Overview of Modern ML Applications: Convolutional Neural Network (CNN)

Learning objectives

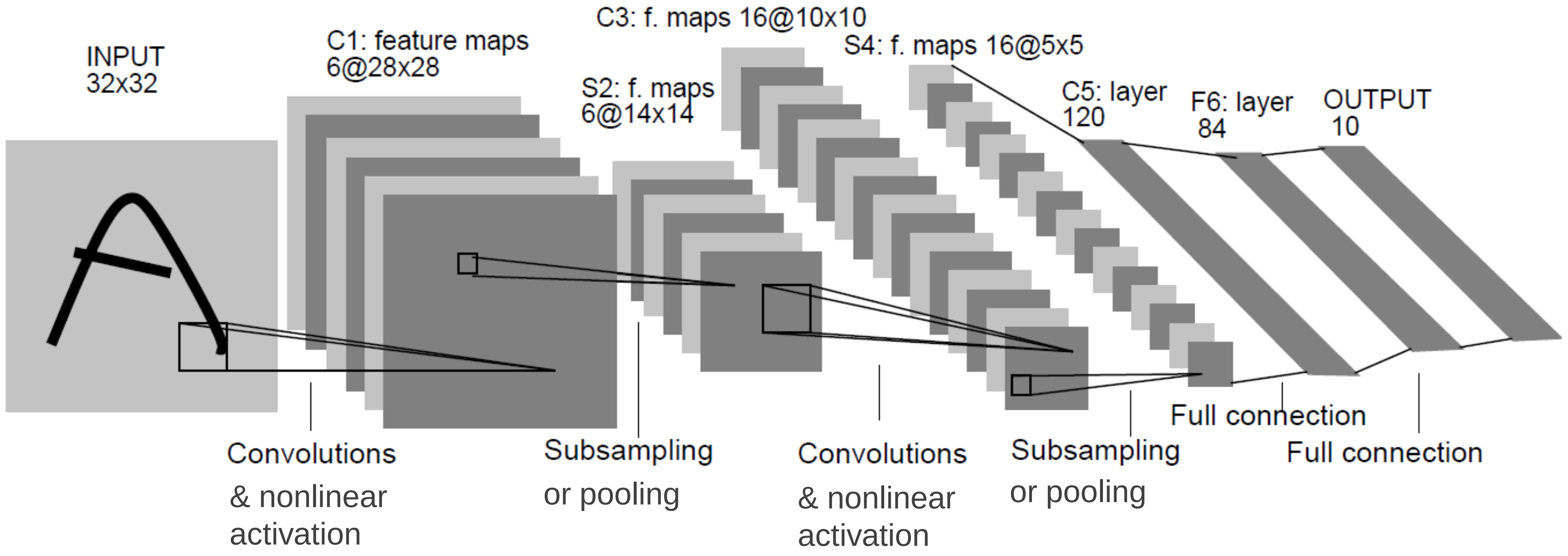
- Describe the structure of CNN
- Build and train simple CNNs using a deep learning package

(Ref: Ch 9 of Goodfellow et al. 2016)

Acknowledgment: Some graphics and slides were adapted from Stanford's CS231n by Fei-Fei Li et al.: http://cs231n.stanford.edu/

Convolutional Neural Network (CNN)

The **single** most important technology that fueled the rapid development of ***deep learning*** and ***big data*** in the past decade.



LeCun, Bottou, Bengio, Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE*, **1998**.

Why is Deep Learning so Successful?

1. **Improved model:** convolutional layer, more layers (“deep”), simpler activation (i.e., ReLU), skip/residual connection (i.e., ResNet), attention (i.e., Transformer)
2. **Big data:** huge dataset, transfer learning
3. **Powerful computation:** graphical processing units (GPUs)

◆ Example of big data: ImageNet (22K categories, 15M images)

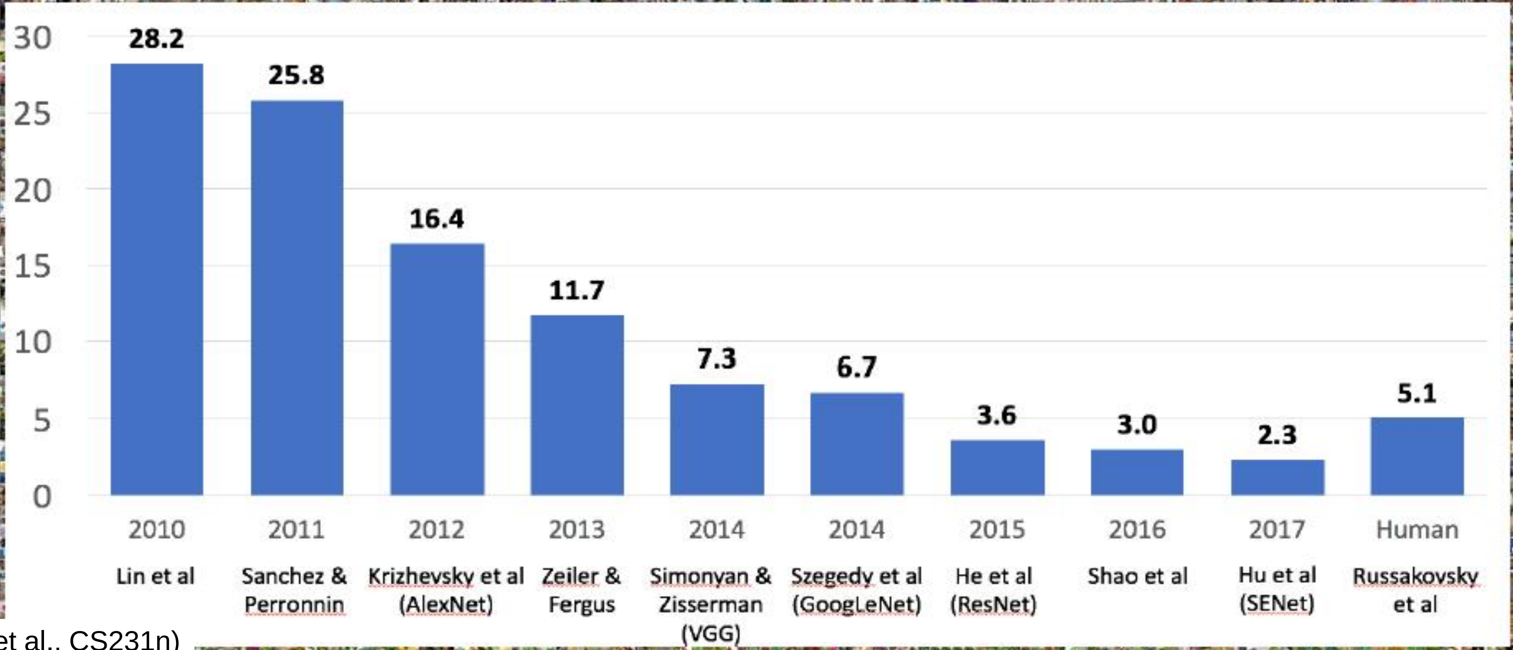
Deng, Dong, Socher, Li, Li & Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” *IEEE CVPR*, 2009.

IMAGENET Large Scale Visual Recognition Challenge

The Image Classification Challenge:

1,000 object classes

1,431,167 images



(Fei-Fei Li et al., CS231n)

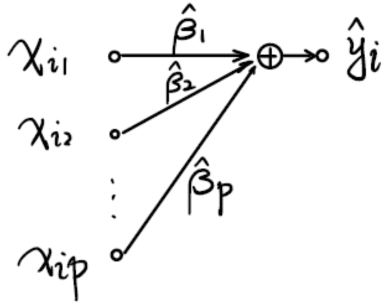
Linear Model to Neural Network

Recall linear model w/ multiple predictors / features / inputs.

$$\underbrace{y_i}_{\text{true output}} = \sum_{j=1}^{p} x_{ij}\beta_j + e_i = \underbrace{[\beta_1, \dots, \beta_p]}_{\text{true weights}} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} + e_i \quad , i = 1, \dots, n.$$

$$\underbrace{\hat{y}_i}_{\text{predicted output}} = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j = \underbrace{[\hat{\beta}_1, \dots, \hat{\beta}_p]}_{\text{estimated weights}} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \quad , i = n+1, \dots, n+m.$$

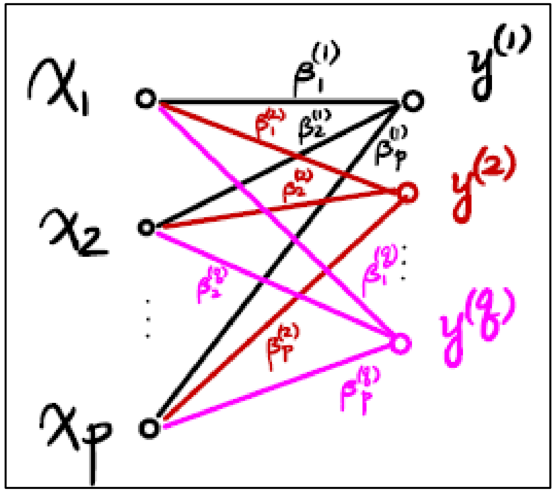
Graphically we have:



① Use multiple linear models

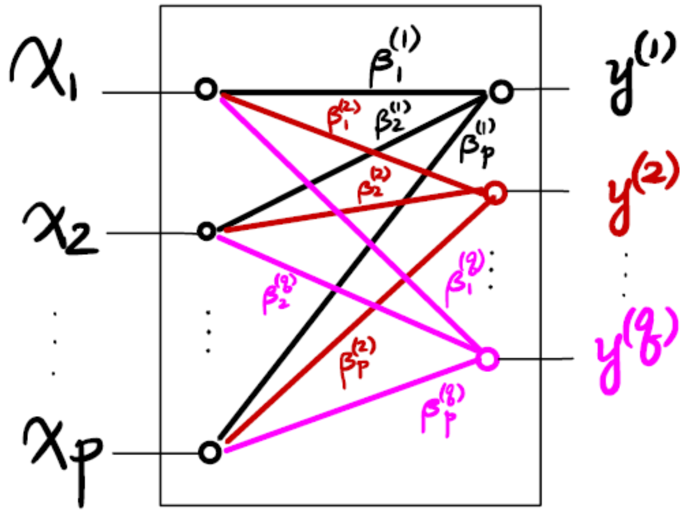
$\longrightarrow$

② Simplify the notations.



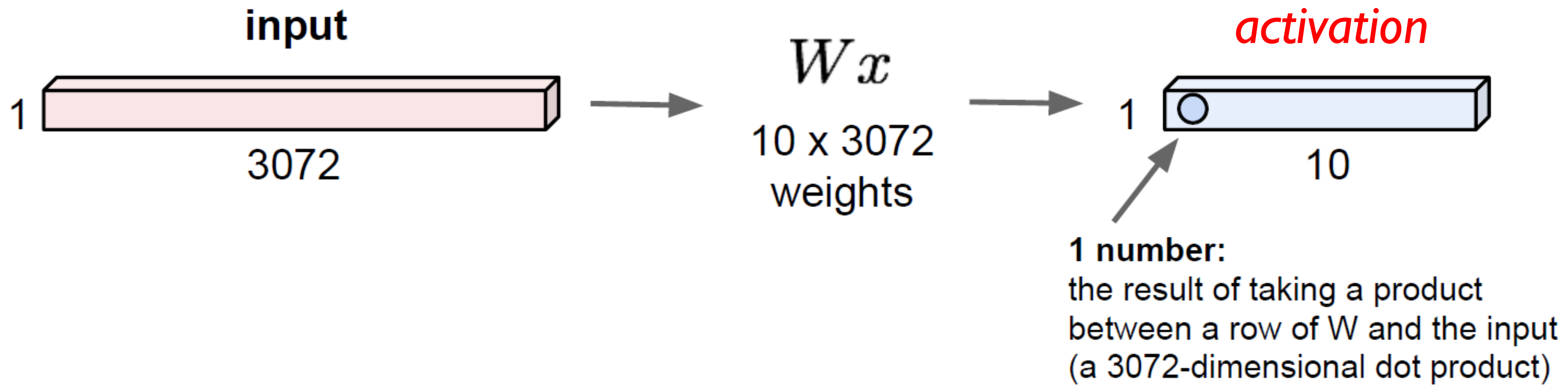
Fully-Connected Layer for 1D Signal

$$\underbrace{\begin{bmatrix} y^{(1)} \\ \hdashline y^{(2)} \\ \hdashline \vdots \\ \hdashline y^{(q)} \end{bmatrix}}_{\substack{\text{layer output,} \\ \underset{\sim}{y} \in \mathbb{R}^{q}}} = \underbrace{\begin{bmatrix} \beta_1^{(1)} & \beta_2^{(1)} & \cdots & \beta_p^{(1)} \\ \hdashline \beta_1^{(2)} & \beta_2^{(2)} & \cdots & \beta_p^{(2)} \\ \hdashline \vdots & & & \vdots \\ \hdashline \beta_1^{(q)} & \beta_2^{(q)} & \cdots & \beta_p^{(q)} \end{bmatrix}}_{\substack{\text{dense weight matrix} \\ B \in \mathbb{R}^{q \times p}}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \\ x_p \end{bmatrix}}_{\substack{\text{layer input,} \\ \underset{\sim}{x} \in \mathbb{R}^{p}}}$$



Fully-Connected Layer for RGB Image

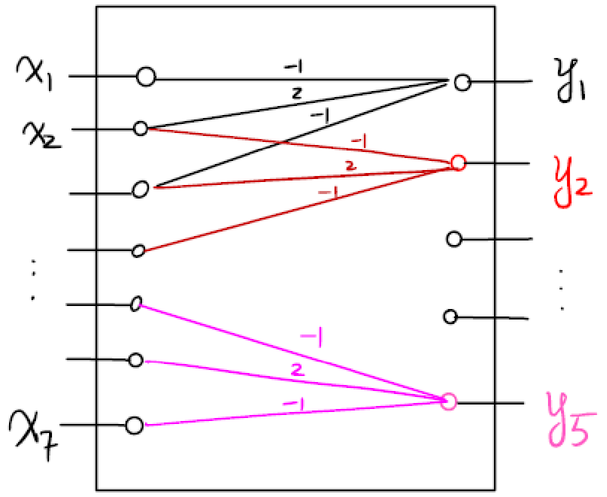
32x32x3 image -> stretch to 3072 x 1


input
1
3072
Wx
10 x 3072
weights
activation
1
10
1 number:
the result of taking a product between a row of W and the input (a 3072-dimensional dot product)

