Evolving of Deep Neural Networks: Algorithm and Applications

Ming Yang
Co-founder & VP, Software

Horizon Robotics
Define the brain of things

NJSD 全球软件大会 2017
IT大咖说 知识分享平台

Who am I : a short bio

Co-founder & VP of Software

AI Research

Multimedia Analytics

PhD., Dept. of EECS

BE. And ME., Dept. of EE

Dr. Ming Yang is one of the founding member of the Facebook Artificial Intelligence Research (FAIR) and a former senior researcher at NEC Labs America. Dr. Yang is a well-recognized researcher in computer vision and machine learning. He co-authored 14 US patents, and over 20 publications in top conferences like CVPR and ICCV and 8 publications in the top international journal T-PAMI with more than 4750 citations.

During his tenure at Facebook, Dr. Yang led the deep learning research project “DeepFace”, which had a significant impact in the deep learning research community and got widely reported by various media including Science Magazine, MIT Tech Review and Forbes.

He received his B.Eng. and M.Eng. Degree from the Dept. of Electrical Engineering at Tsinghua University and Ph.D. degree from the Dept. of Electrical Engineering and Computer Science at Northwestern University.

DeepFace: THE largest-scale face recognition system

Science Magazine, Jan. 2015

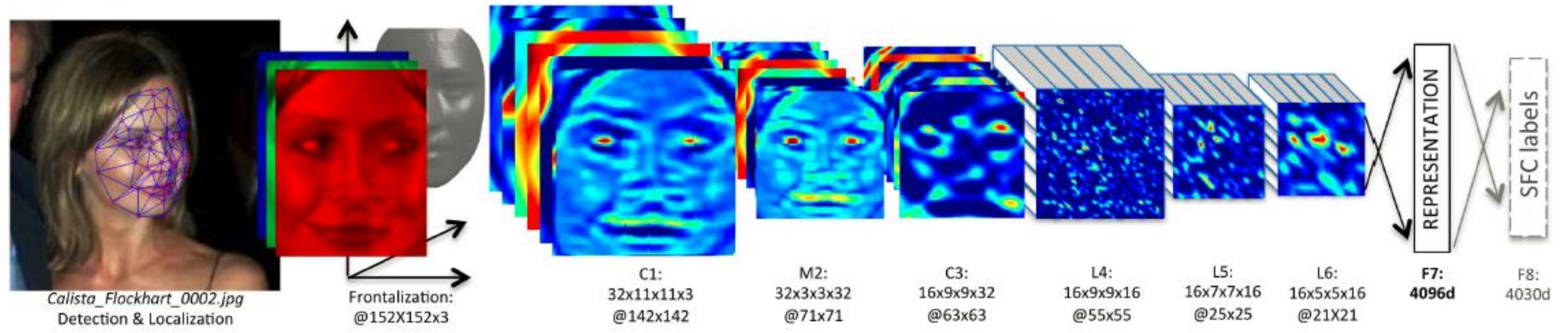


Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

SPECIAL SECTION THE END OF PRIVACY

UNMASKED

Facial recognition software could soon ID you in any photo *By* John Bohannon

Appear in a photo taken at a protest march, a gay bar, or an abortion clinic, and your friends might recognize you. But a machine probably won't—at least for now. Unless a computer has been tasked to look for you, has trained on dozens of photos of your face, and has high-quality images to examine, your anonymity is safe. Nor is it yet possible for a computer to scour the Internet and find you in random, uncaptioned photos. But within the walled garden of Facebook, which contains by far the largest collection of personal photographs in the world, the technology for doing all that is beginning to blossom.

Catapulting the California-based company beyond other corporate players in the field, Facebook's DeepFace system is now as accurate as a human being at a few constrained facial recognition tasks. The intention is not to invade the privacy of Facebook's more than 1.3 billion active users, insists Yann LeCun, a computer scientist at New York University in New York City who directs Facebook's artificial intelligence research, but rather to protect it. Once DeepFace identifies your face in one of the 400 million new photos that users upload every day, "you will get an alert from Facebook telling you that you appear in the picture," he explains. "You can then choose to blur out your face from the picture to protect your privacy." Many people, however, are troubled by the prospect of being identified at all—especially in strangers' photographs. Facebook is already using the system, although its face-tagging system only reveals to you the identities of your "friends."

DeepFace isn't the only horse in the race. The U.S. government has poured funding into university-based facial recognition research. And in the private sector, Google and other companies are pursuing their own projects to automatically identify people who appear in photos and videos.

Exactly how automated facial recognition will be used—and how the law may limit it—is unclear. But once the technology matures, it is bound to create as many privacy problems as it solves. "The genie

492 30 JANUARY 2015 • VOL 347 ISSUE 6221 sciencemag.org SCIENCE

DeepFace: closing the gap to human level performance in face verification, Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf, IEEE Conference on Computer Vision and Pattern Recognition, CVPR, June 2014.

The most cited paper (1200+ citations) on face recognition using deep neural networks.

Horizon Robotics

What is “Artificial Intelligence”

Artificial intelligence (AI) is defined as “the study and design of intelligent agents, in which an intelligent agent perceives its environment and takes actions that maximize its chance of success.

cyber space

physical environment → sensing → thinking → acting → physical environment

User search/clicks	Search	Ads recommendation
User behaviors	Ranking	Service personalization
Text/language	Inference	Language understanding
Speech/audio	Regression	Speech recognition
Image/video	Classification	Image/video recognition
Sensor signals	Understanding	Drone/vehicle control

Horizon Robotics

Why AI matters

PC + Connection

Mobile computing

IoT + intelligence

computation

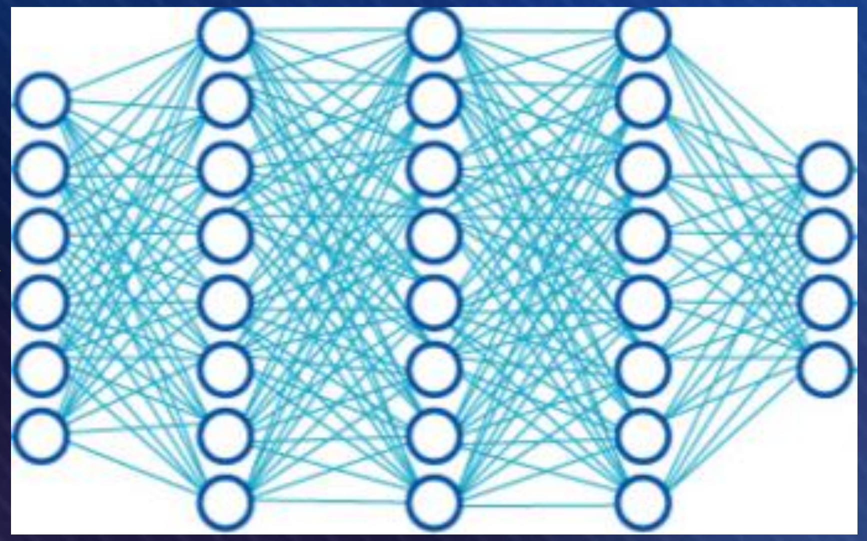
connectivity ↔ intelligence

Horizon Robotics

What is “deep Learning”?

raw data → representations

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen



end-to-end learning

Deep learning, Y. LeCun, Y. Bengio, G. Hinton, Nature, 2015

Horizon Robotics

Deep learning: a brief (incomplete) history

ANN/CNN/RNN/LSTM < Neural networks < Deep learning < AI

TPU, 2017

GAN, 2016

RL, AlphaGo, 2016.3.9

Autonomous driving, 2015

RNN/LSTM 2014-2015

Face recognition, 2014

Facebook AI Research, 2013

ImageNet+GPU
A. Krizhevsky, etc., 2012

Google brain / DistBelief
J. Dean, A. Ng, etc., 2011

DNN for speech recognition
G. Hinton, etc., 2010

Layer-by-layer training
G. Hinton, etc., 2006

Convolutional neural networks, 1989-1998

Back propagation (BP)
P. Werbos, Y.LeCun, etc., 1980s

Artificial neural networks (ANN)
K. Fukushima, etc., 1980

Perceptrons vs. XOR
M. Minsky, 1969

Perceptron, artificial neuron
F. Rosenblatt, 1957

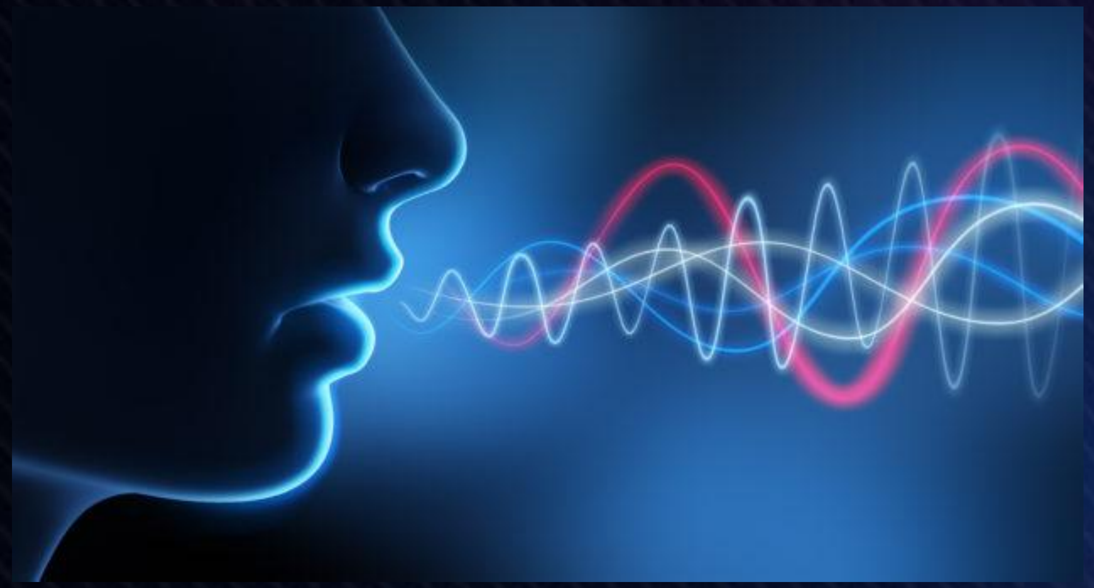
On every smart device / thing!

1950 | 2000 | 2017 | 2020

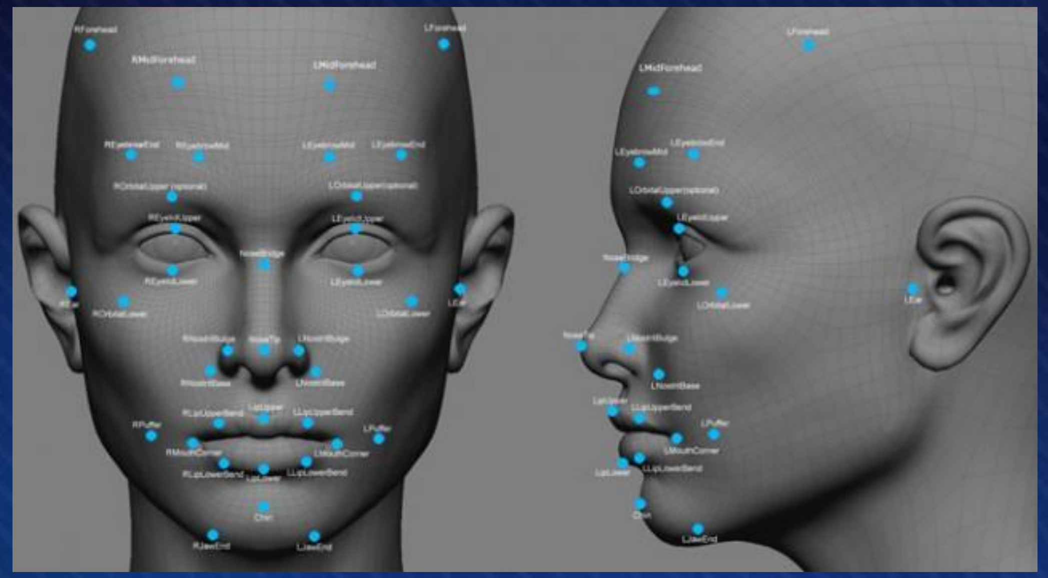
Horizon Robotics

Application fields: quick penetration

Speech



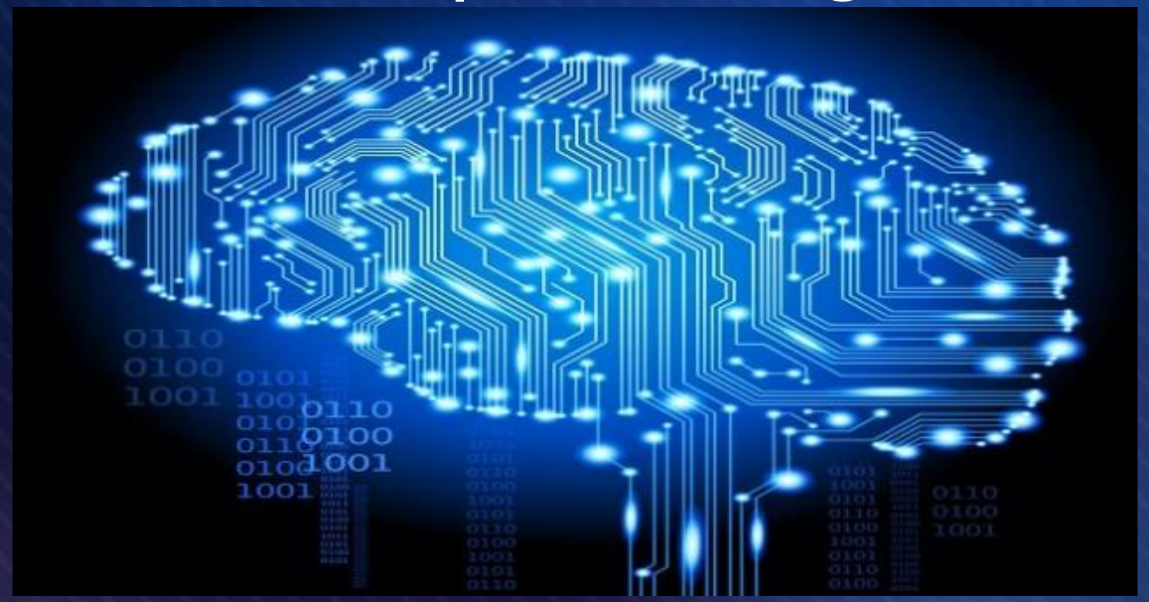
Vision



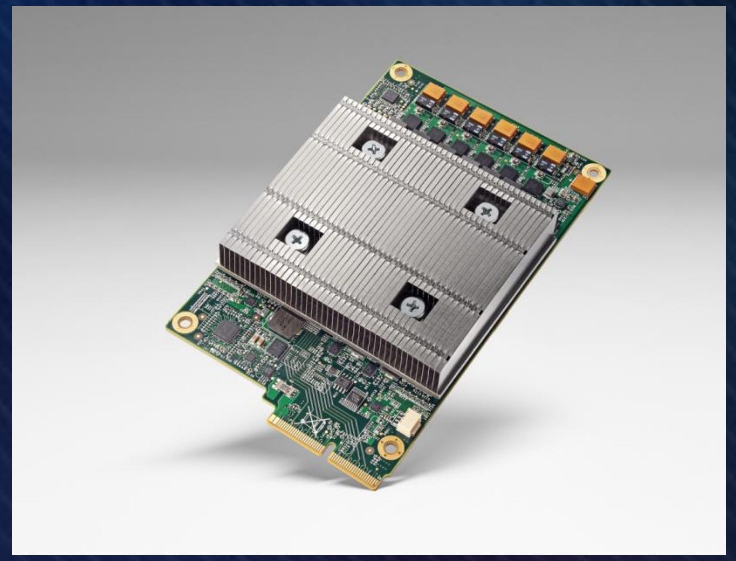
Automotive

Ads/Search

Deep learning



Hardware



Fintech

Industry

Healthcare

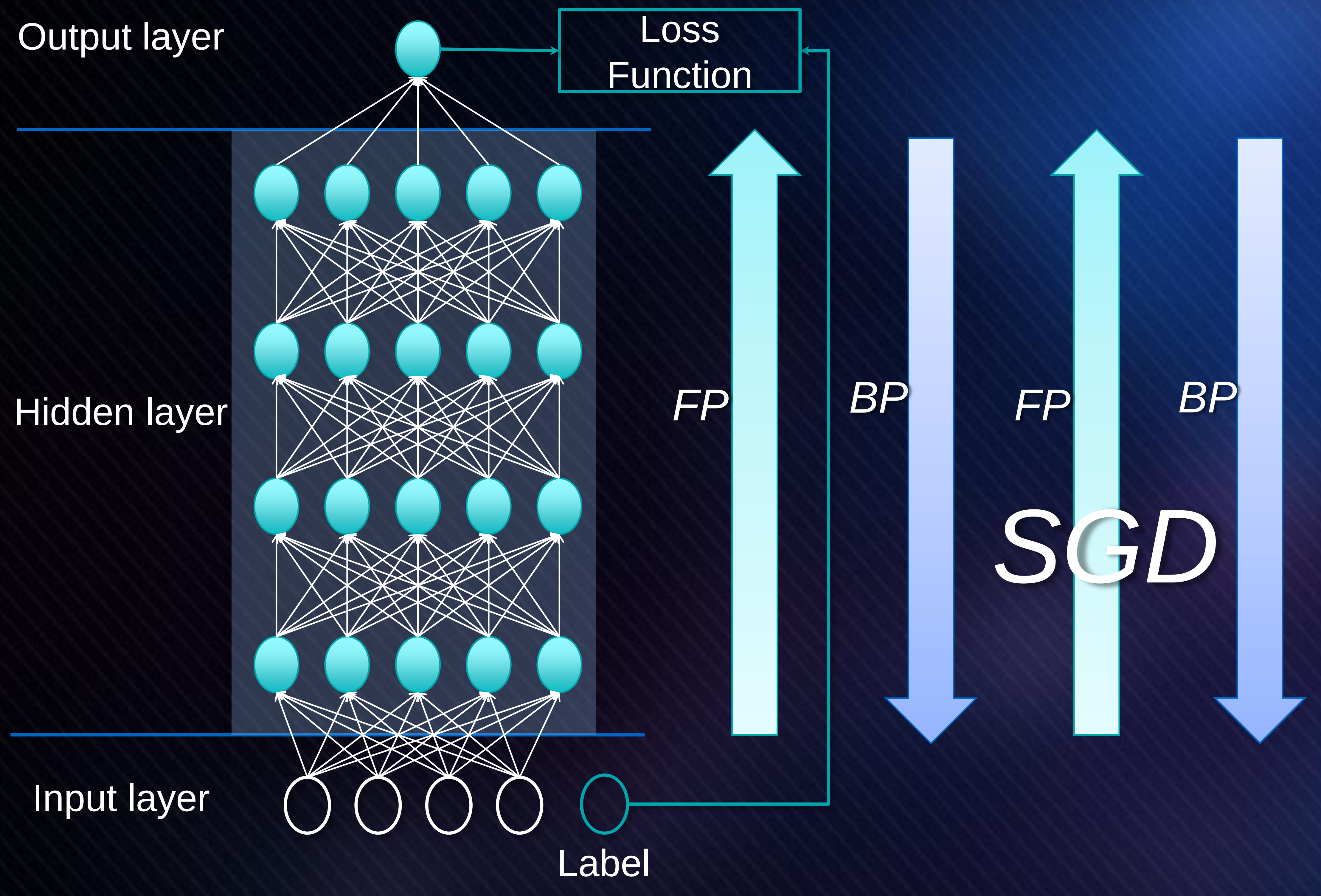
A thousand of AI startups (incomplete)

1,727 AI startups across 13 categories, with a combined funding amount of 14.5B$ (by VentureScanner, 2017/4)

957 startups with 4.8B$ investment, in 2016/3

AI winter => deep learning spring, why?

(Deep) Multi-layer neural networks



What's new?

