

# Deep Learning

## *Shallow vs Deep*

Understanding Machine Learning

# Recap: Goal of ML

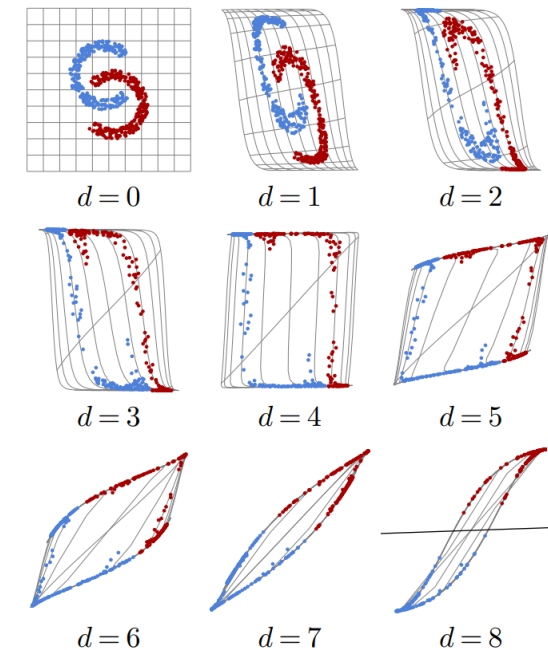
generalization from optimization on training data set  
(approximation of true data generating probability distribution by empirical risk minimization)

- fitting: complex function approximation
- for generalization: learning of good abstraction/representation of data/concepts

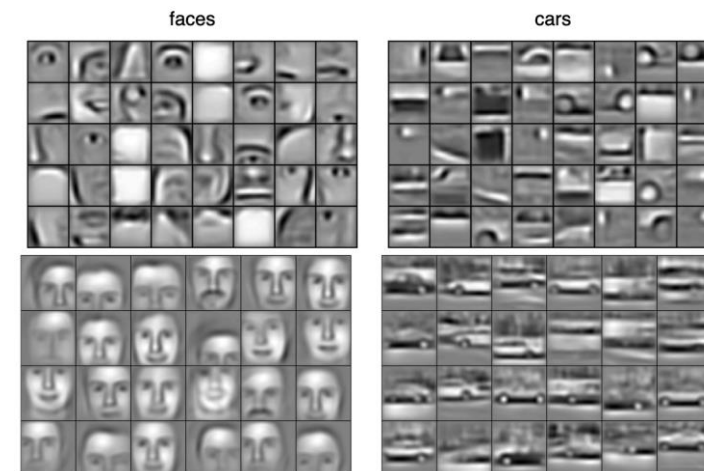
→ deep learning methods (MLP, CNN, ...) optimal candidates

e.g., CNNs can learn hierarchical representation by means of many convolutional and pooling layers

the deeper the better (accuracy, hierarchical representation)



[source](#)



[source](#)

# Deep Learning

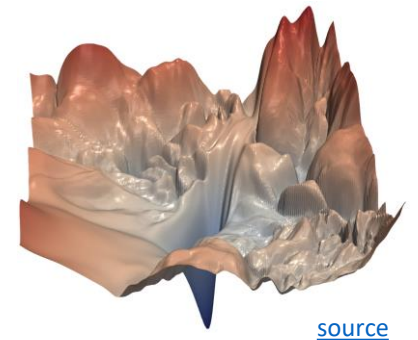
# But ... How to Train Deep Neural Networks?

optimization and regularization difficult

- non-convex optimization problem (e.g., local vs global minima, saddle points), easily overfitting
- many hyperparameters to tune

many methods to get it working in practice (despite partly patchy theoretical understanding)

typical loss surface:



optimization

- activation and loss functions
- weight initialization
- stochastic gradient descent
- adaptive learning rate
- batch normalization

explicit regularization

- weight decay
- dropout
- data augmentation
- weight sharing

implicit regularization

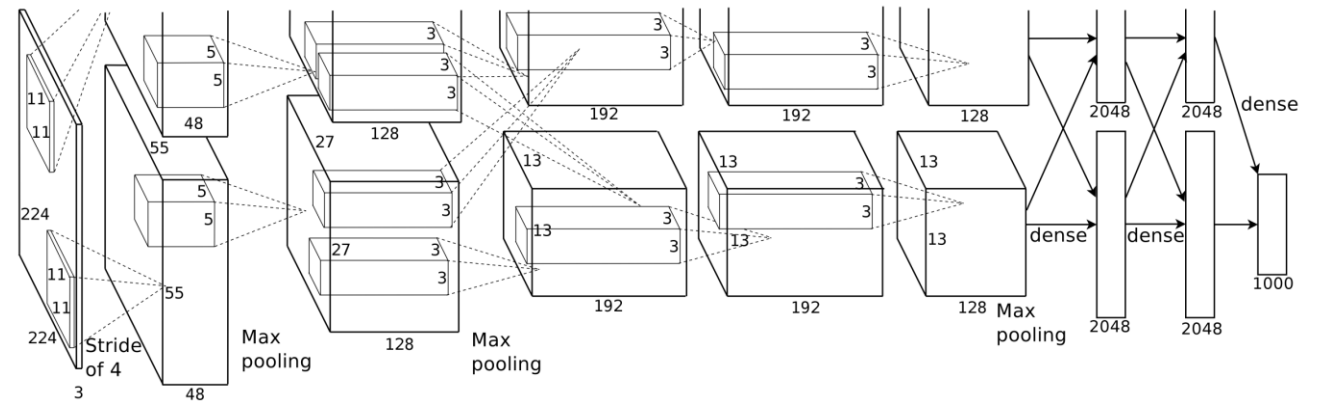
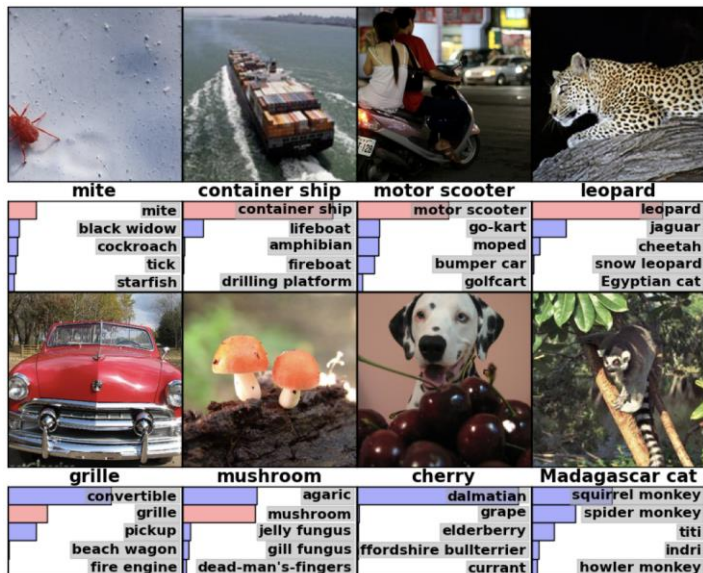
- early stopping
- batch normalization
- stochastic gradient descent

# History: Rise of Deep Learning

a little bit oversimplified:

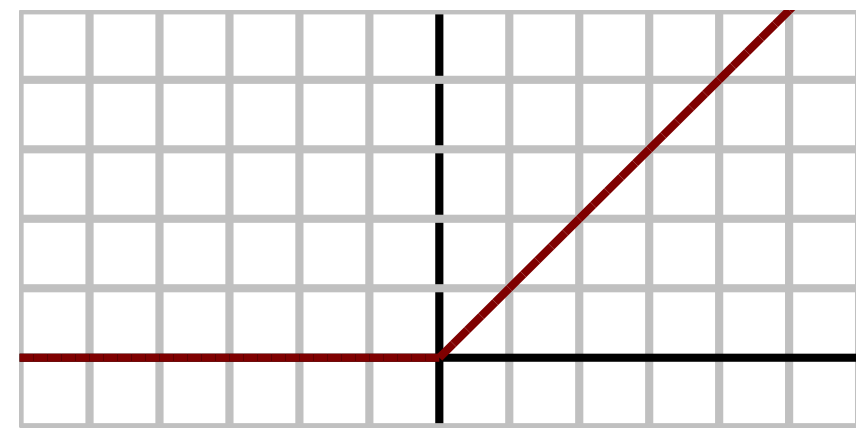
deep learning = lots of training data + parallel computation + smart algorithms

AlexNet: ImageNet (with data augmentation) + GPUs + ReLU, dropout, SGD



[source](#)

# Rectified Linear Unit (ReLU)



reminder: activation function non-linear transformation of summed weighted input of a node (linear), output to be used as input for nodes of subsequent layer

needs to be differentiable for back-propagation (ReLU at 0 no issue, just set to 0 or 1)

neural network model with [ReLU activation](#) can be interpreted as exponential number of linear models that share parameters

main advantages (leading to enablement of deeper networks by better optimization):

