Training Neural Networks II

Milan Straka

March 18, 2019
Let us have a dataset with a training, validation and test sets, each containing examples \((\mathbf{x}, y)\).

Depending on \(y\), consider one of the following output activation functions:

\[
\begin{align*}
\text{none} & \quad \text{if } y \in \mathbb{R} \\
\sigma & \quad \text{if } y \text{ is a probability of an outcome} \\
\text{softmax} & \quad \text{if } y \text{ is a gold class}
\end{align*}
\]

If \(\mathbf{x} \in \mathbb{R}^d\), we can use a neural network with an input layer of size \(d\), hidden layer of size \(h\) with a non-linear activation function, and an output layer of size \(o\) (either 1 or number of classification classes) with the mentioned output function.
We have

\[ h_i = f^{(1)} \left( \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)} \right) \]

where

- \( W^{(1)} \in \mathbb{R}^{h \times d} \) is a matrix of weights,
- \( b^{(1)} \in \mathbb{R}^h \) is a vector of biases,
- \( f^{(1)} \) is an activation function.

The weights are sometimes also called a kernel.

The biases define general behaviour in case of zero/very small input.

Transformations of type \( wx + b \) are called affine instead of linear.
Similarly with another matrix of weights, another vector of biases, being an output activation function.

\[ o_i = f^{(2)} \left( \sum_j W_{i,j}^{(2)} h_j + b_i^{(2)} \right) \]

with
- \( W^{(2)} \in \mathbb{R}^{o \times h} \) another matrix of weights,
- \( b^{(2)} \in \mathbb{R}^o \) another vector of biases,
- \( f^{(2)} \) being an output activation function.
Putting It All Together

The parameters of the model are therefore $W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}$ of total size $d \times h + h \times o + h + o$.

To train the network, we repeatedly sample $m$ training examples and perform SGD (or any its adaptive variant), updating the parameters to minimize the loss.

$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial L}{\partial \theta_i}$$

We set the hyperparameters (size of the hidden layer, hidden layer activation function, learning rate, ...) using performance on the validation set and evaluate generalization error on the test set.
Practical Issues

- Processing all input in *batches*.
- Vector representation of the network.

Considering \( h_i = f^{(1)} \left( \sum_j W_{i,j}^{(1)} x_j + b_i^{(1)} \right) \), we can write

\[
 h = f^{(1)} \left( W^{(1)} x + b^{(1)} \right)
\]

\[
 o = f^{(2)} \left( W^{(2)} h + b^{(2)} \right) = f^{(2)} \left( W^{(2)} \left( f^{(1)} \left( W^{(1)} x + b^{(1)} \right) \right) \right) + b^{(2)}
\]

The derivatives

\[
 \frac{\partial f^{(1)}}{\partial x} \left( W^{(1)} x + b^{(1)} \right), \quad \frac{\partial f^{(1)}}{\partial W^{(1)}} \left( W^{(1)} x + b^{(1)} \right)
\]

are then matrices (called *Jacobians*) or even higher-dimensional tensors.
Computation Graph

\[
x \cdot W^1 + b^1 \\
\rightarrow \\
\cdot b^2 \\
\rightarrow f^1 \cdot W^2 + b^2 \\
\rightarrow f^2 + L
\]
## High Level Overview

<table>
<thead>
<tr>
<th>Classical ('90s)</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>CNN, RNN, Transformer, VAE, GAN, ...</td>
</tr>
<tr>
<td>Activation func.</td>
<td>tanh, ReLU, PReLU, ELU, SELU, Swish, ...</td>
</tr>
<tr>
<td>Output function</td>
<td>none, σ, softmax</td>
</tr>
<tr>
<td>Loss function</td>
<td>NLL (or cross-entropy or KL-divergence)</td>
</tr>
<tr>
<td>Optimization</td>
<td>SGD, RMSProp, Adam, ...</td>
</tr>
<tr>
<td>Regularization</td>
<td>L2, Dropout, Label smoothing, BatchNorm, LayerNorm, ...</td>
</tr>
</tbody>
</table>
Derivative of MSE Loss

