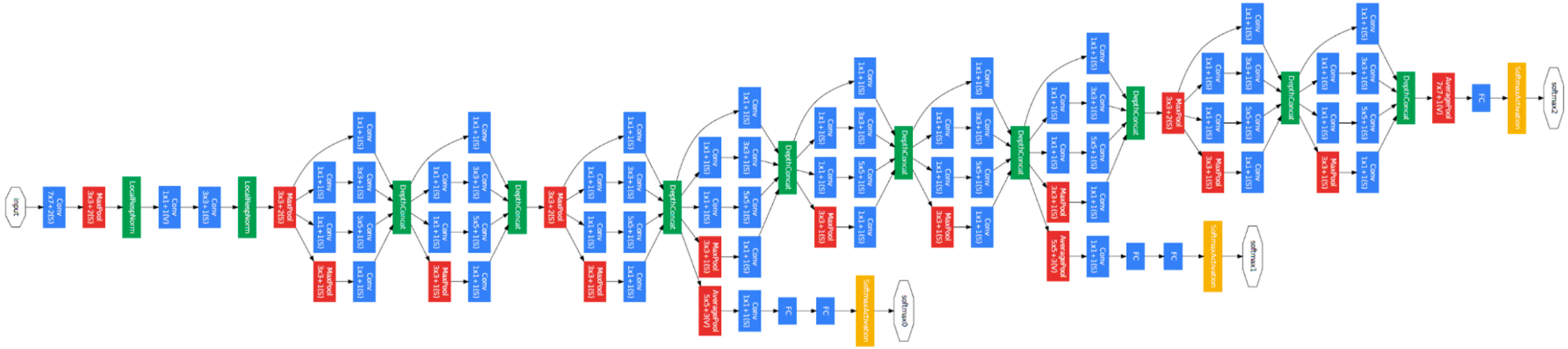
Learning Transferable Architectures for Scalable Image Recognition

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, Quoc V. Le

Google Brain

CVPR 2018

Motivation



GoogLeNet (2014): ImageNet Top-5 accuracy 93%

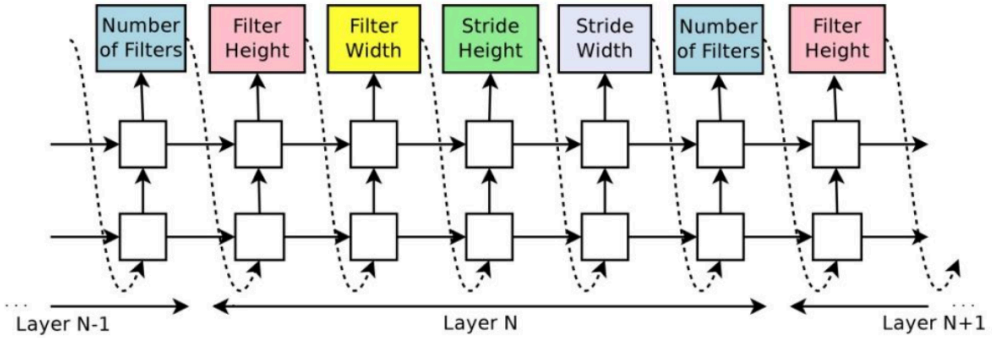
- CNN models require significant architecture engineering
- Can we design an algorithm to design the model architecture?

Previous Work

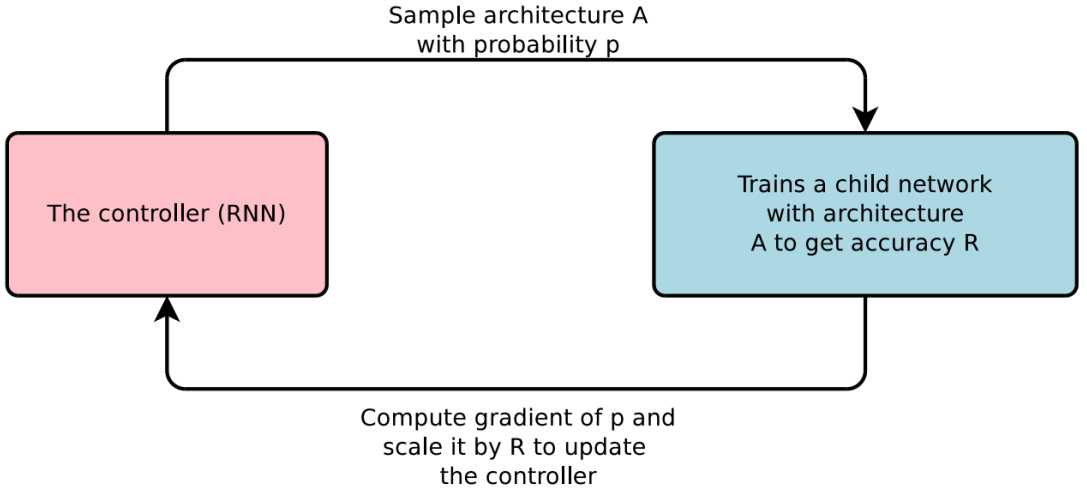
- Hyper-parameter Optimization
  - Included in NASNet
- Transfer Learned Architecture
  - Notably worse than other SOTA methods
- Meta-learning
  - Not applicable to large scale dataset (e.g. ImageNet)

