Paper reading

Deep learning: Review

Jie Gui

Computational Medicine and Bioinformatics,
University of Michigan

27 July 2018

Outline

1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

Introduction

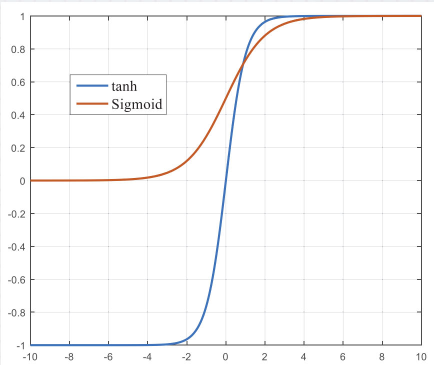
- Deep learning methods have dramatically improved the state-of-the-art in such as speech recognition, drug discovery and genomics.
- Deep convolutional nets and recurrent nets have brought about breakthroughs in many areas.

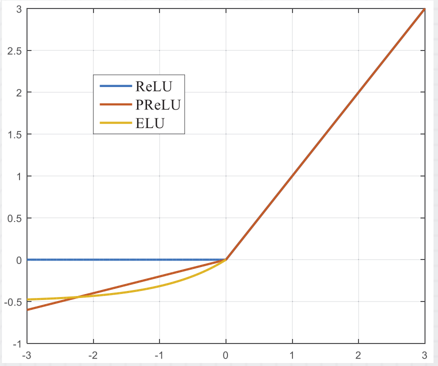
Outline

1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

The non-linear function

- Sigmoid: $S(x) = 1/(1 + e^{-x})$.
- tanh: $S(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$.
- ReLU: $S(x) = \max(0, x)$.
- PReLU: $S(x) = \max(\alpha x, x)$.

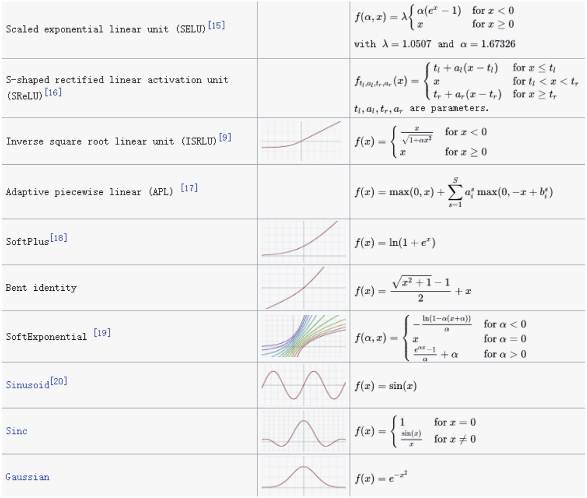




ReLU

- At present, the most popular non-linear function is the rectified linear unit (ReLU).
- In past decades, neural nets used smoother non-linearities, such as sigmoid and tanh.
- However, the ReLU typically learns much faster in networks with many layers, allowing training without unsupervised pre-training.

Other activation functions

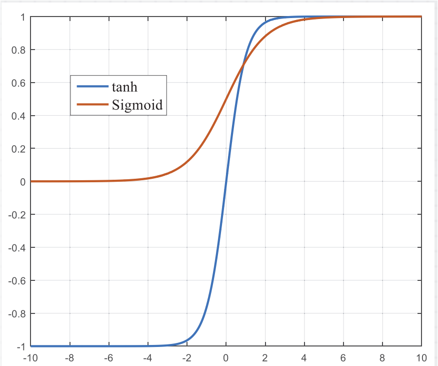


Scaled exponential linear unit (SELU)[15]		$f(\alpha, x) = \lambda \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$ with $\lambda = 1.0507$ and $\alpha = 1.67326$
S-shaped rectified linear activation unit (SReLU)[16]		$f_{t_l, a_l, t_r, a_r}(x) = \begin{cases} t_l + a_l(x - t_l) & \text{for } x \leq t_l \\ x & \text{for } t_l < x < t_r \\ t_r + a_r(x - t_r) & \text{for } x \geq t_r \end{cases}$ t_l, a_l, t_r, a_r are parameters.
Inverse square root linear unit (ISRLU)[9]		$f(x) = \begin{cases} \frac{x}{\sqrt{1+\alpha x^2}} & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$
Adaptive piecewise linear (APL) [17]		$f(x) = \max(0, x) + \sum_{s=1}^{S} a_i^s \max(0, -x + b_i^s)$
SoftPlus[18]		$f(x) = \ln(1 + e^x)$
Bent identity		$f(x) = \frac{\sqrt{x^2 + 1} - 1}{2} + x$
SoftExponential [19]		$f(\alpha, x) = \begin{cases} -\frac{\ln(1-\alpha(x+\alpha))}{\alpha} & \text{for } \alpha < 0 \\ x & \text{for } \alpha = 0 \\ \frac{e^{\alpha x} - 1}{\alpha} + \alpha & \text{for } \alpha > 0 \end{cases}$
Sinusoid[20]		$f(x) = \sin(x)$
Sinc		$f(x) = \begin{cases} 1 & \text{for } x = 0 \\ \frac{\sin(x)}{x} & \text{for } x \neq 0 \end{cases}$
Gaussian		$f(x) = e^{-x^2}$

