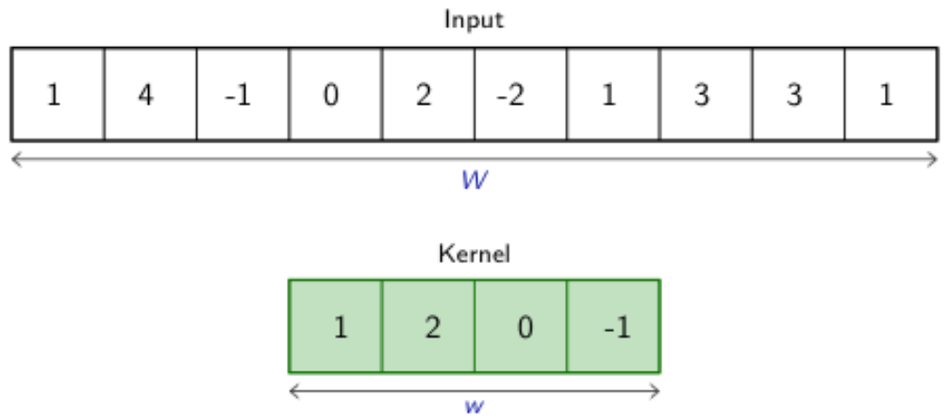
Convolutional neural networks

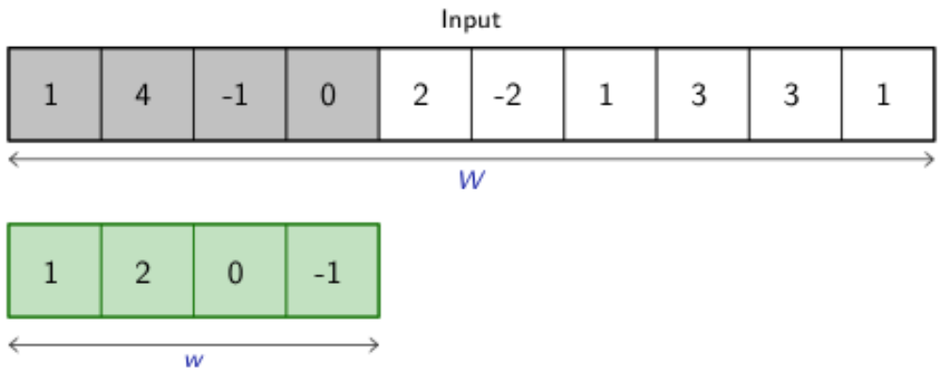
Lecture 11 of ML for physicists

Convolutional layers

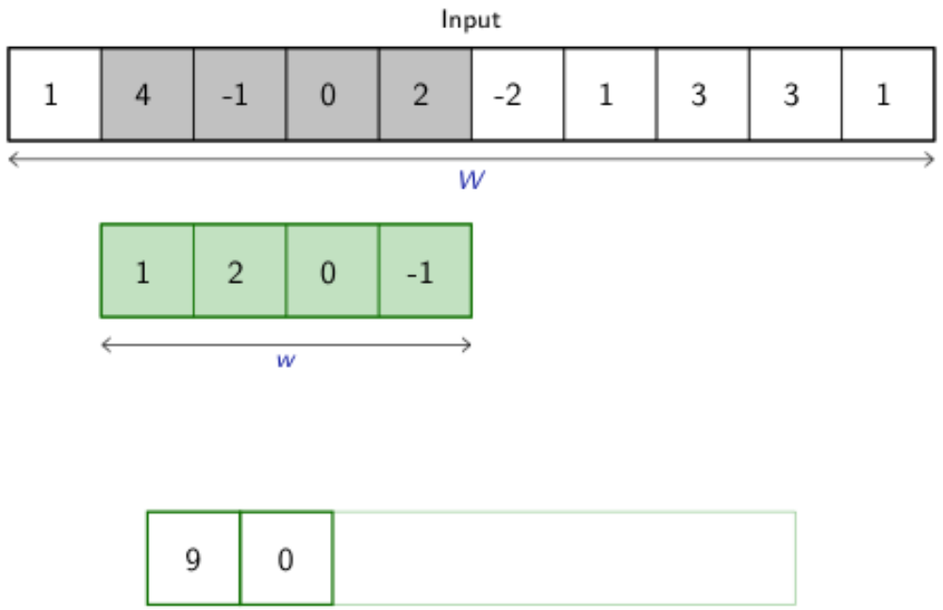
Convolution 1d



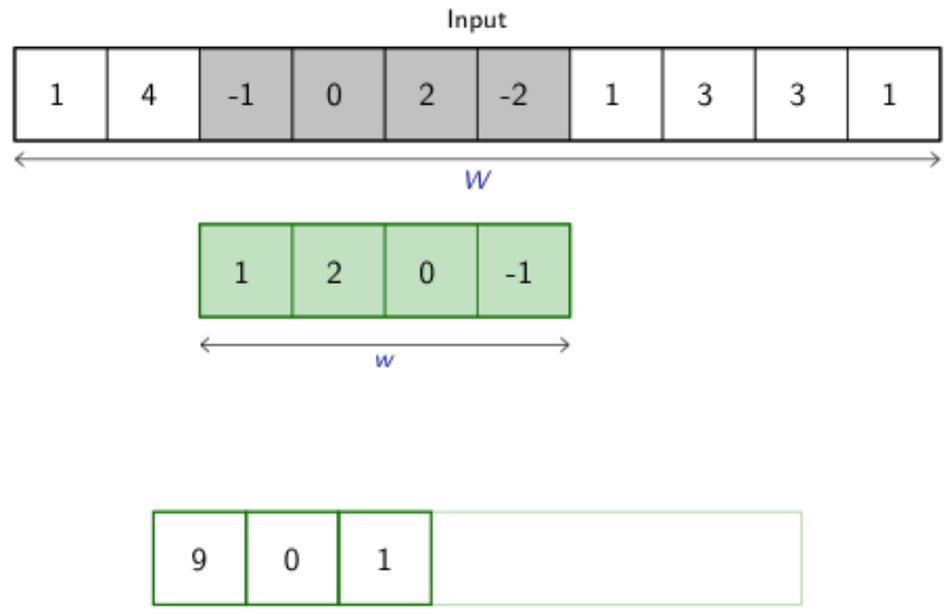
Convolution 1d



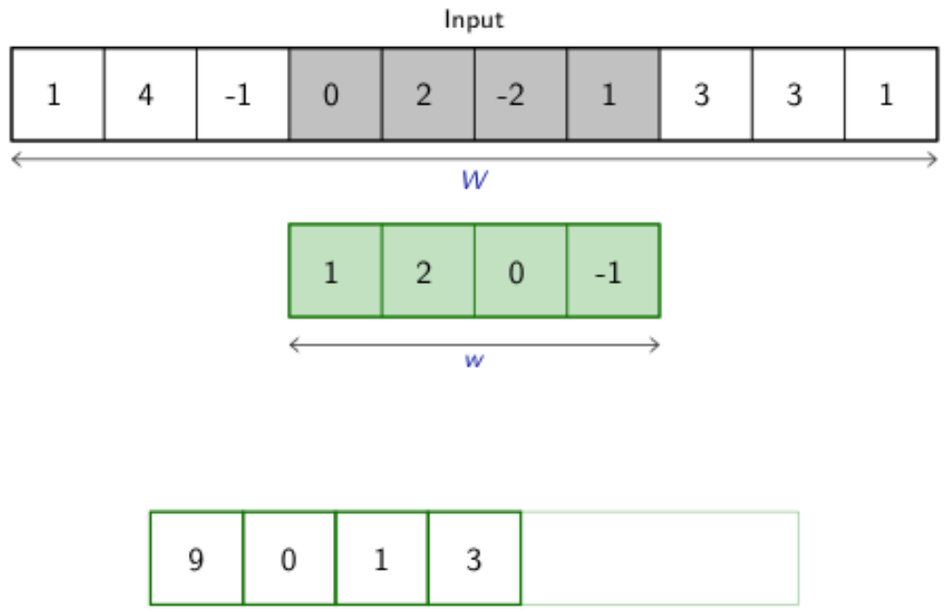
Convolution 1d



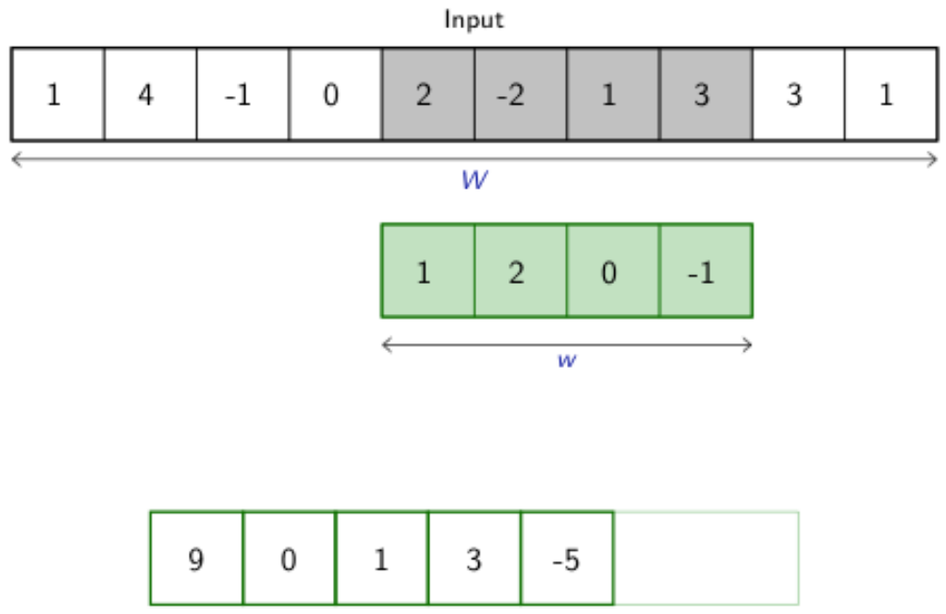
Convolution 1d



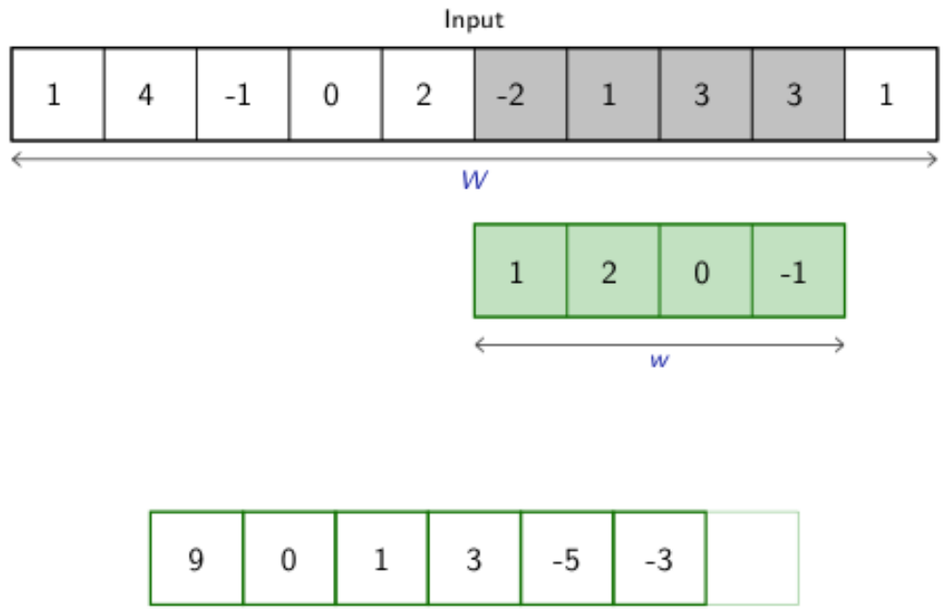
Convolution 1d



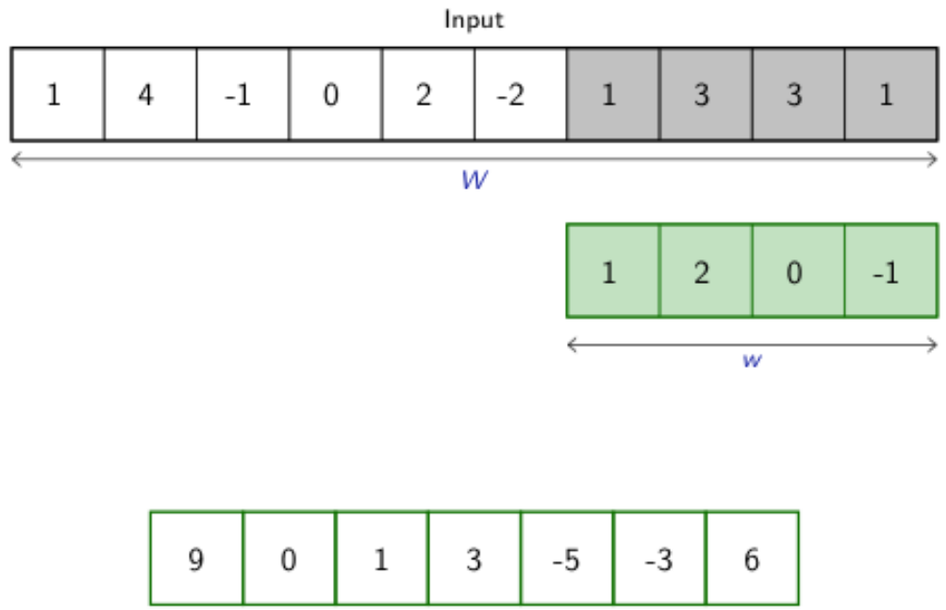
Convolution 1d



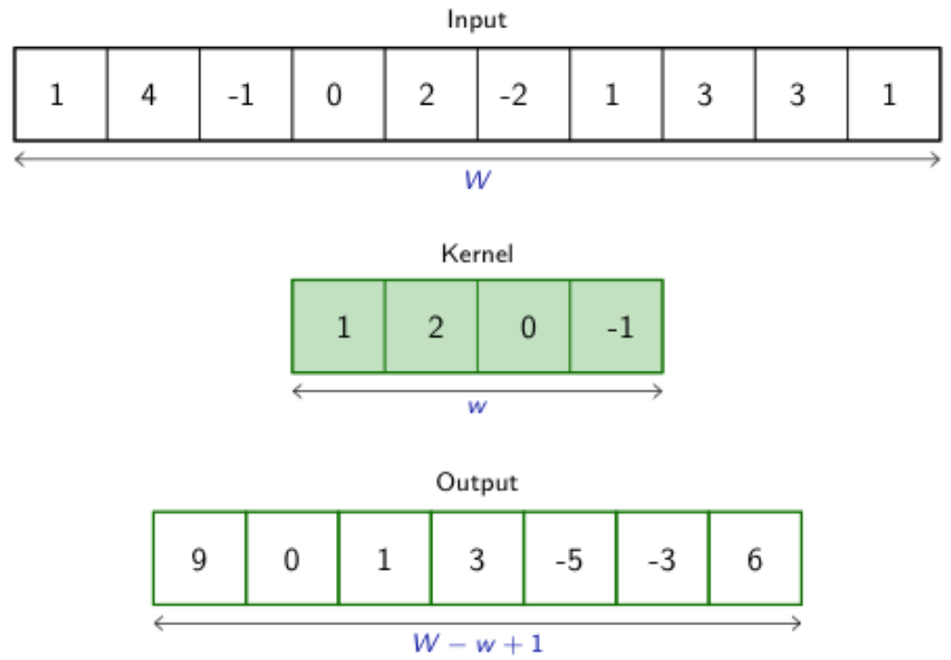
Convolution 1d



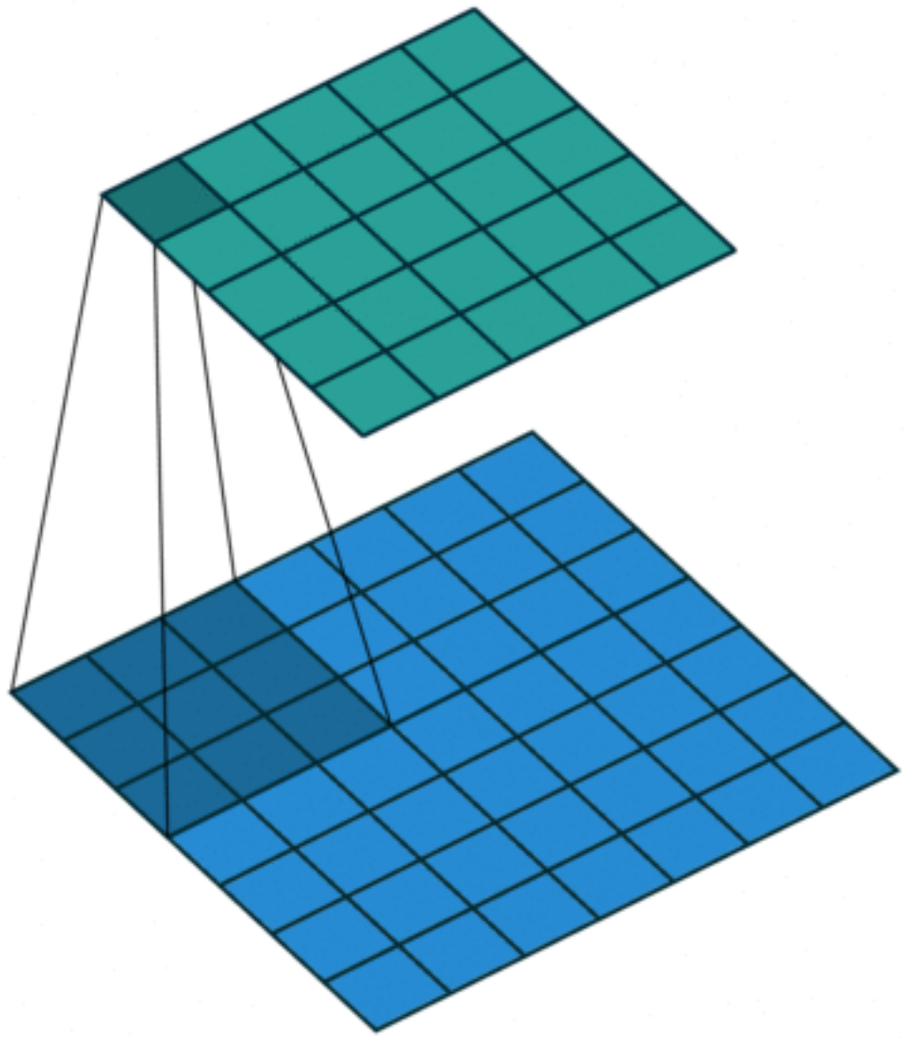
Convolution 1d



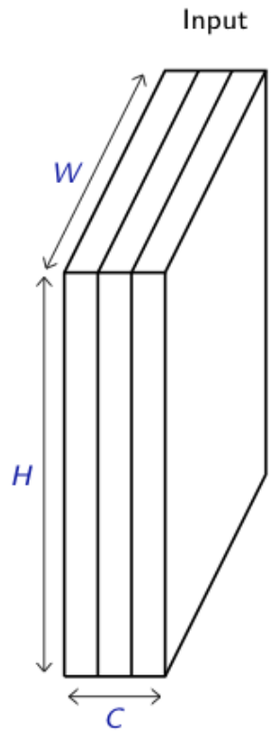
Convolution 1d



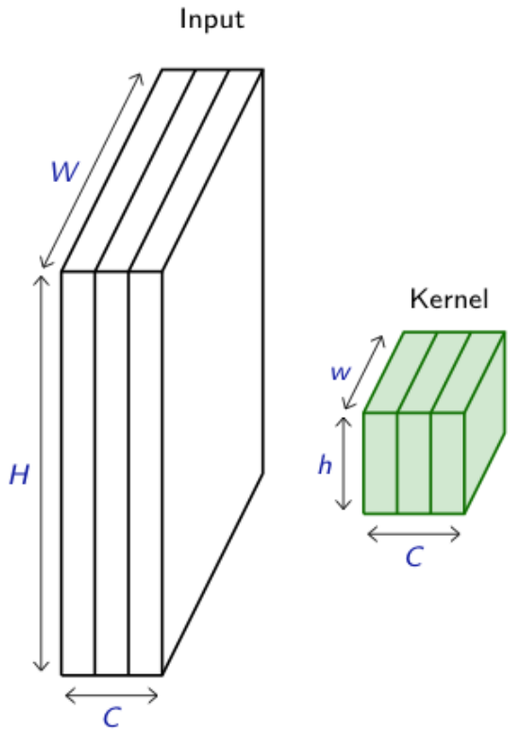
Convolutional filter



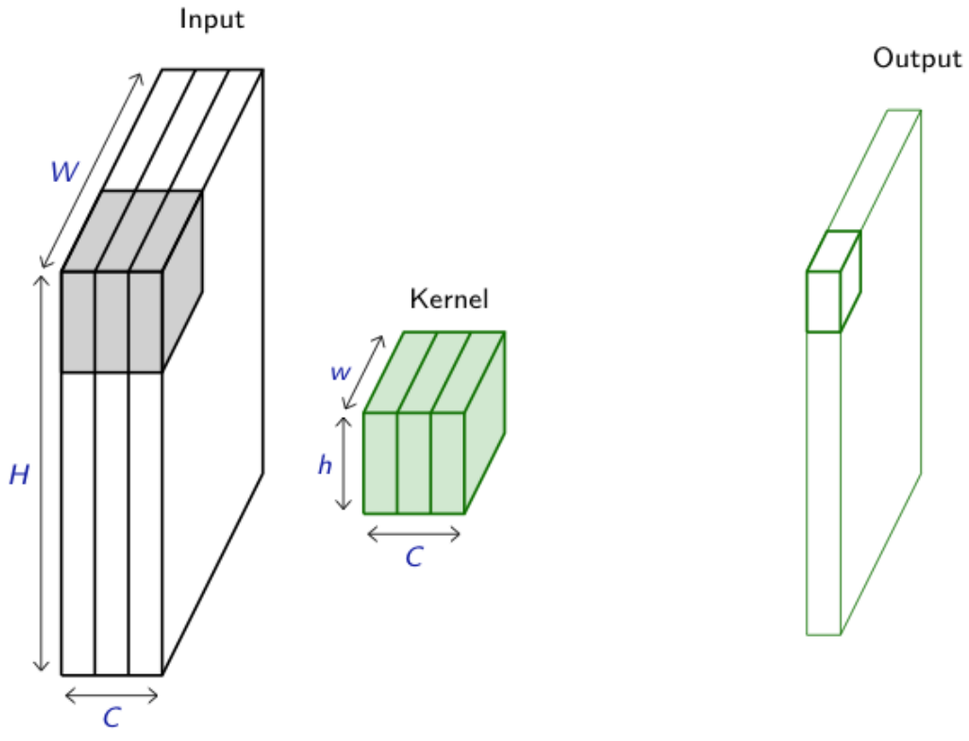
Convolution 2d



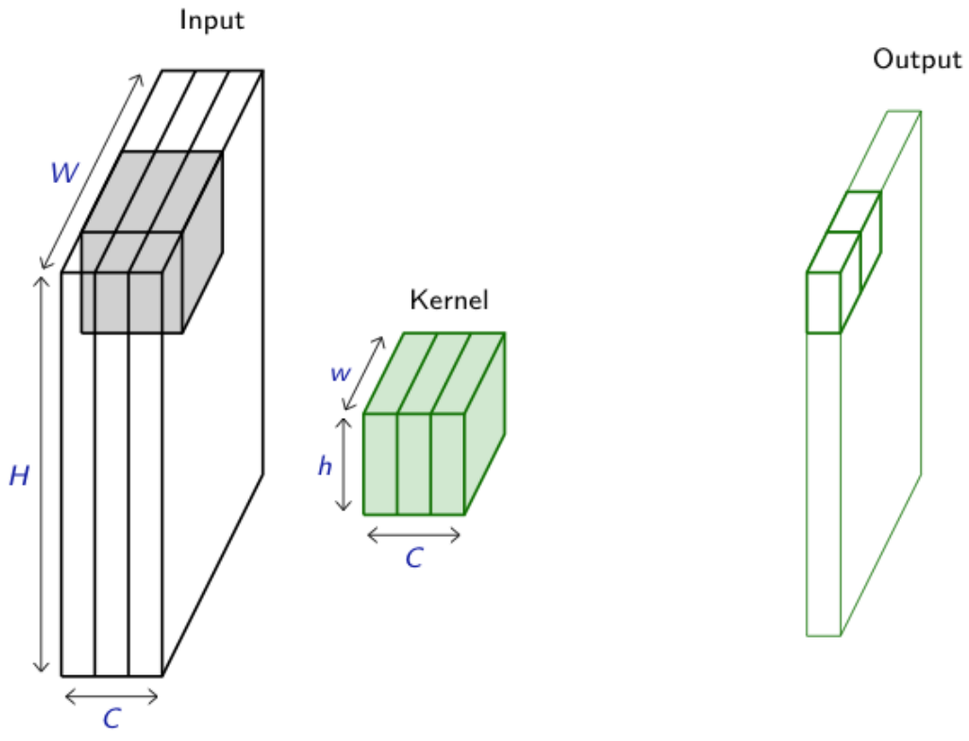
Convolution 2d



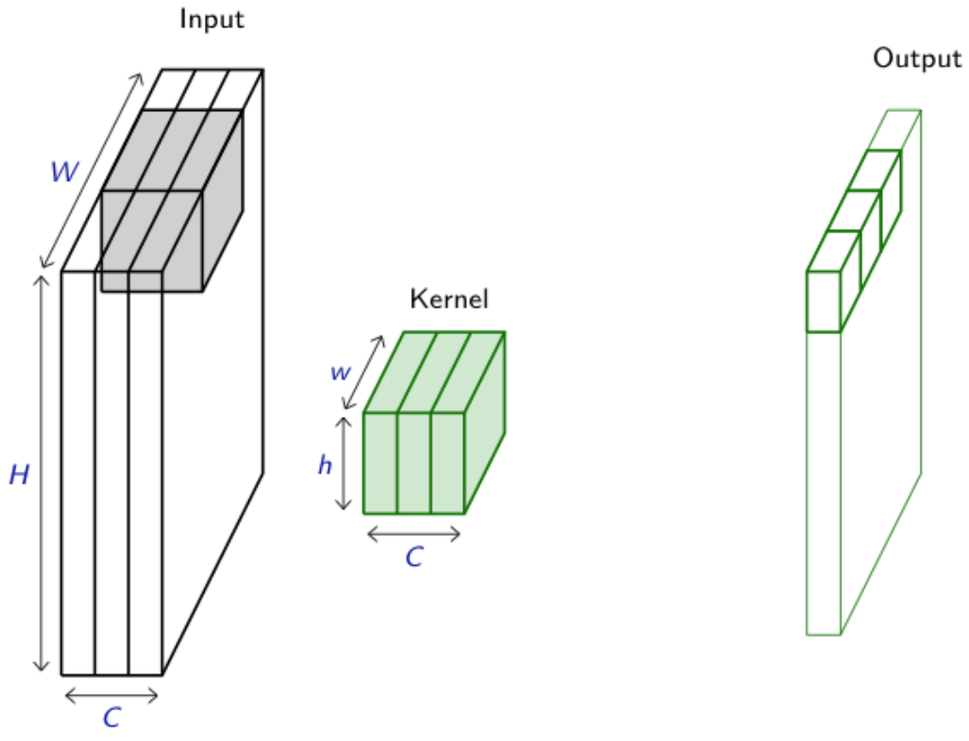
Convolution 2d



