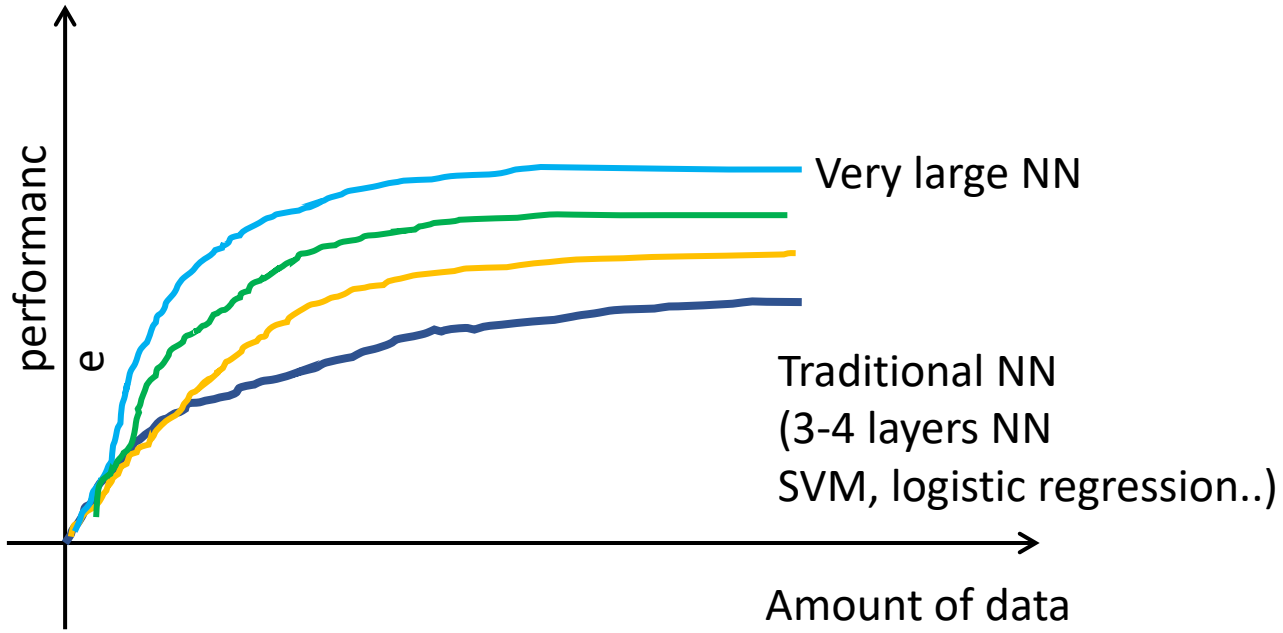
EEE511
Deep Neural Networks/
Deep Learning

What drives deep learning progress?

- The idea of neural computation dates back to the 1940s, but key breakthroughs were all in recent years.
- For a long time, neural networks were academic curiosities, have met with serious skepticism.
- Faster and cheaper computers + new ideas have finally led to rapid progress.



- Large data
- Large networks
- New algorithm capabilities (e.g., ReLu of AlexNet, dropout (Srivastava et al., 2014), attention-learnable pointer structure (Bahdanau et al., 2014), batch normalization: re-centering and re-scaling (Ioffe & Szegedy, 2015)

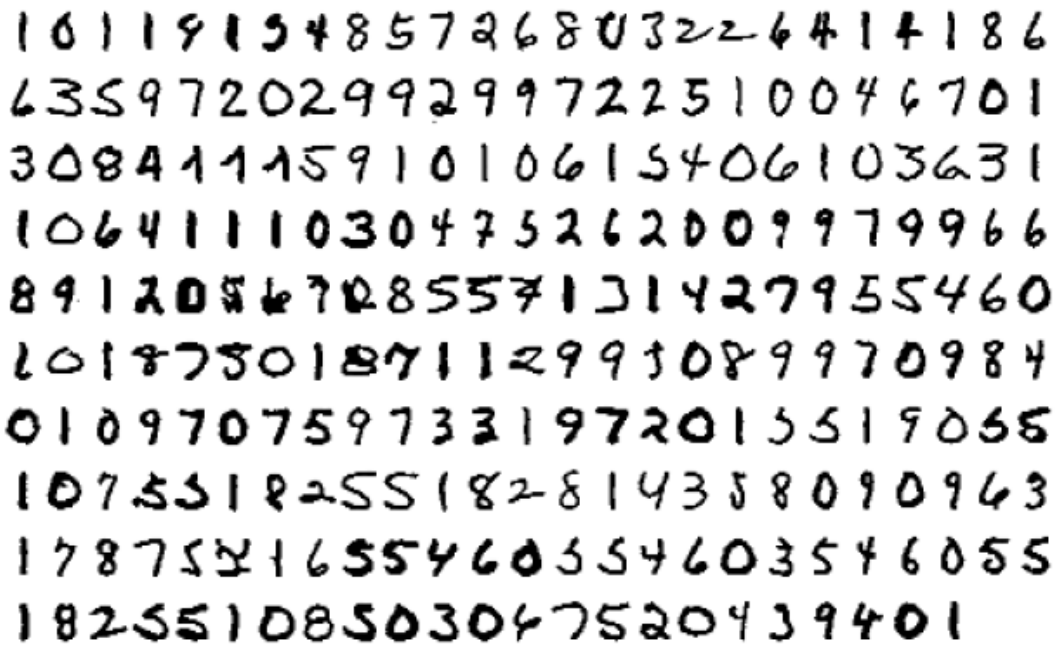
LeNet – A Classic CNN Architecture (for hand-written digit recognition) by Le Cun
(inspired by Neocognitron by Fukushima (1980)

