

Development and Evaluation of a Visual Attention Model with Python and Tensorflow

Oleg Yarin

A thesis presented for the degree of
Bachelor of Science

First supervisor: Prof. Dr. Christian Herta
Second supervisor: Dr. Vera Hollink



**Hochschule für Technik
und Wirtschaft Berlin**

University of Applied Sciences

Applied computer science
HTW Berlin
Germany
31-05-2017

Contents

1	Introduction	4
1.1	Motivation	5
2	Theory	7
2.1	Artificial Neural Networks	7
2.2	Recurrent Neural Network	14
2.2.1	Long Short-Term Memory (LSTM)	16
2.3	Reinforcement Learning	18

List of Figures

2.1	A perceptron [1].	8
2.2	Two layer perceptron network[1].	9
2.3	The architecture of neural networks[1].	11
2.4	Unrolling recurrent neural network(source: [2])	15
16figure.2.5		
2.6	LSTM's output gate layer (Source: ??)	18

List of Tables

Chapter 1

Introduction

Neural network approaches have received much attention in the last several years. It's becoming a popular choice for performing various tasks like speech and image recognition, object detections etc. as these methods have dramatically increased accuracy compared to traditional machine learning approaches. However, achieving high accuracy on recognition tasks is still computationally expensive and needs improvements in performance. This study will be a close resemblance of the recurrent neural network of visual attention which is able to extract necessary information from an image by looking at it in low resolution, and then adaptively select parts that are most relevant for a task [3].

The idea of visual attention was inspired by how human perception works. Humans do not perceive a visual scene as a whole but focus on parts of the scene that gives the most useful information to them. Humans are also capable of combining information from different parts of a picture. They then connect it to build a subjective knowledge of the picture (or sequence of pictures) [4]. Taking into account these properties, researchers from google Deepmind build a model which can be described as follows:

Instead of processing an entire image or even bounding box at once, at each step, the model selects the next location to attend to; based on past information and the demands of the task.... The model is a recurrent neural network (RNN) which processes inputs sequentially, attending to different locations within the images (or video frames) one at a time, and incrementally combines information from these fixations to build up a dynamic internal representation of the scene or environment.[3]

One of the main advantage of this model, is that the computation required is controlled and is independent of the input image size. Deepmind’s researchers evaluated their model on several image classification and dynamic visual control problems which showed a better performance when compared with convolution neural network[4].

The evidence from this study suggests application of this model on large scale object recognition as well as classification of sequence of images, which will be a great fit since the model’s performance is not dependent on the size of an input object.

The main aim of this study is to extend the current knowledge of the work mentioned above and build a model which will be able to classify a set of images and develop appropriate prototype system since it can be useful in a variety of areas. However, the current work is limited by low-resolution images and mostly will concentrate on classifying a group of objects as this restriction will reduce complexity of the task and therefore reach better results on a task of classifying a group of images.

1.1 Motivation

This approach to classify a group of images has a potential to help with automated detection and classification of breast cancer metastases, which is the main concern of camelyon challenge [5]. Camelyon challenge is an inspiration for this work since pathologist’s efforts along with the assistance of automated detection system will reduce significantly not only the workload of pathologists but the human error rate in diagnosis as well.

This work will be the first step in building software that will be capable of classifying whole-slide images of histological lymph node at the patient level. That is, bringing together estimations from multiple lymph node slides into a single outcome.

Digital pathology is a very attractive field for machine learning researchers since whole-slide images have a very high resolution and are typically about 200000 x 100000 pixels. To give you some sense of data, camelyon challenge provides data for 200 patients, where each patient has 5 different slides. It means that in total they release about 1000 slides and that is 55.88gb of uncompressed data [5].

It is quite clear that using CNN for this task is computationally very expensive. Applying model of visual attention promises to solve the issue of high-resolution pictures at a computational level. Therefore making an extensible piece of software, that will allow further improvements is also one of the main concerns of this work.