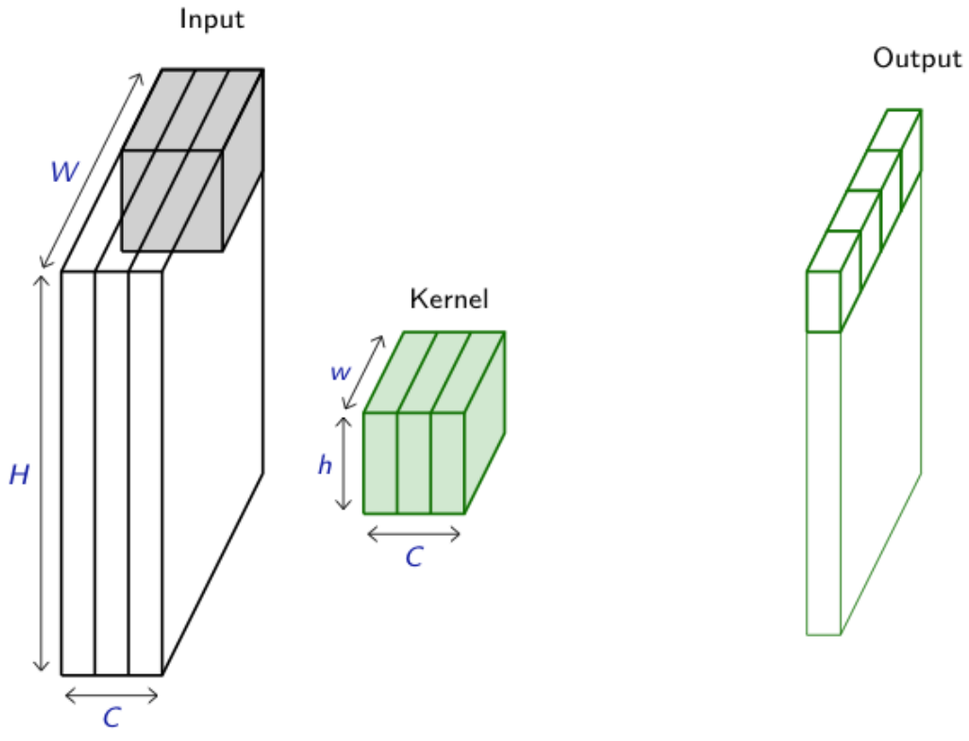
Convolution 2d



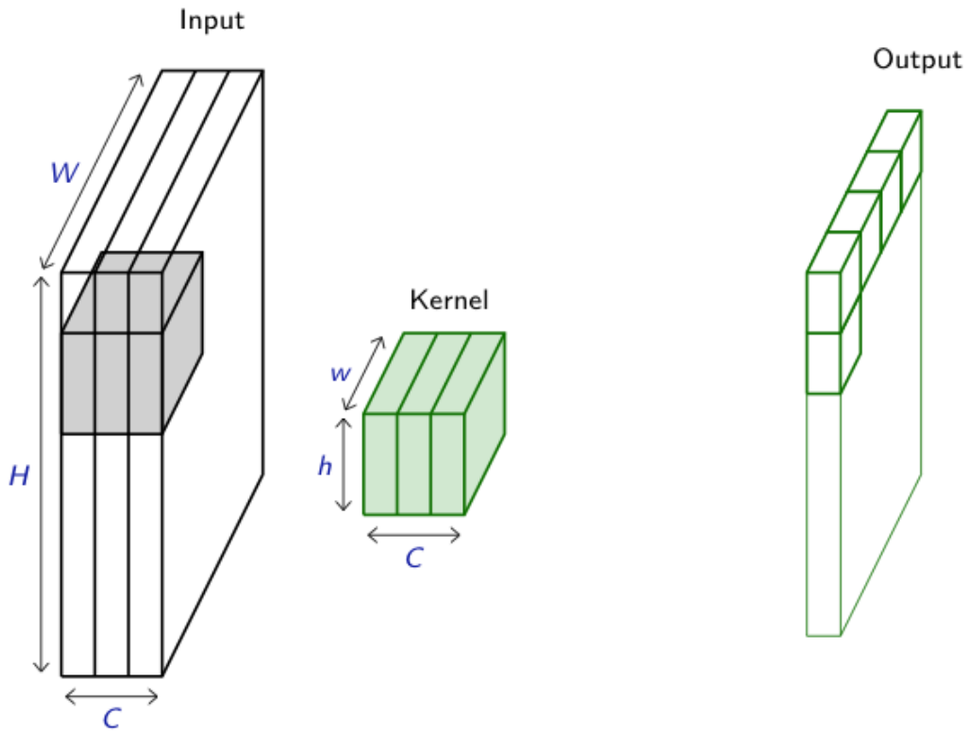
Convolution 2d



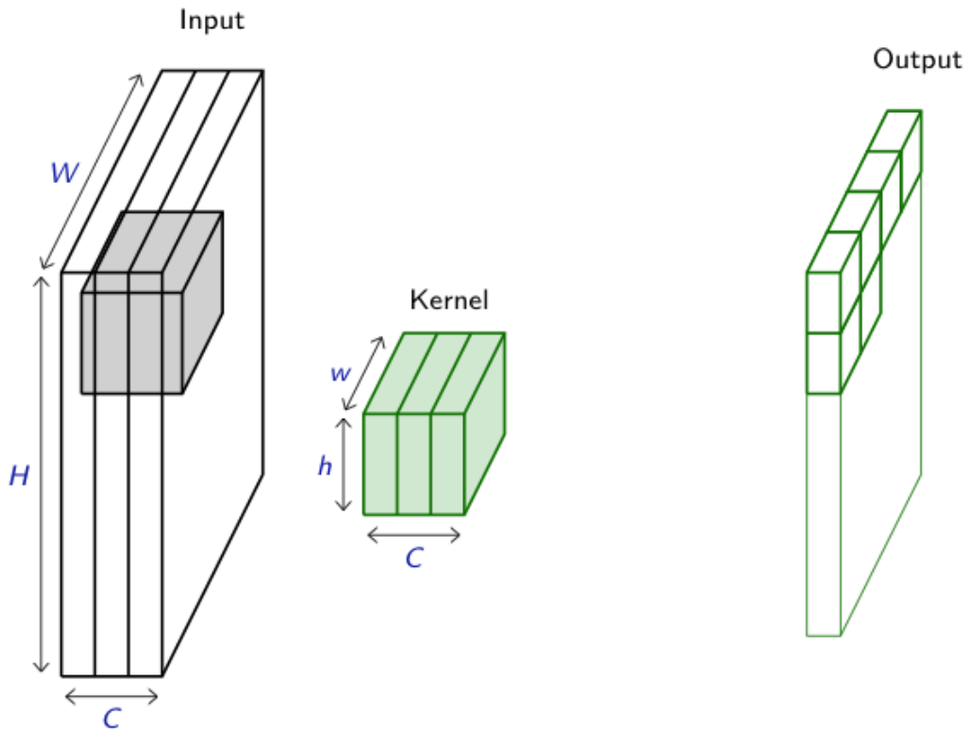
Convolution 2d



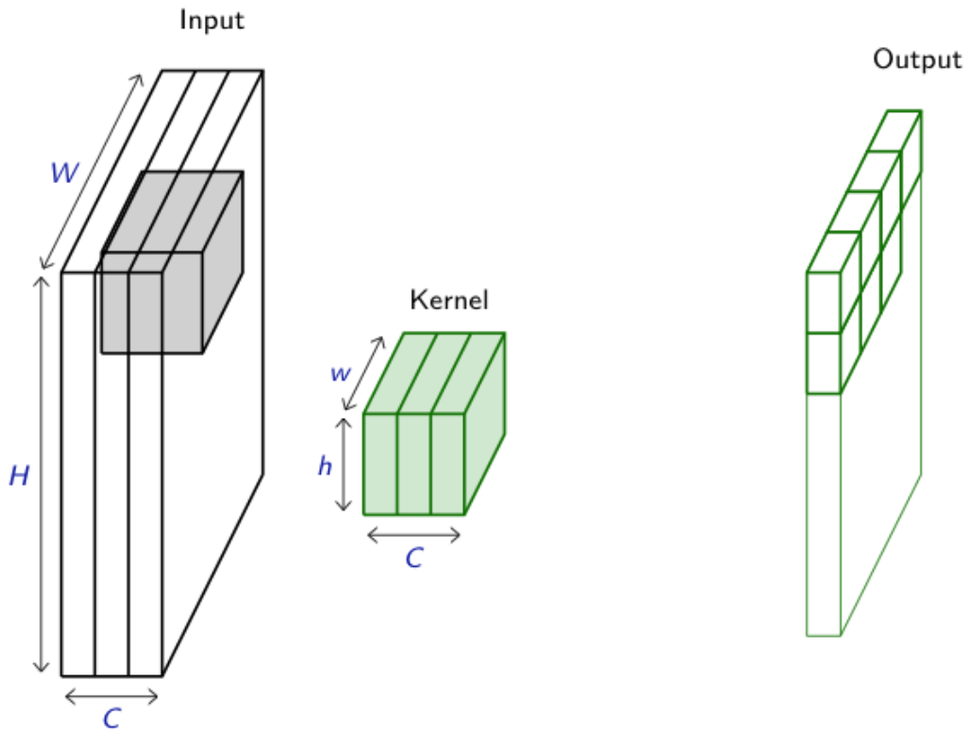
Convolution 2d



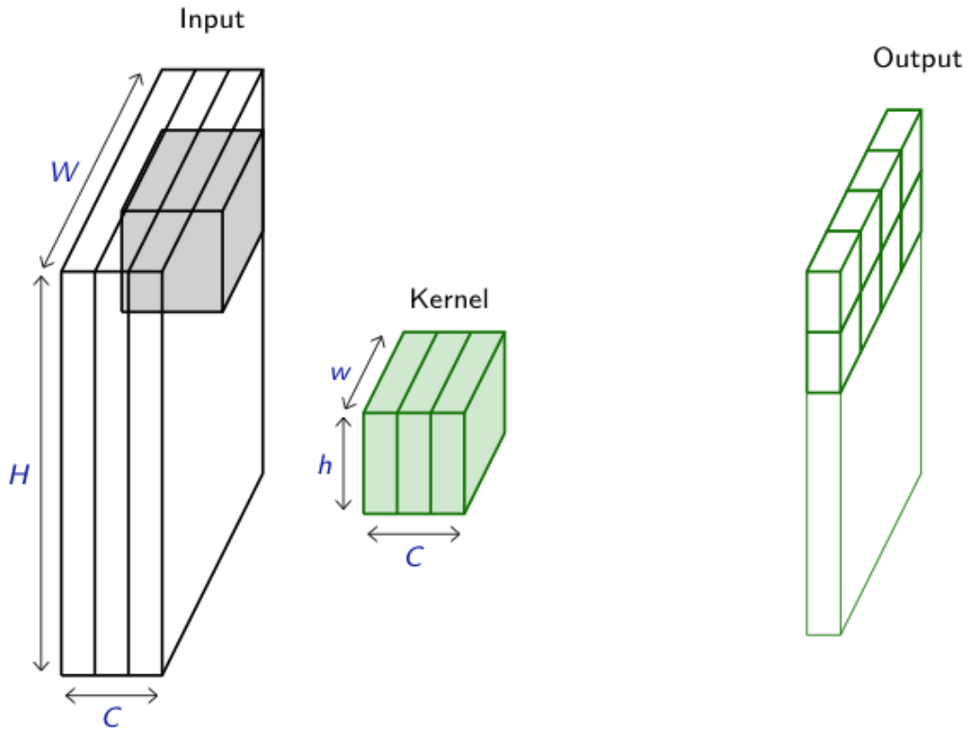
Convolution 2d



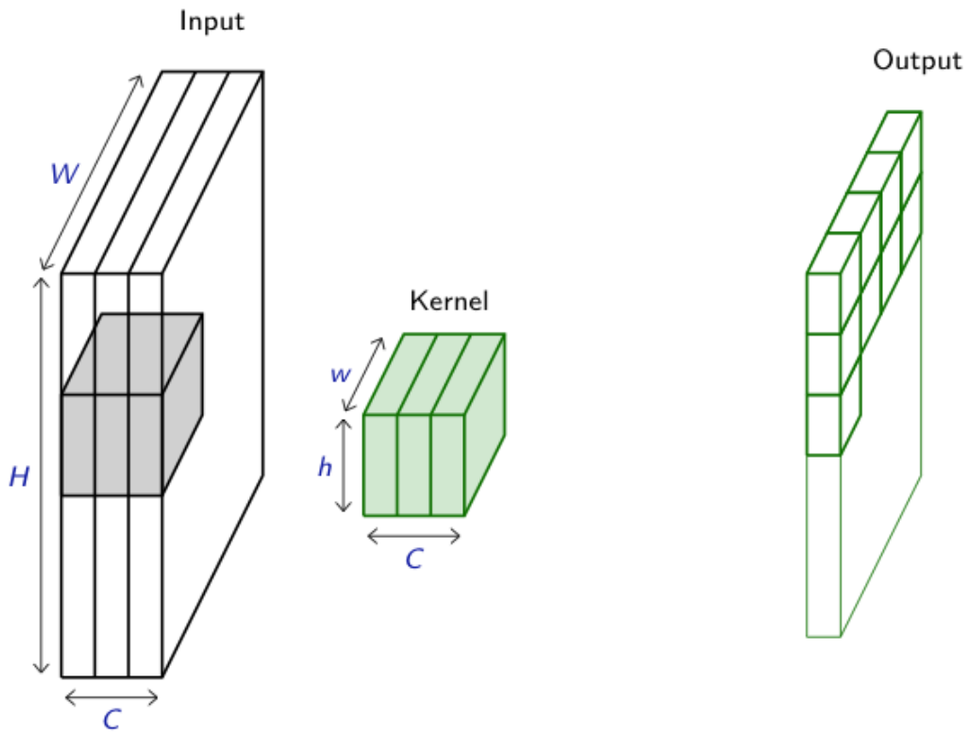
Convolution 2d



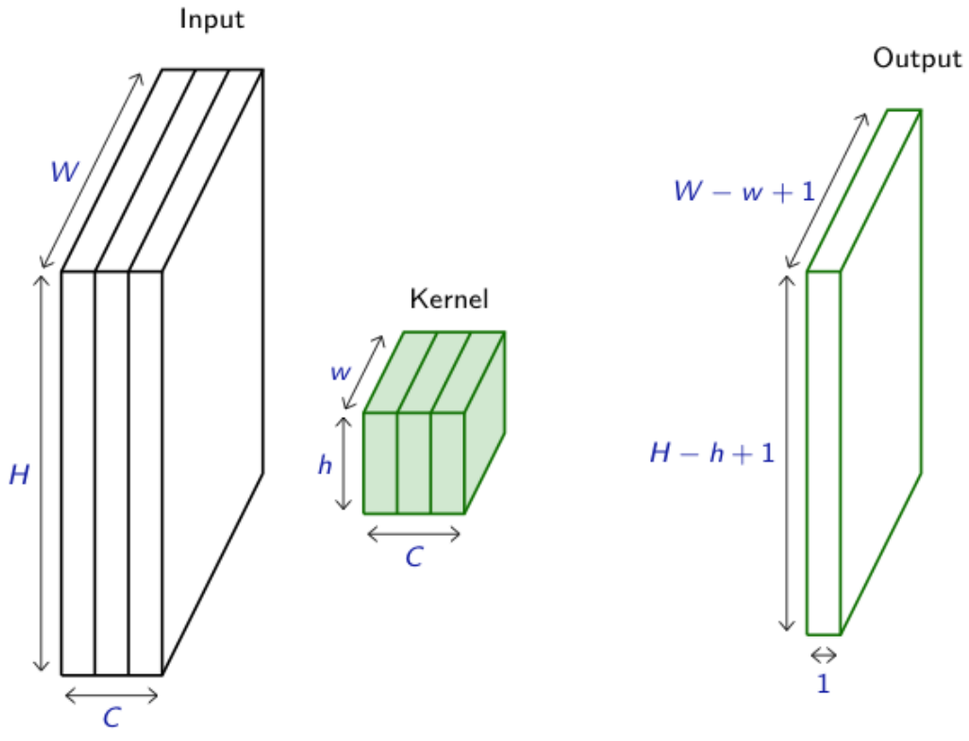
Convolution 2d



Convolution 2d

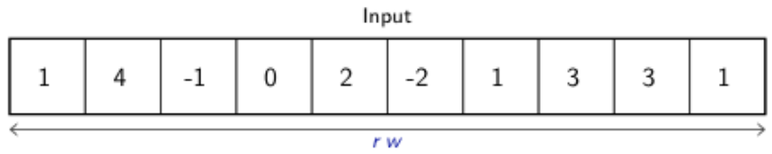


Convolution 2d

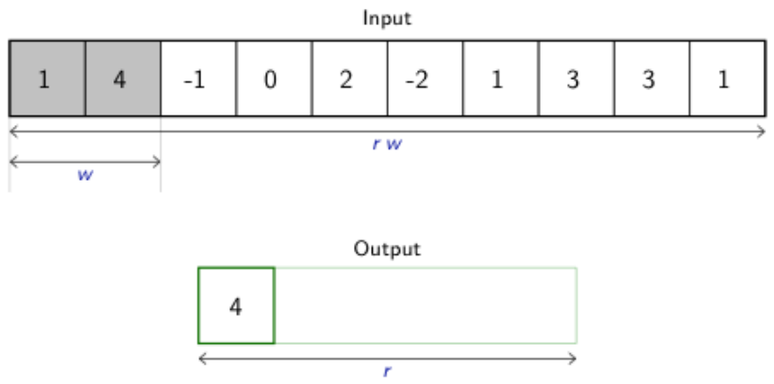


Pooling layers

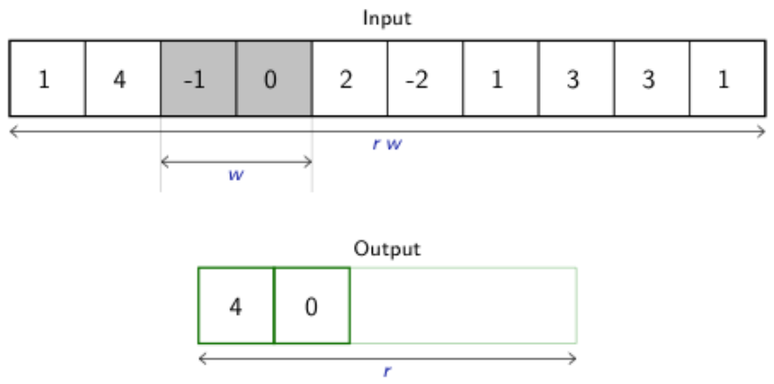
Max-Pooling 1d



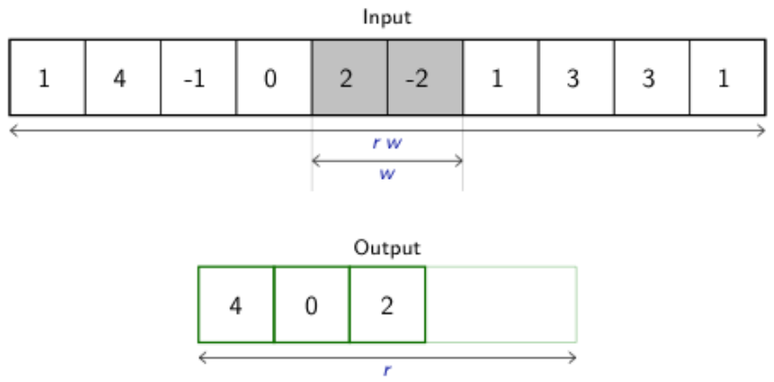
Max-Pooling 1d



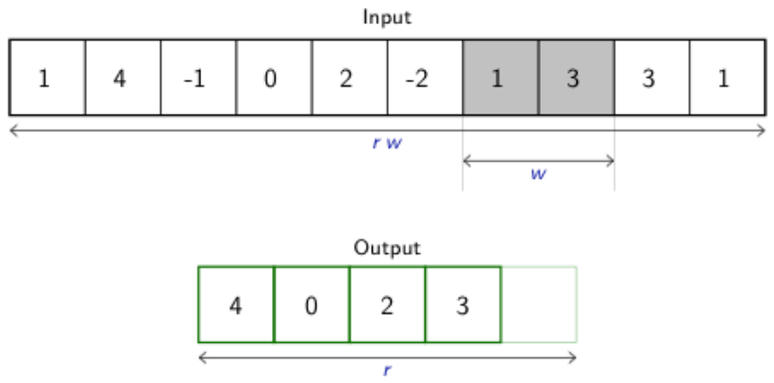
Max-Pooling 1d



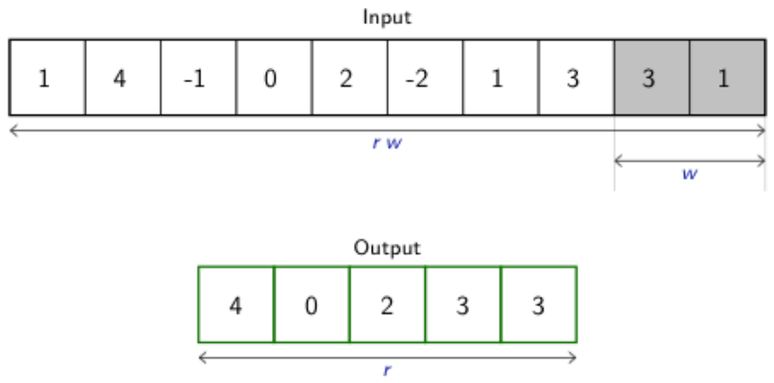
Max-Pooling 1d



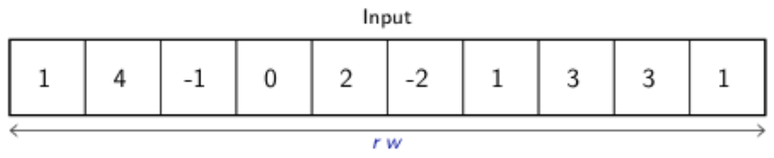
Max-Pooling 1d

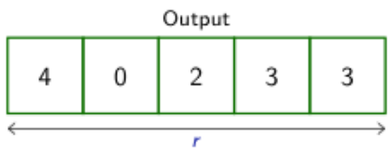


Max-Pooling 1d

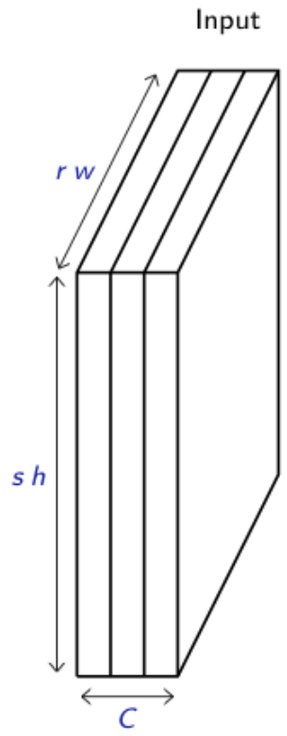


Max-Pooling 1d

