Neural Parameter Allocation Search

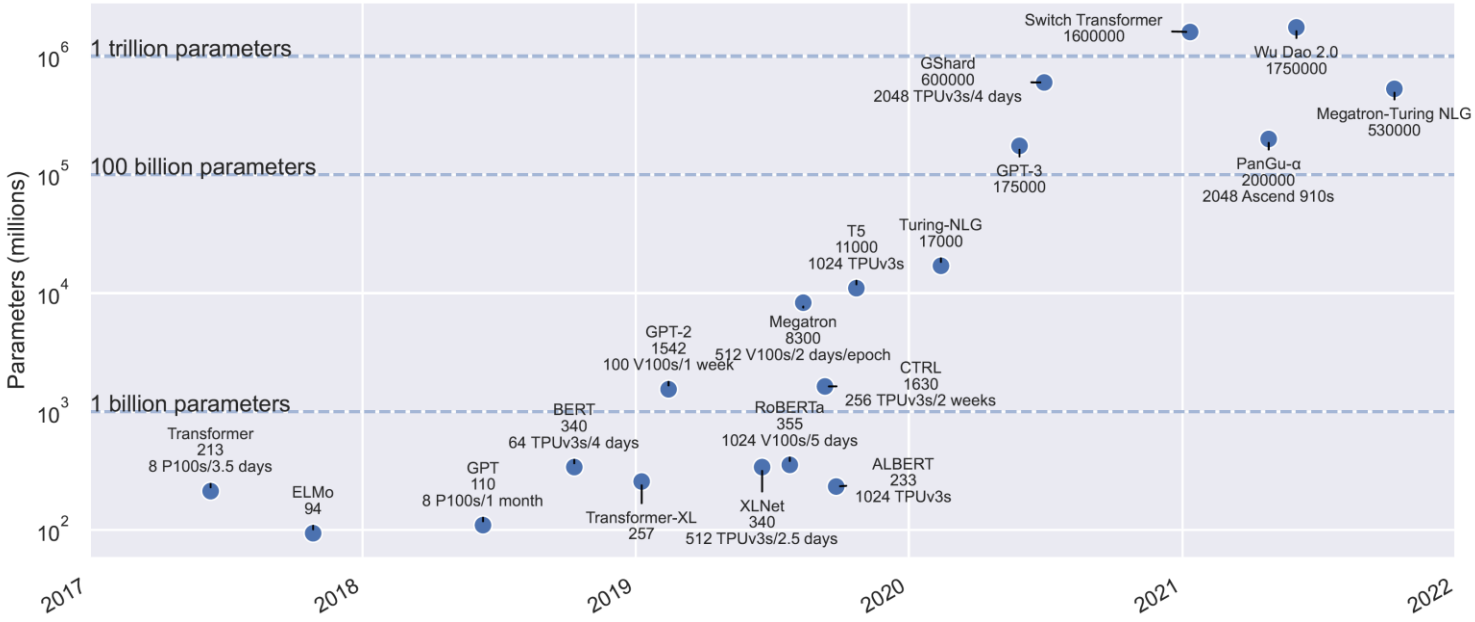
Bryan A. Plummer* Nikoli Dryden* Julius Frost Torsten Hoefler Kate Saenko

BOSTON UNIVERSITY

ETH zürich

MIT-IBM Watson AI Lab

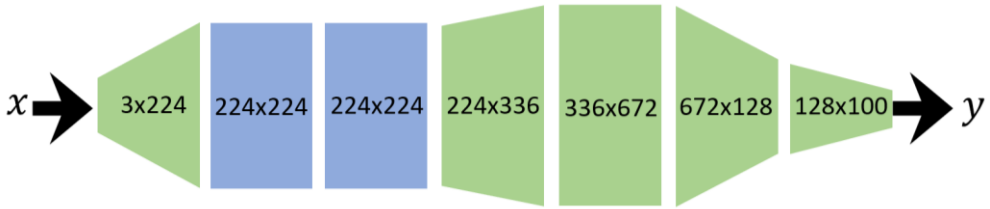
Significant growth in model size... and still going!



Cross-layer parameter sharing can help!

