Neural networks

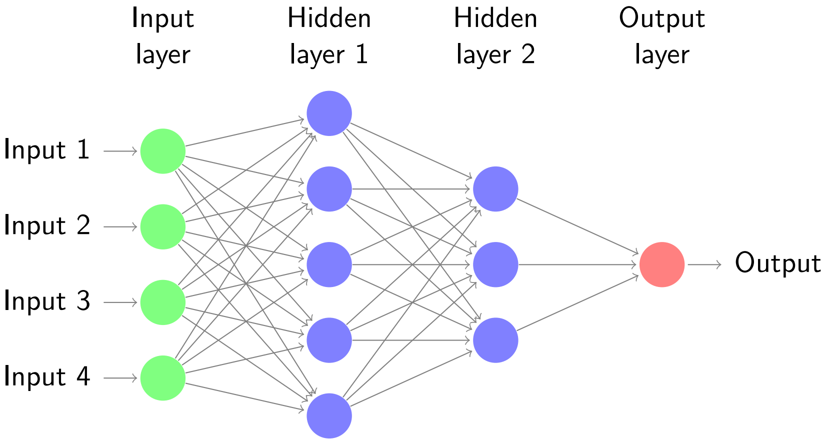
Architectures and training tips

Sebastian Björkqvist

IPRally Technologies

09.01.2019

What is a neural network?



Modified from `http://www.texample.net/tikz/examples/neural-network/`

What is a neural network?

At each hidden layer node i the output value is calculated by

$$o_i = \sigma(\sum w_{ki} o_{ki-1} + b_i).$$

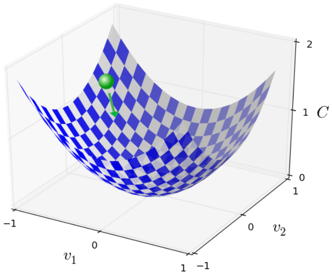
What is a neural network?

At each hidden layer node i the output value is calculated by

$$o_i = \sigma(\sum w_{ki} o_{ki-1} + b_i).$$

The function σ is called the activation function. It must be non-linear to allow the network to learn non-linear dependencies.

Training neural networks using SGD



[Nielsen, 2015], Chapter 1.

The training data is processed in small batches, and the weights of the model are iteratively updated by going in the direction of the negative gradient of the loss function.

Why neural networks?

- Can approximate any function [Hornik, 1991]

Why neural networks?

- Can approximate any function [Hornik, 1991]
- May learn to respond to unexpected patterns

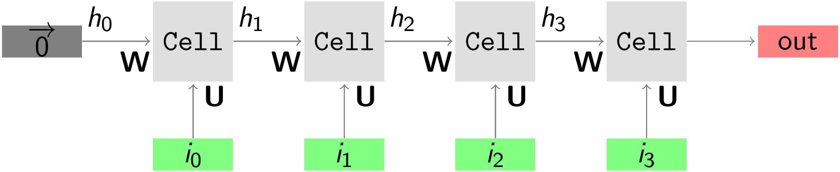
Why neural networks?

- Can approximate any function [Hornik, 1991]
- May learn to respond to unexpected patterns
- Useful especially when the amount of data is large

Why neural networks?

- Can approximate any function [Hornik, 1991]
- May learn to respond to unexpected patterns
- Useful especially when the amount of data is large
- Less need for feature engineering compared to traditional ML methods

Recurrent neural network (RNN)



Processes each element of the input sequence in order, and keeps information about the past elements in a hidden state vector.

Recurrent neural network (RNN)

At each timestep t the new hidden state is calculated using the new input at this timestep and the existing hidden state. The most basic version is the following:

$$h_t = \sigma(Wh_{t-1} + Ui_t + b).$$

Recurrent neural network (RNN)

At each timestep t the new hidden state is calculated using the new input at this timestep and the existing hidden state. The most basic version is the following:

$$h_t = \sigma(Wh_{t-1} + Ui_t + b).$$

Other RNN architectures (for instance LSTM or GRU) use more complicated ways of updating the hidden state to control the flow of information to and from the hidden state.

RNN pros and cons

RNN pros and cons

+ Accepts input of variable size, i.e. sequences (time series, sentences etc)

RNN pros and cons

- + Accepts input of variable size, i.e. sequences (time series, sentences etc)
    - Can even be used to process tree-structured inputs by using Tree-LSTMs [Tai et. al., 2015]

RNN pros and cons

- + Accepts input of variable size, i.e. sequences (time series, sentences etc)
  - Can even be used to process tree-structured inputs by using Tree-LSTMs [Tai et. al., 2015]
- + May learn long-term dependencies

RNN pros and cons

+ Accepts input of variable size, i.e. sequences (time series, sentences etc)
    - Can even be used to process tree-structured inputs by using Tree-LSTMs [Tai et. al., 2015]
+ May learn long-term dependencies

– Training may be slow when sequence length is large

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

1. Patents are transformed to graphs by extracting the relevant information from the patent claims and specifications

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

1. Patents are transformed to graphs by extracting the relevant information from the patent claims and specifications
2. The graphs are then embedded to vectors by using a Tree-LSTM model