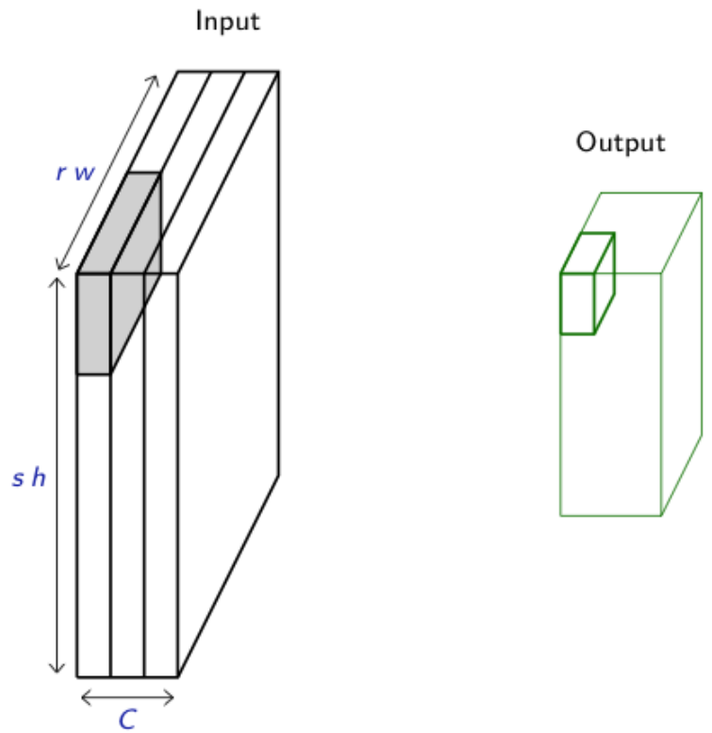




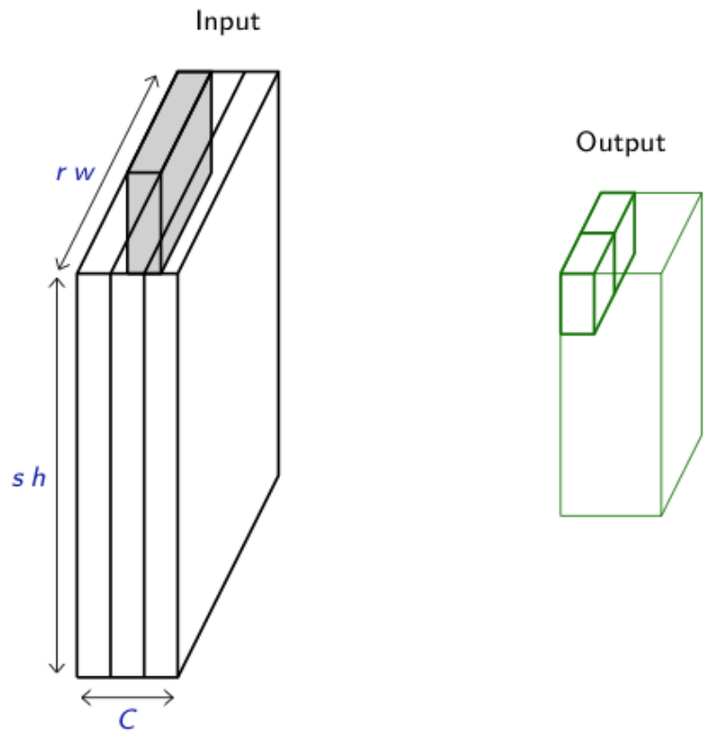
Max-Pooling 2d



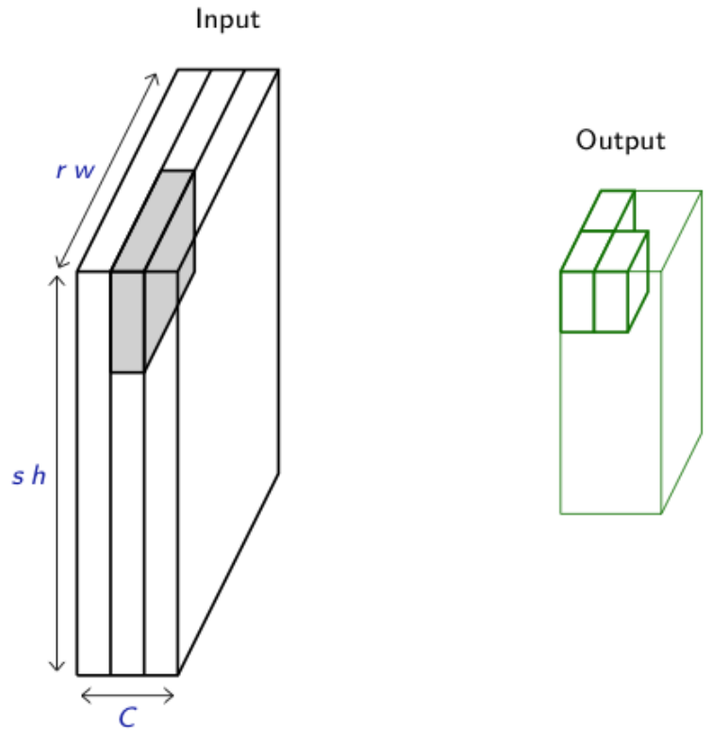
Max-Pooling 2d



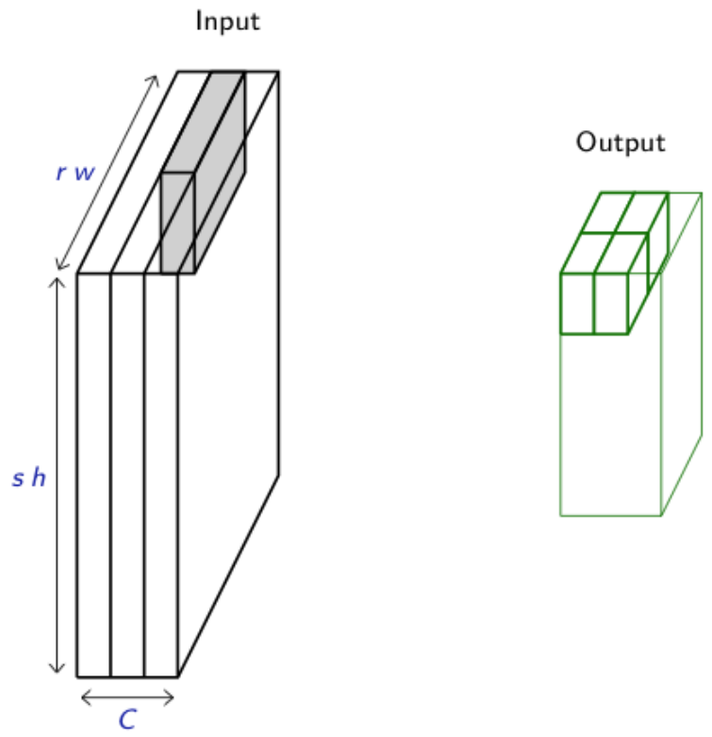
Max-Pooling 2d



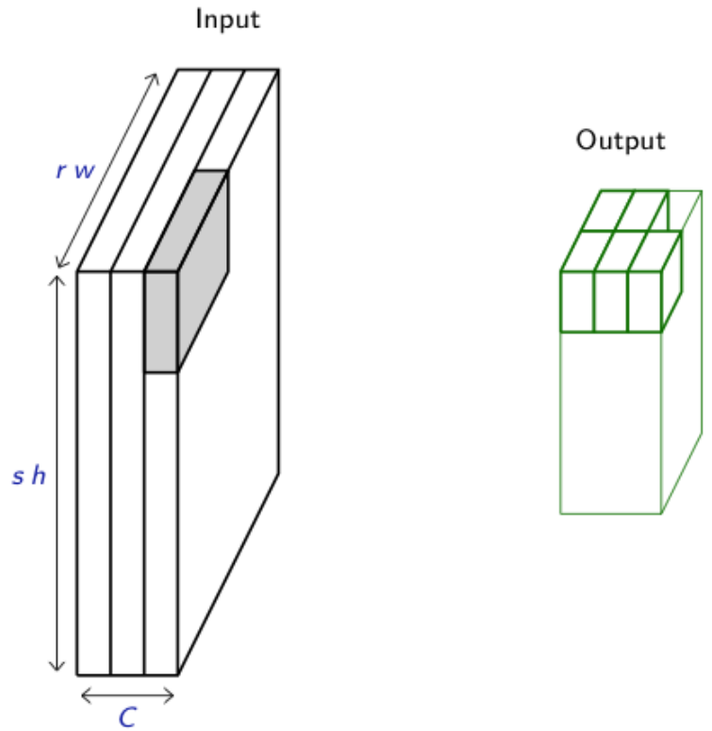
Max-Pooling 2d



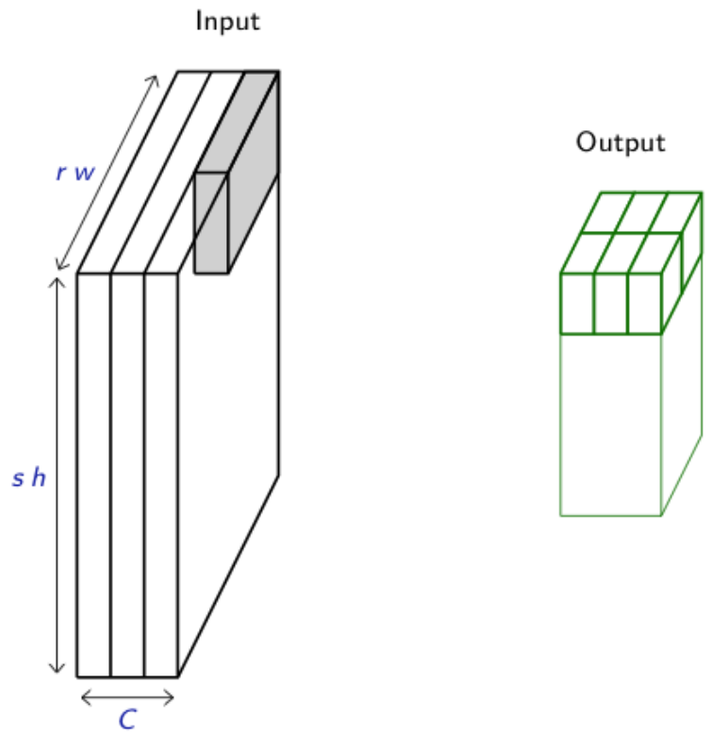
Max-Pooling 2d



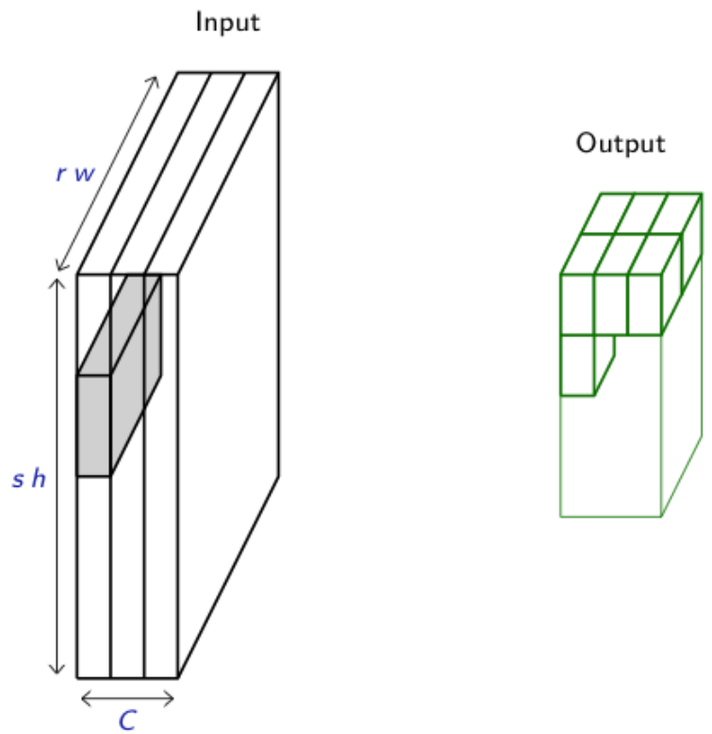
Max-Pooling 2d



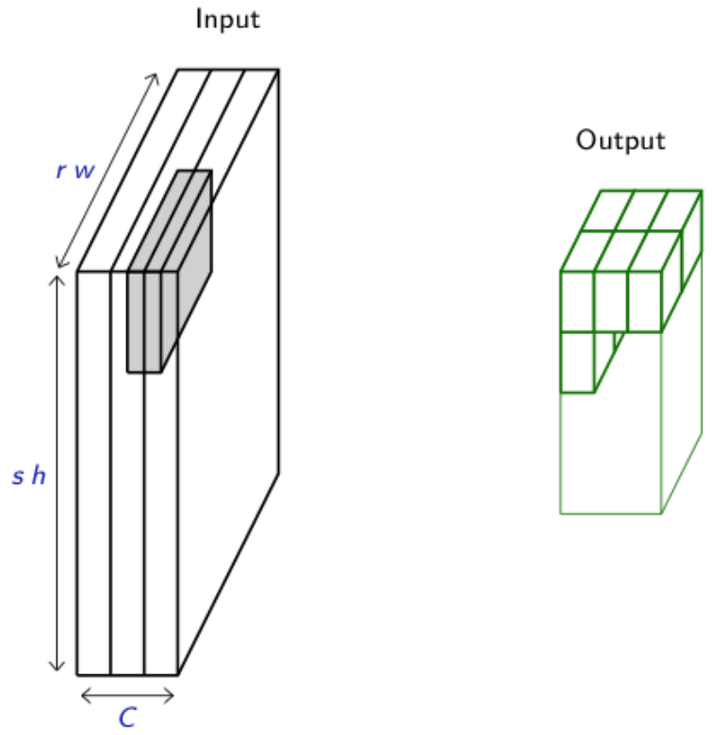
Max-Pooling 2d



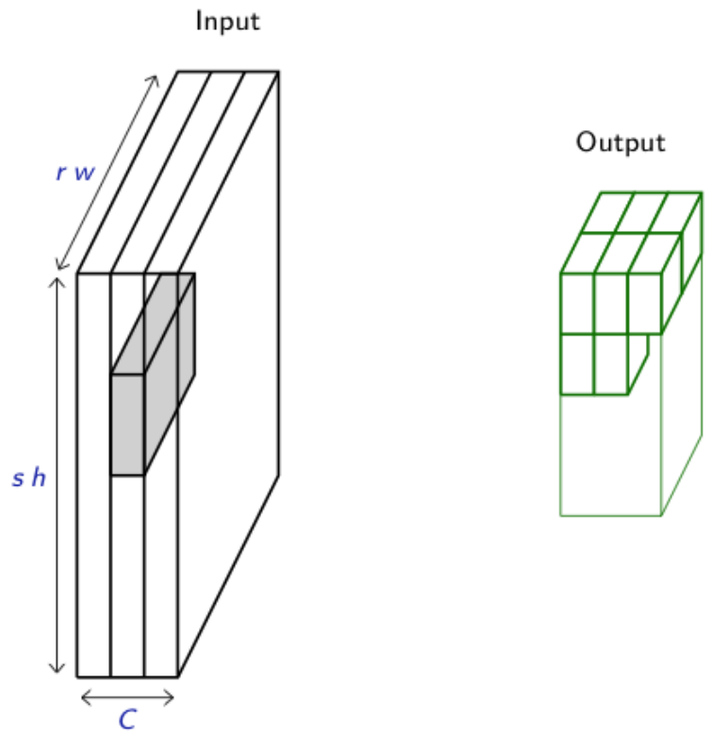
Max-Pooling 2d



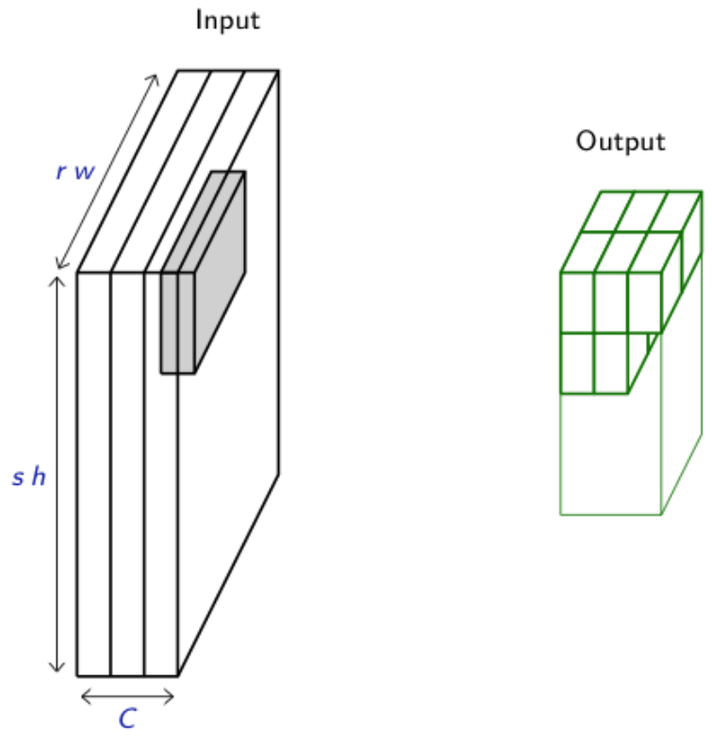
Max-Pooling 2d



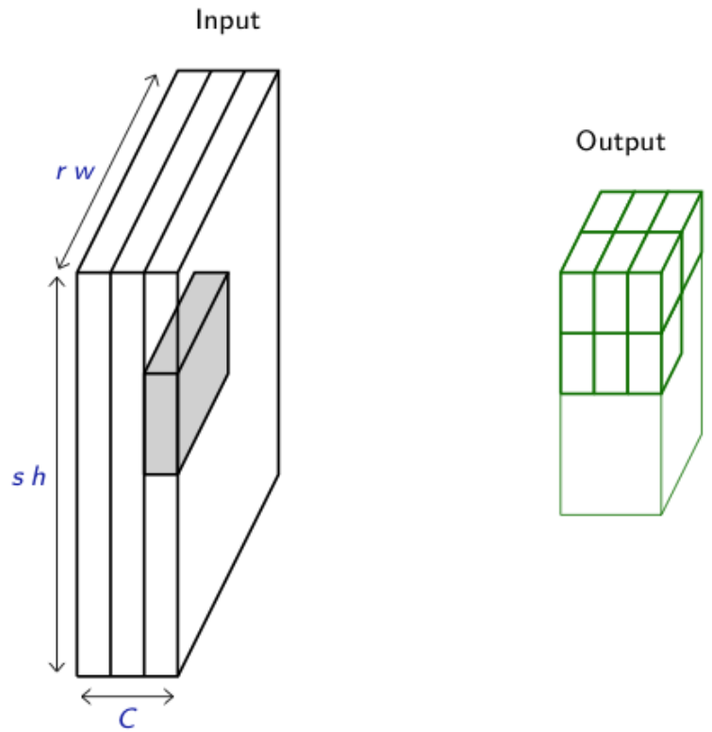
Max-Pooling 2d



Max-Pooling 2d

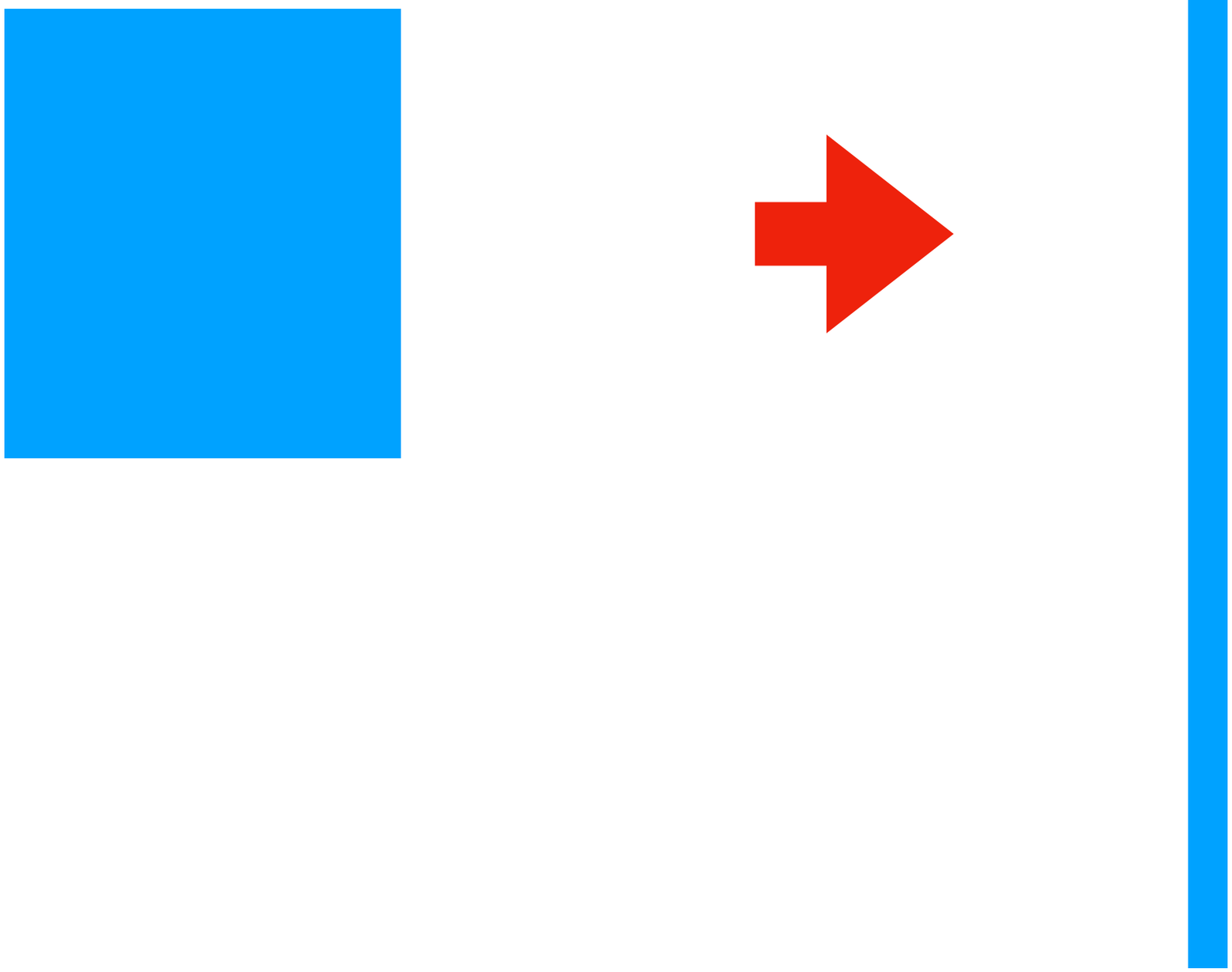


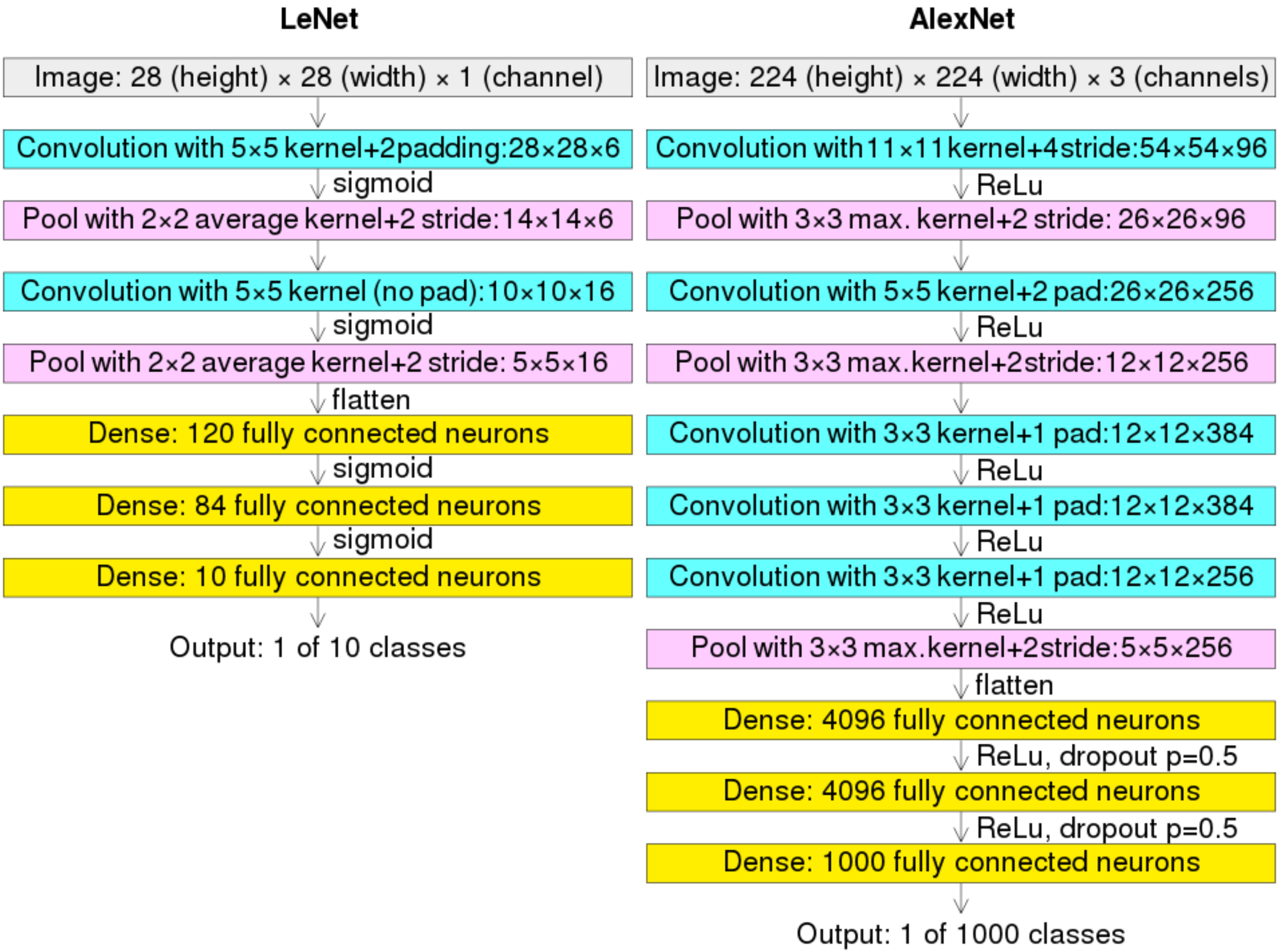
Max-Pooling 2d

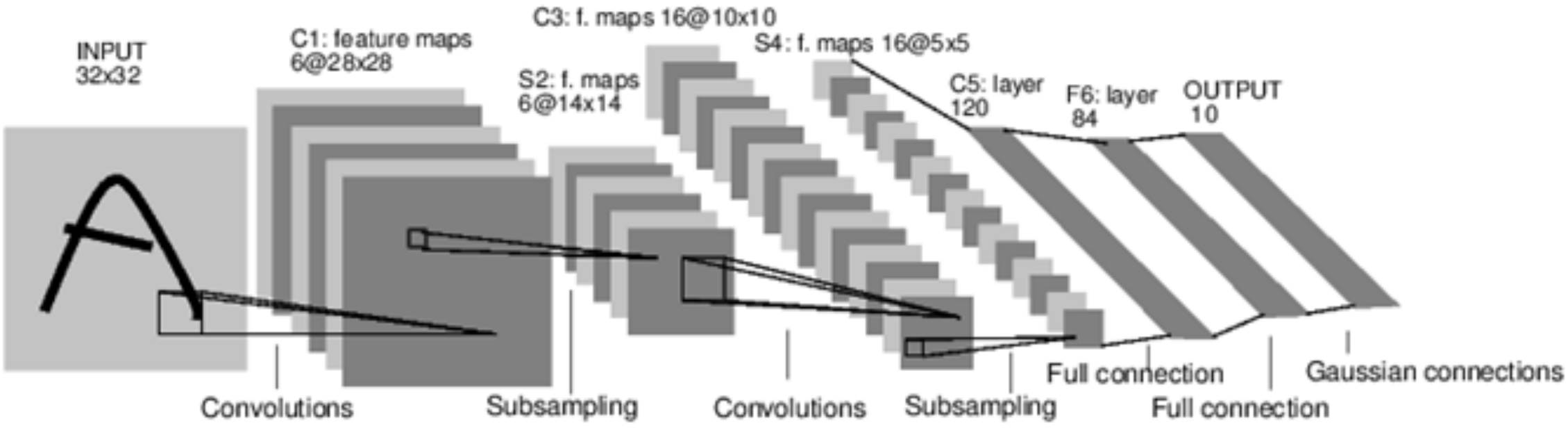


Convolutional architectures

Flatten
2D->1D

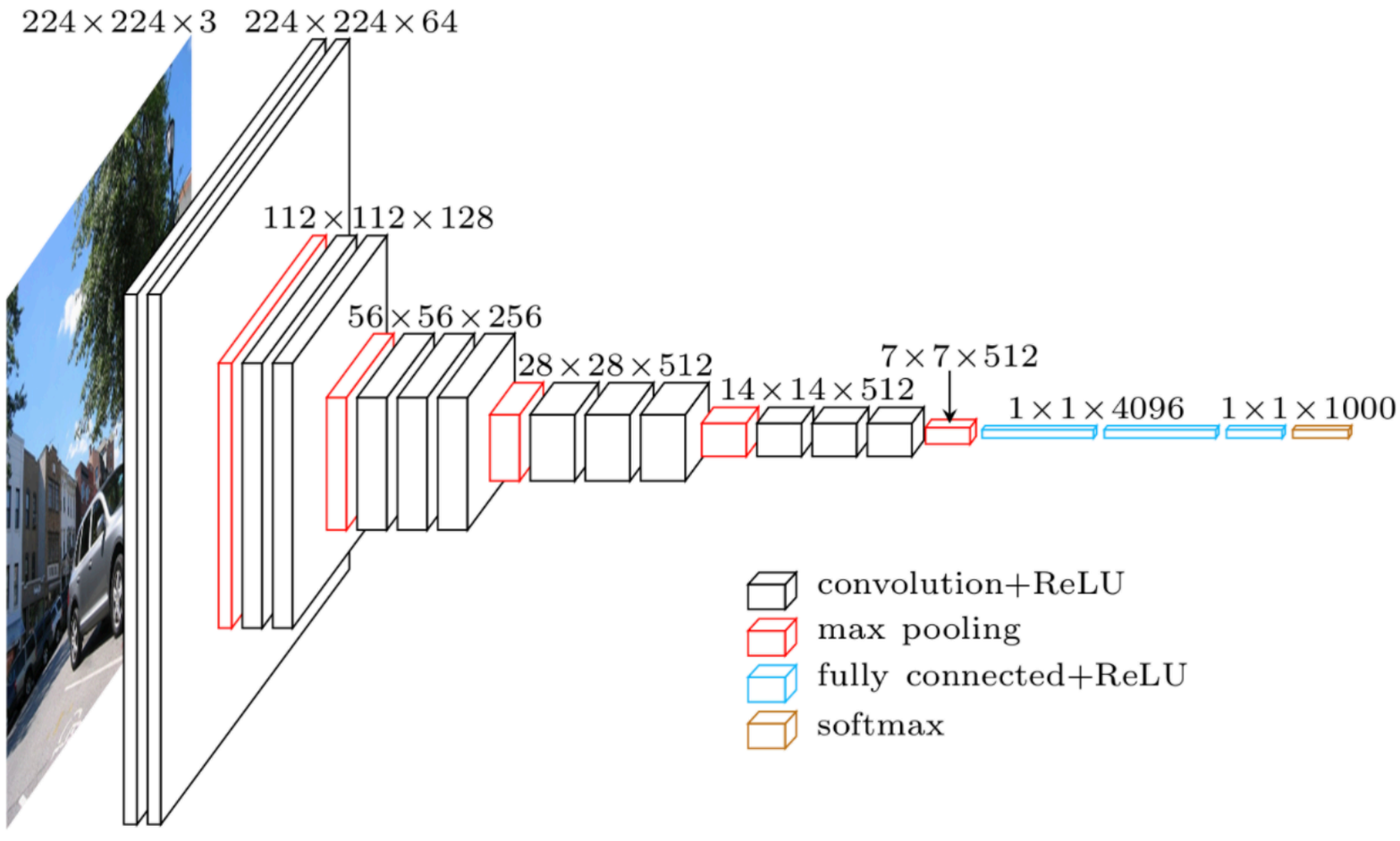



LeNet
Image: 28 (height) × 28 (width) × 1 (channel)
Convolution with 5×5 kernel+2padding:28×28×6
