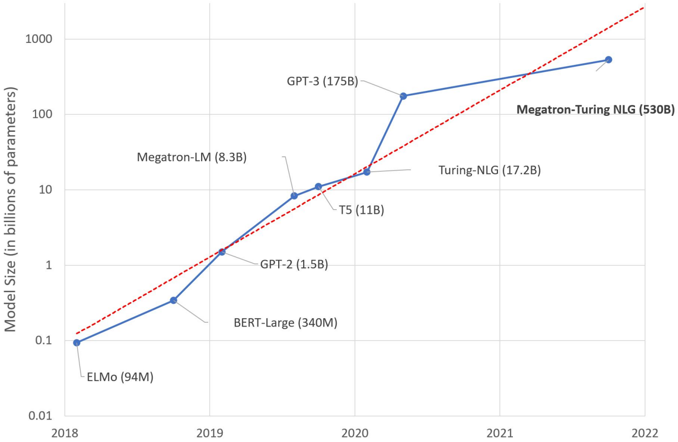
Compressing Self-Supervised Models

Jack Lin[1], Hao Tang[2], Hung-yi Lee[1]

[1]National Taiwan University, [2]The University of Edinburgh

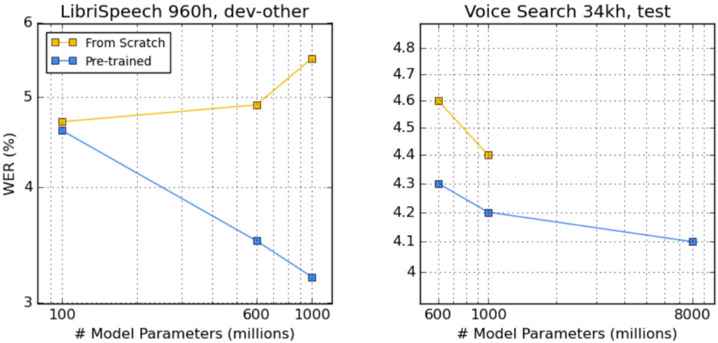
Self-supervised models are getting bigger.



(Simmon, 2021)

Similar things are happening in speech.

CPC	2M
APC	4M
wav2vec	33M
12-layer Transformer	
wav2vec 2.0 Base	95M
HuBERT Base	95M
WavLM Base	95M
24-layer Transformer	
wav2vec 2.0 Large	317M
HuBERT Large	317M
WavLM Large	317M
36-layer Transformer	
ConformerG	8000M



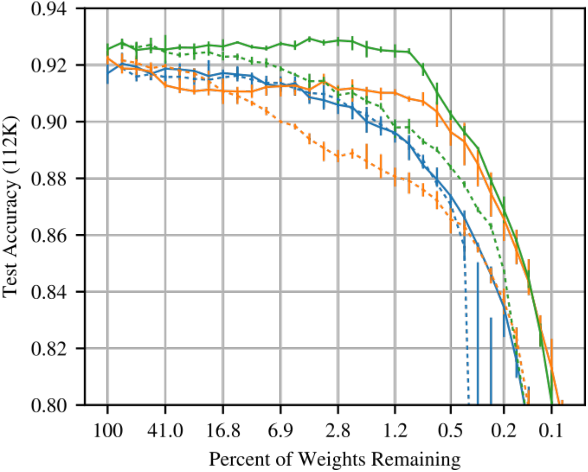
(Zhang et al., 2021)

Large neural networks can be compressed.

Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
LeNet-300-100 Ref	1.64%	-	267K	
LeNet-300-100 Pruned	1.59%	-	**22K**	**12**×
LeNet-5 Ref	0.80%	-	431K	
LeNet-5 Pruned	0.77%	-	**36K**	**12**×
AlexNet Ref	42.78%	19.73%	61M	
AlexNet Pruned	42.77%	19.67%	**6.7M**	**9**×
VGG-16 Ref	31.50%	11.32%	138M	
VGG-16 Pruned	31.34%	10.88%	**10.3M**	**13**×

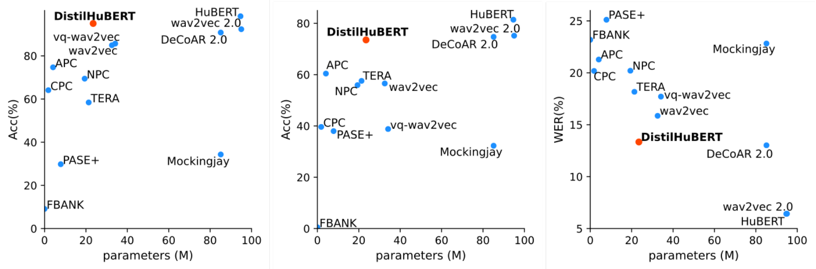
(Han et al., 2015)

Large neural networks can be compressed.



