Artificial Neural Networks - Improving training and performance

Agenda

- The Challenges: Over fitting & local optima
- The Training
  - Epochs, Batch Size, Iterations
  - Gradient Descent (GD) Vs Stochastic GD (SGD) Vs Mini-Batch GD
  - SGD with momentum
  - Learning rates and adaptive learning rates
  - Weight Initialization
  - Batch Normalization
- Guarding against over-fitting
  - L1/L2 Regularization
  - Data Augmentation
  - Drop outs
- Neural Network Architectures

Epochs, Batch size, Iterations

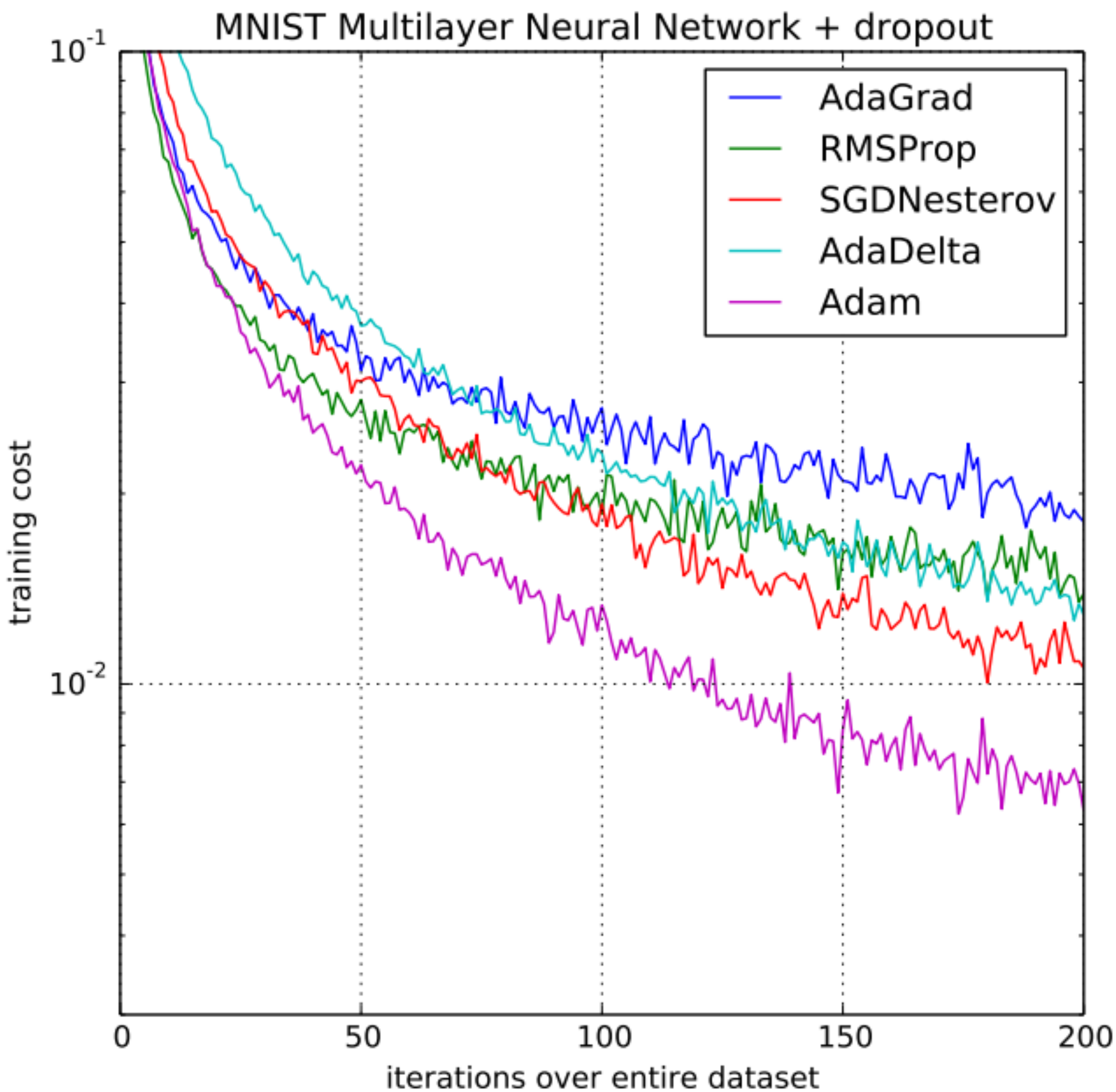
- When dataset is too large, passing all the data through a Neural net before we make weight updates is computationally expensive
- Instead we would create data batches with smaller batch size.
- After each batch is passed and weights updated, we will count it as one iteration.
- When an entire dataset is passed forward and backward (weights updated) through the neural network, we will count it as one epoch.
  - Too few epochs: under fitting, Too many: overfitting
  - Batch training: All of the training samples pass through the neural net, before weights are updated
  - Sequential training: Weights are updated after each training vector is passed through the neural net.