sigmoid
Pool with 2×2 average kernel+2 stride:14×14×6
Convolution with 5×5 kernel (no pad):10×10×16
sigmoid
Pool with 2×2 average kernel+2 stride: 5×5×16
flatten
Dense: 120 fully connected neurons
sigmoid
Dense: 84 fully connected neurons
sigmoid
Dense: 10 fully connected neurons
Output: 1 of 10 classes
AlexNet
Image: 224 (height) × 224 (width) × 3 (channels)
Convolution with 11×11 kernel+4stride:54×54×96
ReLu
Pool with 3×3 max. kernel+2 stride: 26×26×96
Convolution with 5×5 kernel+2 pad:26×26×256
ReLu
Pool with 3×3 max.kernel+2stride:12×12×256
Convolution with 3×3 kernel+1 pad:12×12×384
ReLu
Convolution with 3×3 kernel+1 pad:12×12×384
ReLu
Convolution with 3×3 kernel+1 pad:12×12×256
ReLu
Pool with 3×3 max.kernel+2stride:5×5×256
flatten
Dense: 4096 fully connected neurons
ReLu, dropout p=0.5
Dense: 4096 fully connected neurons
ReLu, dropout p=0.5
Dense: 1000 fully connected neurons
Output: 1 of 1000 classes



