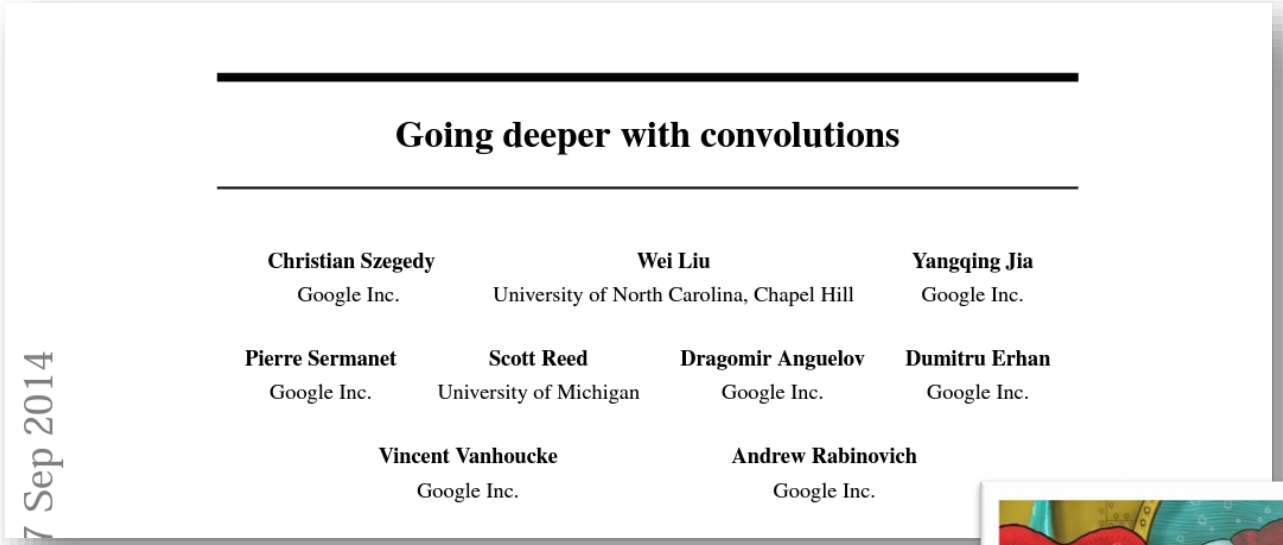
Going deeper with convolutions

2023.06.26

Youngjae Kim

1. Author & about the paper



Going deeper with convolutions

Christian Szegedy, Google Inc.
Wei Liu, University of North Carolina, Chapel Hill
Yangqing Jia, Google Inc.
Pierre Sermanet, Google Inc.
Scott Reed, University of Michigan
Dragomir Anguelov, Google Inc.
Dumitru Erhan, Google Inc.
Vincent Vanhoucke, Google Inc.
Andrew Rabinovich, Google Inc.

7 Sep 2014

- Computer Vision and Pattern Recognition(CVPR 2015)
- SOTA on ILSVRC 2014
- 7 out of 9 authors from Google Inc.

1. abstract

- Codename : Inception
- Deep-convolutional neural network architecture

- GoogLeNet
- ImageNet Large Scale Visual Recognition Challenge 2014 (2등 VGGNet)
  - SOTA on 1) classification 2) detection

- Main hallmark of this architecture?
  - Computing resource <-> Depth and Width

2. Introduction

1. Synergy of deep architecture + algorithm

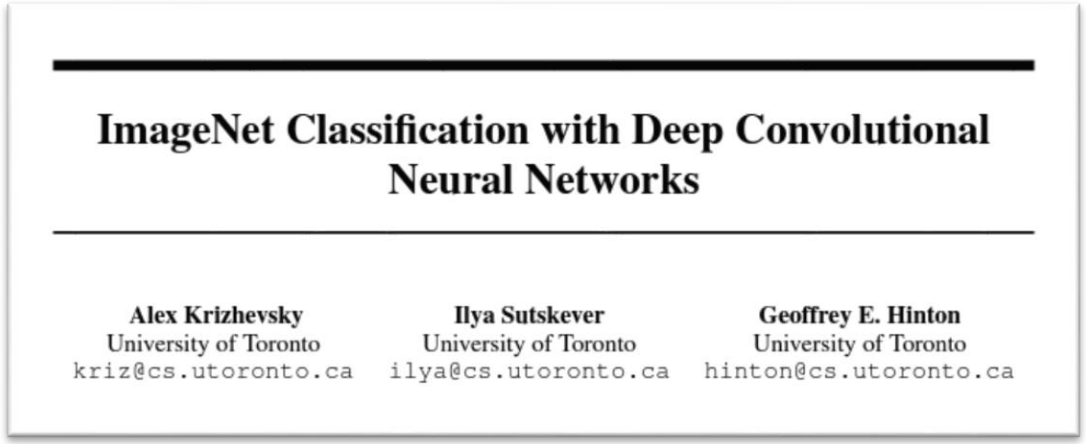
- ILSVRC 2014's top entries used ONLY classification dataset
  - Hardware & bigger datasets?
  - Algorithms & network architecture

