	Models	Notes	Dataset(s) for benchmarking			Parameters (M)	Performance (Speed / FLOPS)	OpenVino Optmizied Speed	Pros	Cons	Efficiency Remarks	Ideal use cases for the models	Reference resource/ notebooks	Link to research paper	Code Complexity	Code Template
Image Classification																
				TOP 1 ACCURACY	TOP 5 ACCURACY											
	Vision Transformer		ImageNet	88.55%		632			(State of the art) High Accuracy with Less Computation Time for Training compared to Noisy Studen	It does not use convolutional neural networks, instead vision transformer uses small patches to apply self attention; the efficacy of using small patches is debatable at the time of writing as it might miss edges		Accuracy is the most critical metric		https://arxiv.org/pdf/2010.11929v1.pdf	Hard	
	NoisyStudent (EfficientNet-L2)		ImageNet	88.40%	98.70%	480			Augments labeled dataset with unlabeled examples	Adding "noise" requires careful consideration and configuration ; computationally very expensive		Semi supervised learning where getting labelled data is hard		https://arxiv.org/pdf/1911.04252v4.pdf	Hard	
	EfficientNet-B7		ImageNet	84.4		66			Uses Neural Architecture Search to optmize architecture					https://arxiv.org/abs/1905.11946	Medium	
	EfficientNet-B4		ImageNet	82.6		19			Uses Neural Architecture Search to optmize architecture; excellent accuracy compared to speed			Both accuracy and performance are critical		https://arxiv.org/abs/1905.11946	Medium	
	EfficientNet-B2		ImageNet	79.8		9.2			Uses Neural Architecture Search to optmize architecture			Both accuracy and performance are critical		https://arxiv.org/abs/1905.11946	Medium	
	EfficientNet-B0		ImageNet	76.3		5.3			Uses Neural Architecture Search to optmize architecture; highly optimized for ML on edge with increased accuracy			ML on the edge where accuracy is important		https://arxiv.org/abs/1905.11946		
	Inception V3		ImageNet	78.8	95.10%	55.8			Uses filters of varying sizes to detect different image features for greater accuracy	Computationally expensive; better models exist				https://arxiv.org/pdf/1602.07261v2.pdf	Easy	
	MobileNet v2		ImageNet	71.88	90.29	3.4			Highly optimized for performance on ML on the edge applications; great backbone for CV on the edge tasks	Not as robust as larger models		ML on the edge		https://arxiv.org/pdf/2006.10702v1.pdf	Medium	
	SqueezeNet		ImageNet	58.1	80.42	1.25			Very fast and hence great for performance critical applications; works out of the box	Poor performance when used for classification and detection of small objects	Highly efficient	ML on the edge; low latency		https://arxiv.org/abs/1602.07360	Medium	
	ResNet 18		ImageNet	69.76	90.08	11.174			Most tried and tested method. Works well.	Increased complexity of architecture due to skip connections; Implementation of Batch normalization layers since ResNet heavily depends on it				https://arxiv.org/pdf/1512.03385.pdf	Easy	
	ResNet 34		ImageNet	73.3	91.42	21.282			Most tried and tested method. Works well.	Increased complexity of architecture due to skip connections; Implementation of Batch normalization layers since ResNet heavily depends on it				https://arxiv.org/pdf/1512.03385.pdf	Easy	
	ResNet 50		ImageNet	76.15	92.87	25.6			Most tried and tested method. Works well.	Increased complexity of architecture due to skip connections; Implementation of Batch normalization layers since ResNet heavily depends on it				https://arxiv.org/pdf/1512.03385.pdf	Easy	
	ResNet 101		ImageNet	79.20%	94.70%	42.513			Most tried and tested method. Works well.	Increased complexity of architecture due to skip connections; Implementation of Batch normalization layers since ResNet heavily depends on it				https://arxiv.org/pdf/1512.03385.pdf	Easy	
	VGG 19		ImageNet	74.50%	92.00%	144			Easy to understand architecture	Computationally expensive and gives much lower accuracy in comparison with much smaller models	Low efficiency	Google deepdream		https://arxiv.org/pdf/1409.1556v6.pdf	Easy	
	VGG 16		ImageNet	74.40%	91.90%	138			Easy to understand architecture	Computationally expensive and gives much lower accuracy in comparison with much smaller models	Low efficiency	Google deepdream		https://arxiv.org/pdf/1409.1556v6.pdf	Easy	