Convolutional Layer for 1D Signal

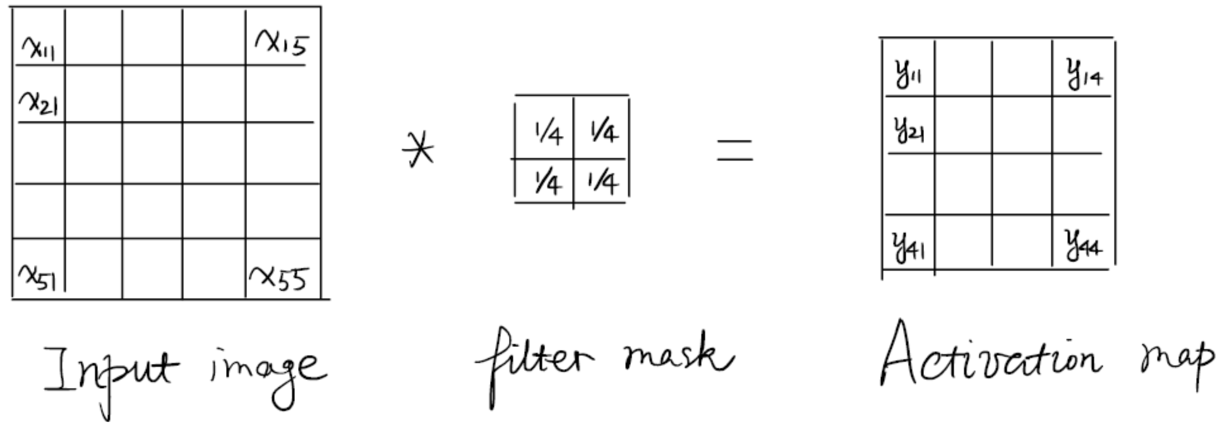
$$\begin{bmatrix} y_1 \\ \vdots \\ y_5 \end{bmatrix} = \begin{bmatrix} -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ \vdots & & & -1 & 2 & -1 & \vdots \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_7 \end{bmatrix}$$

Sparse weight matrix

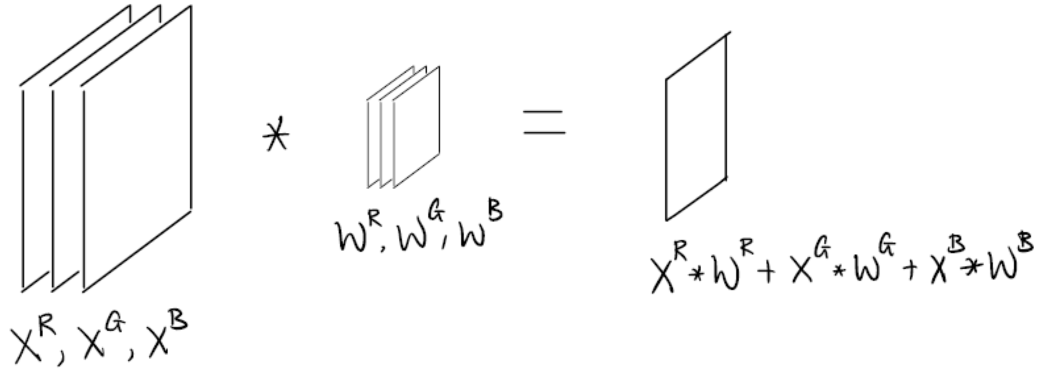


Input	x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7	length 7
Convolution filter mask	$*$ -1 | 2 | -1 $\longrightarrow$	length 3
Output (w/o boundary elements)	y_1 | y_2 | y_3 | y_4 | y_5	length $7-(3-1)=5$

Convolutional Layer for 2D Matrix/Image

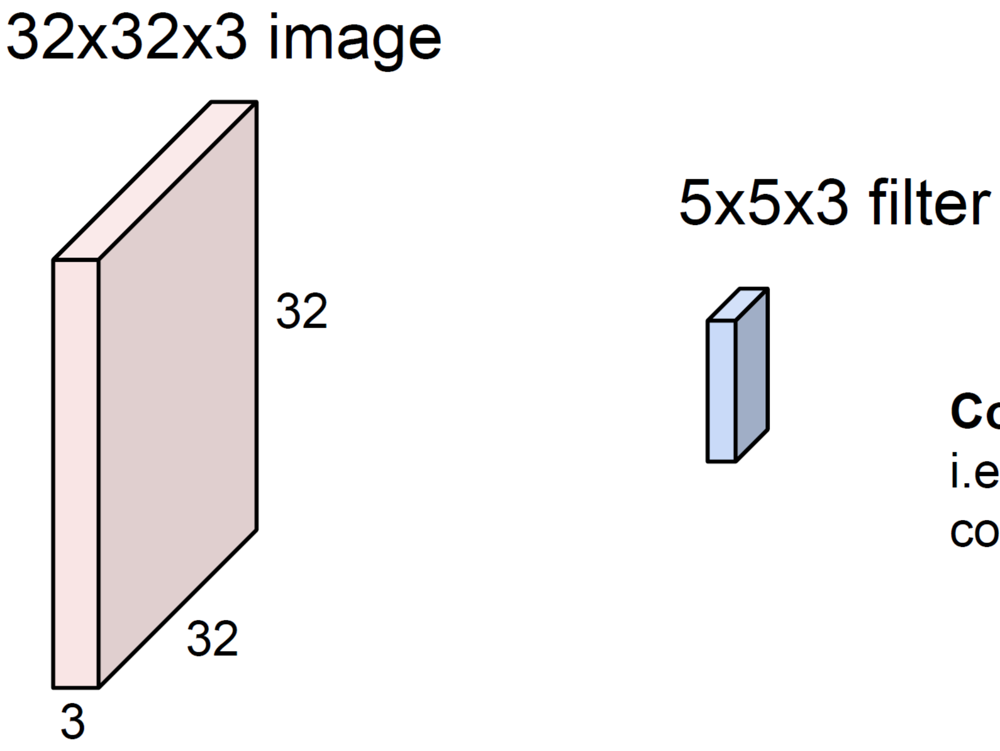


2D Convolution

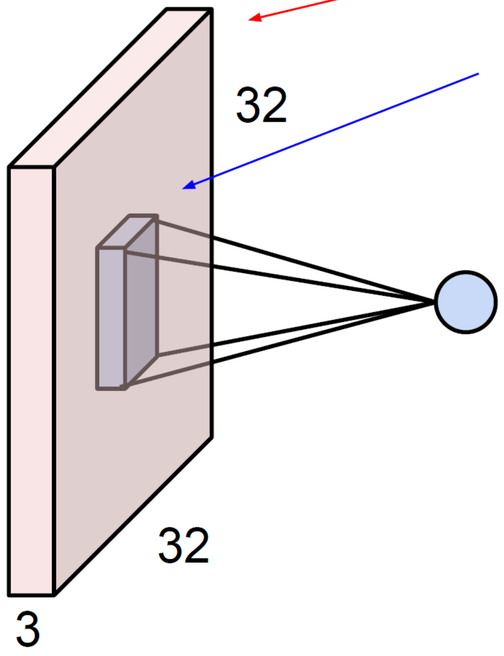


Multiple color channels need multiple filter masks

Convolutional Layer for RGB Image



Convolve the filter with the image
i.e. “slide over the image spatially, computing dot products”



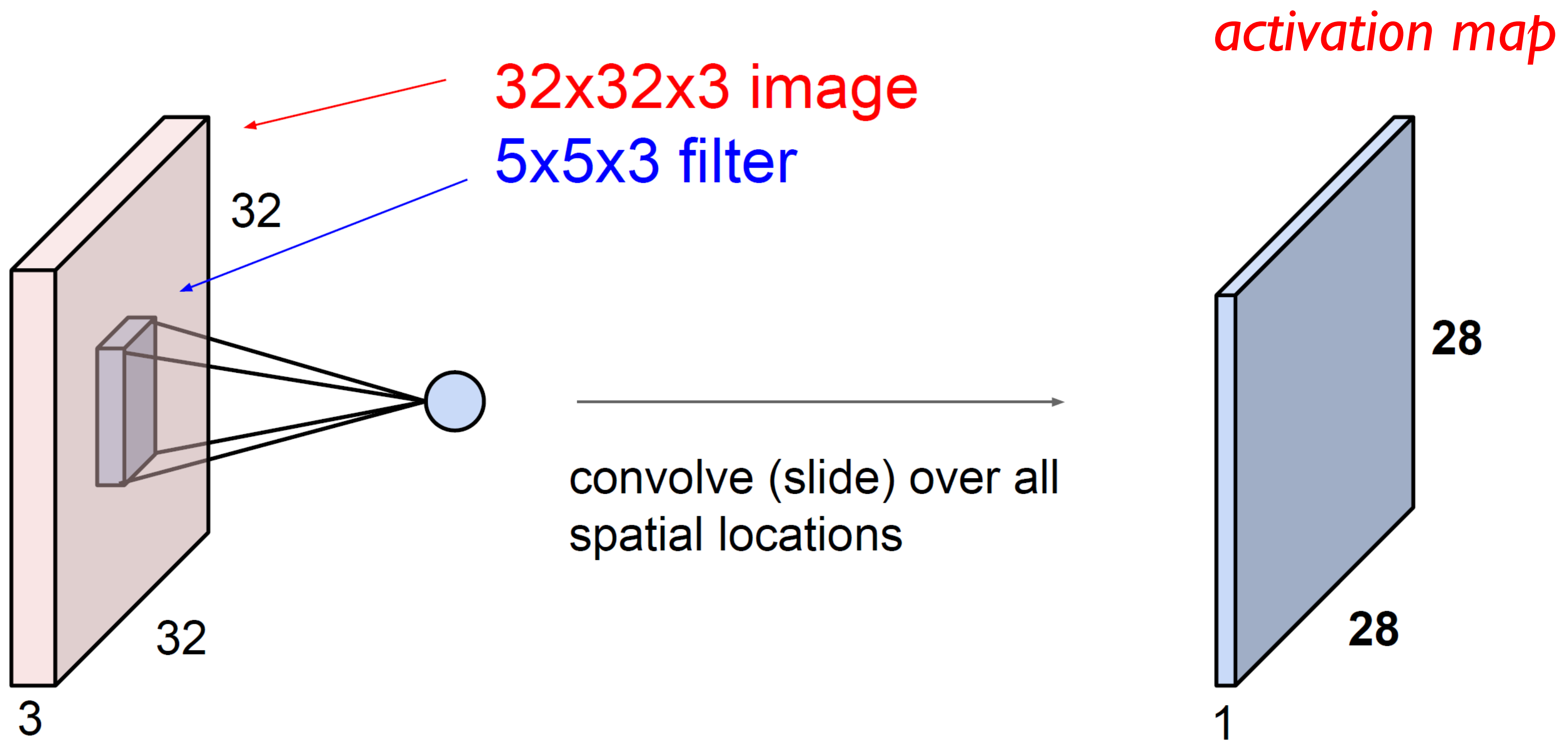
32x32x3 image

5x5x3 filter w

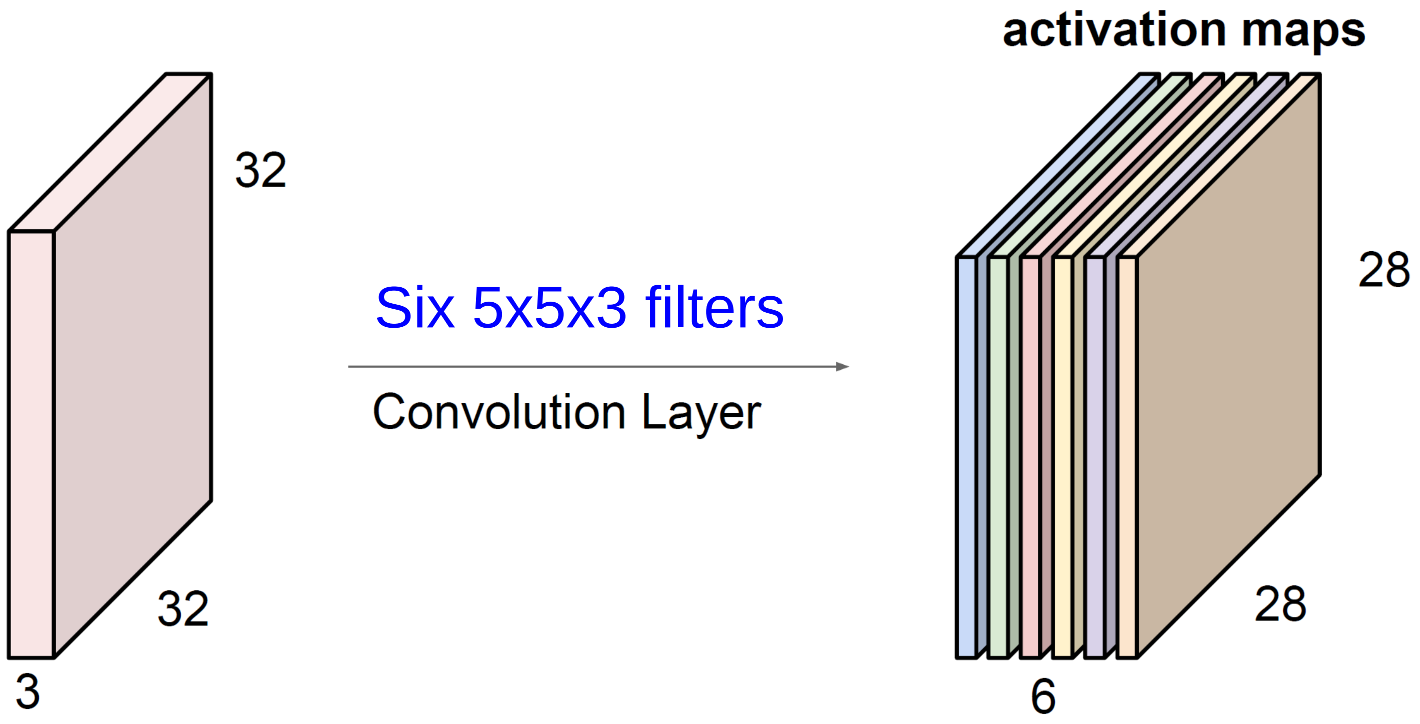
1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

$$w^T x + b$$

A closer look at spatial dimensions:

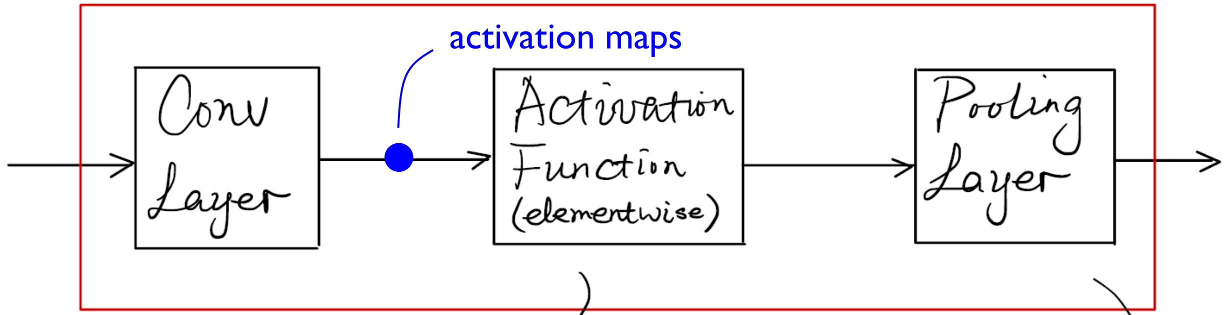


For example, if we had six 5x5 filters, we'll get six separate *activation maps*:



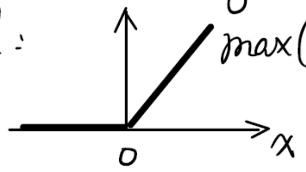
We stack these up to get a "new image" of size 28x28x6!