Previous Work **(NAS)**

- Neural Architecture Search with Reinforcement Learning [Barret & Quoc 2017]



Controller RNN

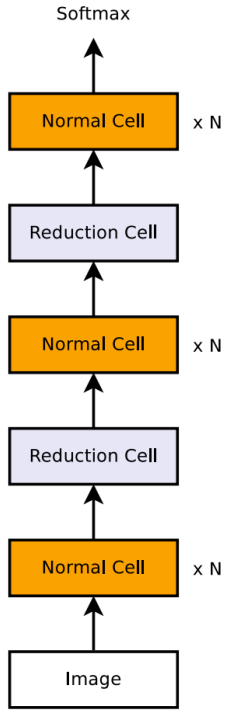


- NAS (2017) Limitations:
- Computationally expensive for only small datasets (e.g. CIFAR-10)
- No transferability between datasets

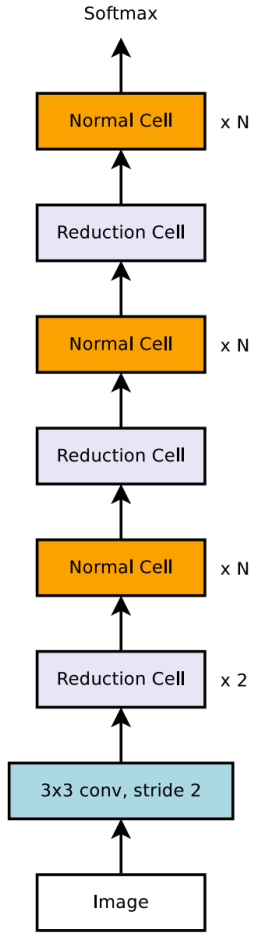
- NASNet (2018) : Re-designing the search space
- Computation cost is reduced
- Transferable from small dataset to large dataset

NASNet: Convolution Cells

- The overall architectures are manually predetermined
- Composed by two repetitive **Convolution Cells**:
- Normal Cell:
  - Output same-dimension feature map
- Reduction Cell:
  - Hight & Width of the output are halved
  - More reduction cells on ImageNet Architecture