Metrics Legend																
Top1	The model is considered to have classified a given image correctly if the target label is the model's top prediction															
Top5	The model is considered to have classified a given image correctly if the target label is the model's top 5 predictions															
Semantic segmentation				MEAN IOU												
	FPN		Encoder (ImageNet)			23.15M	Nil	Nil	i) Same network can perform Image Segmentation, Object detection and Pose Estimation	i) Heavy network cannot be used for real time inference		High performan	https://github.com/NimbleBoxAI/	http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf	Easy	
	Linknet		Encoder (ImageNet)	class IoU(76.4)		11.5M	21.2G	Nil	Learns without any significant increase in number of parameters, fast.	Not the most accurate. compromises accuracy for less params and faster fps.		Mid performance	https://github.com/NimbleBoxAI/	https://arxiv.org/abs/1707.03718	Easy	
	PAN		Cityscapes	84.00%		21.4M	Nil	Nil	Exploits the impact of global contextual information in semantic segmentation by combining attention mechanism and spatial pyramid to extract precise dense features.	i) Even though a much lighter network than using the Inception or VGG backbones, this is still heavy for real-time inference		Use in medical a	https://github.com/NimbleBoxAI/	https://arxiv.org/abs/1805.10180	Easy	
	PSPNet		Cityscapes	80.20%		21.4M	Nil	Nil	Exploits the capability of global context information by differentregion-based context aggregation through our pyramid pooling	i) Heavy model, needs a powerful system to run		Where you need	https://github.com/NimbleBoxAI/	https://arxiv.org/abs/1612.01105	Easy	
	U-Net		Brain MRI segmentation			77.6M	Nil	Nil	This network can localize. the training data in terms of patches is much larger than the number of training images	It is quite slow because the network must be run separately for each patch , and there is a lot of redundancy due to overlapping patches.		When you have t	https://github.com/NimbleBoxAI/	https://arxiv.org/abs/1505.04597	Medium	
Metrics Legend																
Mean IOU : IOU = true_positive / (true_positive + false_positive + false_negative) PCKH-0.5: Probability of 50% pose points matching																
Human Pose Estimation				PCKH-0.5												
	MSPN		MPII Human Pose	92.60%			9.6G	Nil	i) Light model that runs fast on standard chips as well					https://arxiv.org/abs/1901.00148v4	Medium	
	HRNet-W32		MPII Human Pc	92.30%		28.5 M	16.0G	Nil	i) Maintain high res. representations throughout the the whole process ii) Fuse multi-res representations repeatedly	i) Model is heavier compared to others in the segment				https://arxiv.org/abs/1902.09212v1	Easy	
	Pyramid Residual Modules		MPII Human Pose	92.00%				Nil						https://arxiv.org/abs/1708.01101v1	Hard	
	Multi-Context Attention		MPII Human Pc	91.50%				Nil						https://arxiv.org/abs/1702.07432v1	Hard	
	CU-Net	Pretrained weights not available. Code for running ready.	MPII Human Pc	91.20%		(4 Stack) 3.9M (8) 7.9 M (16) 15.9 M		Nil	i) Introduces and uses dense u-net that uses ~70% lower parameters compareed to tradional U-Net stack ii) Since this is a U-Net the input and output size if the same and there is no need to upscaling					https://arxiv.org/abs/1808.02194v2	Medium	
	Stacked hourglass + Inception-resnet		MPII Human Pc	91.20%				Nil						https://arxiv.org/abs/1705.02407v2	Hard	
	EfficientPose IV		MPII Human Pc	91.20%		6.56 M		Nil	Very small orders of magnitude smaller than comparable models	i) Single person only ii) Relies on upscaling which can reduce the effectiveness.		Real Time Inference		https://arxiv.org/abs/2004.12186v1	Easy	
	EfficientPose RT		MPII Human Pc	88.40%		0.46 M		--"--	Order of mag. smaller than IV version	--"--		Real Time Inference		--"--	--"--	
Activity Recognition																
Metrics Legend																
3-Fold Accuracy: The data is split into three parts and 2 are used for training while the third is used for testing. This accuracy is the mean of three.																