Building Block for Modern CNN



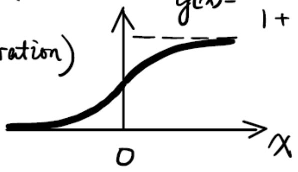
(Rectified Linear Unit)

ReLU: (modern)



$g(x) = \max(0, x)$

Logistic: (prior generation)



$g(x) = \frac{1}{1+e^{-x}}$

Ex:

x_{11}	x_{12}
x_{21}	x_{22}

: $\{x_{ij}\}_{i,j=1}^{2}$

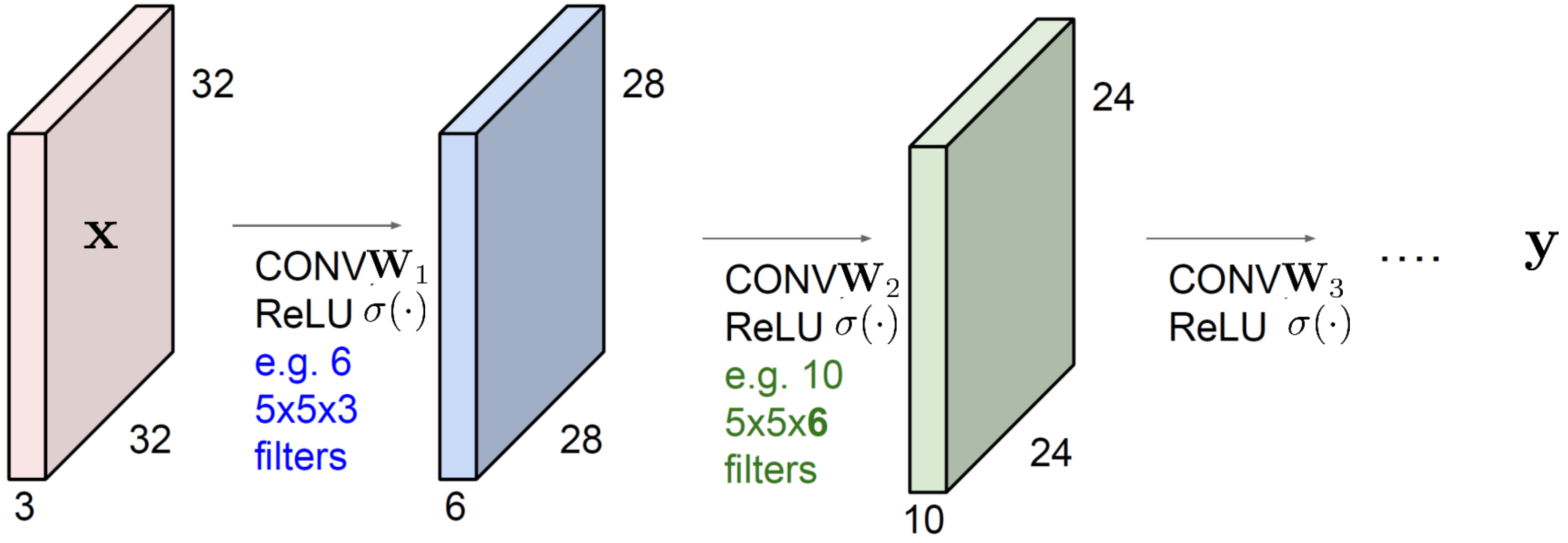
Max pooling:

$$g(\{x_{ij}\}) = \max(\{x_{ij}\})$$

Average pooling:

$$g(\{x_{ij}\}) = \frac{1}{|\{x_{ij}\}|} \sum_{i,j} x_{ij}$$

CNN is composed of a sequence of convolutional layers, interspersed with activation functions (ReLU, in most cases).

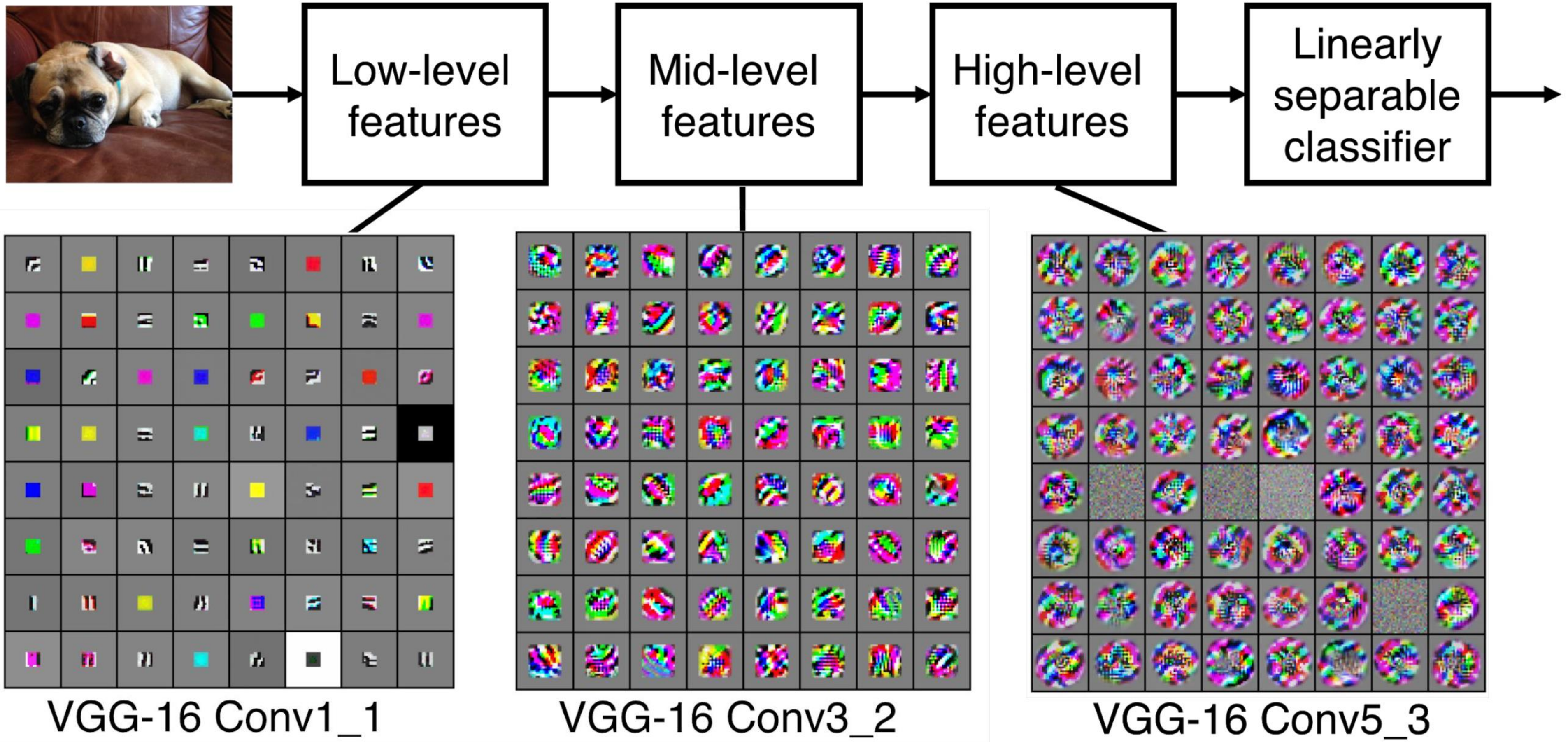


$$\mathbf{y} = \cdots \sigma\Big(\mathbf{W}_3\, \sigma\big(\mathbf{W}_2\, \sigma(\mathbf{W}_1 \mathbf{x})\big)\Big) \cdots$$

Source of nonlinearity, ReLU: $\sigma(x) := \max(0, x)$

[Zeiler and Fergus 2013]

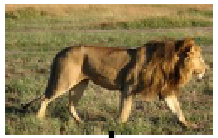
Visualization of VGG-16 by Lane McIntosh. VGG-16 architecture from [Simonyan and Zisserman 2014].

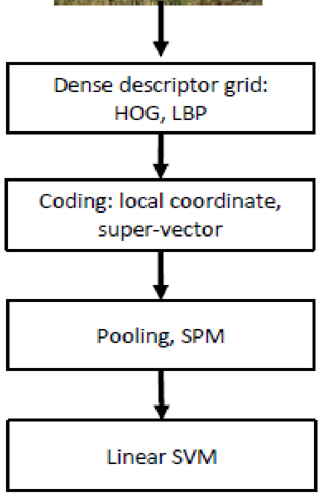


IMAGENET Large Scale Visual Recognition Challenge

Year 2010

NEC-UIUC




Dense descriptor grid: HOG, LBP
Coding: local coordinate, super-vector
Pooling, SPM
Linear SVM

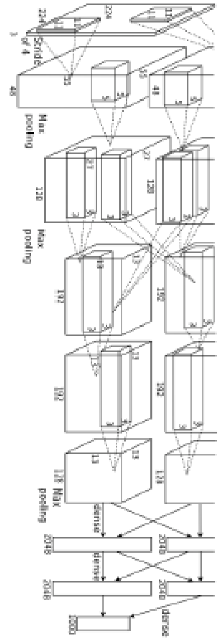
[Lin CVPR 2011]

Lion image by Swissfrog is licensed under CC BY 3.0

AlexNet

Year 2012

SuperVision

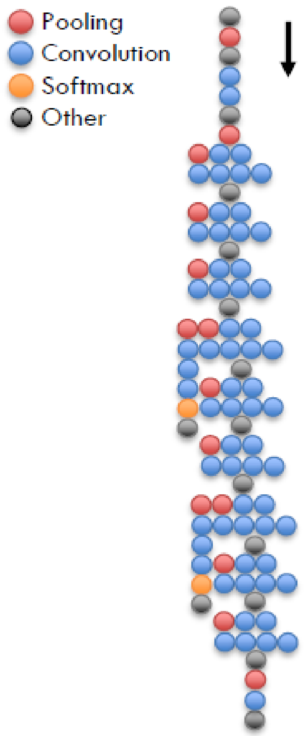


[Krizhevsky NIPS 2012]

Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

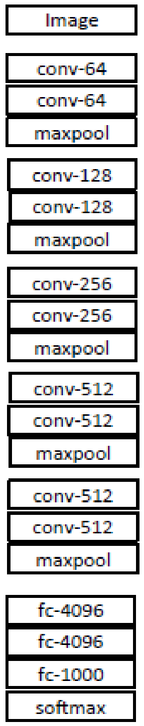
Year 2014

GoogLeNet



[Szegedy arxiv 2014]

VGG

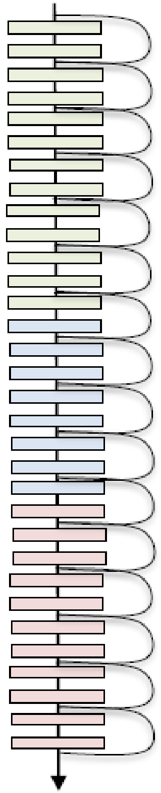


[Simonyan arxiv 2014]

ResNet

Year 2015

MSRA

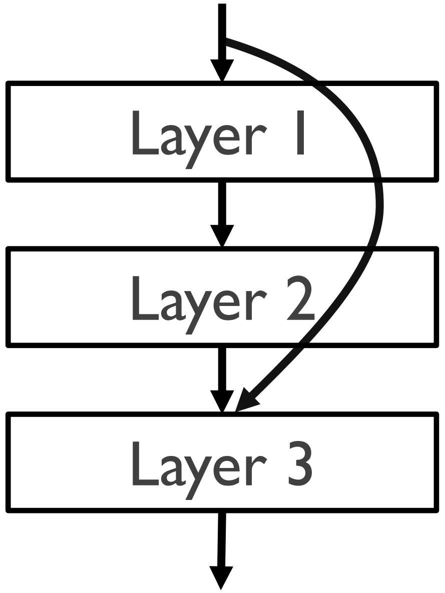


[He ICCV 2015]

(Fei-Fei Li et al., CS231n)

Residual Neural Network (ResNet) (Kaiming He et al., 2015)

- ***Skip connections*** or ***shortcuts*** are added.
- They can
  - avoid "vanishing gradients", and
  - make optimization landscape flatter.
- From Taylor expansion perspective, the neural network only learns the higher-order error terms beyond the linear term $\mathbf{x}$.
- Has interpretations in PDE.
- Preferred modern NN structure.



$$\mathbf{y} = \cdots \sigma\Big(\mathbf{W}_3 \underbrace{\left[\mathbf{x} + \sigma(\mathbf{W}_2\, \sigma(\mathbf{W}_1 \mathbf{x}))\right]}_{g(\mathbf{x})}\Big) \cdots$$

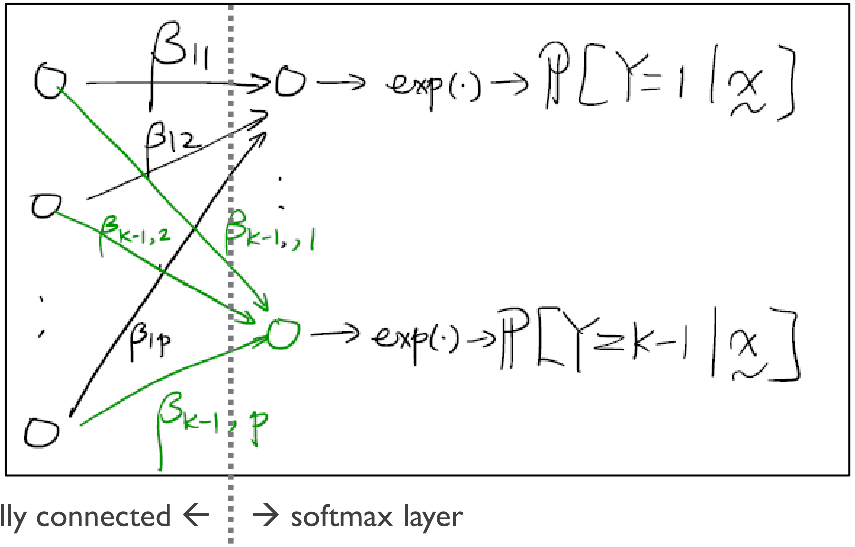
When Output is Categorical / Qualitative

- A ***softmax layer*** is needed:
- *Softmax function:*

$$\sigma_i(\underset{\sim}{z}) = \frac{e^{\beta z_i}}{\sum_{j=1}^{K} e^{\beta z_j}}$$

- Ex:

$$K=2. \quad \sigma_1 = \frac{e^{\beta z_1}}{e^{\beta z_1} + e^{\beta z_2}} = \frac{1}{1+e^{\beta(z_2-z_1)}}$$



When β very large,

$z_2 > z_1$ leads to $\begin{cases} \sigma_1 = 0 \\ \sigma_2 = 1 \end{cases}$

Winner takes all!