2. Introduction

2. Computing efficiency

- mobile / embedded computing

- Used 12 x fewer parameters than ILSVRC 2012



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

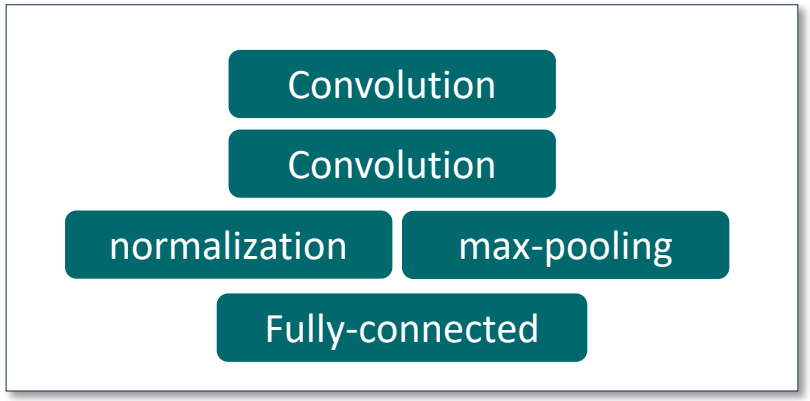
- Keep 1.5 billion multiply-adds at inference time

  - academic curiosity?

  - real world use

3. Related Work

- **LeNet-5 & CNN's standard structure**



- **Increase layer number / size + drop-out to solve overfitting**
- **Concerns that Max-pooling layers result in loss of spatial information**

  - localization / object detection / human pose estimation

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition.*

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.

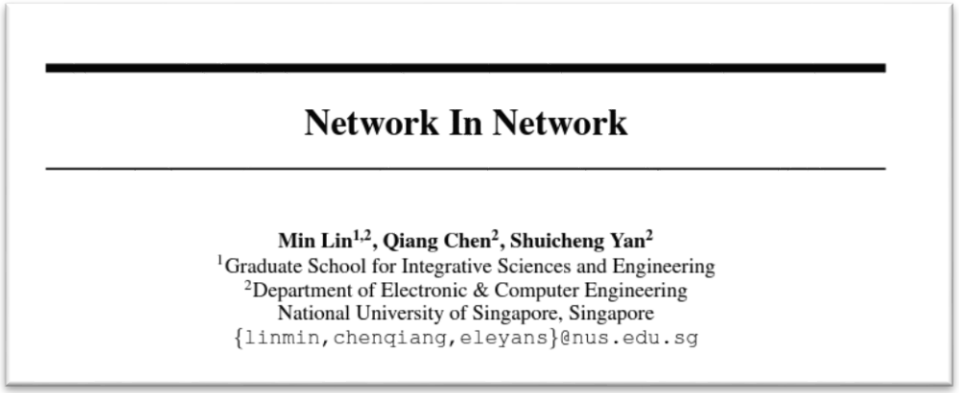
3. Related Work

[15] Thomas Serre, Lior Wolf, Stanley M. Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007.

- **Neuroscience model of the primate visual cortex**
  - **Fixed Gabor filters of different sizes to handle multiple scales**
- **Fixed 2-layer deep model <-> “learning” repeated inception layers (22 layers GoogLeNet model)**

3. Related Work

- Network-in-Network



Network In Network

Min Lin[1,2], Qiang Chen[2], Shuicheng Yan[2]
[1]Graduate School for Integrative Sciences and Engineering
[2]Department of Electronic & Computer Engineering
National University of Singapore, Singapore
{linmin,chenqiang,eleyans}@nus.edu.sg

- 1x1 convolutional layers followed by ReLu
  - Non-linearity = complicated pattern recognitions
  - as dimension reduction module (computational bottleneck)
  - increase in depth / width without performance penalty
- Use of average pooling at last layer (explained later)

3. Related Work

[6] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition. 2014. CVPR 2014. IEEE Conference on*. 2014.

