M2177.0043 Introduction to Deep Learning

Lecture 7: Model fitting / Computation graph and Backpropagation[1]

Hyun Oh Song[1]

[1]Dept. of Computer Science and Engineering, Seoul National University

April 7, 2020

[1]Many slides and figures adapted Justin Johnson

Last time

- Score functions
- Loss functions
- Regularization

Outline

Bias-Variance tradeoff[2]

- Supppose we have a training set of data and labels $D = \{(x_i, y_i)\}_{i=1}^n$
- Assume there is a function with noise $y = f(x) + \epsilon$, where the ϵ has zero mean and variance σ^2
- Want to find a function $\widehat{f}(x; D)$ that approximates the true function $f(x)$ as well as possible
- Concretely, we want MSE $(y - \widehat{f}(x; D))^2$ to be minimal for the training data $x_1, \ldots, x_n$ and for unseen sample x outside the training set.

[2] https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

- The expected error on an unseen sample x is,

$$\begin{aligned}
\mathbb{E}_D[(y - \widehat{f}(x;D))^2] &= \mathbb{E}_D[(f(x) + \epsilon - \widehat{f}(x;D))^2] \\
&= \mathbb{E}_D[(f(x) + \epsilon - \widehat{f}(x;D) + \mathbb{E}_{D'}[\widehat{f}(x;D')] - \mathbb{E}_{D'}[\widehat{f}(x;D')])^2] \\
&= \mathbb{E}_D[(f(x) - \mathbb{E}_{D'}[\widehat{f}(x;D')])^2] + \mathbb{E}_D[\epsilon^2] + \mathbb{E}_D[(\mathbb{E}_{D'}[\widehat{f}(x;D')] - \widehat{f}(x;D))^2] \\
&\quad + 2\mathbb{E}_D[(f(x) - \mathbb{E}_{D'}[\widehat{f}(x;D')])\epsilon] + 2\mathbb{E}_D[\epsilon(\mathbb{E}_{D'}[\widehat{f}(x;D')] - \widehat{f}(x;D))] \\
&\quad + 2\mathbb{E}_D[(\mathbb{E}_{D'}[\widehat{f}(x;D')] - \widehat{f}(x;D))(f(x) - \mathbb{E}_{D'}[\widehat{f}(x;D')])] \\
&= (f(x) - \mathbb{E}_{D'}[\widehat{f}(x;D')])^2 + \mathrm{Var}[\epsilon] + \mathrm{Var}[\widehat{f}(x;D)] \\
&= (\mathrm{Bias}[\widehat{f}(x;D)])^2 + \sigma^2 + \mathrm{Var}[\widehat{f}(x;D)]
\end{aligned}$$

- Finally, MSE loss is obtained by taking the expectation over x

$$\mathrm{MSE} = \mathbb{E}_x[(\mathrm{Bias}[\widehat{f}(x;D)])^2 + \mathrm{Var}[\widehat{f}(x;D)]] + \sigma^2$$

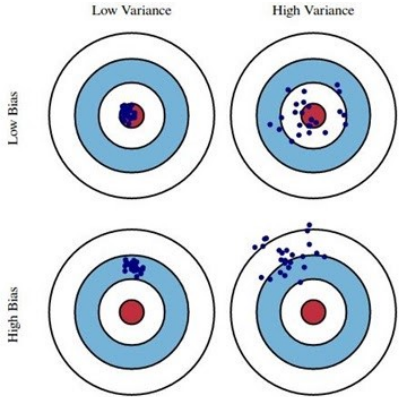
- Summarizing, the expected error on an unseen sample x,

$$\begin{aligned}\mathbb{E}_D[(y - \widehat{f}(x; D))^2] &= (f(x) - \mathbb{E}_{D'}[\widehat{f}(x; D')])^2 + \mathrm{Var}[\epsilon] + \mathrm{Var}[\widehat{f}(x; D)] \\ &= (\mathrm{Bias}[\widehat{f}(x; D)])^2 + \sigma^2 + \mathrm{Var}[\widehat{f}(x; D)]\end{aligned}$$

- The error can be decomposed as the sum of the following terms
  - The square of the *bias*: error caused by the simplifying assumptions built into the method
  - *variance*: how much the learning method $\widehat{f}(x; D)$ will move around its mean
  - *irreducible error*: σ^2

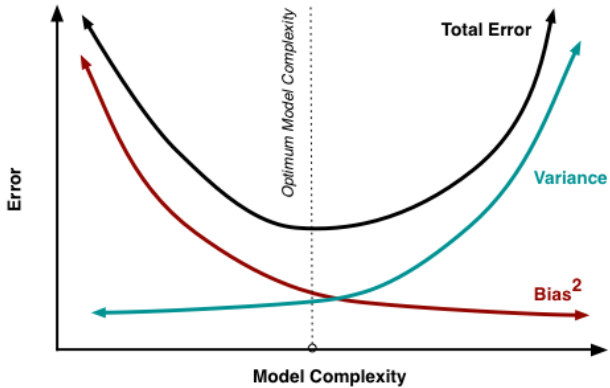
- Bias-variance decomposition also applies to classification and other ML tasks. Look at the wiki page for more details.