Chapter 2

Theory

2.1 Artificial Neural Networks

Why Neural Networks? Before going into what actually artificial neural networks are, let's first try to face the question why do we need it in this paper. The problem that we give to our application to solve can be shortly summarized in the following statement: “ Given a group of images, find the patterns in them that are more influential on your belief that an image(or a group of images) belongs to a specific class. ”

This problem is known as pattern recognition problem or in our case visual pattern recognition problem [6]. To solve this problem it's required to develop ability for a machine to recognize patterns that will help to make a decision about a class. The obstacles that can appear by solving this problem can be more visible if we will try to write a conventional computer program, i.e. bunch of rules to identify these patterns. What seems to be easy for us, is really hard to describe algorithmically. In these system the computational steps are deterministic hence not flexible enough for the whole variety of input data [1].

Solving problem differently Artificial Neural Networks(and machine learning in general) are looking at the problem in a different way. They don't execute programs instructions, they don't have sequential semantic and normally are not deterministic. They acquire their "knowledge" by extracting patterns from raw data, which normally called training data(which normally is a set of tuple (input, label)) This approach also know as concept of statistical pattern recognition. [6] Artificial Neural networks have recently shown an excellent performance and accuracy at recognizing objects compared with other machine learning techniques [7].

What is Neural Network? Artificial Neural Network(ANN), often referred just as Neural Network(NN), in simply words is a computational model, which was inspired by how human/animal brain works. Artificial NN is modeled based on the neuronal structure in the brain's cortex. Though the inspiration was from the brain, it's indeed much much simpler than brain in terms of number of neurons that is used in ANN [8]. To understand how neural networks works it is crucial to understand first the way a *perceptron* work.

Perceptron is a simple type of an artificial neuron. Given several binary inputs, perceptron is meant to produce a single binary output.

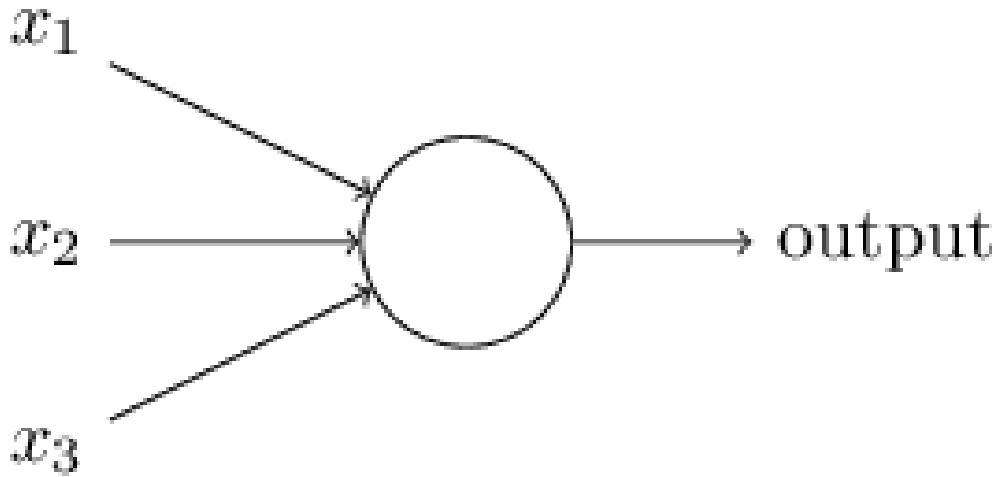


Figure 2.1: A perceptron [1].

In the figure ?? the perceptron has three inputs: x_1, x_2, x_3 . To produce an output the perceptron posses of *weights*: w_1, w_2, w_3 which represents connection between input and output. Weights determine how important is an input to the output. That said, the perceptron output is determined by whether the weighted sum $\sum_j w_j x_j$ is more or less than some *threshold* value:

$$output = \begin{cases} 0, & \text{if } \sum_j w_j x_j \leq threshold \\ 1, & \text{if } \sum_j w_j x_j > threshold \end{cases} \quad (2.1)$$

In shortly, this is a computational model which make a decision by weighting up the evidence(input data) [9].

Of course such a model is not capable of making complicated decisions, but by extending the model to more complex network of perceptrons, we might improve the model.

