Deep Learning

Vazgen Mikayelyan

October 27, 2020

FAST DISCOVERING THE FUTURE

Outline

Stochastic Gradient Descent

Let L be a loss function that we know:

Stochastic Gradient Descent

Let L be a loss function that we know:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(f_w(x_i) - y_i\right)^2,$$

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_I \log f_w(x_i) - (1-y_i)\log\left(1 - f_w(x_i)\right)\right),$$

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i^T \log f_w(x_i)\right).$$

Stochastic Gradient Descent

Let L be a loss function that we know:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(f_w(x_i) - y_i\right)^2,$$

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_I \log f_w(x_i) - (1-y_i)\log\left(1 - f_w(x_i)\right)\right),$$

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i^T \log f_w(x_i)\right).$$

Do you see problems in finding minimum of these functions using GD?

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

SGD algorithm is the following:

- Choose an initial vector of parameters w and learning rate α.

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

SGD algorithm is the following:

- Choose an initial vector of parameters w and learning rate α.
- Repeat until an approximate minimum is obtained.

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

SGD algorithm is the following:

- Choose an initial vector of parameters w and learning rate α.
- Repeat until an approximate minimum is obtained.
  - Randomly shuffle examples in the training set.

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

SGD algorithm is the following:

- Choose an initial vector of parameters w and learning rate α.
- Repeat until an approximate minimum is obtained.
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, ..., n$, do $w \leftarrow w - \alpha \nabla L_i(w)$.

Stochastic Gradient Descent

Note that in each case we can represent the loss function by the following form:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} L_i(w).$$

SGD algorithm is the following:

- Choose an initial vector of parameters w and learning rate α.
- Repeat until an approximate minimum is obtained.
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, ..., n$, do $w \leftarrow w - \alpha \nabla L_i(w)$.

Do you see problems in this optimization method?

Mini-Batch Gradient Descent

MBGD algorithm is the following:

- Choose an initial vector of parameters w, learning rate α and batch size B.

Mini-Batch Gradient Descent

MBGD algorithm is the following:

- Choose an initial vector of parameters w, learning rate α and batch size B.
- Repeat until an approximate minimum is obtained.

Mini-Batch Gradient Descent

MBGD algorithm is the following:

- Choose an initial vector of parameters w, learning rate α and batch size B.
- Repeat until an approximate minimum is obtained.
  - Randomly shuffle examples in the training set.

Mini-Batch Gradient Descent

MBGD algorithm is the following:

- Choose an initial vector of parameters w, learning rate α and batch size B.
- Repeat until an approximate minimum is obtained.
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, ..., \lceil \frac{n}{B} \rceil$, do

$$w \leftarrow w - \alpha \nabla \frac{1}{B} \sum_{k=(i-1)\cdot B+1}^{i \cdot B} L_k(w).$$

Outline

Back-Propagation

Question: How to calculate the derivative of the function $\sin x^2$?

Back-Propagation

Question: How to calculate the derivative of the function $\sin x^2$?

