FractalNet: Ultra-deep Neural Networks Without Residuals

(ICLR 2017)

Gustav Larsson[1],Michael Maire[2] ,Gregory Shakhnarovich[2]
1: University of Chicago 2: TTI Chicago
https://qdata.github.io/deep2Read
Presenter: Arshdeep Sekhon

Fall 2018

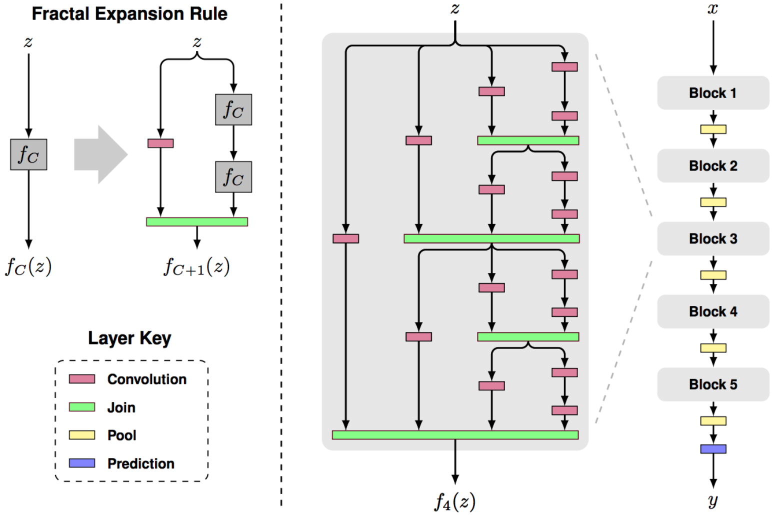
Introduction

- ResNets improve depth and accuracy
- ResNets learn to predict residual outputs not absolute mappings
- a type of deep supervision as near-identity layers effectively reduce distance to the loss.

Fractal Nets

- subnetworks of many depths
- does not rely on residuals
- following characteristics not hard wired: modular, student-teacher learning, deep supervision
- DropPath: regularization techniques

Method: Fractal Nets



FractalNet: Ultra-dee

Method: Fractal Networks

- networks structure, connections and layer types, is defined by $f_C()$.
- successive fractals$=$ $f_{C+1}(z) = [f_C \odot f_C(z)] + [conv(z)]$
- $\odot$ denotes composition and $+$ denotes join/concat operation; C number of columns
- Depth: scales as 2^{C-1}

Regularization via Drop-path

Related: change interactions to discourage co-adaptation

- dropout
- drop-connect

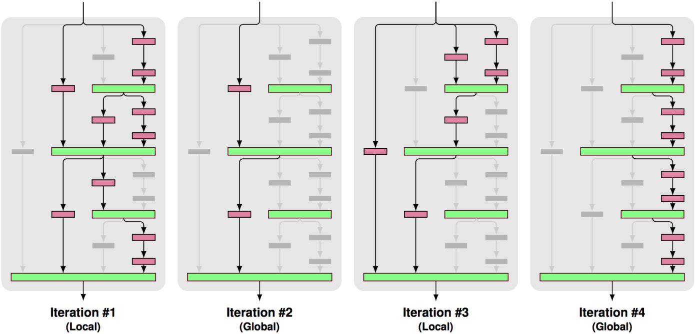
drop path

- prevent co-adaptation of parallel paths
- randomly drop operands of join

Drop-Path

- local: randomly remove inputs from join
- global: select single path for entire net, to allow for individual columns to act as good predictors

local and global drop-path



- A global sampling strategy returns a single column as a subnetwork.
- Alternating it with local sampling encourages the development of individual columns as performant stand-alone subnetworks.

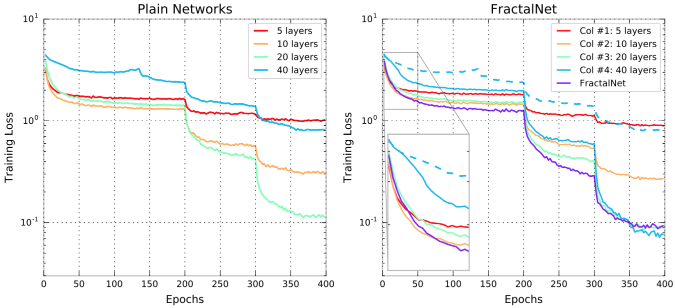
Drop-path

- sample a new subnetwork each mini-batch.
- With sufficient memory, we can simultaneously evaluate one local sample and all global samples for each mini-batch by keeping separate networks and tying them together via weight sharing.
- global drop-path forces the use of many paths whose lengths differ by orders of magnitude (powers of 2).
- The subnetworks by drop-path exhibit large structural diversity.

Implementation and Results

Method	C100	C100+	C100++	C10	C10+	C10++	SVHN
Network in Network (Lin et al., 2013)	35.68	-	-	10.41	8.81	-	2.35
Generalized Pooling (Lee et al., 2016)	32.37	-	-	7.62	6.05	-	1.69
Recurrent CNN (Liang & Hu, 2015)	31.75	-	-	8.69	7.09	-	1.77
Multi-scale (Liao & Carneiro, 2015)	27.56	-	-	6.87	-	-	1.76
FitNet Romero et al. (2015)	-	35.04	-	-	8.39	-	2.42
Deeply Supervised (Lee et al., 2014)	-	34.57	-	9.69	7.97	-	1.92
All-CNN (Springenberg et al., 2014)	-	33.71	-	9.08	7.25	4.41	-
Highway Net (Srivastava et al., 2015)	-	32.39	-	-	7.72	-	-
ELU (Clevert et al., 2016)	-	24.28	-	-	6.55	-	-
Scalable BO (Snoek et al., 2015)	-	-	27.04	-	-	6.37	1.77
Fractional Max-Pool (Graham, 2014)	-	-	26.32	-	-	3.47	-
FitResNet (Mishkin & Matas, 2016)	-	27.66	-	-	5.84	-	-
ResNet (He et al., 2016a)	-	-	-	-	6.61	-	-
ResNet by (Huang et al., 2016b)	44.76	27.22	-	13.63	6.41	-	2.01
Stochastic Depth (Huang et al., 2016b)	37.80	24.58	-	11.66	5.23	-	1.75
Identity Mapping (He et al., 2016b)	-	22.68	-	-	4.69	-	-
ResNet in ResNet (Targ et al., 2016)	-	22.90	-	-	5.01	-	-
Wide (Zagoruyko & Komodakis, 2016)	-	20.50	-	-	4.17	-	-
DenseNet-BC (Huang et al., 2016a)[1]	19.64	17.60	-	5.19	3.62	-	1.74
FractalNet (20 layers, 38.6M params)	35.34	23.30	22.85	10.18	5.22	5.11	2.01
+ drop-path + dropout	28.20	23.73	23.36	7.33	4.60	4.59	1.87
↳ deepest column alone	29.05	24.32	23.60	7.27	4.68	4.63	1.89
FractalNet (40 layers, 22.9M params)[2]	-	22.49	21.49	-	5.24	5.21	-

Implementation and Results



- Evolution of loss for plain networks
- with mixed drop-path, monitoring its loss as well as the losses of its four subnetworks corresponding to individual columns of the same depth as the plain networks.
- As the 20-layer subnetwork starts to stabilize, drop-path puts pressure on the 40-layer column to adapt, with the rest of the network as its teacher.