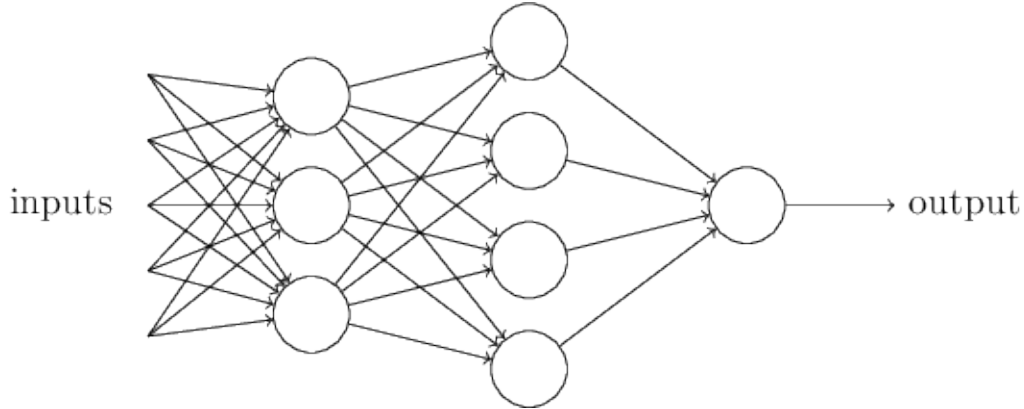


Figure 2.2: Two layer perceptron network[1].

In the network shown on figure ??, we can observe two layer of perceptrons. First layer will correspond to the first column of perceptrons and will be responsible for the weighting up input, in contrast to second layer (second column of perceptrons) which determines output by weighting up the result from the first layer. Therefore second layer is located on more abstract level from input data compared and can make more complicated decisions. Further layers might be capable of making even more sophisticated decisions.

Neural Network Now that we know the way perceptrons work, it's fairly easy to understand Neural Network. However we need to change the mathematical notation a bit. For the sake of convenience, let's move *threshold* in equation ?? to the left part and replace it with a variable known as the bias : $b = -threshold$. Let's also simplify the sum sign: $\sum_j w_j x_j$ by writing weights and input as vectors and use a dot product to multiply them: $\sum_j w_j x_j = W \cdot x$. Using changes described above we can rewrite the equation ?? as following:

$$output = \begin{cases} 0, & \text{if } w \cdot x + b \leq 0 \\ 1, & \text{if } w \cdot x + b > 0 \end{cases} \quad (2.2)$$

Bias is a measure of how influential is a certain neuron on making output 1. Some people also use more biological terms: the bias is a measure of

how easy it is to get an neuron to fire. To devise the neural network next improvement over the perceptron network is that network should not be limited to have an input only binary value, but any value. The same applies on the output. Output being only binary value will limit the ability to make sophisticated decision. Therefore we introduce a function known as *an activation function* before actually outputting a value: $output = g(W \cdot x + b)$ [1] .

Activation function $g(\cdot)$ is known as an activation function. Activation function helps to control the output and non-linearity of the network. Activation function also plays a crucial role in multi layer architecture, where it helps to prevent the values of each layer from blowing up. For example, let's take a look at logistic sigmoid activation function which has following form:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.3)$$

TODO: write about tahn activation function

The use of the sigmoid activation will make a network to produce output to be interpreted as posterior probabilities. Probability interpretation helps to provide more powerful and more useful results [6]. There are a good variety of activation functions but in this work we mainly will use the sigmoid activation function and *rectified function*. Rectified function is fairly simple. It produces 0 when input is less than 0, and it does not change input value if input is more than 0:

$$R(z) = \max(0, z) \quad (2.4)$$

Now that we derived a concept of Neural Network, we can talk more about what is called feedforward neural network and the terms related to it.

Feedforward neural network is a network where the output from one layer is used as input to the next layer. In feedforward NN information is always fed forward in contrast to recurrent neural network(RNN) where information can go in a loop. We will take a closer look at RNN in section ??.

