Review: ZFNet — Winner of ILSVRC 2013 (Image Classification)

Sik-Ho Tsang
Aug 19, 2018 · 7 min read

In this story, ZFNet [1] is reviewed. ZFNet is a kind of winner of the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2013, which is an image classification competition, which has significantly improvement over AlexNet [2], the winner of ILSVRC 2012.

Task 2: Classification

Legend:
Dark grey background = outside training data

Team name	Comment	Error
Clarifai	Multiple models trained on the original data plus an additional model trained on 5000 categories.	0.11197
Clarifai	Multiple models trained on the original data plus an additional model trained on other 1000 category data.	0.11537
Clarifai	Average of multiple models on original training data.	0.11743
Clarifai	Another attempt at multiple models on original training data.	0.1215
Clarifai	Single model trained on original data.	0.12535
NUS	adaptive non-parametric rectification of all outputs from CNNs and refined PASCAL VOC12 winning solution, with further retraining on the validation set.	0.12953
NUS	adaptive non-parametric rectification of all outputs from CNNs and refined PASCAL VOC12 winning solution.	0.13303
ZF	5 models (4 different architectrues) trained on original data.	0.13511
Andrew Howard	This is an ensemble of convolutional neural networks combining multiple transformations for training and testing and models operating at different resolutions.	0.13555
Andrew Howard	This method explores re weighting the predictions from different data transformation and ensemble members in the previous submission.	0.13564
ZF	5 models trained on original data, 1 big.	0.13748

ILSVRC Ranking

Some people/articles think that ZFNet is not the winner, this conclusion maybe come from the ranking of ILSVRC, which as shown above. However, Clarifai is the company founded by the author of ZFNet, Zeiler. In addition, according to ImageNet Large Scale Visual Recognition Challenge, it mentioned:

> *"There were 24 teams participating in the ILSVRC2013 competition, compared to 21 in the previous three years combined. Following the success of the deep learning-based method in 2012, the vast majority of entries in 2013 used deep convolutional neural*

> *networks in their submission.* ***The winner of the classification task was Clarifai****, with several large deep convolutional networks averaged together.* ***The network architectures were chosen using the visualization technique of (Zeiler and Fergus, 2013)****,…”*

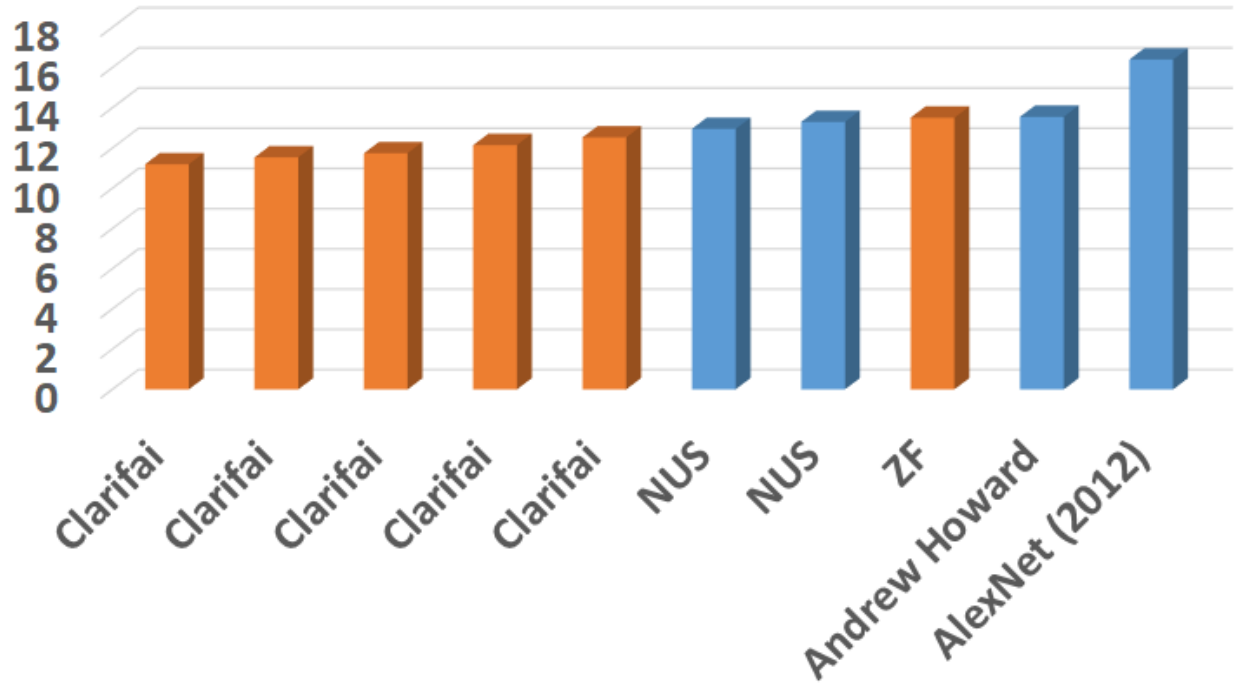
The reference (Zeiler and Fergus, 2013) cited as in the above passage is ZFNet. Thus, it is officially announced that ZFNet is the winner!

