SMASH: One-Shot Model Architecture Search through Hypernetworks

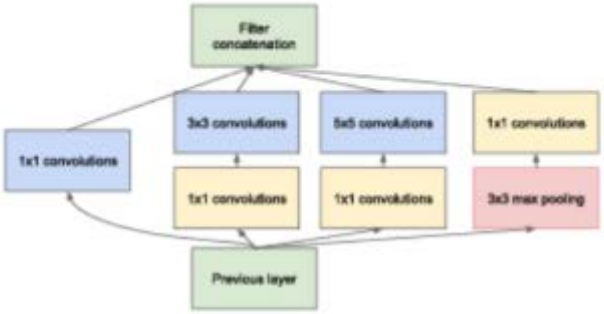
Andrew Brock

EDINBURGH CENTRE FOR ROBOTICS

Innovation Ready

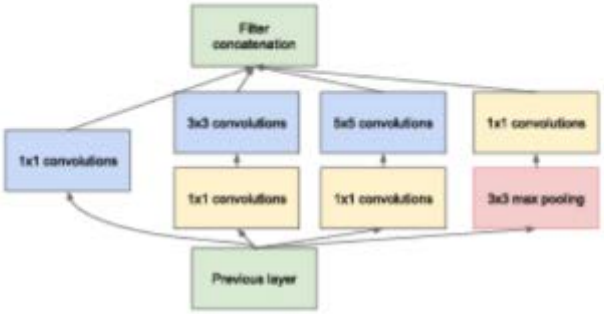
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

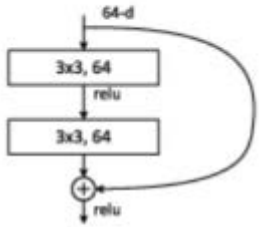
Designing Neural Nets is Hard!

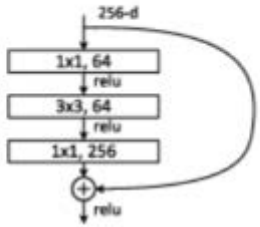


EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Designing Neural Nets is Hard!

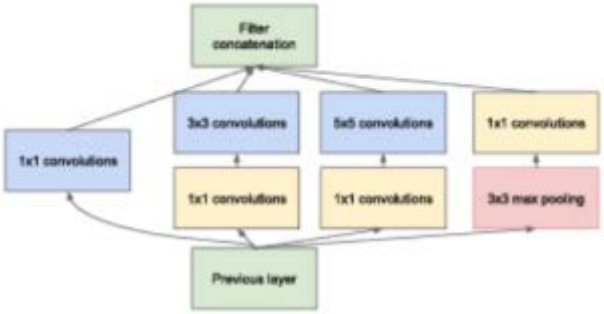


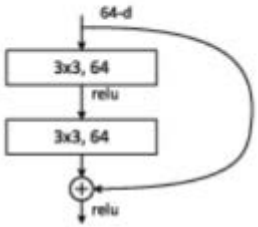


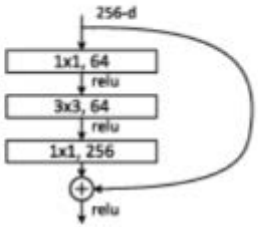


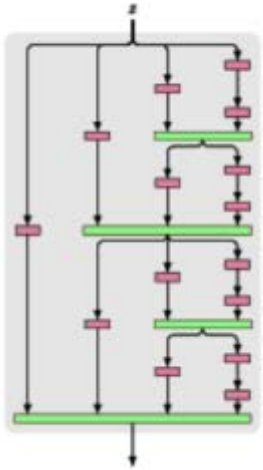
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Designing Neural Nets is Hard!



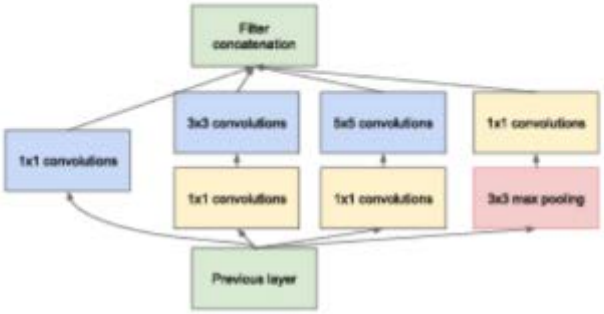


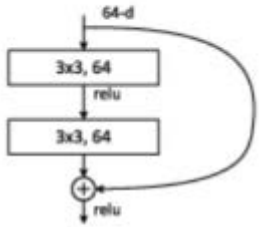


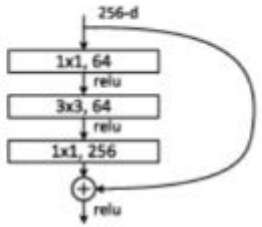


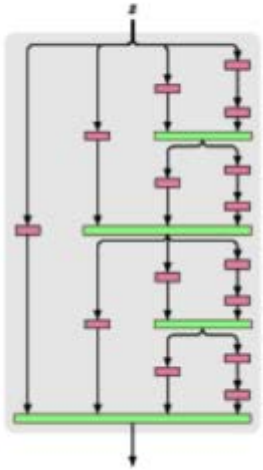
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

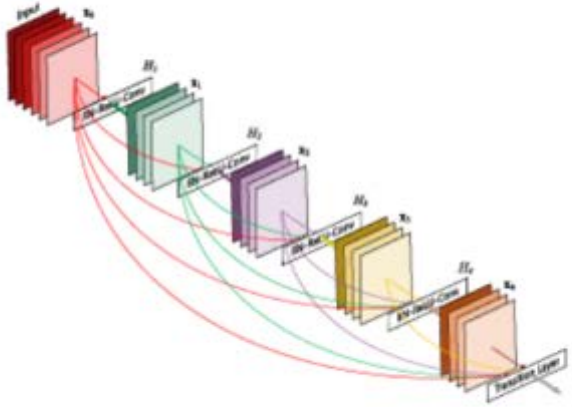
Designing Neural Nets is Hard!