Graphical illustration[3]



[3] http://scott.fortmann-roe.com/docs/BiasVariance.html

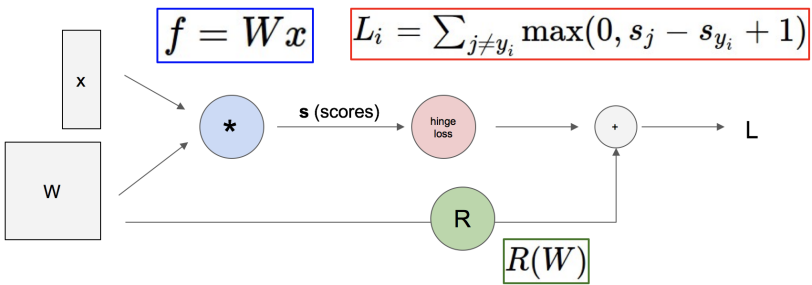
Graphical illustration[4]



[4] http://scott.fortmann-roe.com/docs/BiasVariance.html

Outline

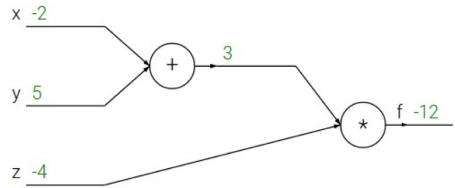
Computational graph



Backpropagation example

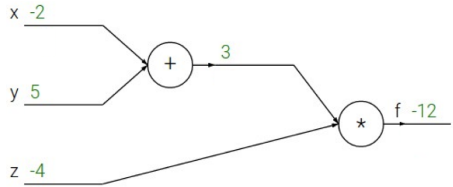
$f(x, y, z) = (x + y)z$

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



Backpropagation example

$$f(x, y, z) = (x + y)z$$
$$e.g.\ x = -2, y = 5, z = -4$$

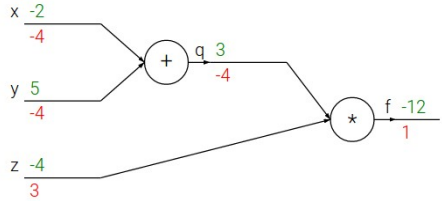


$$q = x + y, \quad \frac{\partial q}{\partial x} = 1, \ \frac{\partial q}{\partial y} = 1$$
$$f = qz, \quad \frac{\partial f}{\partial q} = z, \ \frac{\partial f}{\partial z} = q$$

- Want to compute $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation example

$$f(x, y, z) = (x + y)z$$
$$e.g.\ x = -2, y = 5, z = -4$$

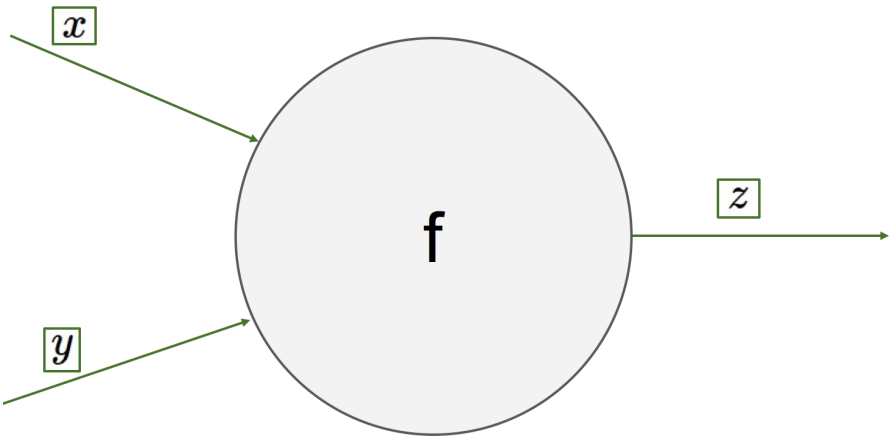


$$q = x + y, \quad \frac{\partial q}{\partial x} = 1, \ \frac{\partial q}{\partial y} = 1$$
$$f = qz, \quad \frac{\partial f}{\partial q} = z, \ \frac{\partial f}{\partial z} = q$$

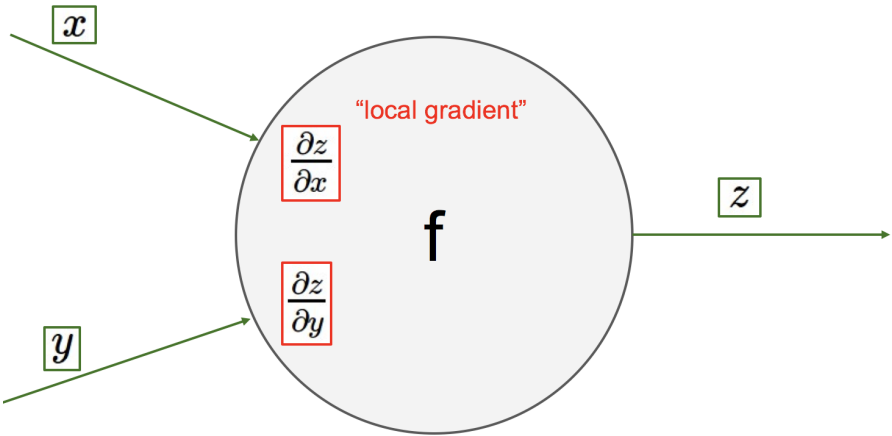
- Want to compute $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