Gradient Descent and its Variants

- Gradient Descent
- Stochastic Gradient Descent
- Mini-Batch GD
- SGD with Momentum

Learning rate

- Choosing the Learning rate (η)
  - Too small, we will need too many iterations for convergence
  - Too large, we may skip the optimal solution
- Adaptive Learning Rate :
  - start with high learning rate and
  - gradually reduce the learning rate with each iteration.
  - Moreover, having different learning rates for different weight updates will help: Adagrad, RMS Prop



Diederik P. Kingma and Jimmy Ba (2014). Adam: A Method for Stochastic Optimization.

Weight Initialization Techniques

- Essential for efficient neural network training (esp. deep Neural Nets)
- Impacts:
    - gradient stability,
    - symmetry breaking,
    - reduced over fitting
    - and convergence speeds.
- Key Techniques:
    - Random Initialization: Small random numbers to break symmetry.
    - Xavier/Glorot Initialization: Scaled according to input and output size, for tanh/sigmoid activation.
    - He Initialization: Similar to Xavier but for ReLU activations.

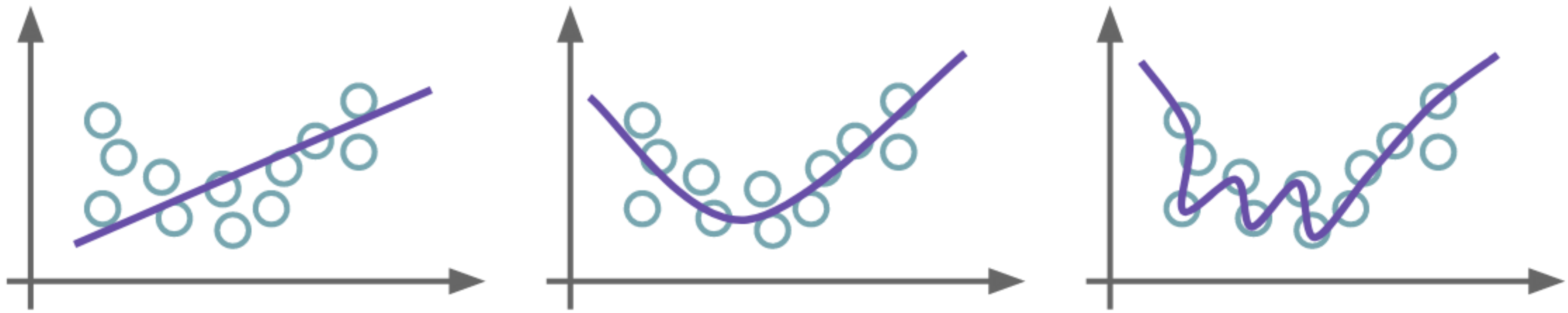
Batch Normalization

- Moving forward in neural nets, activations can become large or very small, and hence training can become unstable> weight initiations, input normalizations etc.
- What is Batch Normalization?
  - Batch…. Normalize
- Why does this work?
  - Covariate shifts> last layer > dist of inputs > keeps changing during training!
  - Regularization: each row when seen as a part of different batches, looks different in training. Disallows overfitting. Almost like data augmentation
- Normalization when testing or deploying?