Reference: `https://en.wikipedia.org/wiki/Activation_function`

Andrew NG's remark on activation functions

- An activation function that almost works better than the sigmoid function is the tanh function.
- Tanh function is a shifted version of the sigmoid function.



Andrew NG's remark on activation functions

- Tanh function almost works better than the sigmoid function because with values between plus one and minus one. The mean of the activations that come out of the hidden layer are closer to having a zero mean. Sometimes you train a neural network. You might center the data and have your data have zero mean. Using a tanh instead of a sigmoid function has the effect of centering your data so that the mean of the data is close to the zero rather than a 0.5 and this actually makes learning for the next layer a little bit easier.
- I pretty much never use the sigmoid activation function any more. The one exception is for the output layer. y is either zero or one. It makes sense for y to be a number that you want to output just between zero and one rather than -1 and 1.

History

- In the late 1990s, neural nets and backpropagation were largely forsaken by the machine-learning community and ignored by the computer-vision and speech-recognition communities.
- It was widely thought that learning useful, multistage, feature extractors with little prior knowledge was infeasible.
- In particular, it was commonly thought that simple gradient descent would get trapped in poor local minima —weight configurations for which no small change would reduce the average error.
- In practice, poor local minima are rarely a problem with large networks.

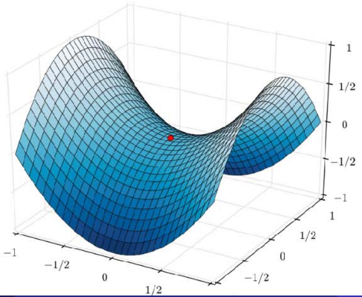
Saddle point

Saddle point is not equal to local minima.

Mathematical condition:

- The first-order derivative is zero.
- The Hessian matrix is indefinite.

Take $y = x_1^2 - x_2^2$ as an example,(0, 0) is a saddle point rather than local minima.



Outline

1. Introduction

2. Backpropagation to train multilayer architectures

3. Convolutional neural networks

4. Image understanding with deep convolutional networks

5. Distributed representations and language processing

6. Recurrent neural networks

7. The future of deep learning

Four key ideas behind ConvNets

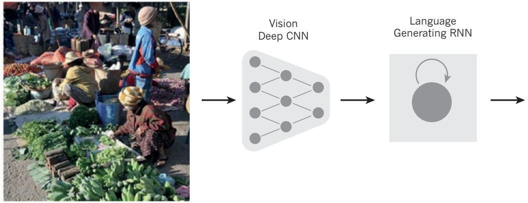
- local connections
- shared weights
- pooling
- the use of many layers

Outline

1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

- Despite these successes, ConvNets were largely forsaken by the mainstream computer-vision and machine-learning communities until the ImageNet competition in 2012.
- ConvNets are now the dominant approach for almost all recognition and detection tasks.
- A recent stunning demonstration combines ConvNets and recurrent net modules for the generation of image captions.

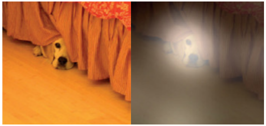
Image caption: from image to text



A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

A woman is throwing a **frisbee** in a park.



A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background

A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

- Recent ConvNet architectures have 10 to 20 layers of ReLUs, hundreds of millions of weights, and billions of connections between units. Whereas training such large networks could have taken weeks only two years ago, progress in hardware, software and algorithm parallelization have reduced training times to a few hours.
- The performance of ConvNet-based vision systems has caused most major technology companies, including Google, Facebook, Microsoft, etc, to initiate research and development projects and to deploy ConvNet-based image understanding products and services.
- ConvNets are easily amenable to efficient hardware implementations in chips or field-programmable gate arrays (FPGA). A number of companies such as NVIDIA, Mobileye, Intel, Qualcomm and Samsung are developing ConvNet chips to enable real-time vision applications in smartphones, cameras, robots and self-driving cars.