Given the MSE loss of

\[ L = (y - \hat{y}(x; \theta))^2 = (\hat{y}(x; \theta) - y)^2, \]

the derivative with respect to \( \hat{y} \) is simply:

\[ \frac{\partial L}{\hat{y}(x; \theta)} = 2(\hat{y}(x; \theta) - y). \]
Let us have a softmax output layer with

\[ o_i = \frac{e^{z_i}}{\sum_j e^{z_j}}. \]
Derivative of Softmax MLE Loss

Consider now the MLE estimation. The loss for gold class index \textit{gold} is then

\[ L(\text{softmax}(z), \text{gold}) = -\log o_{\text{gold}}. \]

The derivation of the loss with respect to \( z \) is then

\[
\frac{\partial L}{\partial z_i} = \frac{\partial}{\partial z_i} \left[ -\log \frac{e^{z_{\text{gold}}}}{\sum_j e^{z_j}} \right] = - \frac{\partial z_{\text{gold}}}{\partial z_i} + \frac{\partial \log(\sum_j e^{z_j})}{\partial z_i}
\]

\[
= - [\text{gold} = i] + \frac{1}{\sum_j e^{z_j}} e^{z_i}
\]

\[
= - [\text{gold} = i] + o_i.
\]

Therefore, \( \frac{\partial L}{\partial z} = -1_{\text{gold}} + o \), where \( 1_{\text{gold}} \) is 1 at index \textit{gold} and 0 otherwise.
Gold distribution

Model distribution

Loss derivative with respect to the softmax inputs.
In the previous case, the gold distribution was \textit{sparse}, with only one target probability being 1. In the case of general gold distribution \( g \), we have

\[
L(\text{softmax}(z), g) = -\sum_i g_i \log o_i.
\]

Repeating the previous procedure for each target probability, we obtain

\[
\frac{\partial L}{\partial z} = -g + o.
\]

\textbf{Sigmoid}

Analogously, for \( o = \sigma(z) \) we get \( \frac{\partial L}{\partial z} = -g + o \), where \( g \) is the target gold probability.
Gold distribution

Model distribution

Loss derivative with respect to the softmax inputs.
As already mentioned, regularization is any change in the machine learning algorithm that is designed to reduce generalization error but not necessarily its training error.

Regularization is usually needed only if training error and generalization error are different. That is often not the case if we process each training example only once. Generally the more training data, the better generalization performance.

- Early stopping
- L2, L1 regularization
- Dataset augmentation
- Ensembling
- Dropout
- Label smoothing
Regularization – Early Stopping

Figure 7.3, page 246 of Deep Learning Book, http://deeplearningbook.org
L2 Regularization

We prefer models with parameters small under L2 metric.

The L2 regularization, also called weight decay, Tikhonov regularization or ridge regression therefore minimizes

$$\tilde{J}(\theta; \mathbf{X}) = J(\theta; \mathbf{X}) + \lambda \|\theta\|_2^2$$

for a suitable (usually very small) $\lambda$.

During the parameter update of SGD, we get

$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial J}{\partial \theta_i} - 2\alpha \lambda \theta_i$$
L2 Regularization

Figure 7.1, page 233 of Deep Learning Book, http://deeplearningbook.org
Another way to arrive at L2 regularization is to utilize Bayesian inference.

With MLE we have

$$\theta_{\text{MLE}} = \arg \max_{\theta} p(X; \theta).$$

Instead, we may want to maximize \textit{maximum a posteriori (MAP)} point estimate:

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta; X)$$

Using Bayes' theorem

$$p(\theta; X) = p(X; \theta)p(\theta)/p(X),$$

we get

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(X; \theta)p(\theta).$$
L2 Regularization as MAP

The $p(\theta)$ are prior probabilities of the parameter values (our *preference*).

One possibility for such a prior is $\mathcal{N}(\theta; 0, \sigma^2)$.

Then

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(X; \theta) p(\theta)$$

$$= \arg \max_{\theta} \prod_{i=1}^{m} p(x^{(i)}; \theta) p(\theta)$$

$$= \arg \min_{\theta} \sum_{i=1}^{m} -\log p(x^{(i)}; \theta) - \log p(\theta)$$

By substituting the probability of the Gaussian prior, we get

$$\theta_{\text{MAP}} = \arg \min_{\theta} \sum_{i=1}^{m} -\log p(x^{(i)}; \theta) + \frac{1}{2} \log(2\pi\sigma^2) + \frac{\theta^2}{2\sigma^2}$$
L1 Regularization

Similar to L2 regularization, but we prefer low L1 metric of parameters. We therefore minimize

$$\tilde{J}(\theta; X) = J(\theta; X) + \lambda ||\theta||_1$$

The corresponding SGD update is then

$$\theta_i \leftarrow \theta_i - \alpha \frac{\partial J}{\partial \theta_i} - \text{sign}(\theta_i) \alpha \lambda.$$
For some data, it is cheap to generate slightly modified examples.

