WEDNESDAY
FEB '18
7
038-327 WK-06

[Original RESNET Paper]
Deep Residual Learning for Image Recognition

How Deep should we make our Neural Network?

- It Depends on

a) Complexity of task at hand

b) Available computational capacity in the time of training

c) Available computational capacity in the time of Inference

- If the task needs lot of Parameters

a) Can we train very deep networks efficiently using current optimization solver

b) Is training a better model as simple as adding more & more layers?, Answer is no.

February 2018

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28											

The four things that have spoiled your life are: "I", "Mine", "You" and "Yours". Forget them.

THURSDAY
FEB '18
8
039-326 WK-06

Why is it not OK to just add more and more layers?

- Because it causes vanishing gradient problem
- Degradation Problem

Degradation Problem:

- Adding more non-linear layer (adding more nodes) lead to drop in accuracy

- Our current optimization solvers are not able to approximate the identity mapping of a stack of added non-linear layers

Residual Learning

H(x) is true mapping function we want to learn.

$$F(x) = H(x) - x$$

Residual block adds explicitly identity connections throughout the network to help learning the required identity Mappings

March 2018

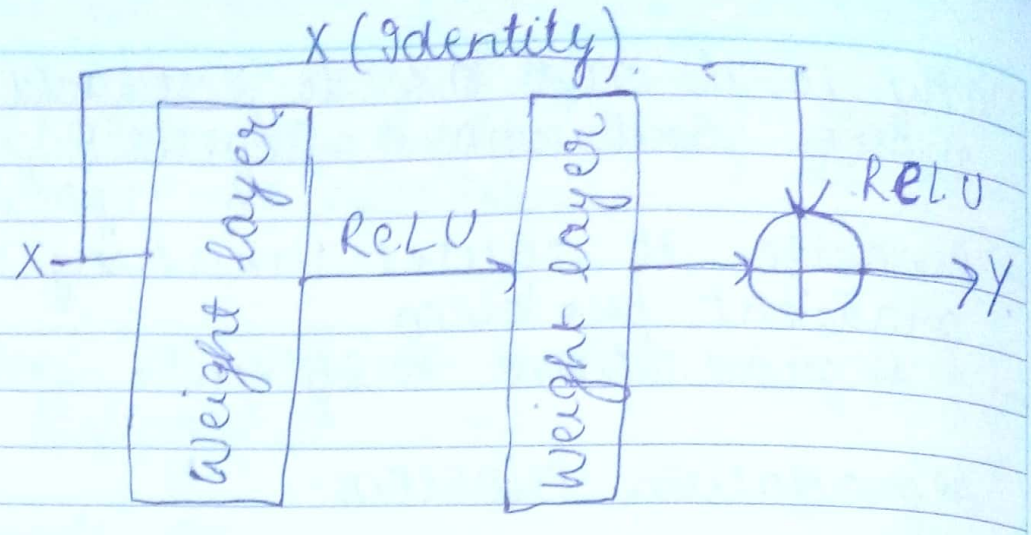
Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28	29	30	31								

Opportunities are usually disguised as hard work, so most people don't recognize them.

08-02

MARCH
APRIL
MAY
JUNE

FRIDAY
FEB '18
9
040-325 WK-06



While moving from one kind of layer to different kind of layer the connection is shown by dotted connection which is linear projection for dimension Mapping.

$$V = F(x\{W_i\}) + W_s X$$

for making two different layer dimension of same dimension, we do padding. We can also use Matrix multiplication in order to project dimension of next layer to previous layer, which add more ~~Data~~ parameters to the data.

Batchnormalization - At end of each layer, we need to normalize the output because is in process of learning, bias and weights, the gradient can go to huge value or

February 2018

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28											

The best argument against democracy is a five-minute conversation with the average voter.

go to zero so no changes happen.

Key takeaways

- Residual Network eases the training process
- Addressing ~~the~~ 'Degradation problem' in training process
- leveraging deeper representations of neural network for image recognition tasks.
- Training neural network with more than 1200 layers.

Other than tackling the vanishing gradient problems, the architecture of ResNet encourage feature reuse making network highly parameter efficient.