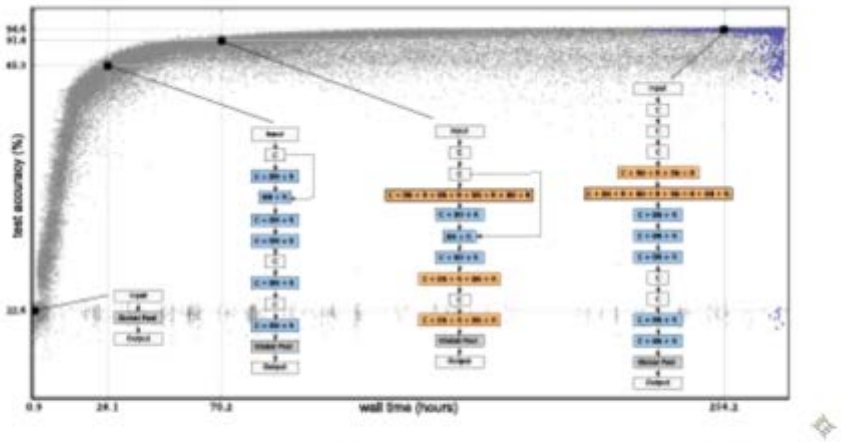






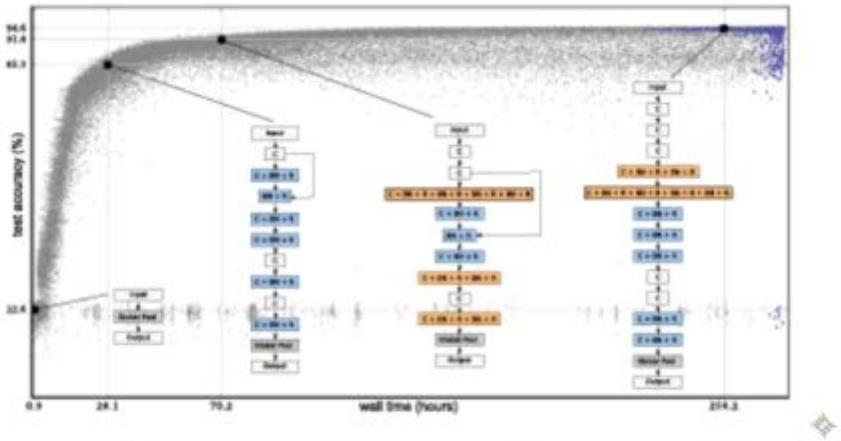
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