Action Recognition				3-FOLD ACCURACY												
	R2+1D BERT	Only following weights are available: https://1drv.ms/u/s!AgKP51Rjkz1Gaifd54VbdRBn6gM?e=7OxYLa	UCF101	98.69%		66.67 M	152.97 G	Nil	i) Uses BERT for Bi-directional embedding ii) BERT output is standard and thus can be utilised with other tasks such as image/video captioning, etc.	i) Super heavy model				https://arxiv.org/abs/2008.01232v3	Medium	

	OmniSource	This discusses more about a new data source augmentation than about the model itsef. Models are simple Teach er Student models	UCF101	98.60%						-	-	Nil	i) Uses a way to convert images to videos that is applicable for other uses as well	i) The videos generated do not look like actual videos			https://arxiv.org/abs/2003.13042v2	Easy
	Two-stream I3D		UCF101	98%						25M	-	Nil	i) Uses two streams 3D conv blocks one for RGB Images and other for Flow	Slow, not suitrable for real time inference			https://arxiv.org/abs/1705.07750v3	Easy
	MARS+RGB/Flow	RGB Implementation complete. FLOW is only an extension that can be implemented further when needed.	UCF101	97.80%						-	-	Nil	i) Reduces a two step process with MERS to a single step process. ii) Demostrates ability to use Cross Entropy and MSE loss together iii) Simple RGB performs better than Flow + RGB or Flow only.				https://openaccess.thecvf.com/content_CVPR_2019/papers/Crasto_MARS_Motion-Augmented_RGB_Stream_for_Action_Recognition_CVPR_2019_paper.pdf	Medium
Accuracy: Absolute Accuracy (Top-1) of top prediction																		
Multimodal Activity recognition				ACCURACY														
	TCN		EV-Action	80.10%								Nil					https://arxiv.org/abs/1704.04516v1	Medium
	ST-GCN		EV-Action	79.60%								Nil					https://arxiv.org/abs/1801.07455v2	Easy
	TSN		EV-Action	73.60%								Nil					https://arxiv.org/abs/1608.00859v1	Hard
Text Recognition				PRECISION	RECALL	F-MEASURE												
	TextFuseNet		ICDAR 2015	93.96	90.56	92.23						Nil					https://www.ijcai.org/Proceedings/2020/72	Hard
	CharNet H-88		ICDAR 2015	92.65	90.47	91.55				89.21M	Nil	Nil	one-stage model that can process two tasks simultaneously in one pass. CharNet directly outputs bounding boxes of words and characters, with corresponding character labels	CharNet H-88 has more params than CharNet H-57 becaue of the backbone used	merges two tasks: bounding box and text prediciton	https://github.com/NimbleBoxAI/box_of_ai_tools/tree/main/Text_Recognition/charnet	https://arxiv.org/abs/1910.07954v1	Medium
	SBD		ICDAR 2015	92.1	88.2	90.1						Nil					https://arxiv.org/abs/1912.09629v2	Medium
	FOTS MS		ICDAR 2015	91.85	87.92	89.84				34.98 M	Nil	Nil					https://arxiv.org/abs/1801.01671v2	Hard
	DB-ResNet-50		ICDAR 2015	91.8	83.2	87.3						Nil					https://arxiv.org/abs/1911.08947v2	Easy