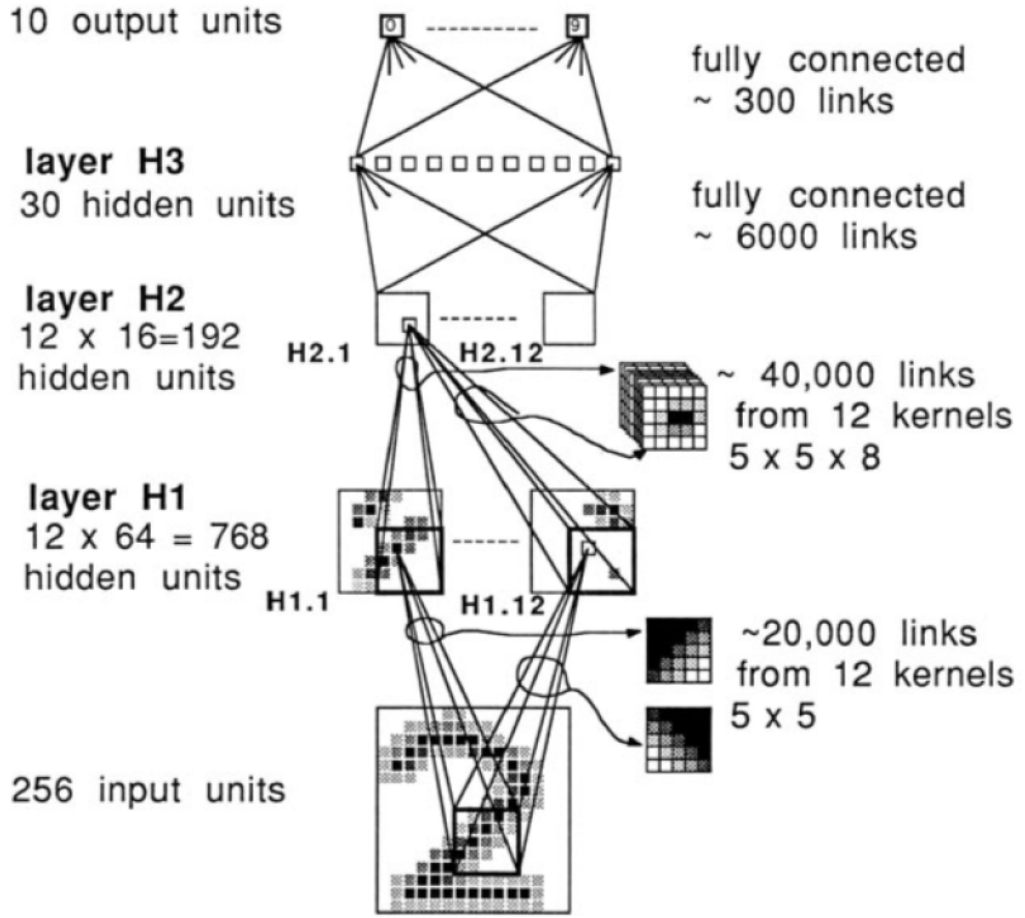
LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition." Neural computation 1, no. 4 (1989): 541-551.

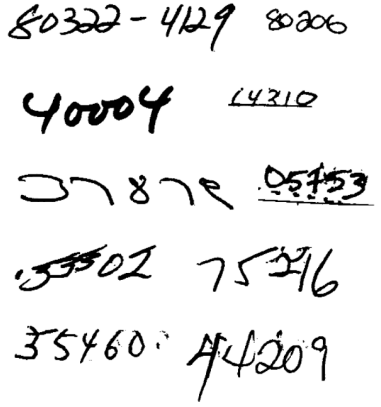
Backpropagation Applied to Handwritten Zip Code Recognition

Y. LeCun
B. Boser
J. S. Denker
D. Henderson
R. E. Howard
W. Hubbard
L. D. Jackel
AT&T Bell Laboratories, Holmdel, NJ 07733 USA

The ability of learning networks to generalize can be greatly enhanced by providing constraints from the task domain. This paper demonstrates how such constraints can be integrated into a backpropagation network through the architecture of the network. This approach has been successfully applied to the recognition of handwritten zip code digits provided by the U.S. Postal Service. A single network learns the entire recognition operation, going from the normalized image of the character to the final classification.