$$\frac{\partial f}{\partial f} = 1, \frac{\partial f}{\partial z} = 3, \frac{\partial f}{\partial q} = -4$$
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y} = -4(1) = -4$$
$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial f}{\partial q}}_{\text{Upstream gradient}} \underbrace{\frac{\partial q}{\partial x}}_{\text{Downstream gradient}} = -4(1) = -4$$

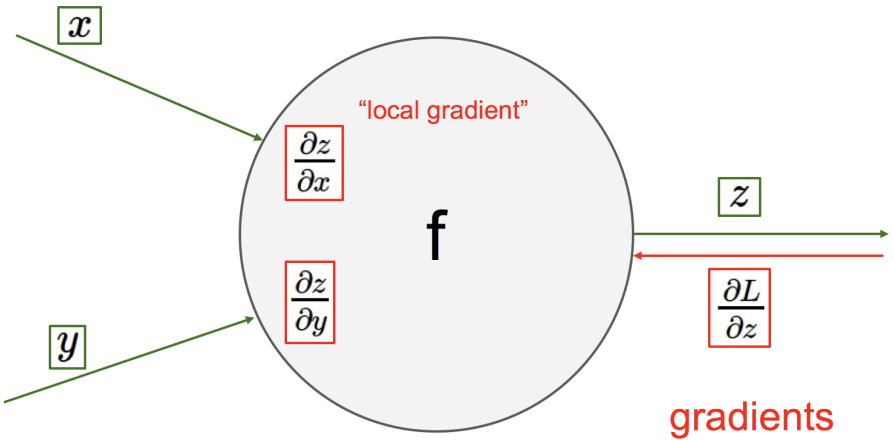
Nodes in computational graph



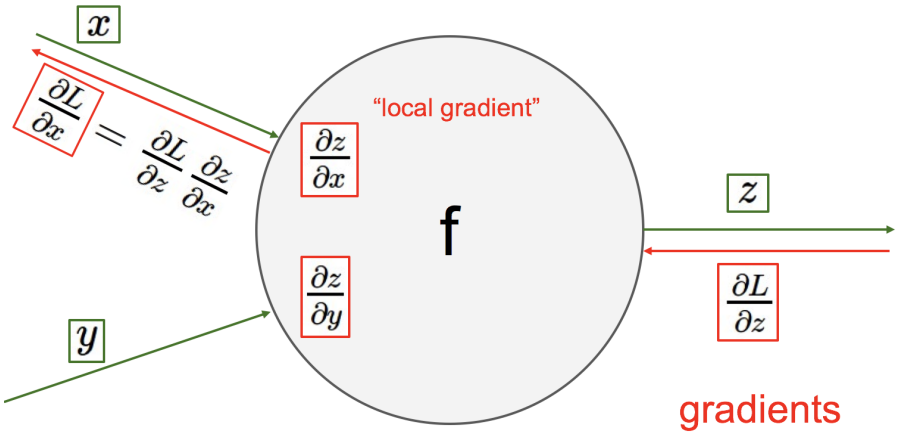
Nodes in computational graph



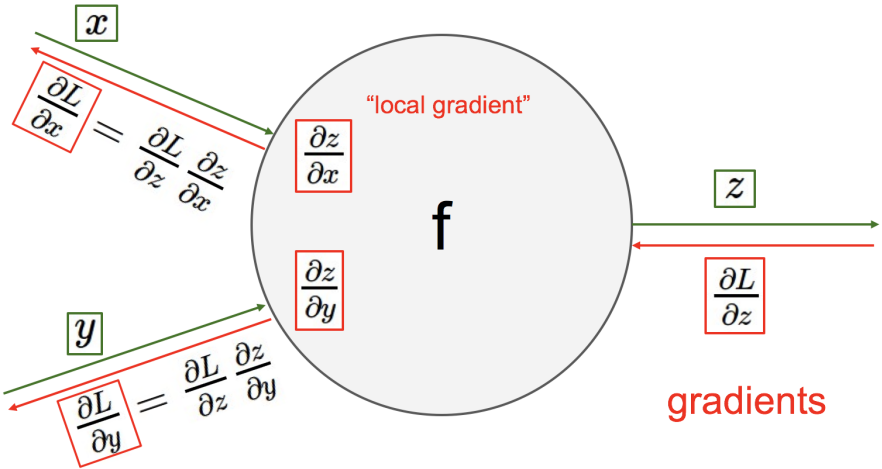
Nodes in computational graph



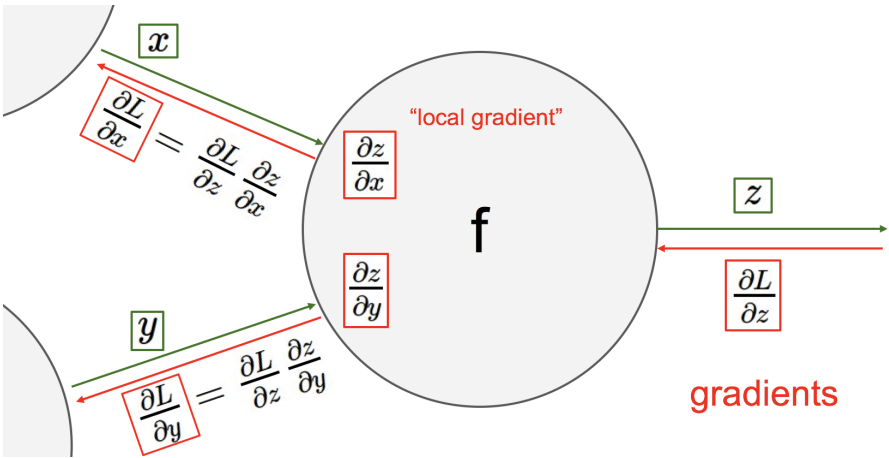
Nodes in computational graph



Nodes in computational graph

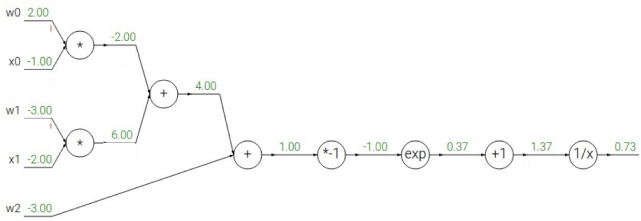


Nodes in computational graph



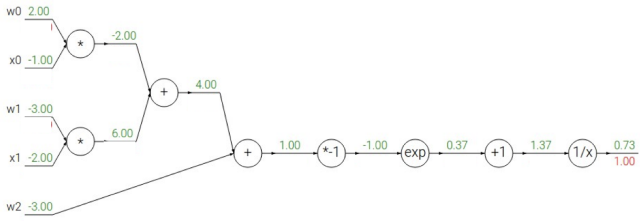
Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

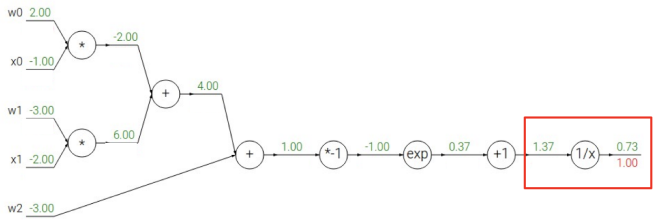
