AMC: AutoML for Model Compression and Acceleration on Mobile Devices

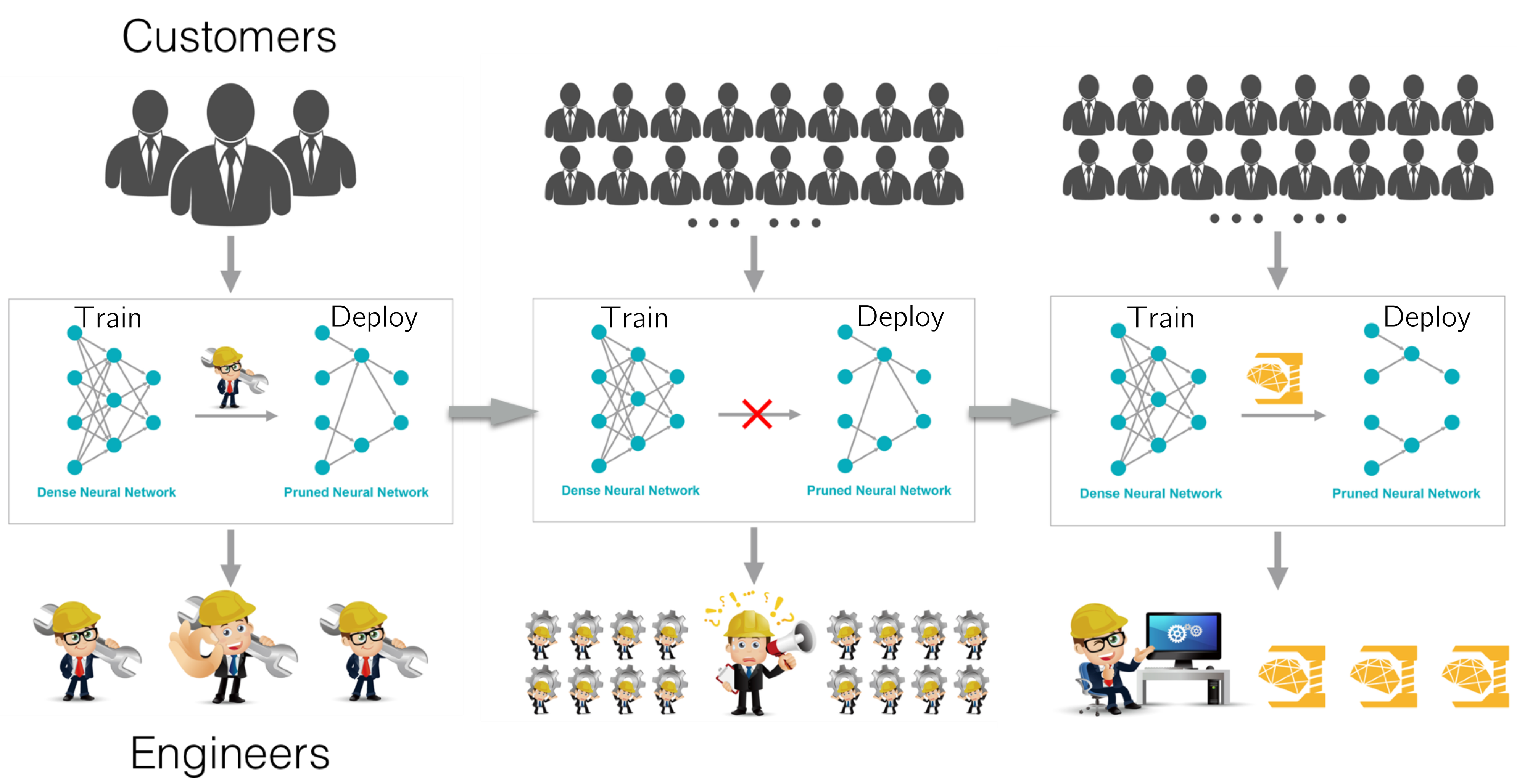
Yihui He[2*] Ji Lin[1*] Zhijian Liu[1] Hanrui Wang[1] Li-Jia Li[3] Song Han[1]
1 Massachusetts Institute of Technology 2 Xi'an Jiaotong University 3 Google (* equal contributions)

Project Page

Automated Compression via AutoML

Model compression is an important technique facilitating efficient inference, while human expert needs to find a good set of hyper-parameters (e.g., compression ratio of each layer), which requires domain expertise and many trials and errors, and is usually time-consuming and sub-optimal.

Goal: Automate the compression pipeline and free human labor. "Model compression by AI", which is automated, faster and enjoys higher performance.