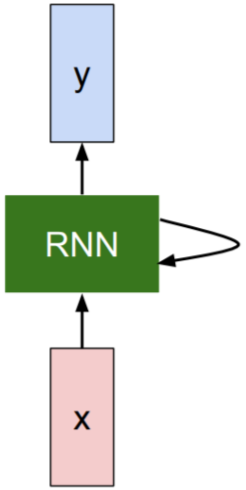
Other Essential Aspects of CNN

- Due to time constraints, this overview lecture covered only the structural elements of CNNs. Other essential aspects are:
  - Cost function/loss, e.g., MSE, cross entropy.
  - How to train CNNs or estimate the weights (will only give practice code), i.e., *backpropagation* (will cover in the next two lectures).
  - Practical training considerations including
    - How to determine *number of hidden units/channels* to be used,
    - How to tune *learning rate* and *batch size*, and
    - When to stop training (number of *epochs*).
- For a more complete treatment on CNN, refer to the dedicate courses such as CS231n CNNs for Visual Recognition or ECE 542/492.

Machine Learning (ML) and Data Science (DS)

- Follow-up machine learning / data science courses:
  - ECE 542/492 Neural Networks (S'23)
  - ECE 512 Data Science (each fall)
  - ECE 792-61 Advanced Topics in Machine Learning (S'23)
  - ECE 592-103 Large-Scale ML and Optimization
  - ECE 759 Pattern Recognition (S'24)
  - ECE 763 Computer Vision (S'24)
  - Any courses/videos on YouTube, Coursera, etc.
- State-of-the-art theory & applications: ICML, NeurIPS, ICLR, AAAI
- Data science competitions: kaggle.com
- Programming languages for ML/DS: Python, R, Matlab

Overview of Modern ML Applications: Recurrent Neural Network (RNN) and LSTM



Acknowledgment: Some graphics and slides were adapted from Stanford's CS231n by Fei-Fei Li et al.: http://cs231n.stanford.edu/

A Sequence of Identical Neural Network Modules

Force the neural nets (in green) to be the same to lower the complexity!

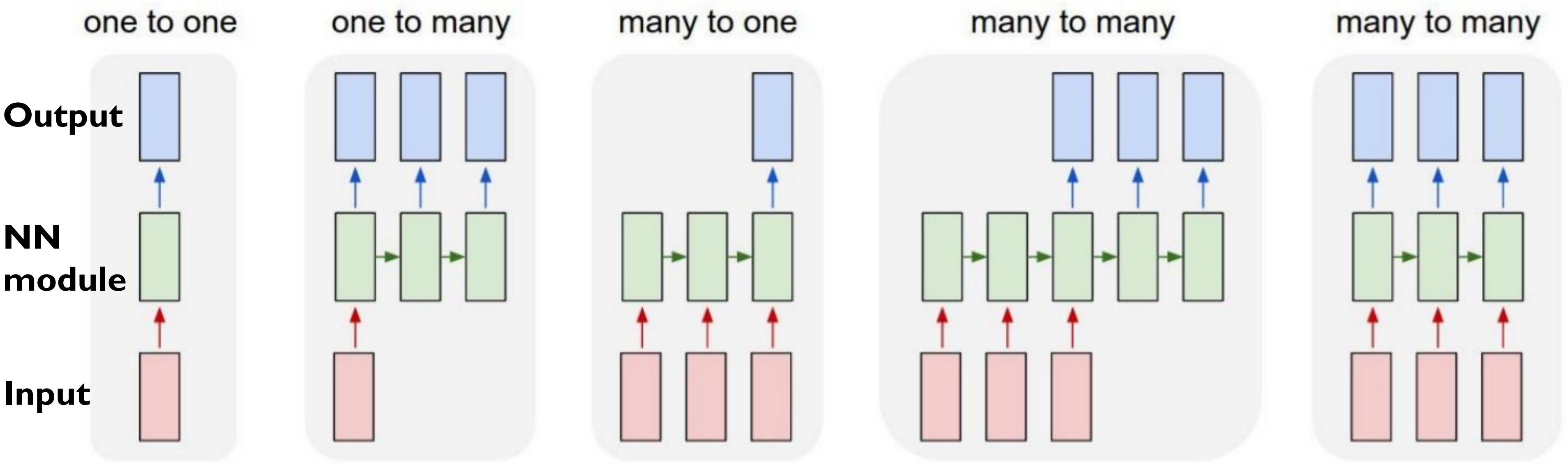


Image Captioning
image -> seq of words

Emotion Classification
seq of words -> emotion

Machine Translation
seq of words -> seq of words

Video classification
at frame level

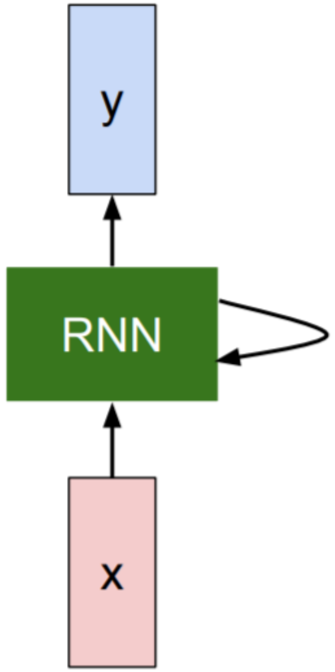
Recurrent Neural Network (RNN): Definition

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

$$h_t = f_W(h_{t-1}, x_t)$$

h_t : state, x_t : input
f_W : neural network

$$y_t = W_{hy} h_t$$



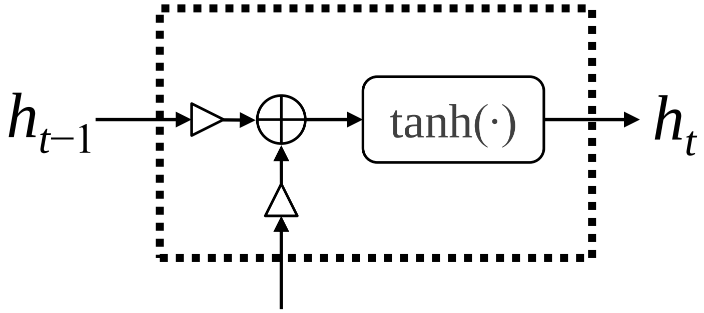
Recurrent Neural Network (RNN): Implementation

$$h_t = f_W(h_{t-1}, x_t)$$

$$\downarrow$$

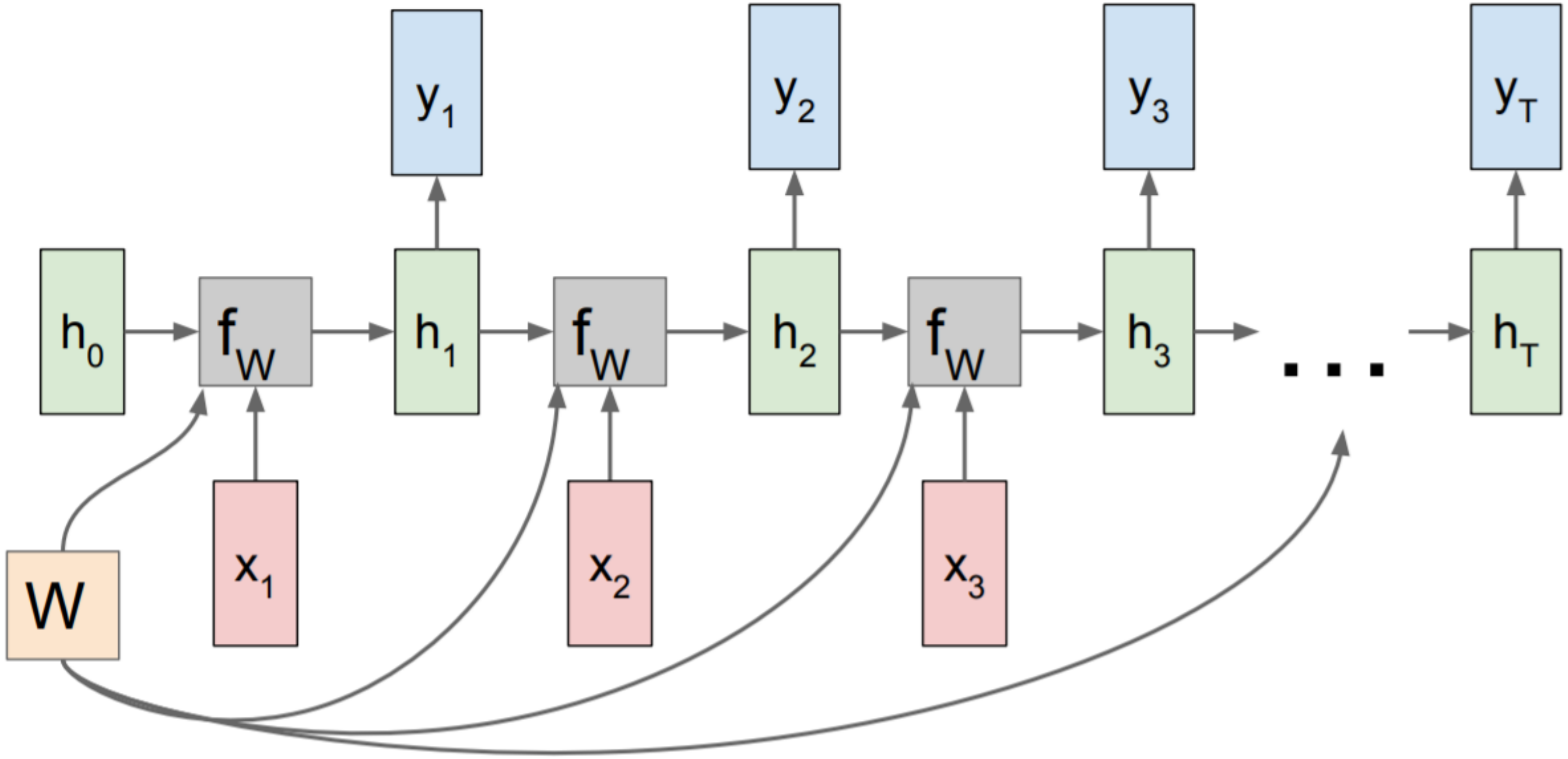
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

Diagram for f_W.
Can you label the details?



f_W is implemented via
- linear transforms W_{hh} and W_{xh} and
- elementwise nonlinear function $\tanh(\cdot)$

Recurrent Neural Network (RNN): Unrolled

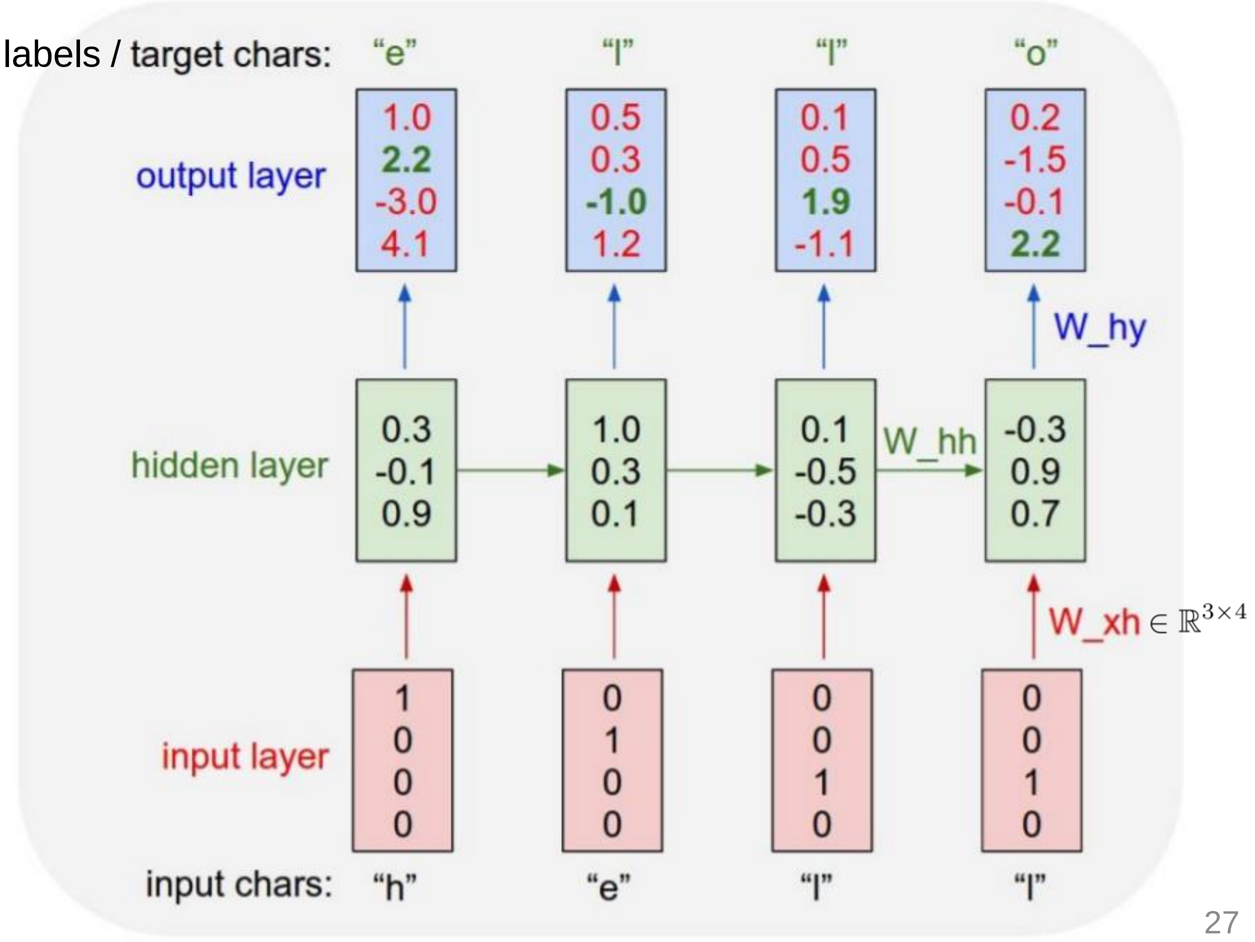


Example: Character-Level Language Model

Vocabulary:
[h, e, l, o]