We know l_th layer will have $K*(l-1) + K\_0$ feature maps where K_0 is number of channels in the Input Image, K is growth rate

SUNDAY
FEB '18
11
042-323 WK-06

1x1 convolutional bottleck layer is also used to reduce the number of feature maps before the expensive 3x3 convolution

~~One major drawback of ResNet is the of deeper~~

One major drawback of training deep neural Network is that it requires weeks for training, making it practically infeasable, hence counter-Intuitive method of Randomly dropping layers during training and using the full network in testing. In training time, each layer has "survival probability" and is Randomly dropped. In testing time, all blocks are kept active and re-calibrated according to its survival probability during training

H_l be the output of l_th residual block, f_l be the mapping defined by the l_th block's weighted Mapping, ~~b~~ b be a Bernoulli random variable ~~(0 or 1)~~ (0 or 1, Indicating whether block is active or not) during training

February 2018

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28											

Truly great people usually don't feel important; they make others feel important.

11-02

$H_\ell = ReLU(b_\ell * f_\ell(H_{\ell-1}) + id(H_{\ell-1}))$

$b_\ell = 1$, the block is normal residual block

$b_\ell = 0$, the above formula becomes

$$H_\ell = ReLU(id(H_{\ell-1}))$$

Let p_ℓ be survival probability of layer ℓ during training, during test we have.

$$H_\ell = ReLU(p_\ell * f_\ell(H_{\ell-1}) + id(H_{\ell-1}))$$

Since earlier layer use extract low level feature, it should not be dropped too frequently, hence resulting rule becomes.

$$p_\ell = 1 - \frac{\ell}{L}(1 - p_L)$$

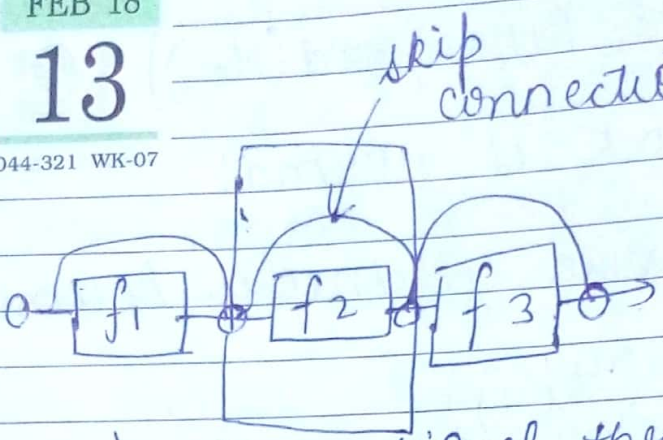
L - denotes total number of blocks

In Dropout, it drops part of hidden units in one layer during training whereas in ResNet, it drops entire layer

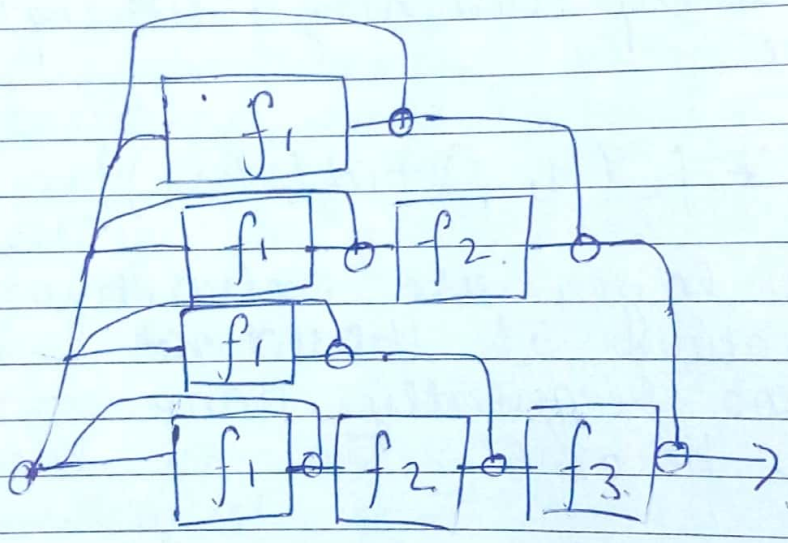
March 2018

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28	29	30	31								

TUESDAY
FEB '18
13
044-321 WK-07



a) conventional three block residual network



b) unravelled view of (a)

It is quite clear removing a couple of layers in ResNet architecture doesn't compromise its performance too much, the architecture had many independent effective paths

February 2018

Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27	28											

and majority of them remain intact after we remove couple of layers.

To investigate the relationship between path length and magnitude of gradient flowing through it. To get the magnitude of gradients in path of length k, the authors first fed a batch of data to the network and randomly sample k residual blocks. When back propogating the gradients they propagated through weight layer only for the sampled residual blocks — It shows that magnitude of gradients decreased rapidly as the path becomes longer, Hence ResNet did not solve vanishing gradient problem for very long paths, and ResNet actually enables training very deep network by shortening its effective paths.