This is a **2014 ECCV** paper with more than **4000 citations** when I was writing this story. This is an important paper which teaches us to visualize the CNN kernels in deep layers. (Sik-Ho Tsang @ Medium)

ImageNet, is a dataset of over 15 millions labeled high-resolution images with around 22,000 categories. ILSVRC uses a subset of ImageNet of around 1000 images in each of 1000 categories. In all, there are roughly 1.3 million training images, 50,000 validation images and 100,000 testing images.

15 millions of images

Some Facts about Ranking



ILSVRC2013 Ranking [3]

In 2013, ZFNet was invented by Dr. Rob Fergus and his PhD student at that moment, Dr. Matthew D. Zeiler in NYU. (Prof. Yann LeCun, the inventor of LeNet is also from NYU. Hence, they also thanks Prof. LeCun for discussions at the acknowledgement in the paper.) That's why it is called ZFNet, based on their surname, Zeiler and Fergus, with the paper in 2014 ECCV, called "**Visualizing and Understanding Convolutional Networks**" [1]. Strictly speaking, ZFNet actually is not the winner of ILSVLC 2013. Instead, Clarifai, which was a new start-up company at that moment, is the winner of ILSVLC 2013 for image classification. And, Zeiler is also the founder and CEO of Clarifai.

As in the figure above, **ZFNet has significantly improved the image classification error rate compared with AlexNet [2], the winner in ILSVRC 2012.** And Clarifai has only small improvement over ZFNet. (For more details about the ranking, please go to [3].) Nevertheless, when we are talking about the deep learning network of the winner of ILSVLC 2013, we usually talk about ZFNet [1].