A Full Convolutional Neural Network (LeNet)

VGG-16


224 × 224 × 3
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
convolution+ReLU
max pooling
fully connected+ReLU
softmax

Correction on attention layer

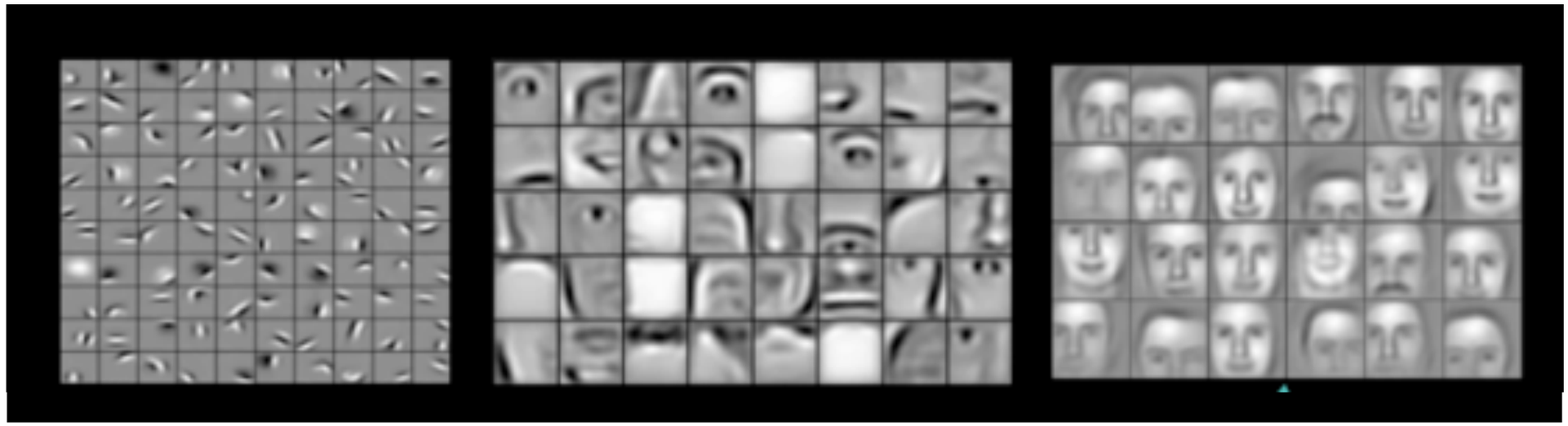
$$t_{ai}^{\mu,(l+1)} = \sum_{b=1}^{T} A_{ab}^{(l)} \sum_{j=1}^{d} t_{bj}^{\mu,(l)} V_{ji}^{(l)}$$

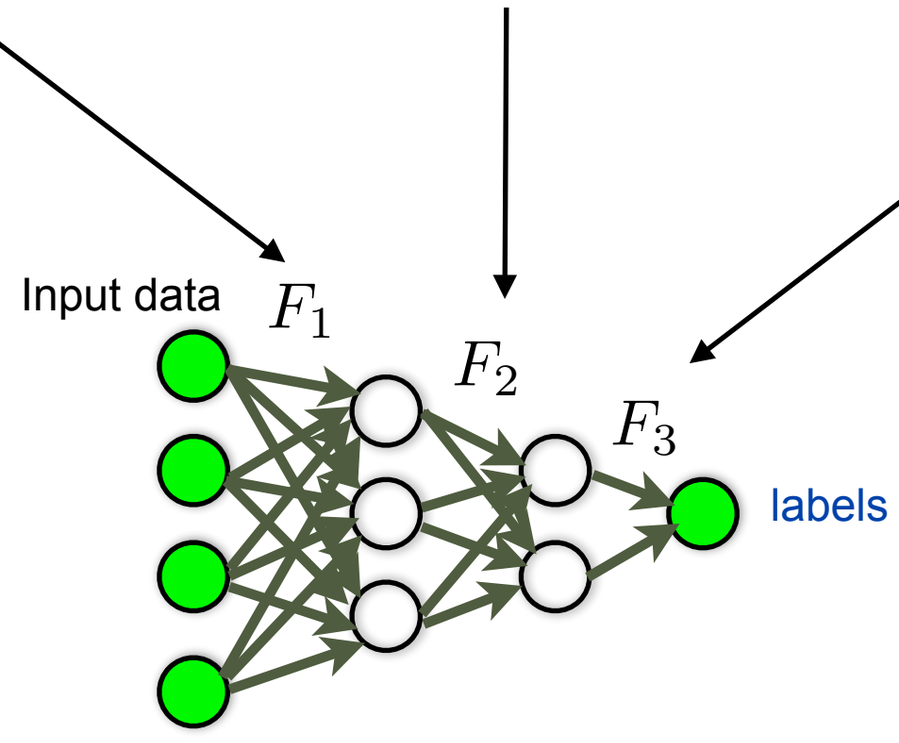
$$A_{ab}^{(l)} = \frac{\exp\left[\frac{1}{\sqrt{r}} \sum_{k=1}^{r} \left(\sum_{j=1}^{d} t_{aj}^{\mu,(l)} Q_{jk}^{(l)}\right)\left(\sum_{j=1}^{d} t_{bj}^{\mu,(l)} K_{jk}^{(l)}\right)\right]}{\sum_{c=1}^{T} \exp\left[\frac{1}{\sqrt{r}} \sum_{k=1}^{r} \left(\sum_{j=1}^{d} t_{aj}^{\mu,(l)} Q_{jk}^{(l)}\right)\left(\sum_{j=1}^{d} t_{cj}^{\mu,(l)} K_{jk}^{(l)}\right)\right]}$$

Modus Operandi of Deep Learning

Lecture 13
of ML for physicists

Hierarchy of features



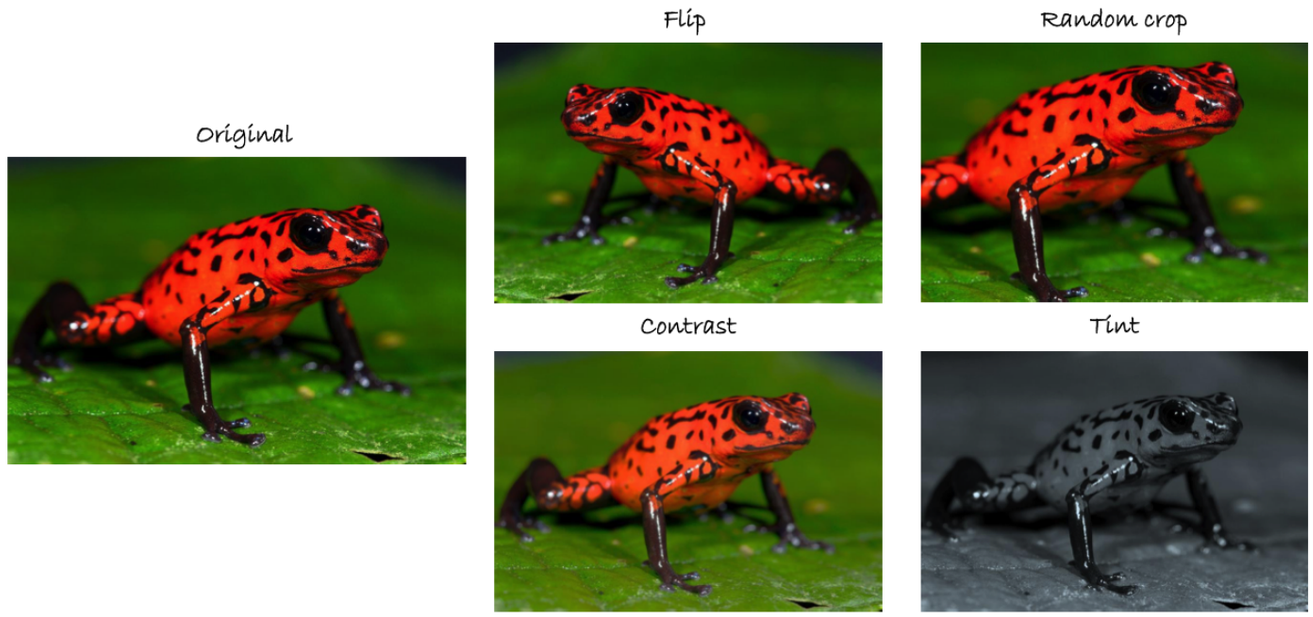


What do you do when do not have enough data?

You create more!