Theorem 1

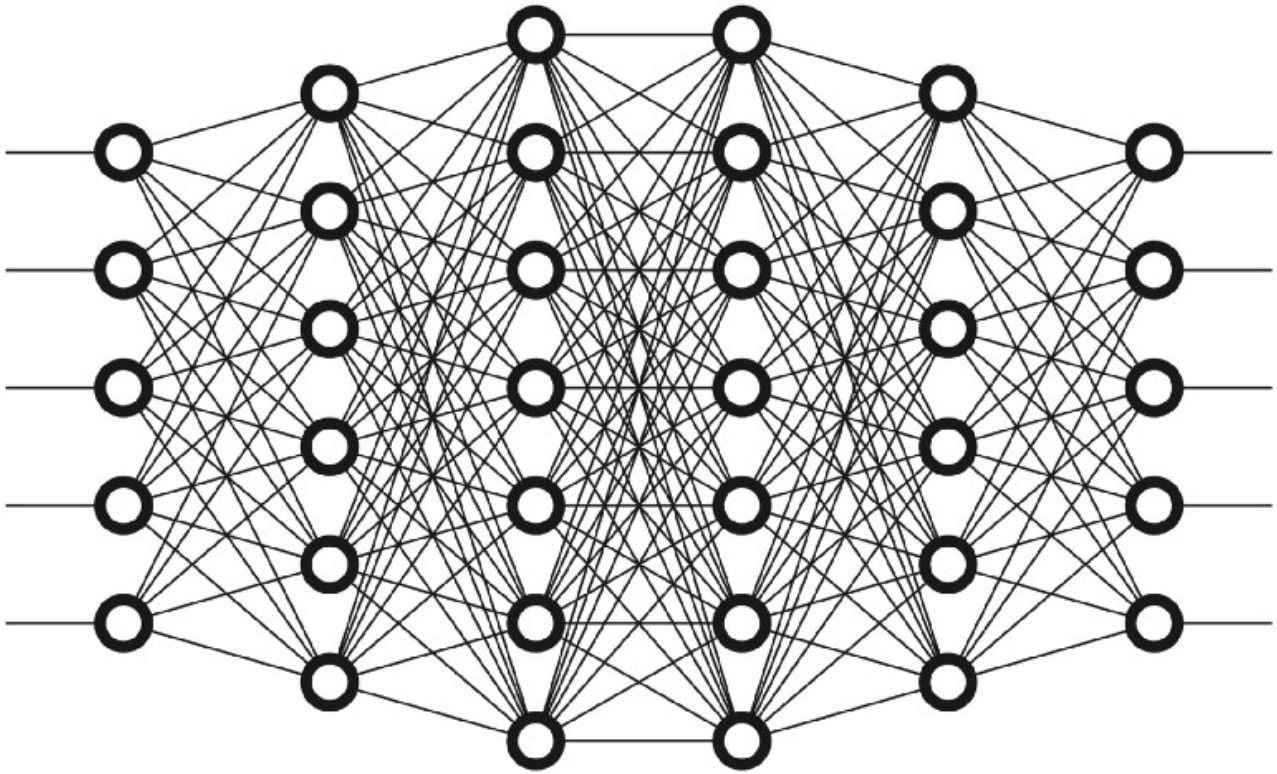
Given n functions $f_1, \ldots, f_n$ with the composite function

$$f = f_1 \circ (f_2 \circ \cdots (f_{n-1} \circ f_n)),$$

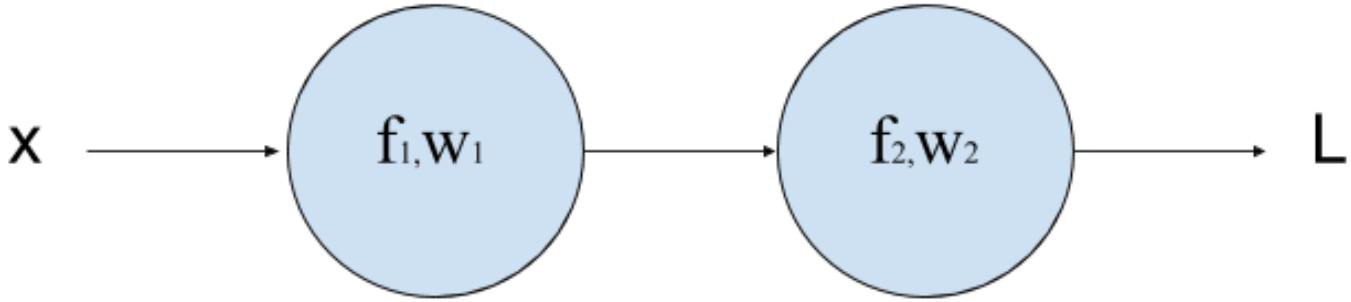
if each function f_i is differentiable at its immediate input, then the composite function is also differentiable by the repeated application of Chain Rule, where the derivative is