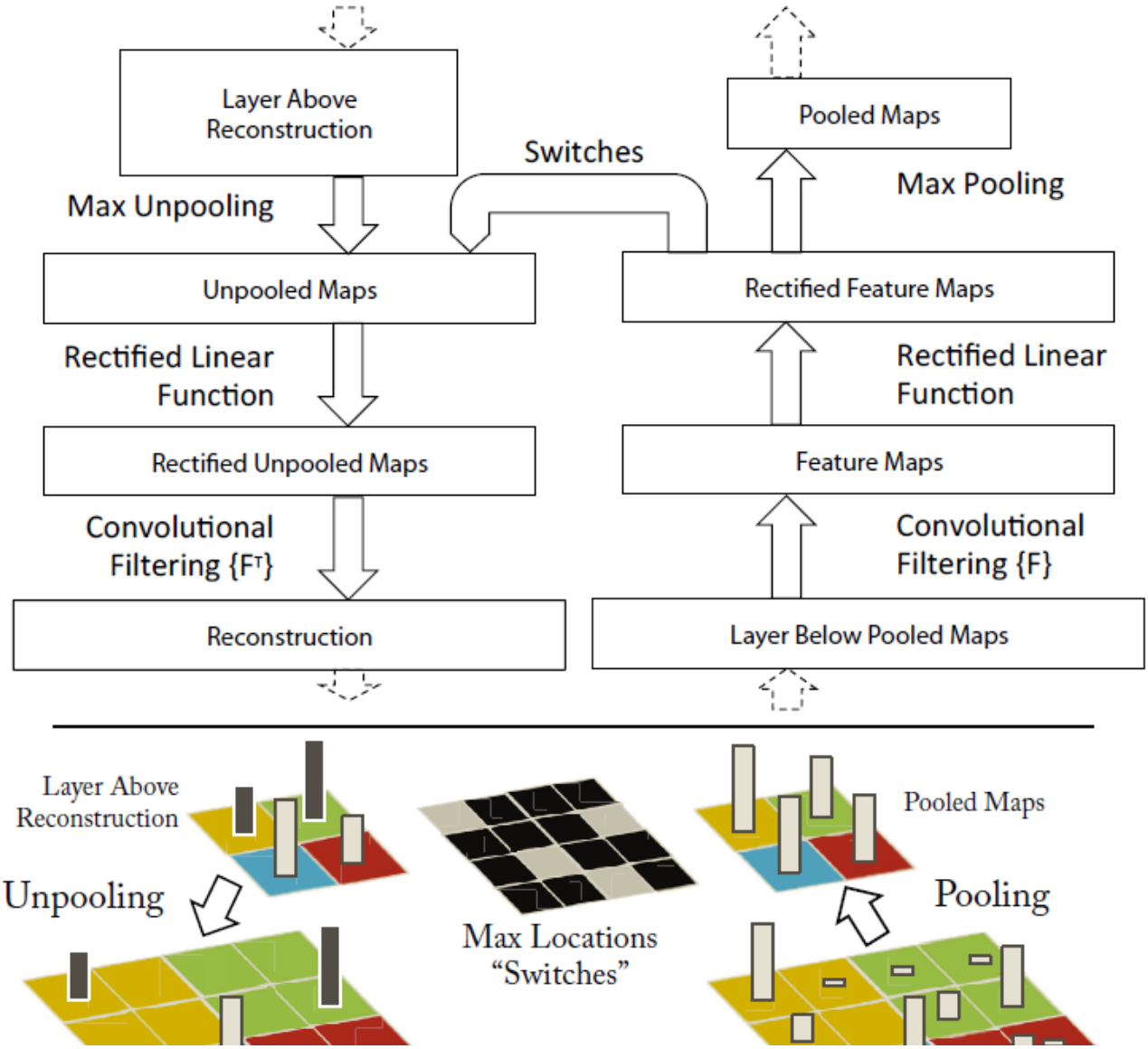
What We'll Cover

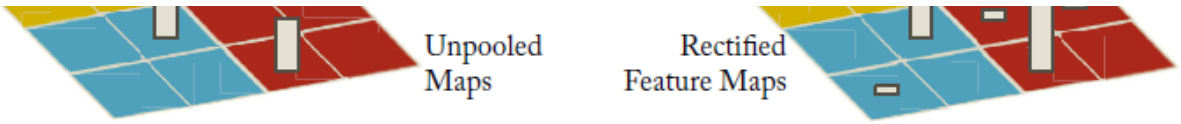
How and why convolutional networks can perform so well is always a mystery. Most of the time, we can only reason by intuitive explanation or empirical experiment. In this story, I will cover how ZFNet visualizes the convolutional network. By visualizing

the convolutional network, ZFNet become the Winner of ILSVLC 2013 in image classification by fine-tuning the AlexNet invented in 2012. Hence, the sections to be covered:

1. **Deconvnet Techniques for Visualization**
2. **Visualization for Each Layer**
3. **Modifications of AlexNet Based on Visualization Results**
4. **Experimental Results**
5. **Conclusions**

1. Deconvnet Techniques for Visualization

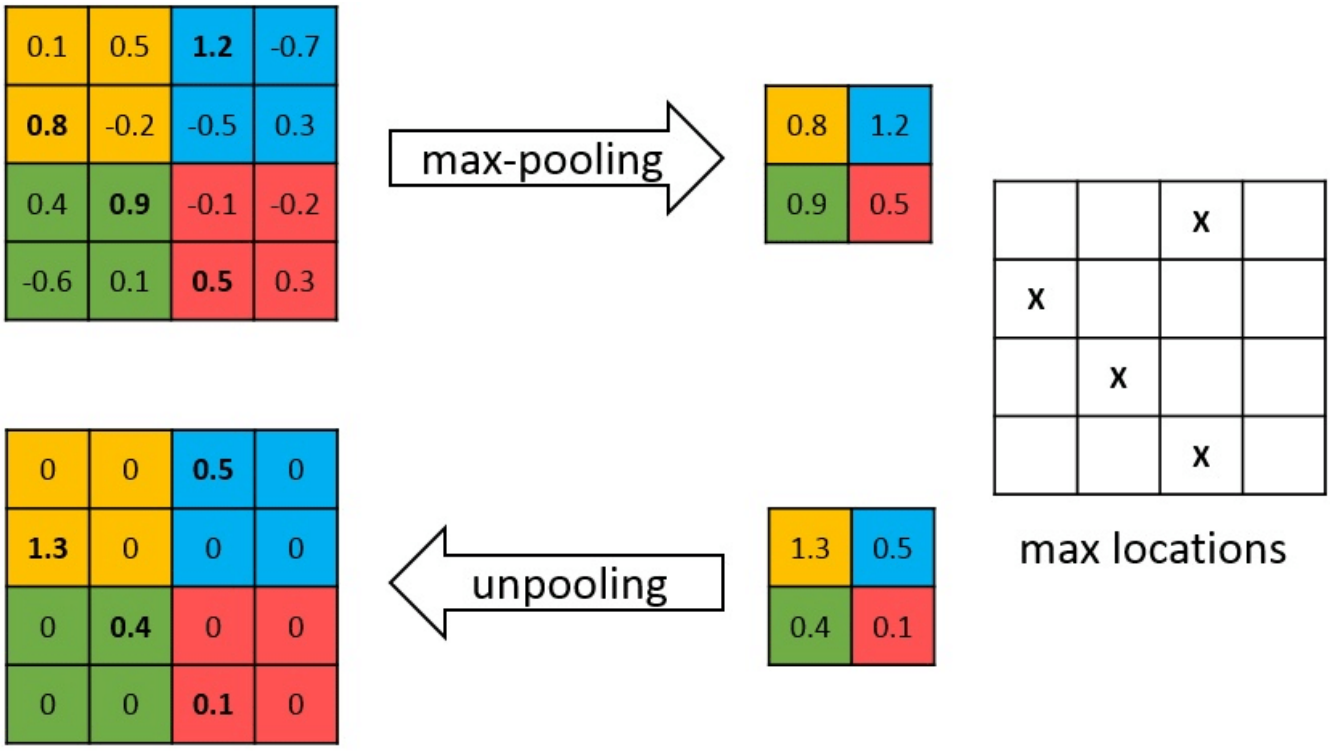




The Process to Deconv a Deep Layer

As we should know, a standard step in deep learning framework is to have a series of **Conv > Rectification (Activation Function) > Pooling**. To visualize a deep layer feature, we need a set of decovnet techniques to reverse the above actions such that we can visualize the feature in pixel domain.