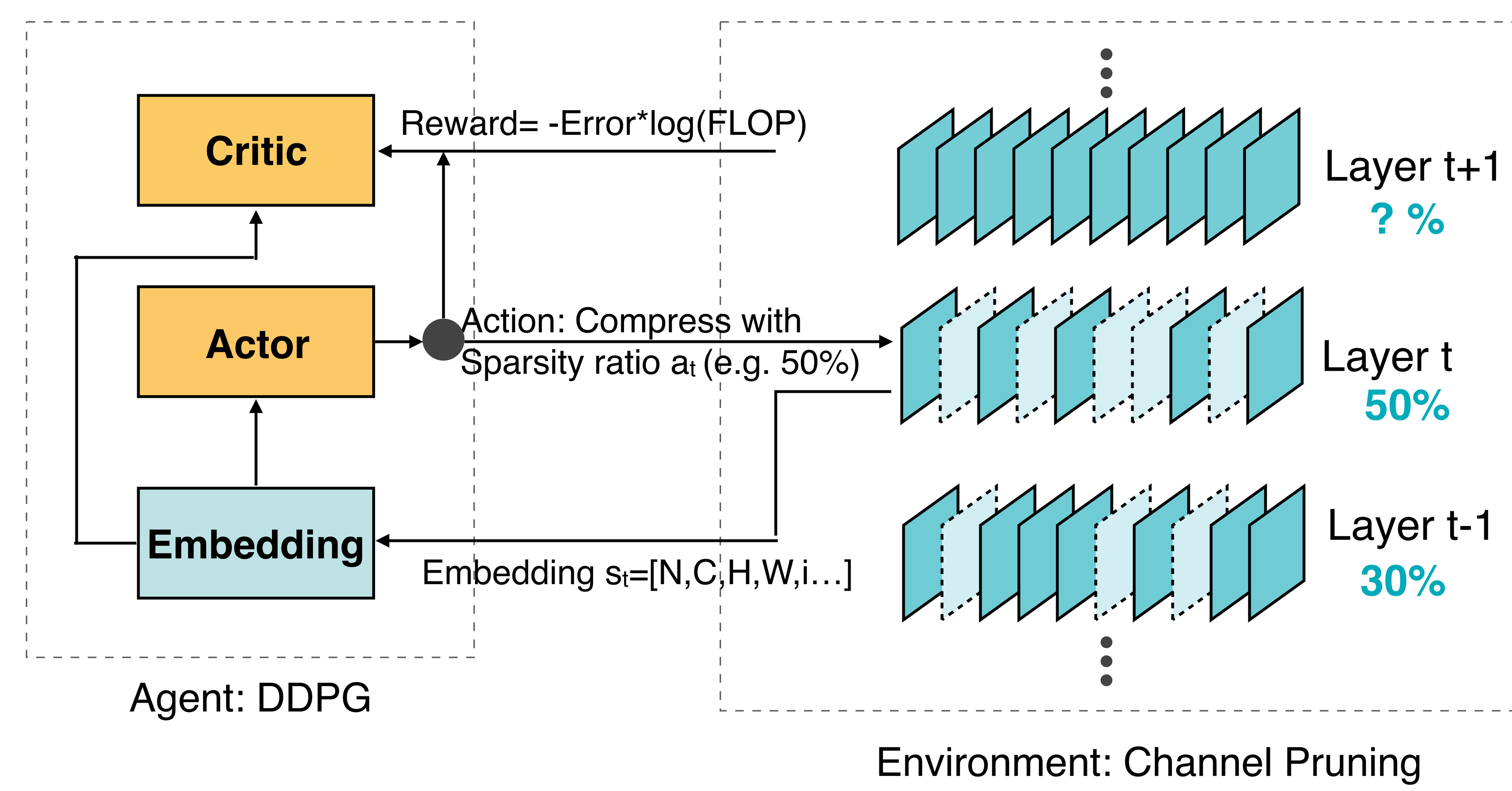


Novelty:
1. Learning based compression > Rule based compression
2. Resource-constrained search
3. Continuous action space for fine-grained surgery
4. Fast exploration with few GPUs (1GPU 4hours on ImageNet)