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

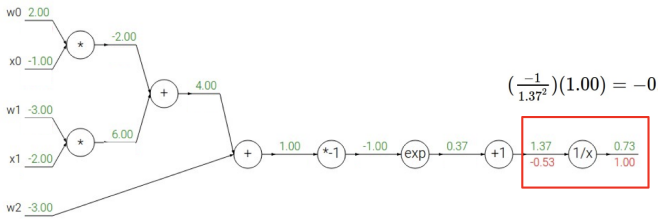
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$(\frac{-1}{1.37^2})(1.00) = -0.53$

$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$

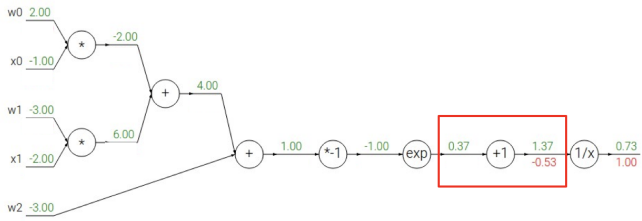
$f_a(x) = ax \rightarrow \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \dfrac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \dfrac{df}{dx} = e^x$

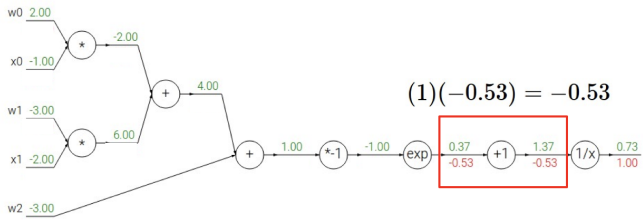
$f_a(x) = ax \quad \rightarrow \quad \dfrac{df}{dx} = a$

$f(x) = \dfrac{1}{x} \quad \rightarrow \quad \dfrac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \dfrac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \dfrac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$(1)(-0.53) = -0.53$

$f(x) = e^x \rightarrow \dfrac{df}{dx} = e^x$

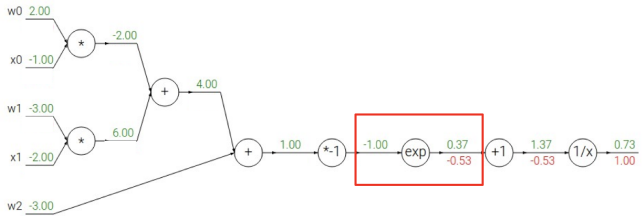
$f_a(x) = ax \rightarrow \dfrac{df}{dx} = a$