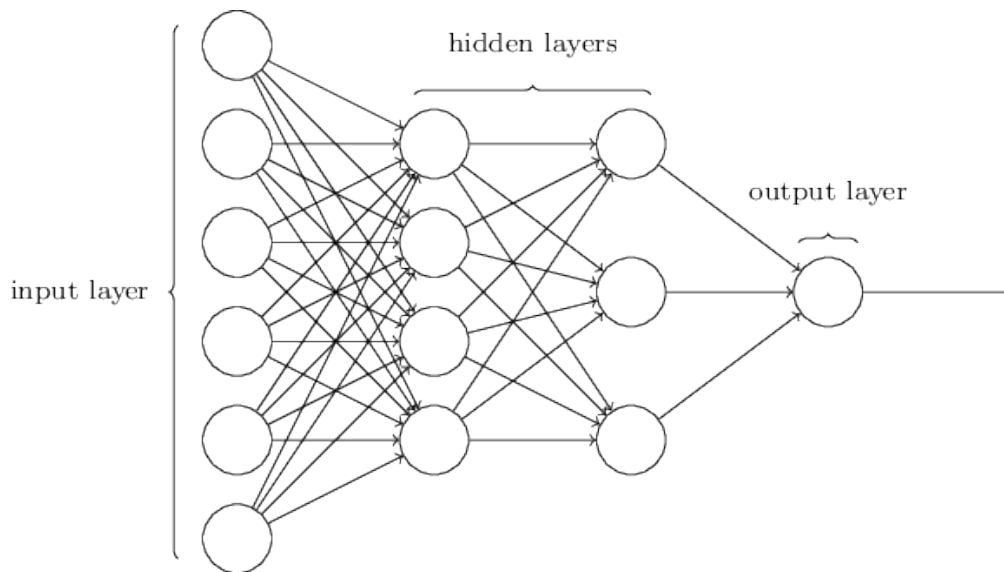


Figure 2.3: The architecture of neural networks[1].

Basically Feedforward neural network is exactly what we described above. Let's name different parts of feedforward NN:

- Input layer - the leftmost layer in the network. The Neurons within input layer, called input neurons.
- Output layer - the rightmost layer in the network. The Neurons within input layer, called output neurons.
- Hidden layers - all the layers excluding input and output layers.

For example, the Neural network in figure ?? consist of

- 6 input neurons
- 2 hidden layers
 - first hidden layer consist of 4 neurons
 - second hidden layer consist of 3 neurons
- output layer consist of one single neuron

Training As mentioned above Neural Network is capable of solving complicated pattern recognition problems. However designing an neural network is not sufficient for this. It's also requiring to train an network. In this paragraph we will introduce learning procedure. But before going into learning, let's recap how our neural network model looks like:

$$y = g(W * x + b) \quad (2.5)$$

Where g is an activation function, W - weights, b -biases, x - input data. The space of different weights and biases values building a space of solution for cetrain problem. The goal of the training is to find the best parameters for neural network (W, b) , that suited our problem.

Training data To solve pattern recognition problem we need to provide a continuous feedback to NN, which NN can use to learn from. This feedback in machine learning called *training data*. Training data consist of the input data samples and appropriated outputs. You can think of training data as of an list of tuples: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$ where x - input, y - output (also known as ground truth), n - amount of training examples. Neural network trained on training data should be able to generalise the output on unseen input data. The goal of the learning is to train network on training data and make it to capable of generalizing the output on unseen input data.

Cost In order to teach the model, it's essential to understand what does it mean for an neural network to be good or to be bad. For this purpose it's required first to define a cost function. Cost(also knows as error) is the metric for the NN(or any other function approximation method), which represents how far of the model is from desired outcome. If cost is big , our network does not work well. With the cost function, it's possible to define our training goal more precisely: the smaller our cost, the better our model works, therefore the goal of the learning is to minimize the cost of our model.

Types of cost There are a plenty of ways to define cost of the model. Let's consider the common type of cost function called *mean squared error function*. Mean squared error function has following form:

$$I_{W'} = 1/2 * n * \sum_{i=1}^n (y_{W'}(x^{(i)}) - y^{(i)})^2 \quad (2.6)$$

- $y_{W'}$ - is our neural network function with the parameters (W', b') ,

- $x^{(i)}, y^{(i)}$ - is input and output of a training sample respectively.