1.1. Unpooling



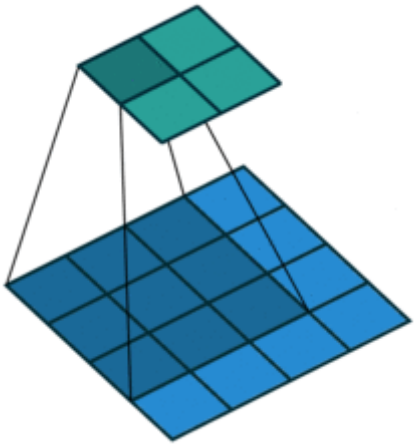
Unpooling

Max pooling operation is non-invertible, however we can obtain an approximate inverse by recording the locations of the maxima within each pooling region, as in the figure above.

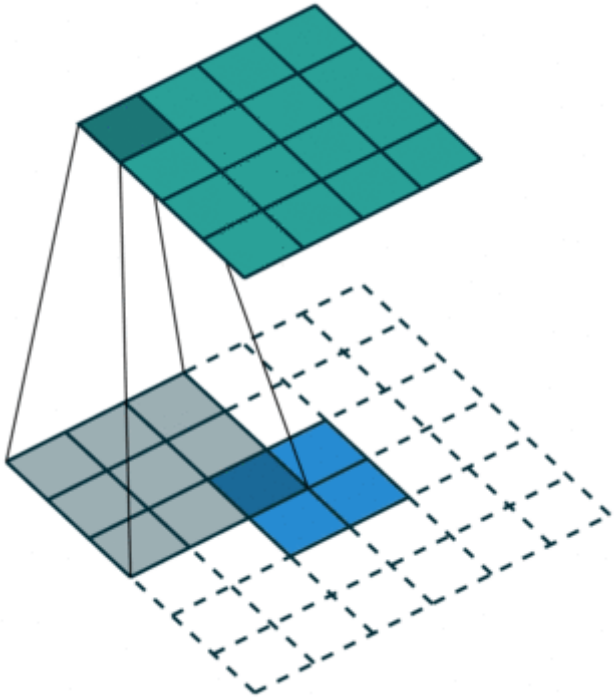
1.2. Rectification (Activation Function)

Since ReLU is used as the activation function, and ReLU is to keep all values positive while make negative values become zero. In the reverse operation, we just need to perform ReLU again.

1.3. Deconv



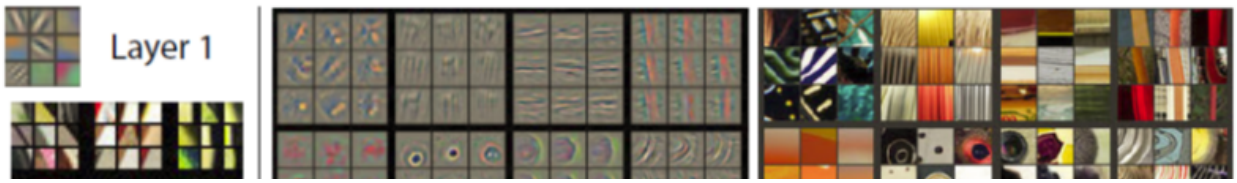
Conv (Blue is input, cyan is output)

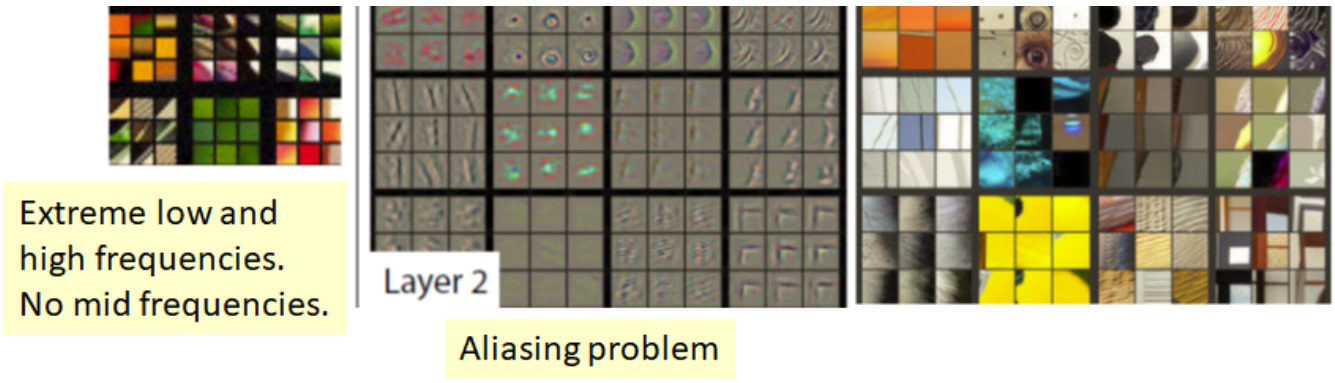


Deconv (Blue is input, cyan is output)

To do the deconv operation, indeed, it is a transposed version of conv.