Overfitting

- Neural Network Models are susceptible to overfitting
  - Large number of weights and biases
  - Excessive learning (Epochs) on training data
- Ways to avoid Overfitting
  - Increase sample size
  - Early stopping
  - Reduce Network Size
  - Regularization

Under Vs Over-fitting



L1/L2 Regularization

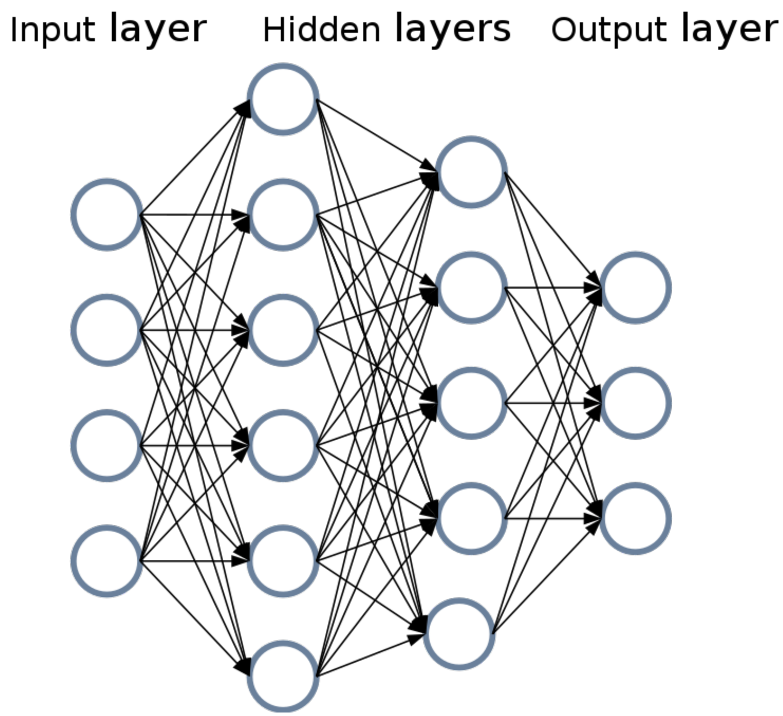
- Regularization techniques are in general methods to avoid this overfitting problem.
- The idea behind regularization is that models that overfit the data are complex models that have for example too many parameters.
- Weight decay based regularization, penalizes the usual loss function by adding a complexity term that would give a bigger loss for more complex models.
- Types of Regularization
  - LASSO (L1)
  - Ridge (L2)
- Optimal value of λ, the decay rate or penalty coefficient, is determined through cross-validation