Prior work

manual sharing between identical layers (in blue)

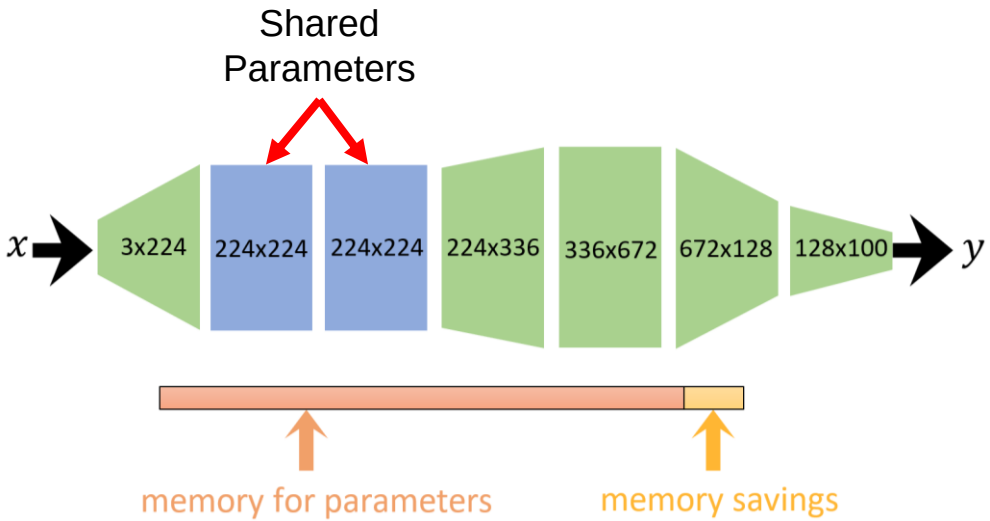


*Jaegle et al, **Perceiver**: General Perception with Iterative Attention*

	Valid	Train	Params	FLOPs
No weight sharing	72.9	87.7	326.2M	707.2B
W/ weight sharing	**78.0**	79.5	44.9M	707.2B

Limitations of cross-layer parameter sharing

- Relies on manually created parameter sharing strategies



Limitations of cross-layer parameter sharing

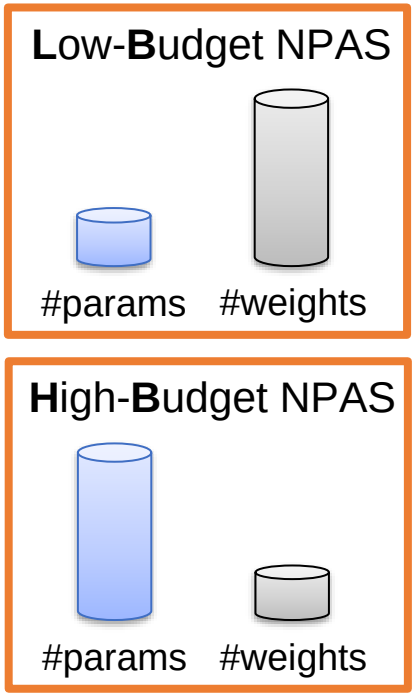
- Relies on manually created parameter sharing strategies
- Assumptions constrain supported architectures (requires many identical layers to be effective)
- Restrictions on parameter savings (proportional to # of identical layers)

Neural Parameter Allocation Search

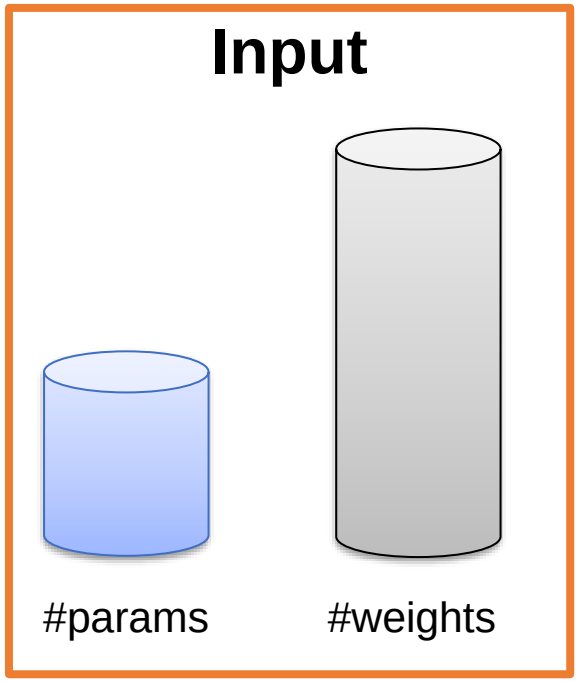
Goal: generate layer weights for any architecture under a fixed parameter budget