- unlike sigmoid or tanh (predominantly used before) activation, no issue with vanishing gradients from saturation effects
- very efficient computation: constant gradients of 0 and 1 below and above input of zero
- sparse activation: many hidden nodes deactivated (output 0) → information disentangling

# Weight Initialization

starting values for weights crucial for convergence of deep learning trainings

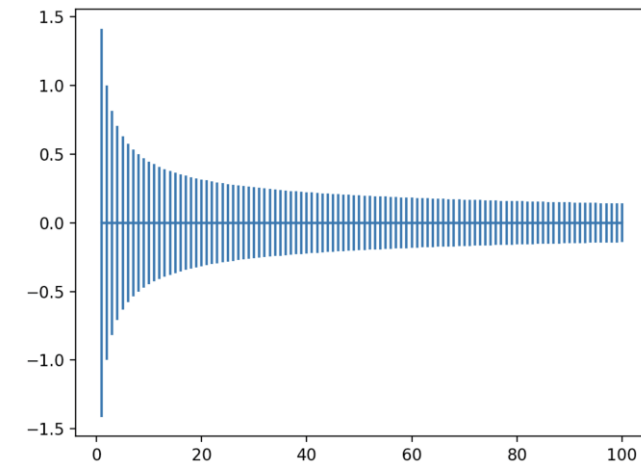
most important: need to break symmetry between different nodes in a hidden layer (same initial weights lead to identical weight updates) ← early issue in back-propagation

→ small random numbers (from Gaussian or uniform distribution) work (only bias weights set to zero by default)

but specific heuristics using information on activation function and number of inputs to a node can improve optimization (remain same expected variance between layers)

for ReLU, [He initialization](#) works well: randomly draw from zero-mean Gaussian with standard deviation  $\sqrt{2/n}$

number of inputs (hidden nodes in previous layer)



Plot of Range of He Weight Initialization With Inputs From One to One Hundred

[source](#)

# (Stochastic) Gradient Descent

using gradient of cost (objective) function with respect to weights:  $\nabla_{\hat{\mathbf{w}}} J(\hat{\mathbf{w}})$

updates  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} - \eta \nabla_{\hat{\mathbf{w}}} J(\hat{\mathbf{w}})$  can be done with whole training data set ( $n$  observations) or small random sample:

- $J(\hat{\mathbf{w}}) = \frac{1}{n} \sum_{i=1}^n J_i(\hat{\mathbf{w}})$       batch (or deterministic) gradient descent
- $J(\hat{\mathbf{w}}) = J_i(\hat{\mathbf{w}})$       stochastic gradient descent (single example)
- $J(\hat{\mathbf{w}}) = \frac{1}{m} \sum_{i=1}^m J_i(\hat{\mathbf{w}})$       mini-batch stochastic gradient descent (size  $m$ )

implicit regularization: (mini-batch) SGD follows gradient of true generalization error, if no examples are repeated (but usually many epochs in training)



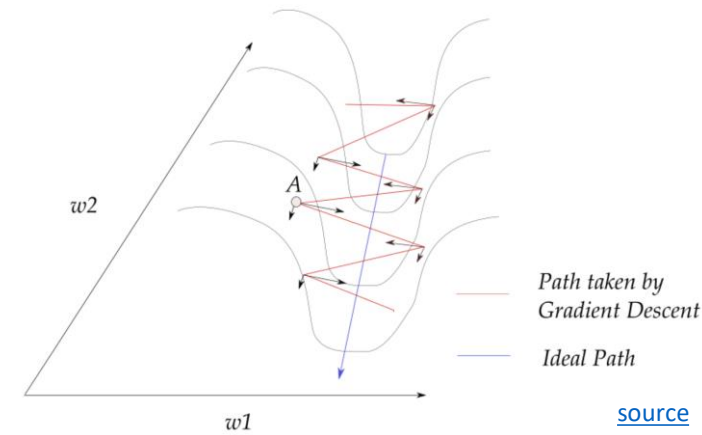
# Mini-Batch Sizes

trade-off:

- larger batches give more accurate gradient estimates → allowing for higher learning rate
- smaller batches have (implicit) regularization effect and better convergence

in practice, also need to consider memory limitations and run times

# Adaptive Learning Rate



strategies for gradient descent learning rate: constant, decaying, with momentum (escape from local minima and saddle points)

better convergence by adapting learning rate for each weight: lower/higher learning rates for weights with large/small updates (avoid direction of oscillations)

popular methods ( $\hat{w}$  here denotes individual weight  $\rightarrow g_{\hat{w}}$  component of gradient):

- Adagrad:  $\hat{w} \leftarrow \hat{w} - \frac{\eta}{\sqrt{\sum_{\tau=1}^t g_{\hat{w},\tau}^2}} g_{\hat{w}}$  with  $t, \tau$  denoting current and past iterations

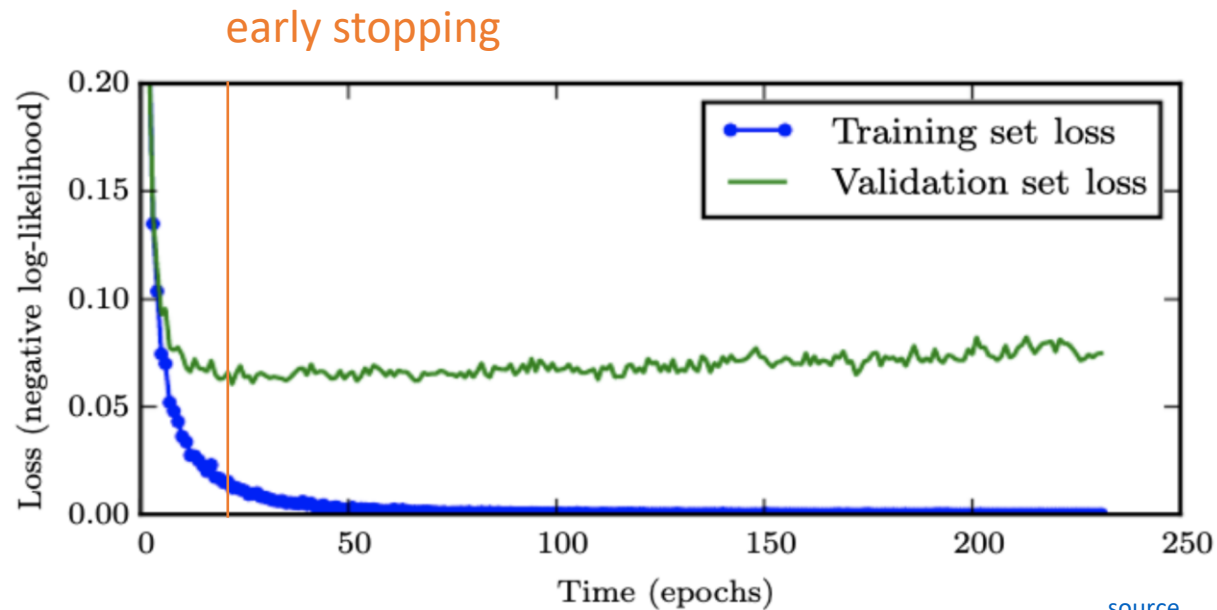
(issue: sum in denominator grows with more iterations  $\rightarrow$  danger of sticking)

- RMSProp:  $\hat{w} \leftarrow \hat{w} - \frac{\eta}{\sqrt{v(\hat{w})}} g_{\hat{w}}$  with  $v(\hat{w}) \leftarrow \gamma v(\hat{w}) + (1 - \gamma) g_{\hat{w}}^2$

- Adam (Adaptive Moment Optimization): combines RMSProp with momentum

# Early Stopping

loss independently measured on validation set  
halting training when overfitting begins to occur



[source](#)

(suppresses double descent)

# Skip/Residual Connections

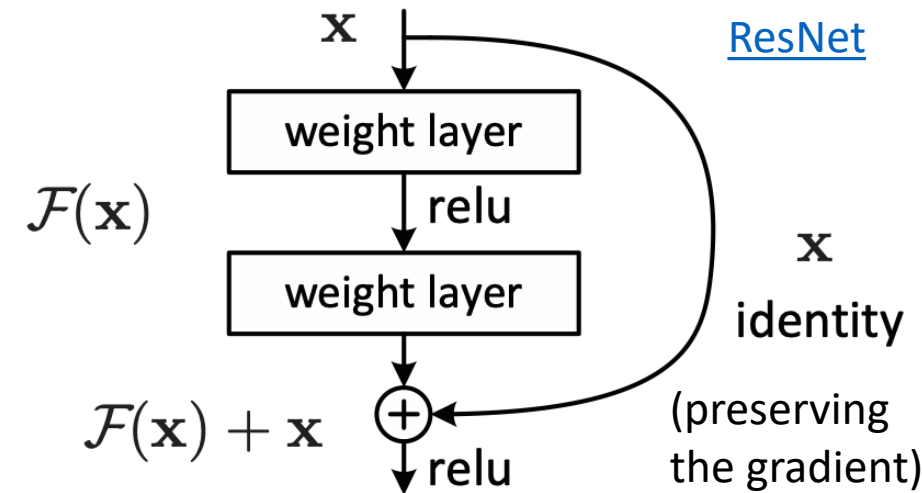
issue: degradation of training and test errors when adding more and more layers → not due to overfitting (but reason controversial)

solution: learning of residuals by means of skip connections (resulting in combination of different paths through computational graph)