Challenges

- ✓ over-fitting
- ✓ noise sensitive
- ✓ a black box
- ✓ black magic

  an art to tune the network architecture

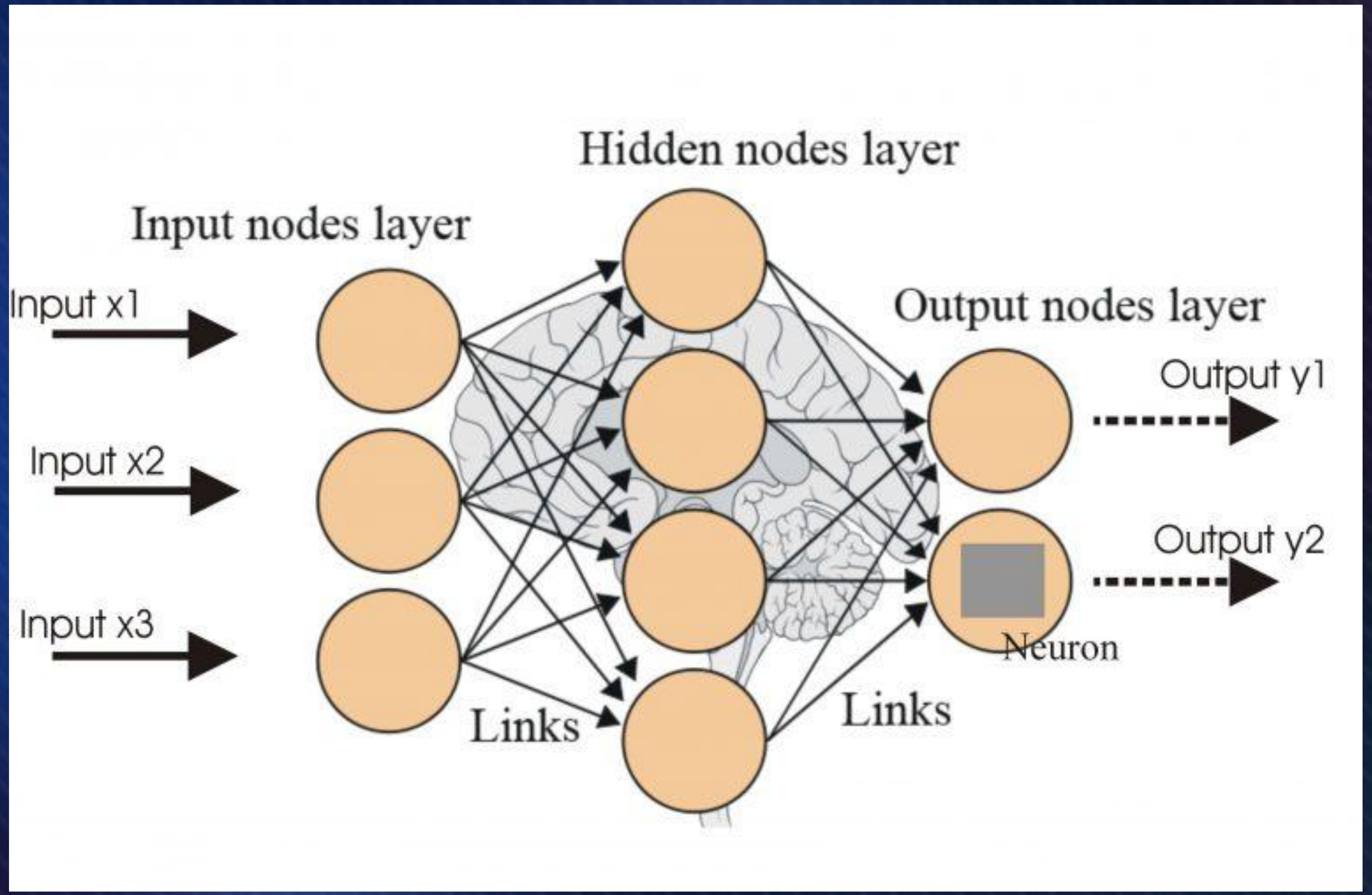
Now

- ✓ big big big data
- ✓ GPU + parallelism
- ✓ training tricks
- ✓ black magic

  an art to tune the network architecture

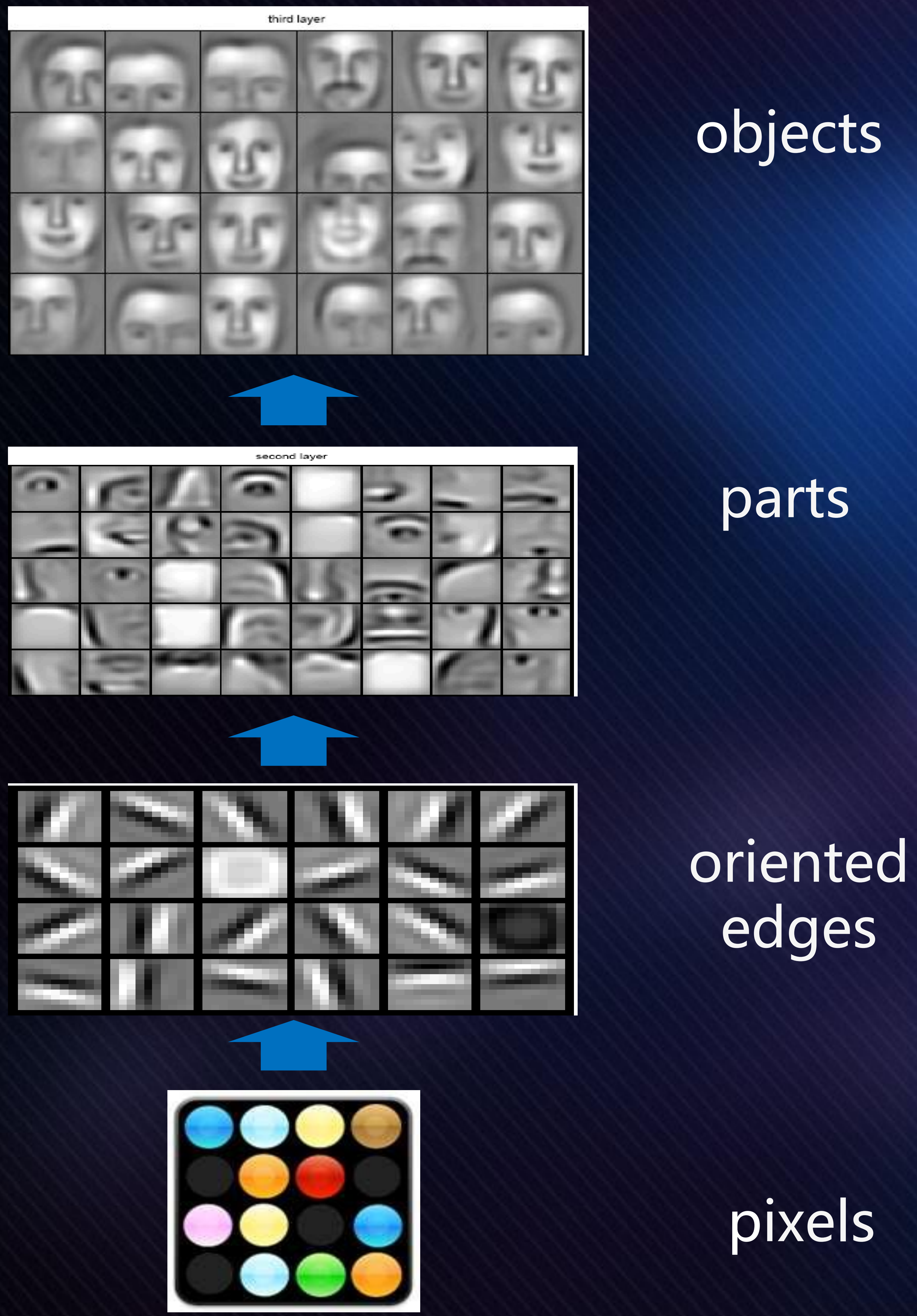
AI winter => deep learning spring, why?

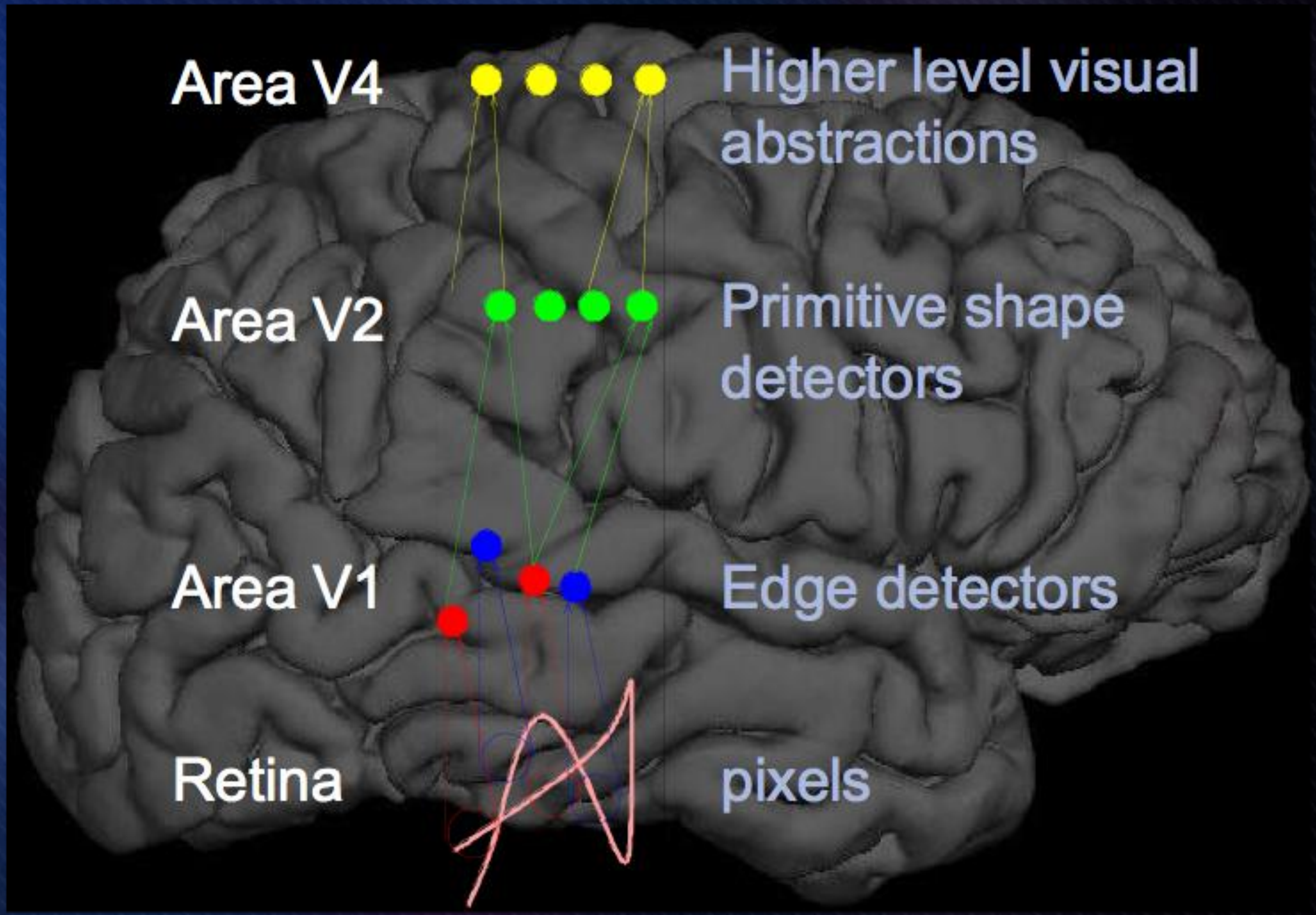
- Biological Inspired
- Big big big data
- End-to-end learning
- A flexible modeling language



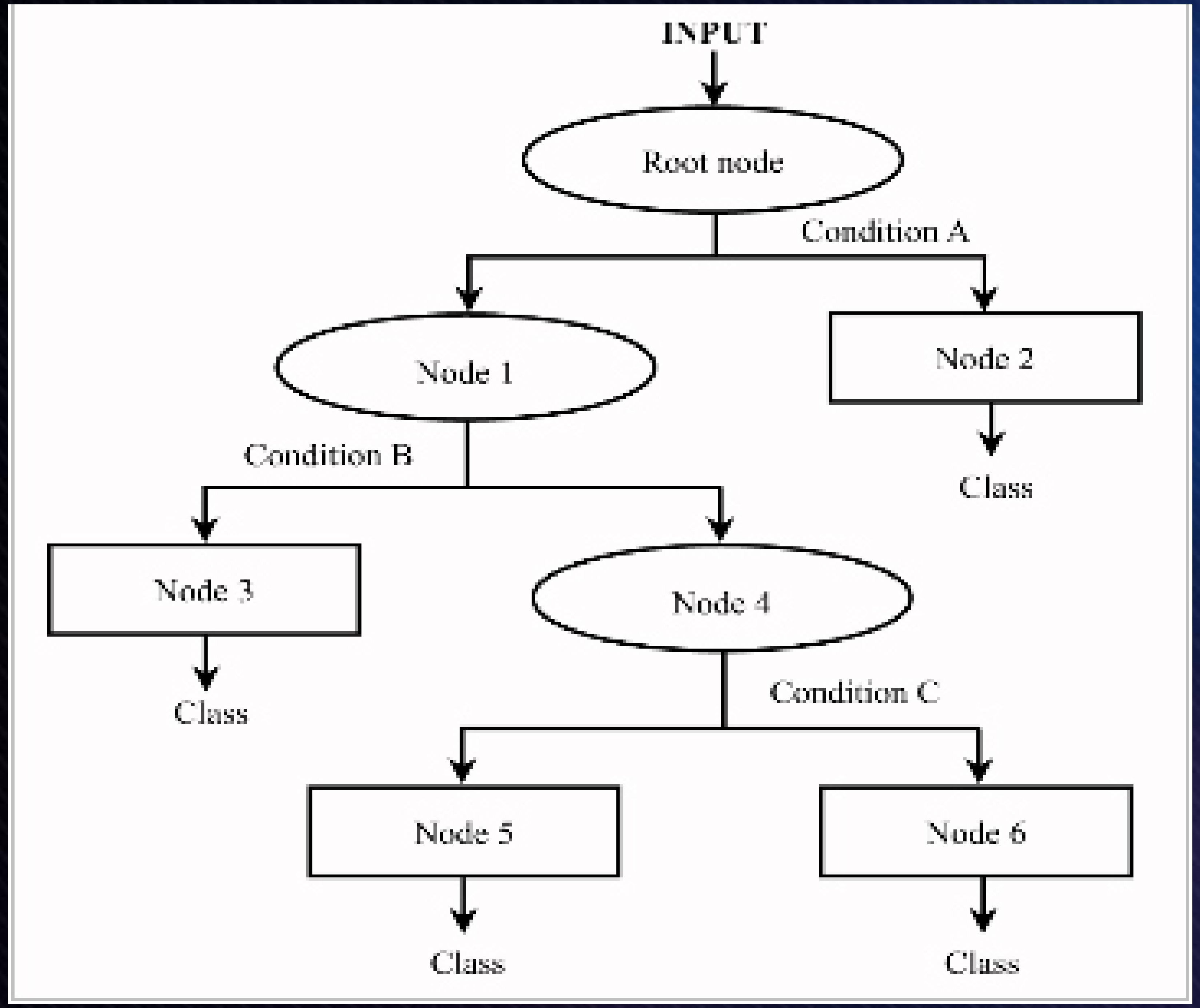
Horizon Robotics

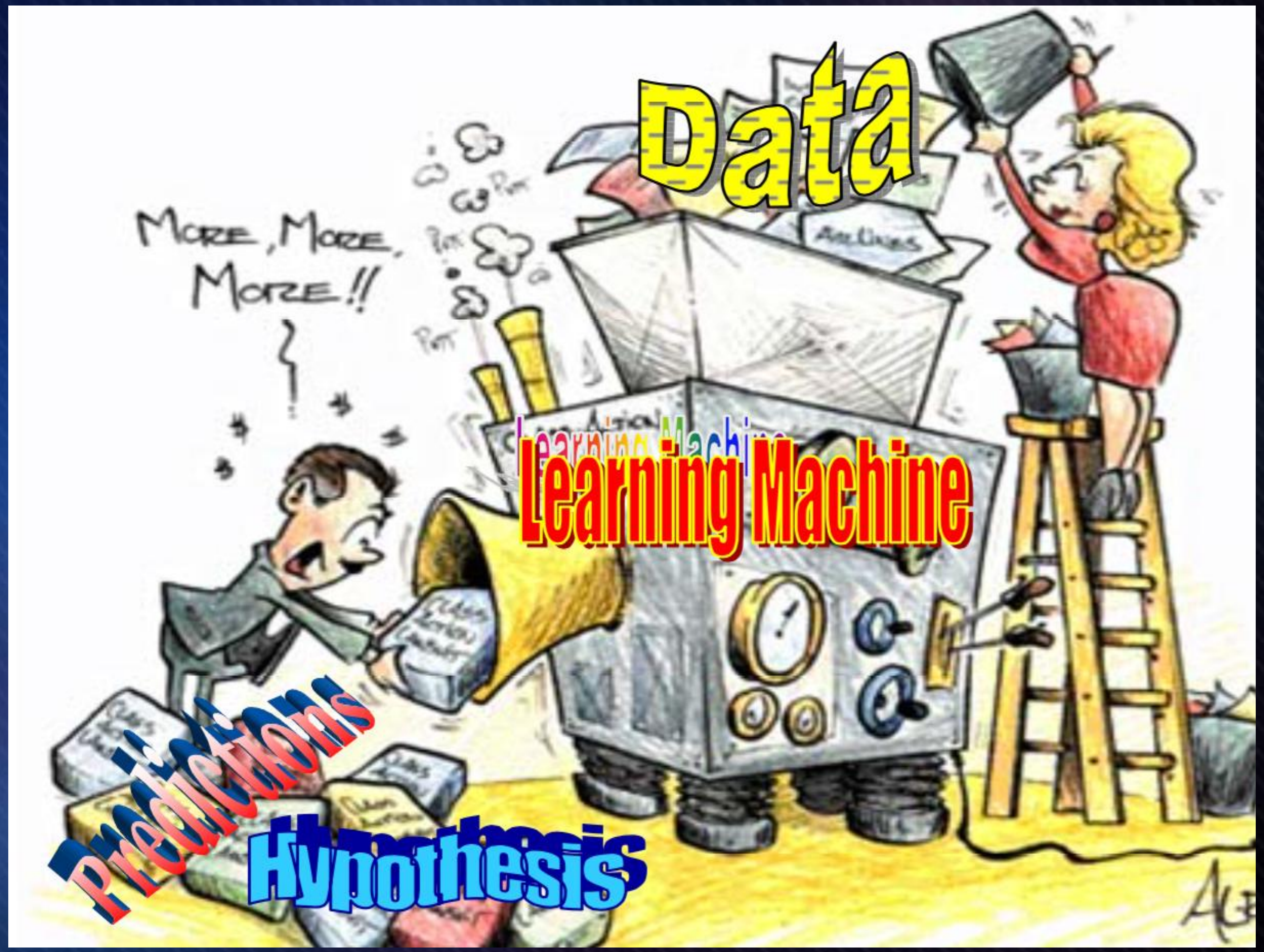
Biological inspired by brain’s structure and behavior





AI : 60 year history in one page

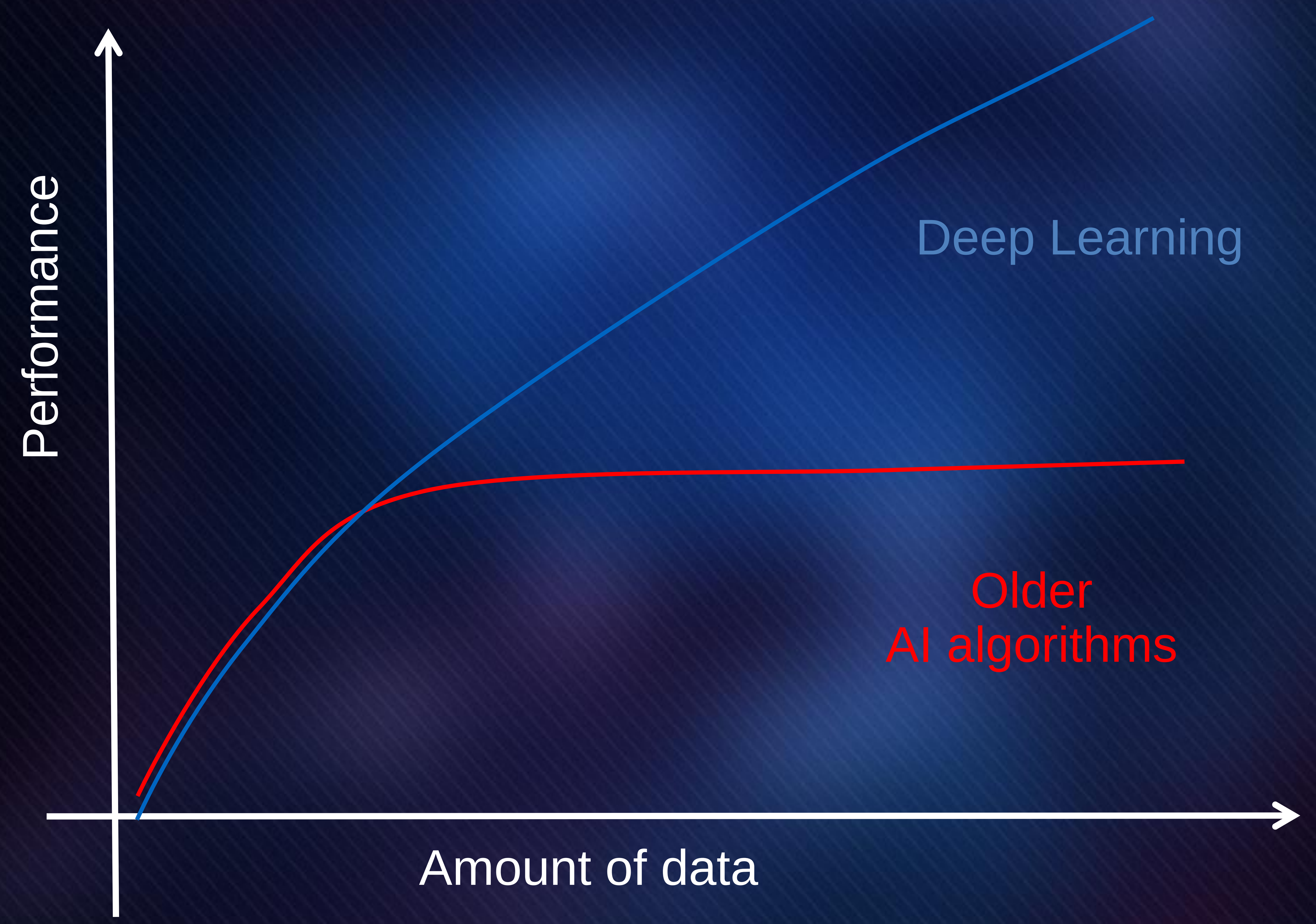




Rule-based AI

Data-driven AI

Big data, the more the better?



Horizon Robotics

End-to-end learning

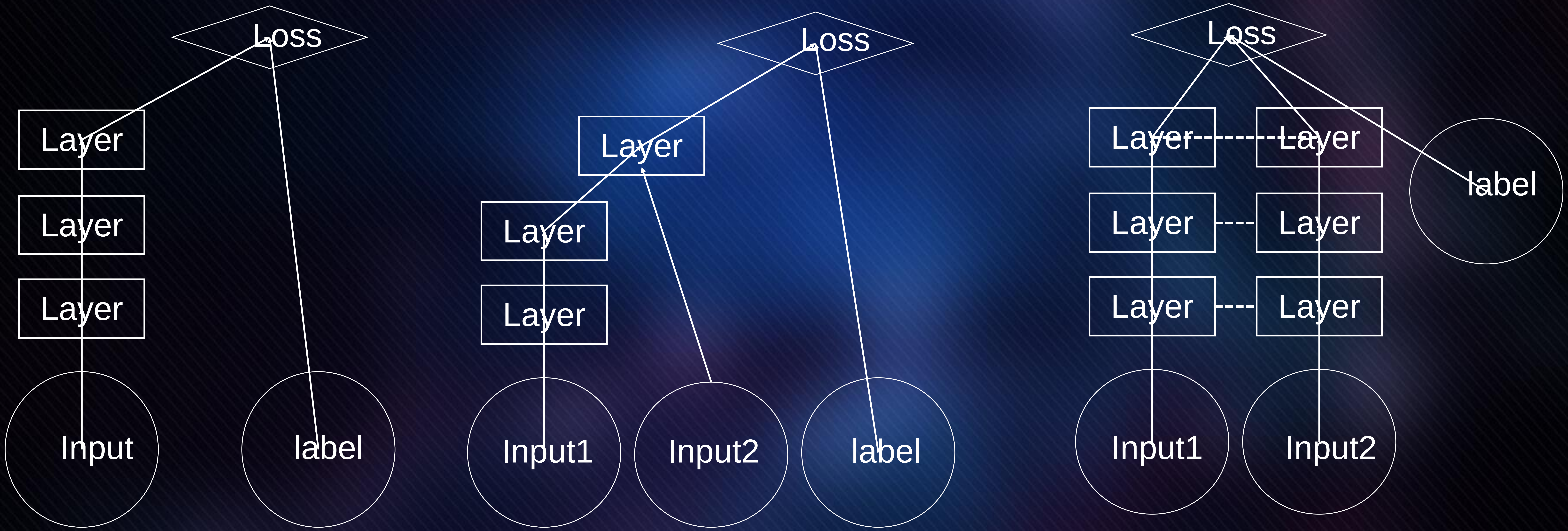
Most Efforts in Machine Learning

Low-level sensing → Pre-processing → Feature extract. → Feature transform → Prediction, recognition

- Most critical for accuracy
- Account for most of the computation for testing
- Most time-consuming in development cycle
- Often hand-craft in practice

Deep Learning

A flexible modeling language



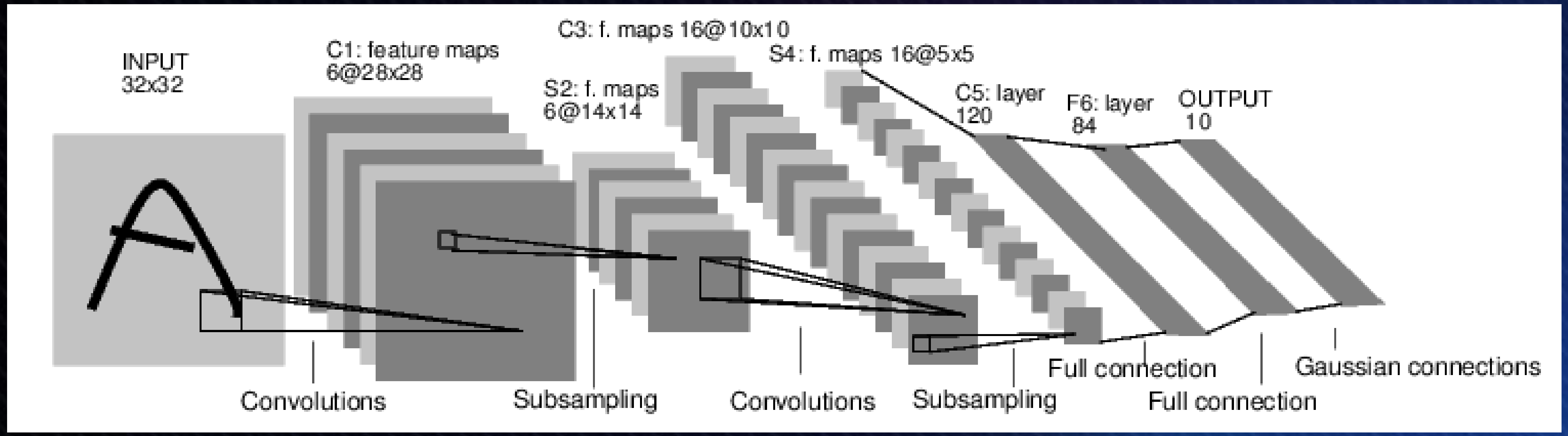
The network architecture is very flexible, so long as it is differentiable and learnable!