2. Visualization for Each Layer



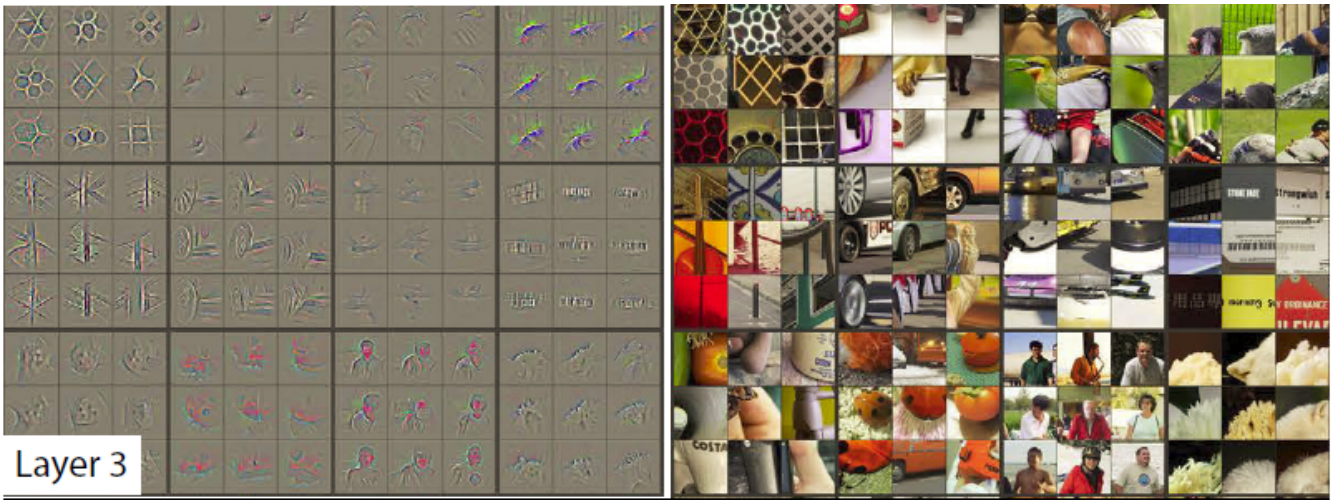


Layer 1 and Layer 2

By using deconv techniques, the top 9 activated patterns in randomly selected feature maps are shown for each layer. And **two problems are observed in layer 1 and layer 2**.

(i) Filters at layer 1 are a mix of extremely high and low frequency information, with little coverage of the mid frequencies. Without the mid frequencies, there is a chain effect that deep features can only learn from extremely high and low frequency information.

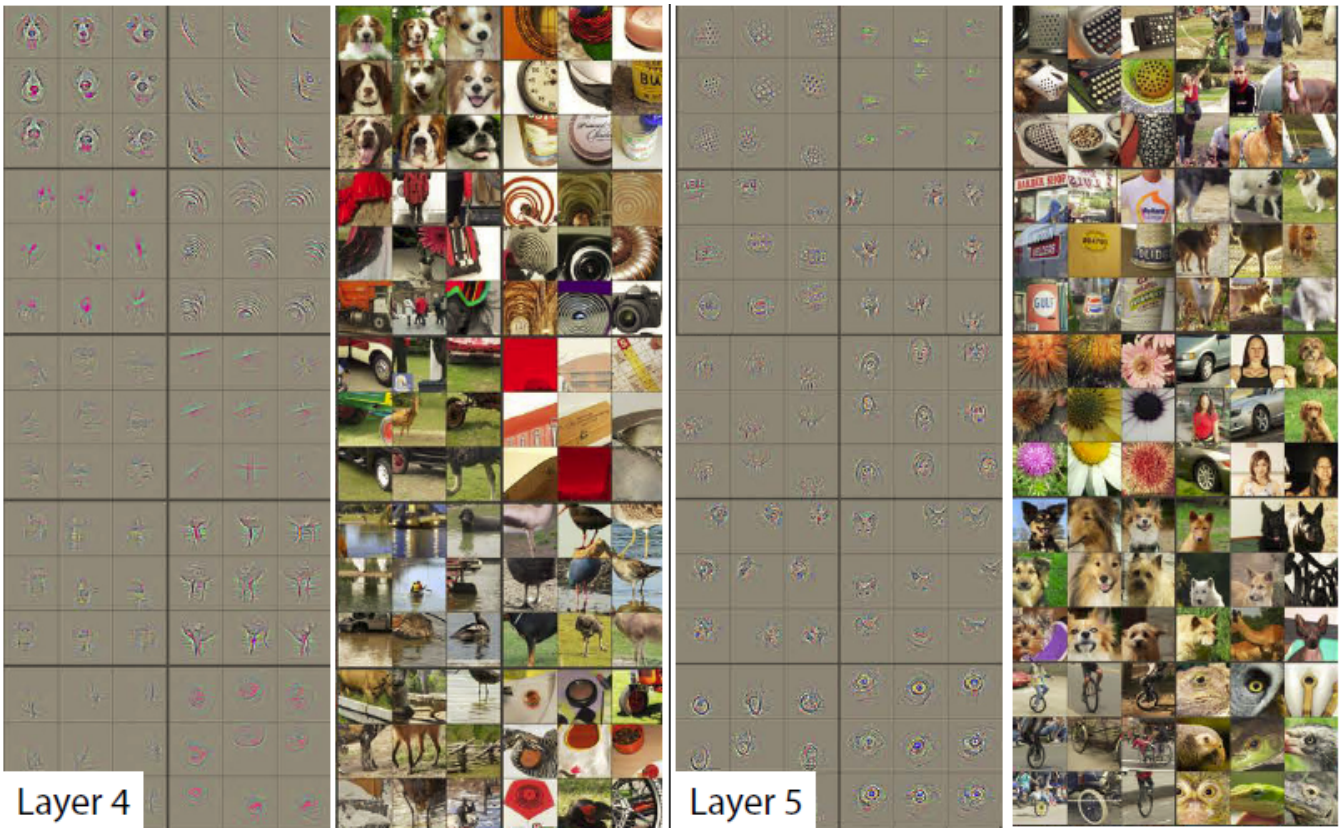
(ii) Layer 2 shows aliasing artifacts caused by the large stride 4 used in the 1st layer convolutions. **Aliasing occurs when sampling frequency is too low.**



Layer 3

Let us observe 3 more layers.