Deep learning driven/enabled by needs associated with large datasets

As of 2022:

- 2.5 quintillion bytes of data is created every day.
- 5 billion Snapchat videos and photos are shared per day.
- 333.2 billion emails are sent per day.
- Google, Facebook, Microsoft, and Amazon store at least 1,200 petabytes of information.
- The world spends almost $1 million per minute on commodities on the Internet.
- 20% of people online watch online games…

Megabyte, Gigabyte, Terabyte, Petabyte, Exabyte (Amazon Redshift data warehouse), Zettabyte (~44 ZB globally early 2020, Seagate 1st 3-ZB storage shipped in 2021), Yottabyte - the largest unit approved as a standard size by the International System of Units (SI)

Deep learning – driven/enabled by ever increasing computing power

Decade	Dataset	Memory	Floating point cal/s	
2010		10 G (advertising)	1 GB	1 TF (Nvidia C2050)
2020		1 T (social network)	100 GB	1 PF (Nvidia DGX-2)

Tensor Processing Unit (TPU) of google - an ASIC accelerator for NN ML based on TensorFlow software, used in AlphaGo versus Lee Sedol man-machine Go games, AlphaZero system (Chess, Shogi and Go playing programs), in Google Photos, a TPU can process over 100 million photos a day.

Single GPU vs large # of (~1024) GPUs – handling ~ 64000 observations in minibatches of ~64, ResNet-50 training time ~7 minutes, reported in 2018 (instead of multiple days) on ImageNet dataset.

Broad Application Domains of Deep learning

- Natural language processing
- Computer vision
- Drug discovery
- Financial engineering
- Customer management
- Robotics
- Self-driving car

and many more…

Deep learning performs well at computer vision tasks

- Biometric data (face) scan for access to smartphone
- Taking photos once detecting smiling faces
- Tesla self-driving cars
- Facebook facial and image recognition
- Spotting defects and irregularities on the assembly line
- Developing software to allow insurance companies to process and categorize photographs of claims automatically
- Categorizing millions of online images and hours of video
- Covid detection using scans

Deep learning performs well at computer vision tasks

- Prerequisites on large dataset, generalization by training on expanded dataset via image augmentations (believed to have benefited AlexNet)
- Computer vision applications leveraging layer-wise representations - features in one layer are combined in many different ways to create more abstract features in the next layer
- Increasingly more abstract representations of (visual) information into deeper layers or of a compositional nature, effectively represent images in multiple levels
- Transfer learning possible (copy source model to target model + additional tuning) – pretrained on source model, transferred to target model (assuming knowledge in source applicable in target, output layer may be different)

Winning DNN architectures for ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

Coding: local coordinate, super-vector

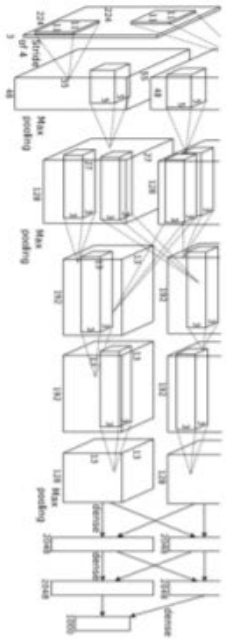
Pooling, SPM

Linear SVM

[Lin CVPR 2011]

Year 2012

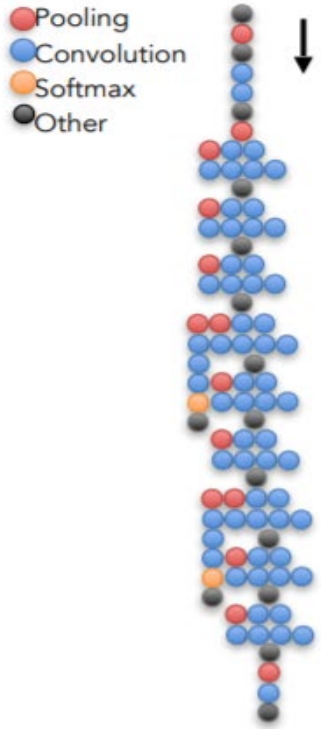
SuperVision



[Krizhevsky NIPS 2012]

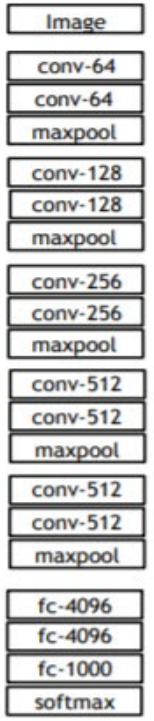
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