- **R-CNN (Regions with Convolutional Neural Networks)**

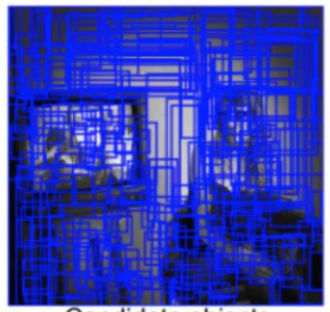
  - two stage approach

  - low-level cues + CNN classifier

  -> multi-box prediction + ensemble

Input Image

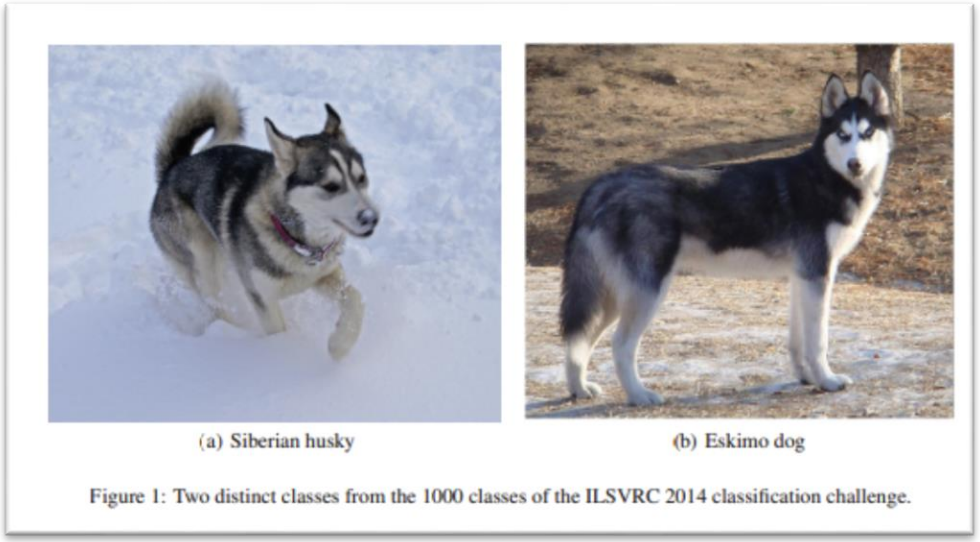
Segmentation



Candidate objects

4. Motivations and High level considerations

- Deep neural network performance == Larger size model (width + height)
  - overfitting -> more high quality data sets needed -> expensive cost
  - dramatically increased use of computational resource



(a) Siberian husky (b) Eskimo dog

Figure 1: Two distinct classes from the 1000 classes of the ILSVRC 2014 classification challenge.

4. Motivations and High level considerations

- Hebbian Principle – “neurons that fire together, wire together”

2] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. *CoRR*. abs/1310.6343. 2013.

- Making connection using correlations among features

- Early LeNet (Sparse) -> AlexNet(FC)
- Fully connected to sparsely connected architecture
  - inefficient (computationally)
  - improvements on fully connected related libraries