$$\frac{df}{dx} = \frac{df_1}{df_2}\frac{df_2}{df_3} \cdots \frac{df_n}{dx}.$$

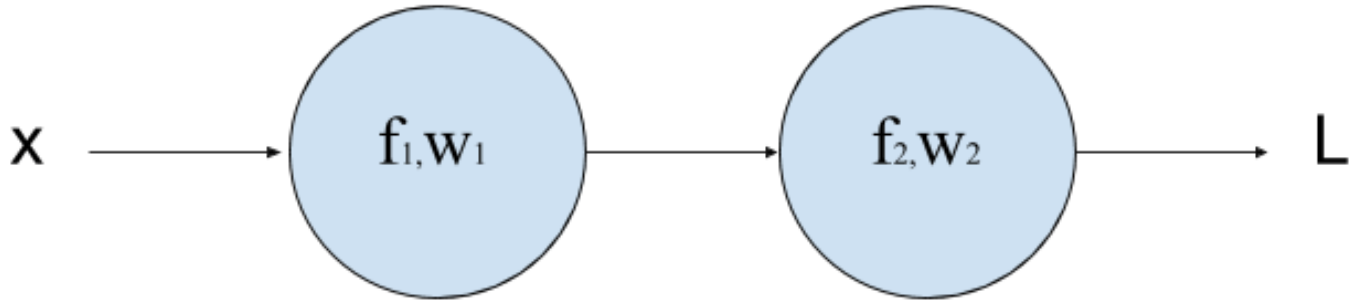
Back-Propagation



Back-Propagation



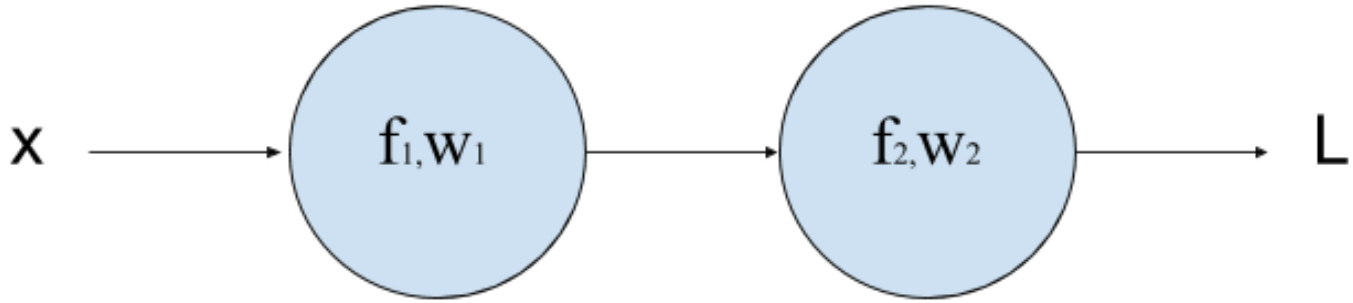
Back-Propagation



In this case we have the following function

$$L(w_1, w_2) = (f_2(w_2 f_1(w_1 x)) - y)^2$$

Back-Propagation

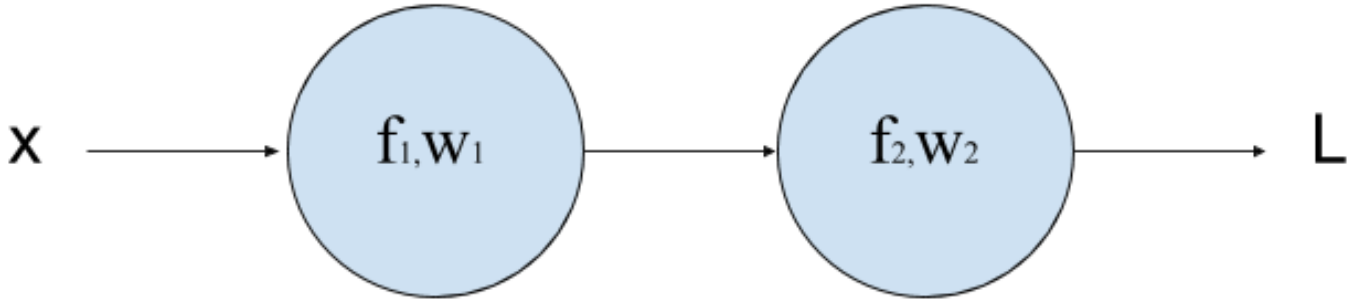


In this case we have the following function

$$L(w_1, w_2) = (f_2(w_2 f_1(w_1 x)) - y)^2$$

We have to calculate the derivatives $\frac{\partial L}{\partial w_1}$ and $\frac{\partial L}{\partial w_2}$:

Back-Propagation



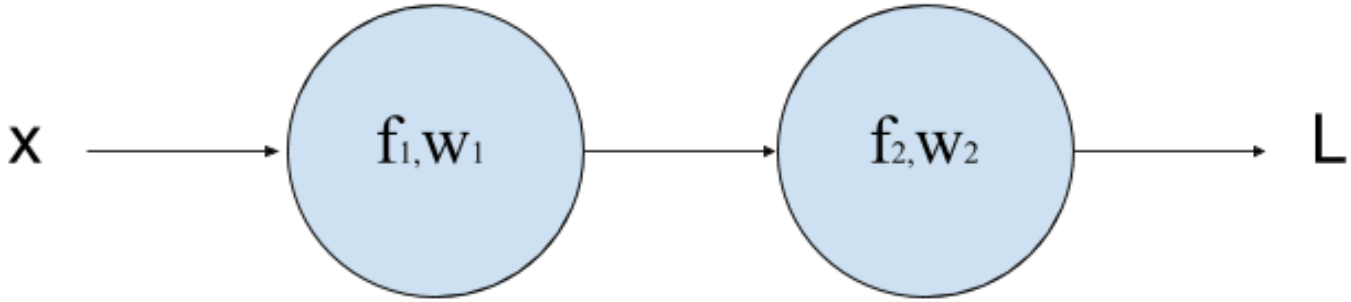
In this case we have the following function

$$L(w_1, w_2) = (f_2(w_2 f_1(w_1 x)) - y)^2$$

We have to calculate the derivatives $\frac{\partial L}{\partial w_1}$ and $\frac{\partial L}{\partial w_2}$:

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f_2} \frac{\partial f_2}{\partial (w_2 f_1)} \frac{\partial (w_2 f_1)}{\partial w_2},$$

Back-Propagation



In this case we have the following function

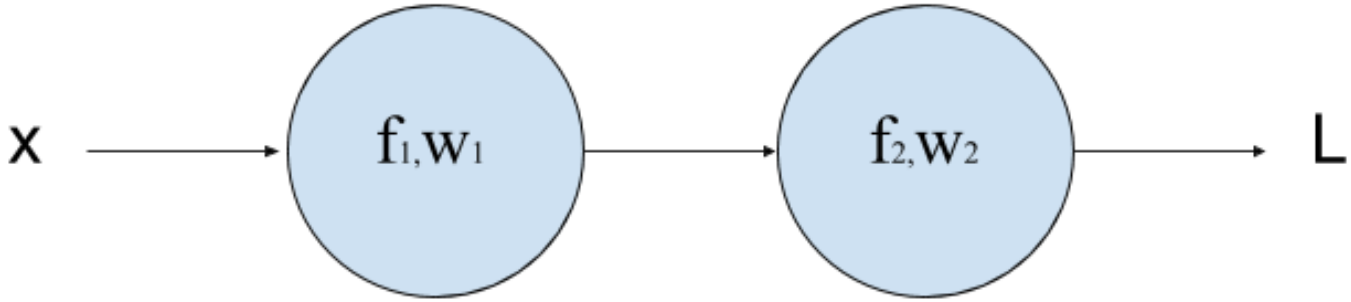
$$L(w_1, w_2) = (f_2(w_2 f_1(w_1 x)) - y)^2$$

We have to calculate the derivatives $\frac{\partial L}{\partial w_1}$ and $\frac{\partial L}{\partial w_2}$:

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f_2} \frac{\partial f_2}{\partial (w_2 f_1)} \frac{\partial (w_2 f_1)}{\partial w_2},$$

$$\frac{\partial L}{\partial w_1} =$$

Back-Propagation



In this case we have the following function

$$L(w_1, w_2) = (f_2(w_2 f_1(w_1 x)) - y)^2$$

We have to calculate the derivatives $\frac{\partial L}{\partial w_1}$ and $\frac{\partial L}{\partial w_2}$:

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f_2} \frac{\partial f_2}{\partial (w_2 f_1)} \frac{\partial (w_2 f_1)}{\partial w_2},$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial f_2} \frac{\partial f_2}{\partial (w_2 f_1)} \frac{\partial (w_2 f_1)}{\partial f_1} \frac{\partial (f_1)}{\partial (w_1 x)} \frac{\partial (w_1 x)}{\partial w_1}.$$