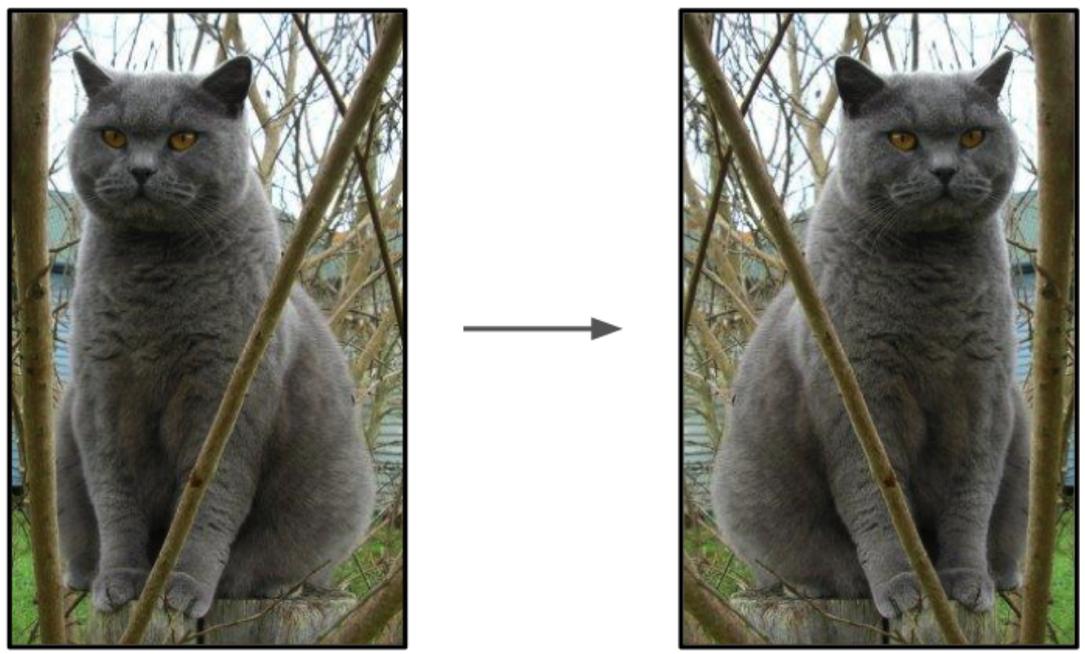
Data augmentation

- Changing the pixels without changing the label
- Train on transformed data
- Widely used



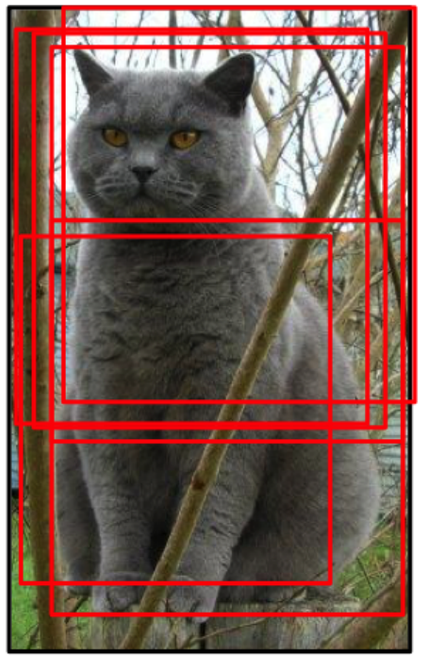
Data augmentation

Horizontal flips



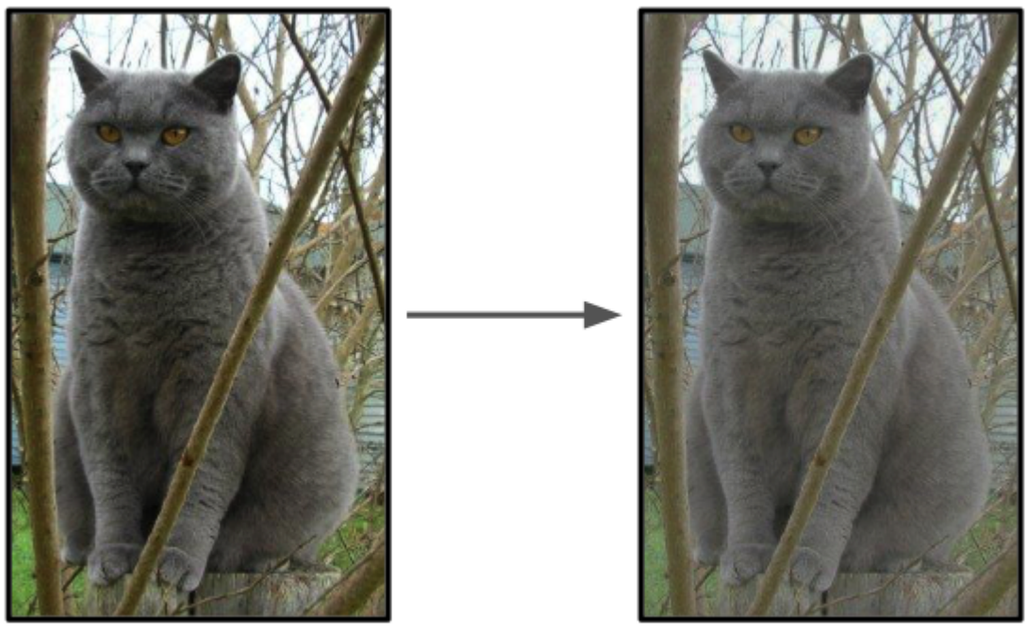
Data augmentation

Random crops/scales



Data augmentation

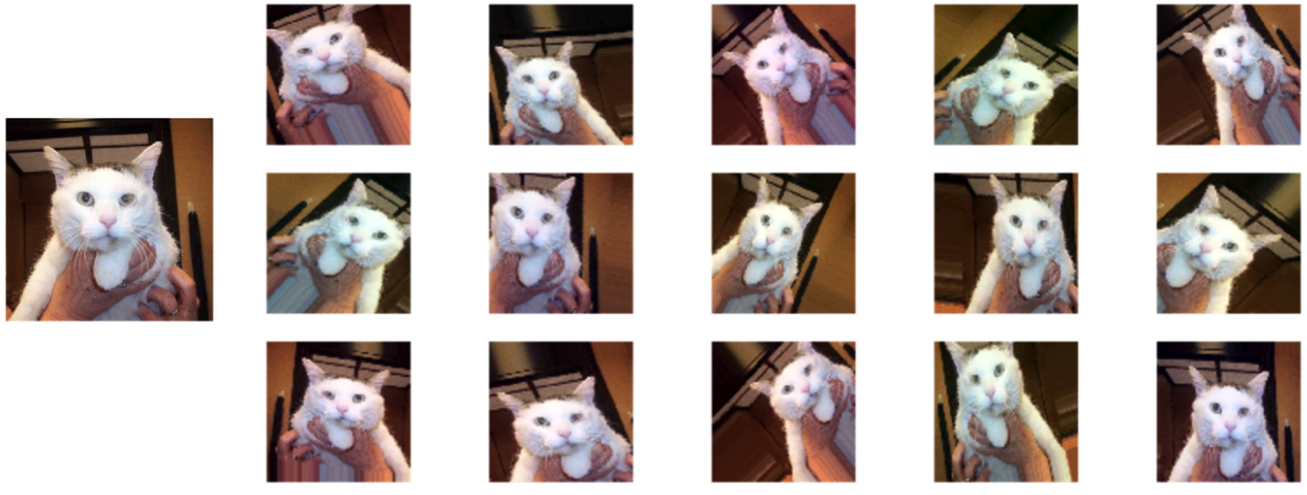
Color jitter



- randomly jitter color, brightness, contrast, etc.

Data augmentation

- Various techniques can be mixed
- Domain knowledge helps in finding new data augmentation techniques
- Very useful for small datasets



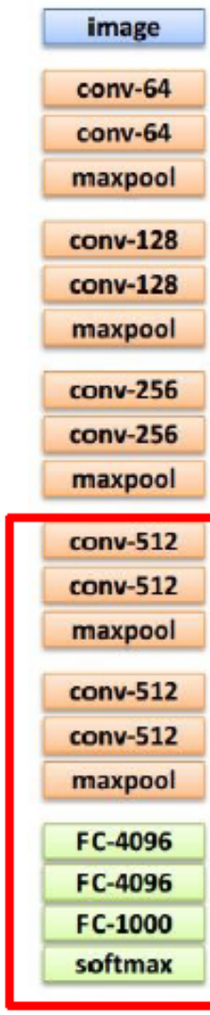
Transfer Learning

Transfer Learning with CNNs

1. Train on ImageNet

2. If small dataset: fix all weights (treat CNN as fixed feature extractor), retrain only the classifier

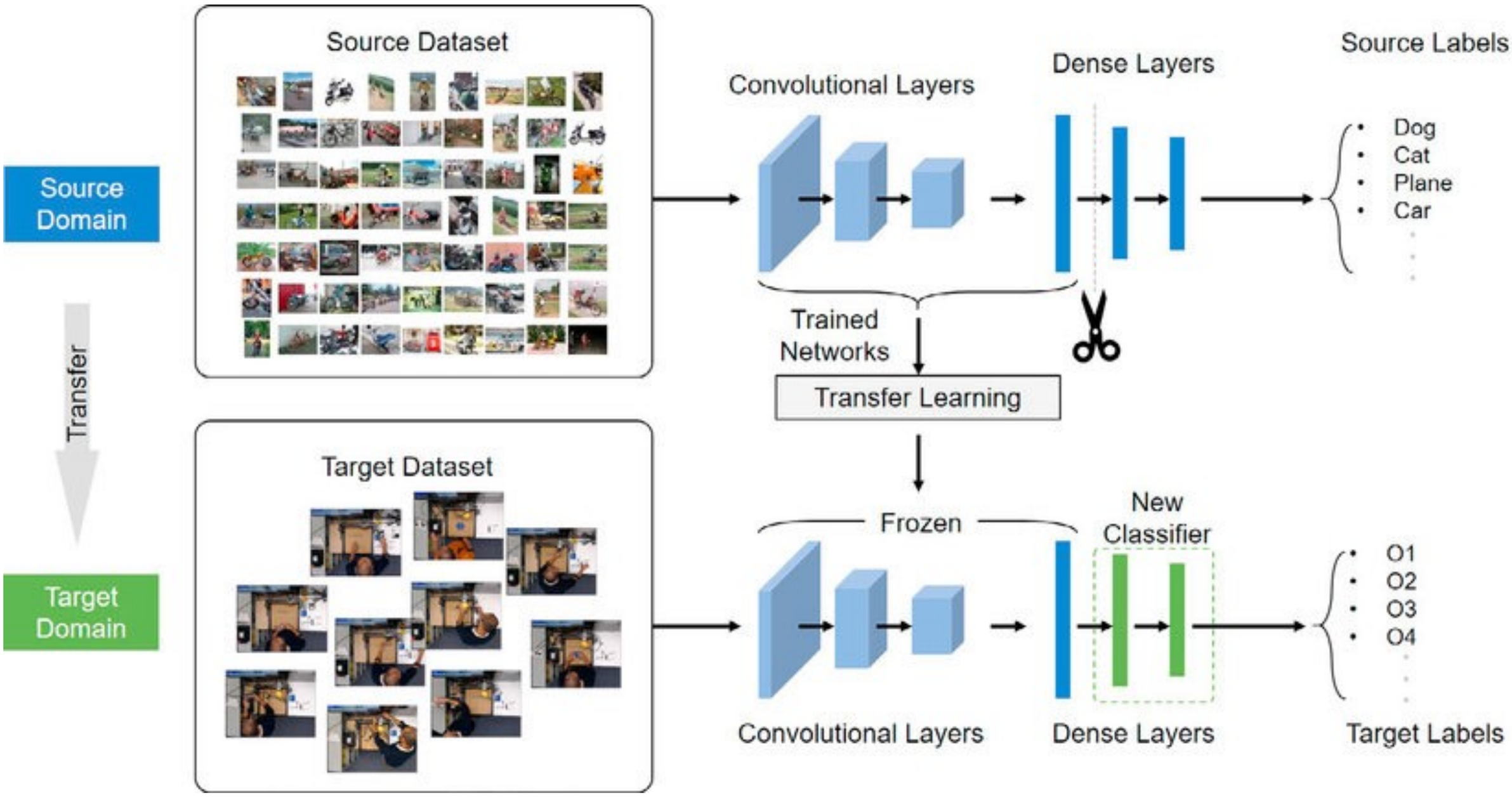
i.e. swap the Softmax layer at the end



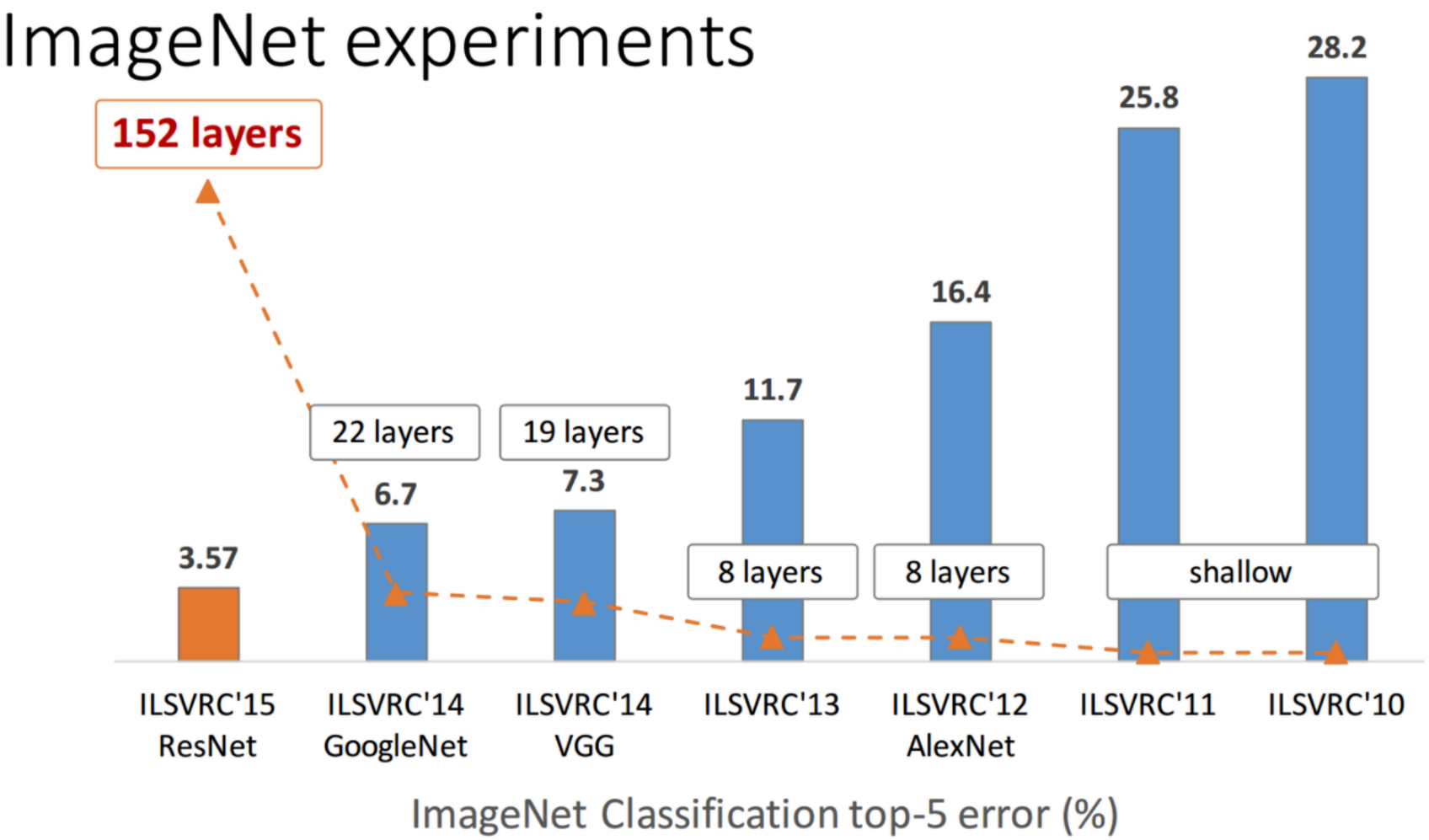
3. If you have medium sized dataset, **"finetune"** instead: use the old weights as initialization, train the full network or only some of the higher layers

retrain bigger portion of the network, or even all of it.

Transfer Learning



Deeper is better ?

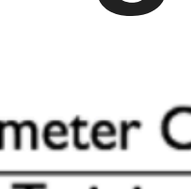


Larger is better

Parameter Count
Num Training Samples

Overparametrization

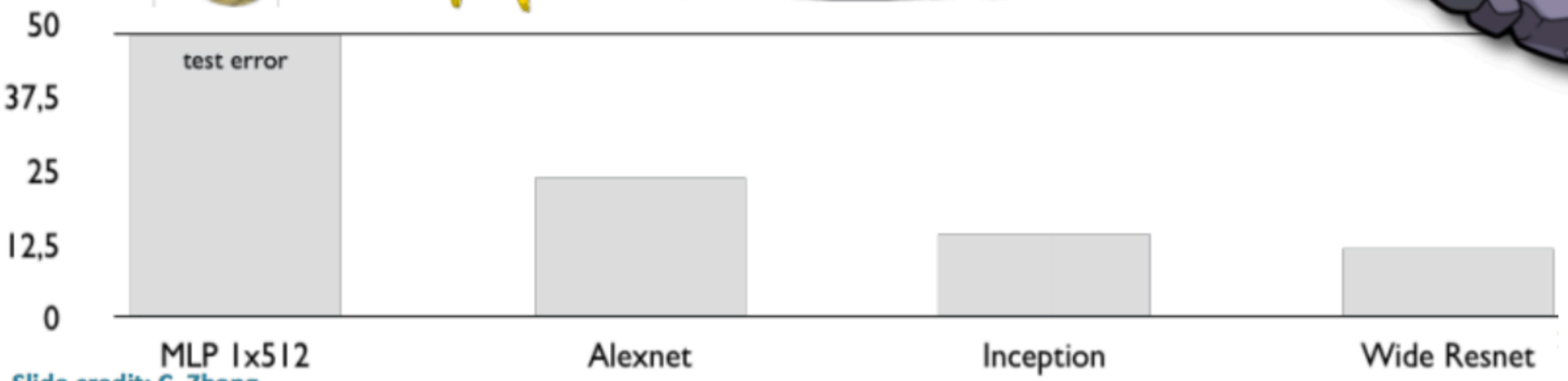
MLP 1x512
p/n: 24



Alexnet
p/n: 28

Inception
p/n: 33

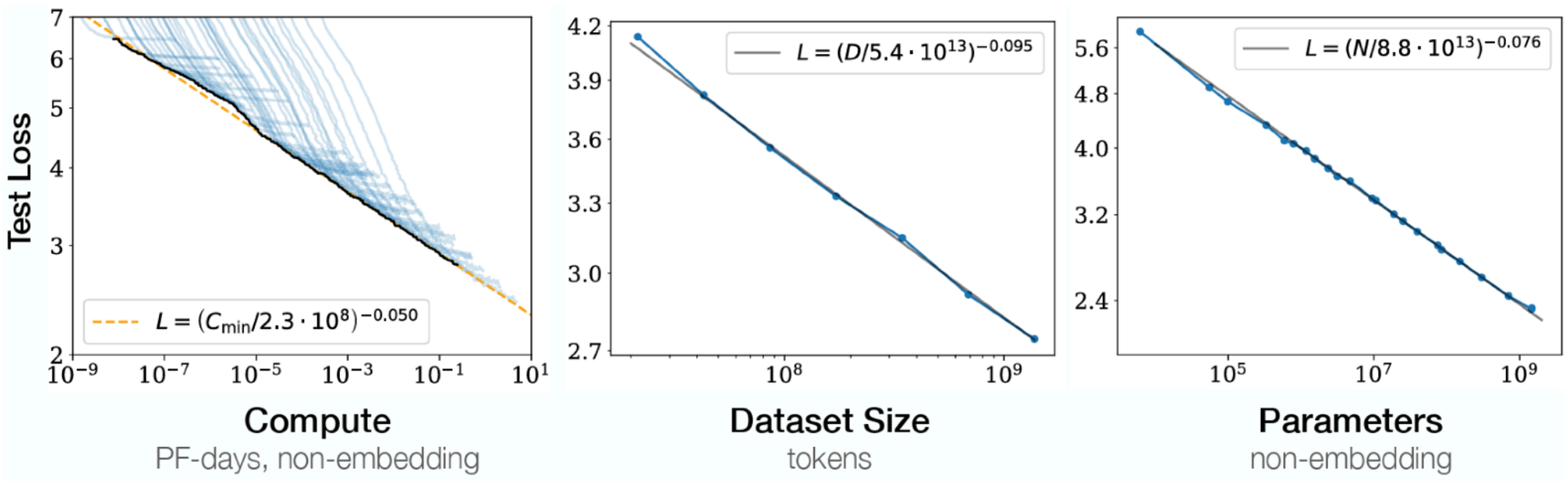
Wide Resnet
p/n: 179



Slide credit: C. Zhang

Scaling drives AI forward …

- More compute (GPUs), larger datasets, more trainable parameters.
- **Scaling-laws are the key drive of current AI industry.**



Kaplan et al.’20

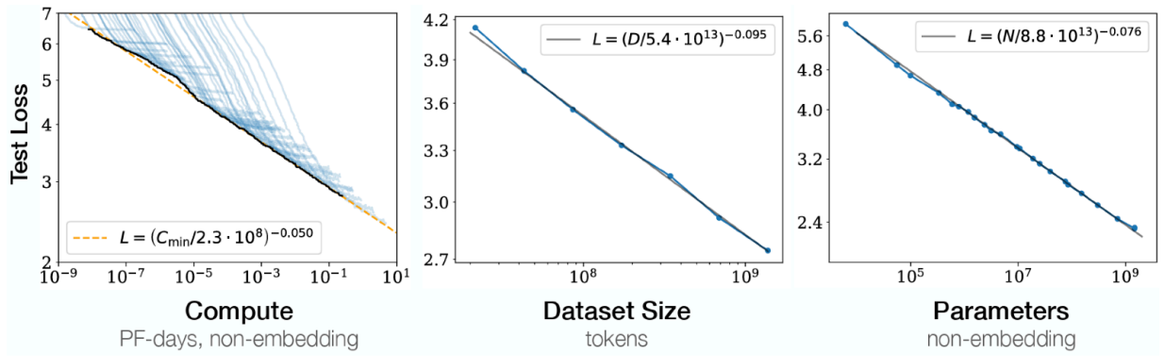
… but related cost pulls AI back!