CIFAR10 Architecture



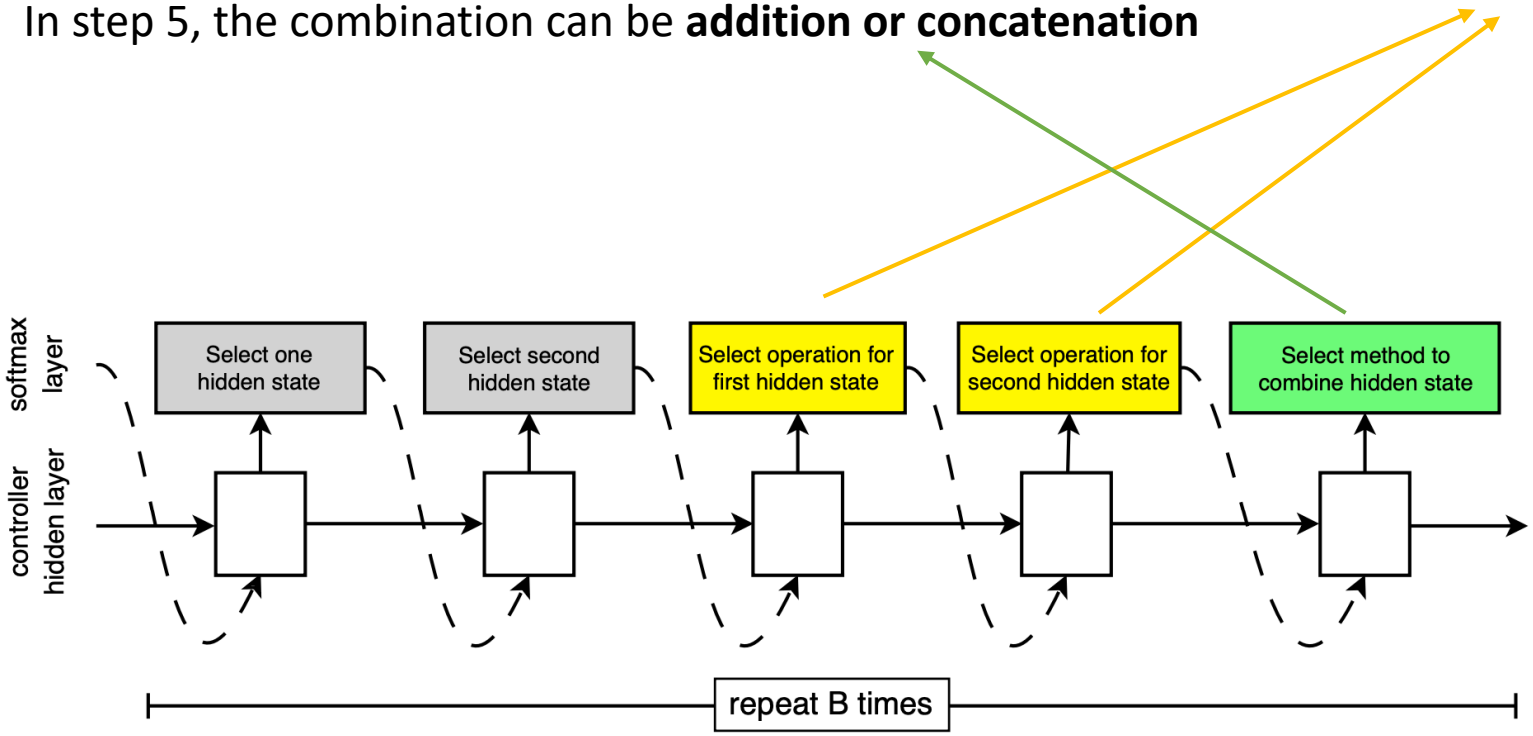
ImageNet Architecture

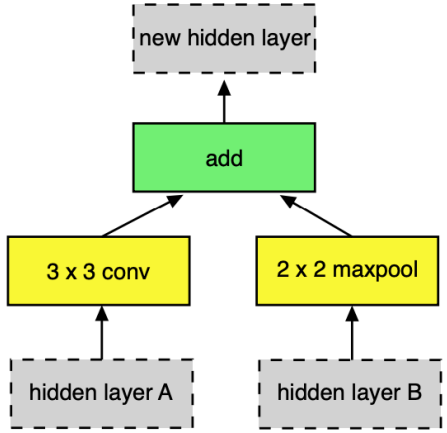
NASNet: Controller

- Predictions for each cell are grouped into ***B*** blocks
- Each block has 5 prediction steps
- In step 5, the combination can be **addition or concatenation**

- identity
- 1x7 then 7x1 convolution
- 3x3 average pooling
- 5x5 max pooling
- 1x1 convolution
- 3x3 depthwise-separable conv
- 7x7 depthwise-separable conv
- 1x3 then 3x1 convolution
- 3x3 dilated convolution
- 3x3 max pooling
- 7x7 max pooling
- 3x3 convolution
- 5x5 depthwise-seperable conv