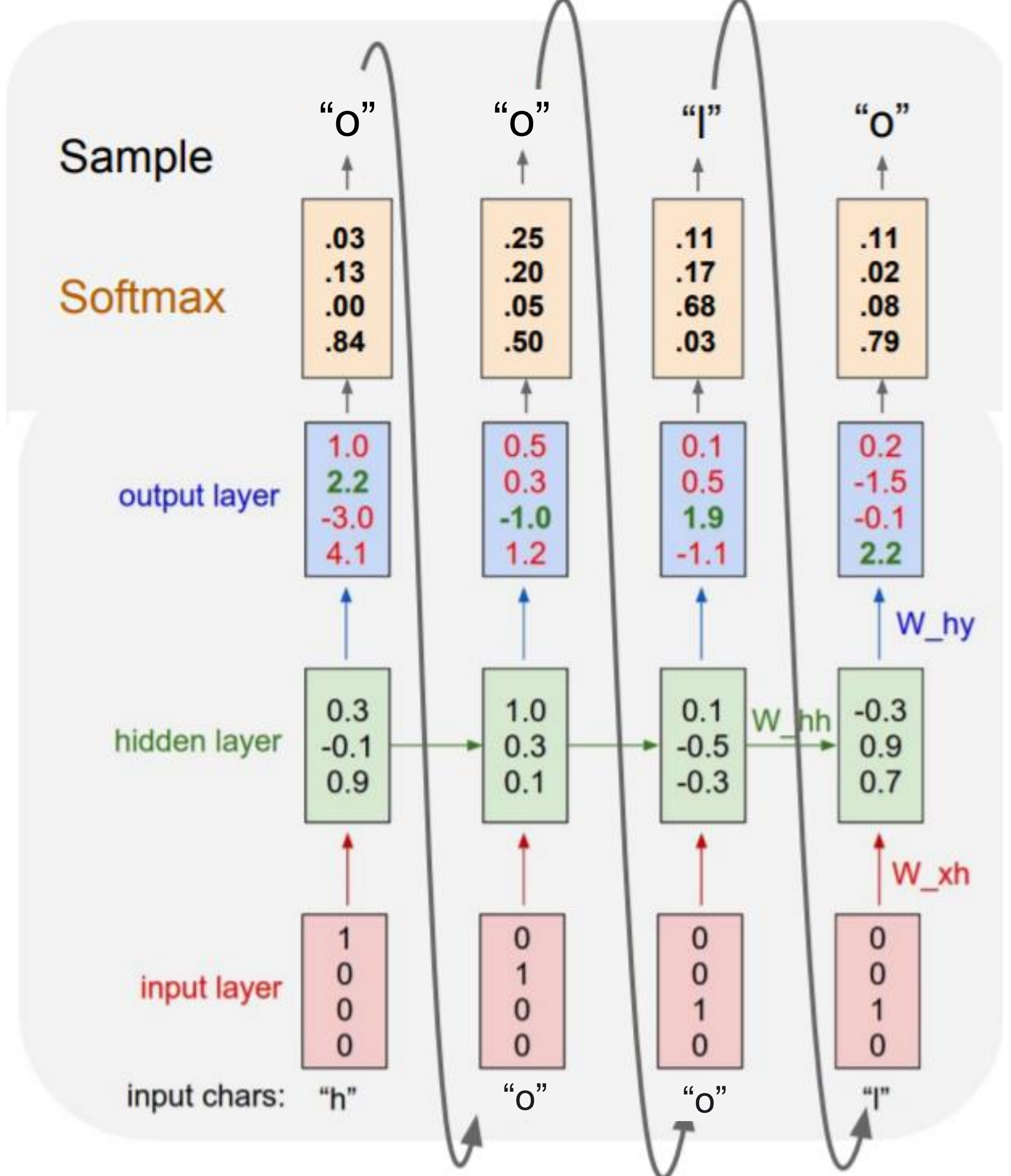
(Character) Embeddings:

Example training sequence: “hello”. Supervised learning pairs:



Vocabulary:
[h, e, l, o]

At test time, sample characters one at a time, feed back to model



Long Short-Term Memory (LSTM) Network

- RNN has the "vanishing gradient" problem!
- Resolved by long short-term memory (LSTM) units.

RNN

$$h_t = \tanh\left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}\right)$$

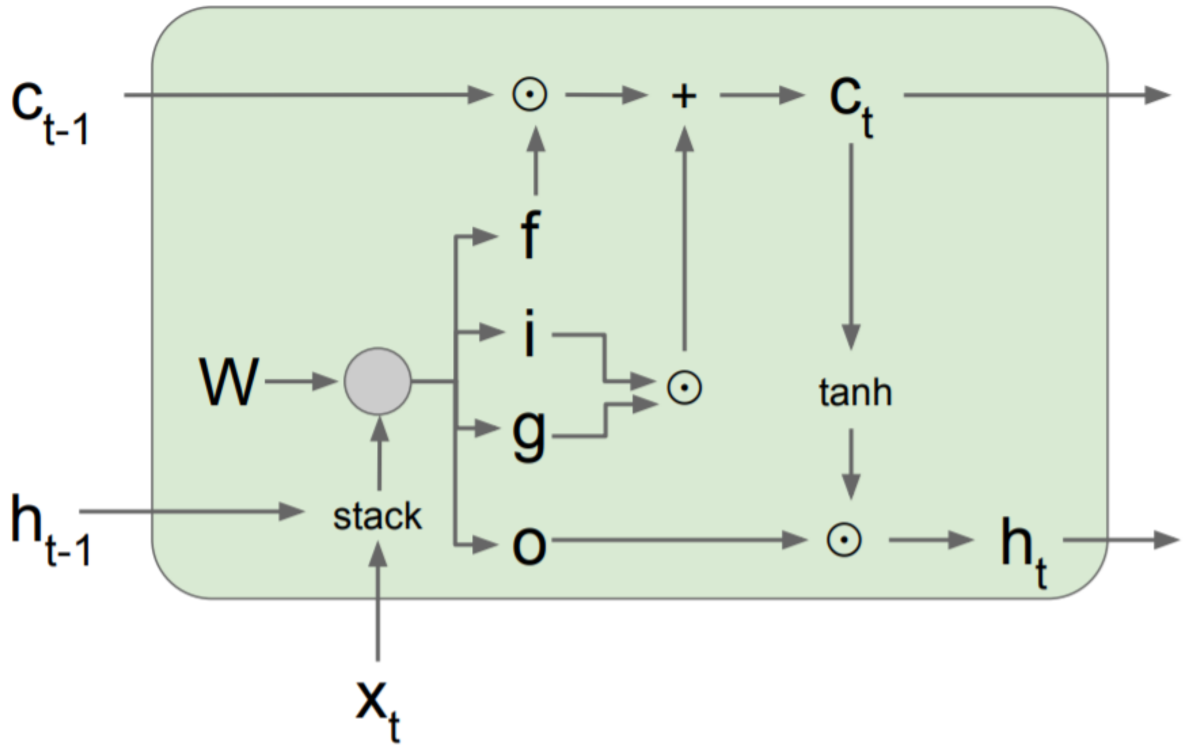
LSTM

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

An LSTM Unit [Hochreiter et al., 1997]

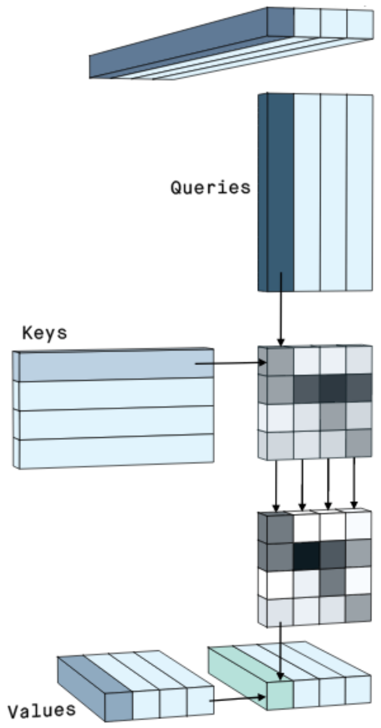


$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

Overview of Modern ML Applications: Transformer (Underlying technology of BERT & GPT)



Acknowledgment: Some graphics and slides were adapted from https://peltarion.com/blog/data-science/self-attention-video

How to make good sense of language?

- Reading comprehension: If you were Google, what result(s) should you return for "brazil traveler to usa need a visa"?
  1. A webpage on U.S. citizens traveling to Brazil
  2. A webpage of the U.S. embassy/consulate in Brazil

- Contextualization/attention is the key!
  - A nice walk by the river bank.
  - Walk to the bank and get cash.

- The transformer paper (Vaswani et al., 2017) showed that only attention is needed; convolution and recurrence are unnecessary.

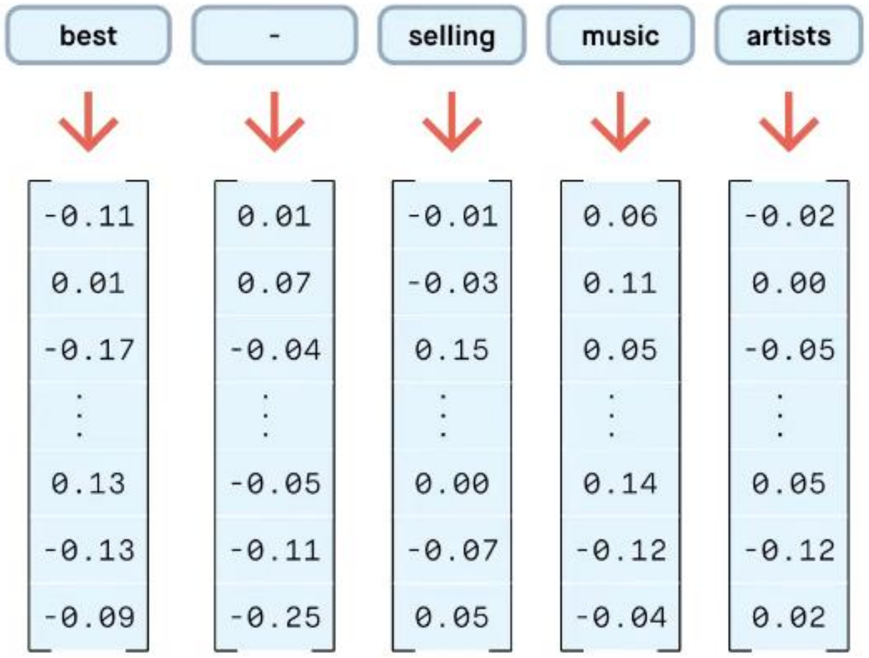
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6000–10, 2017. **[Transformer paper]**

Word Embeddings in Natural Language Processing (NLP)

WordPiece tokens:

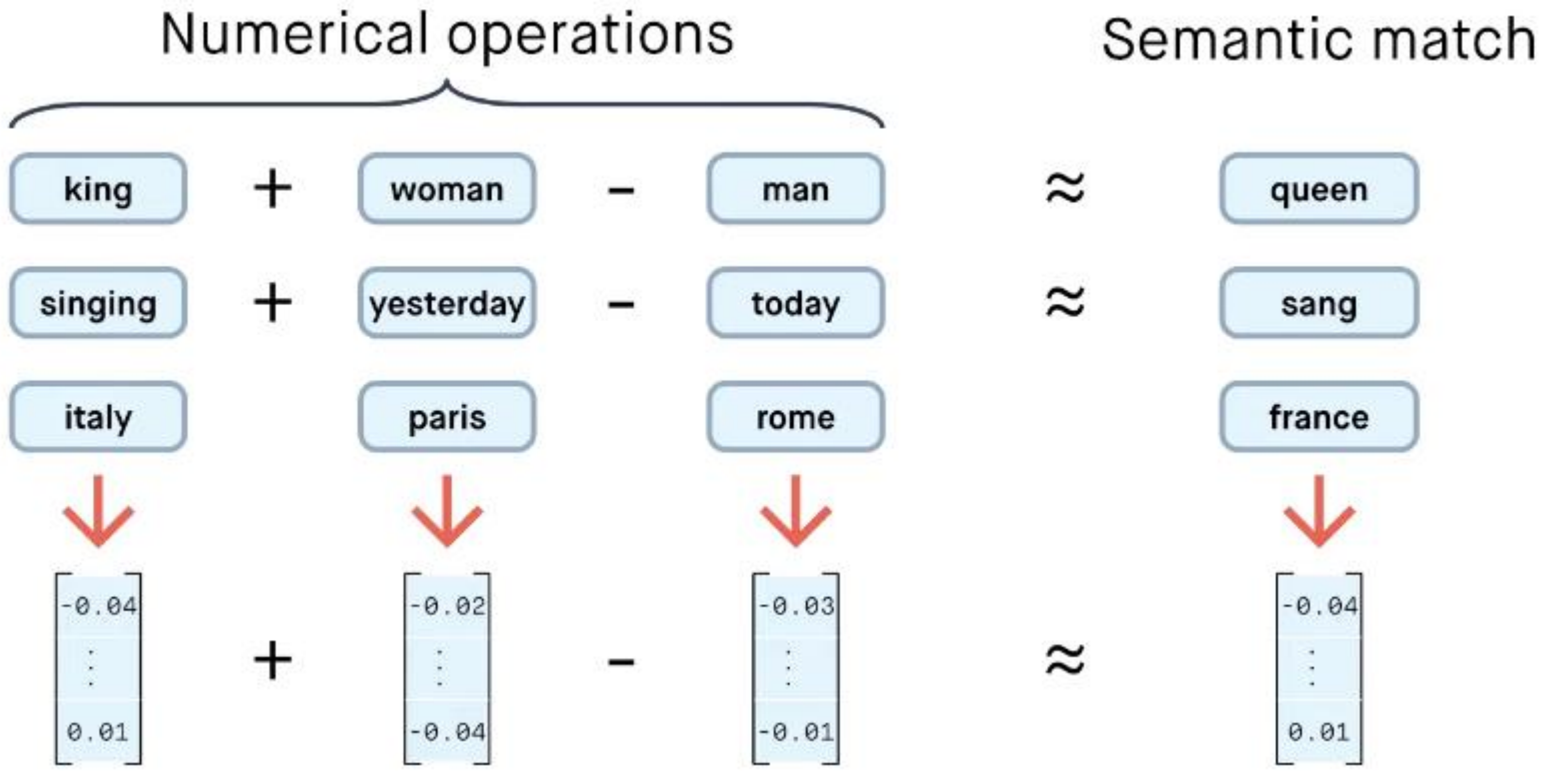
Embedding vectors:

Values are *pretrained*

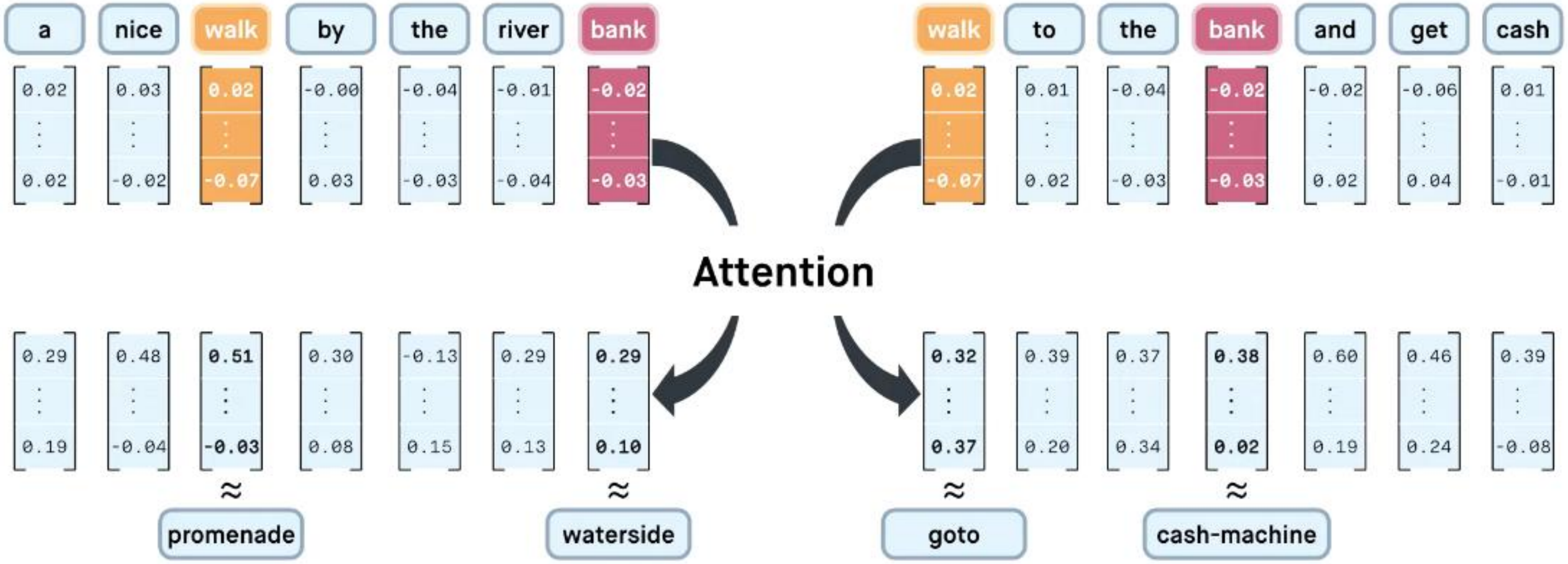


Embedding examples: Bag of Words (BoW), Word2Vec, …

Word Embeddings are Meaningful Under “+” and “–”



Contextualization by "Attention"

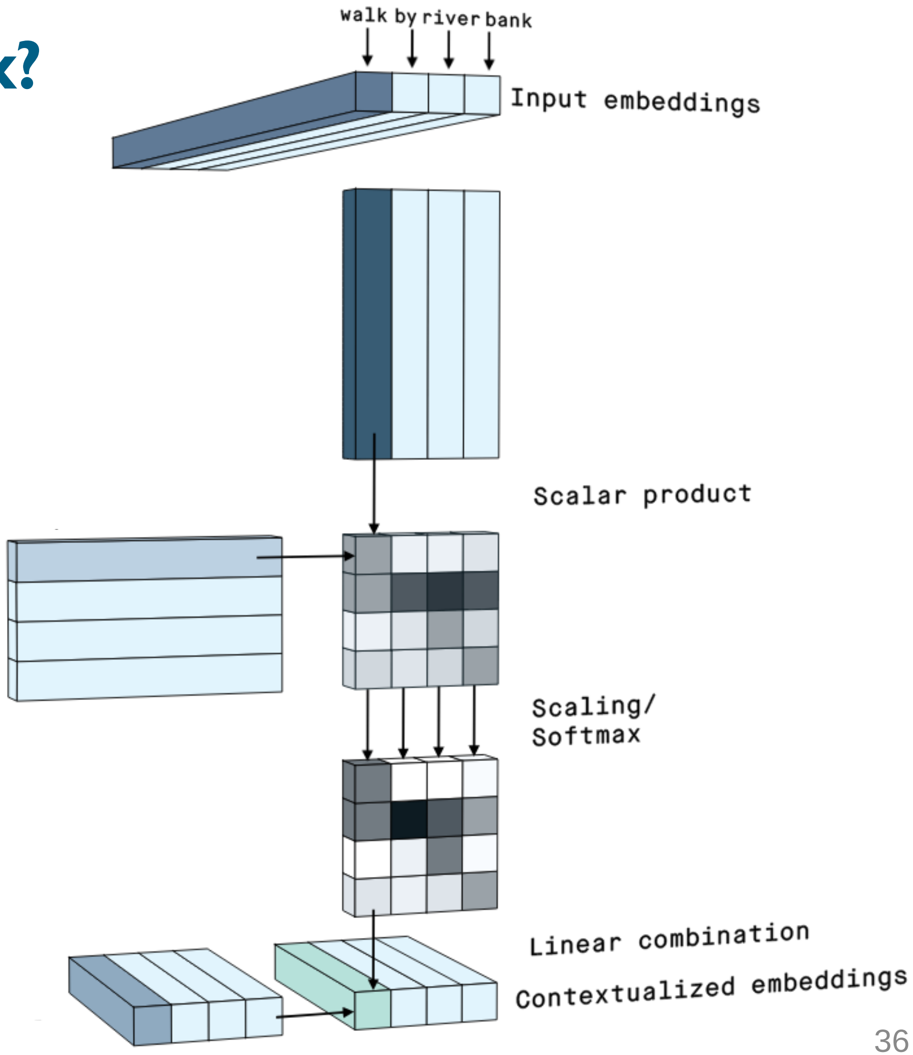


How does "attention" work?