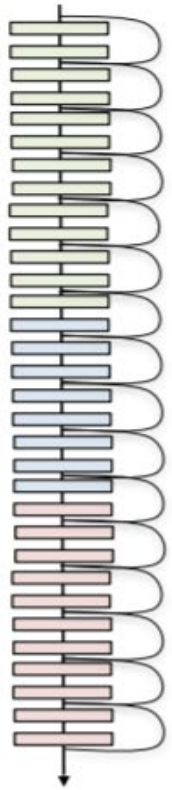
VGG



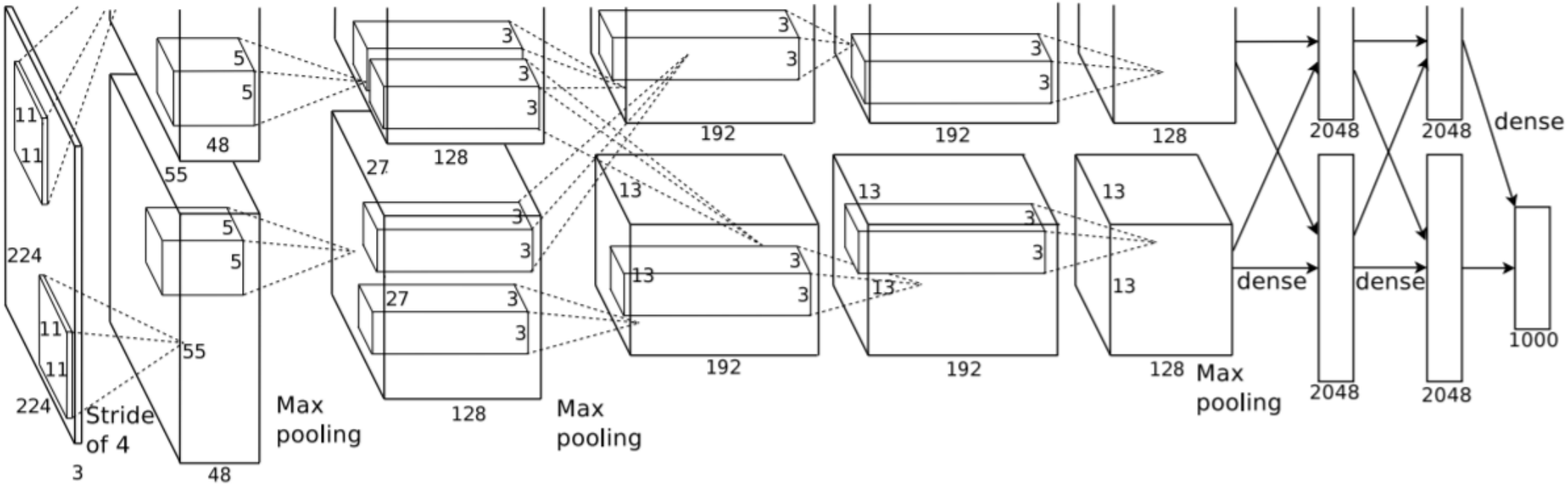
[Simonyan arxiv 2014]