Layer 3 starts to learn some general patterns, such as mesh patterns, and text pattern.

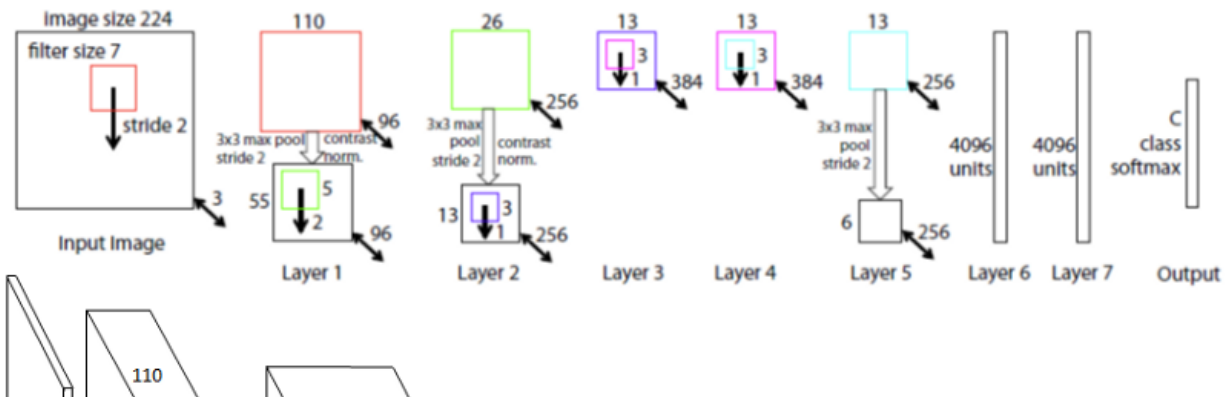


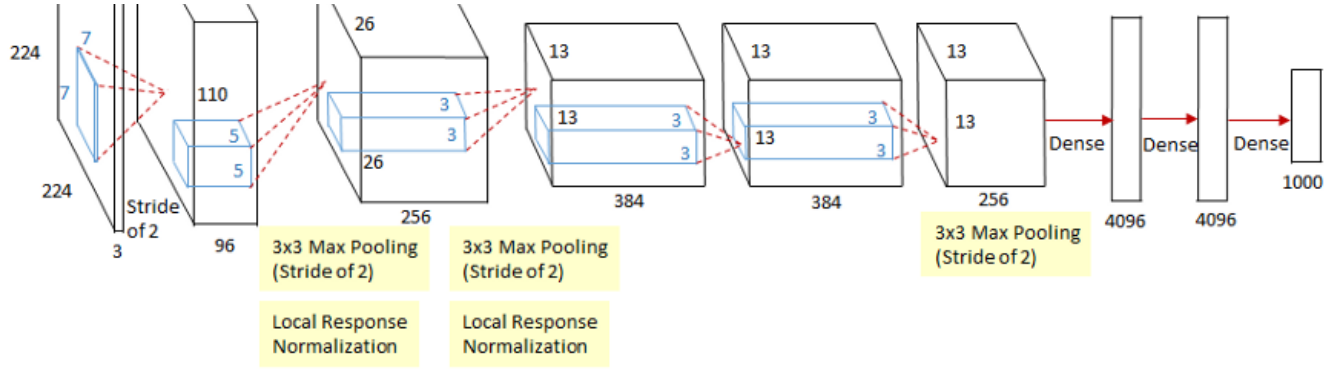
Layer 4 and Layer 5

Layer 4 shows significant variation, and is more class-specific, such as dogs' faces and birds' legs.

Layer 5 shows entire objects with significant pose variation, such as keyboards and dogs.