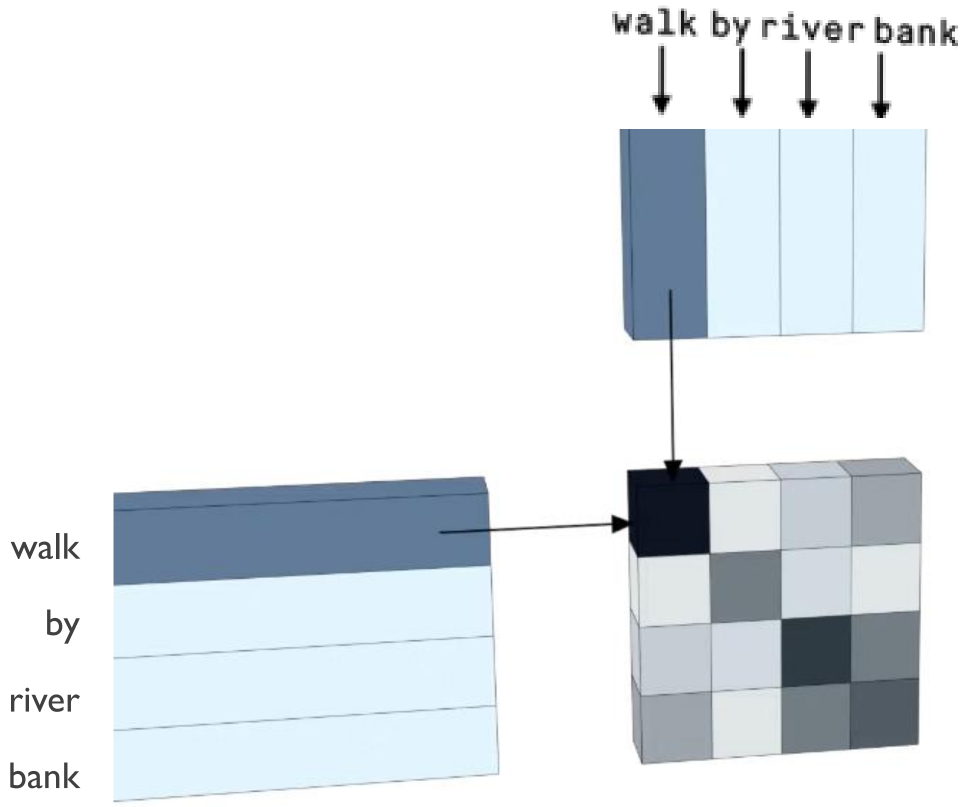
Similarity measure

Exponential saturation & normalization

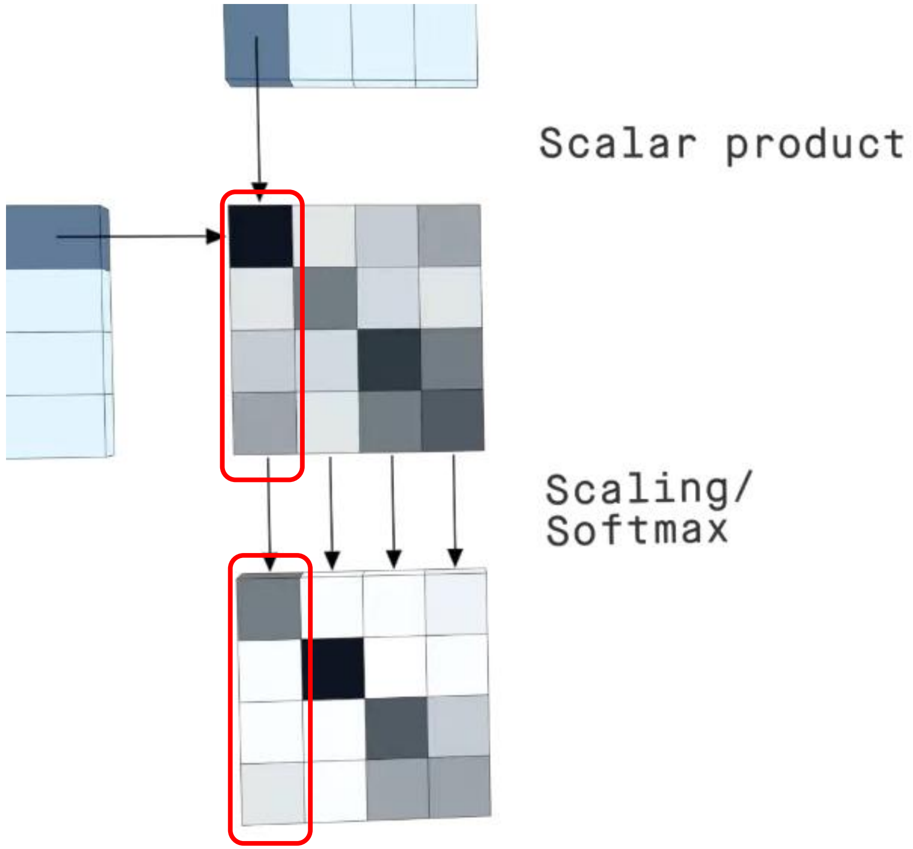
Linear combination

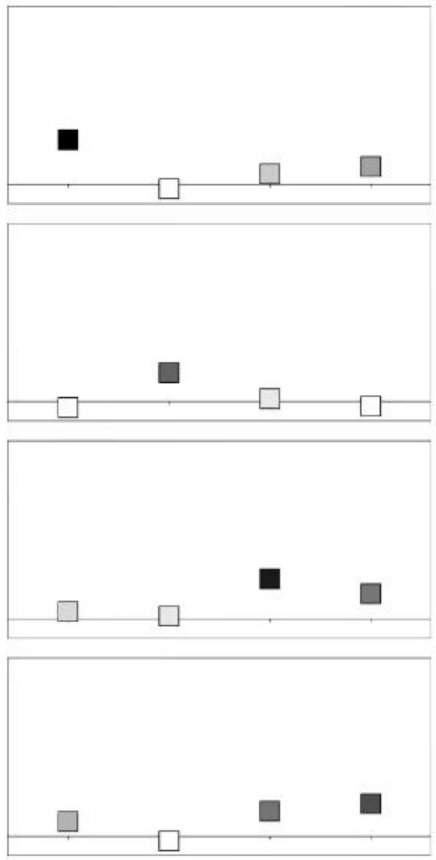


Similarity Measure via Inner Product

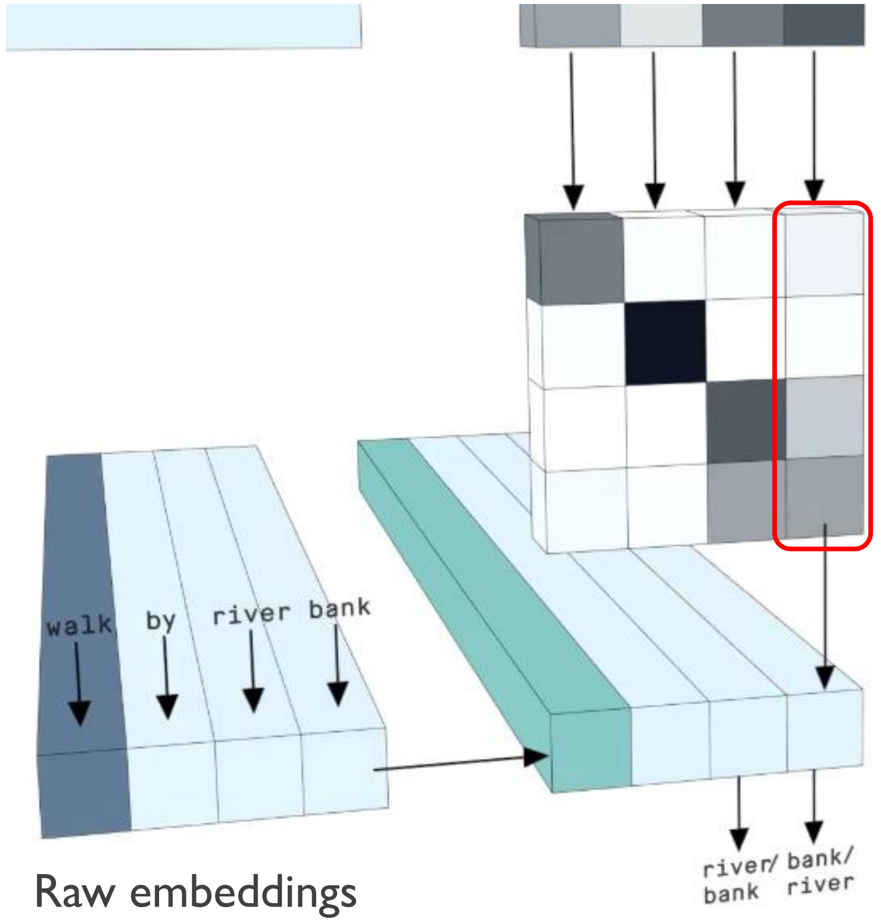


Exponential Saturation & Normalization via Softmax





Contextualization via Linear Combination



Weights for linear combination

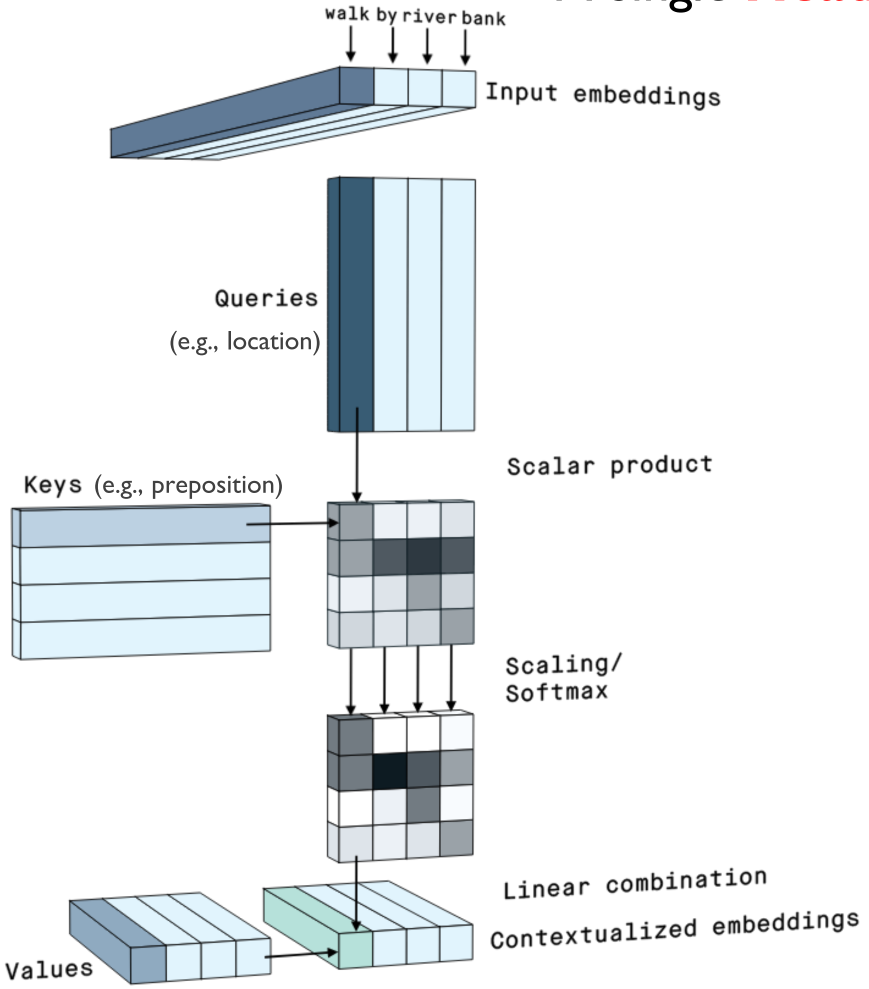
Contextualized embeddings

Raw embeddings

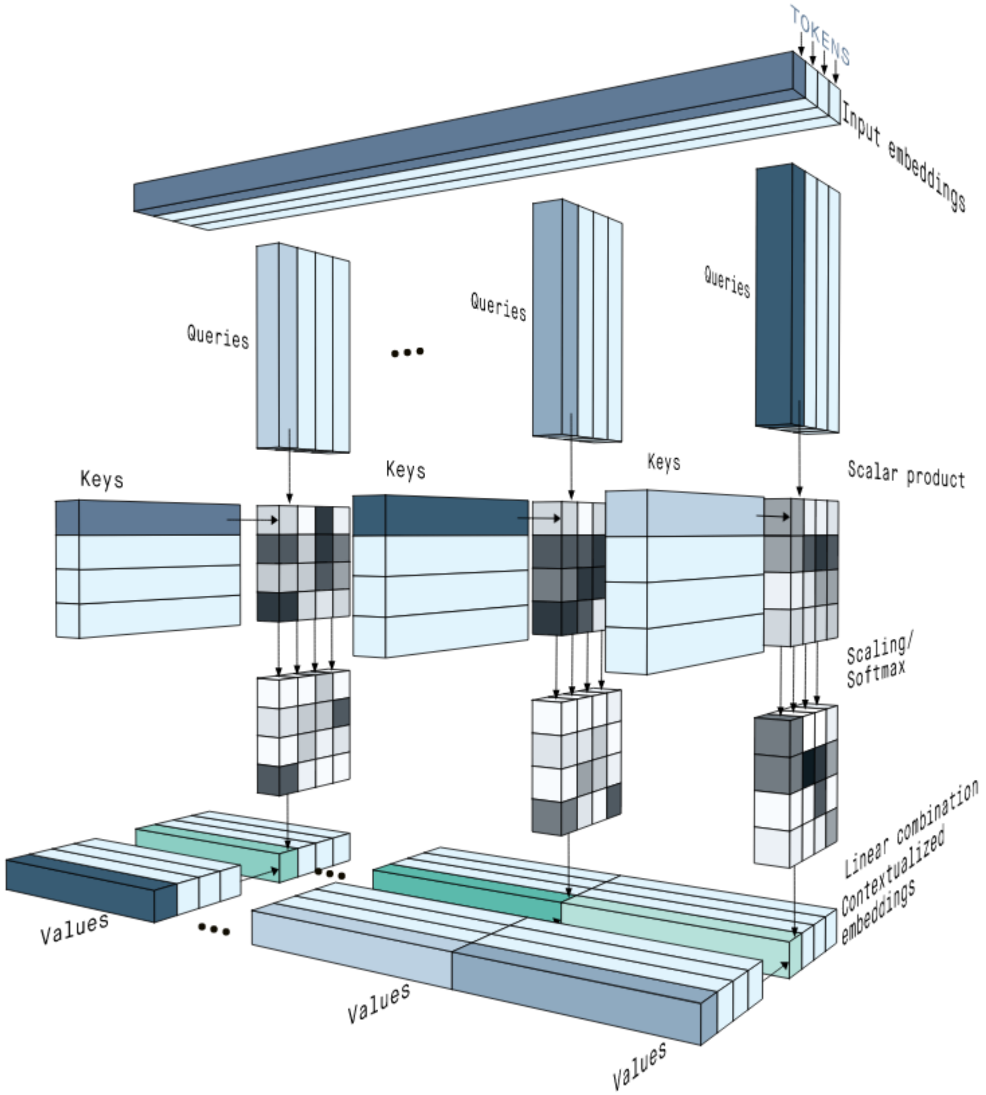
A Single **Head**

Key, Value, and Query

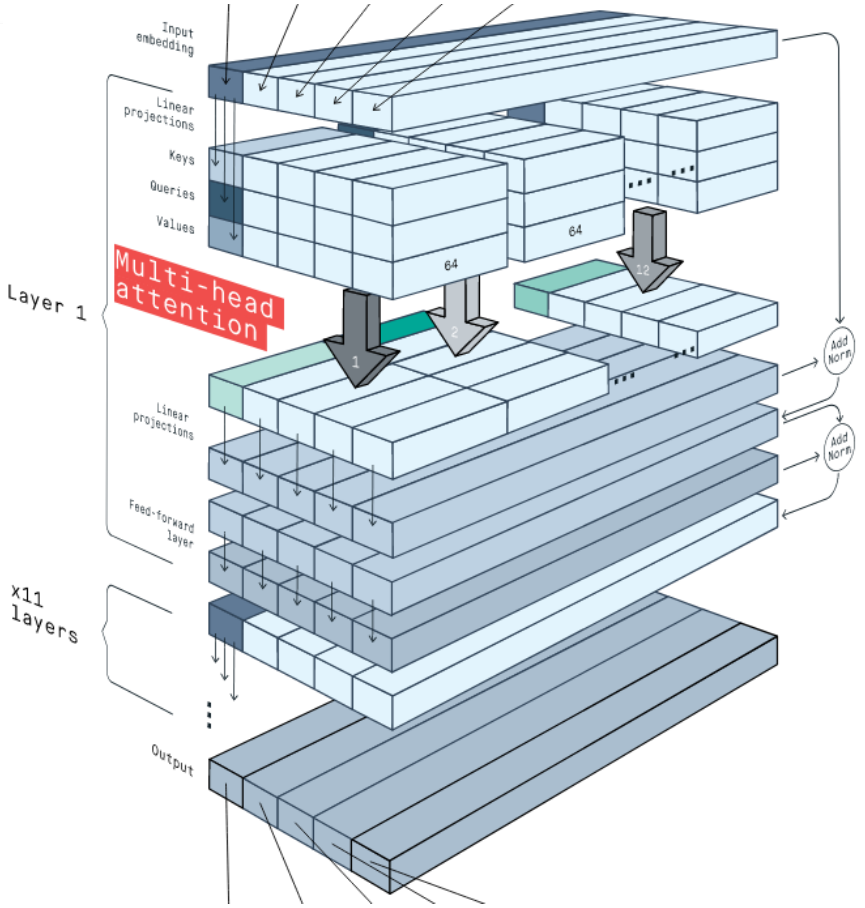
- "Key", "value", and "query" are three projections of an input embedding to three vector subspaces.
- Each subspace represents a unique semantic aspect.
- The projection operators / matrices provide trainable parameters for Transformer neural networks.



Multi-Head Attention



Bidirectional Encoder Representations from Transformers (BERT)



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6000–10, 2017. **[Transformer paper]**

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Association for Computational Linguistics, 2019.* **[BERT paper]**

BERT's strategies to train a language model

- Self-supervised task #1: ***Masked language model*** *(MLM)*
  - Mask 15% of all tokens in each sequence.
  - Predict masked tokens.
- Self-supervised task #2: ***Next sentence prediction*** *(NSP)*
  - Generate sentence pairs (X, Y). 50% chance Y is the actual next sentence that follows X, and 50% chance Y is a random sentence.
  - Predict whether Y is the actual next sentence of X.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *North American Chapter of the Association for Computational Linguistics, 2019.*

Generative pretrained transformers (GPT) by OpenAI

- Uses transformer architecture.
- Training strategy: Next token prediction (autoregressive modeling).
- GPT-3 also trained on computer source code.
- GPT-3 improved via *reinforcement learning from human feedback (RLHF)*.
- ChatGPT fine-tuned with conversational interaction with human users.
- Model scale in billion of parameters: GPT-1 (0.1), GPT-2 (1.5), GPT-3 (175), GPT-4 (~10x more); BERT (0.1–0.3).

- Alternative: Large language model Meta AI (**LLaMA**) 2, released on July 2023. Sizes: 7, 13, and 70 bn. LLaMA Chat available.

NC STATE UNIVERSITY

Neural Network Training: Backpropagation

Acknowledgment: Some graphics and slides were adapted from Profs. Jain (MSU), Min Wu (UMD), Fei-Fei Li (Stanford)
Some figures are from Duda-Hart-Stork textbook, Fei-Fei Li's slides

3-Layer Neural Network Structure

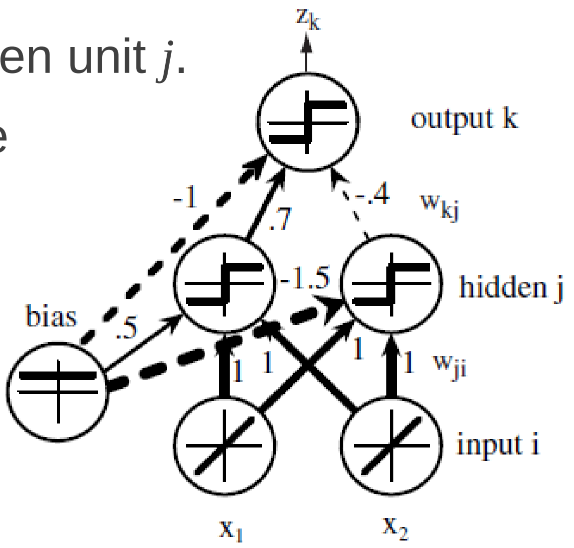
- A single “bias unit” is connected to each unit in addition to the input units
- Net activation: $$\text{net}_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0} = \sum_{i=0}^{d} x_i w_{ji} \equiv \mathbf{w}_j^T \mathbf{x},$$

  where the subscript i indexes units in the input layer, j indexes units in the hidden layer;

  w_{ji} denotes the input-to-hidden layer weights at hidden unit j.

  - In neurobiology, such weights or connections are called “synapses”
- Each hidden unit emits an output that is a nonlinear function of its activation

$$y_j = \sigma(net_j)$$



Training Neural Networks / Estimating Weights

- Notations: t_k ~ the kth target (or desired) output,
  z_k ~ the kth estimated/computed output with $k = 1, \ldots, c$.
  w_{ij} ~ weight of the network
- Squared cost func:

$$J(\mathbf{w}) = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2 = \frac{1}{2}\|\mathbf{t} - \mathbf{z}\|^2$$

- Learning based on gradient descent by iteratively updating the weights:

$$\mathbf{w}(m+1) = \mathbf{w}(m) + \Delta\mathbf{w}(m), \quad m = 1, 2, \ldots$$

  - The weights are initialized with random values, and updated in a direction to reduce the error.

$$\Delta\mathbf{w} = -\eta \cdot \nabla_{\mathbf{w}} J$$

  - **Learning rate**, η, controls the step size of the update in weights.

Walking man image is CC0 1.0 public domain

Figure source: Stanford CS231n by Fei-Fei Li

Gradient Descent

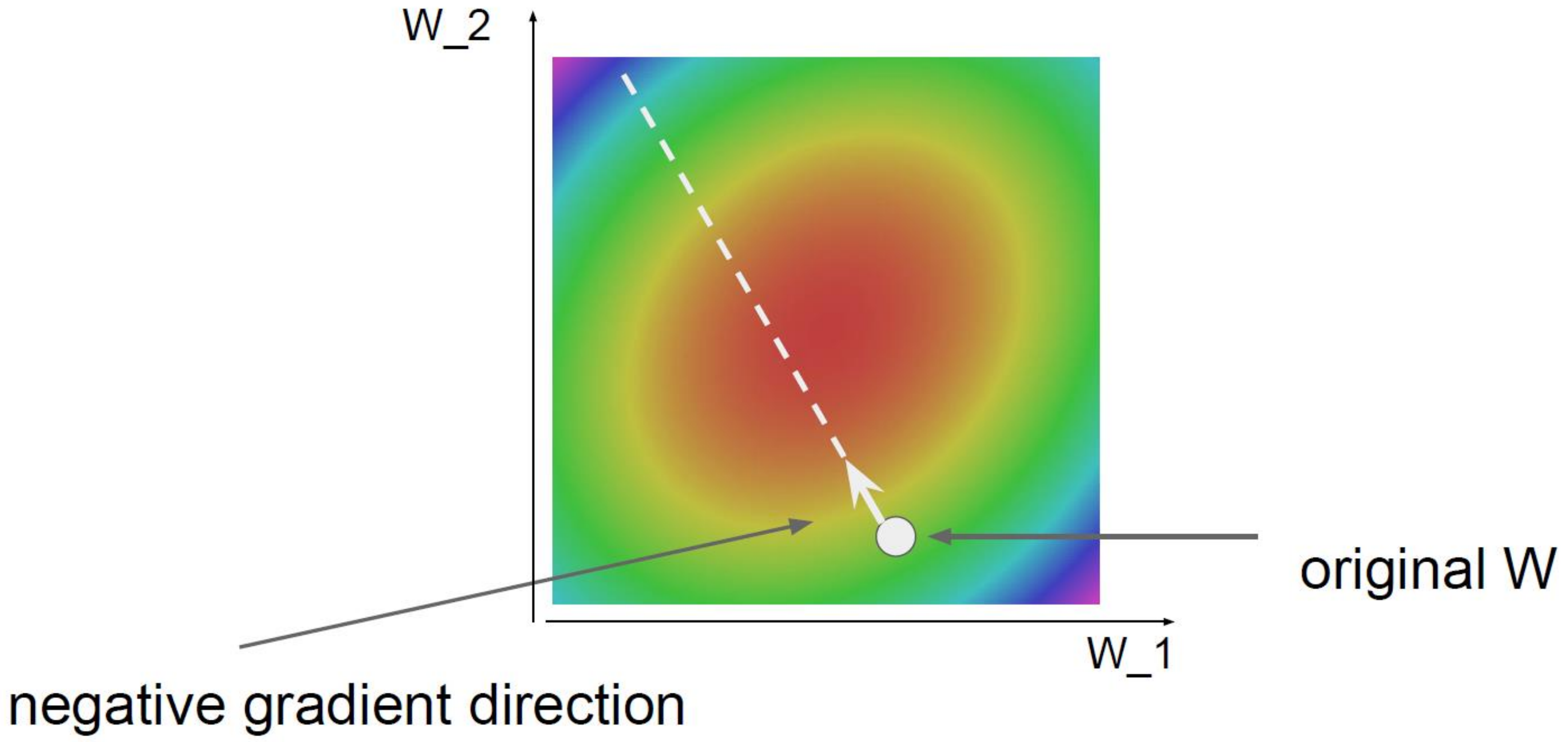


Figure source: Stanford CS231n by Fei-Fei Li

Efficient Gradient Calculation: Backpropagation

- Computes $\partial J / \partial w_{ji}$ for a single input-output pair.
- Exploit the chain rule for differentiation, e.g.,

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j} \cdot \frac{\partial y_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}}$$

- Computed by ***forward*** and ***backward*** sweeps over the network, keeping track only of quantities local to each unit.
- Iterate backward one unit at a time from last layer. Backpropagation avoids redundant calculations.