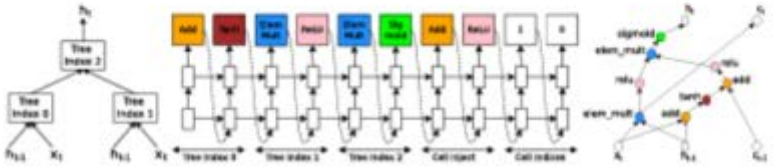
We *can* automate it…

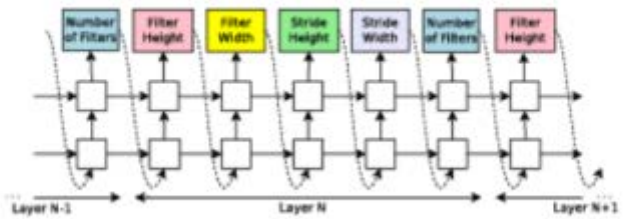


EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

We *can* automate it…

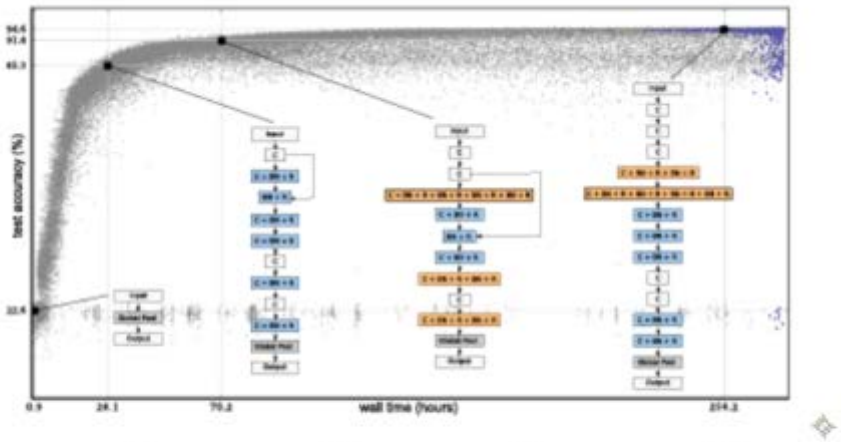


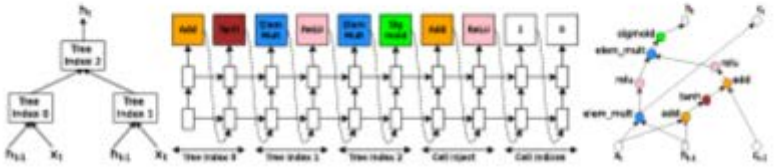


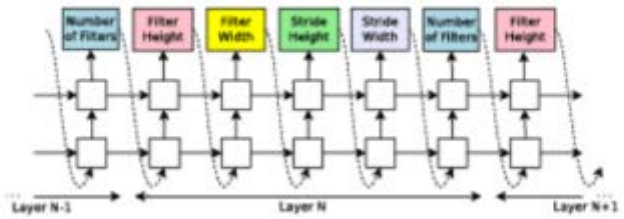


EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

But it can get expensive

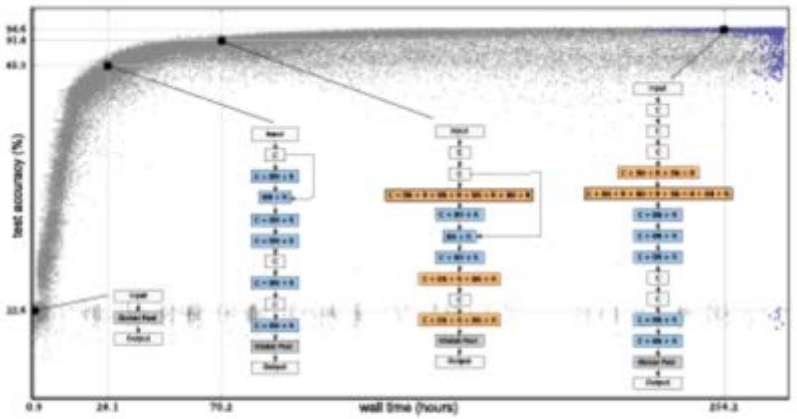


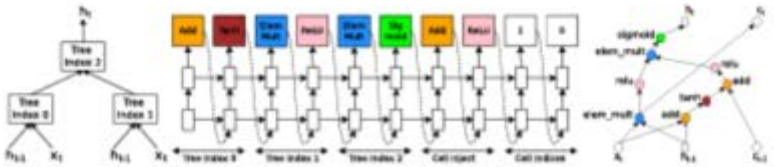


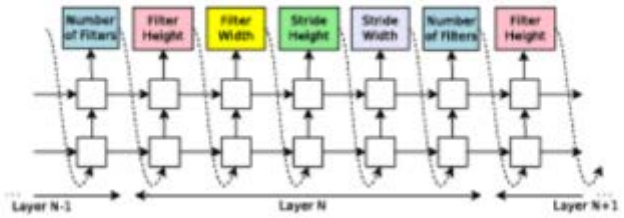


EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

But it can get expensive





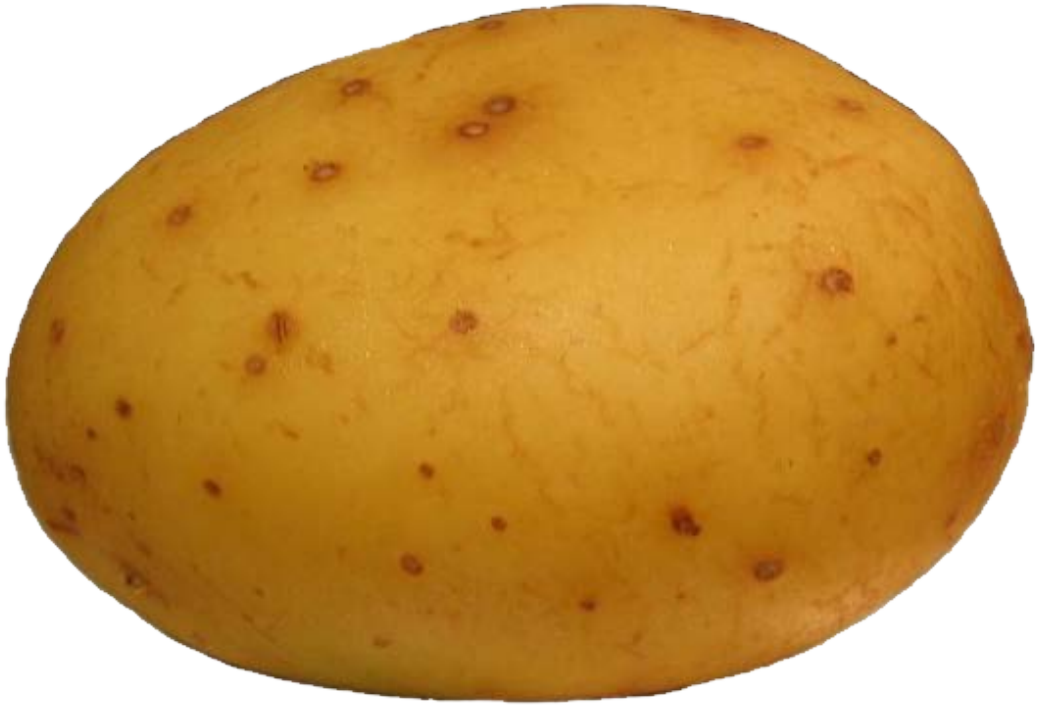


EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Reducing the cost

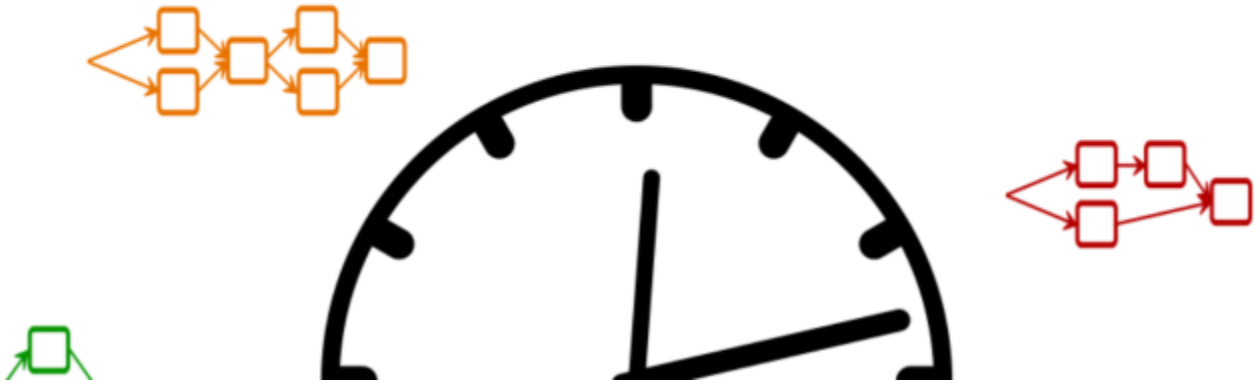
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Reducing the cost



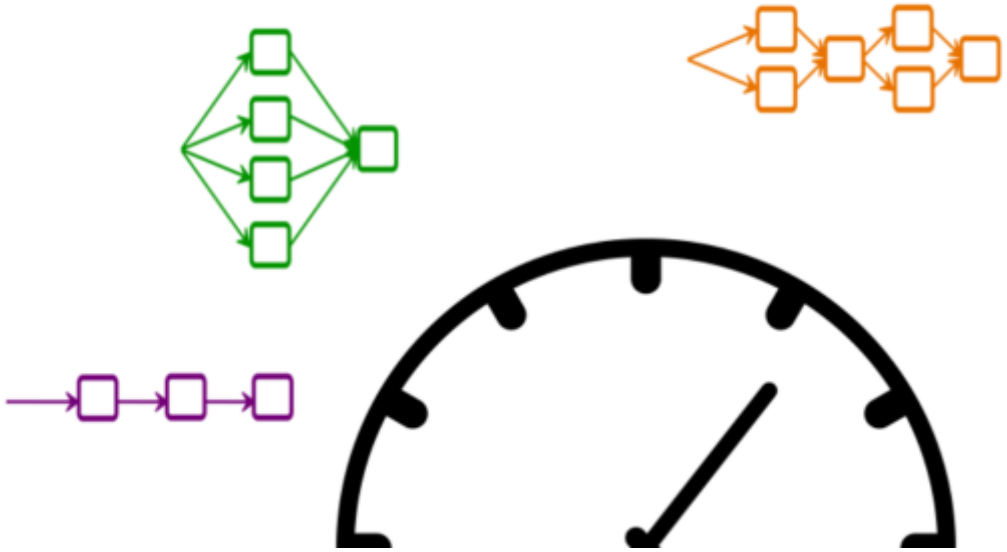
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Reducing the cost