Outline

1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

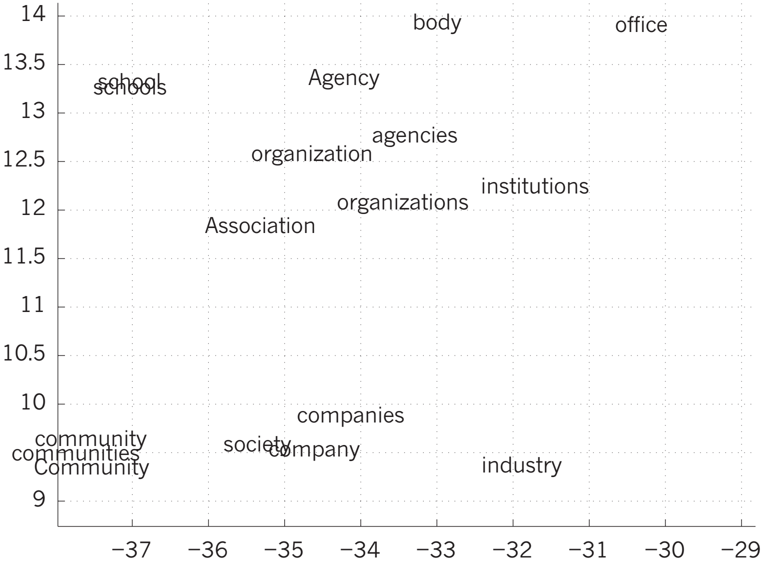
Two advantages

- First, learning distributed representations enable generalization to new combinations of the values of learned features beyond those seen during training (for example, 2^n combinations are possible with n binary features).
- Second, composing layers of representation in a deep net brings the potential for another exponential advantage (exponential in the depth).

Reasons of using NN to predict the next word in a sequence

- The hidden layers of a multilayer neural network learn to represent the network's inputs in a way that makes it easy to predict the target outputs.
- This is nicely demonstrated by training a multilayer neural network to predict the next word in a sequence from a local context of earlier words.

Visualizing the learned word vectors



What are distributed representations?

- When trained to predict the next word, the learned word vectors for Tuesday and Wednesday are very similar.
- Such representations are called **distributed representations** because their elements are not mutually exclusive.
- These word vectors are composed of learned features that were automatically discovered by the neural network.
- Vector representations of words learned from text are now very widely used in natural language applications.

Debate

Logic-inspired paradigm:

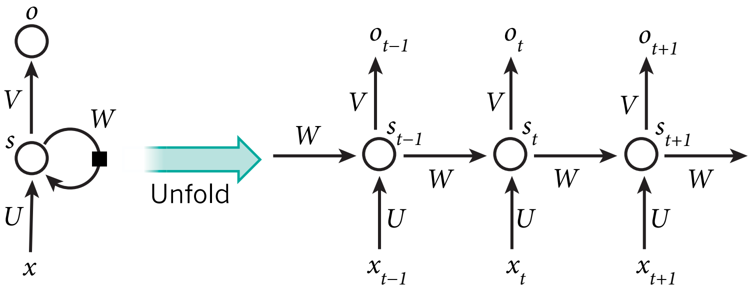
- An instance of a symbol is something for which the only property is that it is identical/non-identical to others.
- It has no internal structure that is relevant to its use.
- To reason with symbols, they must be bound to the variables in judiciously chosen rules of inference.

Neural-network-inspired paradigm just uses big activity vectors to perform the type of fast ‘intuitive’ inference.

Outline

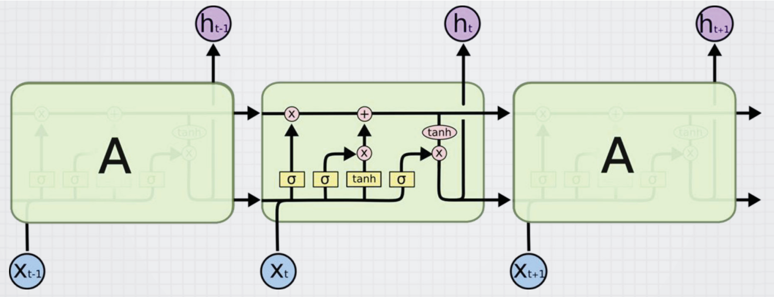
1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

A recurrent neural network



Long short-term memory (LSTM) networks

- Forget gate
- Input gate
- Output gate



Outline

1. Introduction
2. Backpropagation to train multilayer architectures
3. Convolutional neural networks
4. Image understanding with deep convolutional networks
5. Distributed representations and language processing
6. Recurrent neural networks
7. The future of deep learning

The future of deep learning

- Unsupervised learning had a catalytic effect in reviving interest in deep learning.
- Although we have not focused on it in this Review, we expect unsupervised learning to become far more important.
- Human and animal learning is largely unsupervised.
- We discover the structure of the world by observing it, not by being told the name of every object
- Which systems do we expect much of the future progress in vision to come from?

The future of deep learning

- Natural language understanding is another area in which deep learning is poised to make a large impact.
- Ultimately, major progress in AI will come about through systems that combine representation learning with complex reasoning.

Thank You!