	Mask TextSpotter		ICDAR 2015	91.6	81	86				Nil		Nil	takes advantage of simple and smooth end-to-endlearning procedure, in which precise text detection and recognition areacquired via semantic segmentation	proposes region, box classification and word segementation differently	proposes region,	https://github.com/NimbleBoxAI/box_of_ai_tools/tree/main/Text_Recognition/MaskTextSpotter	https://arxiv.org/abs/1807.02242v2	Easy
	CharNet H-57		ICDAR 2015	91.43	88.74	90.06				34.96M	Nil	Nil	one-stage model that can process two tasks simultaneously in one pass. CharNet directly outputs bounding boxes of words and characters, with corresponding character labels	CharNet H-57 has comparatively less params then Charnet H-88 but compromises on the performance	merges two tasks: bounding box and text prediciton	https://github.com/NimbleBoxAI/box_of_ai_tools/tree/main/Text_Recognition/charnet	https://arxiv.org/abs/1910.07954v1	Medium
	PMTD		ICDAR 2015	91.3	87.43	89.33						Nil					https://arxiv.org/abs/1903.11800v1	Medium
	CTPN+CRNN												The CTPN detects a text line in a sequence of fine-scale text proposals directly in convolutiona feature maps.	it is not an end to end solution a CRNN needs to be attached to to predict text.	it is not an end to	https://github.com/NimbleBoxAI/box_of_ai_tools/tree/main/Text_Recognition/CTPN%2BCRNN		
Metrics Legend																		
Precision	true_positive / (true_positive + false_positive)																	
Recall	true_positive / (true_positive + false_negative)																	
F-Measure	1/(1/precision + 1/recall)																	
				BOX AP	AP50	AP75	APS	APM	APL									
Object Detection																		
	YOLO V4 P7		MS COCO	55.8	73.2	61.2	38.8	60.1	68.2	287.4							https://arxiv.org/pdf/2011.08036v1.pdf	Medium
	DetectoRS (ResNeXt-101-64x4d, multi-scale)		MS COCO	55.7	74.2	61.1	37.7	58.4	68.1								https://arxiv.org/pdf/2006.02334v2.pdf	Medium
	EfficientDet-D7x		MS COCO	55.1	74.3	59.9	37.2	57.9	68								https://arxiv.org/pdf/1911.09070v7.pdf	Medium
	YOLO V4 P6		MS COCO	54.3	72.3	59.5	36.6	58.2	65.5								https://arxiv.org/pdf/2011.08036v1.pdf	Medium
	SpineNet-190		MS COCO	54.3													https://arxiv.org/pdf/1912.05027v3.pdf	Medium
	Cascade MaskRCNN		MS COCO	53.3	71.9	58.5	35.5	55.8	66.7								https://arxiv.org/pdf/1909.03625v1.pdf	Medium
	ResNeSt-200DCN		MS COCO	53.3	72	58	35.1	56.2	66.8								https://arxiv.org/pdf/2004.08955v1.pdf	Medium
	Probabilistic Anchor Assignment with IoU Prediction for Object Detection (PAA)		MS COCO	53.5	71.6	59.1	36	56.3	66.9								https://arxiv.org/pdf/2007.08103v2.pdf	High
	Faster RCNN ResNet 101		MS COCO	43.9	65.7	48.1	25.4	46.7	56.3								https://arxiv.org/pdf/1908.04156v3.pdf	Easy
	RetinaNet (ResNext-101)		MS COCO	40.8	61.1	44.1	24.1	44.2	51.2								https://arxiv.org/pdf/1708.02002v2.pdf	Easy
	SSD 512		MS COCO	28.8	48.5	30.3											https://arxiv.org/pdf/1512.02325v5.pdf	Easy