Another common type of function to measure the cost of NN known as *cross entropy*. In short, cross entropy gives a way to express how different two distributions are:

$$I_{W'} = - \sum_{i=1}^n y_{W'}(x^{(i)}) \log(y^{(i)}) \quad (2.7)$$

where:

- $y_{W'}$ - is our neural network function with the parameters (W', b') ,
- $x^{(i)}, y^{(i)}$ - is input and output of a training sample respectively.

??

Gradient Descent Once we defined our cost function, we need to find a set of parameters W, b which make the cost as small as possible. The most common algorithm used to minimize the cost function called *gradient descent*. Let's explain the algorithm on an example function: $f = f(V)$, where $V = \vec{v}_1, \vec{v}_2, \dots$ are variables that we want to minimize. Gradient descent uses partial derivative to iteratively update parameters. Derivative of a function shows how function output will change with respect to very, very small change in input ΔV . For example, partial derivative with respect to variable $\Delta \vec{v}_1$ will tell us, how different the output will be Δf if we change \vec{v}_1 on the small amount. This property of derivative is used in gradient descent algorithm. Essentially, the gradient descent performs updates on the variables to be minimized according to partial derivative of the cost function with respect to this variables.

Gradient descent adopt the following procedure. Beginning with an initial guess for value $v = \vec{v}_1, \vec{v}_2, \dots$, we update the vector v by moving a small distance in v-space in direction in which our function f raises most rapidly, i.e. in the direction of $-\Delta_V * f$. Iterating this process, we can devise the new set of parameters $V^{(new)}$:

$$v_i^{new} = \vec{v}_i - \alpha \frac{df(\vec{v}_i)}{d\vec{v}_i} \quad (2.8)$$

where α - is a small positive number known as *learning rate*. Learning rate determines the smoothness of updates and it's very important to choose it appropriately since if learning rate is too small the learning can be too slow, while, if learning rate is too big, algorithm updates can be too big to achieve the minimum (it can overstep the minimum).

Depends on the conditions this will converge to the parameters V where the function f is minimized. One important things to notice is that, we can use the gradient descent algorithm only if f is differentiable. [6]

Mini-batch Gradient Descent Normally gradient descent algorithm is associated with the update on the loss computed with whole set of training data, while, gradient descent where updates is performed only using loss computed on a small batch of data knows as *mini-batch gradient descent*. Much faster convergence can be achieved in practice using mini-batch gradient descent. [10]

Backpropagation In order to compute gradients in gradient descent algorithm backpropagation algorithm is normally used. Backpropagation is the procedure of computing gradients applying a gradient chain rules and updating the weights accordingly. It performs first an forward update to receive the network's error value. This error value then is back propagated beginning from output layer(neuron) through all neurons till the input in order to associate it with extent of this error(Δ) which an certain neuron is responsible for. Once this extent is calculated, it performs weights update [11].

2.2 Recurrent Neural Network

Why recurrent NN? Motivation for recurrent neural network(RNN) is that in contrast to feedforward neural network, RNNs are capable of having internal memory, i.e. capable of memorizing information, therefore deal better with sequential input data. RNNs are closer to the way human's brain works. We don't start our thinking scratch, all of us have different background, memories and experience and based on this we're making our own decision and actions.

As already mentioned in chapter chapter 1, in this work the network should be able attend to different locations within an image, i.e. choose a location and process only area with respect to this location. The network then incrementally combines the information from different location and based on the knowledge(memory) extracted from a location, network chooses a new location to attend. As you might notice since more steps need to be required, we will have sequential data. Hence, RNNs are underlying concept for this work.

What is RNN RNN is a special type of neural network architecture, which can accept a sequence with an arbitrary length without changing weights of a network. RNN are capable of persisting the information by means of recurrences, i.e. including the output of the network into next computational steps and summarizing this information into an object called *state*. ??.