$f(x) = \dfrac{1}{x} \rightarrow \dfrac{df}{dx} = -1/x^2$

$f_c(x) = c + x \rightarrow \dfrac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

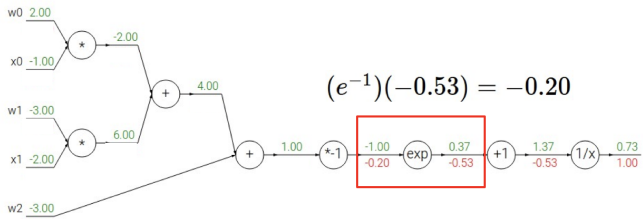
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

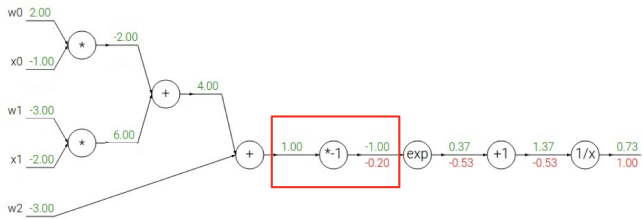
Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x$	$\rightarrow$	$\frac{df}{dx} = e^x$	$f(x) = \frac{1}{x}$	$\rightarrow$	$\frac{df}{dx} = -1/x^2$
$f_a(x) = ax$	$\rightarrow$	$\frac{df}{dx} = a$	$f_c(x) = c + x$	$\rightarrow$	$\frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \dfrac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \dfrac{df}{dx} = e^x$

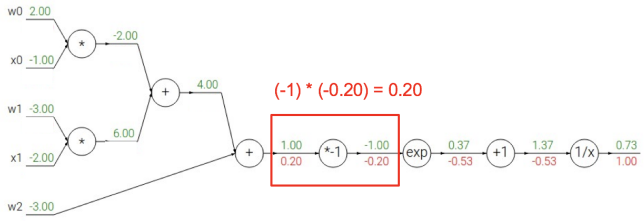
$f_a(x) = ax \quad \rightarrow \quad \dfrac{df}{dx} = a$

$f(x) = \dfrac{1}{x} \quad \rightarrow \quad \dfrac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \dfrac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

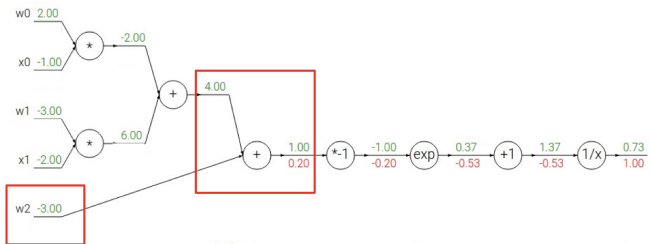
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \dfrac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \dfrac{df}{dx} = e^x$

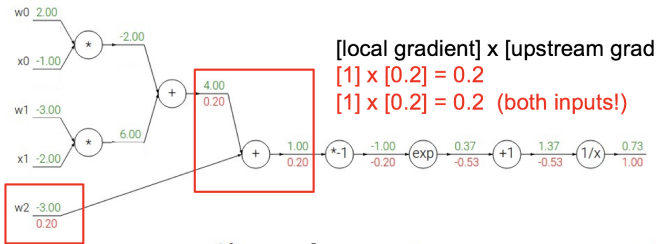
$f_a(x) = ax \quad \rightarrow \quad \dfrac{df}{dx} = a$

$f(x) = \dfrac{1}{x} \quad \rightarrow \quad \dfrac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \dfrac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



[local gradient] x [upstream gradient]

[1] x [0.2] = 0.2

[1] x [0.2] = 0.2 (both inputs!)

$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

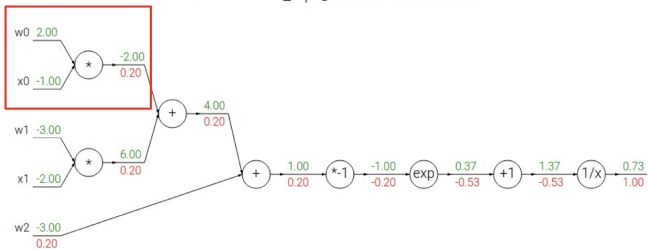
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$

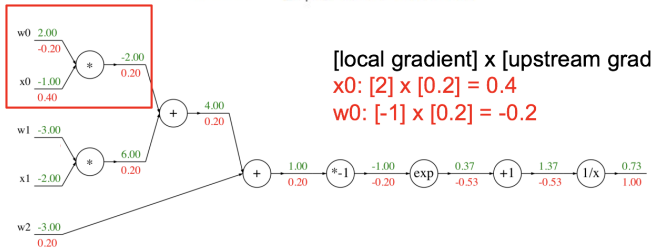
$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$