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Reducing the cost



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

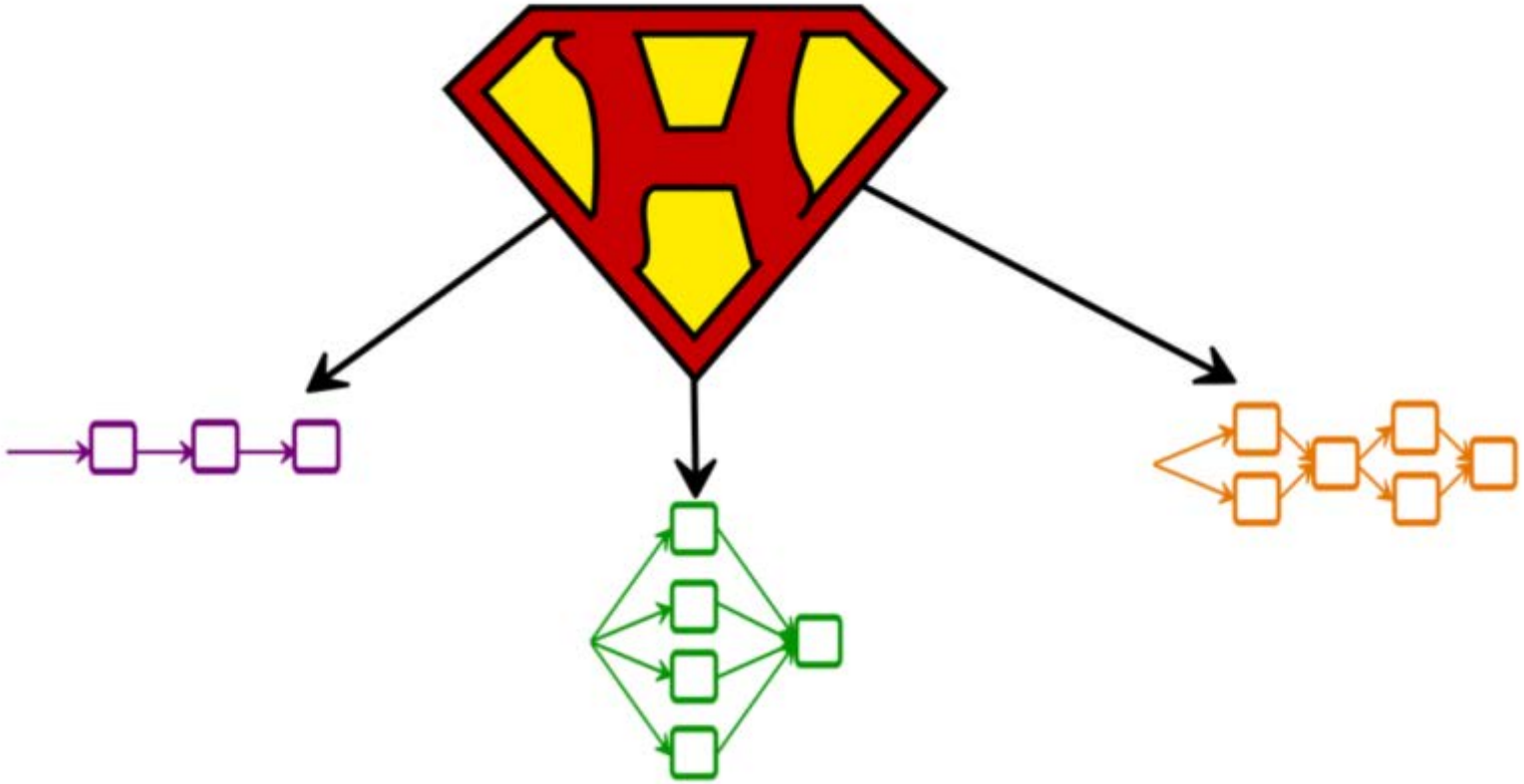
Reducing the cost

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Meta-Modeling

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Meta-Modeling



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

How's it work?

-Sample a random architecture

-Generate weights conditioned on that architecture, $W=H(c)$

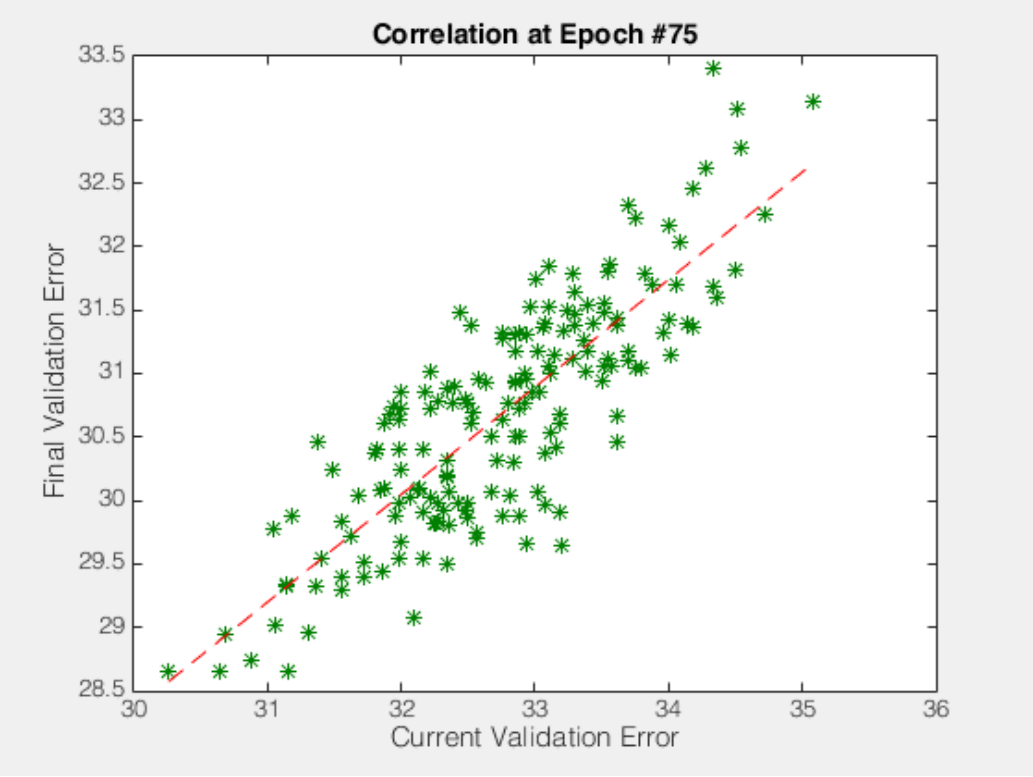
-Train the whole thing end-to-end

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

How's it work?