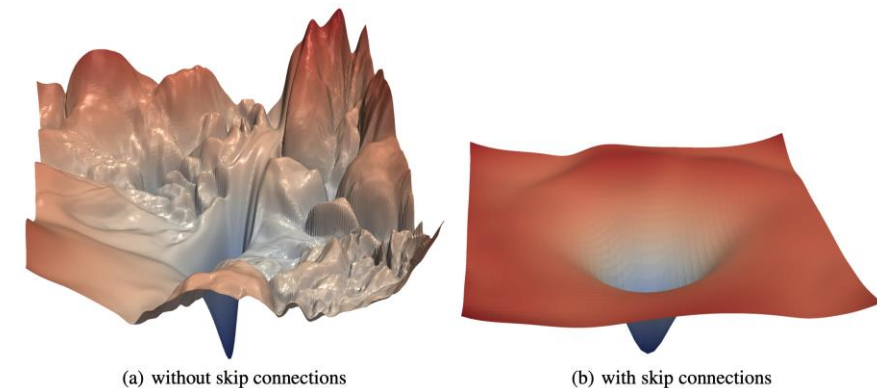
→ produces loss functions that train easier

together with batch normalization (avoiding exploding gradients), skip connections enable extremely deep networks (~1000 layers) without degradation

residual mapping (special kind of skip/shortcut connections):



loss surface:



[source](#)

# Batch Normalization

**Input:** Values of  $x$  over a mini-batch:  $\mathcal{B} = \{x_1 \dots x_m\}$ ;

Parameters to be learned:  $\gamma, \beta$

**Output:**  $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{ mini-batch mean}$$

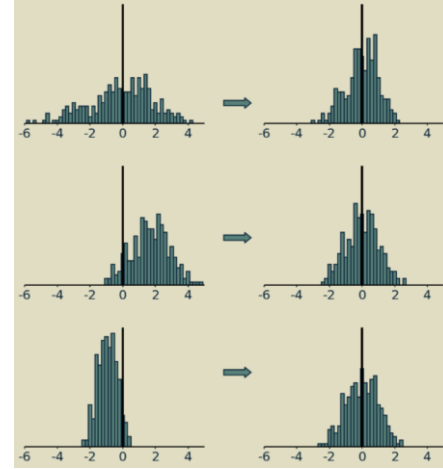
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{ mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{ normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{ scale and shift}$$

**Algorithm 1:** Batch Normalizing Transform, applied to activation  $x$  over a mini-batch.

[source](#)



[source](#)

adaptive reparameterization of inputs to a network layer (before or after activation)

independently for each input/feature

(not to confuse with weight normalization:  
decoupling of length and direction of  
weight vectors)

to maintain expressive power (optional):  
 $\gamma, \beta$  learned together with weights via  
back-propagation

# Benefits from Batch Normalization

- allows higher learning rates
- reduces importance of weight initialization
- alleviates vanishing/exploding gradients
- (implicit) regularization effect: introducing both additive and multiplicative noise, sometimes making dropout (multiplicative noise) unnecessary

reason why batch normalization improves optimization still controversial

most plausible explanation: smoothening of loss landscape (similar to skip connections)

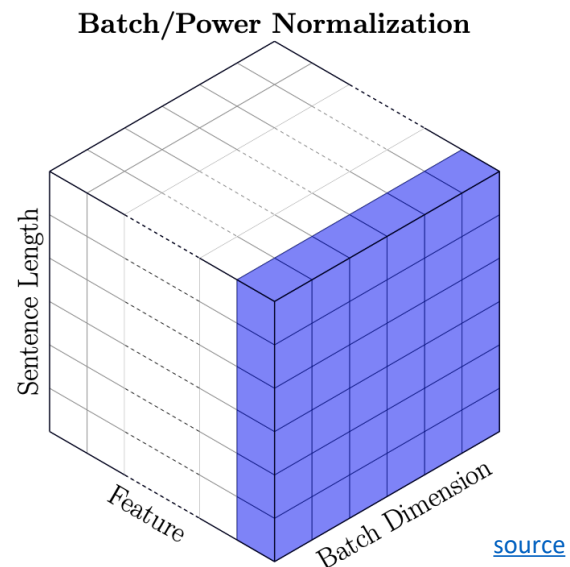
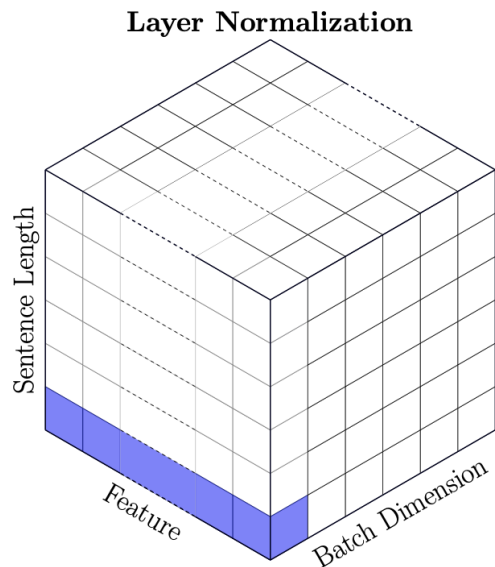
# Layer Normalization

[layer normalization](#)

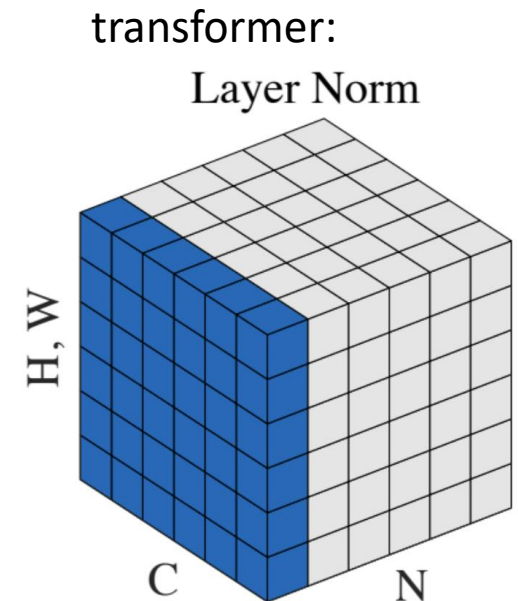
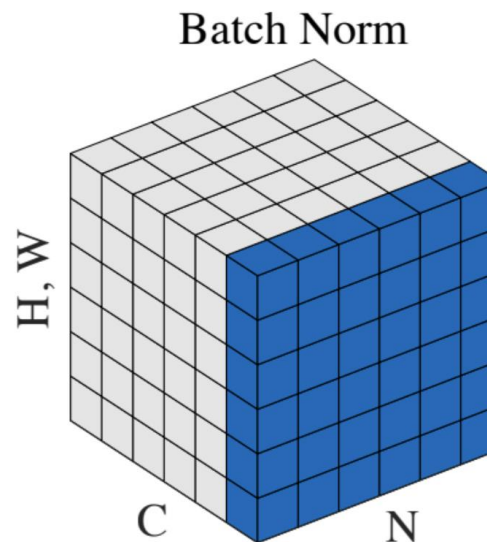
normalization over inputs/features, independently for each data sample

→ mean and variance shared over all hidden nodes of a network layer

batch norm often in computer vision (CNN), layer norm in NLP (variable-sized inputs)



[source](#)



[source](#)