Year 2015

MSRA

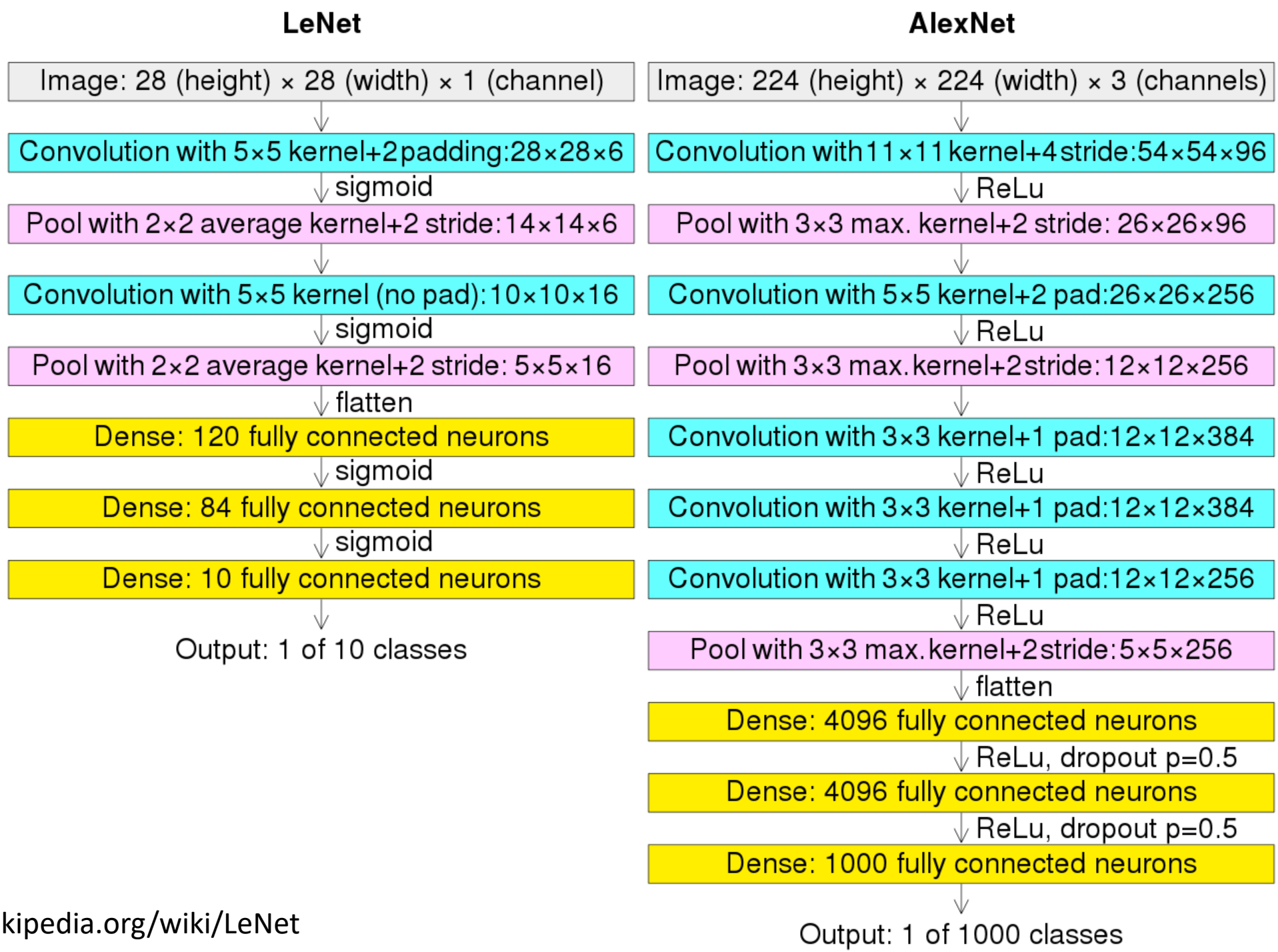


[He ICCV 2015]



AlexNet architecture of CNN, implemented on two GPUs (1 for the top track another for the bottom track. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440-186,624-64,896-64,896-43,264-4,096-4,096-1000.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).



https://en.wikipedia.org/wiki/LeNet

Acclaimed Deep Learning Models

- VGG
- GoogleNet
- ResNet
- DenseNet
- BERT
- GPT
- …

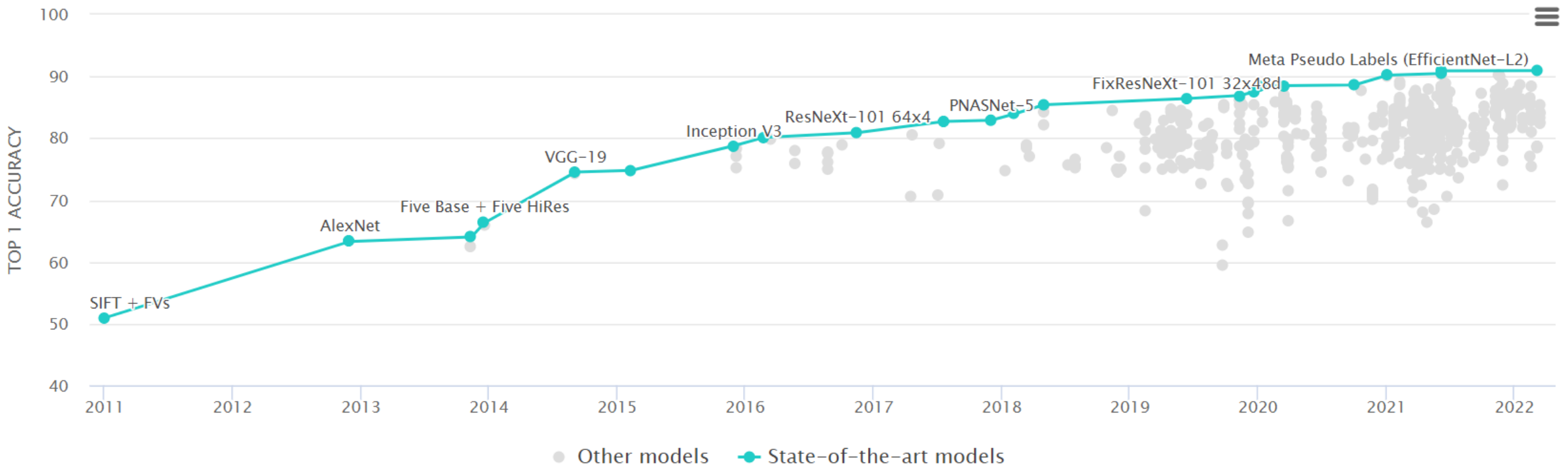
And fast developing… e.g., GPT-2 released in Feb 2019 with 1.5B parameters, GPT-3 released in June 2020 with 175B parameters…