4. Motivations and High level considerations

- Inception model

-> making “sparsely connected architecture” with fully connected architecture

-> firm motivation for future work in this direction

5. Architectural Details

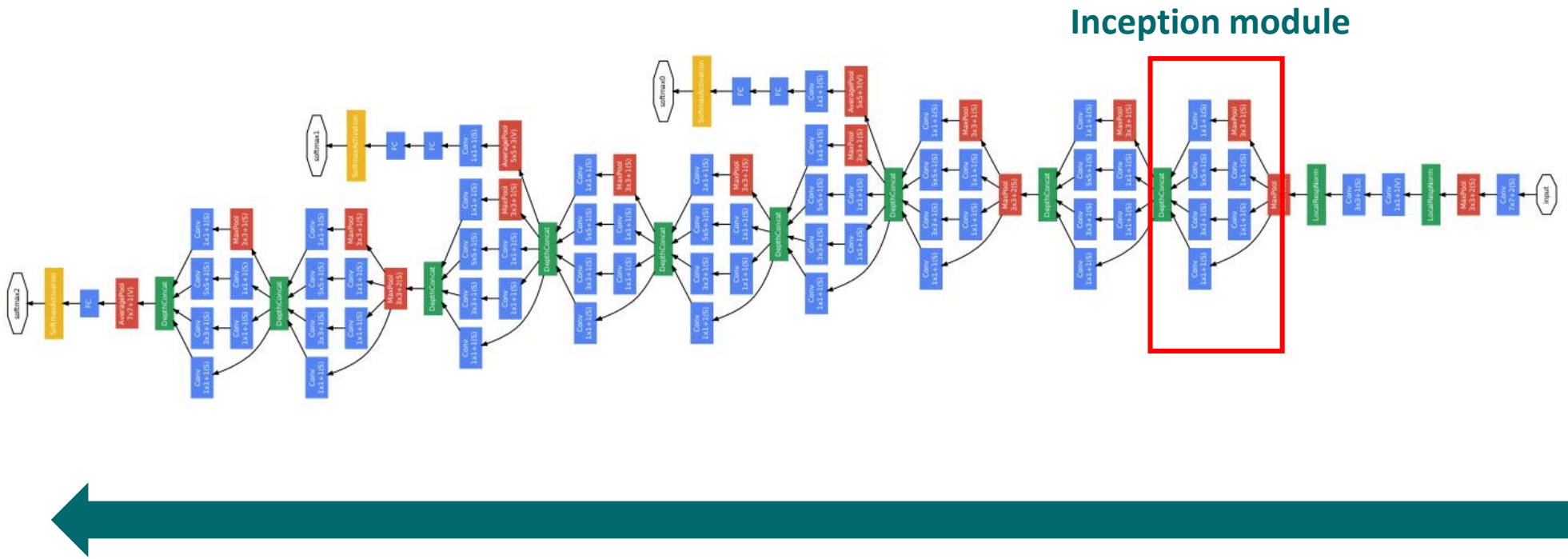
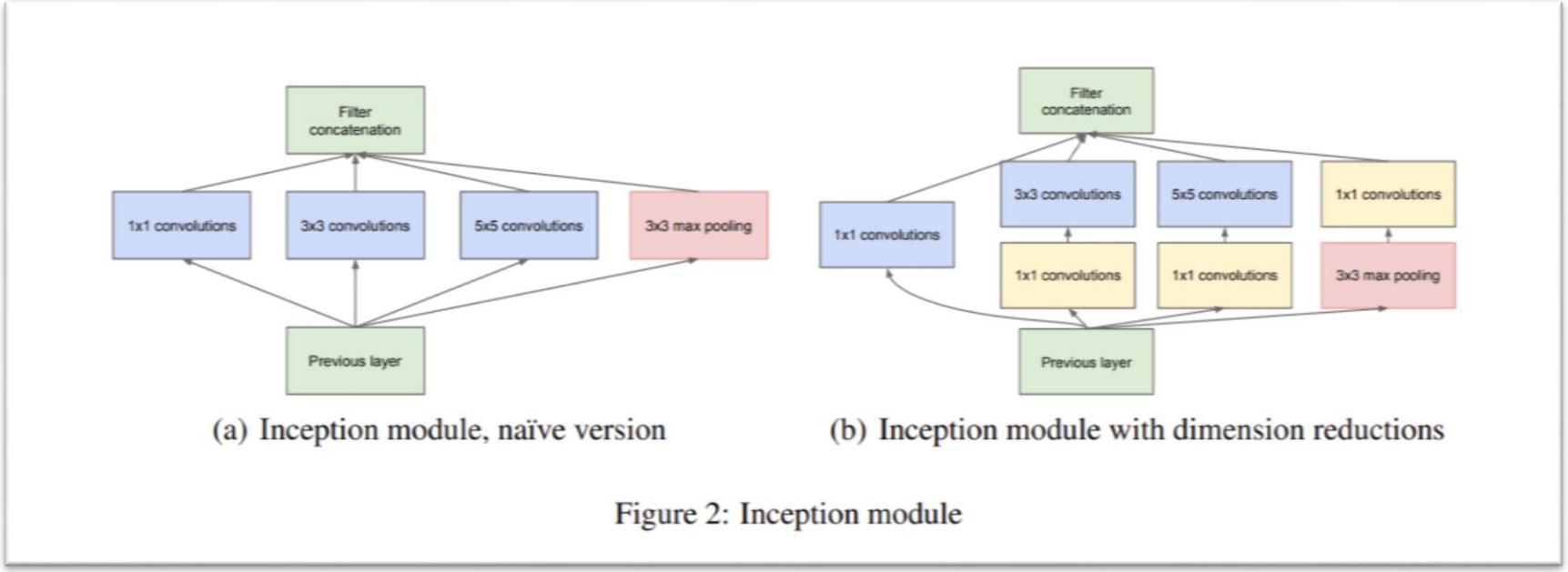