To simplify the understanding you can imagine RNN as an composition of identical feedforward network, one for each step in time, passing the message to a successor. Essentially, it's a computational unit that is repeated over and over again and can be also thought as an for-loop. One neural network in this composition known as *RNN cell*.

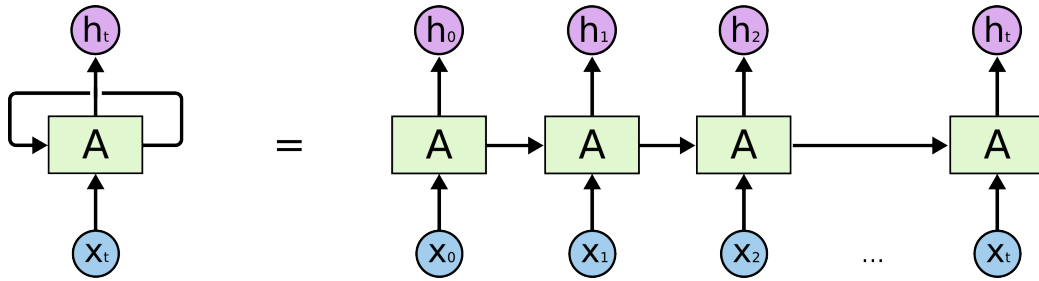


Figure 2.4: Unrolling recurrent neural network(source: [2])

On the right side of the figure ?? you can see an unrolled RNN that accepts as input $x_0, x_1, x_2, \dots, x_t$ and produces following output: h_1, h_2, \dots, h_t . One time step represents a layer in terms of forward neural network. The whole concept can be explained in the following equation:

$$\begin{pmatrix} s_t \\ o_t \end{pmatrix} = f \begin{pmatrix} s_{t-1} \\ x_t \end{pmatrix} \quad (2.9)$$

where:

- s_t, s_{t-1} - are states at time step t and $t - 1$ respectively,
- o_t - is the output at time step t ,
- x_t - is the input at time step t ,
- f - is a recurrent function(normally called as RNN cell).

As you might notice, the all calculations responsible for extracting and memorising information performed in f , which provide knowledge about specific *RNN architecture(RNN cell)*. Thus the choice of recurrent function

$f(\text{RNN cell})$ is essential for RNNs to work and remember information. There are a lot of variations of RNN cells, but we mostly will consider one of the recent and most widely known architecture called *Long Short-Term Memory (LSTM)*.

2.2.1 Long Short-Term Memory (LSTM)

The reason that the vanilla RNN(Elman RNN cell ??) is not considered here, because they are not able to learn long term dependencies due to vanishing and exploding gradient problem ??.

Long Short Term Memory networks(LSTMs) are a special architecture of RNN cell, capable of learning long-term dependencies. ?? LSTMs have the ability to remember new information and forgetting old, unnecessary information using concept of gates. LSTM cell holds information in the object called *state*(C_t) and only the gates are permitted to manipulate and change this state. Gates are represented as an sigmoid layer and pointwise operation and will be explained below.

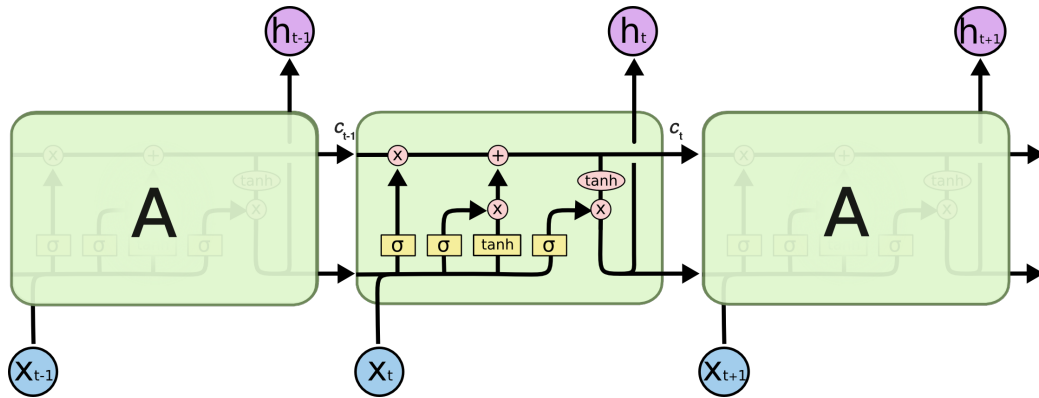


Figure 2.5: Structure of LSTM cell(Source: ??)