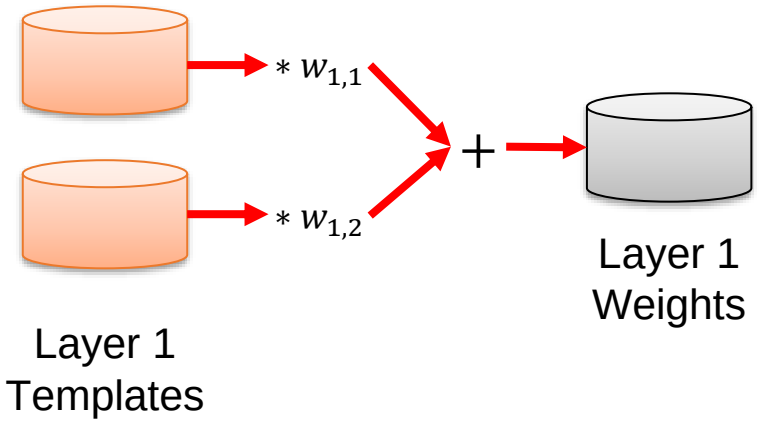
Desirable properties:

1. No architecture assumptions
2. Automated
3. Minimal overhead
4. High performance
5. Supports any parameter budget

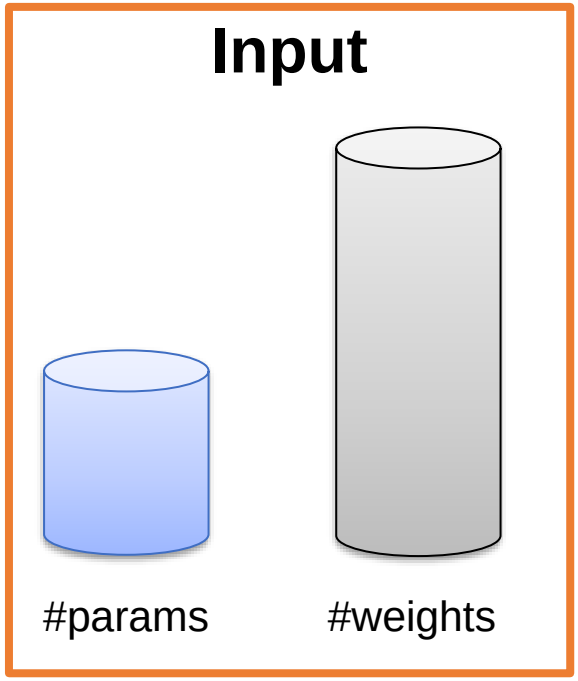


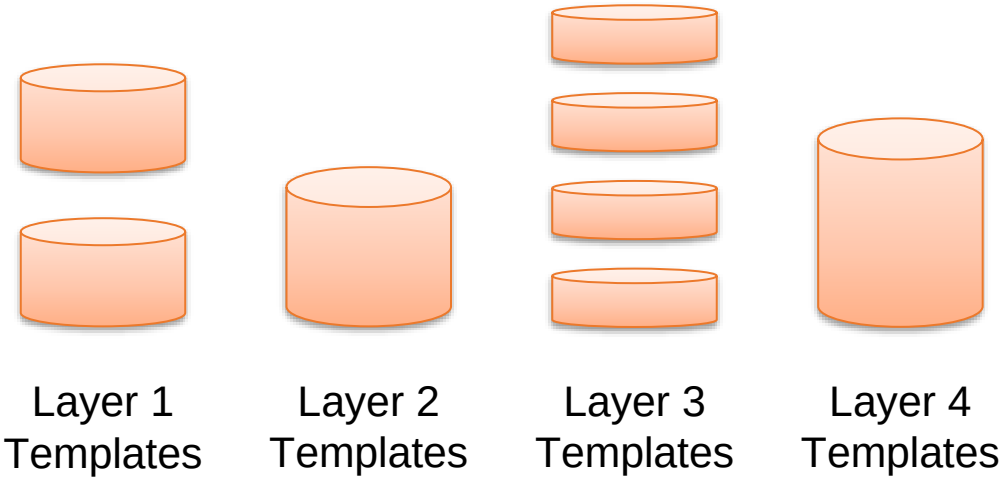
Shapeshifter Networks: Automated Template Mixing