$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$

Sigmoid example

Another example: $f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$



[local gradient] x [upstream gradient]

x0: [2] x [0.2] = 0.4

w0: [-1] x [0.2] = -0.2

$f(x) = e^x \rightarrow \frac{df}{dx} = e^x$

$f_a(x) = ax \rightarrow \frac{df}{dx} = a$

$f(x) = \frac{1}{x} \rightarrow \frac{df}{dx} = -1/x^2$

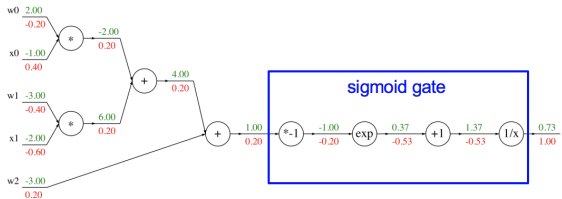
$f_c(x) = c + x \rightarrow \frac{df}{dx} = 1$

Sigmoid example

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$ sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$

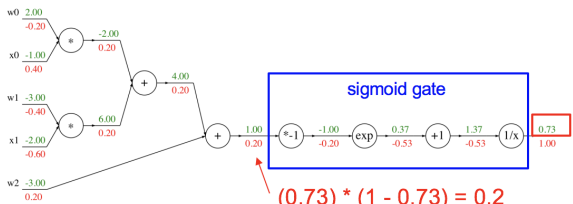


Sigmoid example

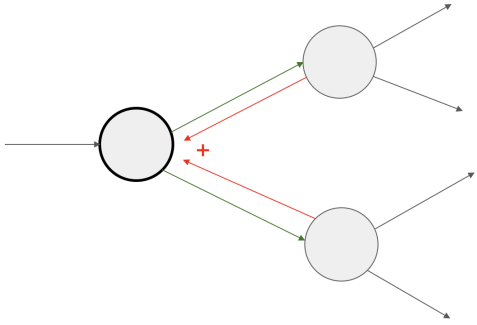
$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

$$\boxed{\sigma(x) = \frac{1}{1+e^{-x}}}$$ sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$



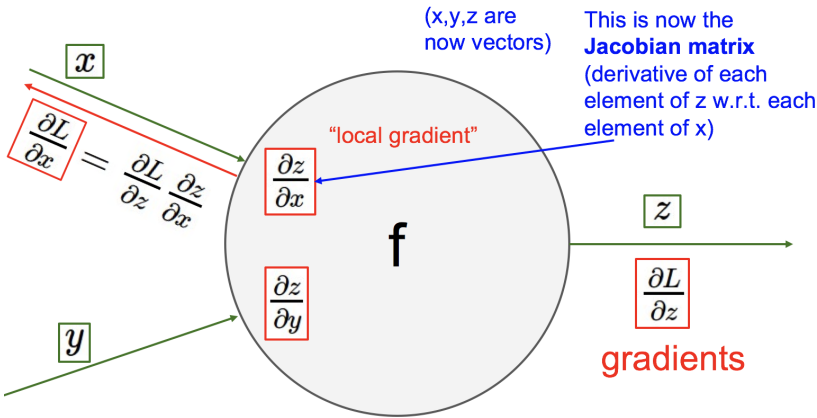
Gradients add at branches



Chain rule

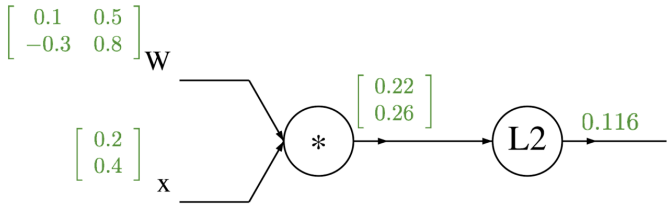
$$\frac{\partial f}{\partial x} = \sum_i \frac{\partial f}{\partial q_i} \frac{\partial q_i}{\partial x}$$

Gradients for vectorized code



Vectorized example

$$f(x, W) = \|Wx\|^2 = \sum_{i=1}^{n} (Wx)_i^2$$

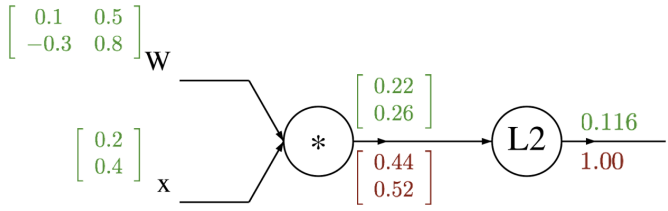


$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = \|q\|^2 = q_1^2 + \cdots + q_n^2$$

Vectorized example

$$f(x, W) = \|Wx\|^2 = \sum_{i=1}^{n} (Wx)_i^2$$



$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

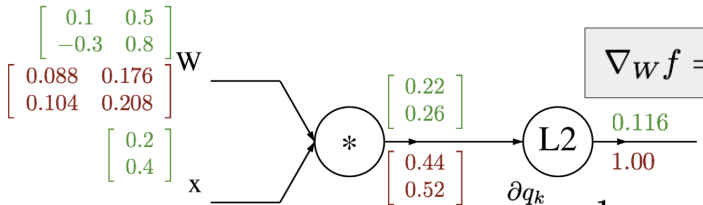
$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\boxed{\nabla_q f = 2q}$$