3. Modifications of AlexNet Based on Visualization Results



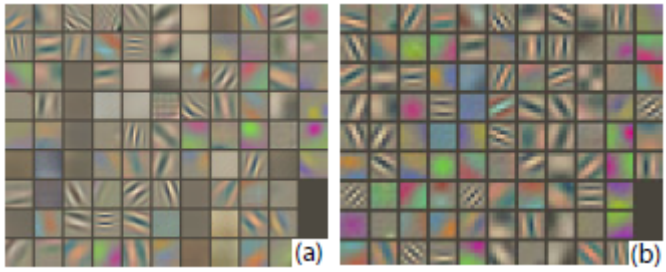


ZFNet

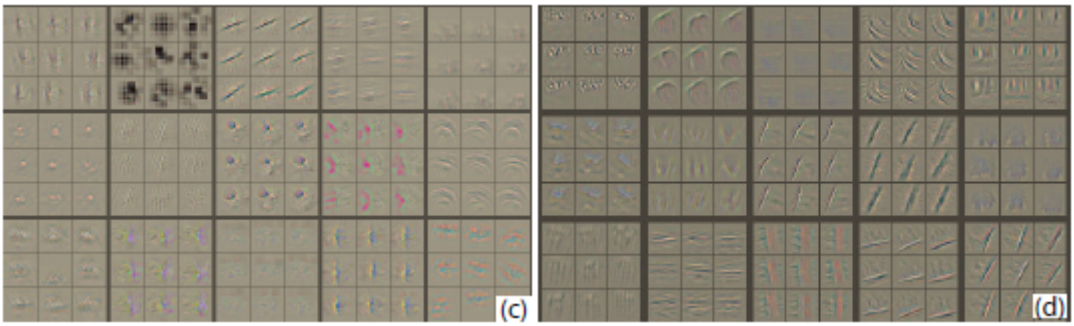
ZFNet is redrawn as the same style of AlexNet for the ease of comparison. To solve the two problems observed in layer 1 and layer 2, ZFNet makes two changes. (To read the AlexNet review, please visit [4].)

(i) Reduced the 1st layer filter size from 11x11 to 7x7.

(ii) Made the 1st layer stride of the convolution 2, rather than 4.



Layer 1: (a) More mid-frequencies in ZFNet, (b) Extremely low and high frequencies in AlexNet



Layer 2: (c) Aliasing artifacts in AlexNet and (d) much cleaner features in ZFNet

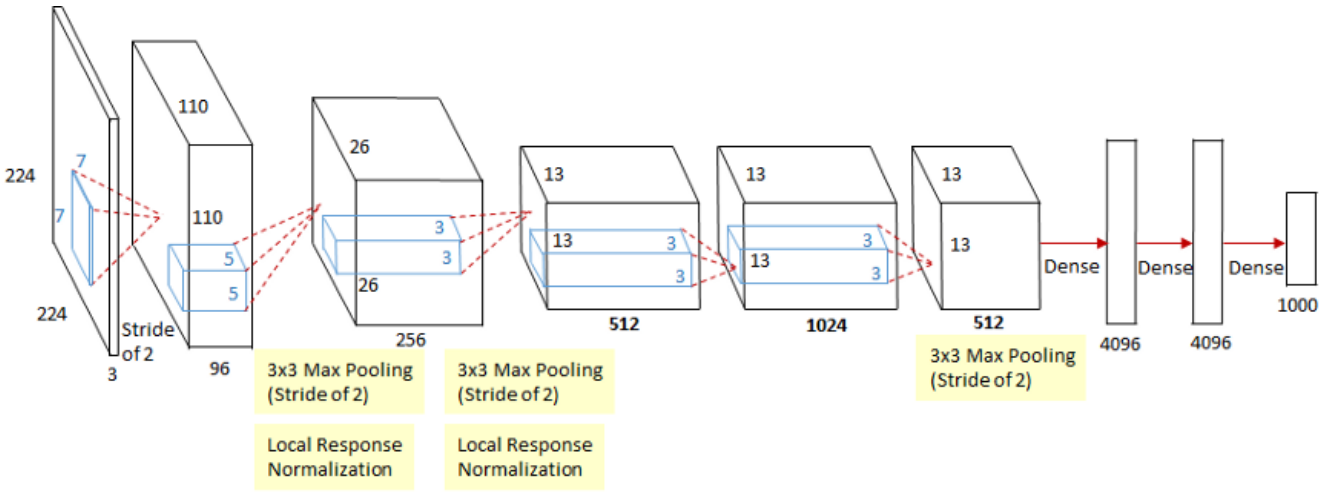
4. Experimental Results

4.1. The Modified ZFNet based on Ablation Study

Error %	Train Top-1	Val Top-1	Val Top-5
Our replication of Krizhevsky *et al.* [18], 1 convnet	35.1	40.5	18.1
Removed layers 3,4	41.8	45.4	22.1
Removed layer 7	27.4	40.0	18.4
Removed layers 6,7	27.4	44.8	22.4
Removed layer 3,4,6,7	71.1	71.3	50.1
Adjust layers 6,7: 2048 units	40.3	41.7	18.8
Adjust layers 6,7: 8192 units	26.8	40.0	18.1
Our Model (as per Fig. 3)	33.1	38.4	16.5
Adjust layers 6,7: 2048 units	38.2	40.2	17.6
Adjust layers 6,7: 8192 units	22.0	38.8	17.0
Adjust layers 3,4,5: 512,1024,512 maps	18.8	**37.5**	**16.0**
Adjust layers 6,7: 8192 units and Layers 3,4,5: 512,1024,512 maps	**10.0**	38.3	16.9

The Modified ZFNet

Ablation Study



The Modified ZFNet based on Ablation Study

There are also ablation study on removing or adjusting layers. **The modified ZFNet can obtain 16.0% on top-5 validation error.**

4.2. Comparison with State-or-the-art Approaches

Error %	Val Top-1	Val Top-5	Test Top-5
Gunji *et al.* [12]	-	-	26.2
DeCAF [7]	-	-	19.2
Krizhevsky *et al.* [18], 1 convnet	40.7	18.2	−−
Krizhevsky *et al.* [18], 5 convnets	38.1	16.4	16.4

Krizhevsky *et al.* *[18], 1 convnets	39.0	16.6	– –
Krizhevsky *et al.* *[18], 7 convnets	36.7	15.4	15.3
Our replication of Krizhevsky *et al.* , 1 convnet	40.5	18.1	– –
1 convnet as per Fig. 3	38.4	16.5	– –
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	36.0	14.7	14.8
Howard [15]	-	-	13.5
Clarifai [28]	-	-	11.7

1 AlexNet by ZF

1 ZFNet

1 Modified ZFNet

5 ZFNet + 1 Modified ZFNet

Error Rate (%)

By using AlexNet, top-5 validation error rate is 18.1%.