- Image processing: translations, horizontal flips, scaling, rotations, color adjustments, ...
  - Mixup (appeared in 2017)

(b) Effect of mixup on a toy problem.

*Figure 1b of paper "mixup: Beyond Empirical Risk Minimization", https://arxiv.org/abs/1710.09412*

- Speech recognition: noise, frequency change, ...
- More difficult for discrete domains like text.
Ensembling (also called *model averaging* or in some contexts *bagging*) is a general technique for reducing generalization error by combining several models. The models are usually combined by averaging their outputs (either distributions or output values in case of a regression).

The main idea behind ensembling is that if models have uncorrelated (independent) errors, then by averaging model outputs the errors will cancel out.

However, ensembling usually has high performance requirements.
There are many possibilities how to train the models to average:

- Generate different datasets by sampling with replacement (bagging).
- Use random different initialization.
- Average models from last hours/days of training.

*Figure 7.5, page 257 of Deep Learning Book, http://deeplearningbook.org*
How to design good universal features?

- In reproduction, evolution is achieved using gene swapping. The genes must not be just good with combination with other genes, they need to be universally good.

Idea of dropout by (Srivastava et al., 2014), in preprint since 2012.

When applying dropout to a layer, we drop each neuron independently with a probability of $p$ (usually called dropout rate). To the rest of the network, the dropped neurons have value of zero.
Regularization – Dropout

Dropout is performed only when training, during inference no nodes are dropped. However, in that case we need to *scale the activations down* by a factor of $1 - p$ to account for more neurons than usual.
Alternatively, we might *scale the activations up* during training by a factor of $1/(1 - p)$.
Regularization – Dropout as Ensembling

Figure 7.6, page 260 of Deep Learning Book, http://deeplearningbook.org
Regularization – Dropout Effect

Figure 7: Features learned on MNIST with one hidden layer autoencoders having 256 rectified linear units.

*Figure 7 of paper “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, http://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf*
def dropout(input, rate=0.5, training=False):
    def do_inference():
        return tf.identity(input)

    def do_train():
        random_noise = tf.random_uniform(tf.shape(input))
        mask = tf.cast(tf.less(random_noise, rate), tf.float32)
        return input * mask / (1 - rate)

    if training == True:
        return do_train()
    if training == False:
        return do_inference()
    return tf.cond(training, do_train, do_inference)
Regularization – Label Smoothing

Problem with softmax MLE loss is that it is never satisfied, always pushing the gold label probability higher (but it saturates near 1).

This behaviour can be responsible for overfitting, because the network is always commanded to respond more strongly to the training examples, not respecting similarity of different training examples.

Ideally, we would like a full (non-sparse) categorical distribution of classes for training examples, but that is usually not available.

We can at least a simple smoothing technique, called label smoothing, which allocates some small probability volume $\alpha$ uniformly for all possible classes.

The target distribution is then

\[
(1 - \alpha)\mathbf{1}_{\text{gold}} + \alpha \frac{1}{\text{number of classes}}.
\]
Regularization – Label Smoothing

Gold distribution

Smoothed distribution
When you need to regularize, then a good default strategy is to:

- use dropout on all hidden dense layers (not on the output layer), good default dropout rate is 0.5 (or use 0.3 if the model is underfitting);
- use L2 regularization for your convolutional networks;
- use label smoothing (start with 0.1);
- if you require best performance and have a lot of resources, also perform ensembling.
The training process might or might not converge. Even if it does, it might converge slowly or quickly.

There are *many* factors influencing convergence and its speed, we now discuss three of them.
Convergence – Saturating Non-linearities

Neural networks usually need random initialization to break symmetry.

- Biases are usually initialized to a constant value, usually 0.
- Weights are usually initialized to small random values, either with uniform or normal distribution.
  - The scale matters for deep networks!
  - Originally, people used $U\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]$ distribution.

The authors theoretically and experimentally show that a suitable way to initialize a $\mathbb{R}^{n \times m}$ matrix is

$$U\left[-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right].$$
Figure 6 of paper "Understanding the difficulty of training deep feedforward neural networks", http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.
Figure 7 of paper "Understanding the difficulty of training deep feedforward neural networks", http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.