Development of neural networks

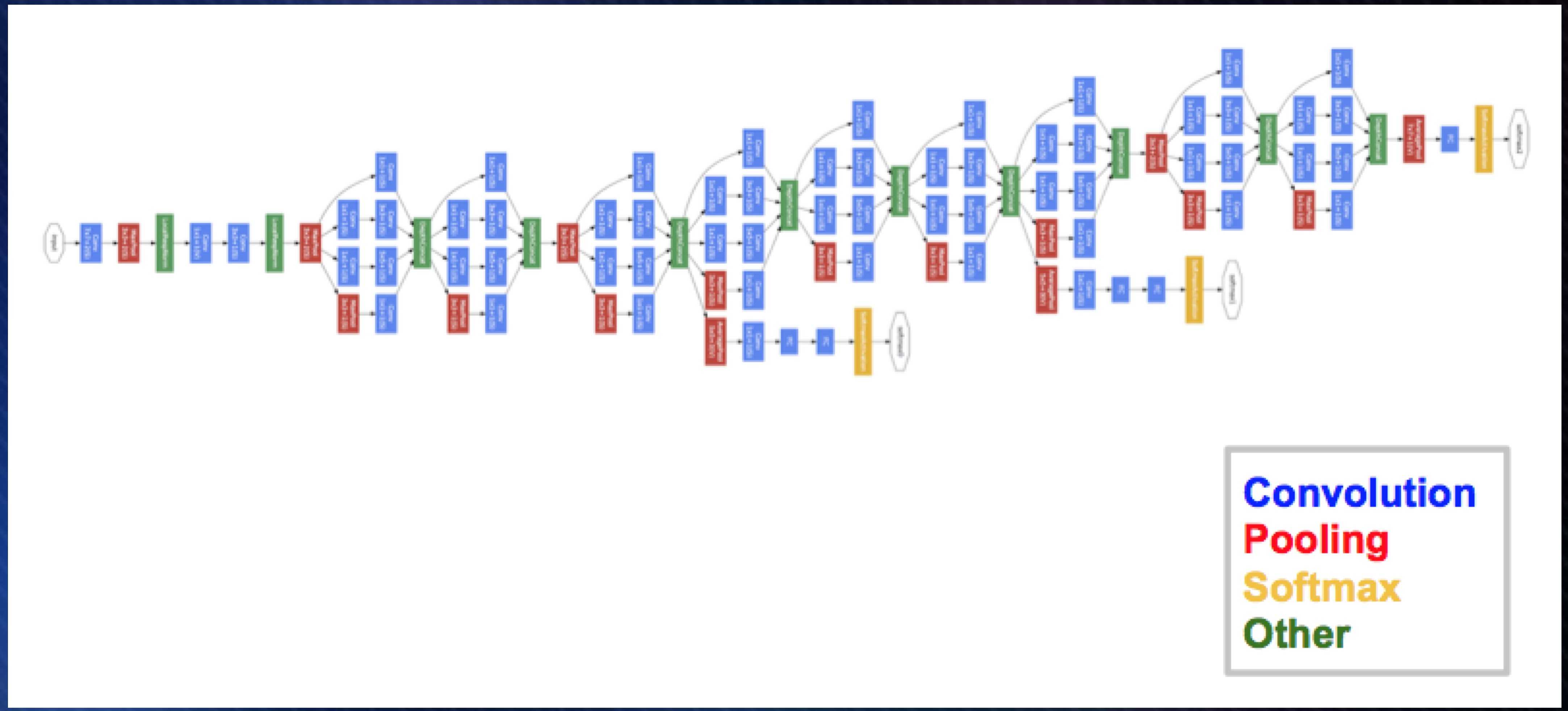
- Network performance, size, operations

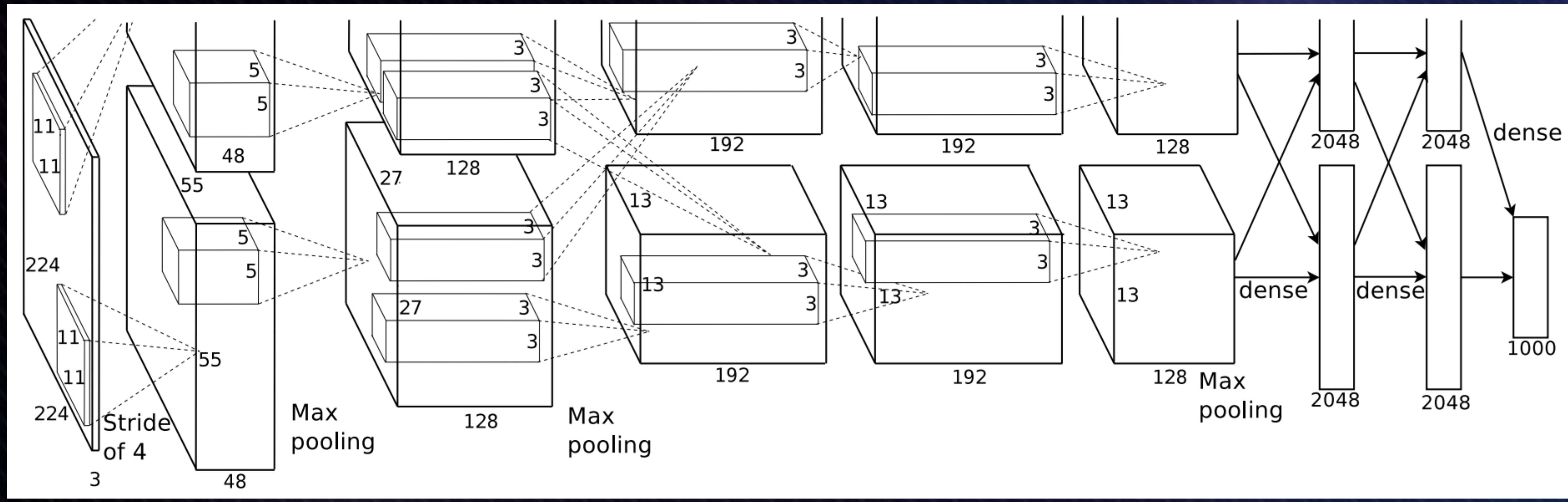
- Network architecture : the deeper the better? spatially or temporally
- Smaller storage: model compression, or with low precision operations
- Faster computation : less operations, or hardware acceleration

Evolving of network architecture

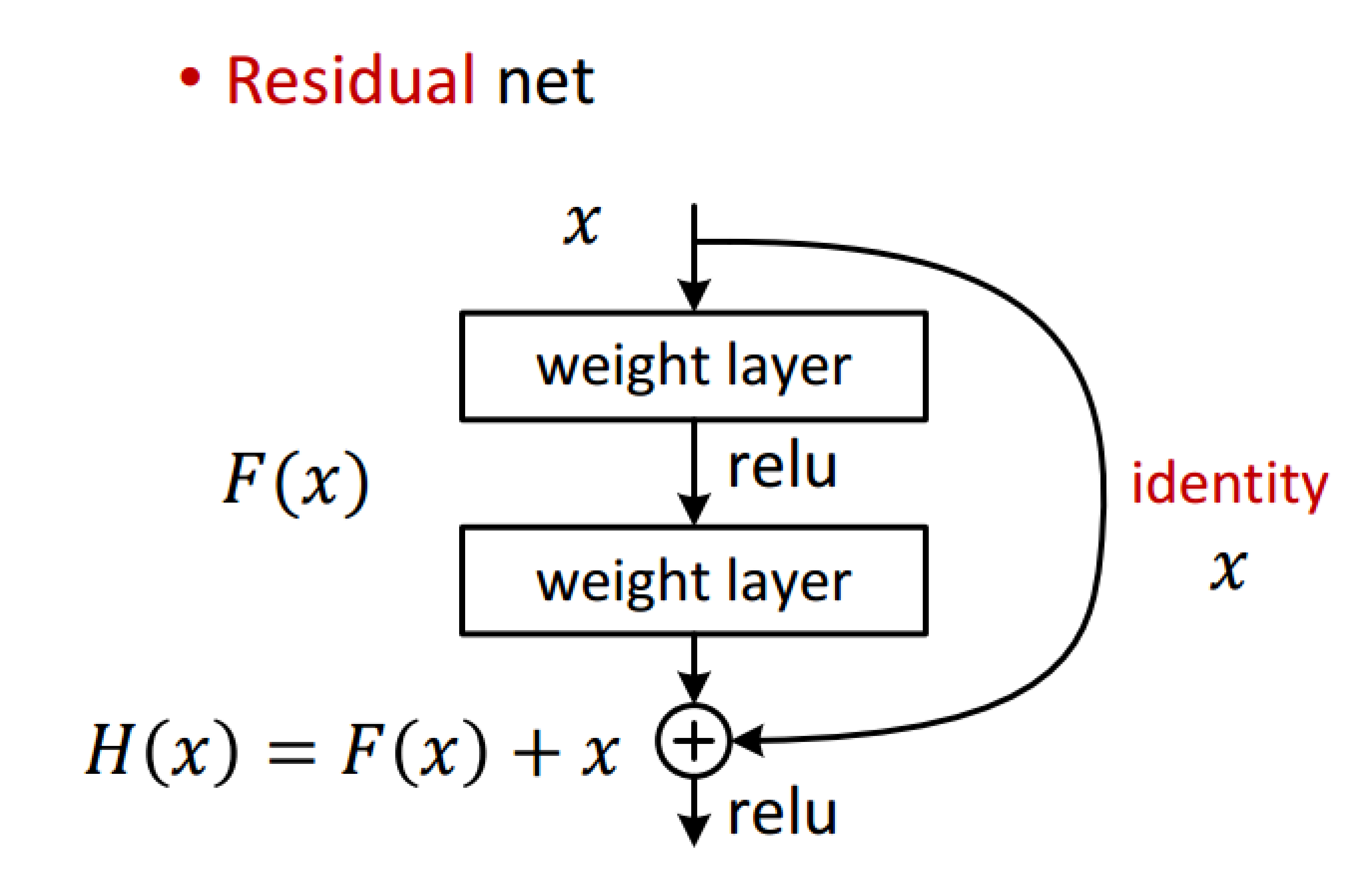


Gradient-based learning applied to document recognition, Y. LeCun, L. Bottou, Y. Bengio, P. Halfner, Proc. of IEEE, 1998

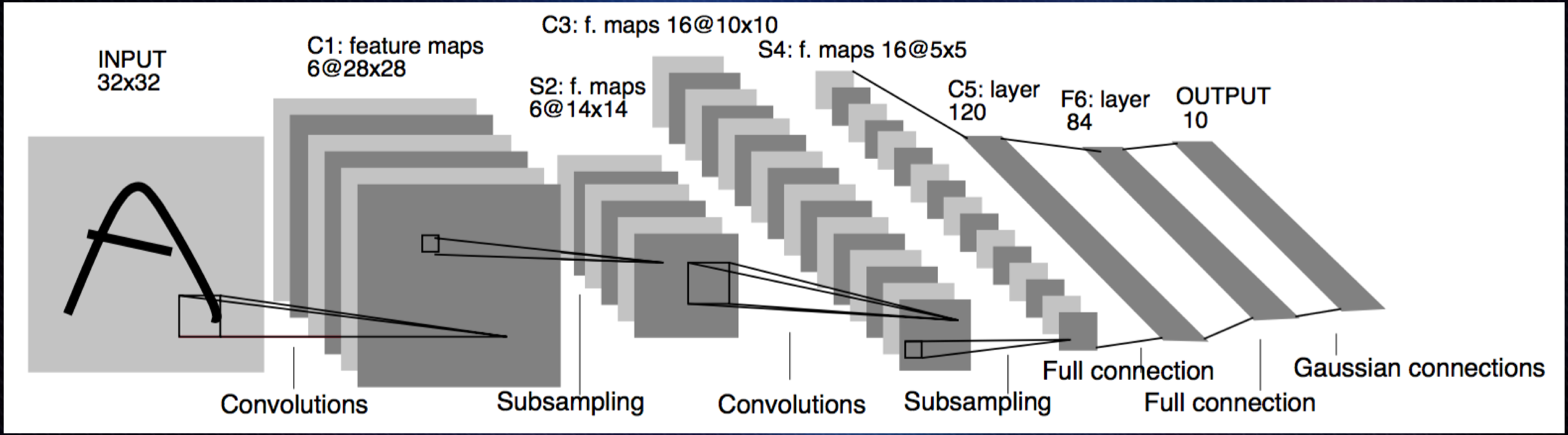




ImageNet classification with deep convolutional neural networks, A. Krizhevsky, L. Sutskever, G. Hinton, NIPS 2012

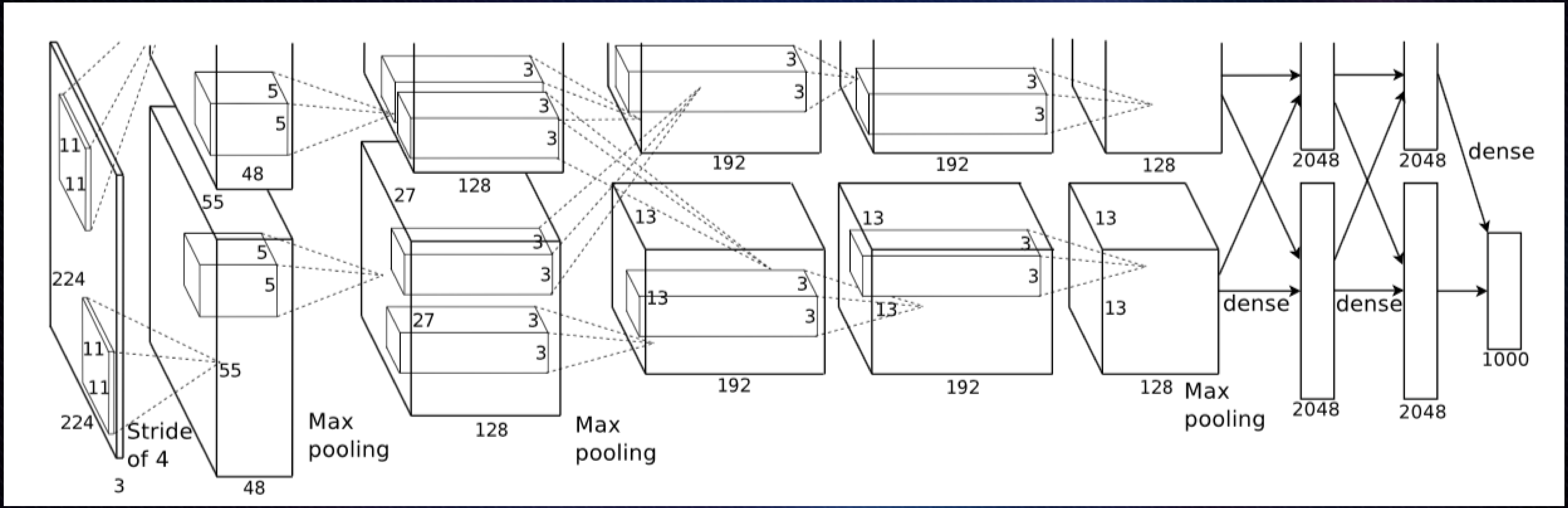


LeNet-5



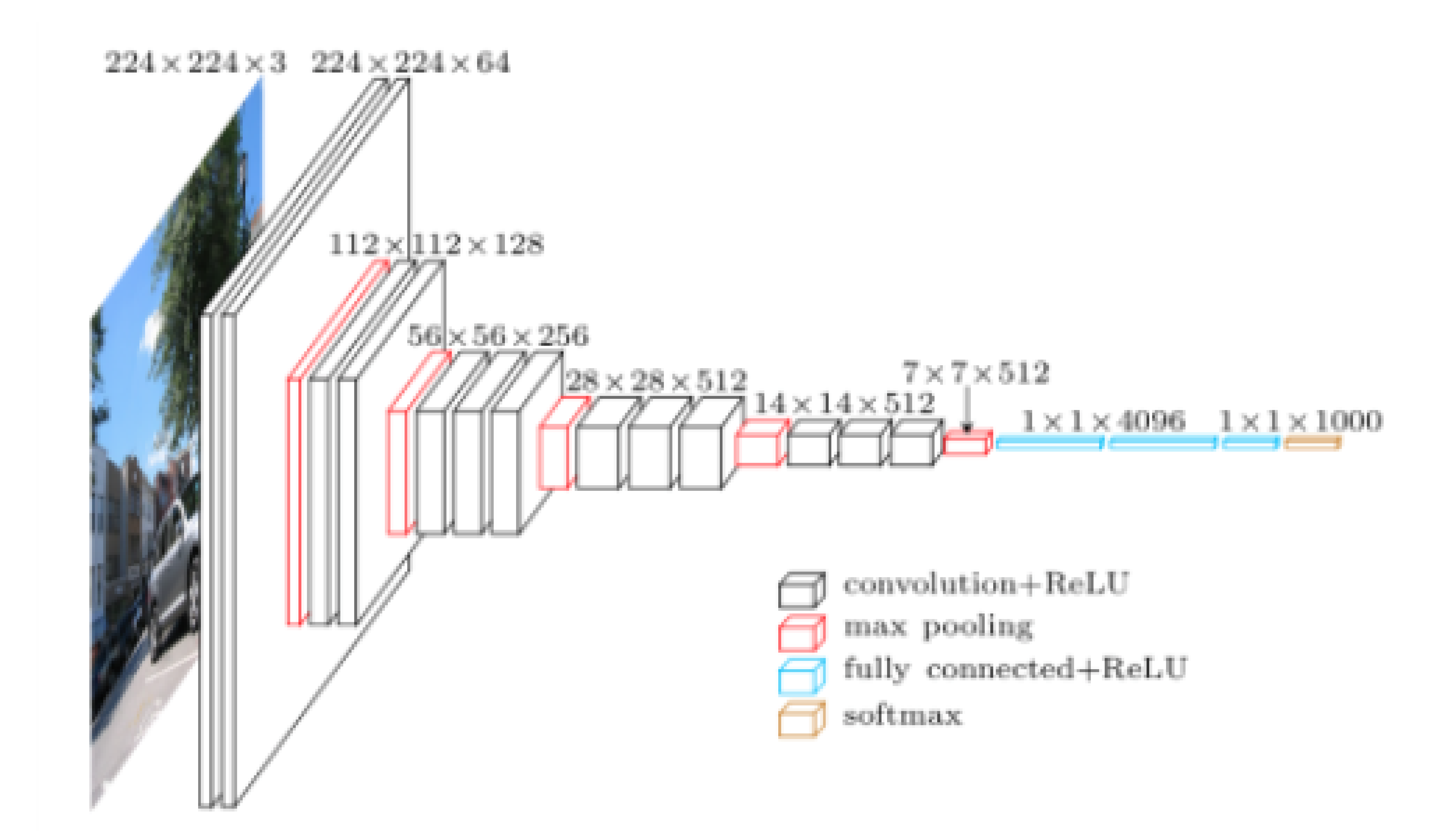
Gradient-based learning applied to document recognition, Y. LeCun, L. Bottou, Y. Bengio, P. Halfner, Proc. of IEEE, 1998

AlexNet



ImageNet classification with deep convolutional neural networks, A. Krizhevsky, L. Sutskever, G. Hinton, NIPS 2012

VGG

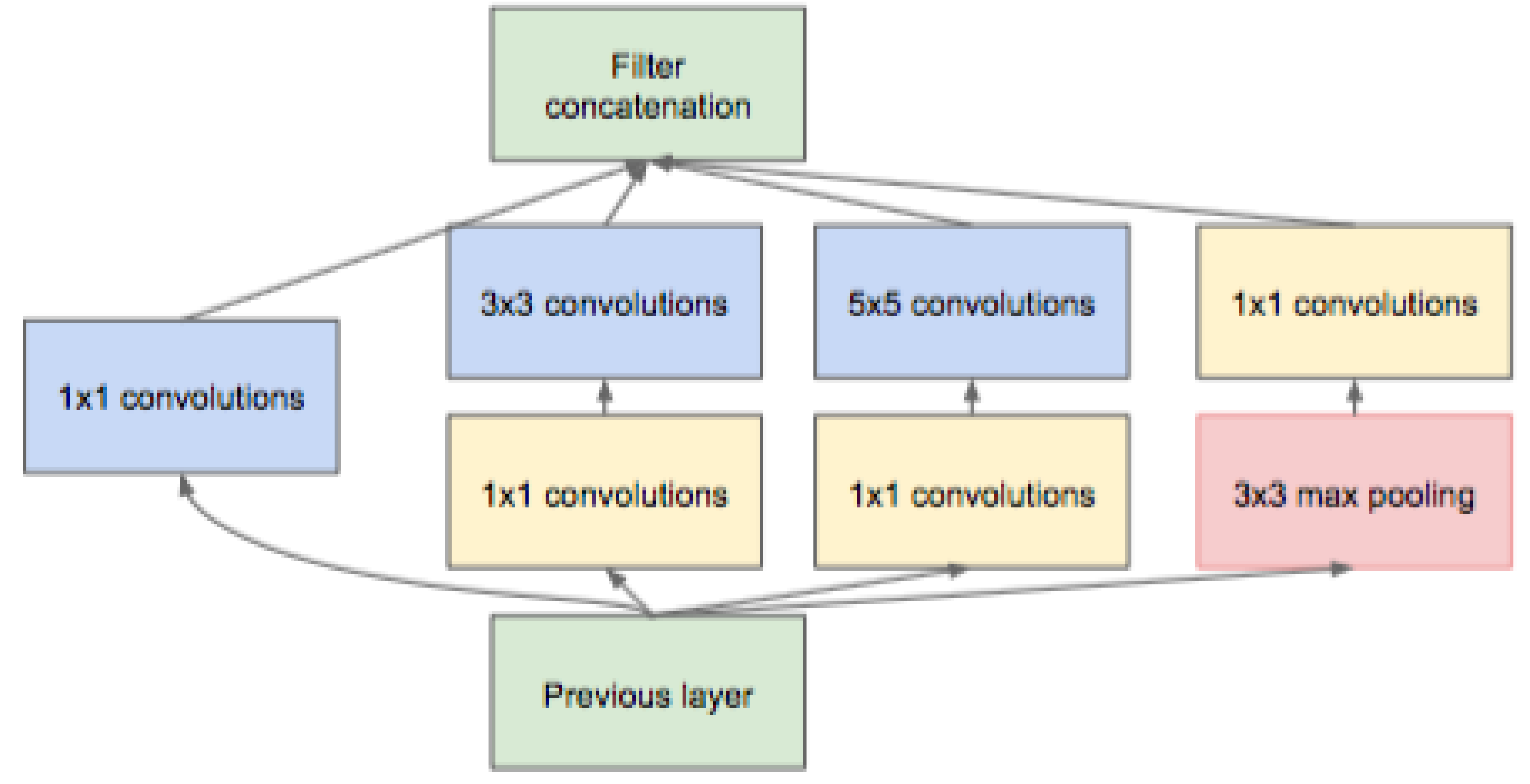


Very deep convolutional networks for large-scale image recognition, K. Simonyan, A. Zisserman, ICLR 2015