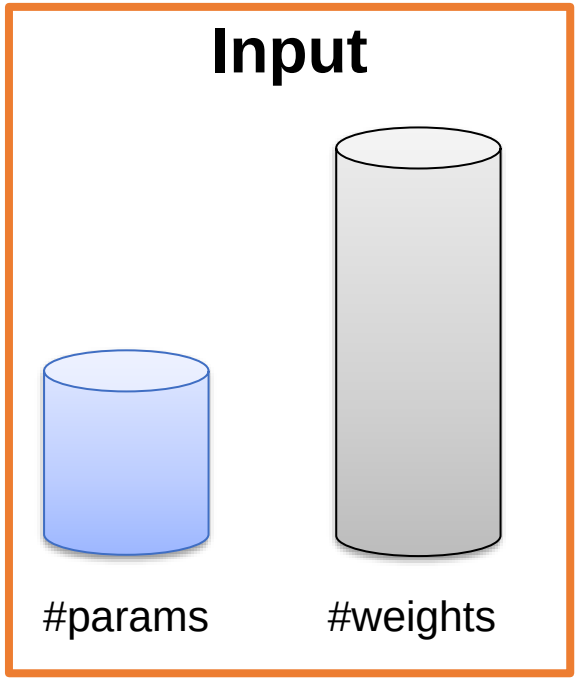


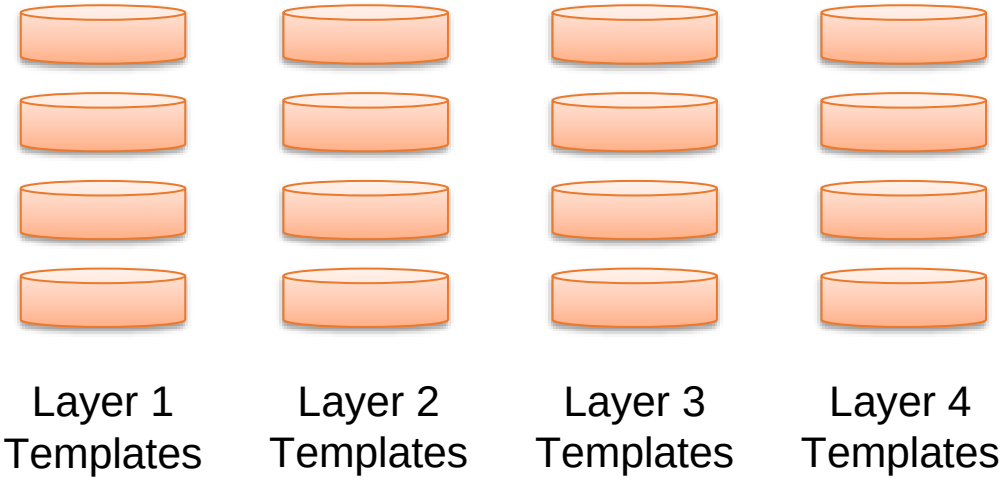
Generating weights from trainable parameters



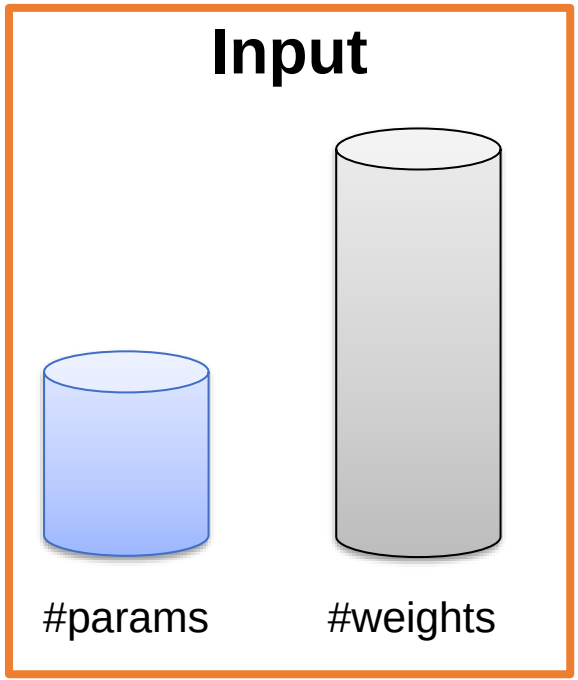


What if parameters cannot be effectively shared between some layers?