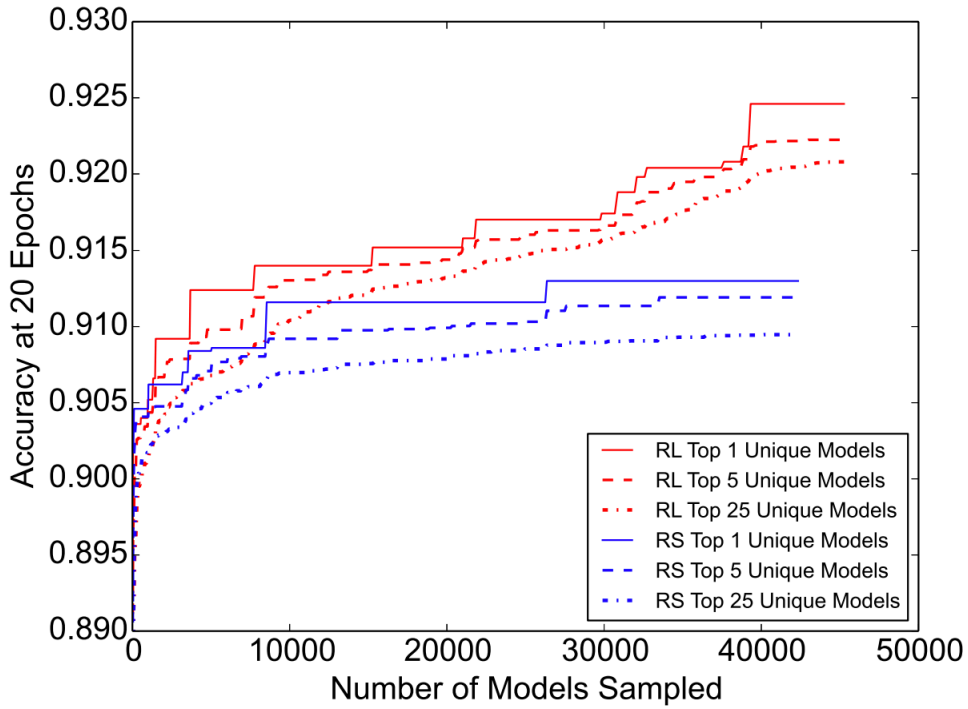
Operation list for step 3, 4





NASNet: Controller

- Controllers have $2 \times 5B$ predictions
- Trained by same reinforcement learning proposal as **NAS**
- **Random search** is applicable, but worse than RL

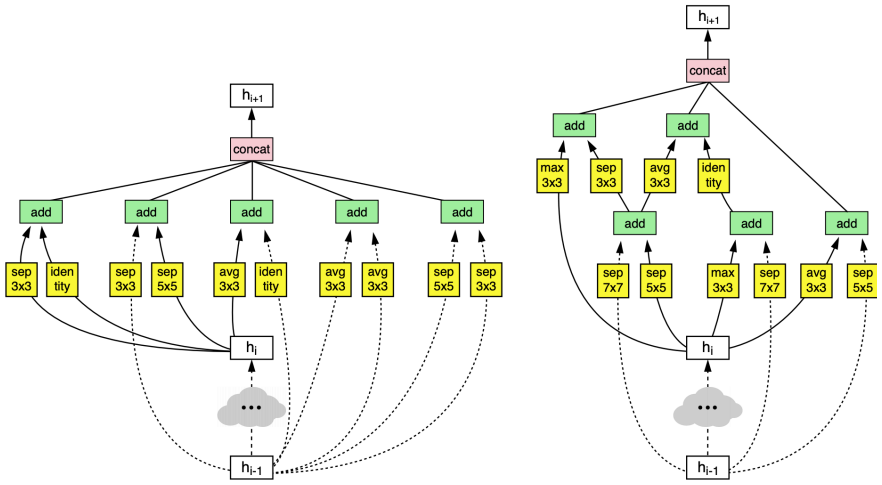


- Advantages:
  - Scalable to the different image datasets
  - Strong transferability in experiments.

- Disadvantages:
  - Training cost expensive: 500 GPUs over 4 days
  - Fixed layers

Results: CIFAR-10

Architecture of the best convolutional cells



Normal Cell

Reduction Cell

model	depth	# params	error rate (%)
DenseNet $(L = 40, k = 12)$ [26]	40	1.0M	5.24
DenseNet$(L = 100, k = 12)$ [26]	100	7.0M	4.10
DenseNet $(L = 100, k = 24)$ [26]	100	27.2M	3.74
DenseNet-BC $(L = 100, k = 40)$ [26]	190	25.6M	3.46
Shake-Shake 26 2x32d [18]	26	2.9M	3.55
Shake-Shake 26 2x96d [18]	26	26.2M	2.86
Shake-Shake 26 2x96d + cutout [12]	26	26.2M	2.56
NAS v3 [71]	39	7.1M	4.47
NAS v3 [71]	39	37.4M	3.65
NASNet-A (6 @ 768)	-	3.3M	3.41
NASNet-A (6 @ 768) + cutout	-	3.3M	2.65
NASNet-A (7 @ 2304)	-	27.6M	2.97
NASNet-A (7 @ 2304) + cutout	-	27.6M	2.40
NASNet-B (4 @ 1152)	-	2.6M	3.73
NASNet-C (4 @ 640)	-	3.1M	3.59

Table 1. Performance of Neural Architecture Search and other state-of-the-art models on CIFAR-10. All results for NASNet are the mean accuracy across 5 runs.

Results: Transfer to ImageNet

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Model size for [25] calculated from open-source implementation.