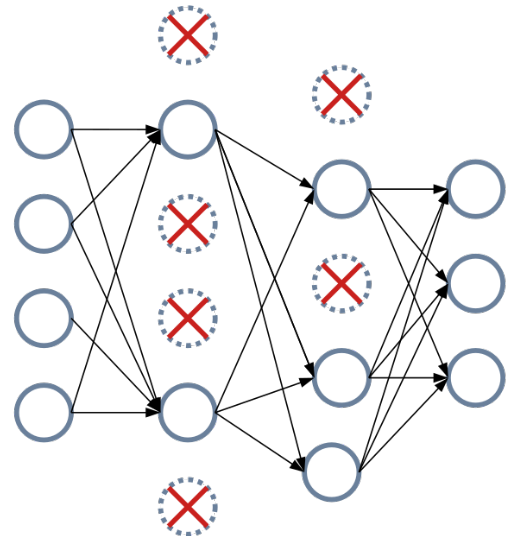
Data Augumentation

-source: towardsdatascience.com, MachineX: Image Data Augmentation Using Keras

Dropout

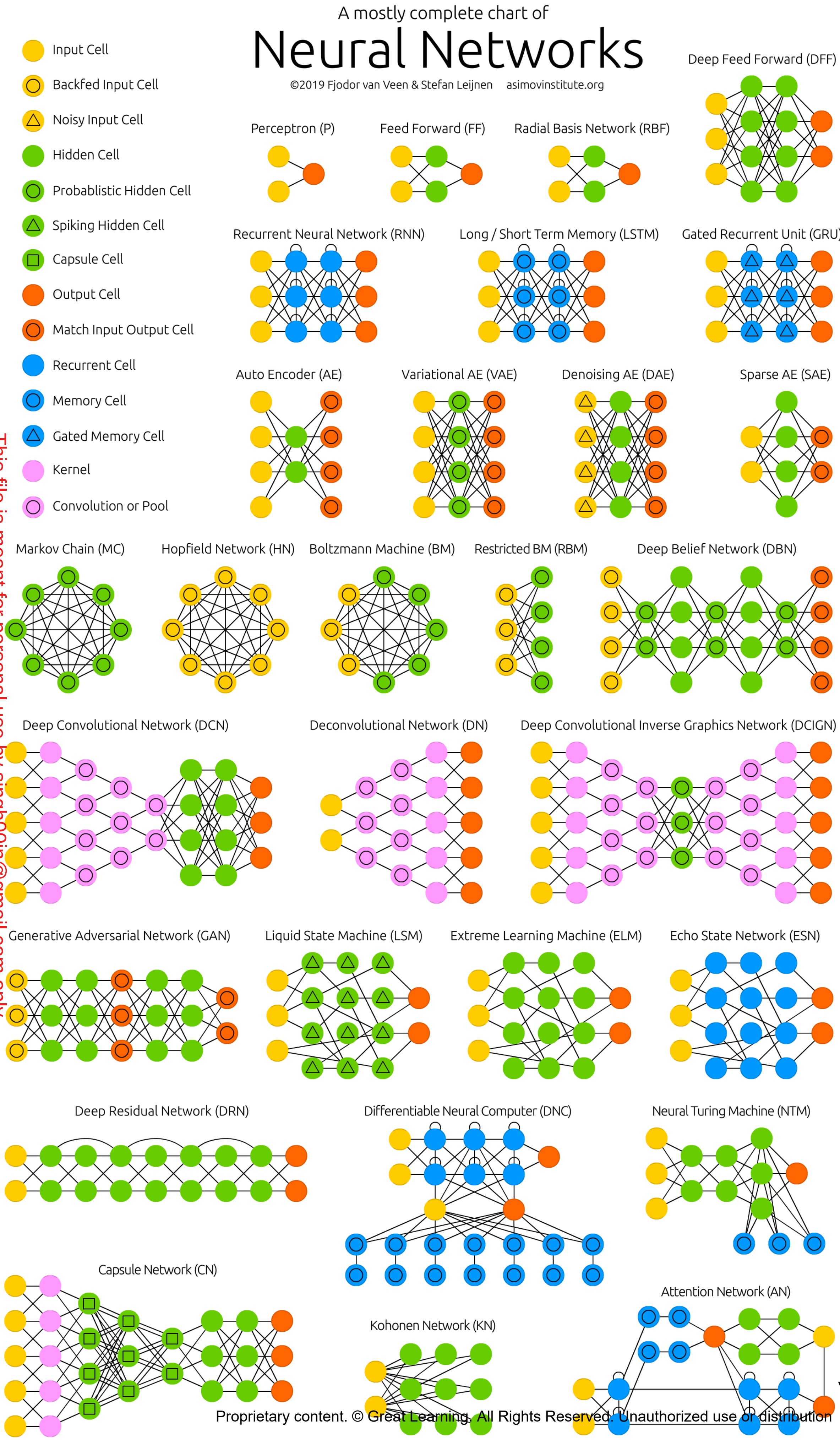
- What is dropout? How is it implemented?
- Only in training
- Why does it work
  - Usual training: some nodes gain more predictive power than others
  - Forces redundancy and robustness
  - Training teams
  - Co-adapt > units change in a way to fix up mistakes induced in previous layers > overfitting to data (does not generalize) > Regularization
- Scaling weights appropriately after training.