AMC Results on CIFAR-10

Model	Policy	Ratio	Val Acc.	Test Acc.	Acc. after FT.
Plain-20 (90.5%)	deep (handcraft)	50% FLOPs	79.6	79.2	88.3
	shallow (handcraft)		83.2	82.9	89.2
	uniform (handcraft)		84.0	83.9	89.7
	AMC (R_{Err})		**86.4**	**86.0**	**90.2**
ResNet-56 (92.8%)	uniform (handcraft)	50% FLOPs	87.5	87.4	89.8
	deep (handcraft)		88.4	88.4	91.5
	AMC (R_{Err})		**90.2**	**90.1**	**91.9**
ResNet-50 (93.53%)	**AMC (R_{Param})**	60% Params	**93.64**	**93.55**	-

Overview of AutoML for Model Compression (AMC) Engine



DDPG Agent

- DDPG Agent for continuous action space (0-1)
- Input state embedding of each layer and output sparse ratio

Compression Methods Studied

- Fine-grained Pruning for model size compression
- Coarse-grained/Channel Pruning for faster inference

Search Protocols

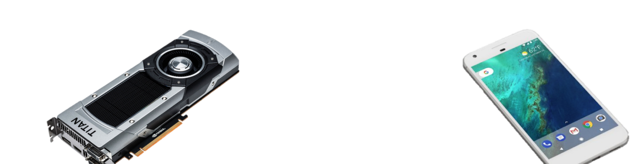
- Resource-Constrained Compression to reach a desired compression ratio while getting highest possible performance.
- Accuracy-Guaranteed Compression to fully preserve the original accuracy while maintain smallest possible model size.

Reward Functions

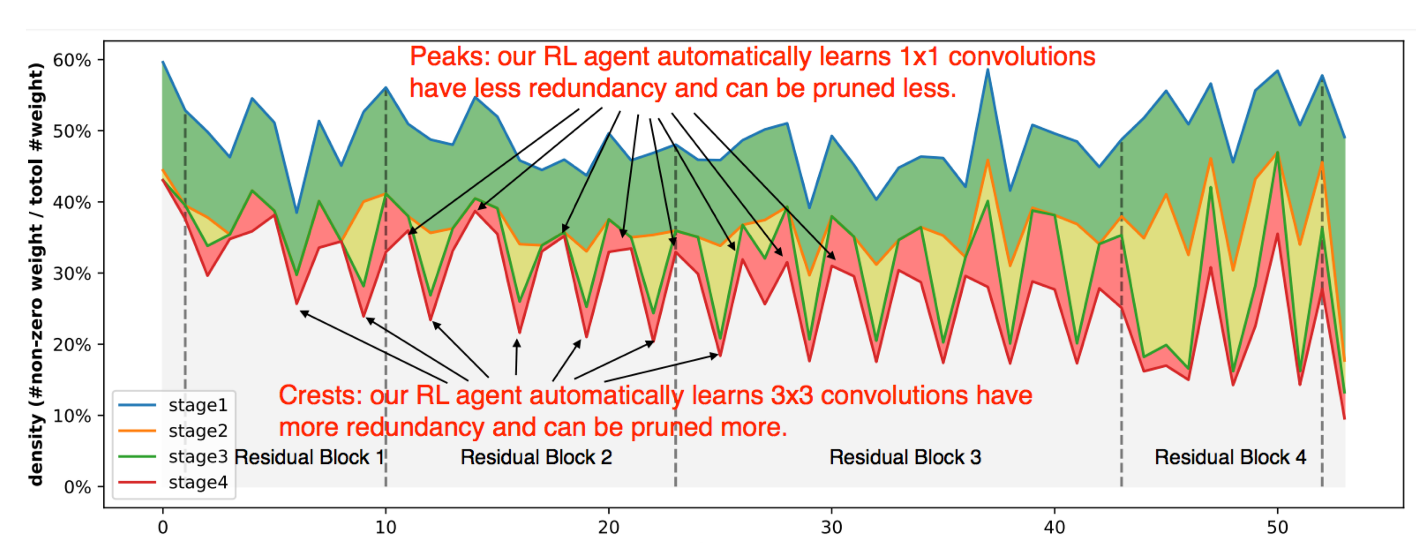
- For Resource-Constrained Compression, simply use $R_{\text{err}} = -Error$
- For Accuracy-Guaranteed Compression, considering both accuracy and resource (like FLOPs): $R_{\text{FLOPs}} = -Error \cdot log(\text{FLOPs})$