Results: Transfer to ImageNet

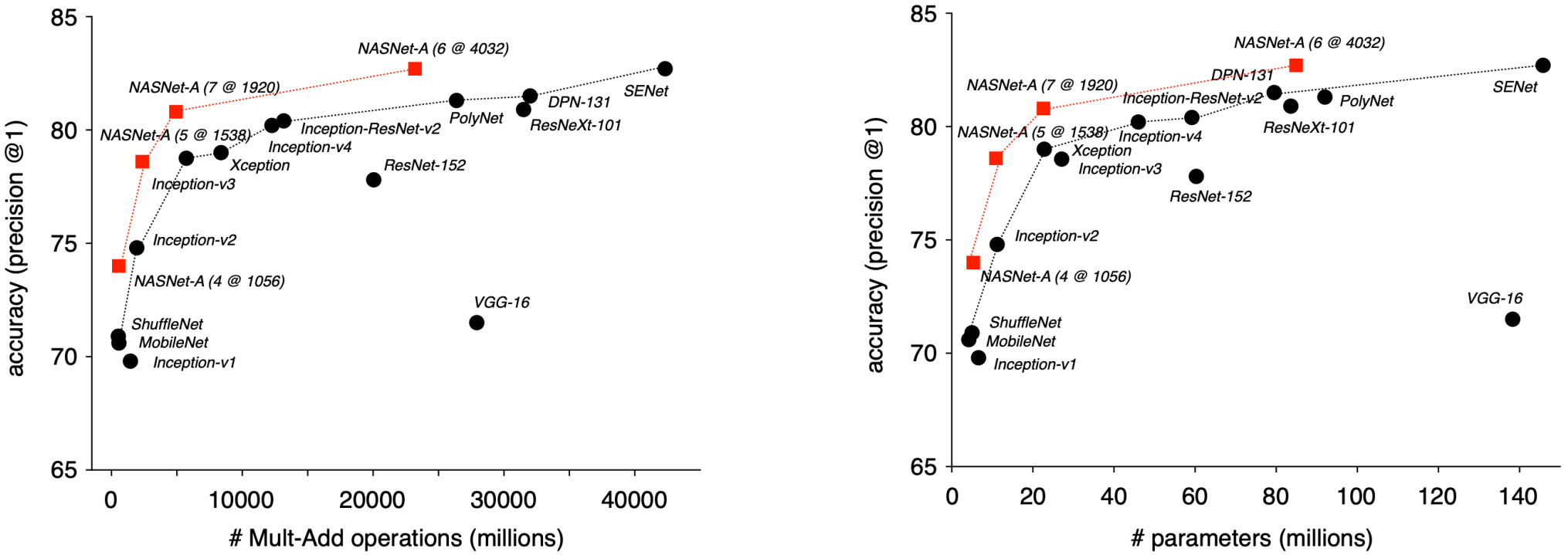


Figure 5. Accuracy versus computational demand (left) and number of parameters (right) across top performing published CNN architectures on ImageNet 2012 ILSVRC challenge prediction task. Computational demand is measured in the number of floating-point multiply-add operations to process a single image. Black circles indicate previously published results and red squares highlight our proposed models.

Results: Object detection

Model	resolution	mAP (mini-val)	mAP (test-dev)
MobileNet-224 [24]	600 × 600	19.8%	-
ShuffleNet (2x) [70]	600 × 600	24.5%†	-
NASNet-A (4 @ 1056)	600 × 600	**29.6%**	-
ResNet-101-FPN [36]	800 (short side)	-	36.2%
Inception-ResNet-v2 (G-RMI) [28]	600 × 600	35.7%	35.6%
Inception-ResNet-v2 (TDM) [52]	600 × 1000	37.3%	36.8%
NASNet-A (6 @ 4032)	800 × 800	41.3%	40.7%
NASNet-A (6 @ 4032)	1200 × 1200	**43.2%**	**43.1%**
ResNet-101-FPN (RetinaNet) [37]	800 (short side)	-	39.1%

Table 4. Object detection performance on COCO on *mini-val* and *test-dev* datasets across a variety of image featurizations. All results are with the Faster-RCNN object detection framework [47] from a single crop of an image. Top rows highlight mobile-optimized image featurizations, while bottom rows indicate computationally heavy image featurizations geared towards achieving best results. All *mini-val* results employ the same 8K subset of validation images in [28].

Conclusion & Discussion:

- Contribution: a novel search space for Neural Architecture Search
- Neural Architecture Search may improve the human-designed models
- Can we use similar method to construct an autoencoder?
- Is it possible to further reduce the training cost and computation cost?

Reference

- Neural Architecture Search with Reinforcement Learning
  - [Barret & Quoc 2017]

- Learning Transferable Architectures for Scalable Image Recognition
  - [Barret et al. 2018]