Input layer
Hidden layers
Output layer

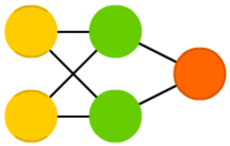


Types of NN

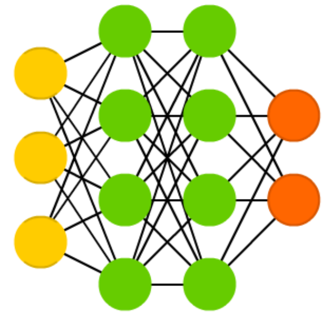
- Feed Forward
  - MLP
  - DNN
  - CNN
- RNN
  - LSTM


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

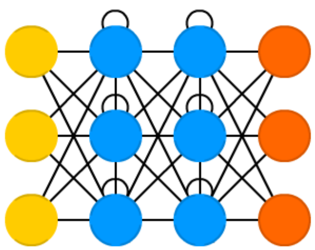
Feed Forward (FF)



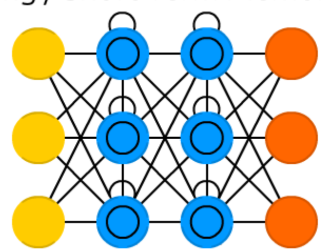
Deep Feed Forward (DFF)



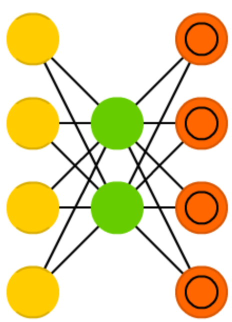
Recurrent Neural Network (RNN)



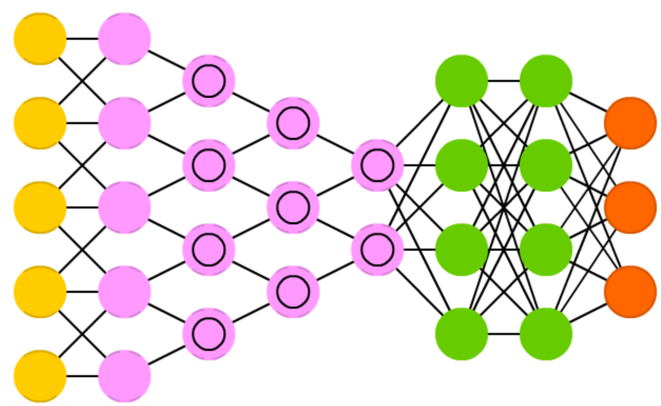
Long / Short Term Memory (LSTM)



Auto Encoder (AE)