Vectorized example

$$f(x, W) = \|Wx\|^2 = \sum_{i=1}^{n} (Wx)_i^2$$



$$\nabla_W f = 2q \cdot x^T$$

$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

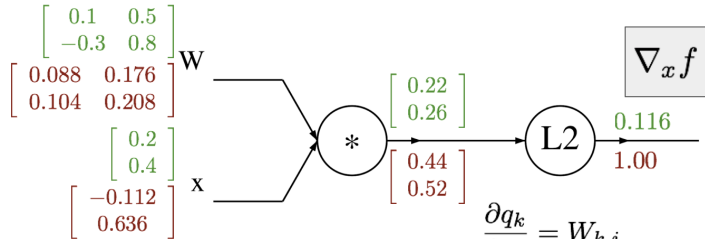
$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial q_k}{\partial W_{i,j}} = \mathbf{1}_{k=i} x_j$$

$$\begin{aligned} \frac{\partial f}{\partial W_{i,j}} &= \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial W_{i,j}} \\ &= \sum_k (2q_k)(\mathbf{1}_{k=i} x_j) \\ &= 2q_i x_j \end{aligned}$$

Vectorized example

$$f(x, W) = \|Wx\|^2 = \sum_{i=1}^{n} (Wx)_i^2$$



$$\nabla_x f = 2W^T \cdot q$$

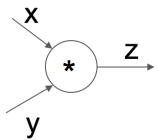
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial q_k}{\partial x_i} = W_{k,i}$$

$$\frac{\partial f}{\partial x_i} = \sum_k \frac{\partial f}{\partial q_k} \frac{\partial q_k}{\partial x_i} = \sum_k 2q_k W_{k,i}$$

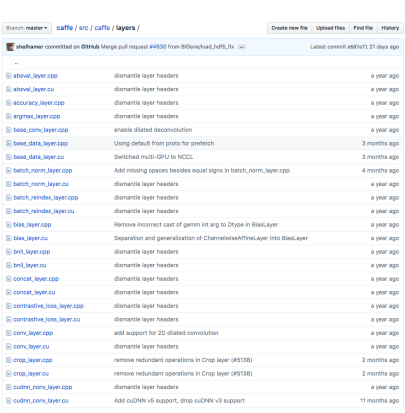
Modularized implementation forward/backward API



(x,y,z are scalars)

```python
class MultiplyGate(object):
    def forward(x,y):
        z = x*y
        self.x = x # must keep these around!
        self.y = y
        return z
    def backward(dz):
        dx = self.y * dz # [dz/dx * dL/dz]
        dy = self.x * dz # [dz/dy * dL/dz]
        return [dx, dy]
```

Example: Caffe layers



Branch: master ▾ caffe / src / caffe / layers / Create new file Upload files Find file History

shelhamer committed on GitHub Merge pull request #4830 from BIDeng/load_hdf5_fix … Latest commit e687a71 21 days ago

..

absval_layer.cpp	dismantle layer headers	a year ago
absval_layer.cu	dismantle layer headers	a year ago
accuracy_layer.cpp	dismantle layer headers	a year ago
argmax_layer.cpp	dismantle layer headers	a year ago
base_conv_layer.cpp	enable dilated deconvolution	a year ago
base_data_layer.cpp	Using default from proto for prefetch	3 months ago
base_data_layer.cu	Switched multi-GPU to NCCL	3 months ago
batch_norm_layer.cpp	Add missing spaces besides equal signs in batch_norm_layer.cpp	4 months ago
batch_norm_layer.cu	dismantle layer headers	a year ago
batch_reindex_layer.cpp	dismantle layer headers	a year ago
batch_reindex_layer.cu	dismantle layer headers	a year ago
bias_layer.cpp	Remove incorrect cast of gemm int arg to Dtype in BiasLayer	a year ago
bias_layer.cu	Separation and generalization of ChannelwiseAffineLayer into BiasLayer	a year ago
bnll_layer.cpp	dismantle layer headers	a year ago
bnll_layer.cu	dismantle layer headers	a year ago
concat_layer.cpp	dismantle layer headers	a year ago
concat_layer.cu	dismantle layer headers	a year ago
contrastive_loss_layer.cpp	dismantle layer headers	a year ago
contrastive_loss_layer.cu	dismantle layer headers	a year ago
conv_layer.cpp	add support for 2D dilated convolution	a year ago
conv_layer.cu	dismantle layer headers	a year ago
crop_layer.cpp	remove redundant operations in Crop layer (#5138)	2 months ago
crop_layer.cu	remove redundant operations in Crop layer (#5138)	2 months ago
cudnn_conv_layer.cpp	dismantle layer headers	a year ago
cudnn_conv_layer.cu	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_lcn_layer.cpp	dismantle layer headers	a year ago
cudnn_lcn_layer.cu	dismantle layer headers	a year ago
cudnn_lrn_layer.cpp	dismantle layer headers	a year ago
cudnn_lrn_layer.cu	dismantle layer headers	a year ago
cudnn_pooling_layer.cpp	dismantle layer headers	a year ago
cudnn_pooling_layer.cu	dismantle layer headers	a year ago
cudnn_relu_layer.cpp	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_relu_layer.cu	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_sigmoid_layer.cpp	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_sigmoid_layer.cu	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_softmax_layer.cpp	dismantle layer headers	a year ago
cudnn_softmax_layer.cu	dismantle layer headers	a year ago
cudnn_tanh_layer.cpp	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
cudnn_tanh_layer.cu	Add cuDNN v5 support, drop cuDNN v3 support	11 months ago
data_layer.cpp	Switched multi-GPU to NCCL	3 months ago
deconv_layer.cpp	enable dilated deconvolution	a year ago
deconv_layer.cu	dismantle layer headers	a year ago
dropout_layer.cpp	supporting N-D Blobs in Dropout layer Reshape	a year ago
dropout_layer.cu	dismantle layer headers	a year ago
dummy_data_layer.cpp	dismantle layer headers	a year ago
eltwise_layer.cpp	dismantle layer headers	a year ago
eltwise_layer.cu	dismantle layer headers	a year ago
elu_layer.cpp	ELU layer with basic tests	a year ago
elu_layer.cu	ELU layer with basic tests	a year ago
embed_layer.cpp	dismantle layer headers	a year ago
embed_layer.cu	dismantle layer headers	a year ago
euclidean_loss_layer.cpp	dismantle layer headers	a year ago
euclidean_loss_layer.cu	dismantle layer headers	a year ago
exp_layer.cpp	Solving issue with exp layer with base e	a year ago
exp_layer.cu	dismantle layer headers	a year ago