Network in network

Network in network, M. Lin, Q. Chen, S. Yan, 2014

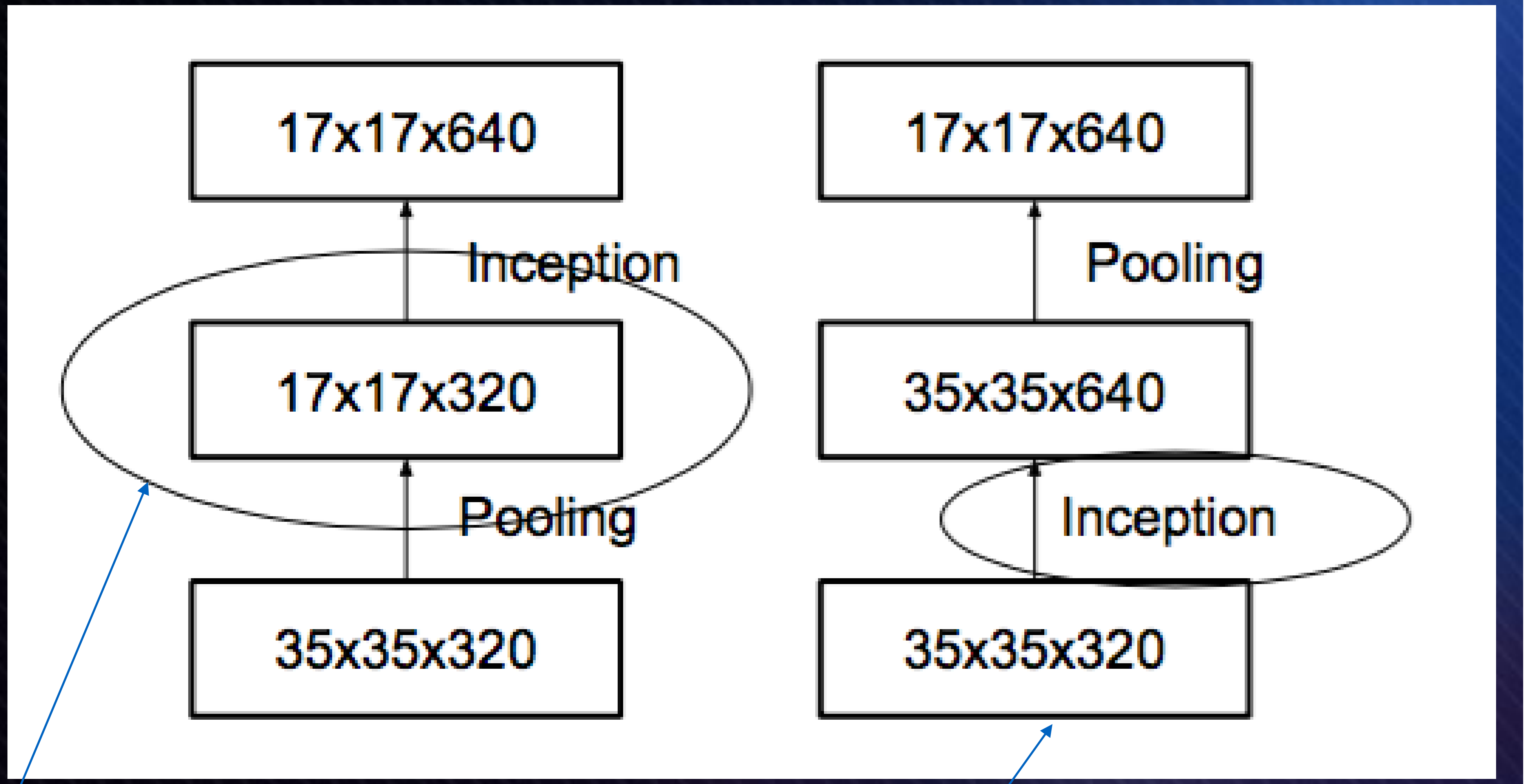
GoogLeNet

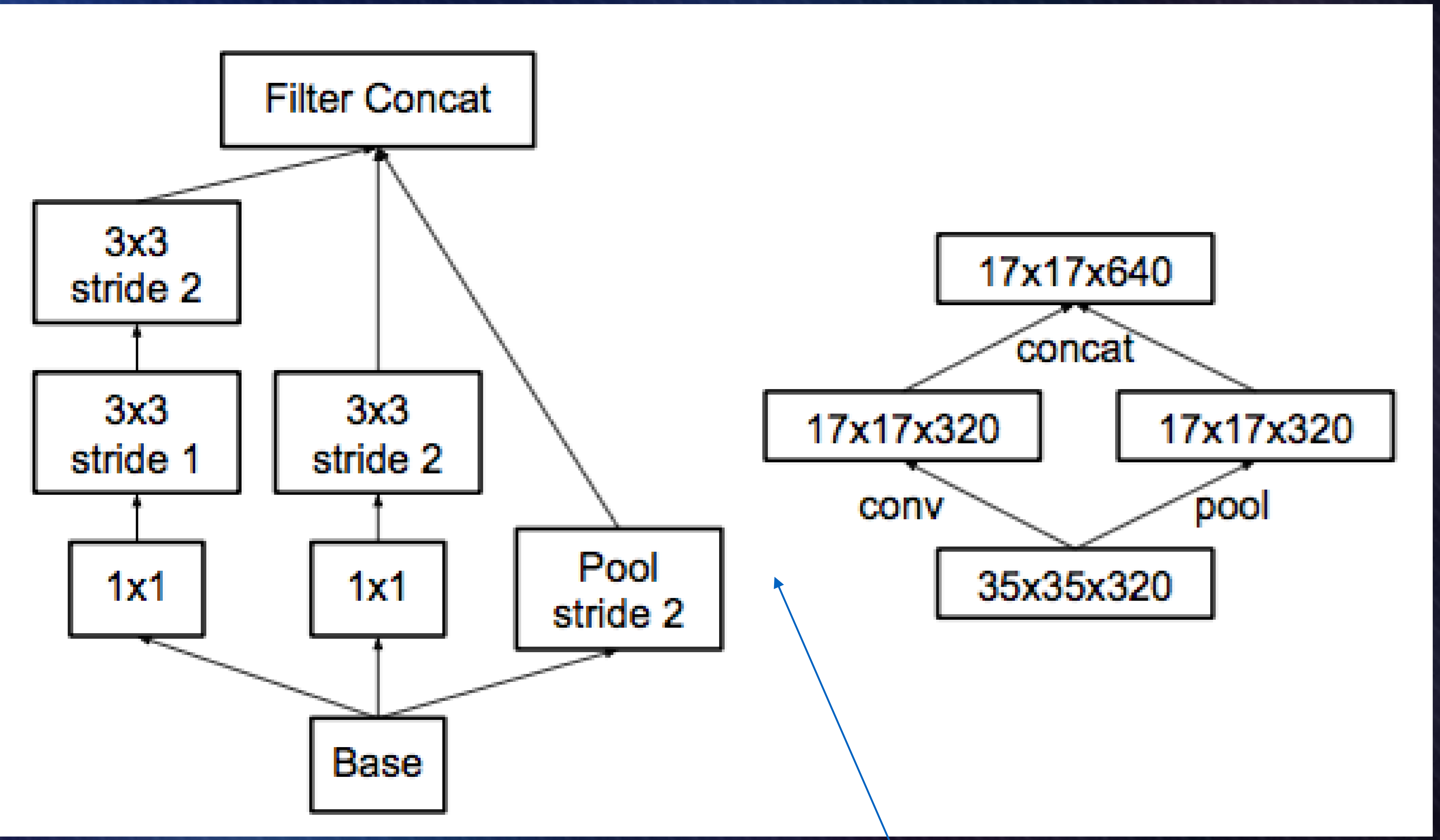


Going deeper with convolutions, C. Szegedy, W. Liu, Y. Jia, etc., CVPR 2015

Design Principles of Inception v3

- Avoid representational bottlenecks





A bottleneck

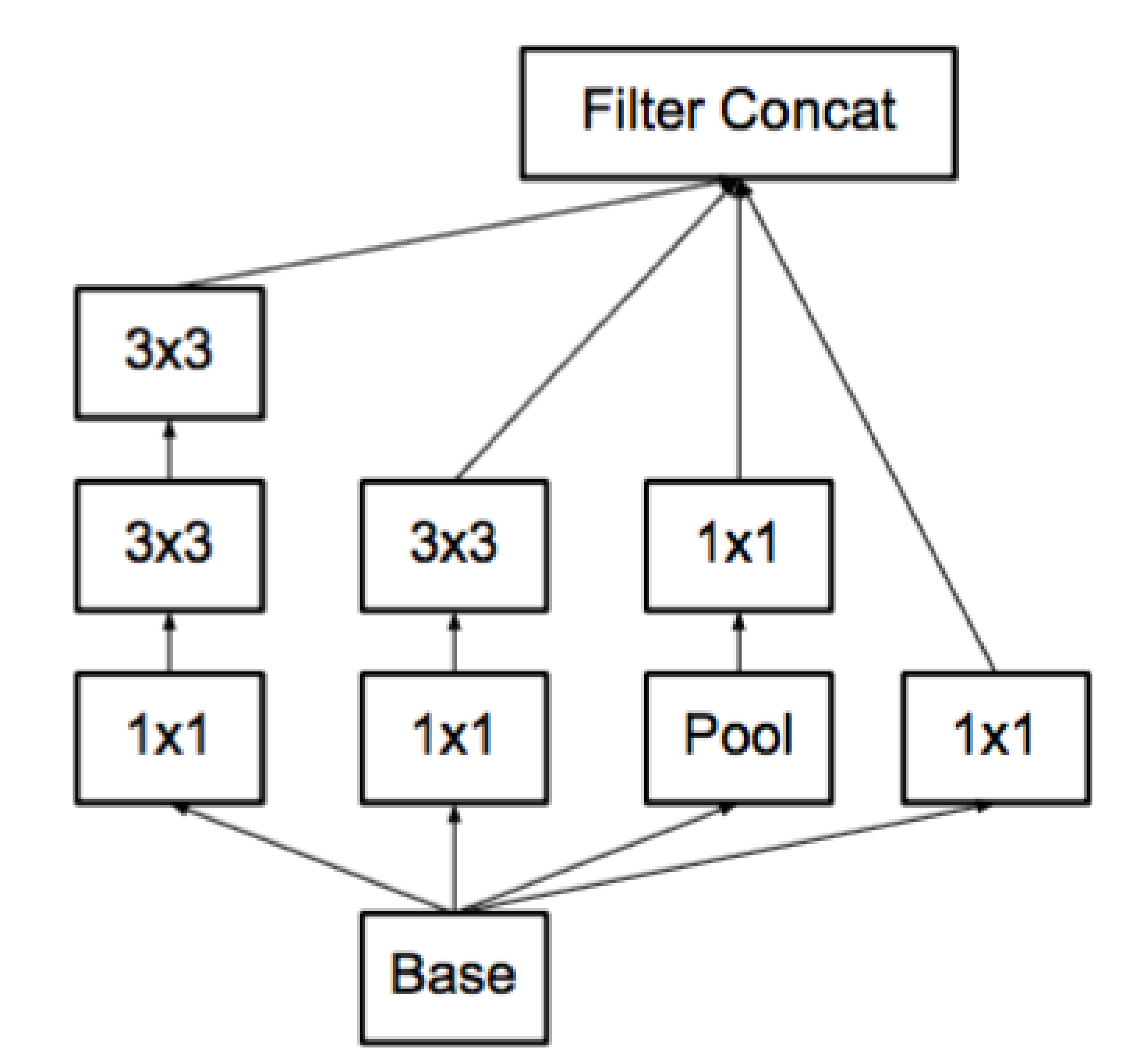
Expensive

Proposed Solution

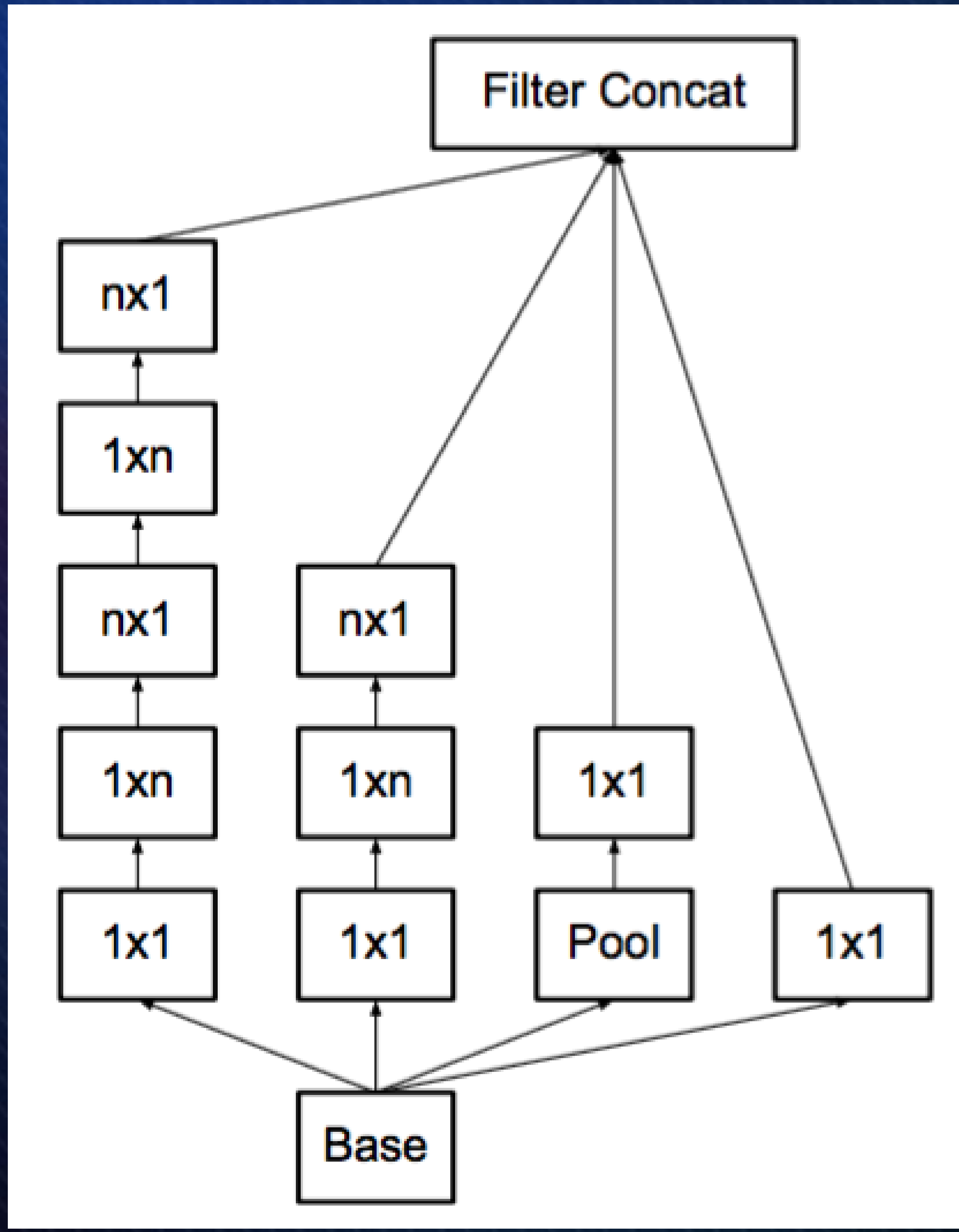
Rethinking the inception architecture for computer vision, C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, 2015

Design Principles of Inception v3

- Avoid representational bottlenecks
- Spatial aggregation can be done over lower dimensional embeddings



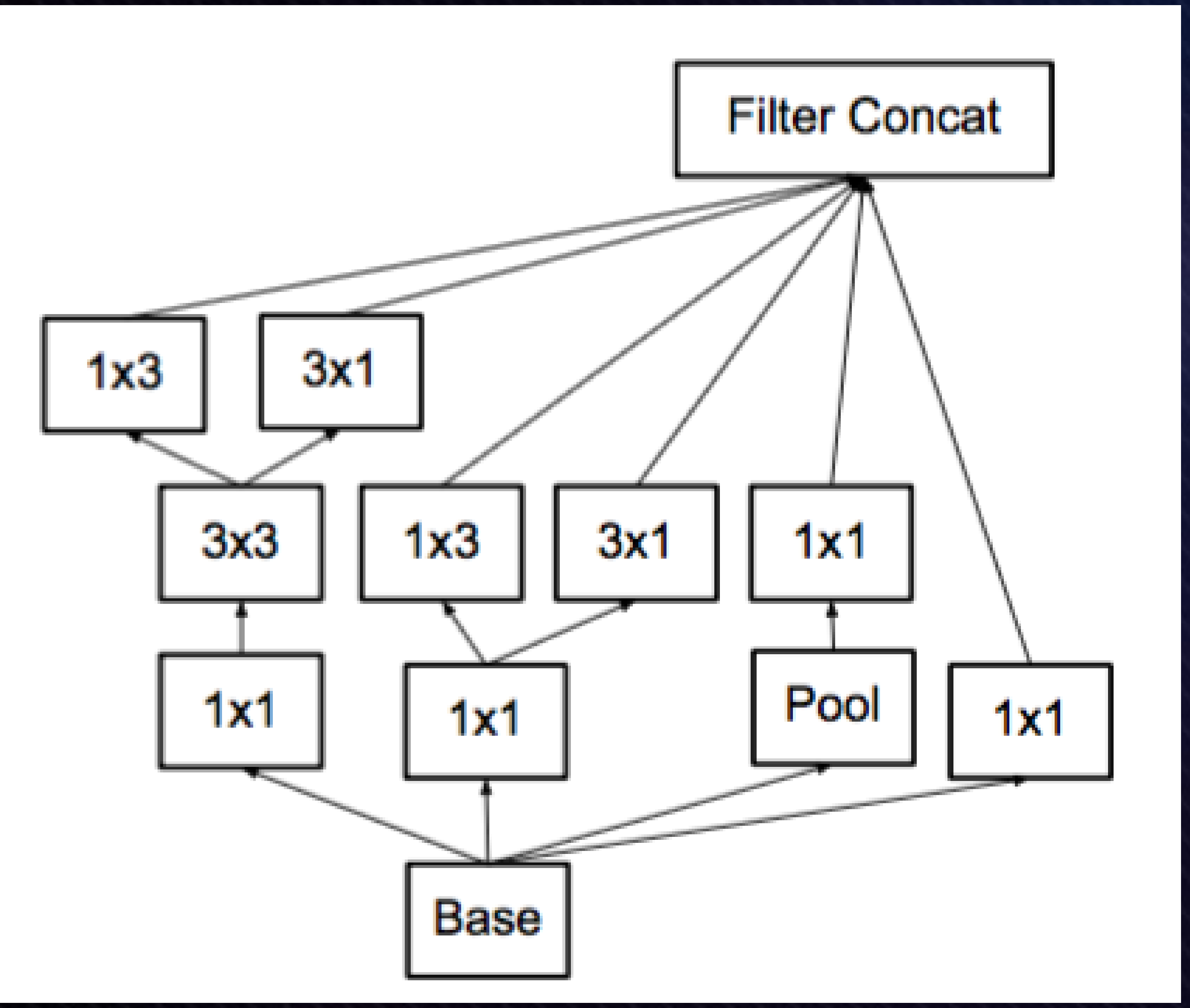
Used on 35 x 35 grids



Used on 17 x 17 grids
n=7

Design Principles of Inception v3

- Avoid representational bottlenecks
- Spatial aggregation can be done over lower dimensional embeddings
- Higher dimensional representations are easier to process locally



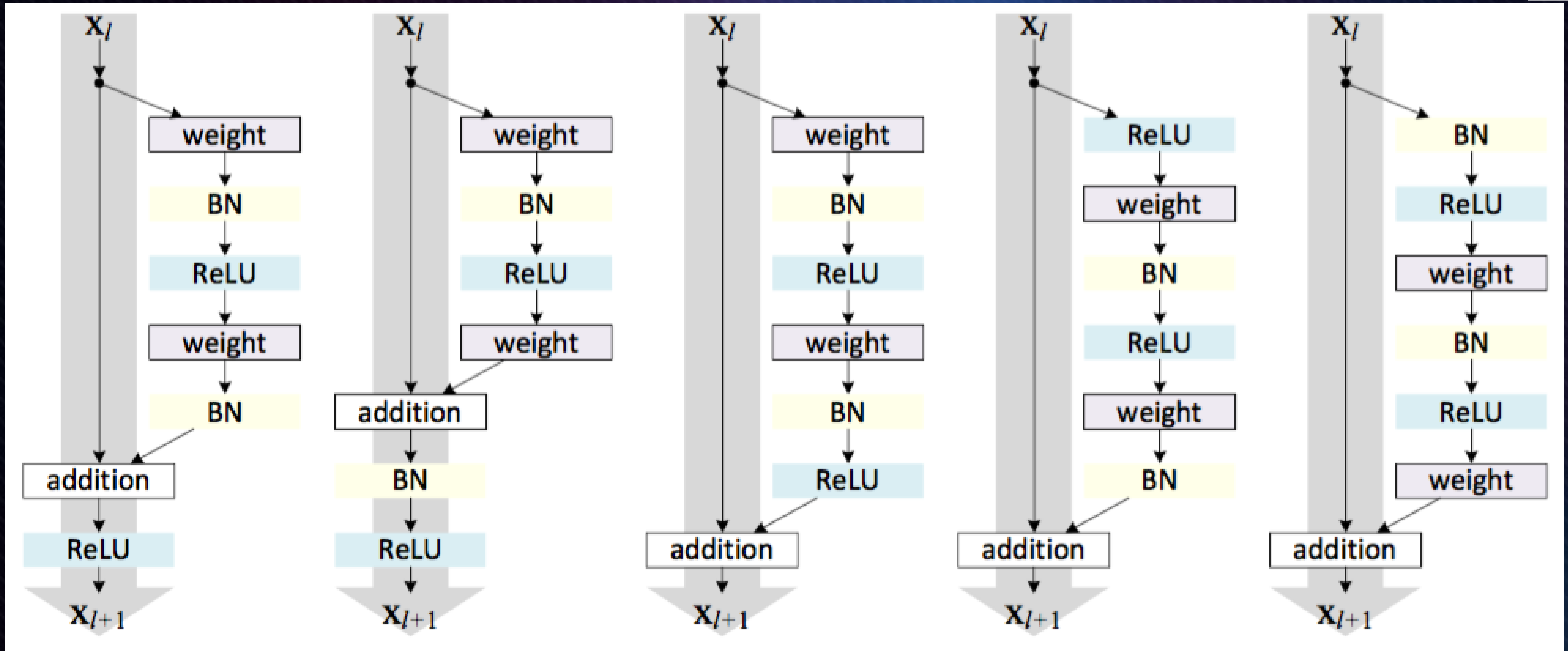
Used on the coarsest (8 x 8) grids

Design Principles of Inception v3

- Avoid representational bottlenecks
- Spatial aggregation can be done over lower dimensional embeddings
- Higher dimensional representations are easier to process locally
- Balance the width and depth of the network