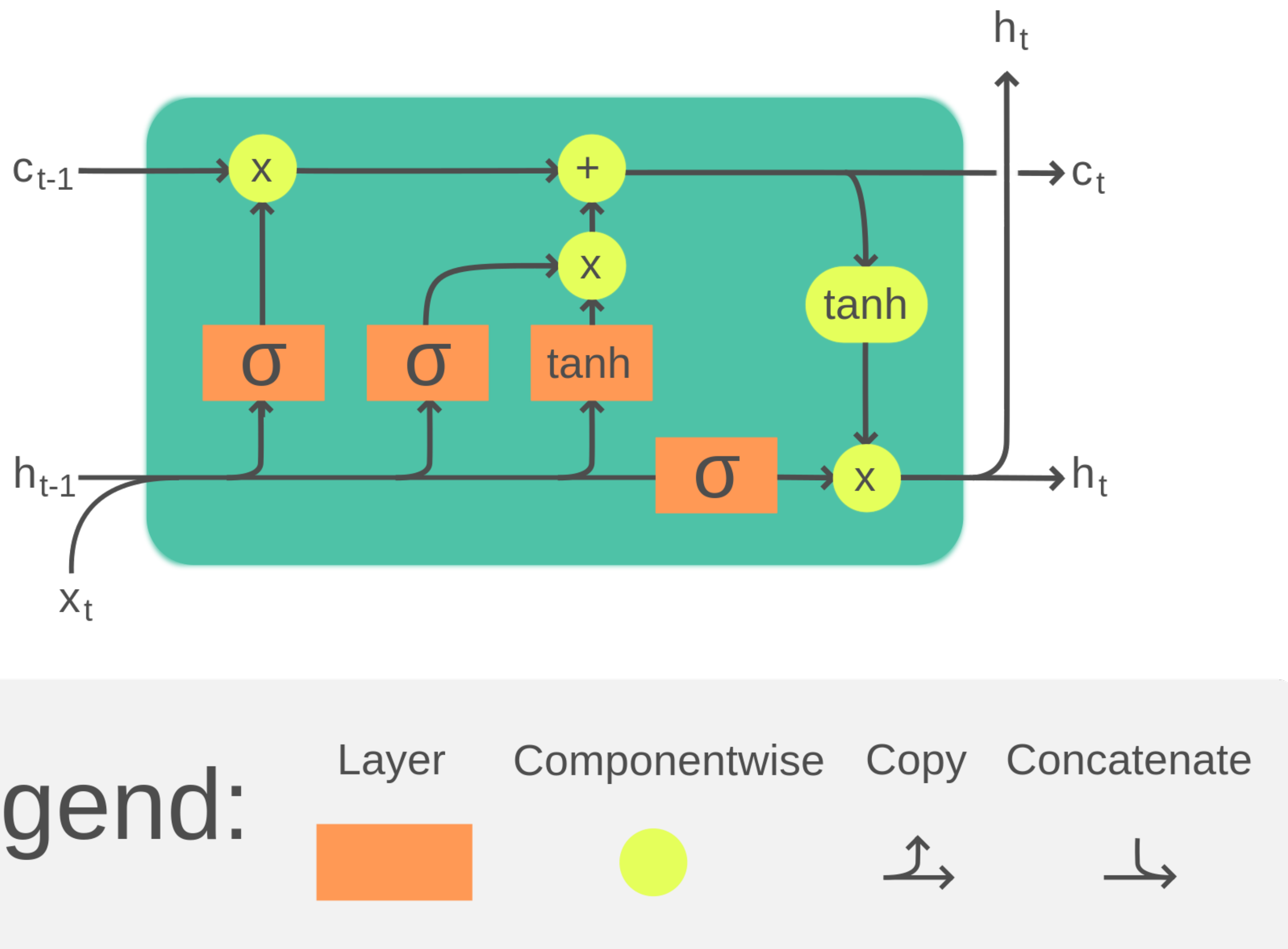
Deep Convolutional Network (DCN)