Scaling brings more questions ….

- **Performance follows scaling laws.**

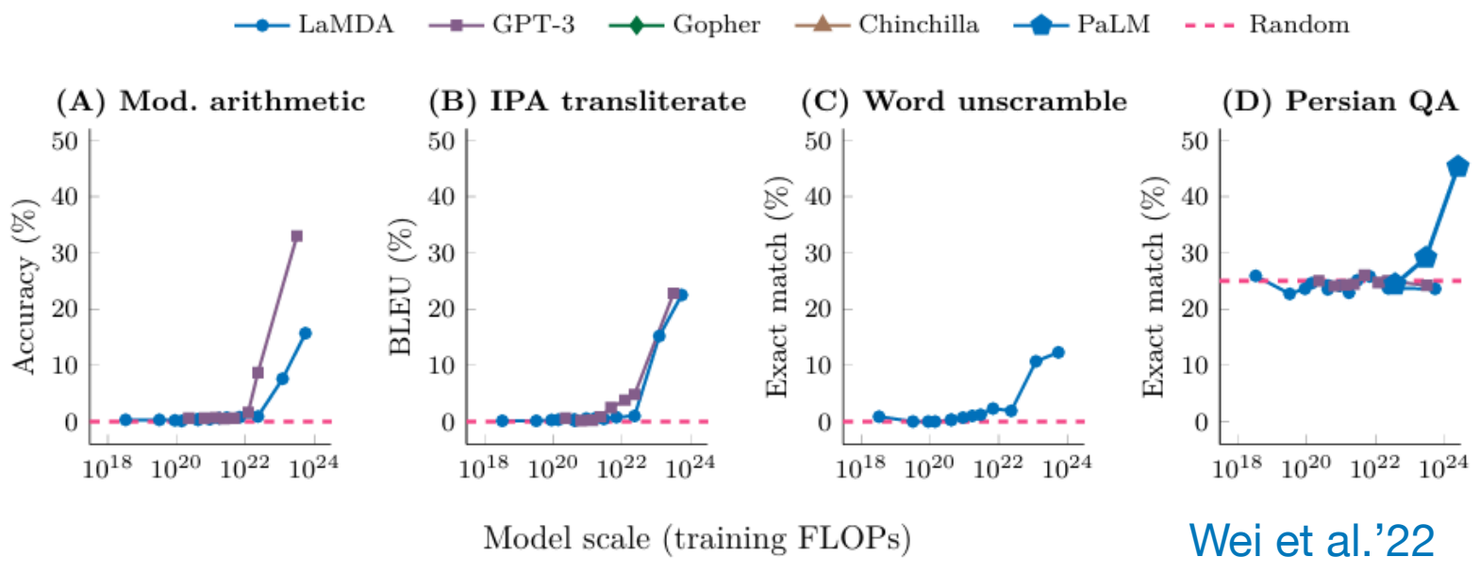
  ➡ Can we derive the exponents?

Kaplan et al.'20



- **Capabilities emerge with scale.**

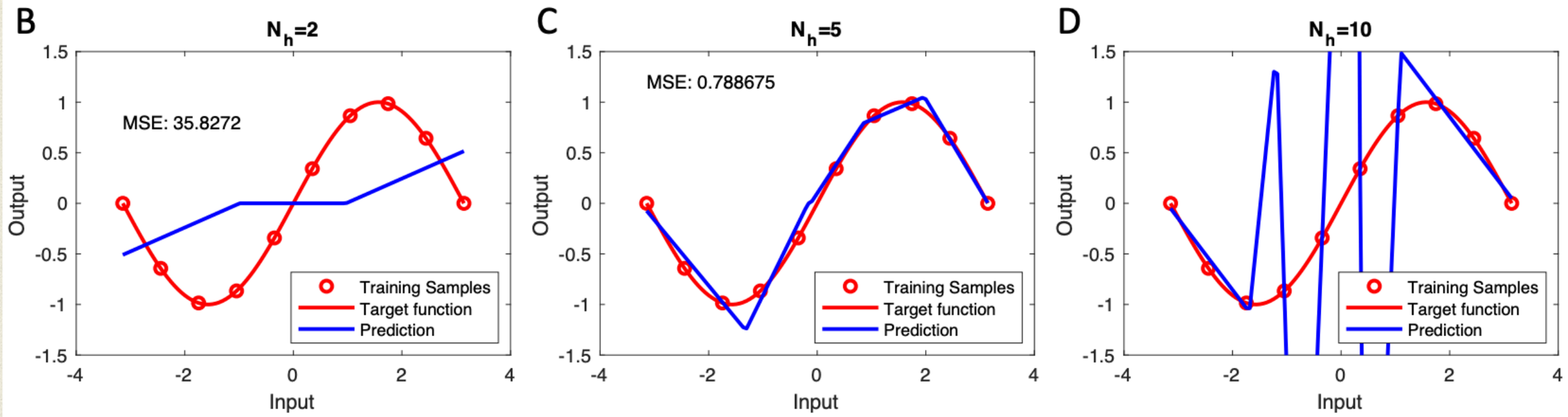
  ➡ Is this a phase transition?

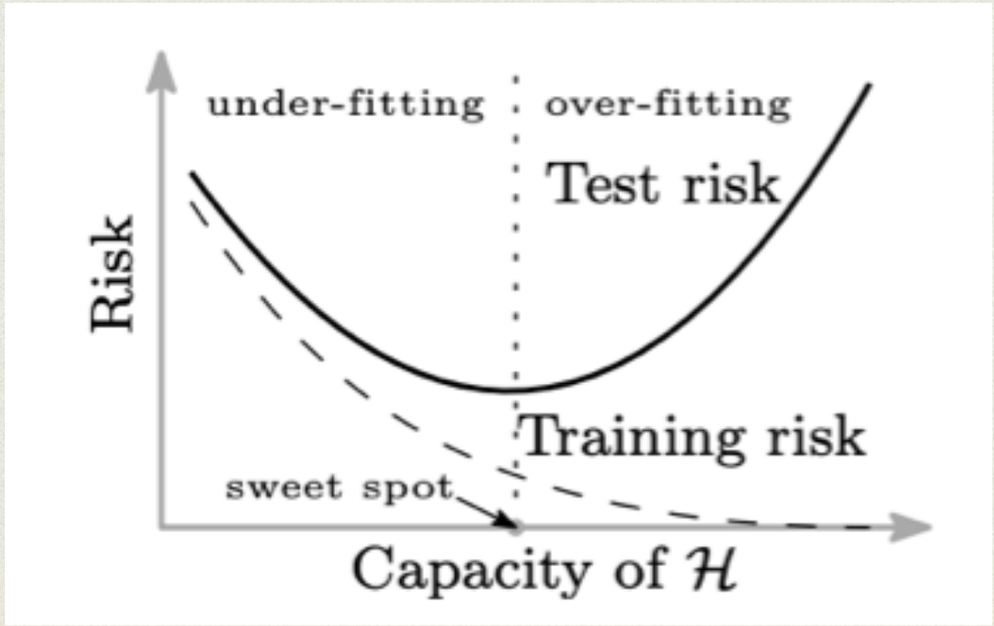


Wei et al.'22

- **The curse of dimensionality =** training neural networks is NP-hard (Blum, Rivest'89).

  ➡ Can we characterize when hardness can be avoided?


B
N_h=2
MSE: 35.8272
C
N_h=5
MSE: 0.788675
D
N_h=10
Output
Input
Training Samples
Target function
Prediction


under-fitting
over-fitting
Test risk
Training risk
Risk
sweet spot
Capacity of $\mathcal{H}$

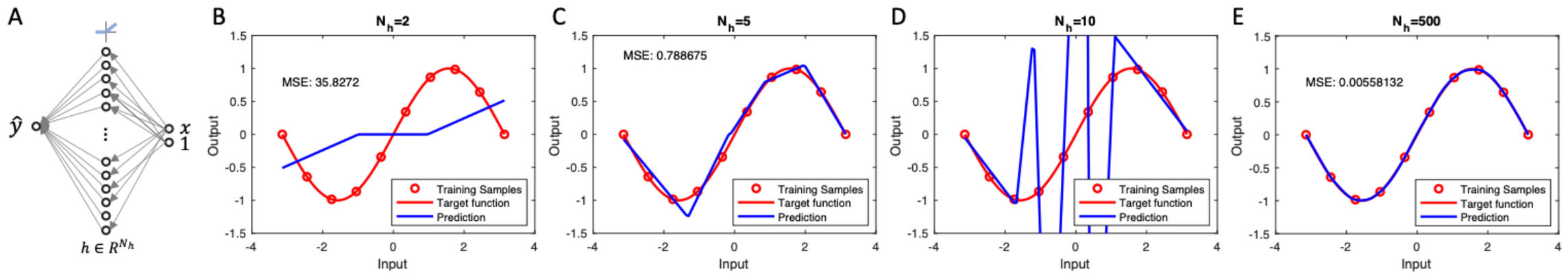
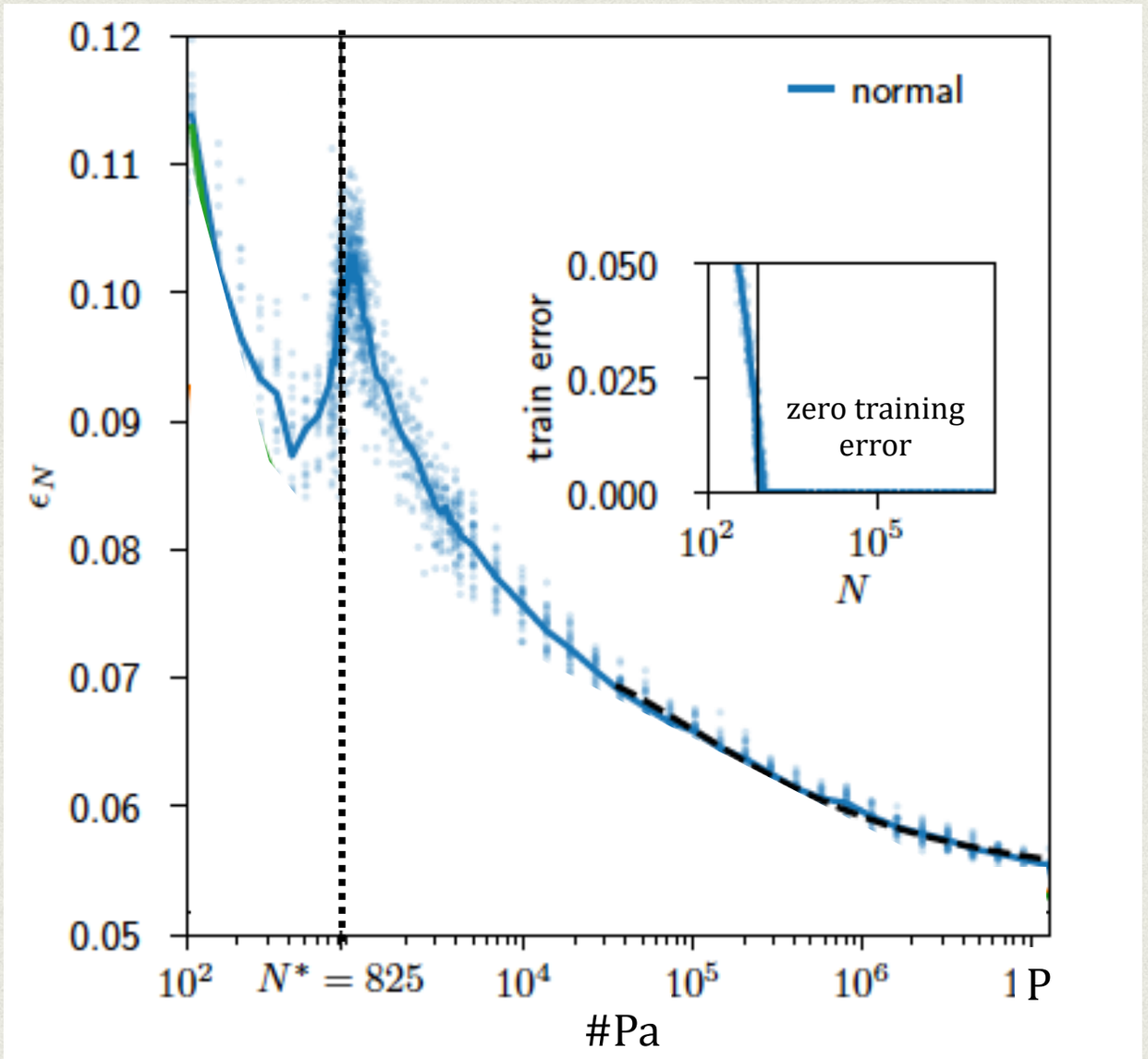


Figure 1: Over-parameterization and generalization via minimum norm solutions. (A) A simple ReLU network with random first layer weights and trained second layer weights. The network receives a scalar input and bias term, and is trained to minimize squared error on ten points (red circles) from a target sinusoid (red curve) shown in panels B-E. (B-E) Blue curves show example functions learned by networks with differing numbers of hidden units $N_h = \{2, 5, 10, 500\}$. Networks in panels B, C, & D show the standard progression from underfitting the training data to overfitting, with a happy medium in panel C. However the large network in panel E generalizes best. This network has $50\times$ more parameters than training examples but generalizes well, because among the infinity of solutions attaining zero training error we have chosen a low norm solution. In this work we derive training algorithms for nonlinear deep networks that explicitly seek minimum norm solutions in the underdetermined, accurate label regime, allowing good generalization in large networks.

Saxe, Advani'17



Parity-MNIST, 5 layers, FCN, hinge loss, no regularisation

[Geiger et al. '18]

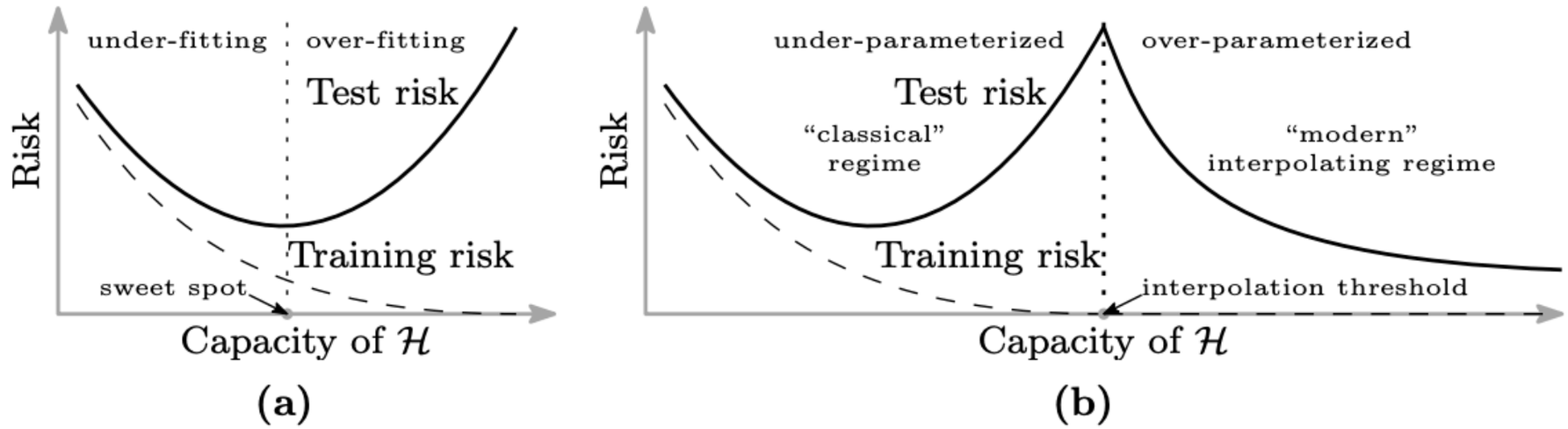


Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

Belkin, Hsu, Ma, Mandal'19

UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

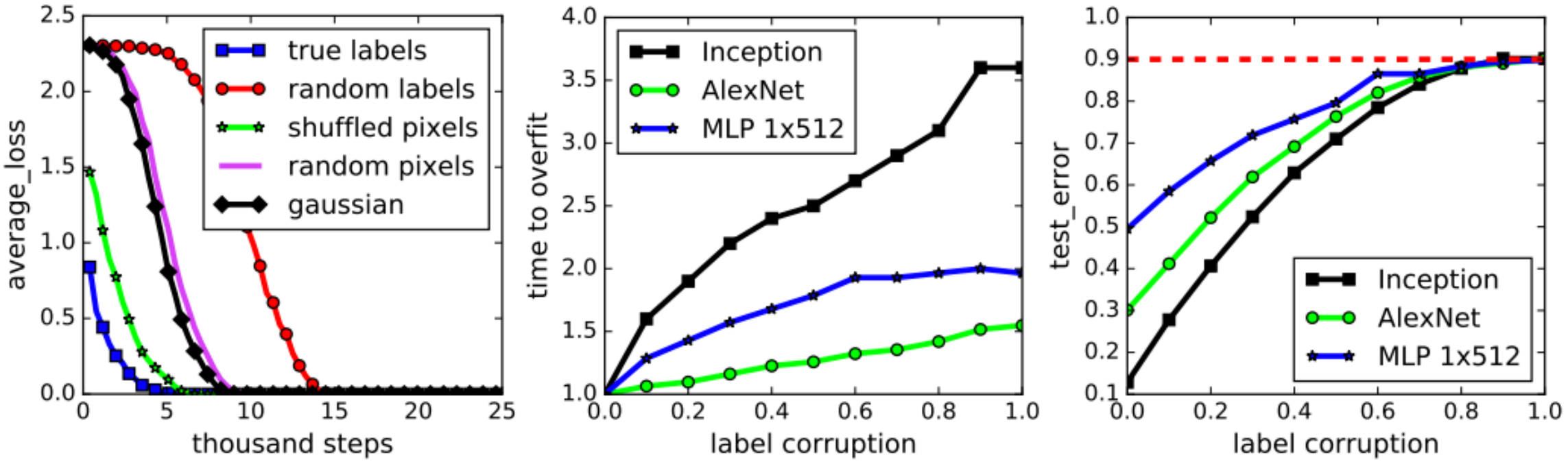
Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

- State-of-the-art neural networks are able to fit random labels.
- Classical bounds on the generalisation error (VC, Rademacher) are void, as they rely on not being able to fit random labels.