ResNet

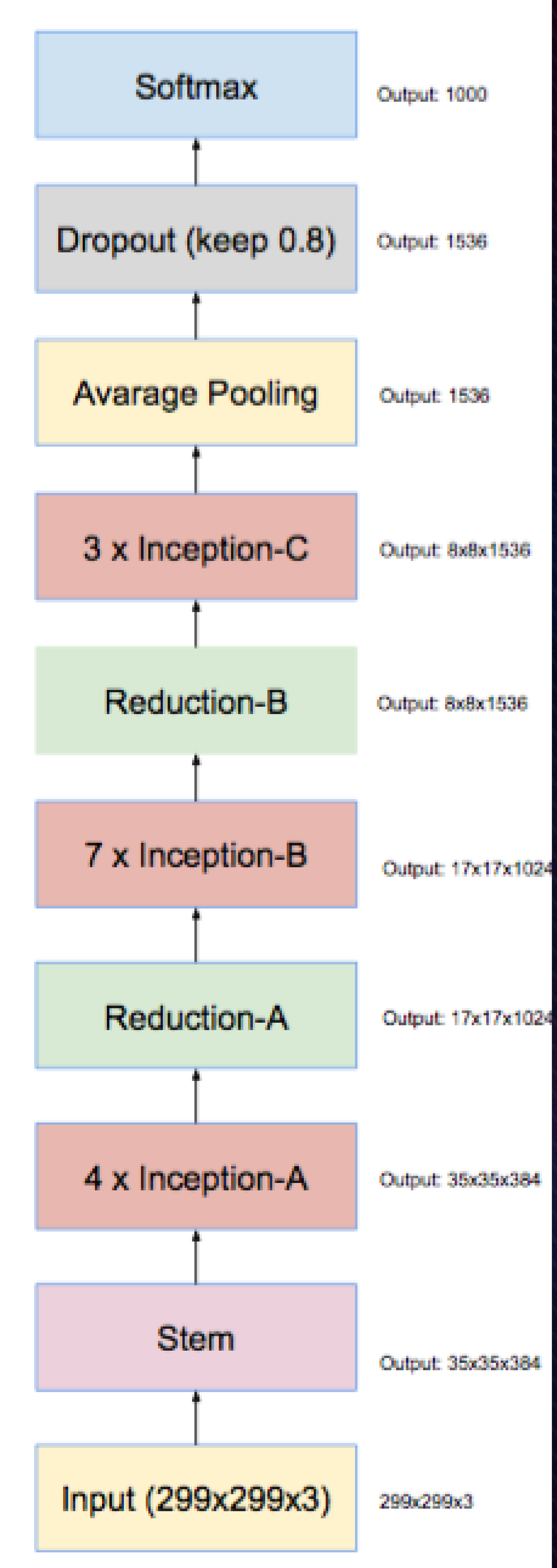
Residual block v1



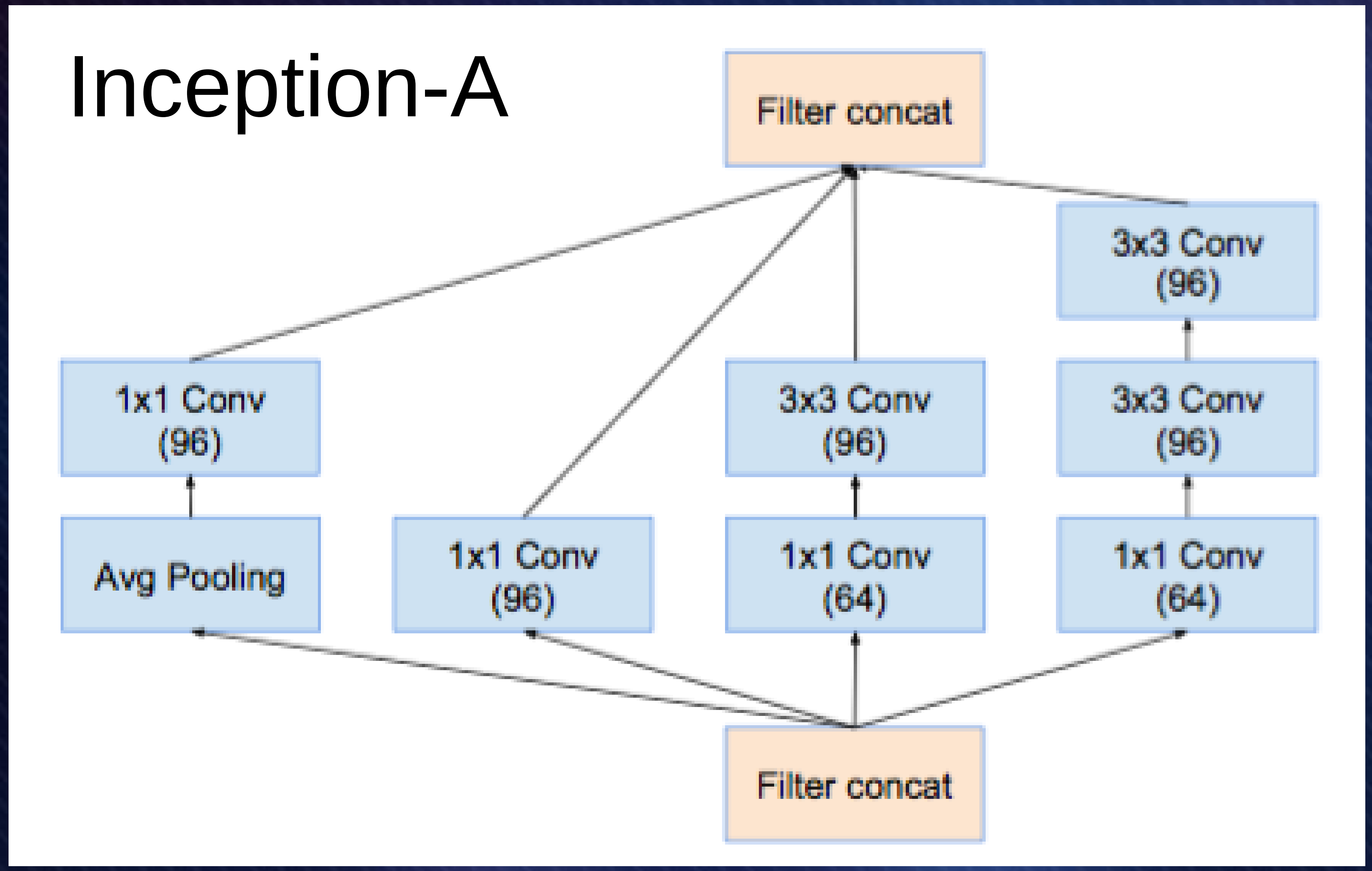
Residual block v2

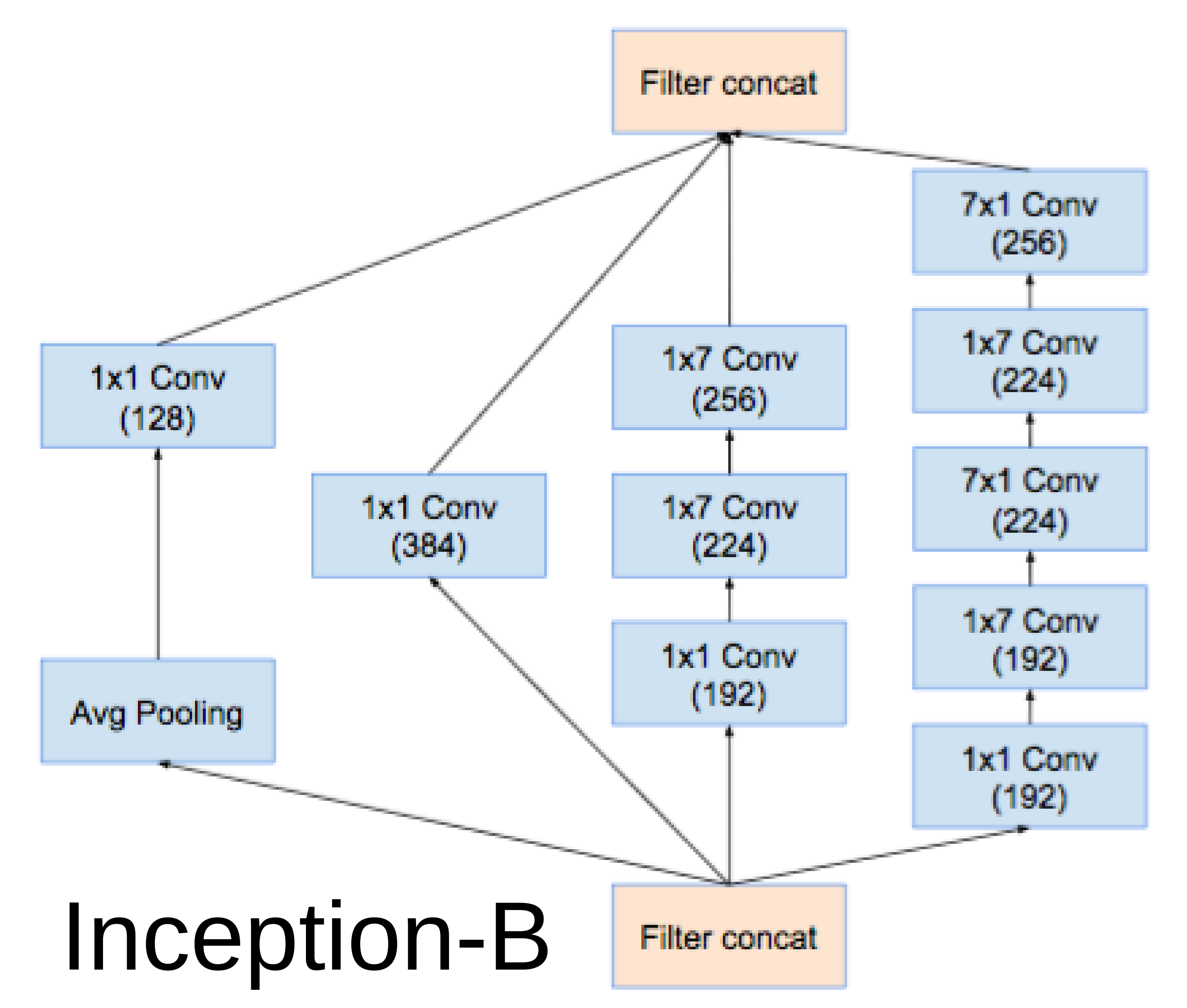
Deep residual learning for image recognition, K. He, X. Zhang, S. Ren, J. Sun