# Comparison to Shallow Methods



# Feature Engineering vs Feature Learning

shallow learning:

representation encoded in features

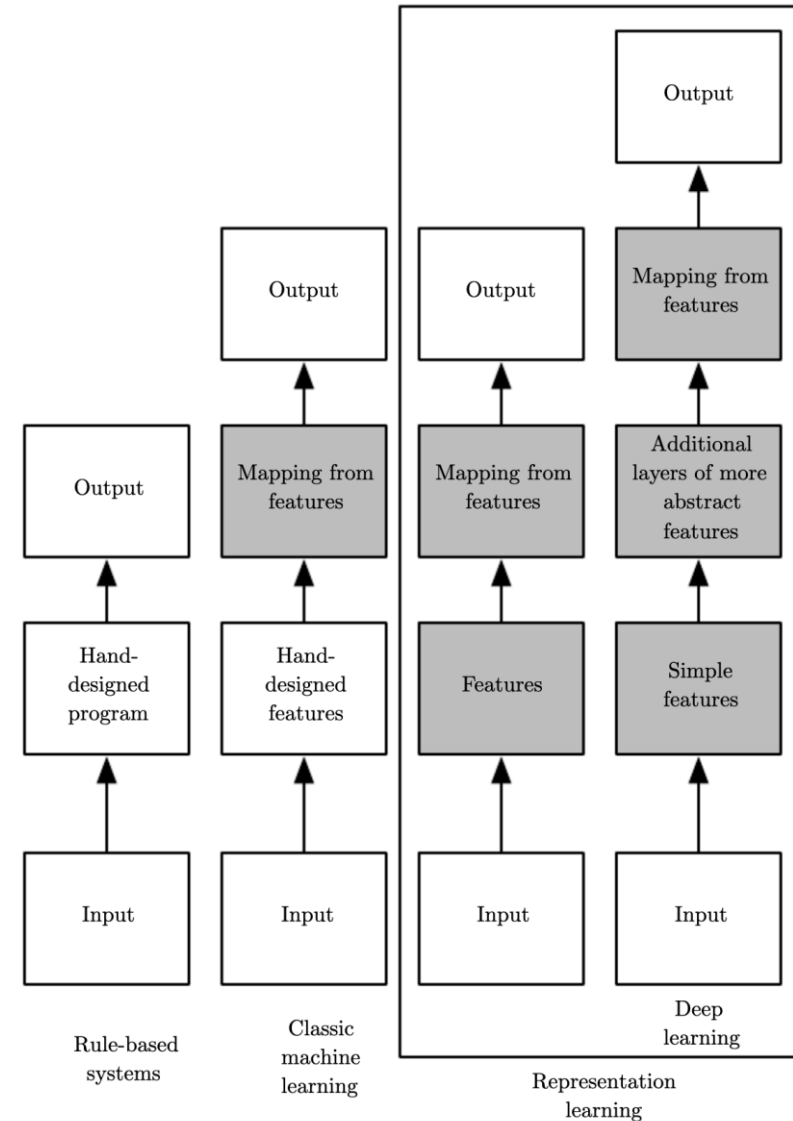
→ feature engineering

deep learning:

representation encoded in network

→ feature/representation learning

(hierarchy of concepts learned from raw data in deep graph with many layers)



[source](#)

# Tabular vs Unstructured Data

deep learning methods dominate applications on unstructured data (like text or images), but not necessarily on tabular data

typical characteristics of tabular data difficult to handle for deep learning:

- irregular patterns in target function (neural networks require piecewise continuous targets)
- uninformative features
- non-rotationally invariant data (linear combinations of features misrepresent the information)

tree-based models (e.g., gradient boosting) can naturally deal with these situations

# Categorical Variables

tabular data usually heterogenous, often with sparse categorical variables (like color of an object)

→ need for an encoding for categorical variables

different possibilities:

- ordinal encoding (introduces artificial order to unordered categories)
- leave-one-out encoding (use mean of target for given category excluding current row, used in CatBoost)
- one-hot encoding (can suffer from curse of dimensionality)
- embeddings (can also alleviate issue of non rotationally-invariant data)

# Embeddings

# Vector Representations

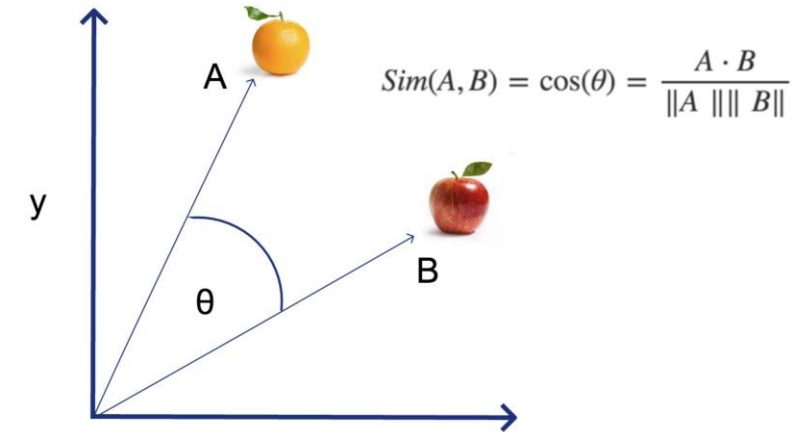
embeddings: representation of entities by vectors

similarity between embeddings by, e.g., cosine similarity  $\rightarrow$  semantic similarity

most famous application: word embeddings  
 $\rightarrow$  associations (natural language processing)

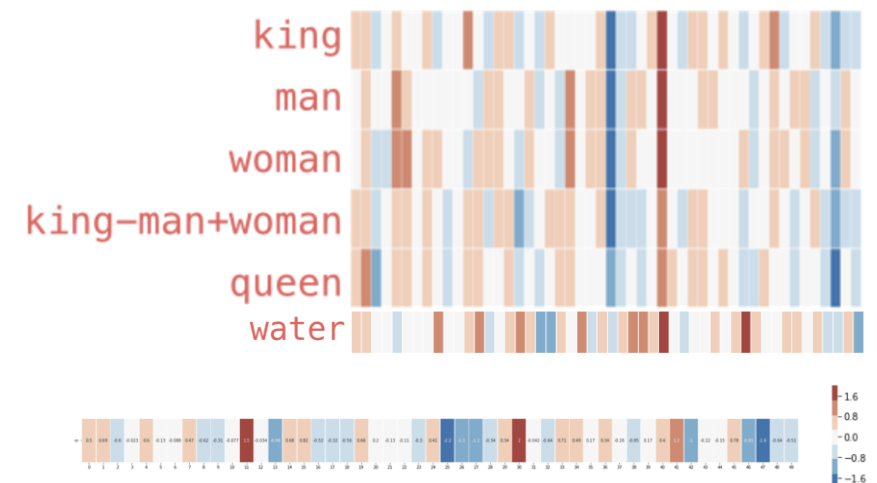
but general concept: embeddings of (categorical) features (e.g., products in recommendation engines)

learned via co-occurrence (e.g., [word2vec](#))



but also direction of difference vectors interesting (analogies):

king - man + woman  $\approx$  queen



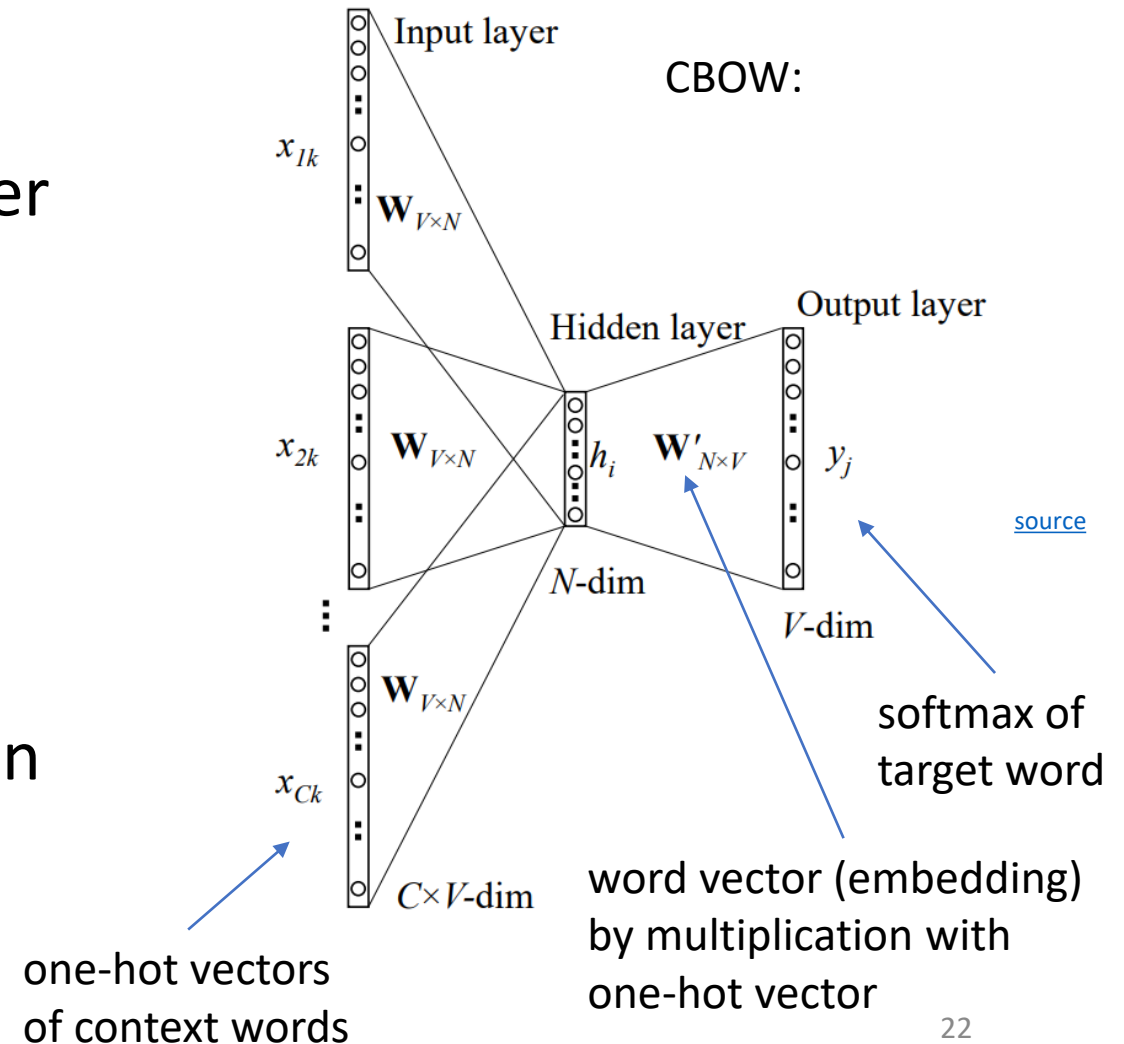
[source](#)