-Fully Convolutional Hypernet

-Weights need only be "close to optimal"



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

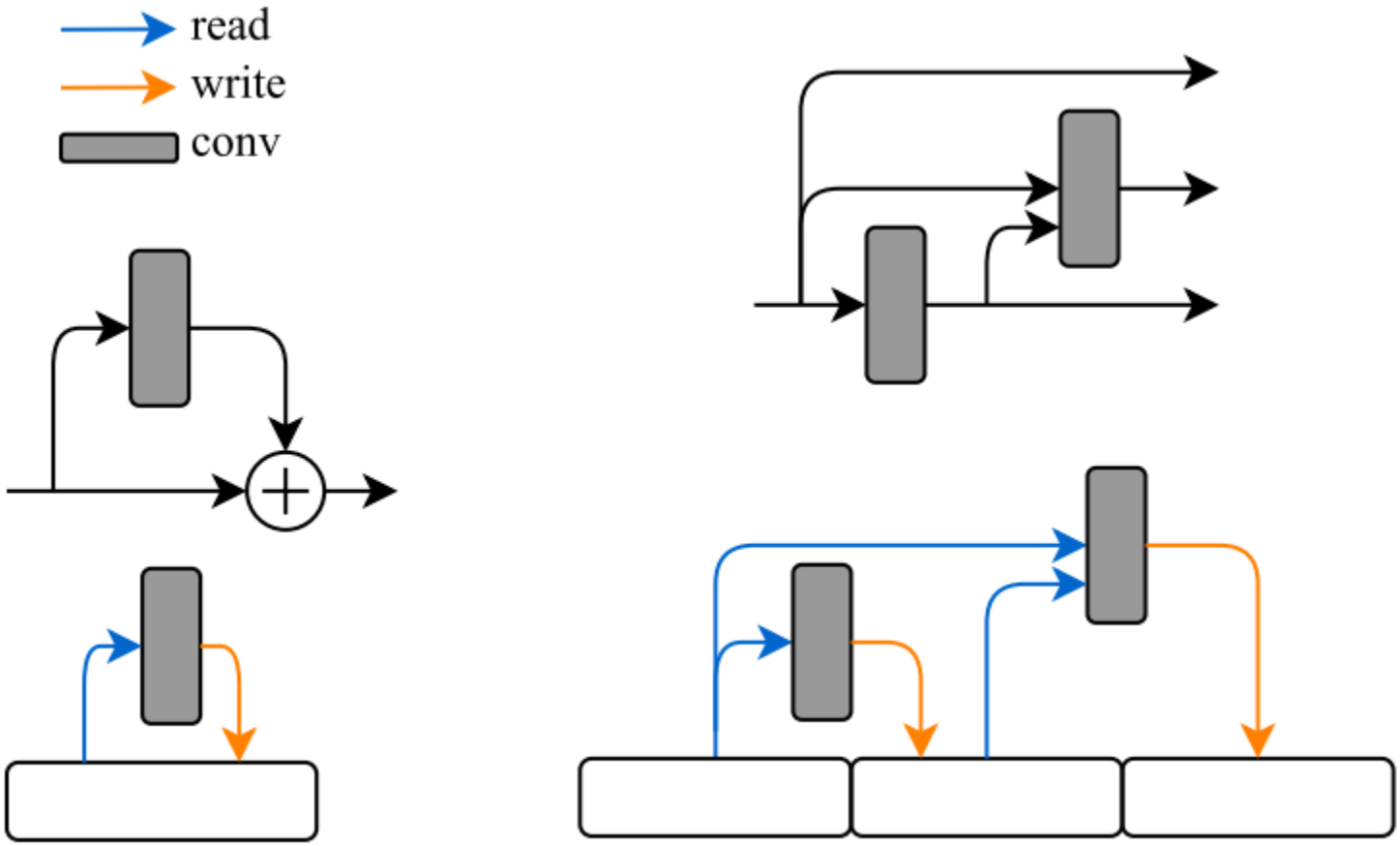
How's it work?

Okay, great.

But how do we sample random architectures?

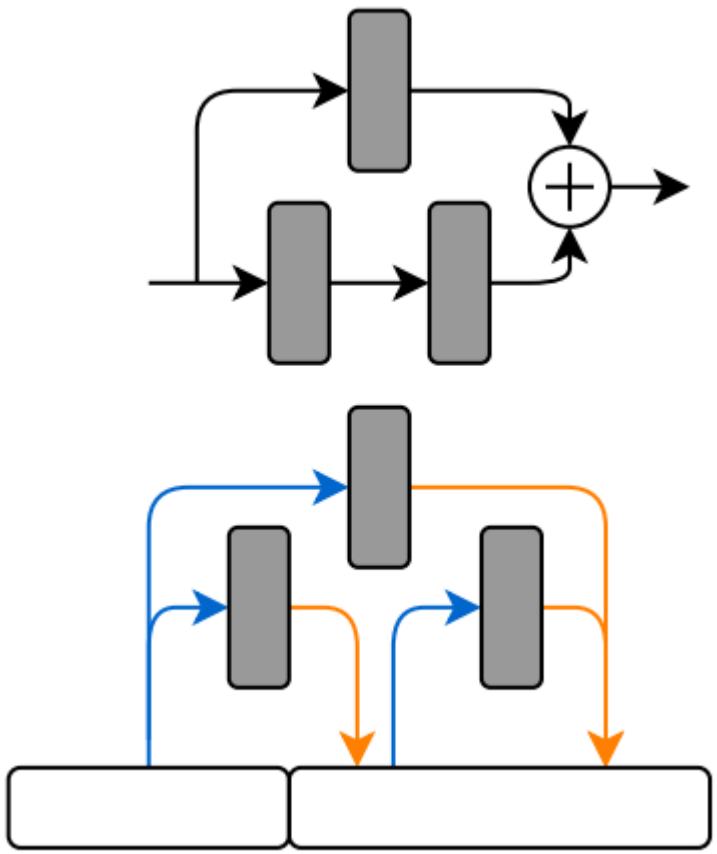
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Memory-Bank View



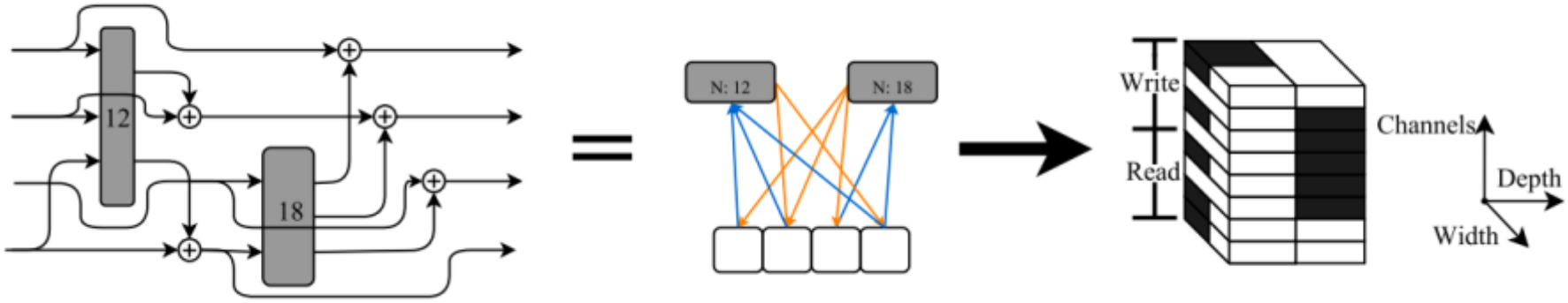
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Memory-Bank View