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

1. Patents are transformed to graphs by extracting the relevant information from the patent claims and specifications
2. The graphs are then embedded to vectors by using a Tree-LSTM model
    - The model is trained by using millions of real-life positive and negative novelty citations from previous patent applications

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

1. Patents are transformed to graphs by extracting the relevant information from the patent claims and specifications
2. The graphs are then embedded to vectors by using a Tree-LSTM model
    - The model is trained by using millions of real-life positive and negative novelty citations from previous patent applications
    - Patents with a positive citation get vectors that are close to each other

RNN real life use case: Patent search

At IPRally we work on automated patent searches. The basic idea is the following:

1. Patents are transformed to graphs by extracting the relevant information from the patent claims and specifications
2. The graphs are then embedded to vectors by using a Tree-LSTM model
   - The model is trained by using millions of real-life positive and negative novelty citations from previous patent applications
   - Patents with a positive citation get vectors that are close to each other
3. A prior art search for a new patent can then be done by searching for the nearest neighbors of the vector created from the new invention

RNN real life use case: Patent search

- table
  - worktop
  - legs
    - telescope pipe
    - motor

IPRally

US patent no. 7 908 981

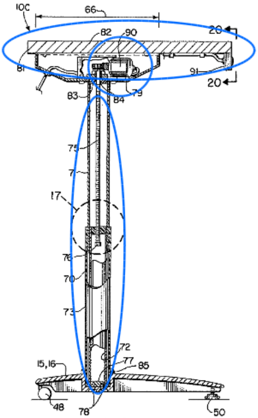


FIG. 16.

Convolutional neural network (CNN)

TODO: Picture here

Extracts features of two-dimensional input (usually an image) using convolutional and pooling layers.

CNN pros and cons

CNN pros and cons

+ Works well with image data

CNN pros and cons

- + Works well with image data
- + Training can be effectively parallelized

CNN pros and cons

+ Works well with image data
+ Training can be effectively parallelized
+ Pre-existing models can be fine-tuned for specific tasks

CNN pros and cons

+ Works well with image data
+ Training can be effectively parallelized
+ Pre-existing models can be fine-tuned for specific tasks
− Does not take into account position or orientation of the object

Challenges when training neural networks

Challenges when training neural networks

- Finding the optimal neural network layout is often time-consuming

Challenges when training neural networks

- Finding the optimal neural network layout is often time-consuming
- The model may be sensitive to changes in hyperparameters

Challenges when training neural networks

- Finding the optimal neural network layout is often time-consuming
- The model may be sensitive to changes in hyperparameters
- A model may take several hours or even days to train
    - This makes hyperparameter searches very expensive

Tips and tricks

- Write unit tests for your model [Roberts, 2017]
  - Check that each layer actually changes weights
  - Make sure that model converges on tiny data set

Tips and tricks

- Write unit tests for your model [Roberts, 2017]
  - Check that each layer actually changes weights
  - Make sure that model converges on tiny data set
- Stick to well-known architectures when starting out (e.g. LSTM/GRU for sequential data)

Tips and tricks

- Write unit tests for your model [Roberts, 2017]
  - Check that each layer actually changes weights
  - Make sure that model converges on tiny data set
- Stick to well-known architectures when starting out (e.g. LSTM/GRU for sequential data)
- Start by using small batch size
  - Usually makes model less sensitive to other hyperparameters

Tips and tricks

- Write unit tests for your model [Roberts, 2017]
  - Check that each layer actually changes weights
  - Make sure that model converges on tiny data set
- Stick to well-known architectures when starting out (e.g. LSTM/GRU for sequential data)
- Start by using small batch size
  - Usually makes model less sensitive to other hyperparameters
- Use normalization (batch, layer, group, weight...)
  - Speeds up convergence significantly
  - Start by trying batch normalization for CNN and feed-forward nets and layer normalization for RNN

The curious case of the batch size

- Training of neural nets can be sped up by increasing the batch size, since then the GPU/TPU can process more training examples in parallel
- Unfortunately increasing the batch size may result in a worse model. In extreme cases the model might not learn anything at all! [Masters et. al., 2018]
- The basic rule is to increase the learning rate linearly when increasing the batch size (e.g. double learning rate when doubling batch size)
  - Otherwise the magnitude of the weight updates decreases
- This means that increasing the batch size trades computational efficiency for stale gradients.

References I

- Nielsen, Michael A. *Neural Networks And Deep Learning*. Determination Press, 2015. http://neuralnetworksanddeeplearning.com/
- Hornik, Kurt. *Approximation Capabilities of Multilayer Feedforward Networks*. Neural Networks, 4(2), 251–257, 1991.
- Roberts, Chase. *How to unit test machine learning code. Medium.com 2017. https://medium.com/@keeper6928/how-to-unit-test-machine-learning-code-57cf6fd81765.*
- *Tai, Kai Sheng et al.* Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. ACL 2015. https://arxiv.org/abs/1503.00075

References II

Masters, Dominic, Luschi, Carlo. *Revisiting Small Batch Training for Deep Neural Networks. arXiv preprint 2018.* *https://arxiv.org/abs/1804.07612*