# Some Thoughts on Word Embeddings

can be implemented as

- neural network with single hidden layer (linear activation)
- using, e.g., bag-of-words approach (predict masked word from its surroundings)

→ not context-aware (need for attention or RNN)

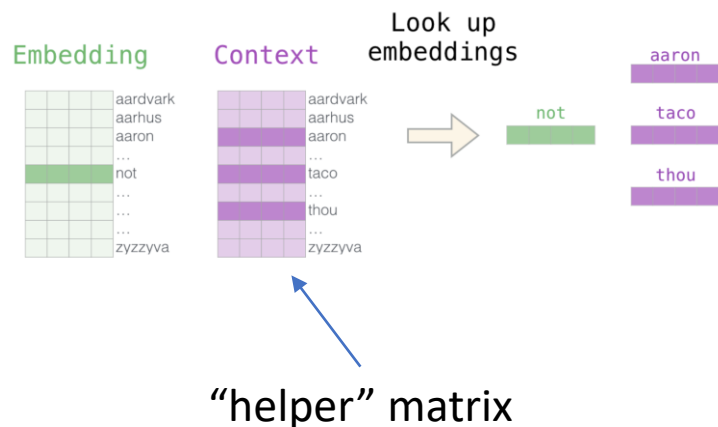


# word2vec

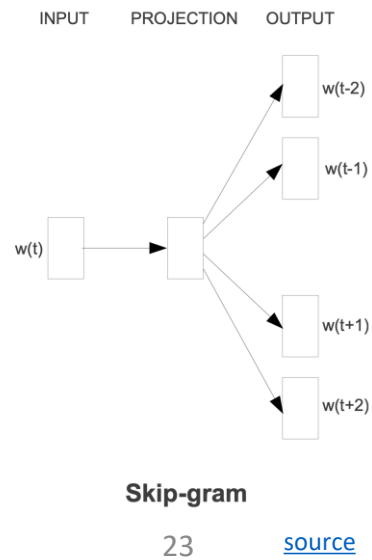
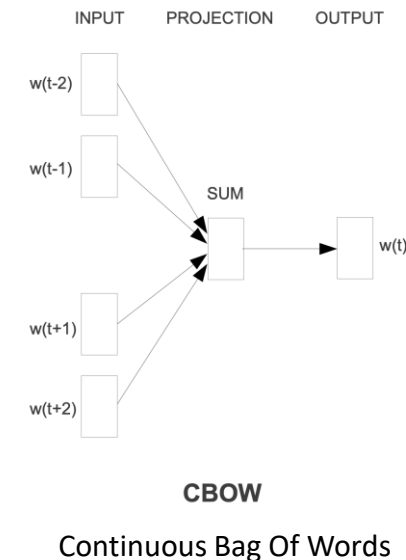
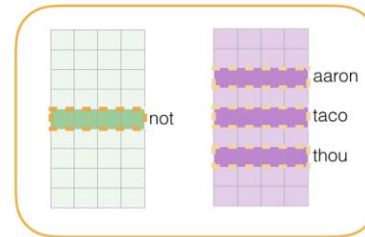
focus on generating word embeddings, not entire language model → negative sampling:

- use input and output words of language model as features, binary target if neighbors (dropping expensive projection to output vocabulary → much faster)
- include random negative samples (samples of words that are not neighbors)

not a deep neural network (just single hidden layer)



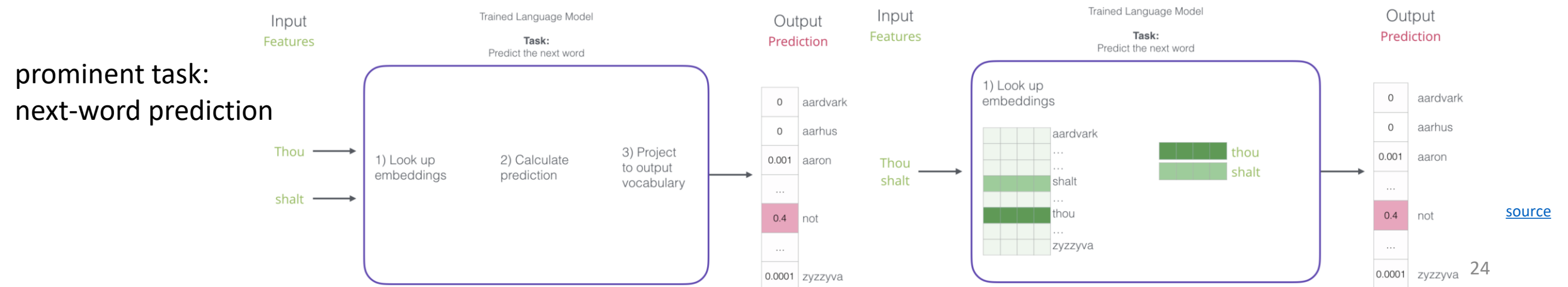
input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68



# Word Embeddings as Part of Language Model

language models contain embedding matrix as part of learned parameters

- can be extracted and subsequently used as pre-trained embeddings for other task
- typically several hundred dimensions for word vectors
- trained on huge data sets (millions in vocabulary)





# Neural Language Models

using neural networks: learning distributed representation of words

self-supervised learning: sliding (with some sliding window) over text to generate training data set

learning of embeddings: kind of feature learning (dimensions of embeddings vectors)

contextually-meaningful embeddings can be learned by means of sequence models (RNN/LSTM, transformer)

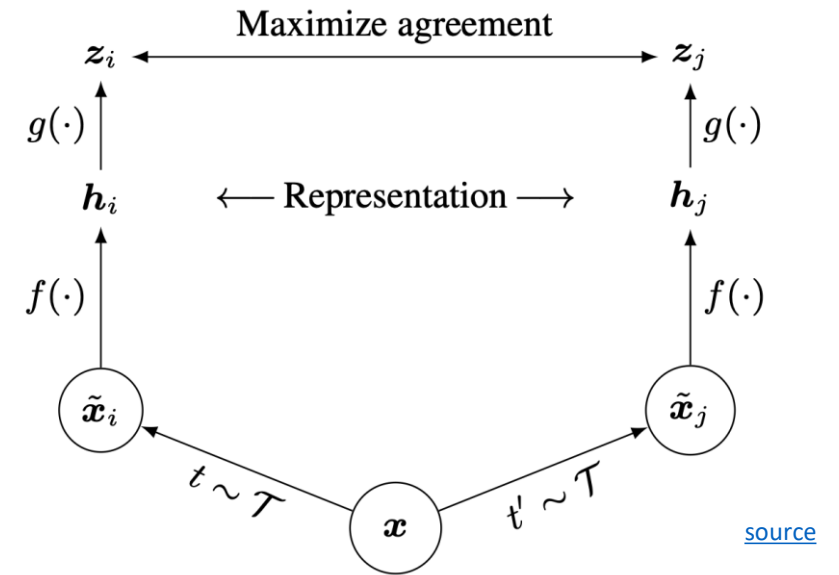
# Contrastive Learning

goal: create embedding space in which similar samples are close to each other and dissimilar ones are far apart