Backpropagation (BP): An Example

Backpropagation: a simple example

$$f(x,y,z) = (x+y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

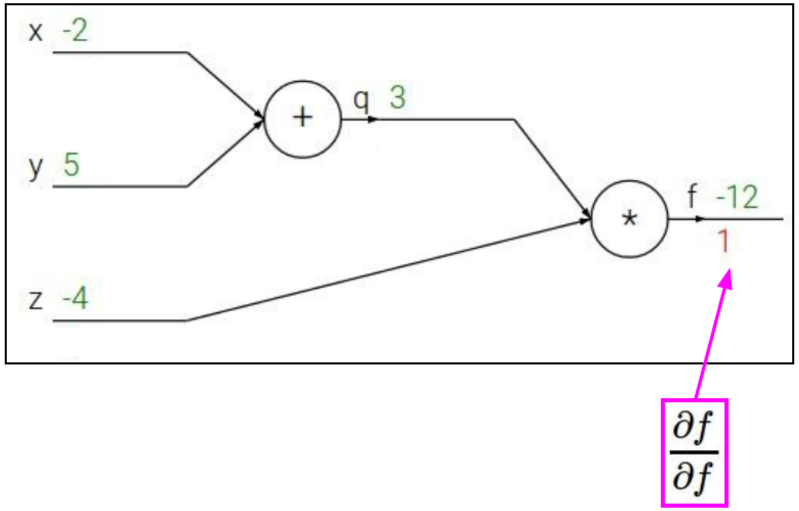
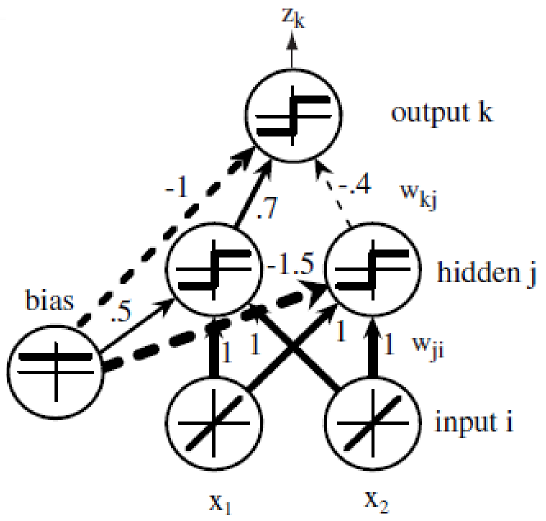


Figure source: Stanford CS231n by Fei-Fei Li

Network Learning by BP (cont'd)



- Error on hidden-to-output weight:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k} \cdot \frac{\partial net_k}{\partial w_{kj}} = -\delta_k \frac{\partial net_k}{\partial w_{kj}}$$

- δ_k, the sensitivity of unit k :
  describes how the overall error changes with the activation of the unit's net activation

$$\delta_k \equiv -\frac{\partial J}{\partial net_k}$$

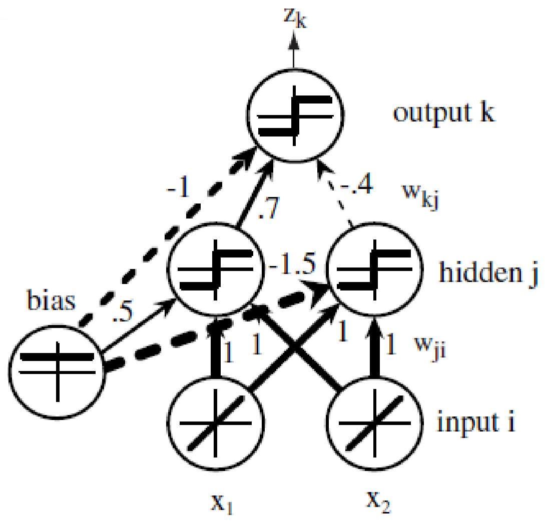
$$\delta_k = -\frac{\partial J}{\partial net_k} = -\frac{\partial J}{\partial z_k} \cdot \frac{\partial z_k}{\partial net_k} = (t_k - z_k) f'(net_k)$$

- Since $net_k = \mathbf{w}_k^T \mathbf{y}$, we have $\frac{\partial net_k}{\partial w_{kj}} = y_j$

- **Summary 1:** weight update (or learning rule) for the hidden-to-output weight is:

$$\Delta w_{kj} = \eta\, (t_k - z_k)\, f'(net_k)\, y_j = \eta\, \delta_k y_j$$

Network Learning by BP (cont'd)



- Error on input-to-hidden weight:
  - chain rule: $$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j} \cdot \frac{\partial y_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}}$$
  - $$\frac{\partial J}{\partial y_j} = \frac{\partial}{\partial y_j}\left[\frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2\right] = -\sum_{k=1}^{c}(t_k - z_k)\frac{\partial z_k}{\partial y_j}$$
    $$= -\sum_{k=1}^{c}(t_k - z_k)\frac{\partial z_k}{\partial net_k} \cdot \frac{\partial net_k}{\partial y_j} = -\sum_{k=1}^{c}(t_k - z_k) f'(net_k) w_{kj}$$
- Sensitivity of a hidden unit:
  (Similarly defined as earlier)
  $$\delta_j \equiv \frac{\partial J}{\partial net_j} = f'(net_j)\sum_{k=1}^{c} w_{kj}\delta_k$$
- **Summary 2:** Learning rule for the input-to-hidden weight is:
  $$\Delta w_{ji} = \eta \underbrace{\left[f'(net_j)\Sigma w_{kj}\delta_k\right]}_{\delta_j} x_i = \eta \delta_j x_i$$

Sensitivity at Hidden Node

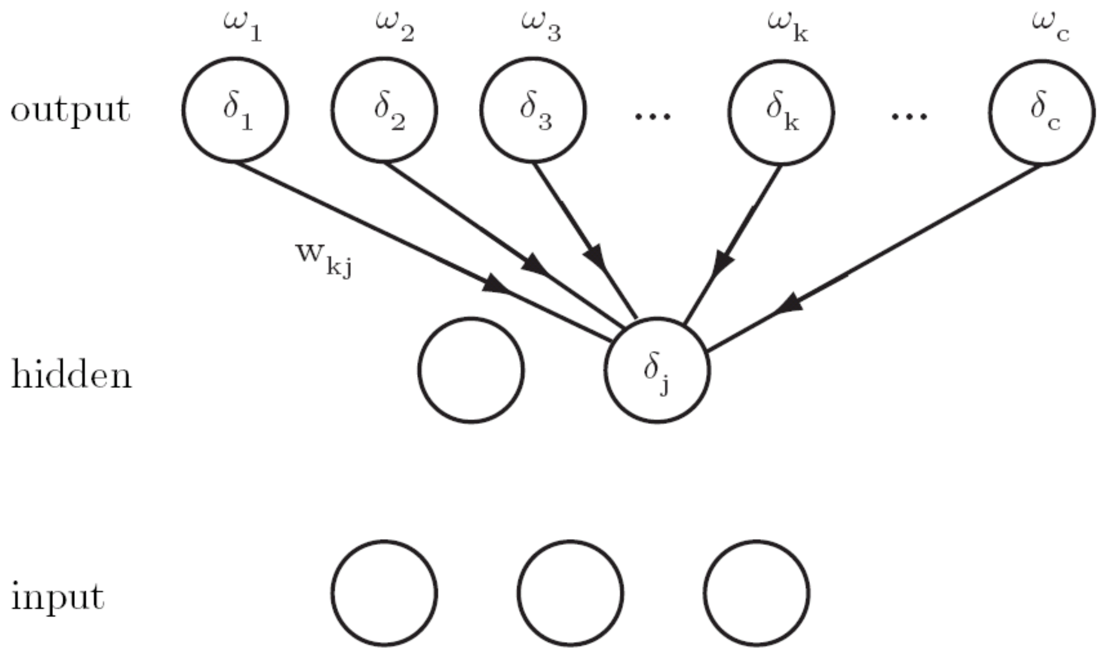


Figure 6.5: The sensitivity at a hidden unit is proportional to the weighted sum of the sensitivities at the output units: $\delta_j = f'(net_j) \sum_{k=1}^{c} w_{kj}\delta_k$. The output unit sensitivities are thus propagated "back" to the hidden units.

BP Algorithm: Training Protocols

1. **Gradient descent**: Use **all** training examples to update weights.
2. **Stochastic gradient descent (SGD)**: Use **one** random training example to iteratively update weights.
3. **Mini-batch gradient descent**: Separate training examples into disjoint ***mini-batches***. Use mini-batches to update weights. One "epoch" = cycling through all mini-batches.

- Note: Making several complete passes of the training data helps the network to learn.

BP Algorithm: Stopping Criteria

- Basic idea: Terminating training when the value or the change of value of the following quantities is below some preset threshold:
  - Loss function $J(\mathbf{w})$,
  - Gradient norm $||\nabla_{\mathbf{w}} J||$.
- Modern deep learning stopping criteria:
  - Lowest validation error,
  - For large models, training as many iterations as resources is available.

Learning Curves

- Before training starts, the training error is high; as the learning proceeds, training error becomes smaller
- Error per epoch depends on the amount of training data and expressive power (e.g., # of weights) of the network
- Average error on an independent test set is always higher than on the training set, and it can decrease or increase

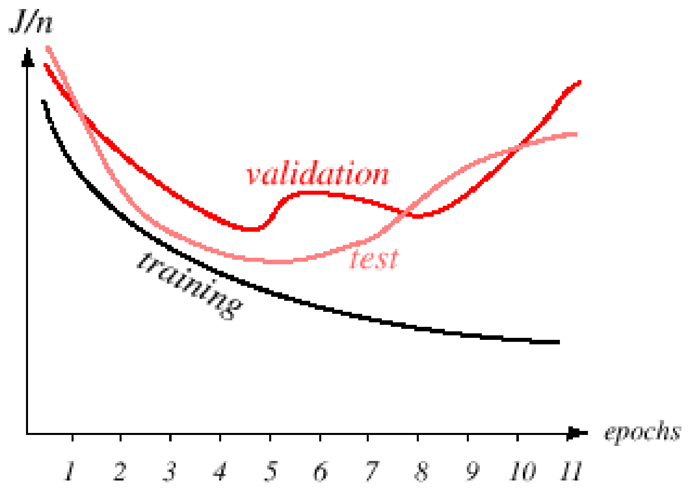


FIGURE 6.6. A learning curve shows the criterion function as a function of the amount of training, typically indicated by the number of epochs or presentations of the full training set. We plot the average error per pattern, that is, $1/n \sum_{p=1}^{n} J_p$. The validation error and the test or generalization error per pattern are virtually always higher than the training error. In some protocols, training is stopped at the first minimum of the validation set. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- A validation set could be used to decide when to stop training. It can avoid overfitting and can ensure the generalization capability of the learned network.

"Stop training when the error on the validation set is minimum"

Practical Considerations: Learning Rate

- Learning Rate
  - Small learning rate: slow convergence
  - Large learning rate: high oscillation and slow convergence

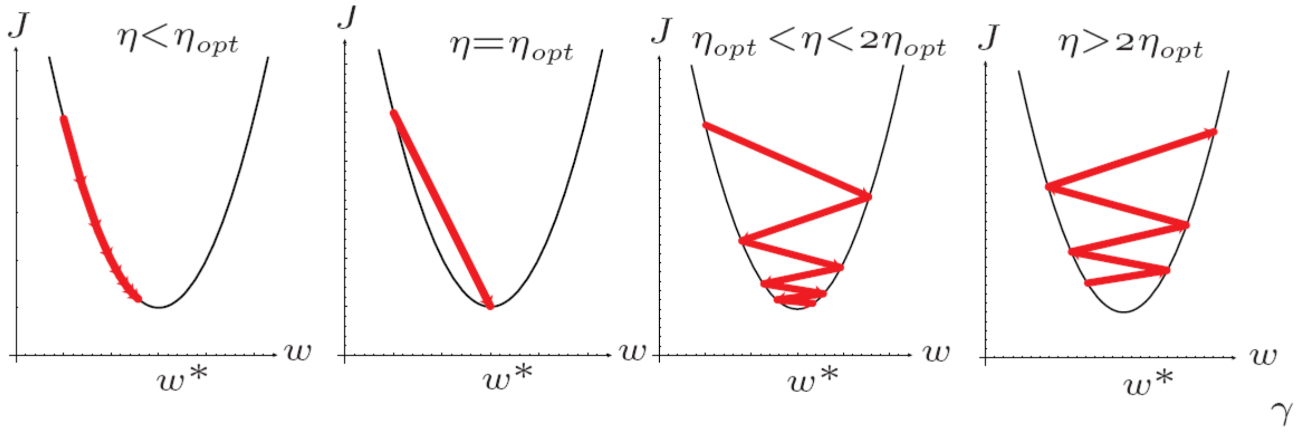


Figure 6.18: Gradient descent in a one-dimensional quadratic criterion with different learning rates. If $\eta < \eta_{opt}$, convergence is assured, but training can be needlessly slow. If $\eta = \eta_{opt}$, a single learning step suffices to find the error minimum. If $\eta_{opt} < \eta < 2\eta_{opt}$, the system will oscillate but nevertheless converge, but training is needlessly slow. If $\eta > 2\eta_{opt}$, the system diverges.