An yellow square with σ inside represents a neural network with sigmoid activation function, while yellow square with *tahn* inside represents a neural network with tahn activation function.

As you can see from the figure ?? LSTM cell has fours layers, which build up three gates to interact with the state:

- Forget gate
- Input gate

- Output gate

Forget gate layer The sigmoid layer that you can see on the right side is called "forget gate layer". As you might notice from the name, this gate is responsible for remembering information(or forgetting). It concatenates output from previous state: h_{t-1} with the input at the timestep t : x_t . Then the result is fed to the neural network with sigmoid activation function which produces an output with values from 0 to 1. Where 0 means to completely forget the information and 1 means to leave the information in the state:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2.10)$$

Input gate layer composed of two networks: networks with sigmoid and tahn activation functions. The first sigmoid network decides which values needed to be updated and till what extent, while the network with tahn activation function create a new candidate state value. Then the outputs from then networks are multiplied with each other to create an update for the LSTM cell's state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) C_t^{(candidate)} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (2.11)$$

Updating the state It's fairly simple now to update the old state C_{t-1} using f_t from equation ?? to forget information and using new state candidate values $C_t^{(candidate)}$ and it's extent i_t from equation ??:

$$C_t = f_t * C_{t-1} + i_t * C_t^{(candidate)} \quad (2.12)$$

Output gate layer is responsible for the cell's output h_t . This layer is allowed to read information from the state C_t and decides what information should be outputted.

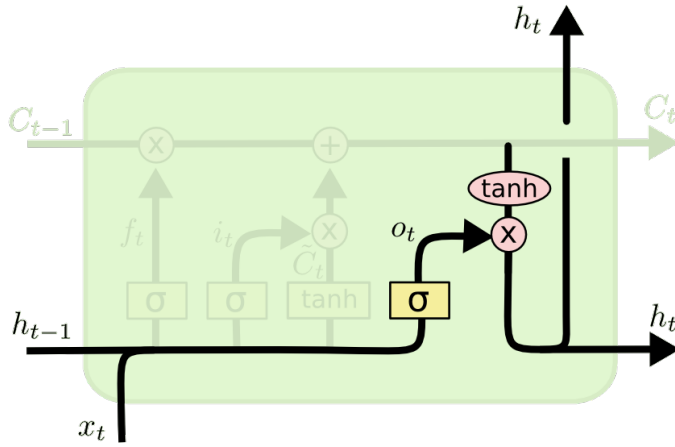


Figure 2.6: LSTM's output gate layer (Source: ??)

As you can see from figure ??: firstly, state cell goes through tahn function and then multiplied with output from the neural network of output gate layer:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)h_t = o_t * \tanh(C_t) \quad (2.13)$$

Backpropagation-Through-Time [8]

2.3 Reinforcement Learning

Bibliography

- [1] M. A. Nielsen, “Neural Networks and Deep Learning,” 2015.
- [2] Colah Christopher, “Understanding LSTM Networks – colah’s blog,” 2015.
- [3] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent Models of Visual Attention,” *Nips-2014*, pp. 1–9, jun 2014.
- [4] P. Goldsborough, “A Tour of TensorFlow Proseminar Data Mining,”
- [5] “CAMELYON17 - Home.”
- [6] C. M. Bishop, “Neural networks for pattern recognition,” *Journal of the American Statistical Association*, vol. 92, p. 482, 1995.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [9] F. Rosenblatt, *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Report (Cornell Aeronautical Laboratory), Spartan Books, 1962.
- [10] Karpathy Andrej, “CS231n Convolutional Neural Networks for Visual Recognition,” 2016.
- [11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536, oct 1986.