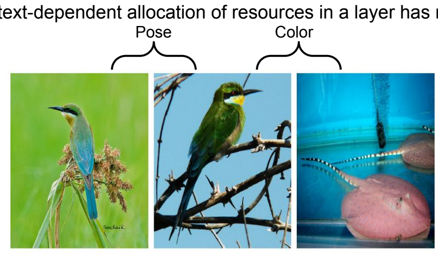
Feature-Dependent Cross-Connections in Multi-Path Networks

Dumindu Tissera[*+], **Kasun Vithanage**[*+], **Rukshan Wijesinghe**[*+], **Kumara Kahatapitiya**[*], **Subha Fernando**[*+] and **Ranga Rodrigo**[*]

[*]University of Moratuwa [+]CODEGEN QBITS Lab **Sri Lanka**

Rich Layer-wise Feature Extraction by Multi-paths

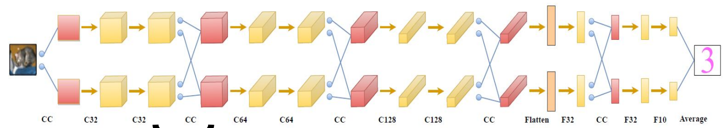
- Neural network deepening is well established to learn complex tasks.
- However, there is still room for powerful feature extraction within layers.
- As opposed to conventional widening, having parallel computations in neural layers improve the efficiency with respect to the number of parameters.
- However, context-dependent allocation of resources in a layer has not been explored

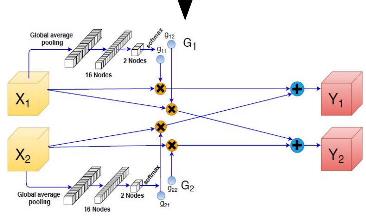


a) Hummingbird b) Hummingbird c) Electric Eel

- 3 ImageNet images, Two Hummingbirds and one Electric Eel
- **a** and **b** similar in pose (abstract detail). **b** and **c** similar in overall color (low detail).
- Image context is distributed along the depth of a neural network.
- In a multi-path network, the nature of resource allocation may change with the depth.
- Therefore, It is intuitive to learn to allocate parallel resources separately, layer-wise.
- In this way, **b** and **c** may get similar resource allocation in initial layers, **a** and **b** may get similar resource allocation in deeper layers.

Feature-Dependent Cross-Connections





- We introduce feature-dependent cross-connections at selected locations in multi-path networks.
- This enable the network to group homogeneous feature maps in each layer to parallel paths and route the input end-to-end through such paths
- Now, parallel resource allocation per layer segment is dependent on the incoming feature maps.
- Here, we show a two path CNN with adaptive cross-connections inserted at selected locations. The cross-connections are weighted by gates which are computed from the input tensors themselves. Connecting (X1,X2) to (Y1,Y2) is described below.
- We feed X to global average pooling, followed by non-linear parametric computation which outputs two gates, the probabilities of X being routed to Y1 and Y2.
- The output Y is constructed by summing the Xs which are weighted by the corresponding gates.

Image Recognition Domain

Compared Network Details (Cn Convolution Fn Dense)

Network	Structure
BaseCNN	$C32\ C32\ C64\ C64\ C128\ C128\ F32\ F32\ F10$
WideCNN	$C64\ C64\ C128\ C128\ C256\ C256\ F32\ F32\ F10$
DeepCNN	$C32\ C32\ C64\ C64\ C128\ C128\ C128$ $C256\ C256\ C256\ F32\ F32\ F10$
BaseCNN Ensemble	Ensemble of 3 BaseCNNs
All Ensemble	Ensemble of BaseCNN, WideCNN and DeepCNN
SENet	SENet on BaseCNN and DeepCNN
Cr-Stitch2	Cross-stitch network with 2 parallel BaseCNNs
BaseCNN-X	BaseCNN - X paths with adaptive cross-connections
ResNet-X	ResNet - X paths with adaptive cross-connections

Ablation study in CIFAR10 - CNN based networks

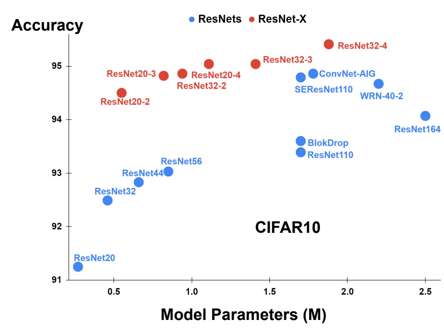
Network	Params (M)	Error %
BaseCNN/WideCNN/DeepCNN	0.55/1.67/2.0	9.26/8.96/8.49
BaseCNN Ensemble	1.66	7.87
All Ensemble	4.27	6.9
SEBaseCNN/SEDeepCNN	0.58, 2.06	8.99,8.15
Cr-Stitch2	1.11	7.89
VGG16	14.9	6.98
BaseCNN-2	1.11	7.03
BaseCNN-3	1.67	**6.51**
BaseCNN-4	2.22	**6.55**