Figure 3: GoogLeNet network with all the bells and whistles

5. Architectural Details



(a) Inception module, naïve version (b) Inception module with dimension reductions

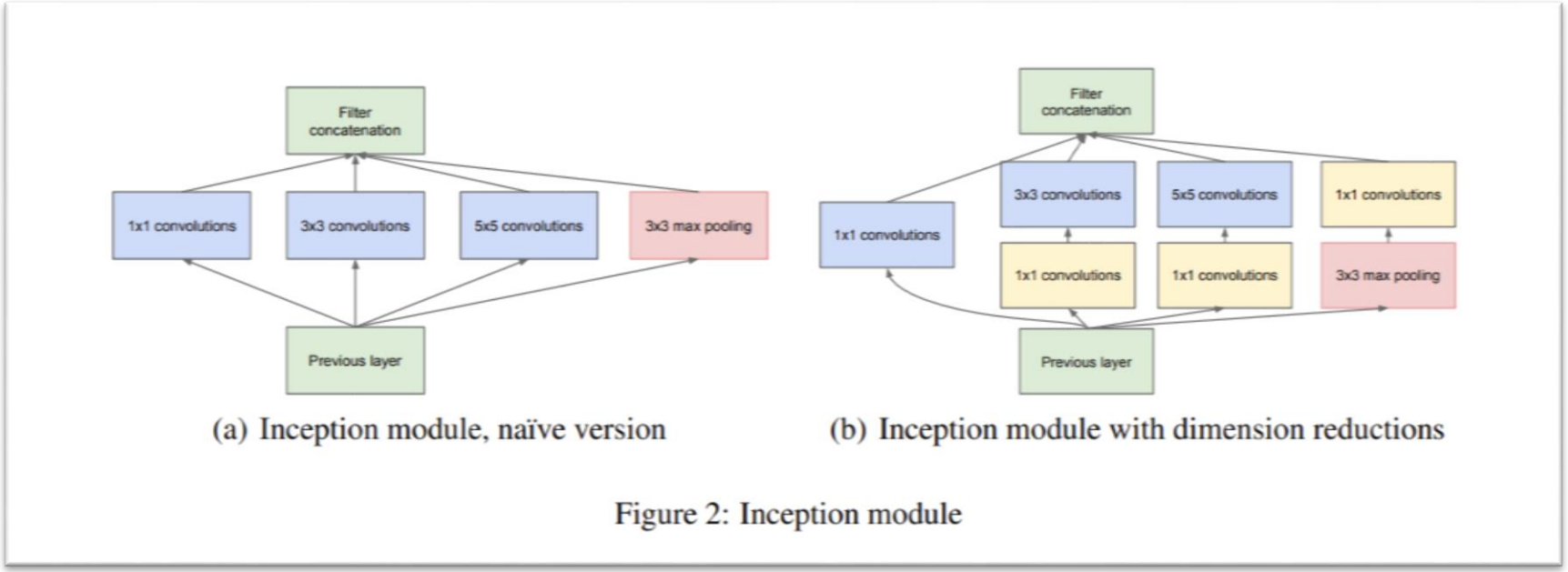
Figure 2: Inception module

- Using 3 filters parallel at one time -> concatenation
- 1 x 1, 3 x 3, 5 x 5 convolutions – Gabor filter

  - avoid patch-alignment issue

  - later GoogLeNet 5x5 -> (3x3) – (3x3) : less parameter & less overfitting

5. Architectural Details



(a) Inception module, naïve version (b) Inception module with dimension reductions