Inception v4



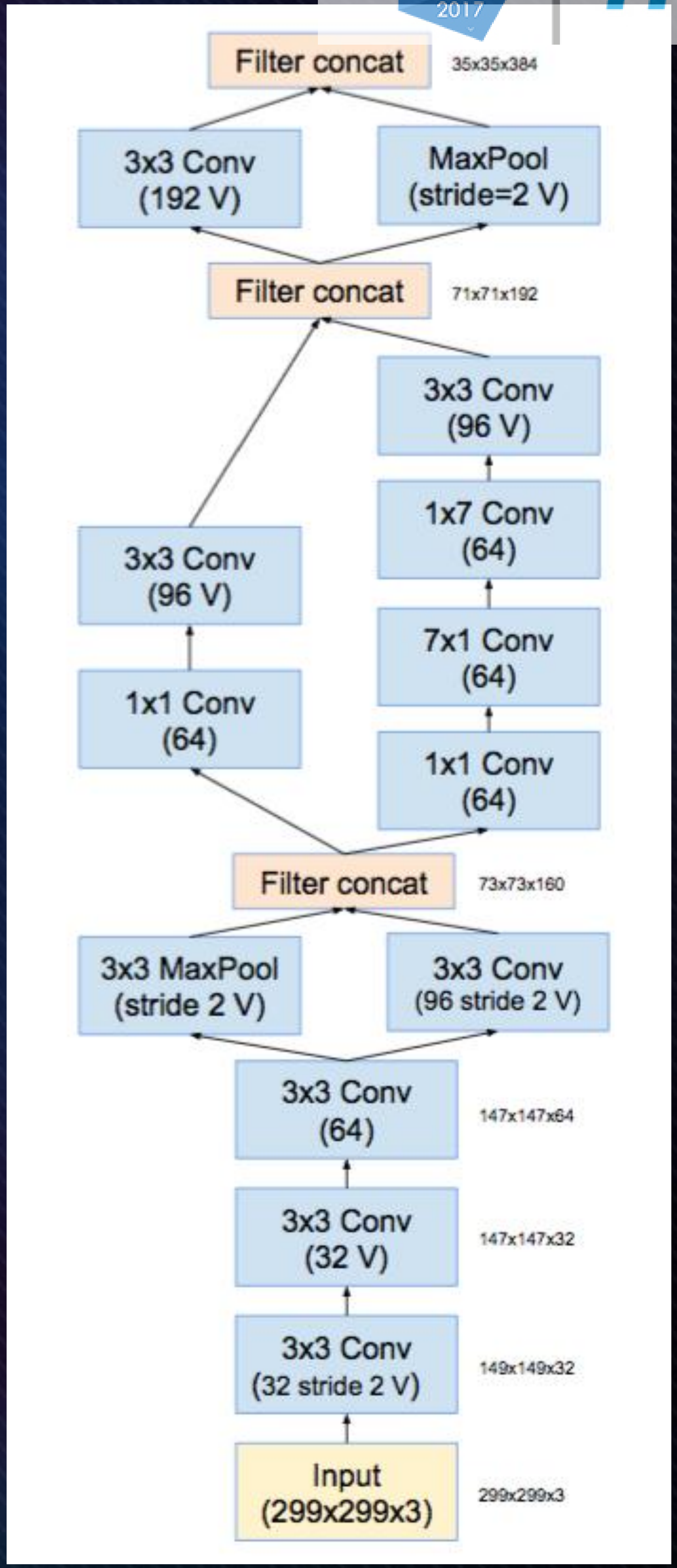
Inception-A



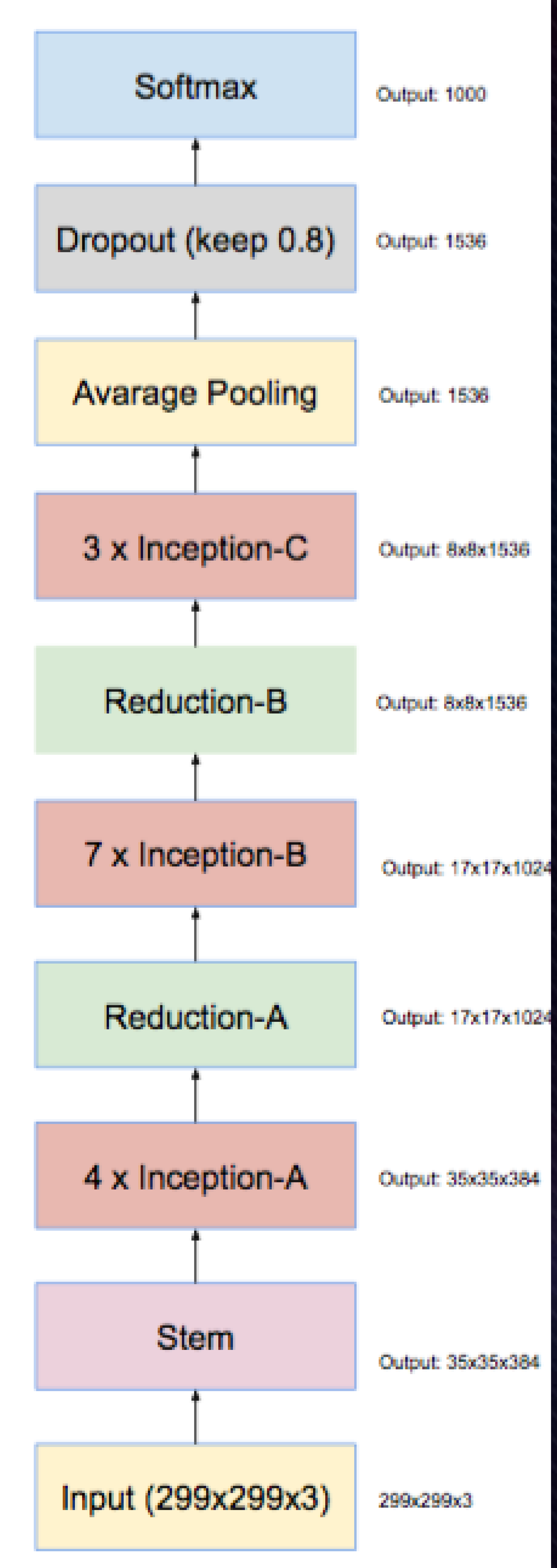


Inception-B

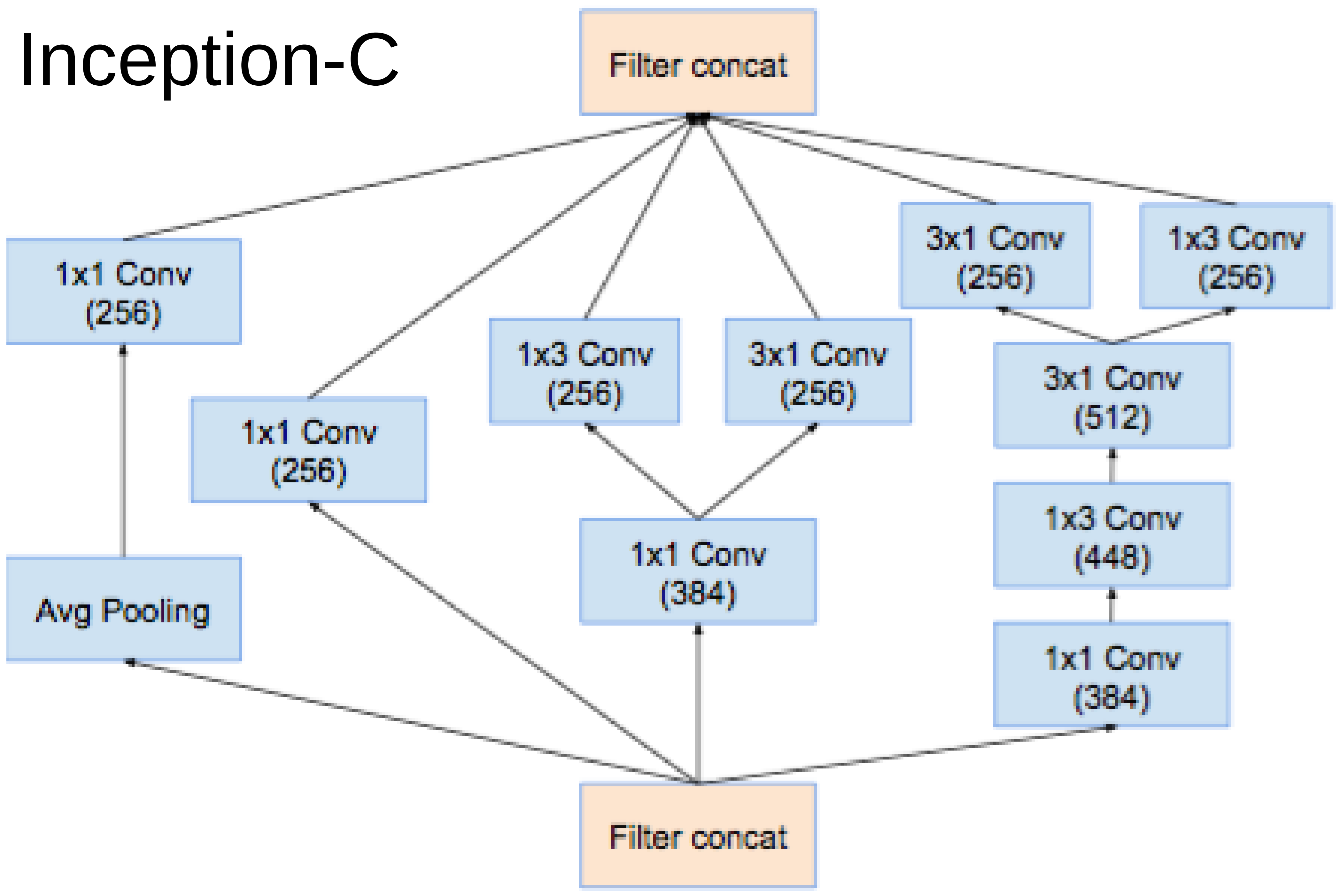
Stem



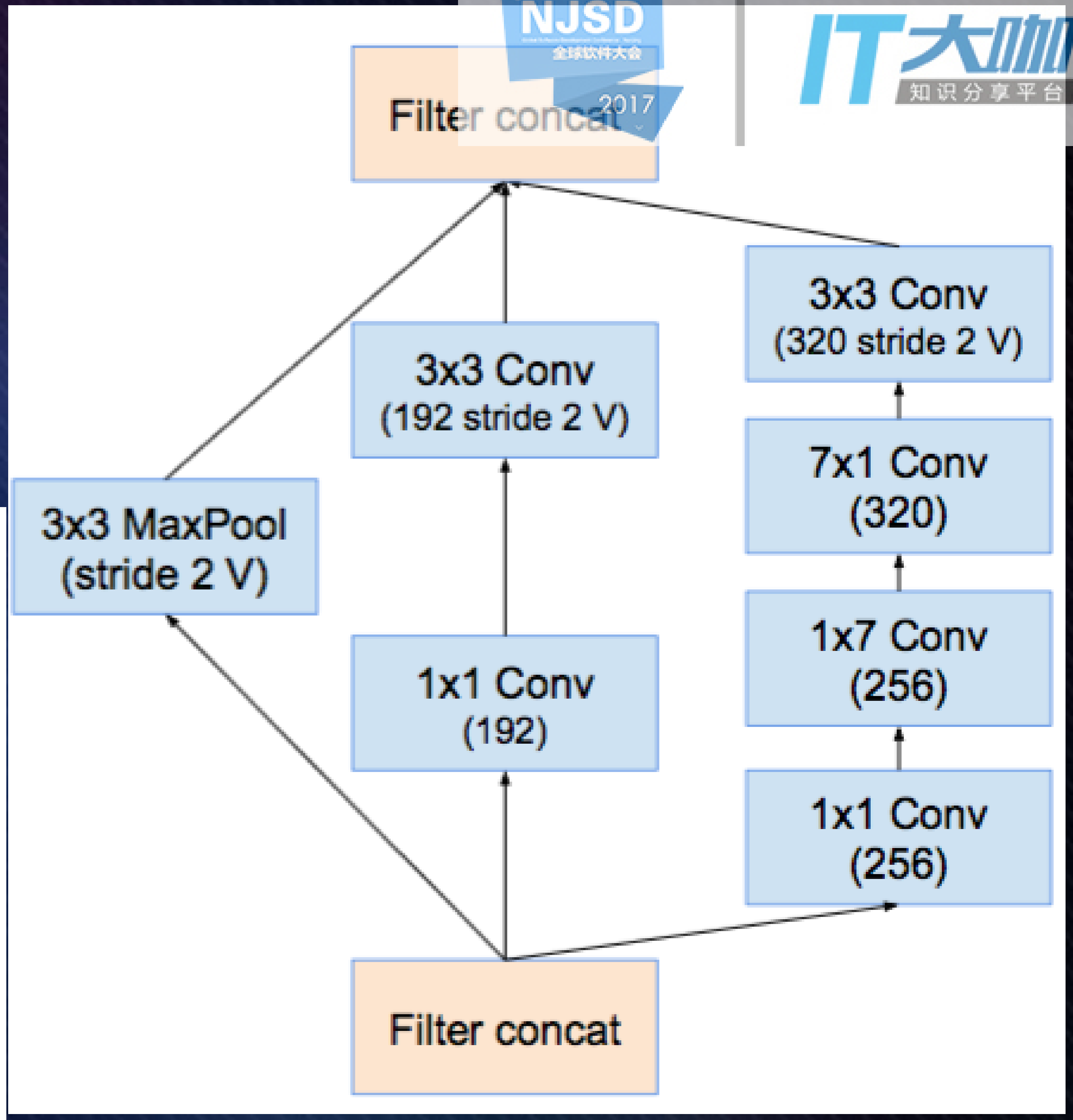
Inception v4



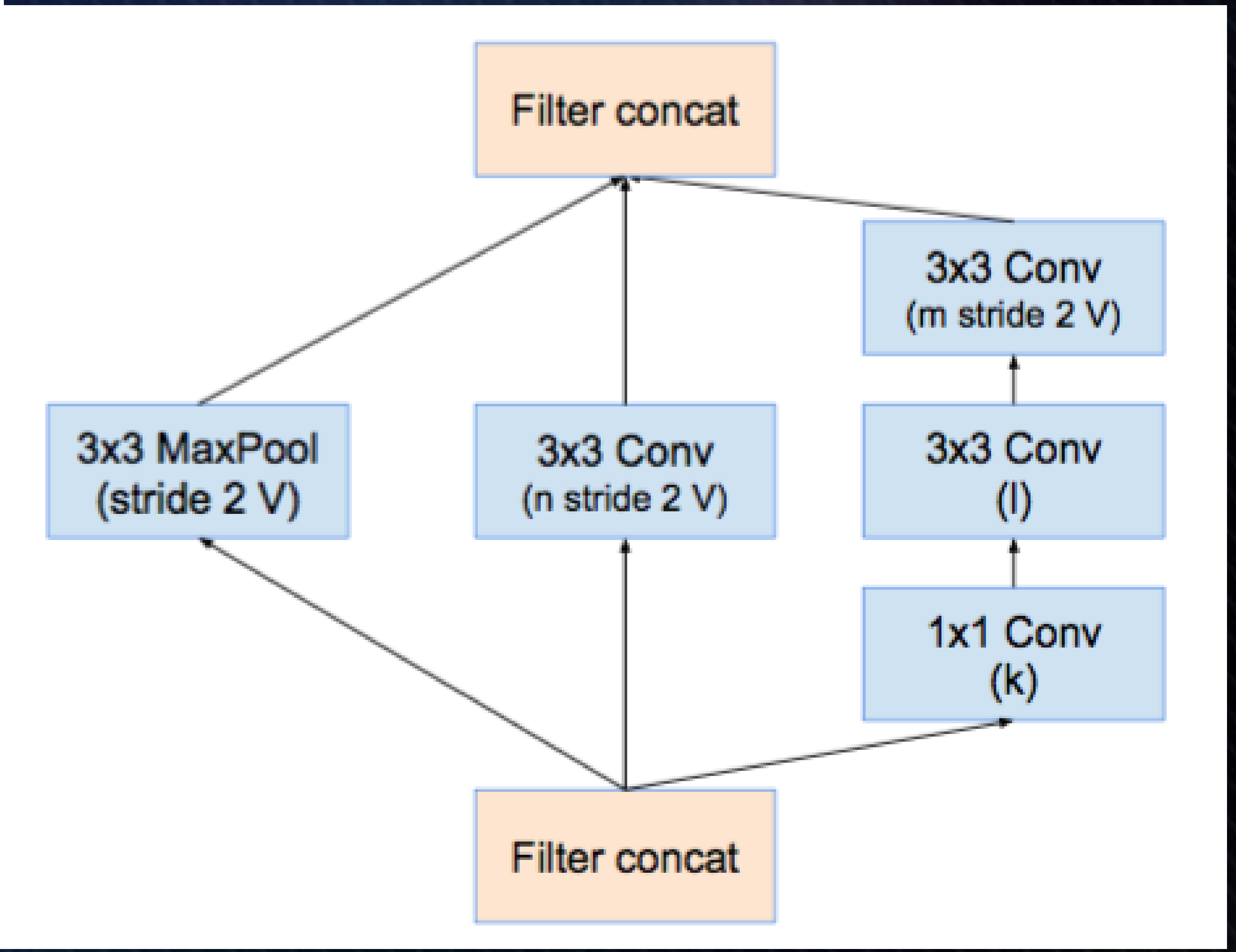
Inception-C



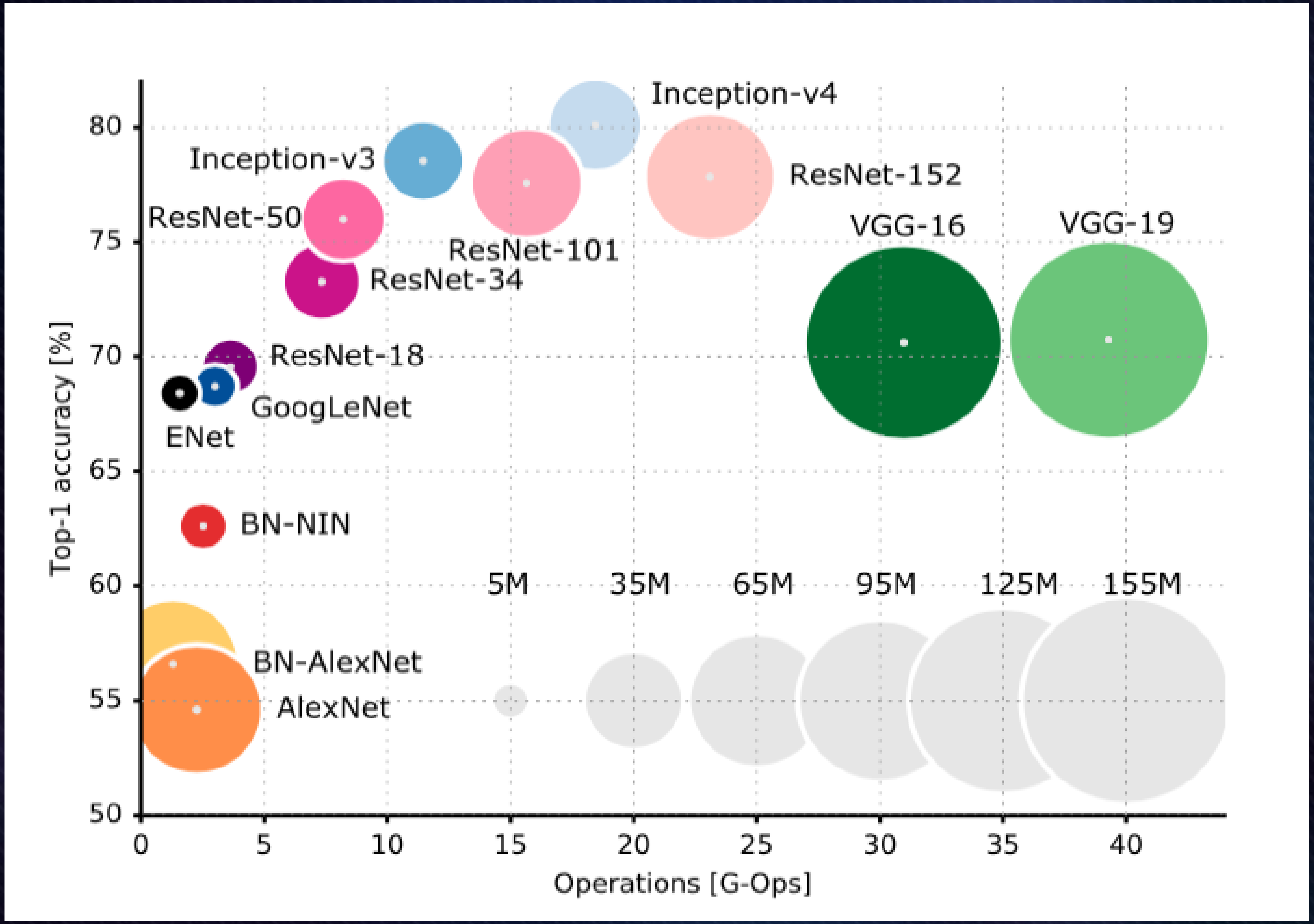
Reduction-A



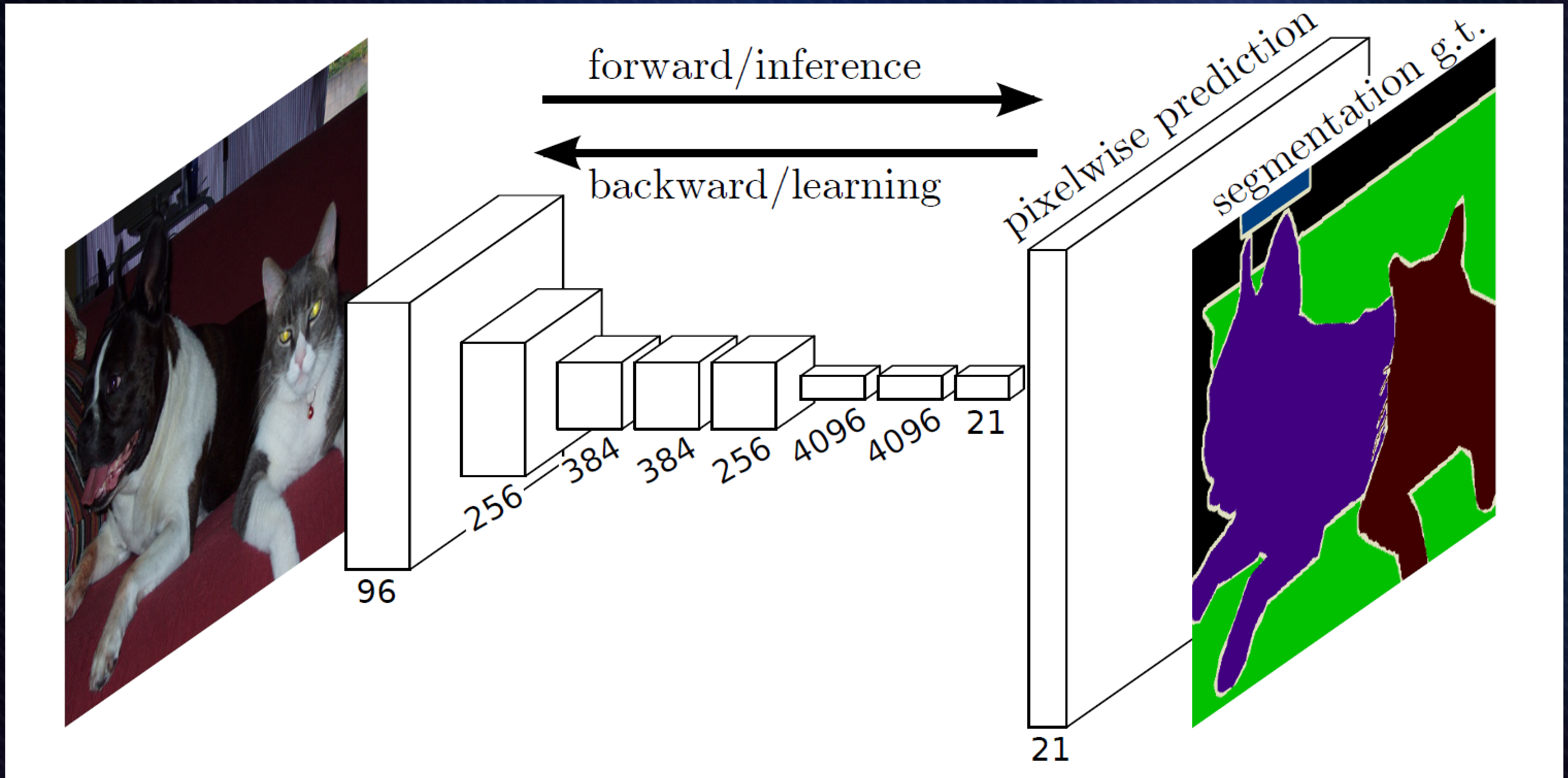
Reduction-B



Performances, Size and Operations

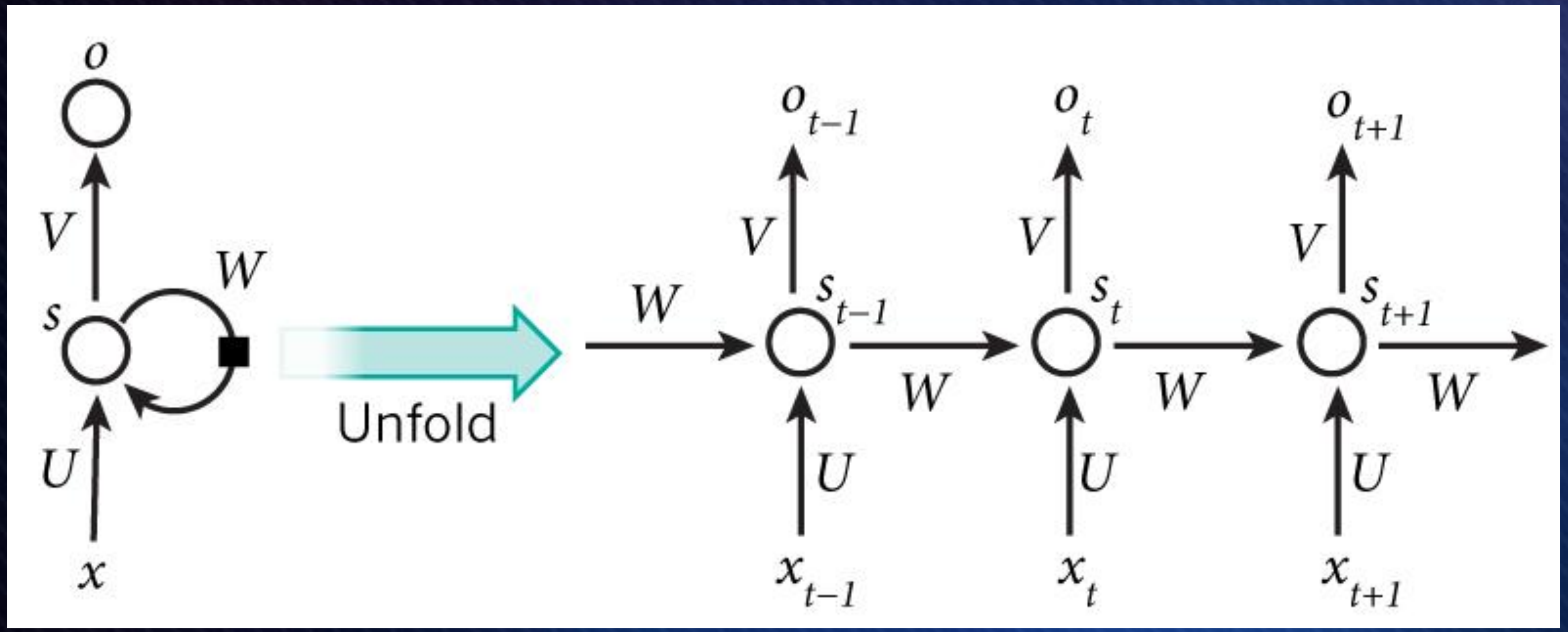


Fully convolutional network

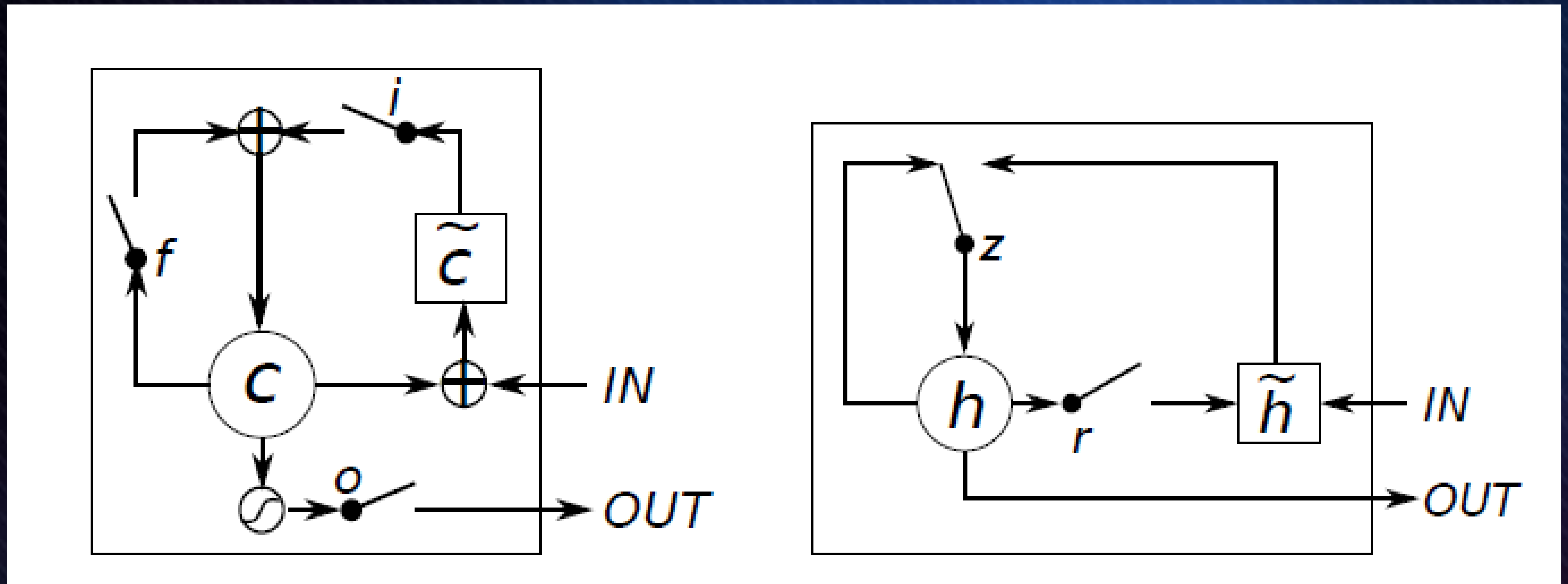


Fully convolutional networks for semantic segmentation, E. Shelhamer, J. Long, and T. Darrell, 2016

From recurrent neural network to LSTM and GRU



LSTM



(a) Long Short-Term Memory
(b) Gated Recurrent Unit

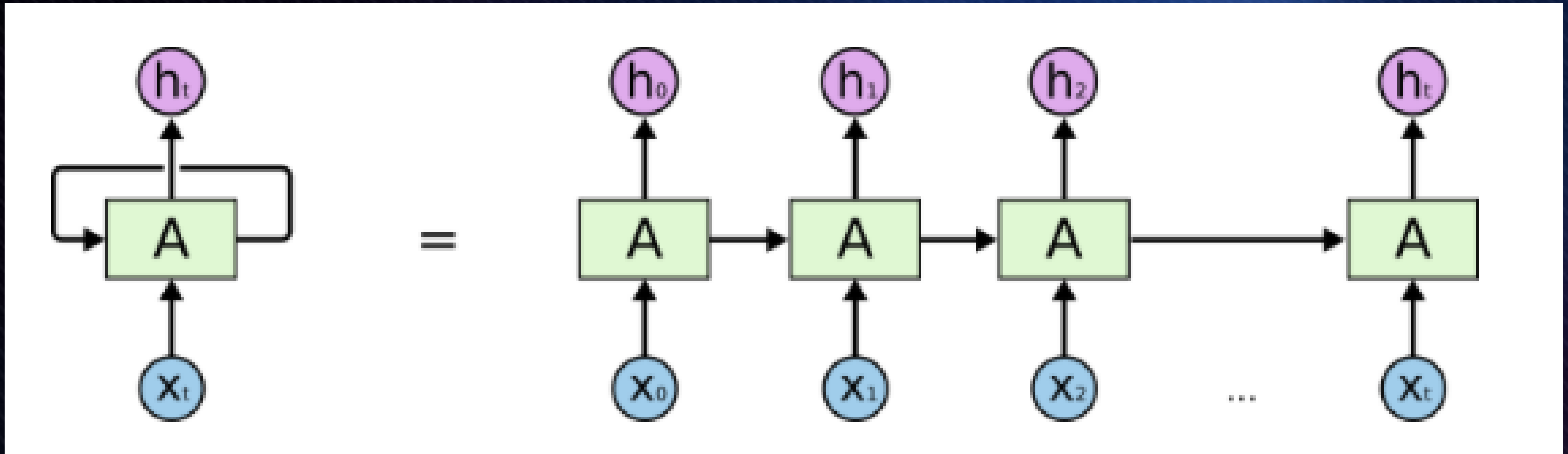
Gated Recurrent Unit

Long short-term memory, S. Hochreiter and J. Schimidhuber, 1999

Learning phrase representations using RNN encoder-decoder for statistical machine translation, K.Cho, etc

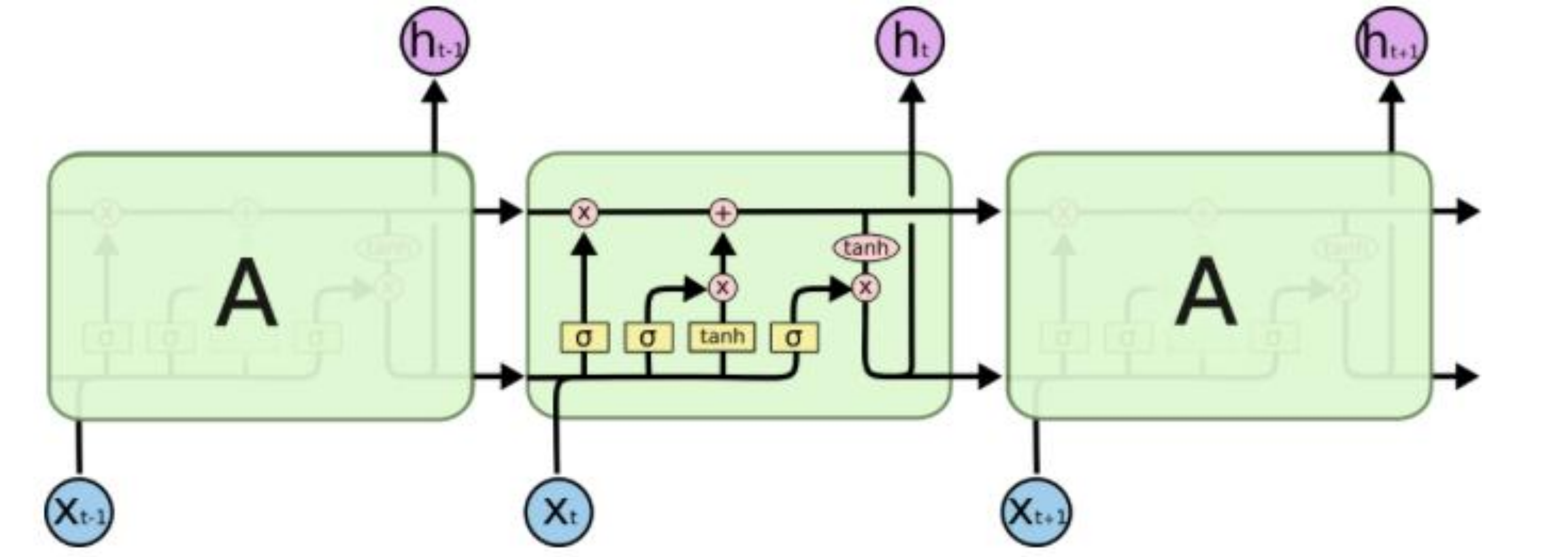
Recurrent Neural Networks

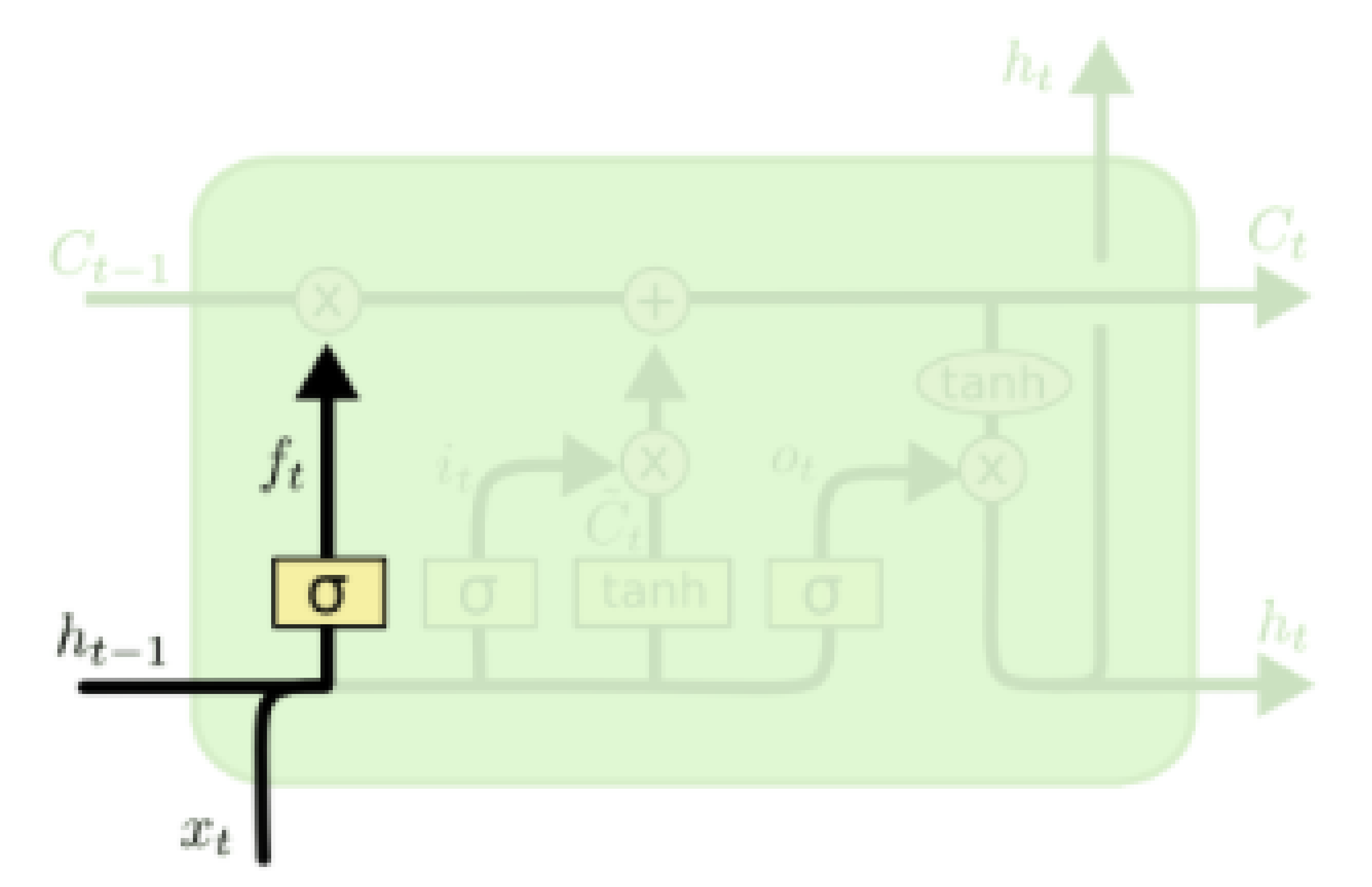
- Training RNNs used to be extremely difficult
  - An Unrolled RNN is equivalent to a very deep net with tiled weights
  - derivatives are susceptible to vanishing or exploding



