Attention

Introduction

In neural networks, **attention** is a technique that **mimics cognitive attention**. The effect enhances some parts of the input data while diminishing other parts — the thought being that the network should devote more focus to that small but important part of the data.

Learning which part of the data is more important than others depends on the context and is trained by gradient descent.

General idea

We have a sequence of tokens labelled by the index i. The neural network computes a soft weight w_i for each token i with the property that $\sum_i w_i = 1$.

Each token is assigned a value vector v_i which is computed from the Word embedding of the ith token. The weighted average $\sum_i w_i v_i$ is the output of the attention mechanism.

A language translation example

To build a machine that translates English-to-French, one starts with an Encoder-Decoder and grafts an attention unit to it.

In the simplest case such as the example below, the attention unit is just lots of dot products of recurrent layer states and does not need training.
In practice, the attention unit consists of 3 fully connected neural network layers that need to be trained.
The 3 layers are called Query, Key, and Value.