Draft version January 5, 2018
Typeset using LaTeX **modern** style in AASTeX61

THE MATHEMATICS OF DEEP LEARNING*

Vishal P. Kasliwal[1]

[1] *Wave Computing*
42 W. Campbell Ave, # 301
Campbell, CA 95008, USA

(Received December 28, 2017; Revised December 28, 2017; Accepted January 5, 2018)

ABSTRACT

Deep Learning is a branch of Machine Learning in which 'deep' neural networks are used for various purposes.

Keywords: neural networks, training

Corresponding author: Vishal Kasliwal
vishal@wavecomp.com, vishal.kasliwal@gmail.com

* This is a draft prepared for the Agent Library group at Wave Computing.

1. PRELIMINARIES

2. NOTATION

Tensors are associated with layers. The $p \times q$ tensor $\mathbf{A}$ when associated with layer l is denoted $\mathbf{A}^{[l]}_{p\times q}$. The transpose of this tensor is denoted by $(\mathbf{A}^{[l]}_{p\times q})^\top$. If the tensor is an input tensor to a layer, it may additionally have a minibatch number, m, and instance within the minibatch, i, associated with it. Thus $(\mathbf{A}^{[l]\{m\}(i)}_{p\times q})^\top$ is the transpose of the i-th training example from the m-th minibatch in the l-th layer from the input tensor $\mathbf{A}$ which is p rows high by q columns long. Furthermore, layers associated with convolutional neural networks may have *channels* associated with them. Channels add extra dimensions to tensors and so $(\mathbf{A}^{[l]\{m\}(i)}_{p\times q\times k_1\times k_2})^\top$ is the transpose of the i-th training example from the m-th minibatch in the l-th layer from the input tensor $\mathbf{A}$ which is p rows high by q columns long and has two sets of channels of depth k_1 and k_2 respectively.

The $\mathrm{n}^{[l]}$ operator returns the number of nodes, i.e. the width, of the l-th layer. Traditionally, the depth of the neural network is denoted by L, and hence we must have $0 \leqslant l \leqslant L$.

We transform the $l-1$-th layer of activations $\mathbf{A}^{[l-1]\{m\}(i)}_{p\times q}$, using neural network operations such as convolution etc... in the l-th layer to produce the l-th layer of activations i.e. $\mathbf{A}^{[l]\{m\}(i)}_{p\times q}$.

3. ACTIVATION FUNCTIONS

3.1. *Sigmoid Activation Function*

The sigmoid activation function takes the form

$$g(z) = \sigma(z) = \frac{1}{1+\mathrm{e}^{-z}}. \tag{1}$$

The derivative of the sigmoid activation function is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}z} = \sigma(1-\sigma). \tag{2}$$

3.1.1. *Derivation*

By definition

$$\sigma(z) = \frac{1}{1+\mathrm{e}^{-z}} = (1+\mathrm{e}^{-z})^{-1},$$

and so

$$\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = -(1+\mathrm{e}^{-z})^{-2}\frac{\mathrm{d}(1+\mathrm{e}^{-z})}{\mathrm{d}z}.$$

But this is just

$$-(1+\mathrm{e}^{-z})^{-2}(-\mathrm{e}^{-z}) = \sigma^2\mathrm{e}^{-z}.$$

Now $\mathrm{e}^{-z} = \frac{1}{\sigma} - 1 = \frac{1-\sigma}{\sigma}$ and so

$$\frac{\mathrm{d}\sigma(z)}{\mathrm{d}z} = \sigma^2\mathrm{e}^{-z} = \sigma^2\frac{1-\sigma}{\sigma} = \sigma(1-\sigma),$$

Q.E.D.

APPENDIX

REFERENCES