Paperswithcode.com

Image Classification on ImageNet

Leaderboard Dataset

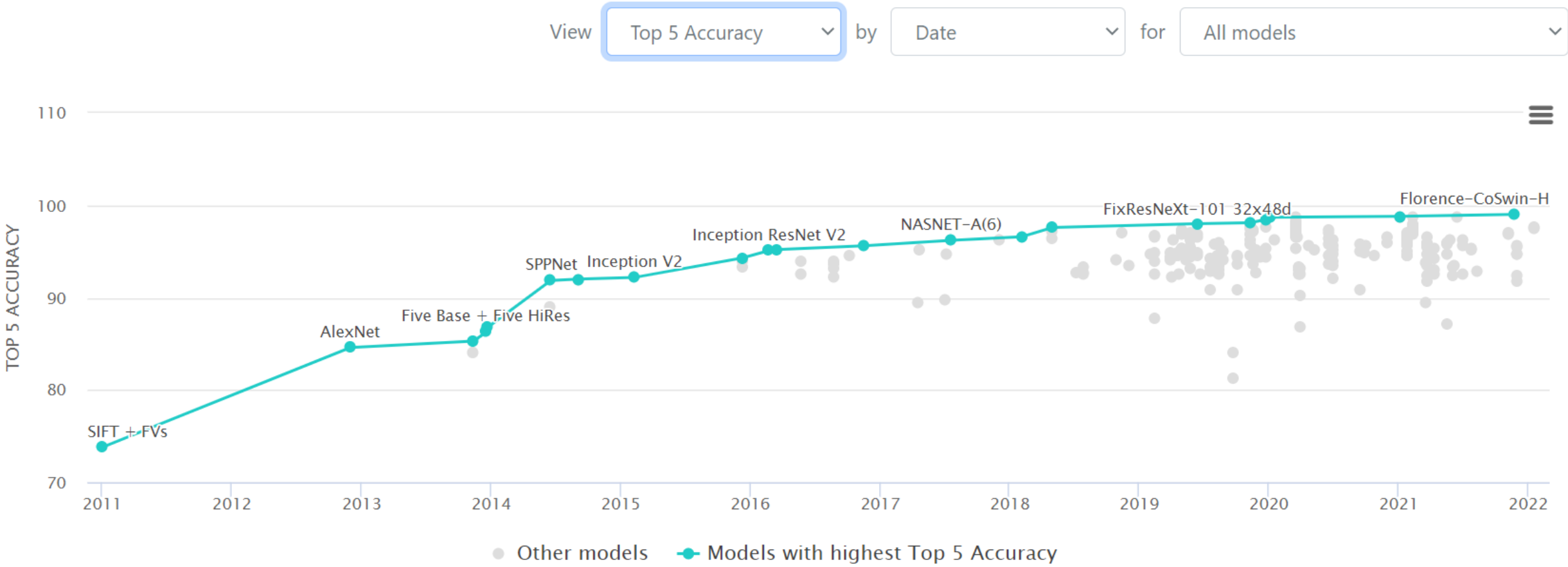
View Top 1 Accuracy by Date for All models