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Bad Global Minima Exist and SGD Can Reach Them

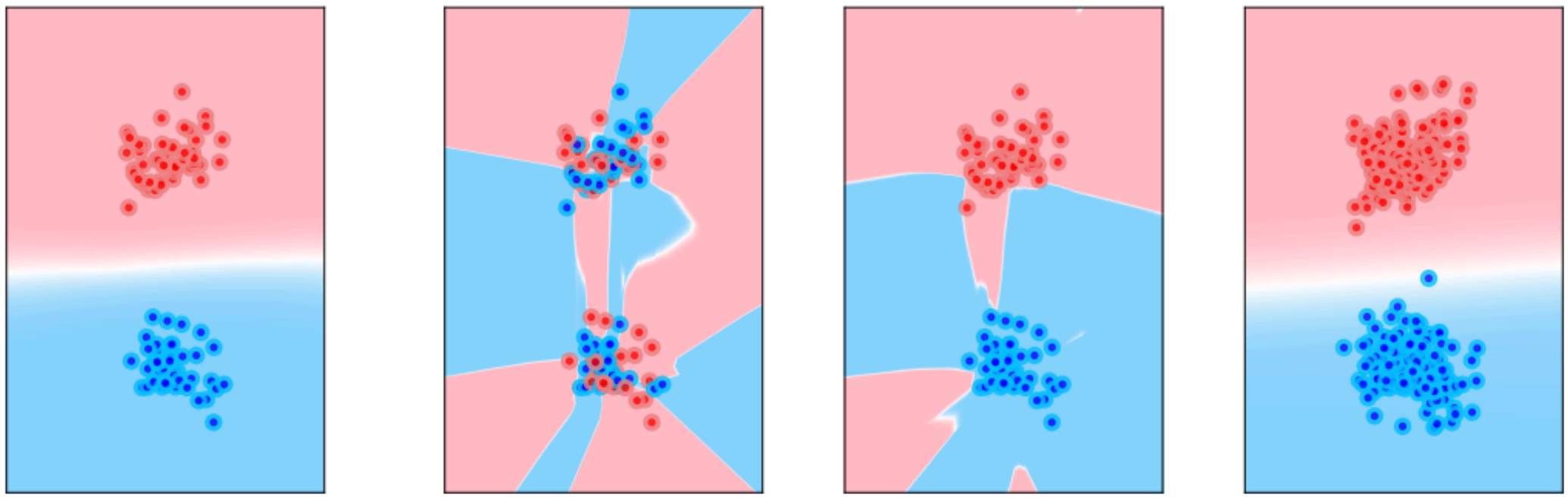
Shengchao Liu, Dimitris Papailiopoulos
University of Wisconsin–Madison

Dimitris Achlioptas
University of California, Santa Cruz

Abstract

Several recent works have aimed to explain why severely overparameterized models, generalize well when trained by Stochastic Gradient Descent (SGD). The emergent consensus explanation has two parts: the first is that there are "no bad local minima", while the second is that SGD performs implicit regularization by having a bias towards low complexity models. We revisit both of these ideas in the context of image classification with common deep neural network architectures. Our first finding is that there exist bad *global* minima, *i.e.*, models that fit the training set perfectly, yet have poor generalization. Our second finding is that given only *unlabeled* training data, we can easily construct initializations that will cause SGD to quickly converge to such bad global minima. For example, on CIFAR, CINIC10, and (Restricted) ImageNet, this can be achieved by starting SGD at a model derived by fitting random labels on the training data: while subsequent SGD training (with the correct labels) will reach zero training error, the resulting model will exhibit a test accuracy degradation of up to 40% compared to training from a random initialization. Finally, we show that regularization seems to provide SGD with an escape route: once heuristics such as data augmentation are used, starting from a complex model (adversarial initialization) has no effect on the test accuracy.

1. Random initialization + Training with true labels.
2. Random initialization + Training with random labels.
3. Random initialization + Training with random labels + Training with true labels.
4. Random initialization + Training with random labels + Training with true labels using data augmentation[1] and l_2 regularization.

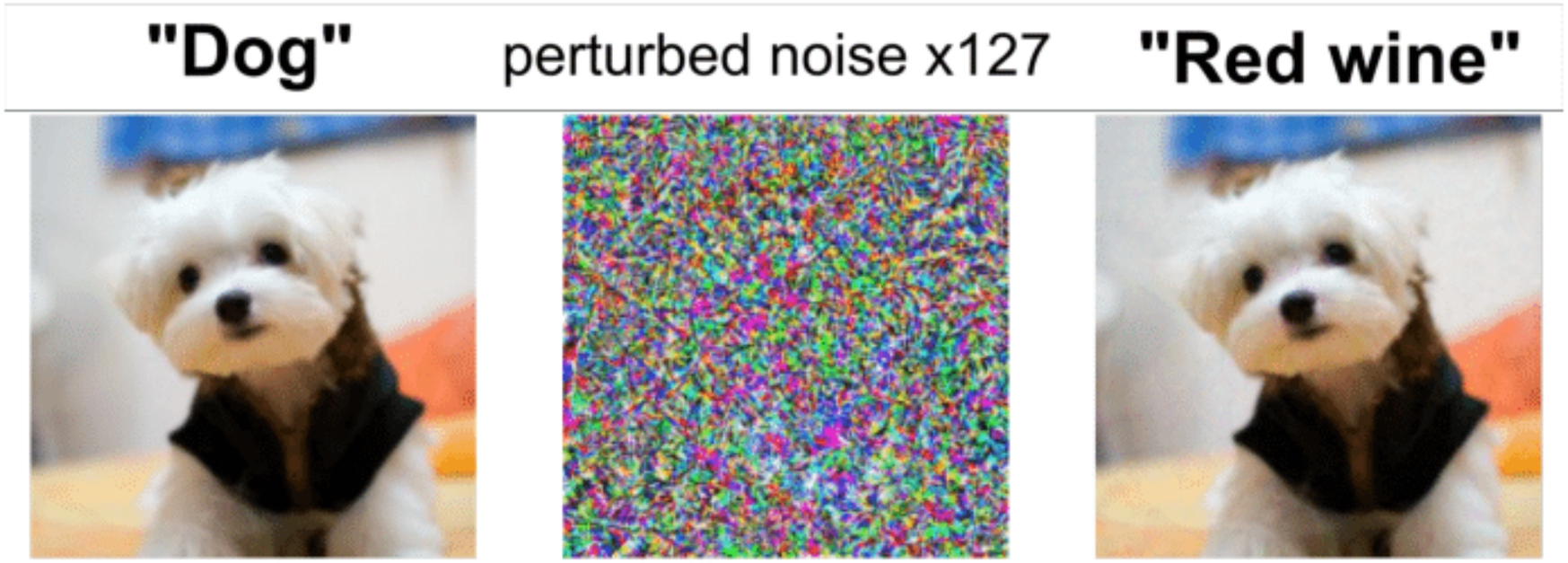


(a) Setting 1 (b) Setting 2 (c) Setting 3 (d) Setting 4

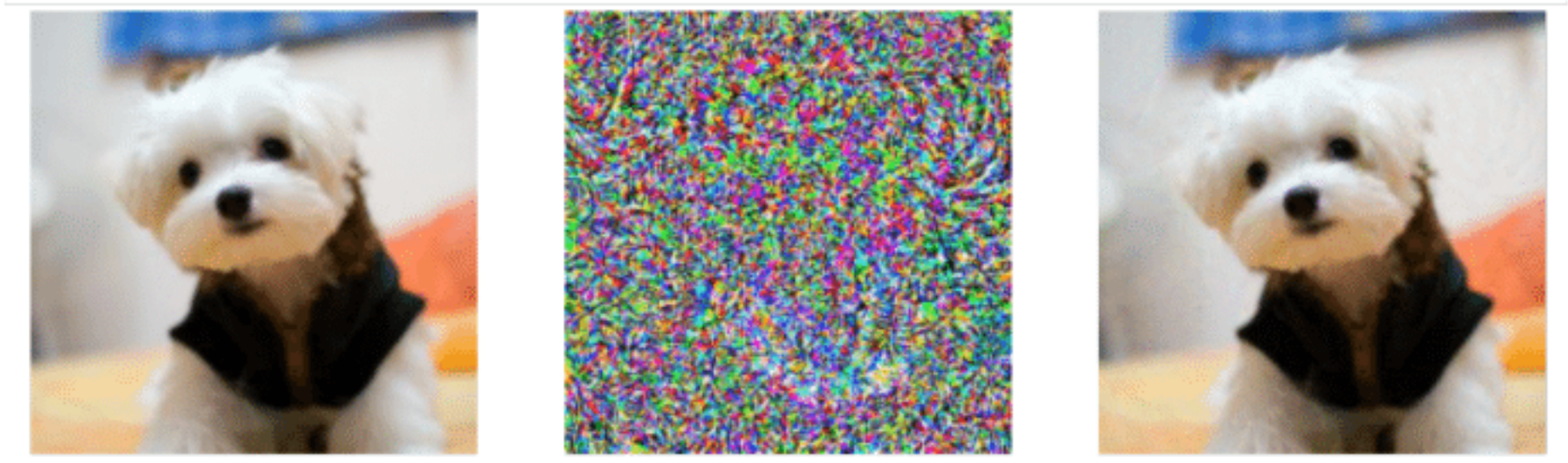
Figure 1: The decision boundary of the model reached by SGD in Settings 1–4, respectively.

Adversarial examples

a)



b)



Adversarial examples

classified as

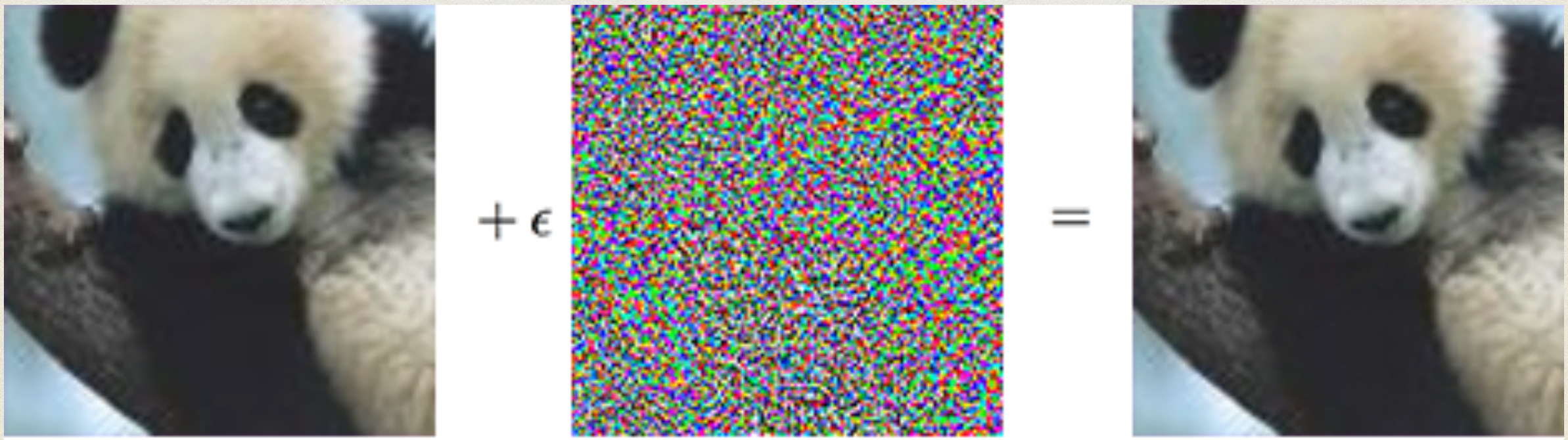
Stop Sign

+

=

classified as

Max Speed 100



What does this mean?

Empirical risk minimisation = minimisation of a loss

$$\min_w \Big[\frac{1}{n}\sum_{\mu=1}^{n} \ell(y_\mu, f_w(X_\mu)) + \lambda\|w\|_2^2\Big]$$

- square loss: $\ell(y,z) = (y-z)^2$,
- logistic loss: $\ell(y,z) = \log_2(1+e^{-yz})$
- cross-entropy loss for K-class classification:

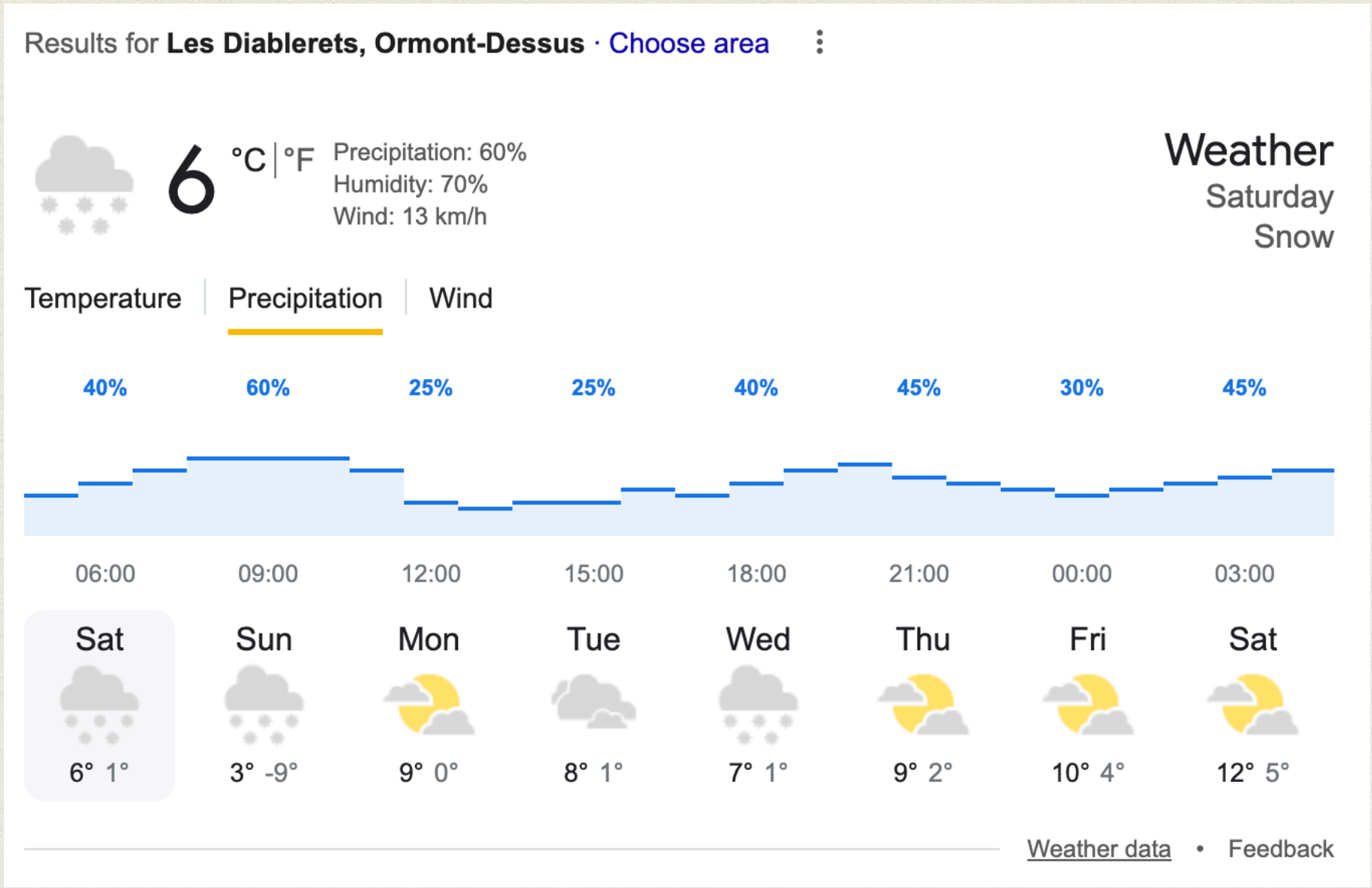
$$\ell(y, z(X)) = -\sum_{a=1}^{K} y_a \log \frac{e^{z_a(X)}}{\sum_b e^{z_b(X)}}$$

$$p_a(X) = \frac{e^{z_a(X)}}{\sum_b e^{z_b(X)}}$$

cross-entropy scores

- Observation: SOTA deep nets are over-parametrized.
- Over-parametrization leads to zero training loss.
- Cross-entropy is zero only if the scores p_a put all probability to one class.
- Early stopping usually prevent cross-entropy to go to zero, but still gives a large score to the most probable class.

over-parametrised = scores overconfident



For times where we predict 25% of rain it should rain 25% of the time.

On calibration of modern neural networks

C Guo, G Pleiss, Y Sun… - … conference on machine …, 2017 - proceedings.mlr.press

Confidence calibration–the problem of predicting probability estimates representative of the true correctness likelihood–is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art architectures …

☆ Save Cite Cited by 3445 Related articles All 9 versions »

Model predicts non-calibrated $p_a(X)$

Define:
$$p_a(X,T) \equiv \frac{e^{\frac{1}{T}z_a(X)}}{\sum_b e^{\frac{1}{T}z_b(X)}}$$

Choose T so as to minimise the validation loss.

$$\mathscr{L}(\{X_\mu, y_\mu\}_{\mu=1}^{n_{\text{val}}}) = -\frac{1}{n_{\text{val}}} \sum_{\mu=1}^{n_{\text{val}}} \sum_{a=1}^{K} y_{\mu,a} \log p_a(X_\mu, T)$$

- Bayesian Deep Learning (Gal, Ghahramani'15, Wilson, Hu, Salakhutdinov, Xing'16)
- Deep Ensembles (Lakshminarayanan, Pritzel, Blundell'16)
- Temperature Scaling (Guo, Pless, Sun, Weinberger'17)
- Bootstrap, Jacknife, Conformal prediction.
- Many many more …

Yet ChatGPT still hallucinates!!