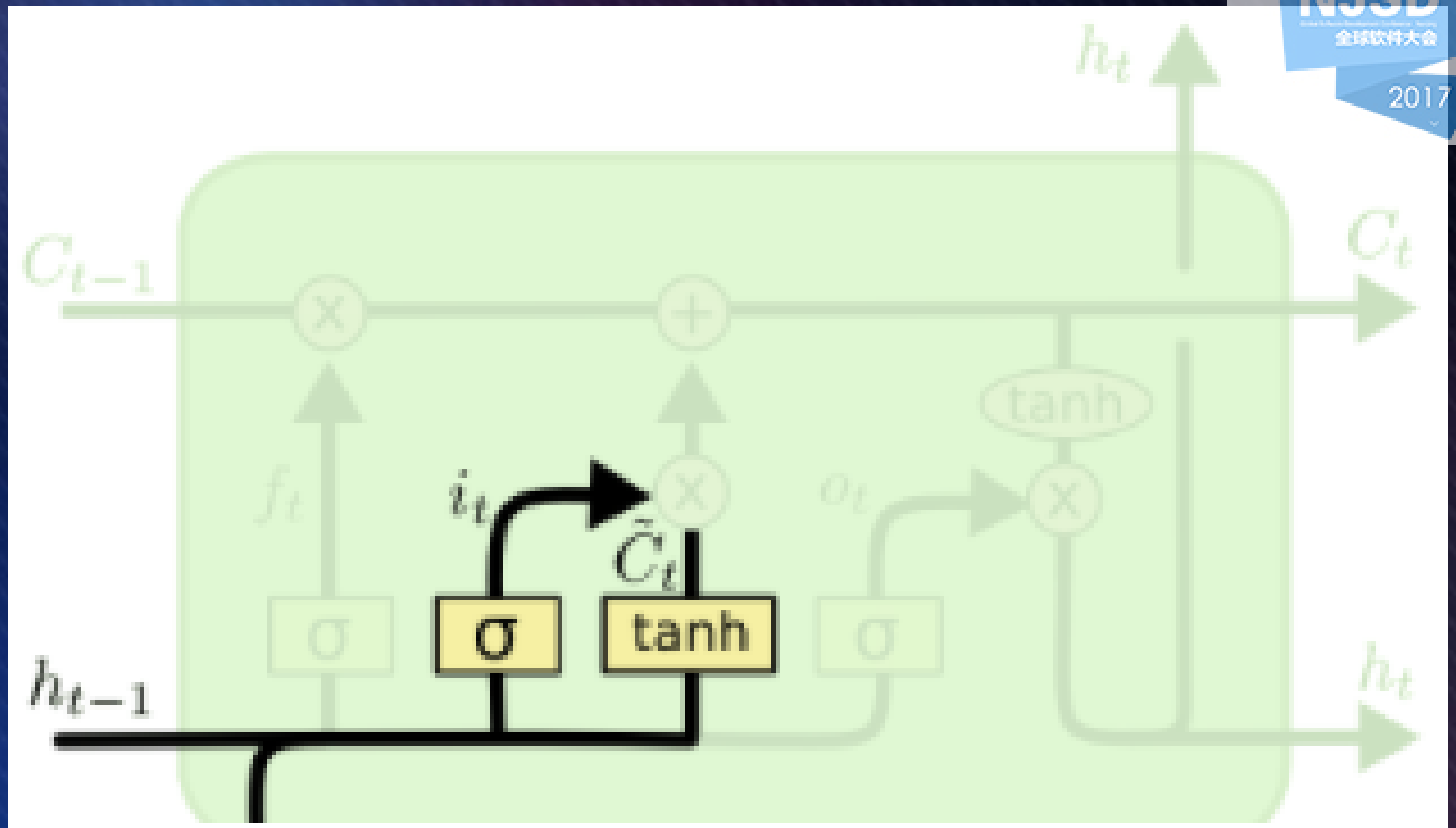
Long Short-Term Memory Units (LSTMs)

- A special kind of RNN have a better memory



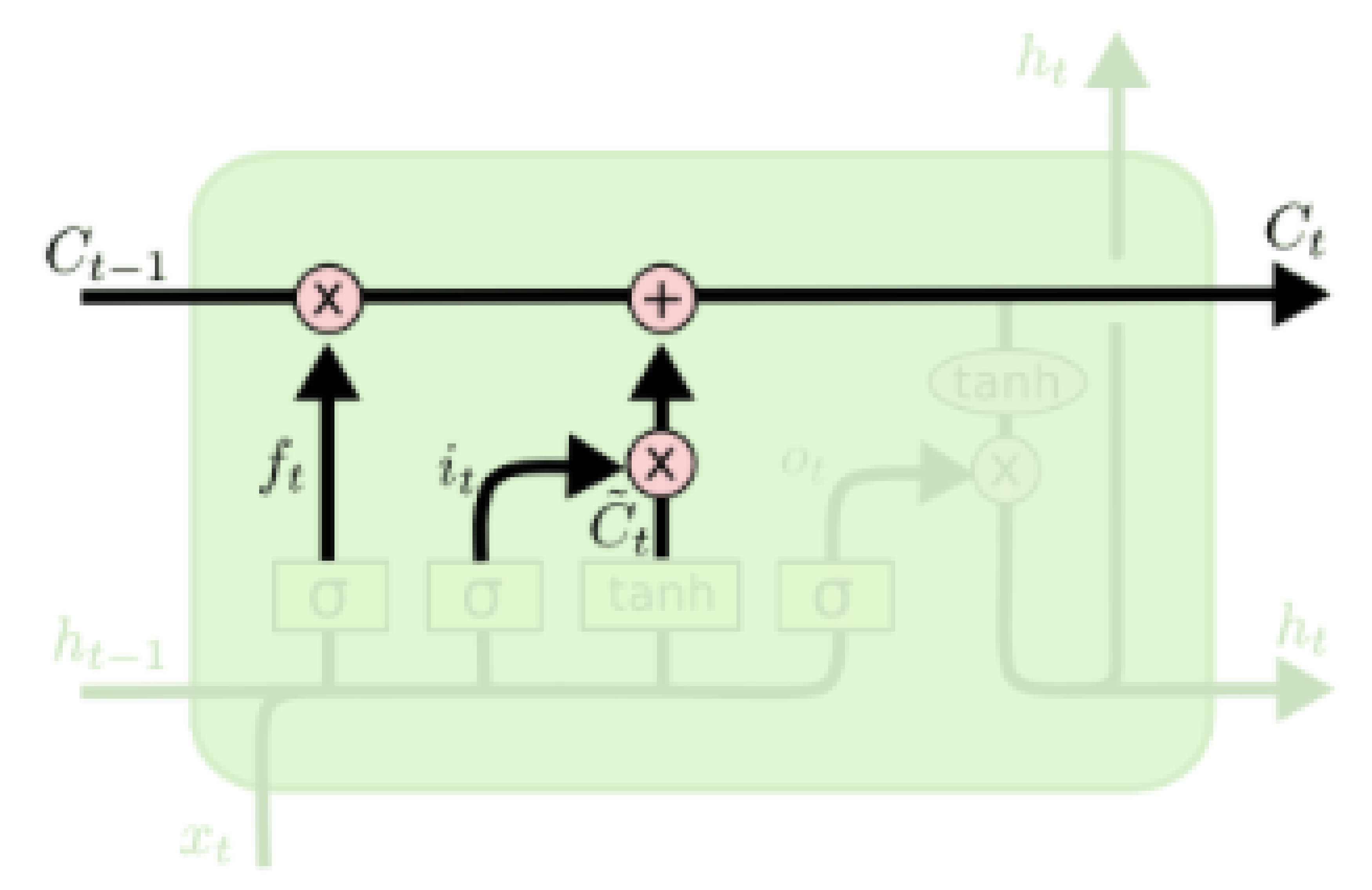


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

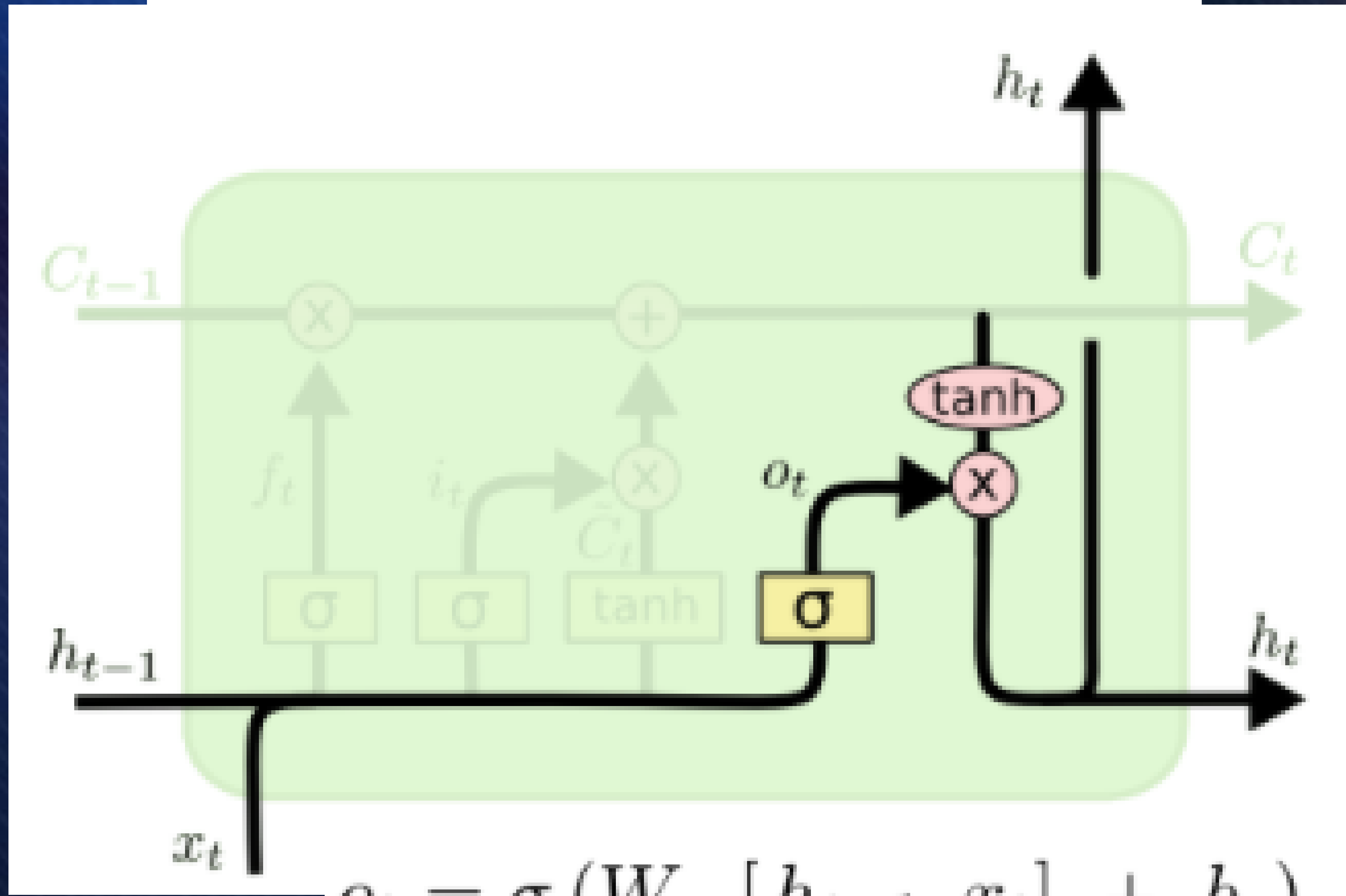


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$



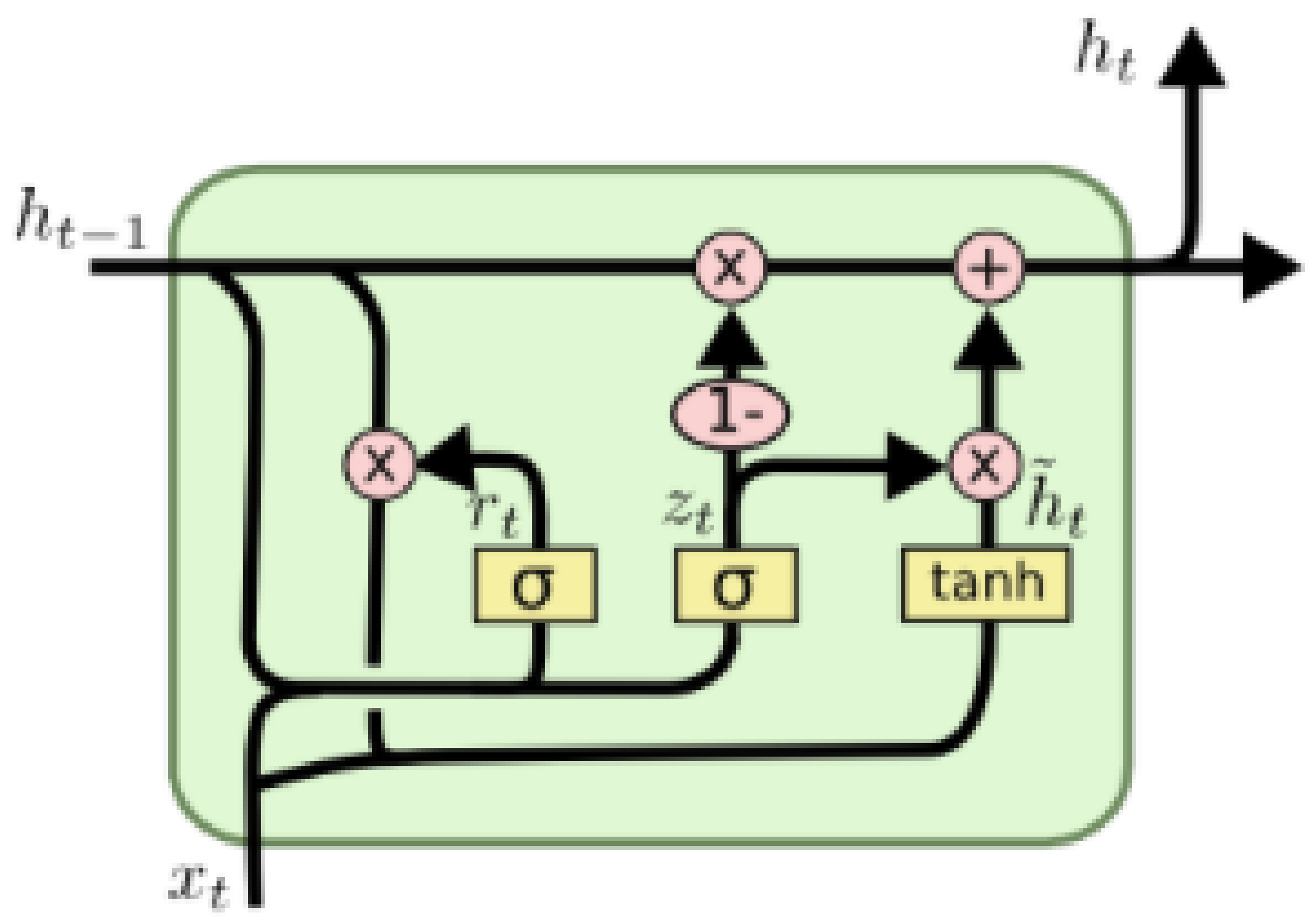
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

Gated Recurrent Unit



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

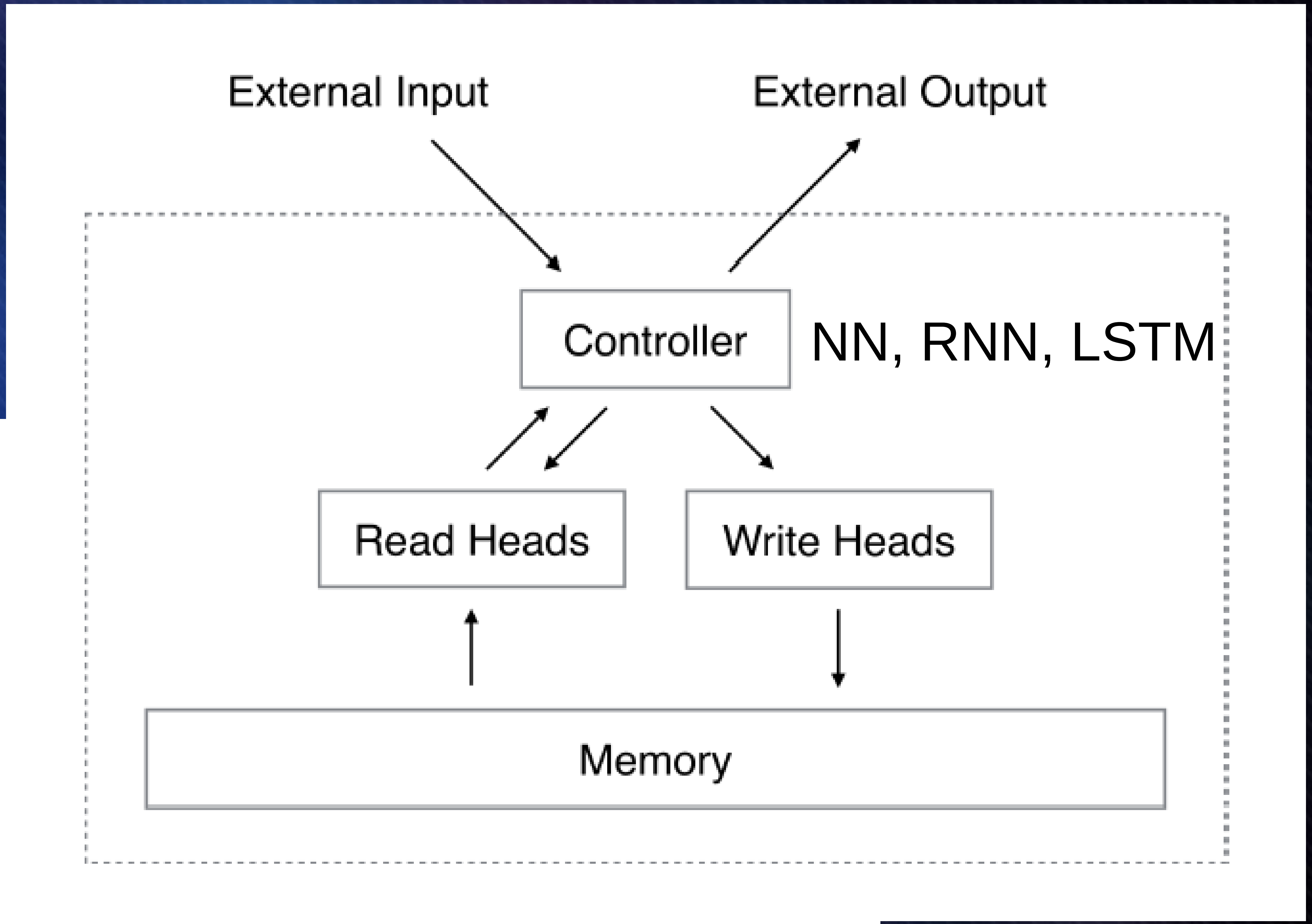
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

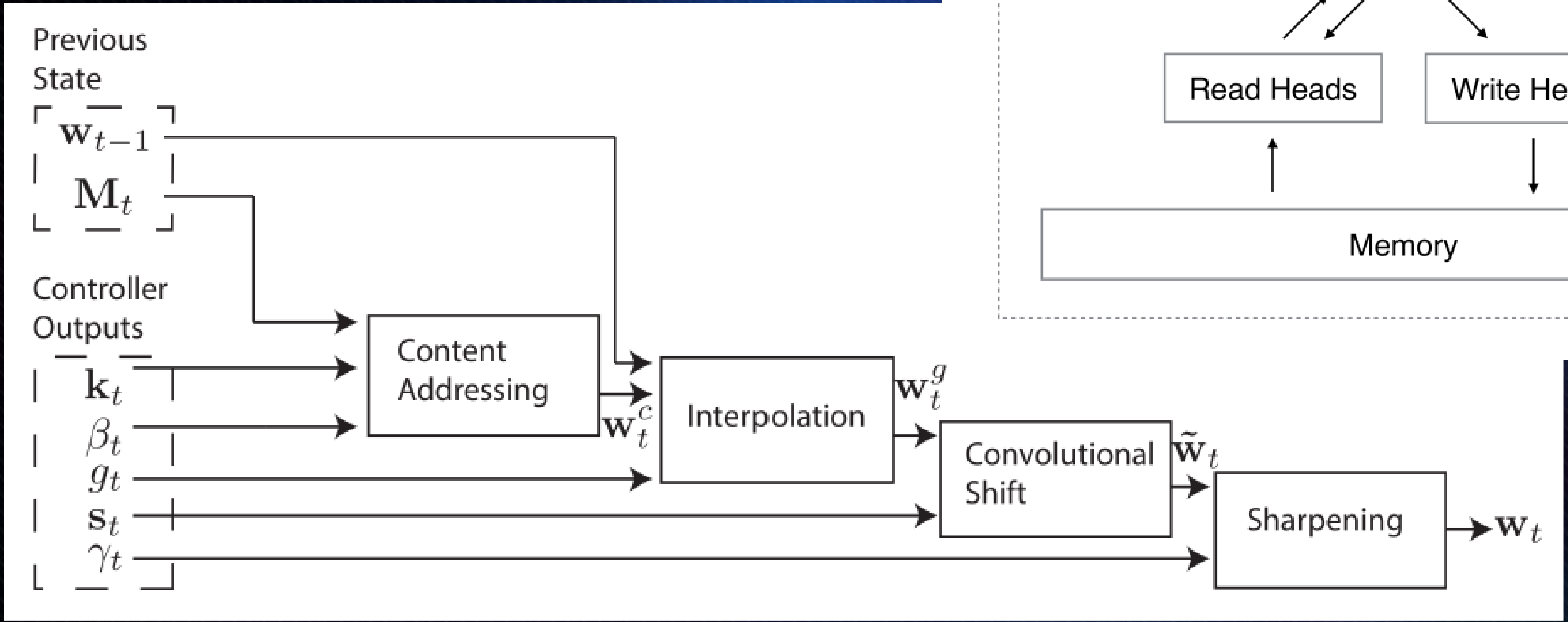
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Neural Turing Machine

Tape-like memory

Learn to copy, sort, etc.