CIFAR10 and CIFAR100 CNN and ResNet based Nets

Network	Params (M)	CIFAR10	CIFAR100
ResNet20/110/164	0.27/1.7/2.5	8.75/6.61/5.93	-/26.88/25.16
WRN-40-2	2.2	5.33	26.04
SEResNet110	1.7	5.21	**23.85**
BlockDrop	1.7	6.4	26.3
ConvNet-AIG	1.78	5.76	-
ConvNet-AIG all	1.78	5.14	-
ResNet20-2 Ours	0.55	5.5	27.36
ResNet20-3 Ours	0.82	5.18	25.76
ResNet20-4 Ours	1.1	**4.96**	**24.81**
ResNet32-2 Ours	0.94	5.14	25.96
ResNet32-3 Ours	1.41	**4.96**	**24.51**
ResNet32-4 Ours	1.88	**4.59**	**23.52**
Sabour *et al.*	8.2	10.6	-
Highway	2.3	7.54	-
HyperNetworks	0.15	7.23	-
BaseCNN-2/3/4 Ours	1.11/1.67/2.22	**6.53/6.09/6.26**	-

ILSVRC2012 (ImageNet) ResNet based Nets

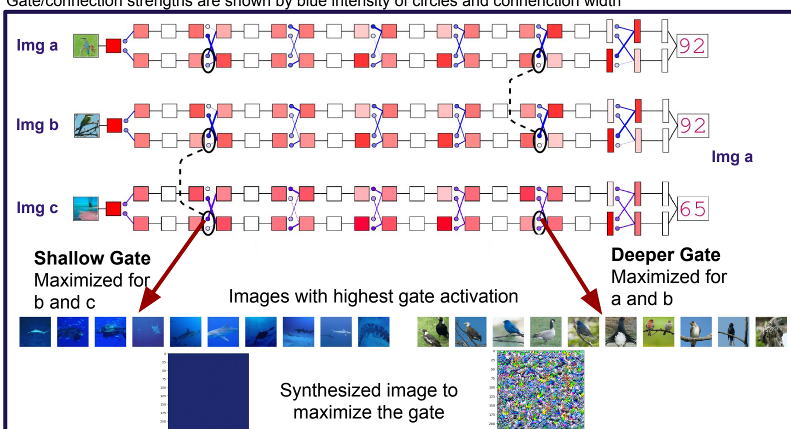
Network	Params(M)	Single-Crop		10-Crop	
		Top-1	Top-5	Top-1	Top-5
ResNet18	11.7	30.4	10.93	28.22	9.42
ResNet34	21.8	26.77	8.77	24.52	7.46
WRN-18-1.5	25.9	27.06	9.0		
ResNet18-2 (Ours)	23.4	**26.48**	**8.6**	**24.5**	**7.34**



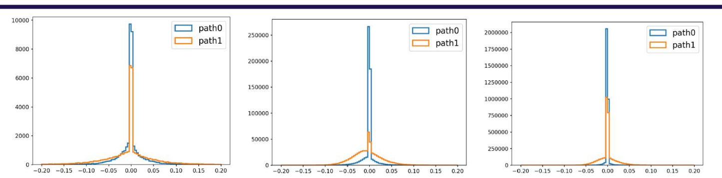
- We conduct experiments in both small scale and large scale image recognition datasets.
- We test multi-path CNNs and multi-path ResNets
- Overall, our networks surpass, conventional widening, model ensembles, existing adaptive feature extraction methods and even deeper networks with similar or less number of parameters.

Routing diagrams. Input/output tensor strengths are denoted in red intensity. Gate/connection strengths are shown by blue intensity of circles and connenction width

Visualizations



- We chose two gates to study the gating patterns. **Shallow gate**, **Deeper gate.**
- The **shallow gate** is maximized for **b** and **c,** although they are two classes. It shows a lower activation for **a.** The highest activated images contain blue color. The synthesized image that maximizes the shallow gate also contains a consistent blue color.
- The **deeper gate** is maximized for two hummingbirds **a** and **b**. It shows a lower activation for the electric eel (**c**). The maximally activated images are all birds. The synthesized image also contains bird patterns.



layer 4 layer 8 layer 11

- We further plot gate histograms of parallel computations in selected layers.
- The histograms of parallel computations are distinct**.** This confirms that parallel computations learn distinct features.

Conclusion

- We selectively introduced adaptive cross-connections between successive pairs of layers in a multi-path network, to group similar feature maps into parallel paths and learn a soft routing between them.
- Our multi-path networks are capable of surpassing state-of-the-art adaptive image classifiers and conventional single path networks of increased width or depth which are of similar or even higher complexity.
- Having parallel paths with adaptive cross-connections is a good alternative to improve neural network performance under resource constraints.