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Encoding an Architecture



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

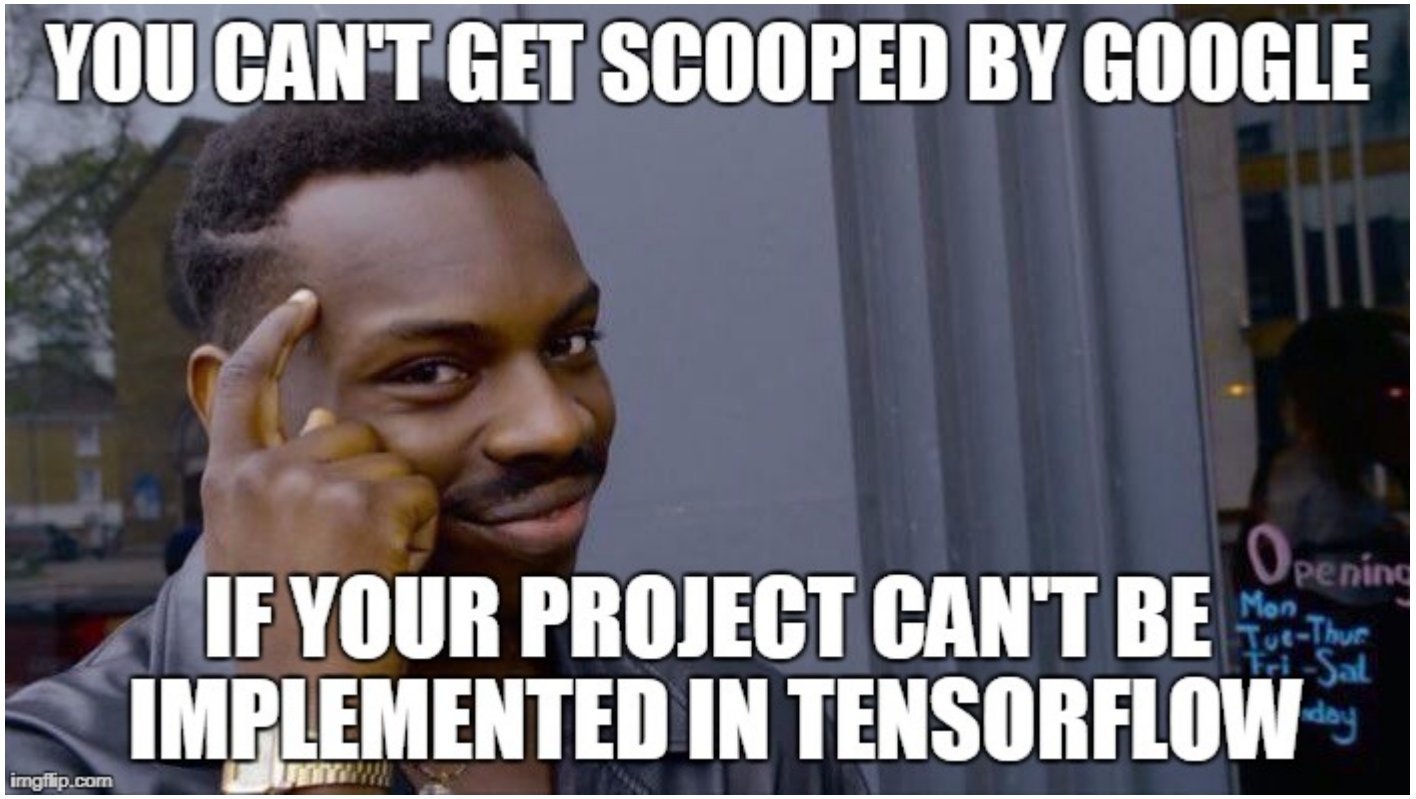
Encoding an Architecture

```
self.hyperconv(embedding).squeeze(0)\
                         .transpose(0,1)\
                         .contiguous()\
                         .view(self.N_max, int((slice_index + 1) * self.depth_compression* self.N))\
                         .index_select(1, V(torch.LongTensor([i for item in
                                                [range(q, (slice_index + 1) * self.depth_compression * self.N, slice_index + 1) for q in range(slice_index + 1)] for i in item]).cuda()))\
                         .transpose(0,1)\
                         .unsqueeze(2).unsqueeze(3)\
                         .contiguous()
```