Paperswithcode.com

Image Classification on ImageNet

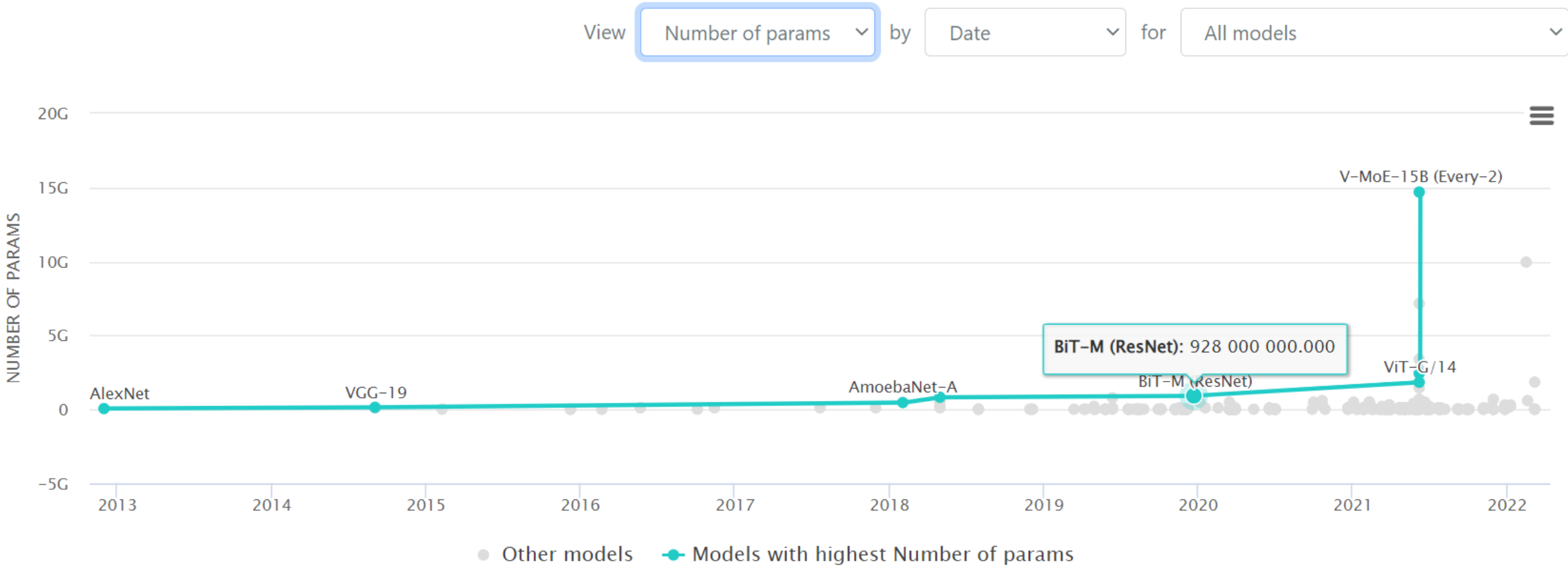
Leaderboard Dataset



Paperswithcode.com

Image Classification on ImageNet

Leaderboard Dataset



Theory

Deep neural networks (DNN) as universal approximators

- The classic universal approximation theorem concerns a single hidden layer of large yet finite size for approximating continuous functions
- The universal approximation theorem for DNN concerns the capacity of networks with bounded width but growing depth

Title	Results
The Expressive Power of Neural Networks: A View from the Width (2017)	Any Lebesgue-integrable function f from R^n to R can be approximated by a fully-connected width-(n + 4) ReLU network to arbitrary accuracy with respect to L_1 distance. For integer k, there exist a class of width-O(k^2) and depth-2 ReLU networks that cannot be approximated by any width-O($k^{1.5}$) and depth-k networks.
The Power of Depth for Feedforward Neural Networks (2016)	The existence of a 3-layer network, which cannot be realized by any 2-layer to more than a constant accuracy if the size is subexponential in the dimension.
Deep Network Approximation for Smooth Functions (2020)	ReLU forward neural networks (FNN) with width O(NlnN) and depth O(LlnL) can approximate functions in the unit ball of $C^s([0,1]^d)$ with approximation rate O($N^{-2s/d}L^{-2s/d}$).
Why deep neural networks for function approximation (2017)	In order to approximate a function which is Θ(log$(1/\varepsilon)$)-order derivable with ε error universally, a deep network with O(log$(1/\varepsilon)$) layers and O(poly log$(1/\varepsilon)$) weights can do but Ω(poly $1/\varepsilon$) weights will be required if there is only o(log $1/\varepsilon$) layers.
Error bounds for approximations with deep ReLU networks (2017)	C^n-functions on R^d with a bounded domain can be approximated with ε error universally by a ReLU network with O(log$(\frac{1}{\varepsilon})$) layers and O($(\frac{1}{\varepsilon})^{d/n}log(\frac{1}{\varepsilon})$)weights.
Representation Benefits of Deep Feedforward Networks (2015)	For any integer k explicitly constructed networks with O(k^3) layers and constant width which cannot be realized by any network with O(k) layers whose size is smaller than 2^k.
On the Expressive Power of Deep Learning: A Tensor Analysis (2016)	proved the existence of classes of deep convolutional ReLU networks that cannot be realized by shallow ones if its size is no more than an exponential bound.
Optimal Approximation of Continuous Functions by Very Deep ReLU Networks (2018)	We prove that constant-width fully-connected networks of depth provide the fastest possible approximation rate that cannot be achieved with less deep networks
Exponential Convergence of the Deep Neural Network Approximation for Analytic Functions (2018)	Deep ReLU networks can approximate functions of d variables as well as linear approximation by algebraic polynomials with a comparable number of parameters.
Deep Network Approximation Characterized by Number of Neurons (2020)	It is shown by construction that ReLU FNNs with width $C_1 \max\{d\lfloor N^{\frac{1}{d}}, N+1\rfloor\}$ and depth $12L + C_2$can approximate an arbitrary Holder continuous function of order α with a constant λ on $[0,1]^d$ with a nearly tight approximation rate $19\sqrt{d}\lambda N^{-2\alpha/d}L^{-2\alpha/d}$measured in L_p-norm for any given $N, L \in N^+$ and $p \in [1,\infty]$.

Theory – probabilistic interpretation

- The theory derives from machine learning field
- Probabilistic interpretation considers the activation nonlinearity as a CDF, it leads to the idea of dropout as a regularization approach in neural networks. Refer to work by Hopfield, Widrow, Narendra.
- Inference based on optimization concepts of training and testing, related to fitting and generalization.

