1. Explain the basic architecture of RNN cells

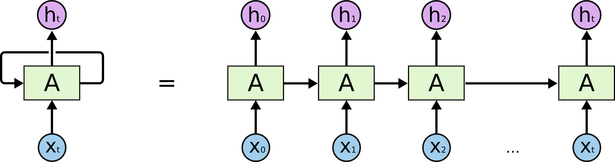
The fundamental feature of a Recurrent Neural Network (RNN) is that the network contains at least one feed-back connection, so the activations can flow round in a loop. That enables the networks to do temporal processing and learn sequences, e.g., perform sequence recognition/reproduction or temporal association/prediction. In sequential tasks such as natural language and speech processing, there is always dependence of present input data upon the previous applied inputs. Task of RNNs is to find the relationship between current input and the previous applied inputs. 

Figure: Basic architecture of Recurrent Neural Networks

Here,

Xt is the input at time step t. Xt is a vector of any size N.

A is the hidden state at time step t. It’s the “memory” of the network. It is calculated based on the previous hidden state and the input at the current step.

Represented by At= f (W Xt +U At-1)

Here W and U are weights for input and previous state value input. And f is the non-linearity applied to the sum to generate the final cell state.

1. Explain Backpropagation through time (BPTT)

Backpropagation Through Time, or BPTT, is the application of the Backpropagation training algorithm to recurrent neural networks applied to sequence data like a time series.A recurrent neural network is shown one input each timestep and predicts one output.

Conceptually, BPTT works by unrolling all input timesteps. Each timestep has one input timestep, one copy of the network, and one output. Errors are then calculated and accumulated for each timestep. The network is rolled back up and the weights are updated.

1. Explain Vanishing and exploding gradients.

**Vanishing:**

During backpropagation, the calculation of (partial) *derivatives/gradients* in theweight update formula follows the Chain Rule, where gradients in earlier layers are the multiplication of gradients of later layers:

Vanishing

where,

Vanishing

As the gradients frequently become SMALLER until they are close to zero, the new model weights (of the initial layers) will be virtually identical to the old weights without any updates. As a result, the gradient descent algorithm never converges to the optimal solution. This is known as the problem of vanishing gradients, and it’s one example of unstable behaviors of neural nets.

### Exploding:

On the contrary, if the gradients get LARGER or even NaN as our backpropagation progresses, we would end up with exploding gradients having big weight updates, leading to the divergence of the gradient descent algorithm.

1. Explain Long short-term memory (LSTM)

* Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.
* This is a behavior required in complex problem domains like machine translation, speech recognition, and more.
* LSTMs are a complex area of deep learning. It can be hard to get your hands around what LSTMs are, and how terms like bidirectional and sequence-to-sequence relate to the field.

1. Explain Gated recurrent unit (GRU)

* GRU or Gated recurrent unit is an advancement of the standard RNN i.e recurrent neural network.
* GRUs are very similar to Long Short Term Memory(LSTM). Just like LSTM, GRU uses gates to control the flow of information. They are relatively new as compared to LSTM. This is the reason they offer some improvement over LSTM and have simpler architecture.
* Another Interesting thing about GRU is that, unlike LSTM, it does not have a separate cell state (Ct). It only has a hidden state(Ht). Due to the simpler architecture, GRUs are faster to train.

1. Explain Peephole LSTM.

In peephole LSTM, we let the gate layers look at the cell state. In this peephole connection we can see that all the gates are having an input along with the cell state.

1. Bidirectional RNNs

Bidirectional long-short term memory(Bidirectional LSTM) is the process of making any neural network to have the sequence information in both directions backwards (future to past) or forward(past to future).

1. Explain the gates of LSTM with equations.

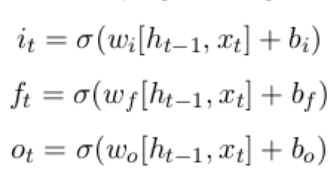
In LSTM we will have 3 gates:

**1) Input Gate.**

**2) Forget Gate.**

**3) Output Gate.**

Gates in LSTM are the sigmoid activation functions i.e they output a value between 0 or 1 and in most of the cases it is either 0 or 1.



* First equation is for the Input Gate which tells us what new information we’re going to store in the cell state(that we will see below).
* Second is for the forget gate which tells the information to throw away from the cell state.
* Third one is for the output gate which is used to provide the activation to the final output of the lstm block at timestamp ‘t’.

1. Explain BiLSTM

A Bidirectional LSTM, or biLSTM, is a sequence processing model that consists of two LSTMs:

* One taking the input in a forward direction
* The other in a backwards direction.

BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow *and* precede a word in a sentence).

1. Explain BiGRU

A Bidirectional GRU, or BiGRU, is a sequence processing model that consists of two GRUs. one taking the input in a forward direction, and the other in a backwards direction. It is a bidirectional recurrent neural network with only the input and forget gates.