Caffe Sigmoid layers

```cpp
#include <cmath>
#include <vector>

#include "caffe/layers/sigmoid_layer.hpp"

namespace caffe {

template <typename Dtype>
inline Dtype sigmoid(Dtype x) {
 return 1. / (1. + exp(-x));
}

template <typename Dtype>
void SigmoidLayer<Dtype>::Forward_cpu(const vector<Blob<Dtype>*>& bottom,
    const vector<Blob<Dtype>*>& top) {
  const Dtype* bottom_data = bottom[0]->cpu_data();
  Dtype* top_data = top[0]->mutable_cpu_data();
  const int count = bottom[0]->count();
  for (int i = 0; i < count; ++i) {
    top_data[i] = sigmoid(bottom_data[i]);
  }
}

template <typename Dtype>
void SigmoidLayer<Dtype>::Backward_cpu(const vector<Blob<Dtype>*>& top,
    const vector<bool>& propagate_down,
    const vector<Blob<Dtype>*>& bottom) {
  if (propagate_down[0]) {
    const Dtype* top_data = top[0]->cpu_data();
    const Dtype* top_diff = top[0]->cpu_diff();
    Dtype* bottom_diff = bottom[0]->mutable_cpu_diff();
    const int count = bottom[0]->count();
    for (int i = 0; i < count; ++i) {
      const Dtype sigmoid_x = top_data[i];
      bottom_diff[i] = top_diff[i] * sigmoid_x * (1. - sigmoid_x);
    }
  }
}

#ifdef CPU_ONLY
STUB_GPU(SigmoidLayer);
#endif

INSTANTIATE_CLASS(SigmoidLayer);


}  // namespace caffe
```

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

$$(1-\sigma(x))\,\sigma(x)$$ * top_diff (chain rule)

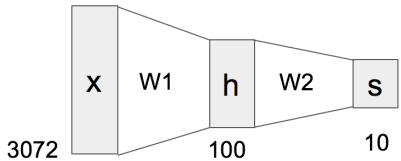
Dynamic programming interpretation

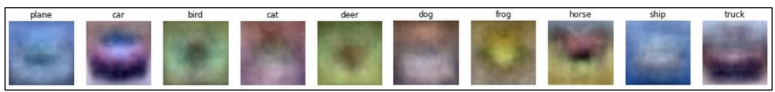
Outline

Neural networks

(**Before**) Linear score function: $f = Wx$

(**Now**) 2-layer Neural Network $f = W_2 \max(0, W_1 x)$





- Linear combination of intermediate templates

Multi-layer perceptron (MLP)

(**Before**) Linear score function: $f = Wx$

(**Now**) 2-layer Neural Network $f = W_2 \max(0, W_1 x)$

or 3-layer Neural Network

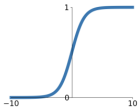
$$f = W_3 \max(0, W_2 \max(0, W_1 x))$$

- Can go many layers deep

Non-linearities

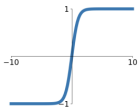
Sigmoid

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



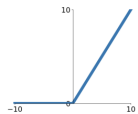
tanh

$\tanh(x)$



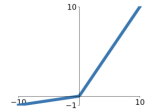
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

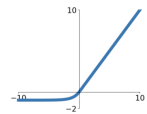


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



- What happens to MLP without non-linearities?