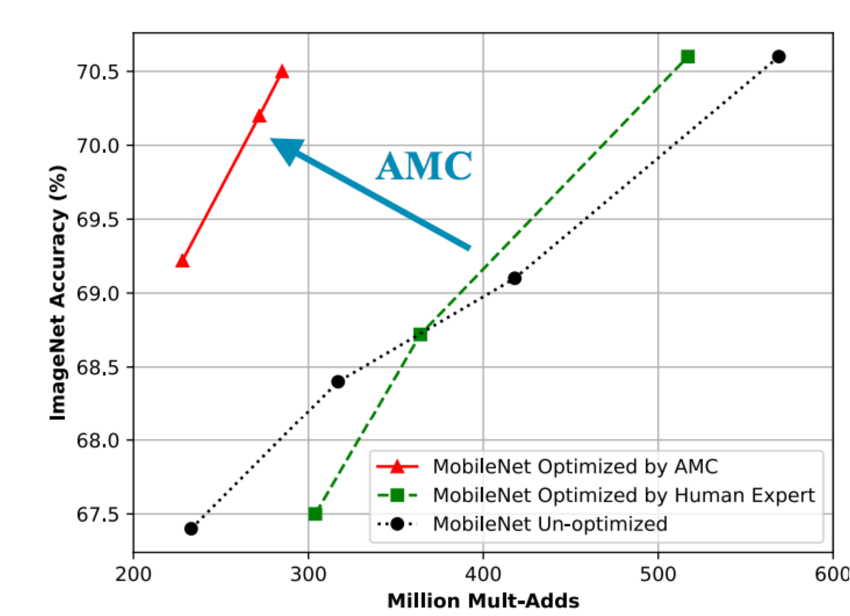
AMC Results on ImageNet

	policy	FLOPs	ΔAcc %
VGG-16	FP (handcraft) [31]	20%	-14.6
	RNP (handcraft) [33]		-3.58
	SPP (handcraft) [49]		-2.3
	CP (handcraft) [22]		-1.7
	AMC (ours)		**-1.4**
MobileNet	uniform (0.75-224) [23]	56%	-2.5
	AMC (ours)	50%	**-0.4**
	uniform (0.75-192) [23]	41%	-3.7
	AMC (ours)	40%	**-1.7**
MobileNet-V2	uniform (0.75-224) [44]	50%	-2.0
	AMC (ours)		**-1.0**

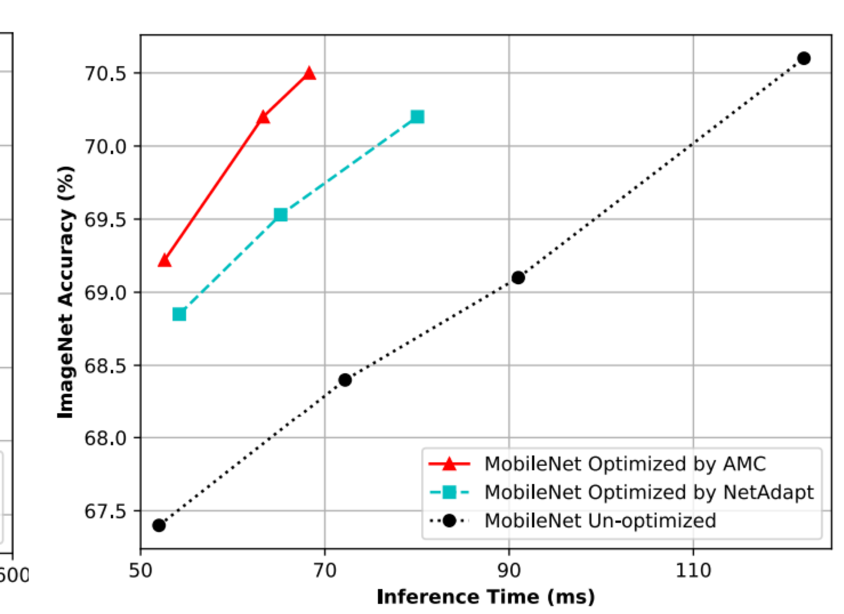


	Million MAC	top-1 acc.	top-5 acc.	GPU latency	GPU speed	Android latency	Android speed	Android memory
100% MobileNet	569	70.6%	89.5%	0.46ms	2191 fps	123.3ms	8.1 fps	20.1MB
75% MobileNet	325	68.4%	88.2%	0.34ms	2944 fps	72.3ms	13.8 fps	14.8MB
NetAdapt [52]	-	69.8%	-	-	-	70.0ms	14.3 fps	-
AMC (50% FLOPs)	285	70.5%	89.3%	0.32ms	**3127 fps (1.43×)**	68.3ms	**14.6 fps (1.81×)**	14.3MB
AMC (50% Latency)	272	70.2%	89.2%	0.30ms	**3350 fps (1.53×)**	63.3ms	**16.0 fps (1.95×)**	13.2MB





(a) Accuracy v.s. MACs



(b) Accuracy v.s. Inference time