DNN design - hyperparameters

Architecture: what layers and how (CNN, pooling, RNN, FC, …)
# of layers
# of nodes per layer
Activation function

Learning rate
# of iterations
Stopping criteria
Minibatch

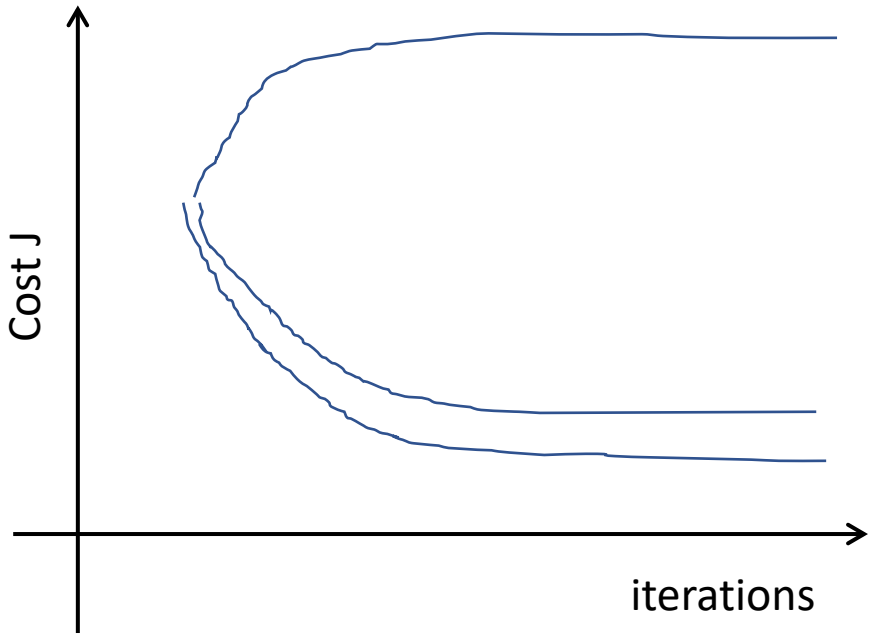
Momentum
Regularization

Transfer learning

It can take days to test a single idea

Designing a deep neural network is an empirical process

Use learning curves to help diagnose!



Datasets for ML research

- Kaggle
- Github
- ImageNet (14 mil images, 20 k categories)
- Microsoft Common Objects in Context (COCO)
- Open Images (~9 mil images, diverse complex scene)
- SIFT10M Dataset (UCI)
- Fashion-MNIST (small)
- KITTI Vision Benchmark (Registry of Open Data on AWS)
- FieldSAFE (obstacle detection, agriculture)

More…

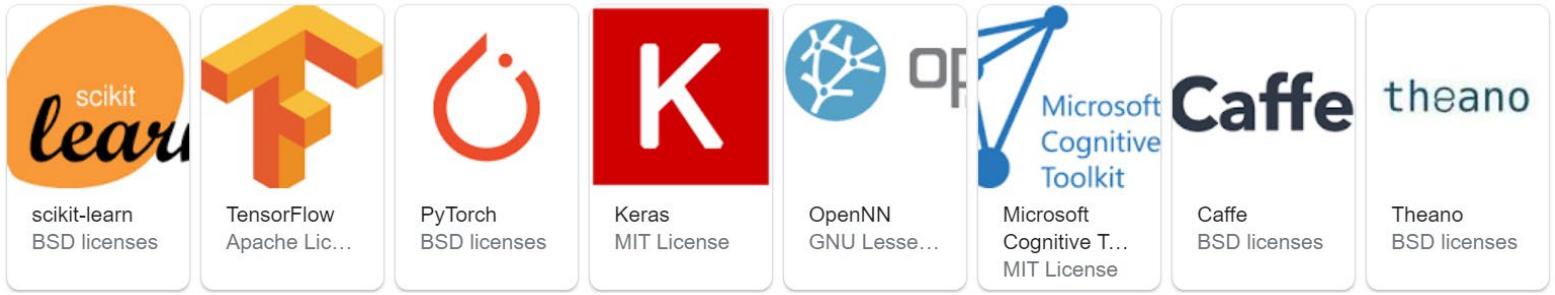
- Registry of Open Data on AWS
- Amazon Sustainability Data Initiative
- Allen Institute for Artificial Intelligence (AI2)
- Digital Earth Africa
- Facebook Data for Good
- NASA Space Act Agreement
- NIH STRIDES
- NOAA Big Data Program
- Space Telescope Science Institute

Deep learning tools

- TensorFlow
- PyTorch
- Caffe2 (merged into PyTorch)
- MxNet
- scikit-learn
- Keras (interface)

Programming language

Python
C++
R
Java
JavaScript
Go
Scala
Julia
Perl



GPU- and CPU-based acceleration: NVIDIA, cuDNN and Intel MKL

Remarks

- To solve complex problems, an organic integration (using hierarchies) of algorithms is helpful
- “Shallow/small” learning still works and is useful – for small problems, little data availability
- Deep reinforcement learning
- Graph neural networks
- Meta-learning