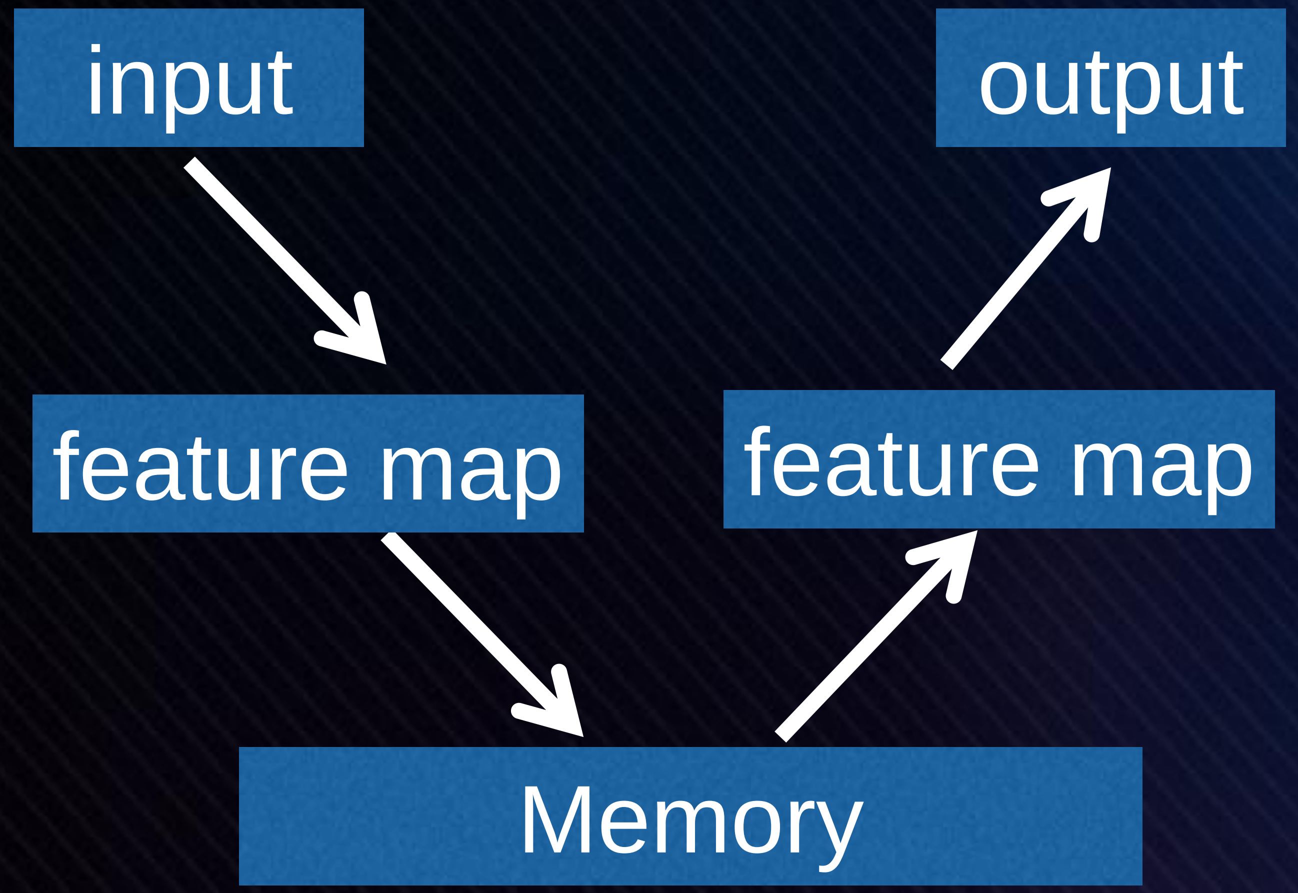


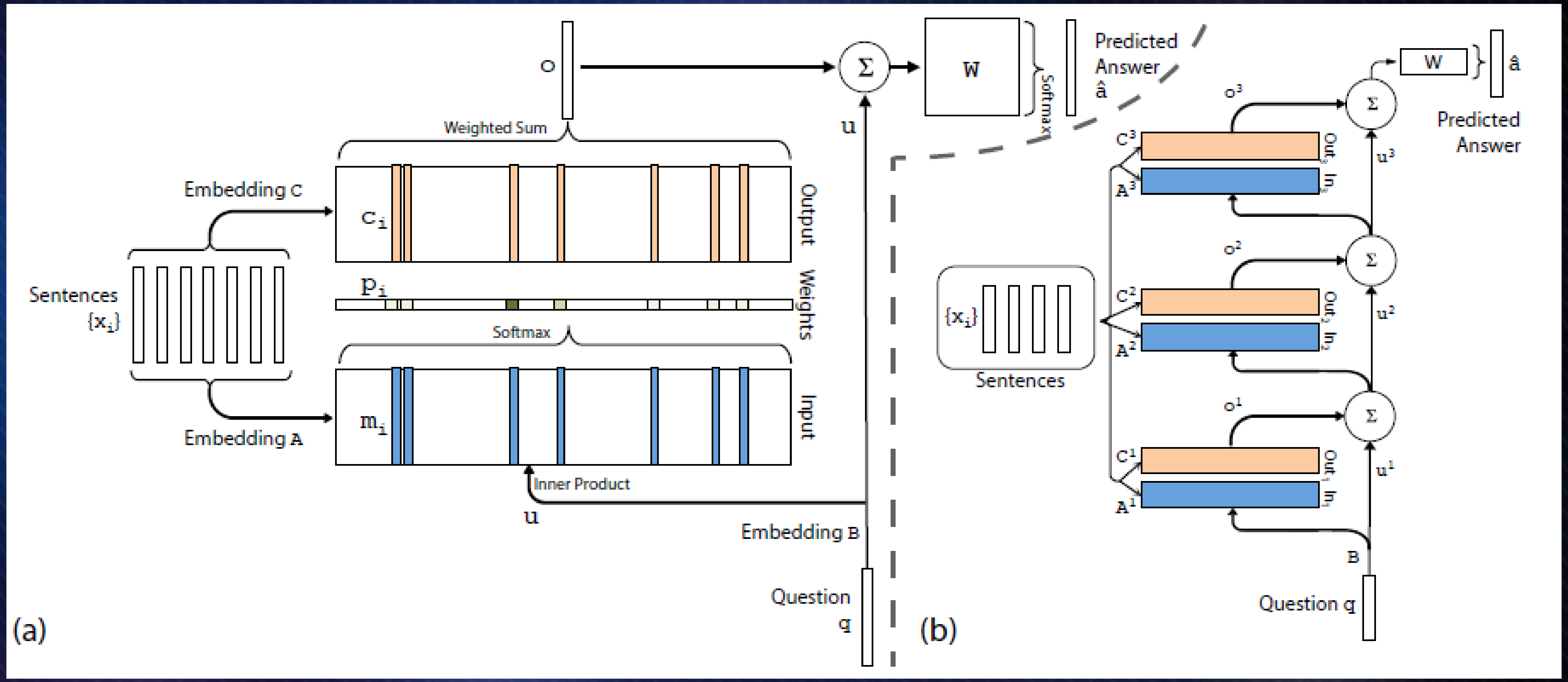


Neural turing machine, A. Grave, G. Wayne, I. Danihelka, 2014

Memory Networks

Associative long-term memory with reasoning





Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen

Memory networks, J. Weston, S. Chopra, A. Bordes, 2014
End-to-End memory networks, S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, 2015

Fixed point neural networks

- Inference of NNs works well on fixed point number
  - 8-bit networks converted from pre-trained float nets without sacrificing accuracy
  - Training from scratch will enable extreme low bit (1-bit) precision networks
- Fixed point are hardware friendly
  - Reduce model size and increase power efficiency
  - Speed-up inference

Binarized neural networks: training neural networks with weights and activations constrained to +1 or -1, 2016

Binary neural networks

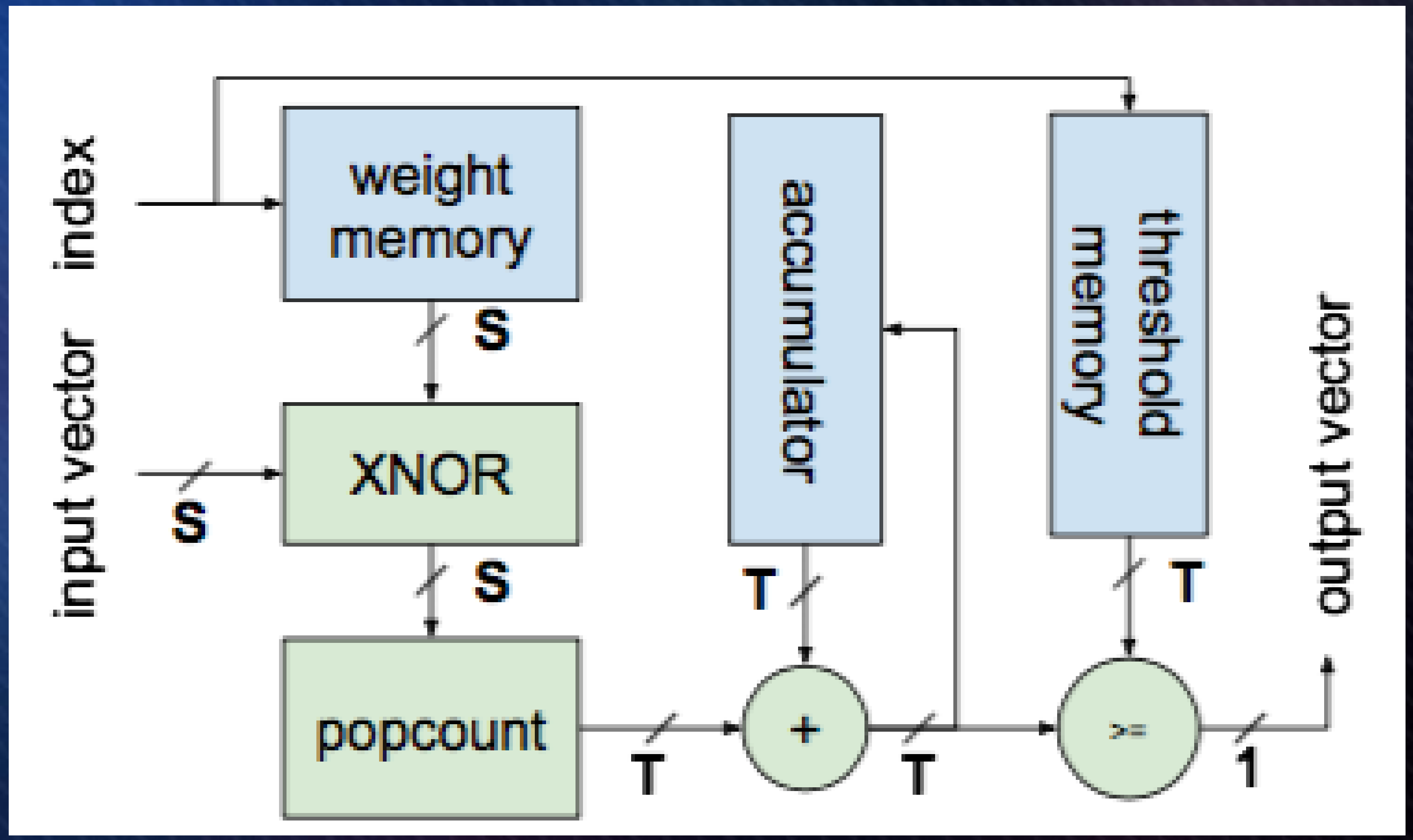
- Binary weights and activations
  - Enable working on large images and keep power efficiency
  - Replace MACC with XNOR and bit-count
- Less unique filters allow further optimization
- BNNs achieved promising results on small datasets

	Main arithmetic operation	2D Unique filters	Memory Saving (inference)	Accuracy on Cifar10
Float Neural Nets	float MAC	100%	1x	89.1%
Binary Neural Nets	XNOR-bitcount	~40%	~32x	88.8%

Binary neural networks

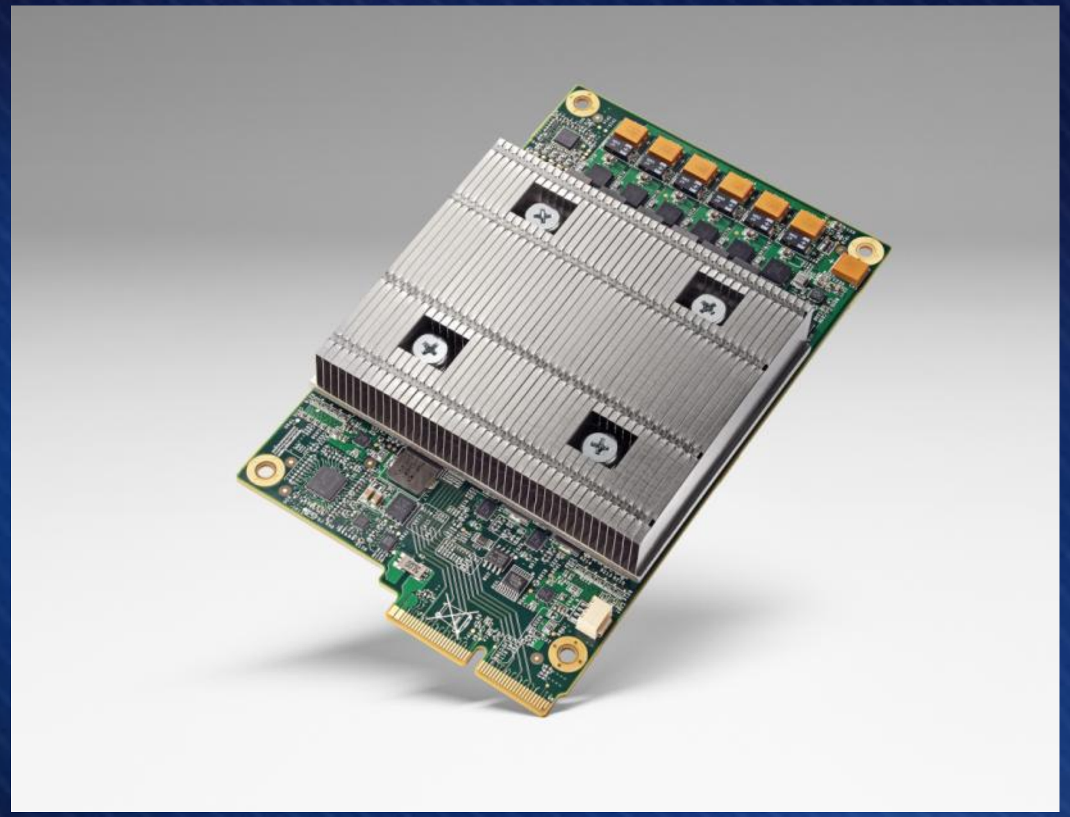
- Some tricks for training:
  - Put batch norm everywhere
  - Avoid network bottle-neck
  - Be patient with the slow converge
- Improve BNNs performances by
  - using more parameters
  - allowing extra high bit layer
  - working with other high bit networks

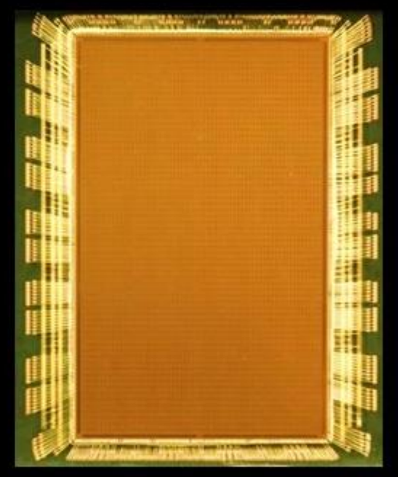
Typical convolutional layer in BNNs
MACC -> XNOR-popcount
Batch norm and binarization merged into a threshold operation



Umuroglu et. al. “FINN: A Framework for Fast, Scalable Binarized Neural Network Inference”

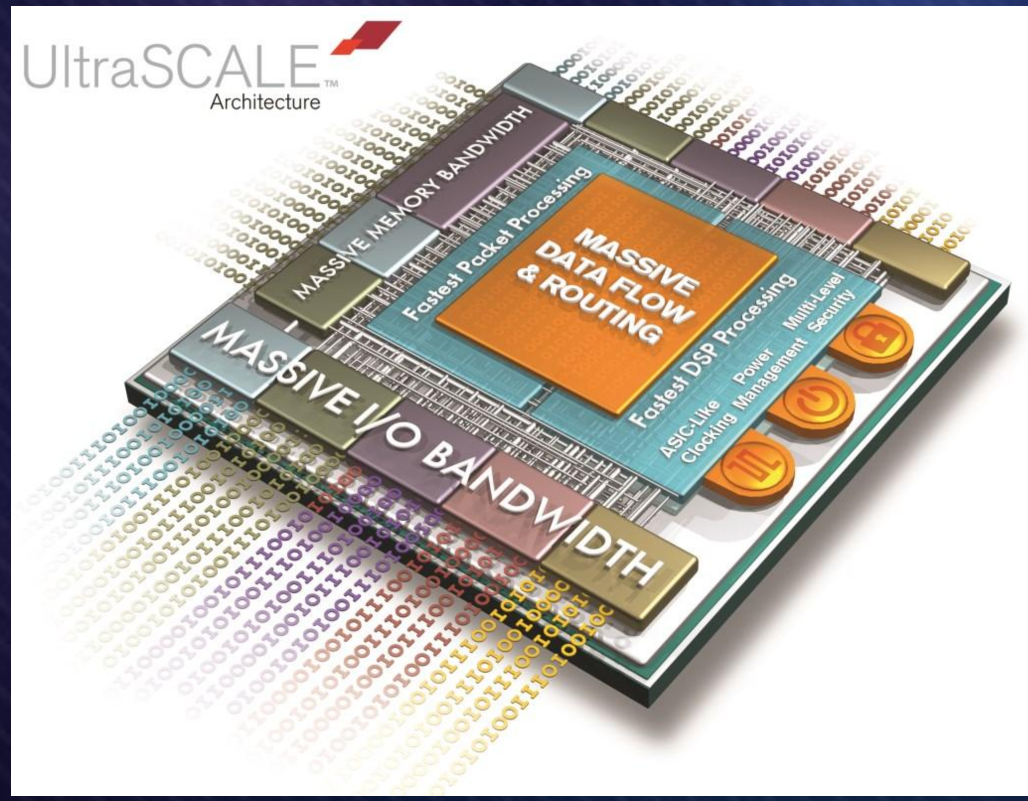
Deep neural networks on silicon

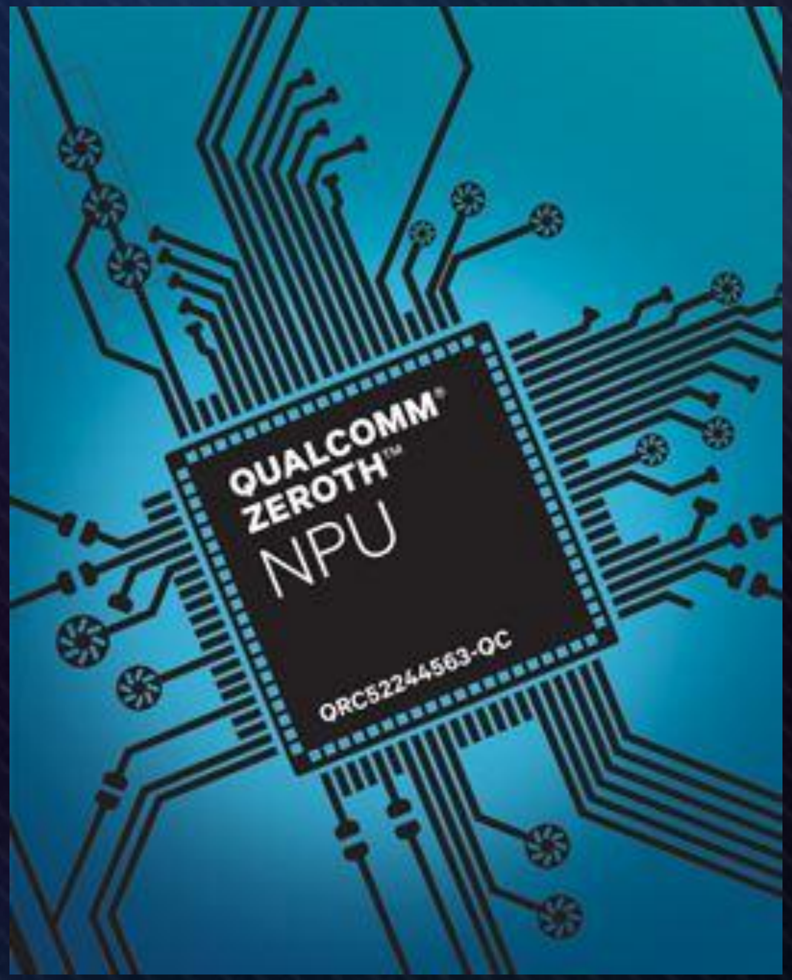




TrueNorth

1 M Neurons
256 M Synapses
5.4 B Transistors
Realtime
73 mW





The Mission of Horizon Robotics

Create the “brain” platform of smart things, to make human life more convenient, safer, and more fun.

Horizon Robotics

Horizon Robotics

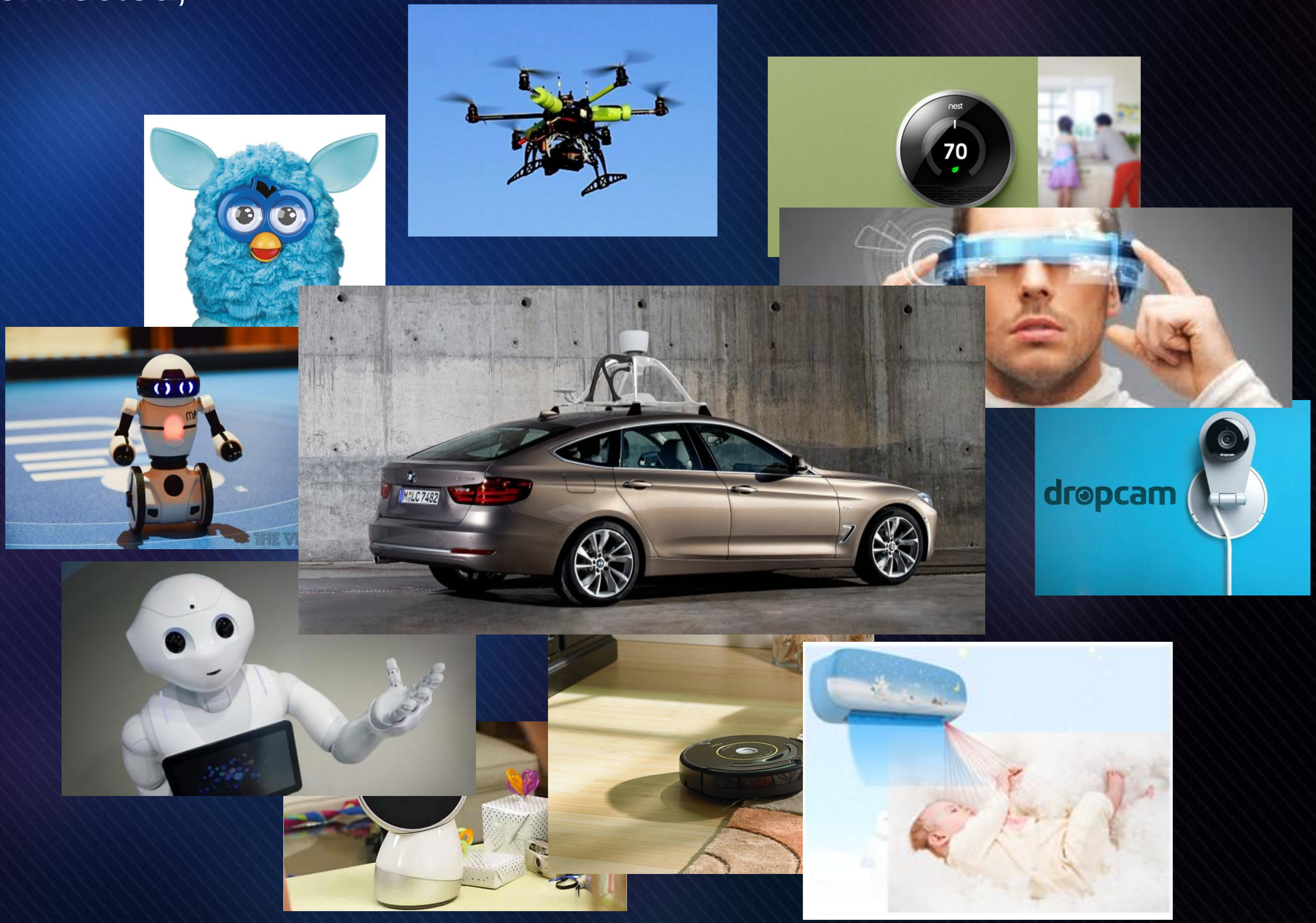
- Founded in July 14th, 2015, HQ in Beijing, and R&D site in Nanjing, an office in Shenzhen and Shanghai
- Experienced engineers from Baidu, Facebook, Google, Huawei, Nokia, Microsoft, TI, NVIDIA
- 40% have oversea experiences, 14% have PhD degrees, 100% are seasoned engineers
- Pioneers of many accomplishments in AI

Nanjing office

Horizon Robotics

IoT？ We think the future is Internet of smart things

- In the future, all the devices are not only connected, but more importantly, “AI inside”.
- sense the environment
- interact with people
- make control decisions
- A local brain on the device
- Perception & HCI
- Low-latency & real-time
- Low-power & low cost
- Privacy protection



Copied from Dr. Jiang Wang