3.85	5.61	7.47	12.56	20.61
	resnet101_tuned_2.2x	602	4	131.0	10.65	3.72	5.23	7.01	11.28	18.55
	resnet152_v1	602	4	240.9	19.58	7.71	11.14	16.21	26.48	44.60
	resnet152_v1b	602	4	240.5	19.54	7.86	11.36	16.41	27.05	45.49
	resnet152_v1c	602	4	240.5	19.55	7.90	11.48	16.64	27.42	46.24
	resnet152_v1d	602	4	240.5	19.55	7.89	11.45	16.59	27.38	46.01
	resnet152_v1s	602	4	241.0	19.58	8.15	11.91	17.50	28.95	49.27

Table 1: Models used for Clockwork experiments. Pre-trained models were sourced from the ONNX Model Zoo [44] and the GluonCV Model Zoo [20], and optimized for NVIDIA Tesla v100 GPUs using TVM v0.7 [10] *Continues on next page.*

Model Family	Model	IO Size (kB)		Weights		GPU Execution Latency (ms)				
		Input	Output	Size (MB)	Transfer (ms)	B1	B2	B4	B8	B16
ResNet v2 [23]	resnet18_v2	602	4	46.7	3.81	1.32	1.81	2.48	4.42	7.12
	resnet34_v2	602	4	87.2	7.11	2.55	3.44	4.83	7.90	14.01
	resnet50_v2	602	4	102.2	8.32	2.73	4.05	5.87	9.93	17.3
	resnet101_v2	602	4	178.1	14.47	5.51	8.05	11.83	18.14	33.57
	resnet152_v2	602	4	240.6	19.56	8.21	11.66	17.03	27.60	48.54
ResNeXt [62]	resnext50_32x4d	602	4	100.0	8.15	2.18	3.23	5.35	9.21	17.42
	resnext101_32x4d	602	4	176.4	14.34	4.65	6.27	10.06	17.75	32.83
	resnext101_64x4d	602	4	333.4	27.18	6.46	10.24	17.13	30.42	60.23
SENet [28]	se_resnext50_32x4d	602	4	110.1	8.95	3.20	4.47	6.87	11.50	20.64
	se_resnext101_32x4d	602	4	195.5	15.89	6.23	8.24	12.53	21.02	37.89
	se_resnext101_64x4d	602	4	352.5	28.75	8.18	12.97	19.93	34.99	66.44
TSN [59]	tsn_inceptionv1_kinetics400	1073	1.6	24.0	1.96	1.95	2.76	4.44	7.51	13.43
	tsn_inceptionv3_kinetics400	1073	1.6	90.4	7.37	4.47	6.87	10.97	16.43	26.12
	tsn_resnet18_v1b_kinetics400	602	1.6	45.5	3.71	1.25	1.72	2.38	3.93	6.83
	tsn_resnet34_v1b_kinetics400	602	1.6	85.9	7.01	2.38	3.38	4.59	7.74	13.37
	tsn_resnet50_v1b_kinetics400	602	1.6	97.2	7.93	2.77	3.94	5.85	9.77	16.52
	tsn_resnet101_v1b_kinetics400	602	1.6	173.1	14.11	5.42	7.80	11.30	18.63	31.15
	tsn_resnet152_v1b_kinetics400	602	1.6	235.6	19.21	7.87	11.35	16.42	27.07	45.44
Wide ResNet [64]	cifar_wideresnet16_10	12	0.04	68.5	5.59	1.27	1.72	2.61	4.07	7.62
	cifar_wideresnet28_10	12	0.04	145.9	11.93	2.21	3.57	5.42	8.41	16.05
	cifar_wideresnet40_8	12	0.04	143.0	11.69	2.49	3.90	5.99	9.86	17.14
Winograd [37] + ResNet v2 [23]	winograd_resnet18_v2	602	4	77.4	6.31	0.95	1.17	1.71	2.81	5.09
	winograd_resnet50_v2	602	4	128.7	10.49	3.39	4.24	6.07	10.28	18.84
	winograd_resnet101_v2	602	4	235.8	19.23	6.36	7.71	10.71	17.26	33.52
	winograd_resnet152_v2	602	4	324.1	26.42	9.40	11.13	15.92	24.42	28.92

Table 1: *Continues from previous page.* Models used for Clockwork experiments. Pre-trained models were sourced from the ONNX Model Zoo [44] and the GluonCV Model Zoo [20], and optimized for NVIDIA Tesla v100 GPUs using TVM v0.7 [10].