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

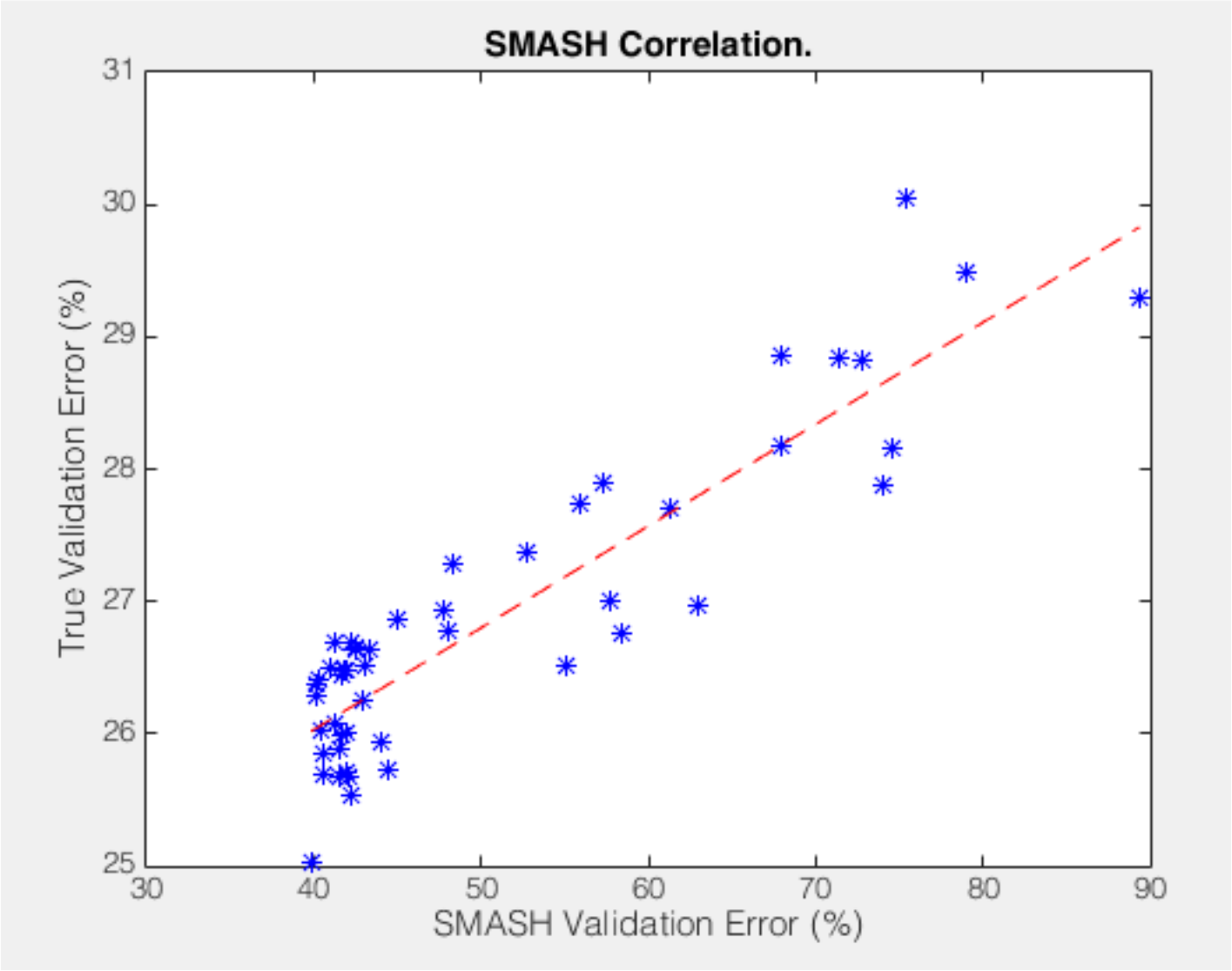
Implementation

PYTORCH



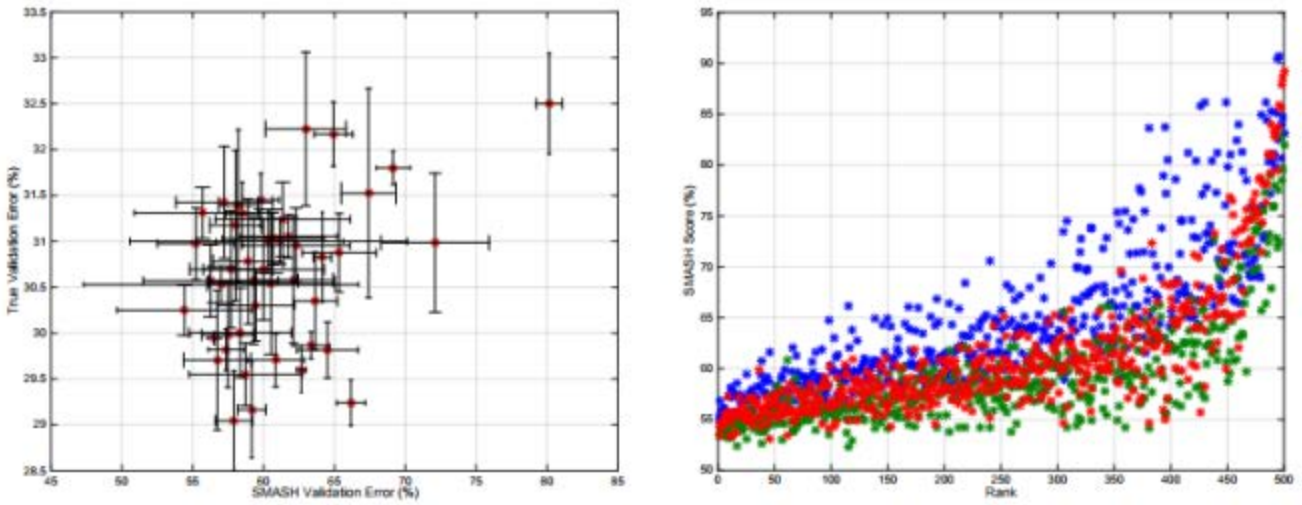
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation



(a) (b)

Figure 5: (a) SMASH correlation with a crippled HyperNet. Error bars represent 1 standard deviation. (b) SMASH scores vs. rank using average scores from three HyperNets with different seeds.

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation: Mapping C to W

-Perturbing c without changing the architecture reduces SMASH scores

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation: Mapping C to W

-Perturbing c without changing the architecture reduces SMASH scores

-If the net really has learned a good mapping W=H(c), does dE/dc provide a meaningful gradient for the architecture?

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation: Mapping C to W

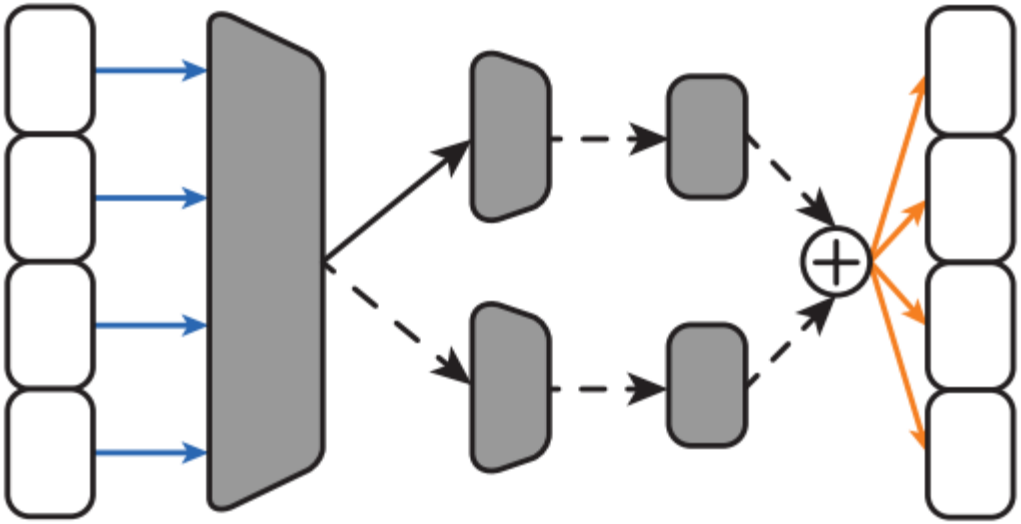
-Perturbing c without changing the architecture reduces SMASH scores

-If the net really has learned a good mapping W=H(c), does dE/dc provide a meaningful gradient for the architecture?