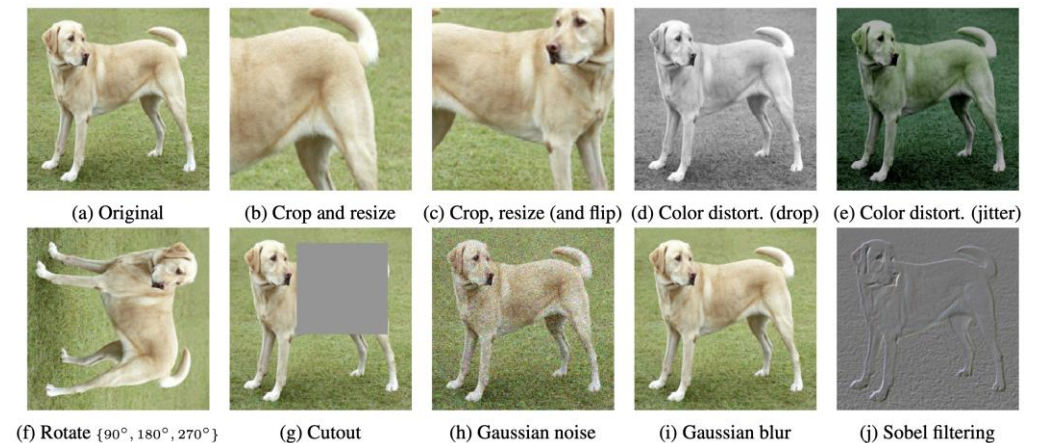
often learned in a self-supervised way

examples:

- natural language processing: word2vec
- computer vision: [SimCLR](#), [SimCLRv2](#) (learning of image representations)



[source](#)



# Autoencoders

# Representation Learning

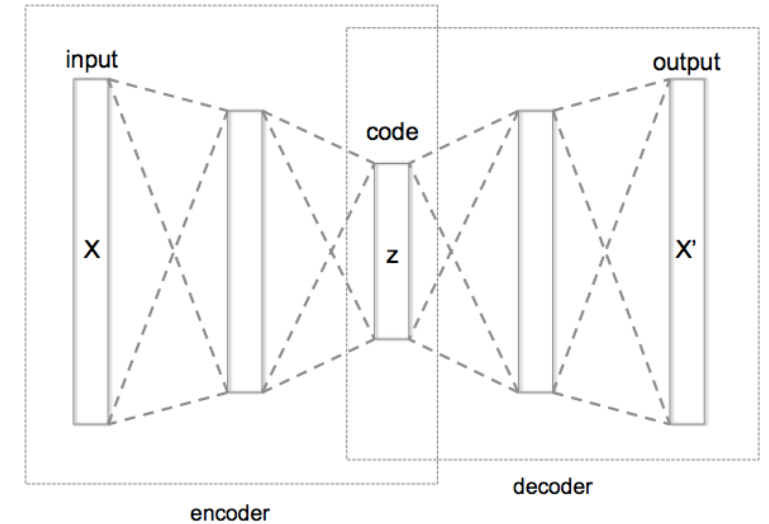
autoencoders as prime example of representation learning  
combination of

- encoder: converting input data into different representation (code)
- decoder: converting learned representation back into original format

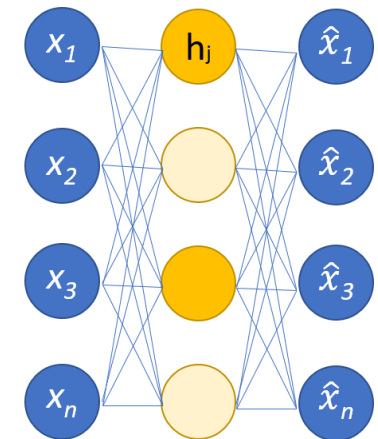
possibilities to avoid simple duplication:

- undercomplete autoencoders: code with smaller dimension (less nodes) than input (generalized PCA: autoencoder is PCA if encoder and decoder are linear transformations)
- sparse autoencoders: sparsity penalty to deactivate hidden nodes (e.g., with help from ReLU activation)

learned in the same way as feed-forward neural networks



from wikipedia



# History: Unsupervised Pre-Training

breakthrough in effectiveness of deep learning training in 2006:

Deep Belief Networks introduced idea of greedily initializing each layer by unsupervised learning

(using an energy-based method called Restricted Boltzmann Machine)

→ commonly seen as actual starting point of deep learning wave

only later, ReLU activation functions (and other improvements) enabled effective deep learning without unsupervised pre-training

but unsupervised pre-training still beneficial in context of semi-supervised learning, using large amounts of unlabeled data

# Stacked Autoencoders

besides dimensionality reduction, autoencoders can also be used (instead of Restricted Boltzmann Machines) for unsupervised pre-training

→ feature learning (internal distributed representations, high-level abstractions of input data), initializing weights in region near a good local minimum

stacking of autoencoders (or RBMs) → industry adoption:

- object recognition ([cat paper](#))
- [speech recognition in industry](#)

stimuli for cat neuron:



[source](#)

# Recurrent Neural Networks (RNN)

# Sequential Structures

speech recognition, natural language processing, time series, ...

problem: need to generalize across different points in time (or multiple positions of words within a sentence) and learn from context

idea: parameter sharing across different parts of model

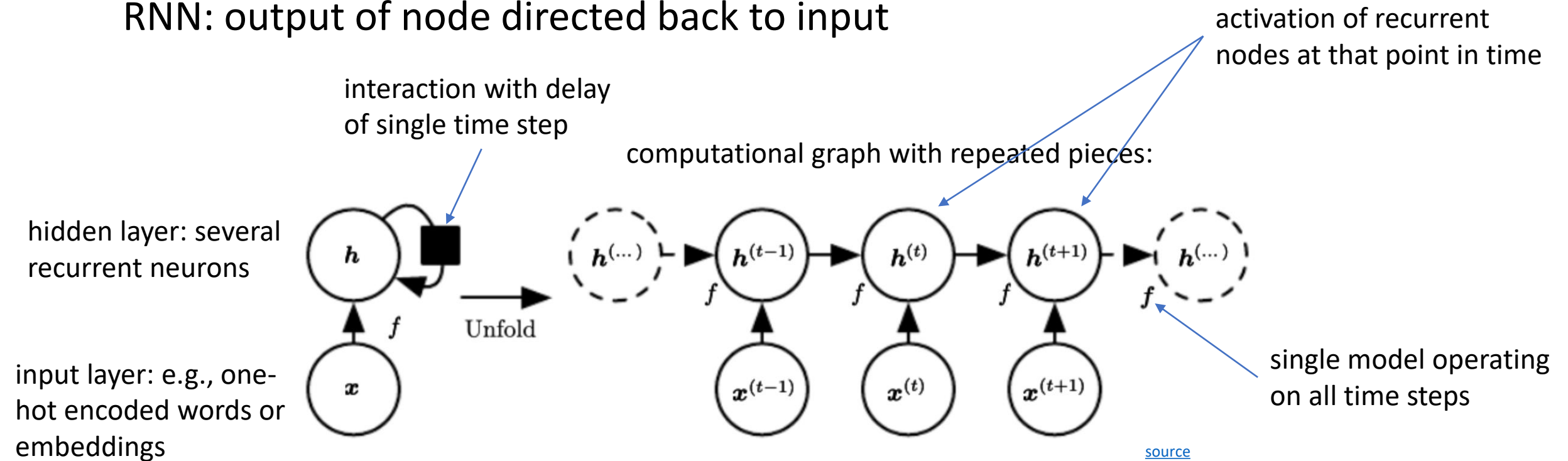
CNNs apply parameter sharing (convolutions) on grid-like structures (including 1-D grids like time series or speech), but this is limited to neighboring inputs (per layer).