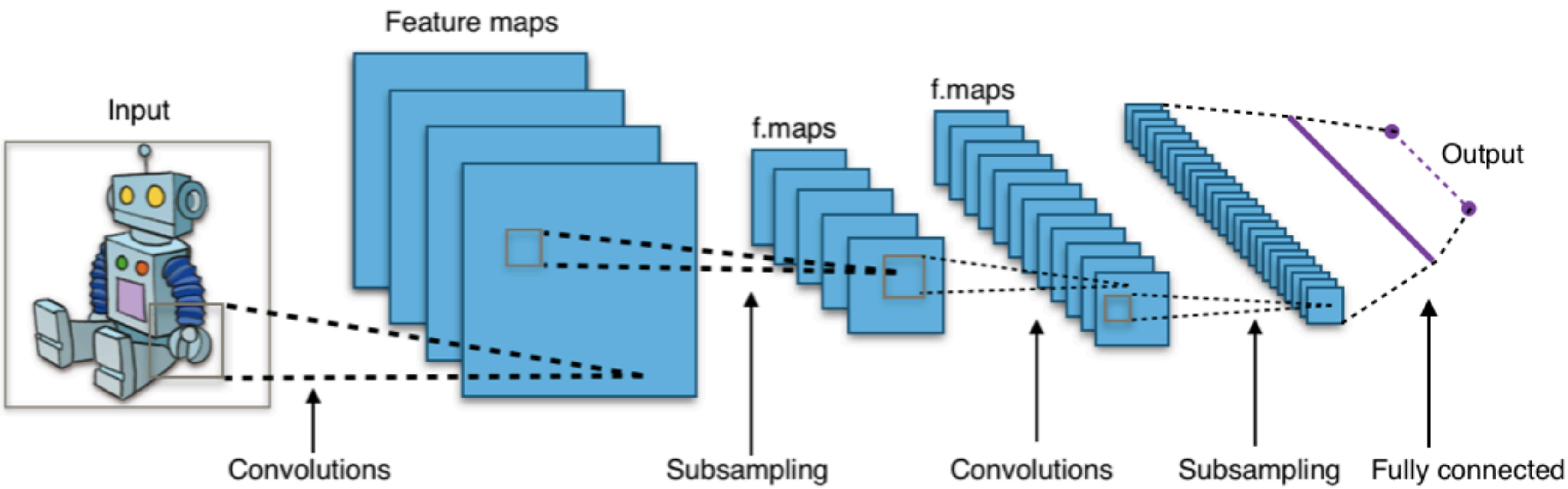
Text generation using an RNN

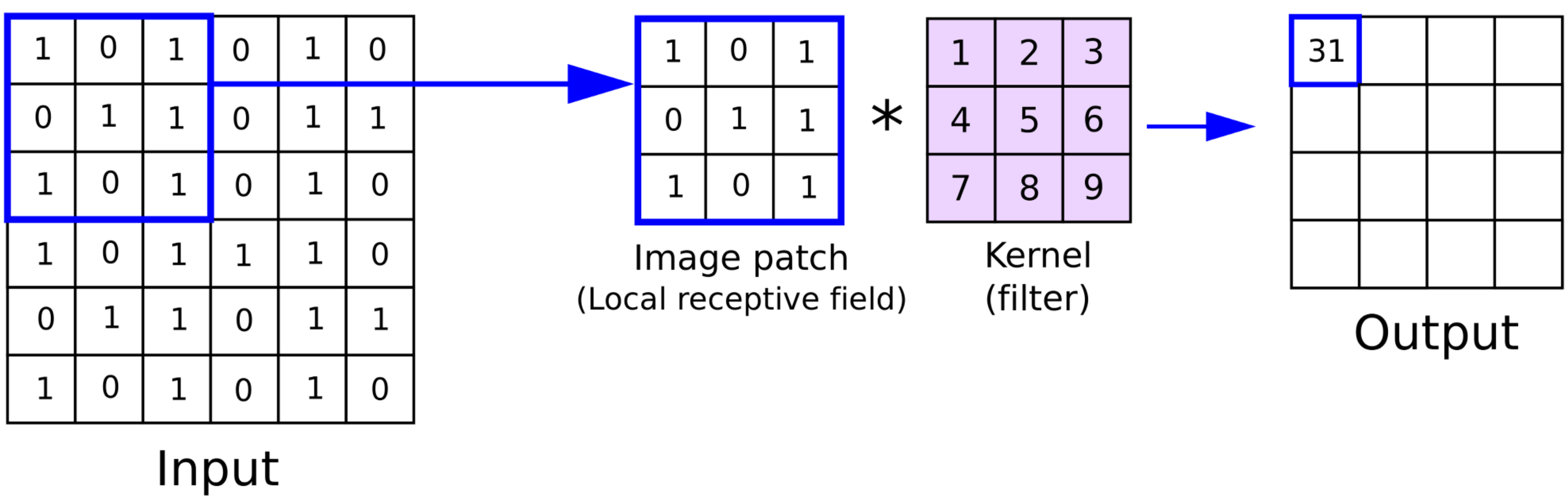
The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the

Ilya Sutskever, et. al. "Generating text with recurrent neural networks", ICML, 2011



Source: wikipedia





Source: wikipedia, iq.opengenus.org