Figure 2: Inception module

- Max-pooling
- 1 x 1 convolutions with ReLu before 3x3 and 5x5 convolutions

5. Architectural Details

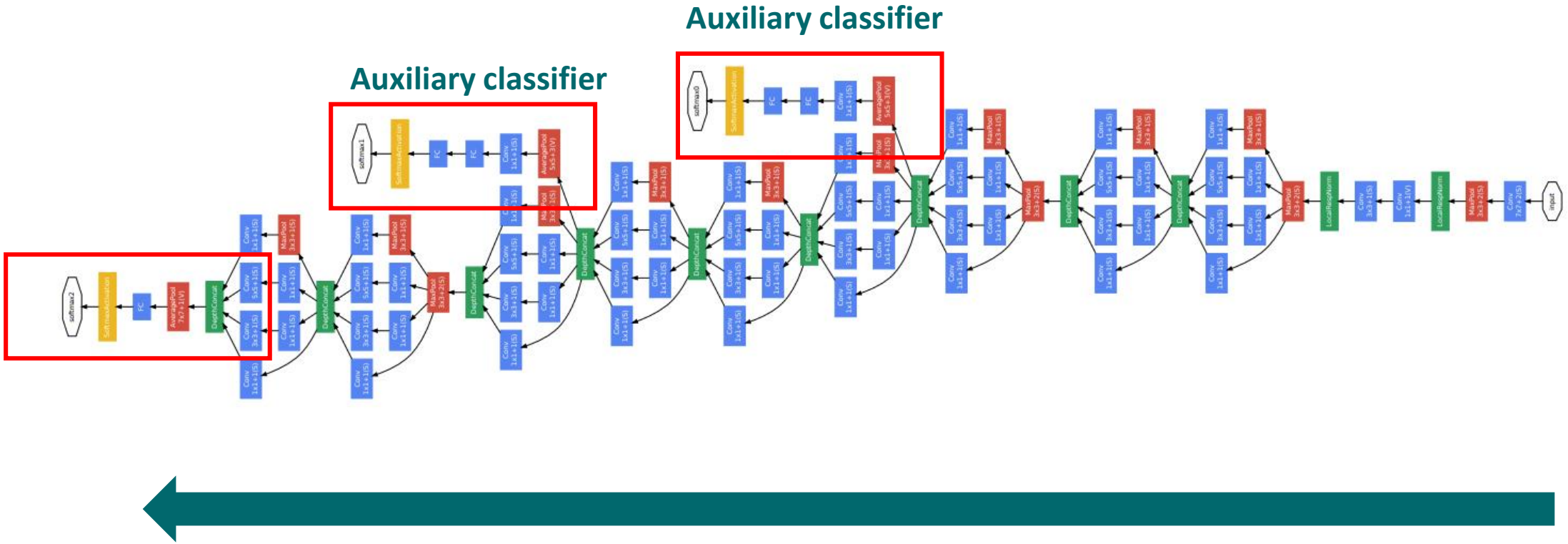


Figure 3: GoogLeNet network with all the bells and whistles

5. Architectural Details

- Auxiliary classifier
  - regularization to prevent overfitting
  - weighted by 0.3
- Average Pooling (Network in Network)
  - adapting and fine tuning for other label set
  - w/ wo/ FC
  - 0.6% improvement

5. Architectural Details

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Table 1: GoogLeNet incarnation of the Inception architecture

6. Results

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

Table 2: Classification performance

Number of models	Number of Crops	Cost	Top-5 error	compared to base
1	1	1	10.07%	base
1	10	10	9.15%	-0.92%
1	144	144	7.89%	-2.18%
7	1	7	8.09%	-1.98%
7	10	70	7.62%	-2.45%
7	144	1008	6.67%	-3.45%

Table 3: GoogLeNet classification performance break down

Team	Year	Place	mAP	external data	ensemble	approach
UvA-Euvision	2013	1st	22.6%	none	?	Fisher vectors
Deep Insight	2014	3rd	40.5%	ImageNet 1k	3	CNN
CUHK DeepID-Net	2014	2nd	40.7%	ImageNet 1k	?	CNN
GoogLeNet	2014	1st	43.9%	ImageNet 1k	6	CNN

Table 4: Detection performance

Team	mAP	Contextual model	Bounding box regression
Trimps-Soushen	31.6%	no	?
Berkeley Vision	34.5%	no	yes
UvA-Euvision	35.4%	?	?
CUHK DeepID-Net2	37.7%	no	?
GoogLeNet	38.02%	no	no
Deep Insight	40.2%	yes	yes

Table 5: Single model performance for detection

- **Resize to 256, 288, 320, 352 and crop**

7. Conclusion

- Solid evidence that approximating the expected optimal sparse structure by dense blocks -> improve neural networks for computer vision
- Efficient computational requirements with wider/deeper network