→ need for approach to learn sequential structures (process input as stream, e.g., speech, rather than one batch, e.g., image)



# Back-Propagation through Time

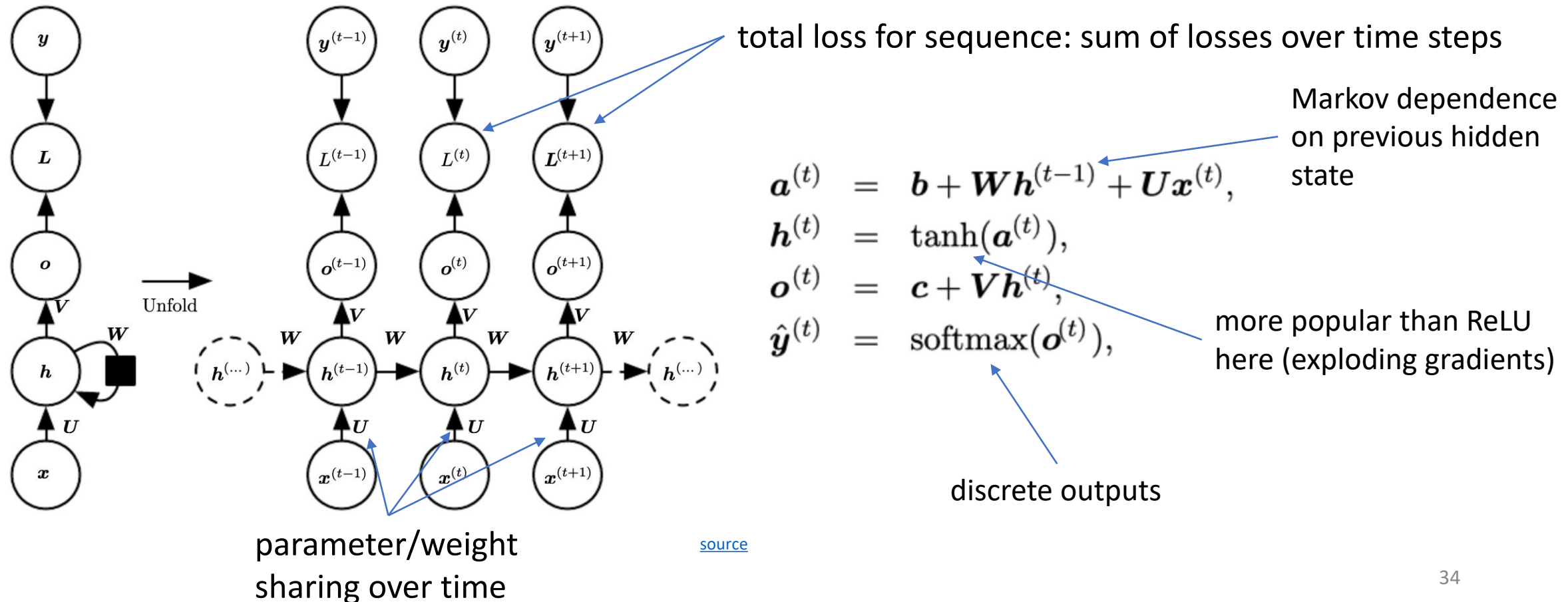
RNN: output of node directed back to input



different kind of depth: one recurrence for each sequential step (e.g., word)

# Weight Sharing across Time

example: input-output mapping at each time step with recurrent hidden nodes



# Further Examples

summarization of sequence:

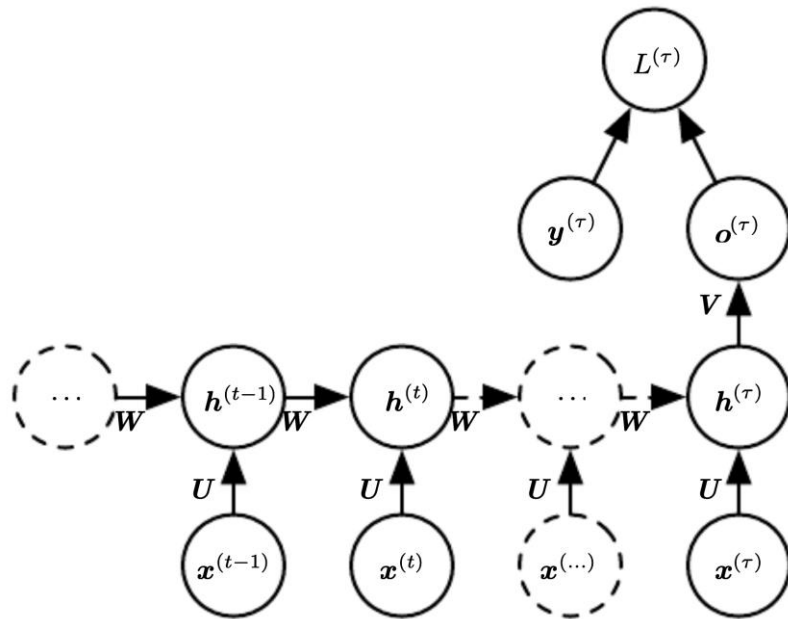
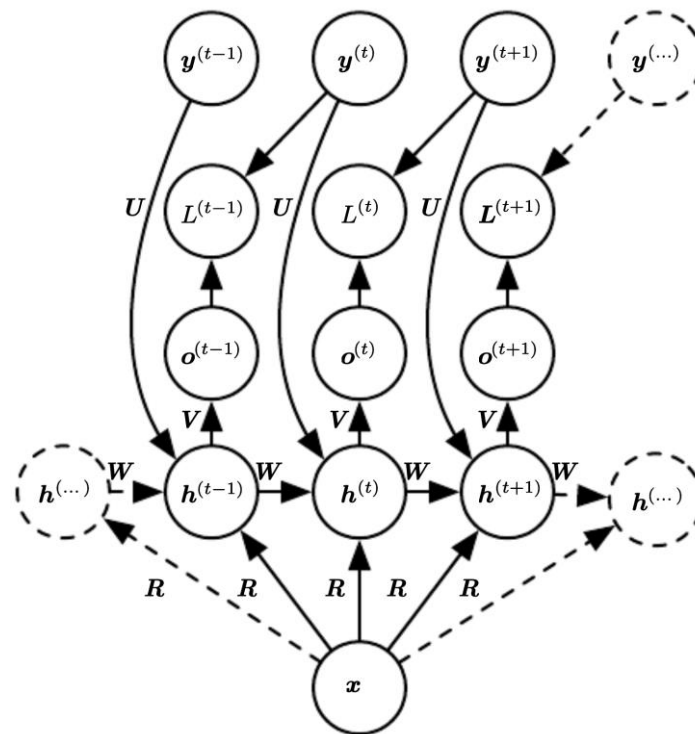
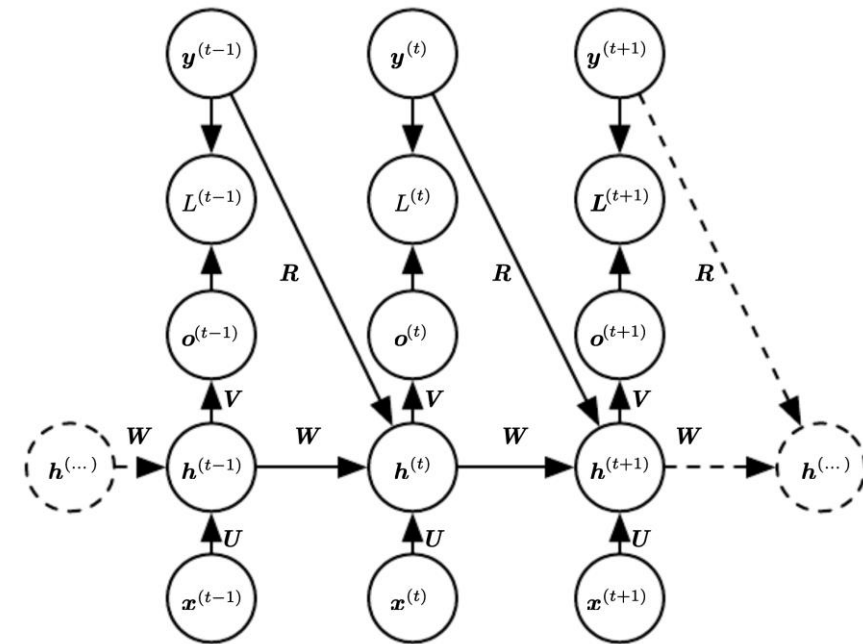


image captioning:



conditional sequence:

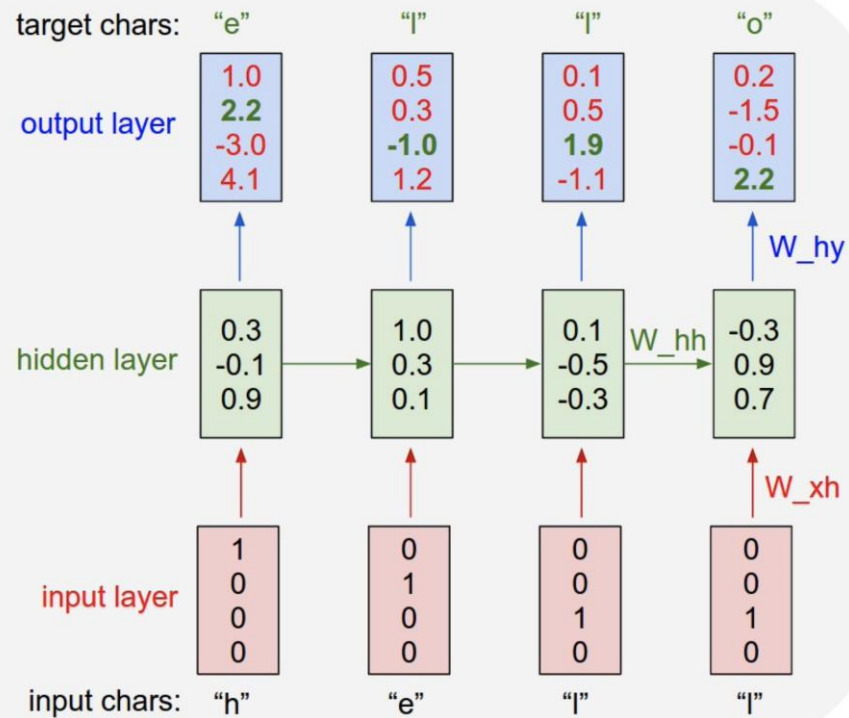


# Visualization

neuron getting excited inside URLs (**excited**, not excited):

t t p : / / w w w . y n e t n e w s . c o m / ] E n g l i s h - l a n g u a g e w e b s i t e o f t  
t p : / / w w w . b a c a h e t s . c o m / - x g l i s h l i n g u a g e s a i r s i t e o f t

next character prediction



[source](#)

high/low activation:

Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

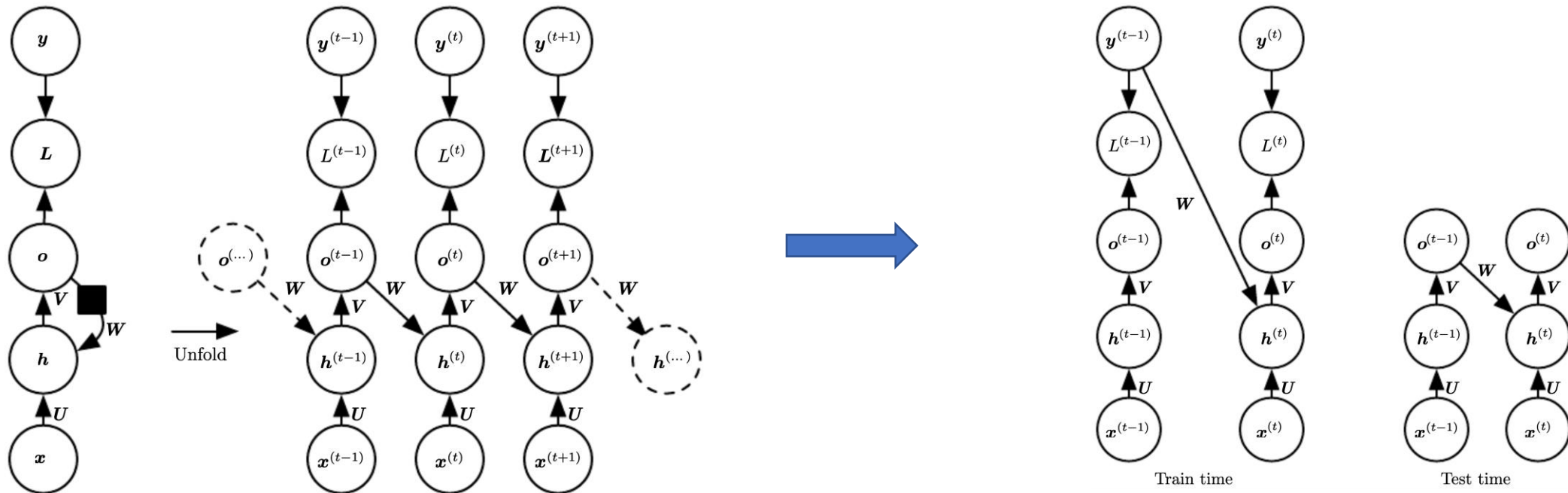
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
                           siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!(current->notifier)(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```

# Teacher Forcing

for feedback from output to hidden layer: instead of feeding model output back into itself, use target values directly



[source](#)



# Gated RNNs

issue with long chain of gradients through recurrences:  
vanishing (mainly) and exploding gradients in back-propagation

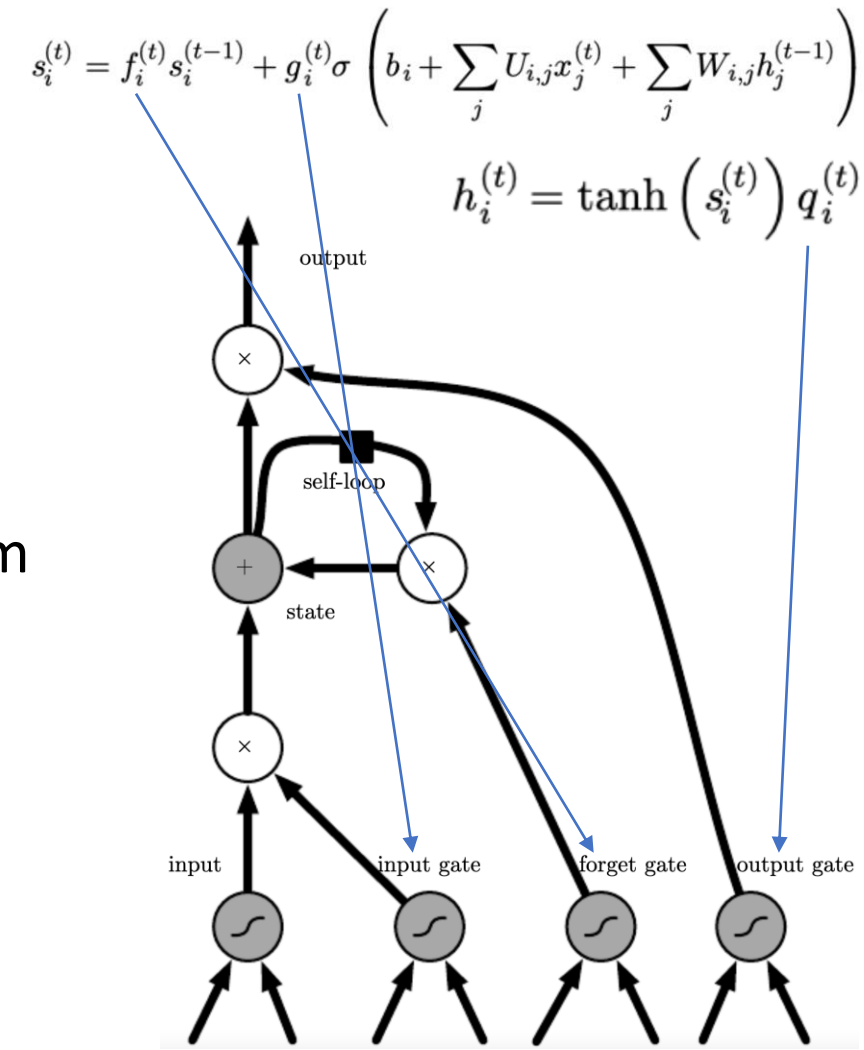
→ need to focus on important sequence elements

replace usual recurrent hidden nodes with long short-term memory (LSTM) cells with internal recurrence (self-loop)

- linear self-loop in addition to outer recurrence of RNN: error carousel preserving (→ long) short-term memory (i.e., activation patterns)
- sigmoid activations with independent weights for input, forget, and output gates

other prominent gated RNN: gated recurrent unit (GRU)

long-term memory: weights  
short term memory: activation patterns



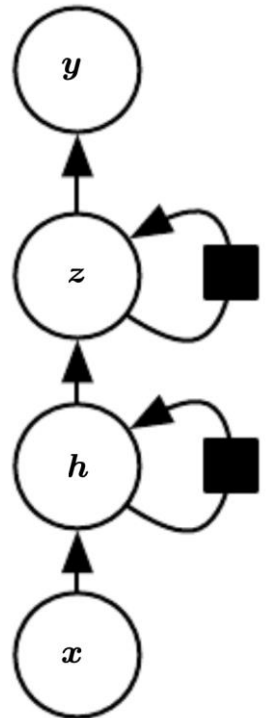
[source](#)

# Issues with RNNs

- especially LSTMs very computationally expensive (with lots of parameters)
- transfer learning (using pre-trained network layers on new task, e.g., popular for CNNs) difficult

(hierarchical) representation learning: go deep by stacking several layers of recurrent nodes (e.g., important for [speech recognition](#))  
→ worsening efficiency even more

(self-)attention and transformers to the rescue ...



[source](#)

# Literature

easily digestible overview:

[The Little Book of Deep Learning](#)

papers:

- [Deep Learning overview](#)
- [ResNet](#)
- [A Neural Probabilistic Language Model](#)
- [word2vec](#)
- [training of Deep Belief Nets](#)

blogs:

- [The Illustrated Word2vec](#)
- [Karpathy on RNNs](#)



# Black-Box Models

To build trust in AI systems, individual predictions/actions need to be fully transparent, i.e., explainable.

Unfortunately, complex models like deep learning methods are difficult to interpret.

→ need for model-agnostic methods to explain black-box models

examples: local surrogates ([LIME](#)), Shapley values ([SHAP](#))

[overview](#)