(Frankle and Carbin, 2019)

Large self-supervised models can be compressed.



(a) IC (b) SID (c) ASR (w/o LM)

(Chang et al., 2021)

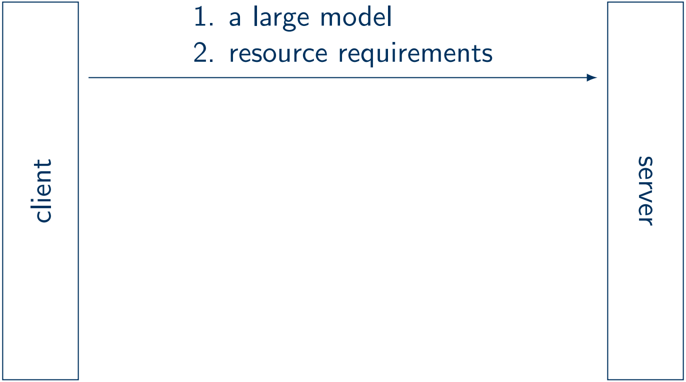
Common Compression Techniques

- Pruning
- Knowledge distillation
- Low-rank approximation
- Quantization (Low-precision floating points)

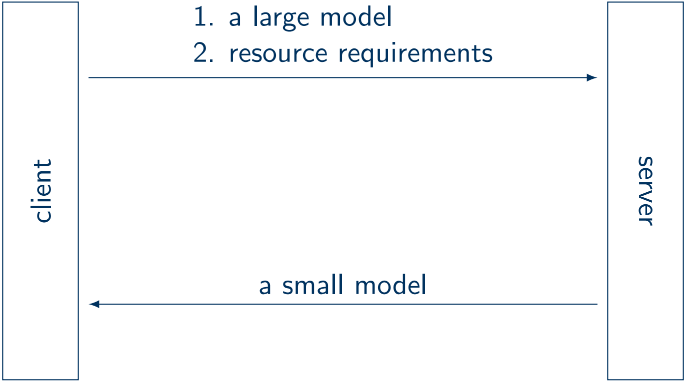
What exactly is model compression?

What exactly is model compression?

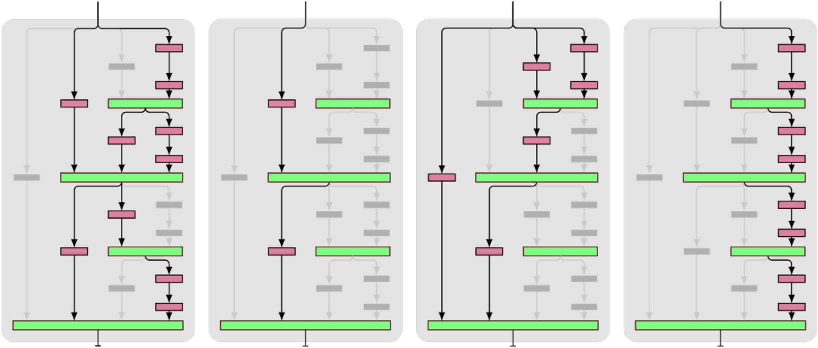
What exactly is model compression?



What exactly is model compression?



Anytime Inference



(Larsson et al., 2017)

Goal

- Understanding the landscape and limit of compressing self-supervised models
- Build anytime transformer

Plan

Mar–Jun

- Train and compress 12-layer transformers on the Librispeech 360-hour subset
- Analyze the impact of compression
- Build an anytime transformer on the Librispeech 360-hour subset

Jun–Aug

- Scale up to Librispeech 960-hour subset
- Scale up analysis to many tasks

Aug–Oct

- Complete analysis
- Summarize findings