By using ZFNet, top-5 validation error rate is 16.5%. We can conclude that the modifications based on the visualization is essential.

By using 5 ZFNet from (a) and 1 modified ZFNet from (b), top-5 validation error rate is 14.7%. This is again a kind of boosting technique which already used in LeNet and AlexNet. (Please visit [5] and [4] for more about the boosting technique.)

4.3. Other relatively small datasets are also tested

# Train	Acc % 15/class	Acc % 30/class
Bo *et al.* [3]	–	81.4 ± 0.33
Yang *et al.* [17]	73.2	84.3
Non-pretrained convnet	22.8 ± 1.5	46.5 ± 1.7
ImageNet-pretrained convnet	$\mathbf{83.8 \pm 0.5}$	$\mathbf{86.5 \pm 0.5}$

Pre-trained ZFNet

Caltech 101 (83.8 to 86.5 mean accuracy)

# Train	Acc % 15/class	Acc % 30/class	Acc % 45/class	Acc % 60/class
Sohn *et al.* [24]	35.1	42.1	45.7	47.9
Bo *et al.* [3]	40.5 ± 0.4	48.0 ± 0.2	51.9 ± 0.2	55.2 ± 0.3
Non-pretr.	9.0 ± 1.4	22.5 ± 0.7	31.2 ± 0.5	38.8 ± 1.4

ImageNet-pretr.	65.7 ± 0.2	70.6 ± 0.2	72.7 ± 0.4	74.2 ± 0.3

Pre-trained ZFNet

Caltech 256 (65.7 to 74.2 mean accuracy)

Acc %	[22]	[27]	[21]	Ours	Acc %	[22]	[27]	[21]	Ours
Airplane	92.0	**97.3**	94.6	96.0	Dining table	63.2	**77.8**	69.0	67.7
Bicycle	74.2	**84.2**	82.9	77.1	Dog	68.9	83.0	**92.1**	87.8
Bird	73.0	80.8	88.2	**88.4**	Horse	78.2	87.5	**93.4**	86.0
Boat	77.5	85.3	60.3	**85.5**	Motorbike	81.0	**90.1**	88.6	85.1
Bottle	54.3	**60.8**	60.3	55.8	Person	91.6	95.0	**96.1**	90.9
Bus	85.2	**89.9**	89.0	85.8	Potted plant	55.9	57.8	**64.3**	52.2
Car	81.9	**86.8**	84.4	78.6	Sheep	69.4	79.2	**86.6**	83.6
Cat	76.4	89.3	90.7	**91.2**	Sofa	65.4	**73.4**	62.3	61.1
Chair	65.2	**75.4**	72.1	65.0	Train	86.7	**94.5**	91.1	91.8
Cow	63.2	77.8	**86.8**	74.4	Tv	77.4	**80.7**	79.8	76.1
Mean	74.3	82.2	**82.8**	79.0	# won	0	**11**	6	3

Pre-trained ZFNet

PASCAL 2012 (79.0 mean accuracy)

From the above tables, we can see that, the accuracy, without pre-training of ZFNet using ImageNet images, i.e. train the ZFNet from the scratch, is low. **With the training (fine-tuning) on top of the pre-trained ZFNet, the accuracy is much high. That means the trained filters are generalized to different images, not just for images for ImageNet.**

Particularly for Caltech 101 and Caltech 256 datasets, ZFNet has overwhelming results.

For PASCAL 2012, the PASCAL images can contain multiple objects and quite different from nature compared with those in ImageNet. Thus, the accuracy is a bit lower but still competitive with state-of-the-art approaches.

5. Conclusions

While only shallow layer features can be observed previously, this paper provides an interesting approach to observe deep features in pixel domain.

By visualizing the convolutional network layer by layer, ZFNet adjusts the layer hyperparameters such as filter size or stride of the AlexNet and successfully reduces the error rates.

It is important to know more about the state-of-the-art approaches in order to understand more about the deep learning. I will write more stories.

Please stay tuned!!!

References

1. [2014 ECCV] [ZFNet]
   Visualizing and Understanding Convolutional Networks
2. [2012 NIPS] [AlexNet]
   ImageNet Classification with Deep Convolutional Neural Networks
3. ILSVRC 2013 Ranking
   http://www.image-net.org/challenges/LSVRC/2013/results.php#cls
4. Review of AlexNet, CaffeNet — Winner of ILSVRC 2012 (Image Classification)
5. Review of LeNet-1, LeNet-4, LeNet-5, Boosted LeNet-4 (Image Classification)