Learn when layers output similar weights!



Cluster mixing parameters, e.g., using K-means

$$\begin{bmatrix} w_{1,1} \\ w_{1,2} \\ w_{1,3} \\ w_{1,4} \end{bmatrix} \quad \begin{bmatrix} w_{2,1} \\ w_{2,2} \\ w_{2,3} \\ w_{2,4} \end{bmatrix} \quad \begin{bmatrix} w_{3,1} \\ w_{3,2} \\ w_{3,3} \\ w_{3,4} \end{bmatrix} \quad \begin{bmatrix} w_{3,1} \\ w_{3,2} \\ w_{3,3} \\ w_{3,4} \end{bmatrix}$$

Layer 1 Mixing Parameters

Layer 2 Mixing Parameters

Layer 3 Mixing Parameters

Layer 4 Mixing Parameters

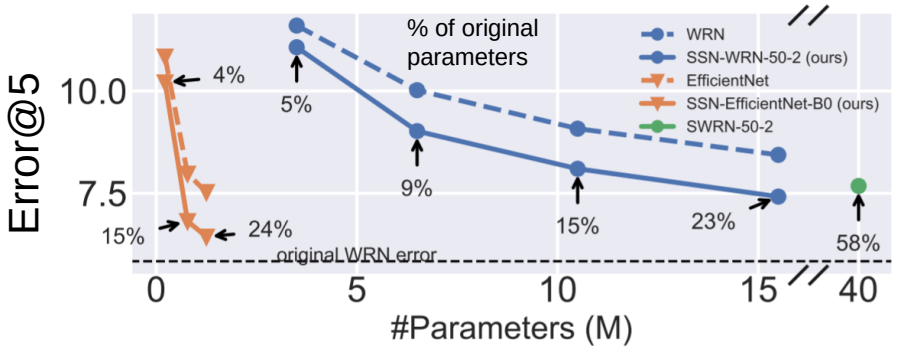
Question Answering F1 Measure

Network	#Params	SquAD v1.1	SquAD v2.0
BERT-Large	334M	92.2	85.0
ALBERT	18M	90.5	82.1
LB-SSN	18M	91.1	83.0

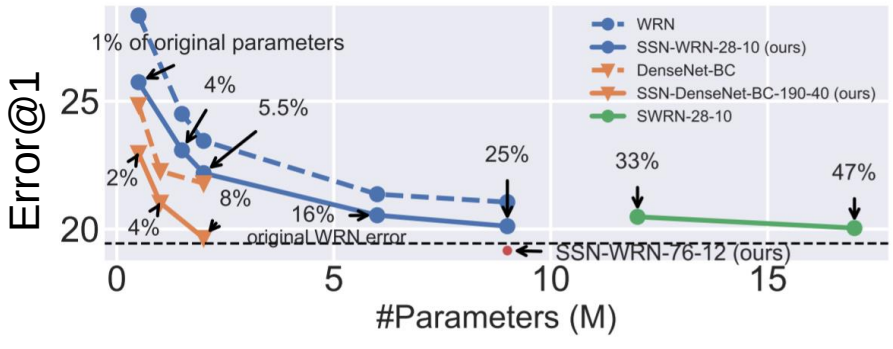
1.4x improvement on training speed using 128 V100 GPUs compared to BERT-Large

Image Classification

ImageNet



CIFAR100



Many more results for tasks like phrase grounding and parameter pruning in our paper!

Summary

- Introduced Neural Parameter Allocation Search, a novel task where the goal is to generate any architecture using a fixed budget of parameters
- Shapeshifter Networks address NPAS using cross-layer parameter sharing to provide an efficient and scalable solution to the task
- Results across nine architectures and four diverse tasks validate the effectiveness of our approach