-Architectural gradient descent by proxy

EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

Evaluation

Table 1: Error rates (%) on CIFAR-10 and CIFAR-100 with standard data augmentation (+).

Method	Depth	Params	C10+	C100+
FractalNet [20]	21	38.6M	5.22	23.30
with Dropout/Drop-path	21	38.6M	4.60	23.73
Wide ResNet [43]	16	11.0M	4.81	22.07
	28	36.5M	4.17	20.50
DenseNet-BC ($k = 24$) [15]	250	15.3M	3.62	17.60
DenseNet-BC ($k = 40$)	190	25.6M	3.46	17.18
Shake-Shake [11]	26	26.2M	**2.86**	**15.85**
Neural Architecture Search w/ RL[44]	39	32.0M	3.84	-
NASNet-A	20	3.3M	3.41	-
MetaQNN [3]	9	11.18M	6.92	27.14
Large-Scale Evolution [26]	-	5.4M	5.40	-
	-	40.4 M	-	23.7
CGP-CNN [38]	-	1.68M	5.98	-
SMASHv1	116	4.6M	5.53	22.07
SMASHv2	211	16M	4.03	20.60

Table 2: Error rates (%) on STL-10.

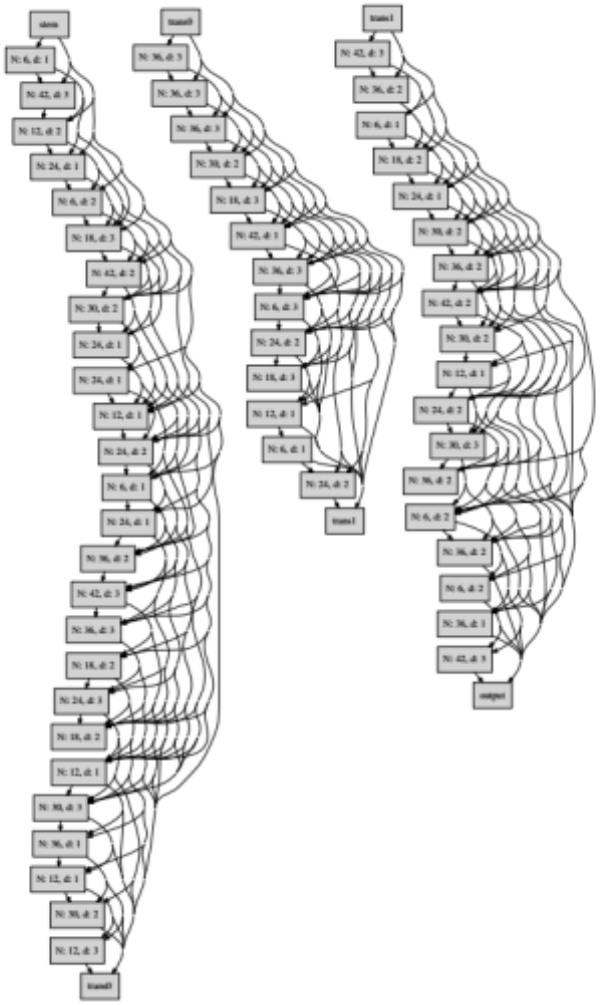
Model	Params	Error
WRN-40-4	8.95M	**35.02 ± 1.14**
WRN-28-10	36.5M	36.69 ± 2.06
SMASHv2	16.2M	41.52 ± 2.10
SMASHv2 (3x3)	4.38M	37.76 ± 0.58

Table 3: Error rates (%) on Imagenet32x32.

Model	Params	Top-1	Top-5
WRN-28-2	1.6M	56.92	30.92
WRN-28-5	9.5M	45.36	21.36
WRN-28-10	37.1M	40.96	18.87
SMASHv2	16.2M	**38.62**	**16.33**

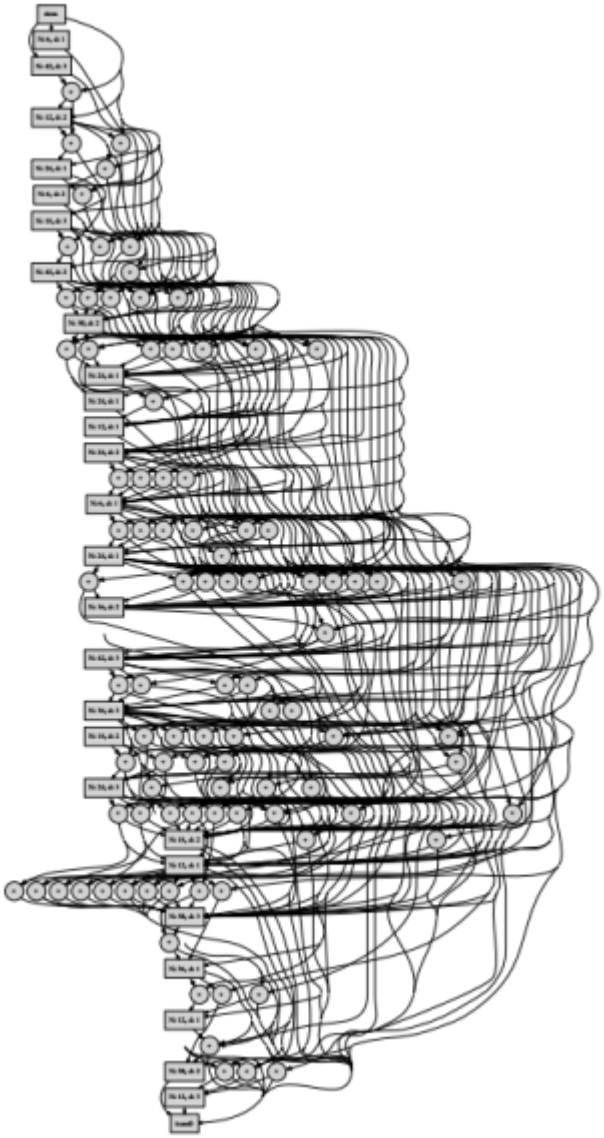
EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

What does a SMASH net look like?



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

What does a SMASH net look like?



EDINBURGH CENTRE FOR ROBOTICS
Innovation Ready

What's Next?

-More rigorous investigation of HyperNet capacity

-Different search spaces

-Better sampling (RNN-based)

-Harder problems!

Thanks!

Andrewjbrock92@gmail.com

EDINBURGH CENTRE FOR
ROBOTICS
Innovation Ready

EDINBURGH CENTRE FOR
ROBOTICS
Innovation Ready

THE UNIVERSITY
of EDINBURGH

HERIOT
WATT
UNIVERSITY

EPSRC
Engineering and Physical Sciences
Research Council