

Personal GPT

A Recurrent Neural Network on chip

Authors:

David Lanzendörfer

Date:

July 4, 2024

Institution or Organization:

Freedom Club

Contact Information:

leviathan@libresilicon.com

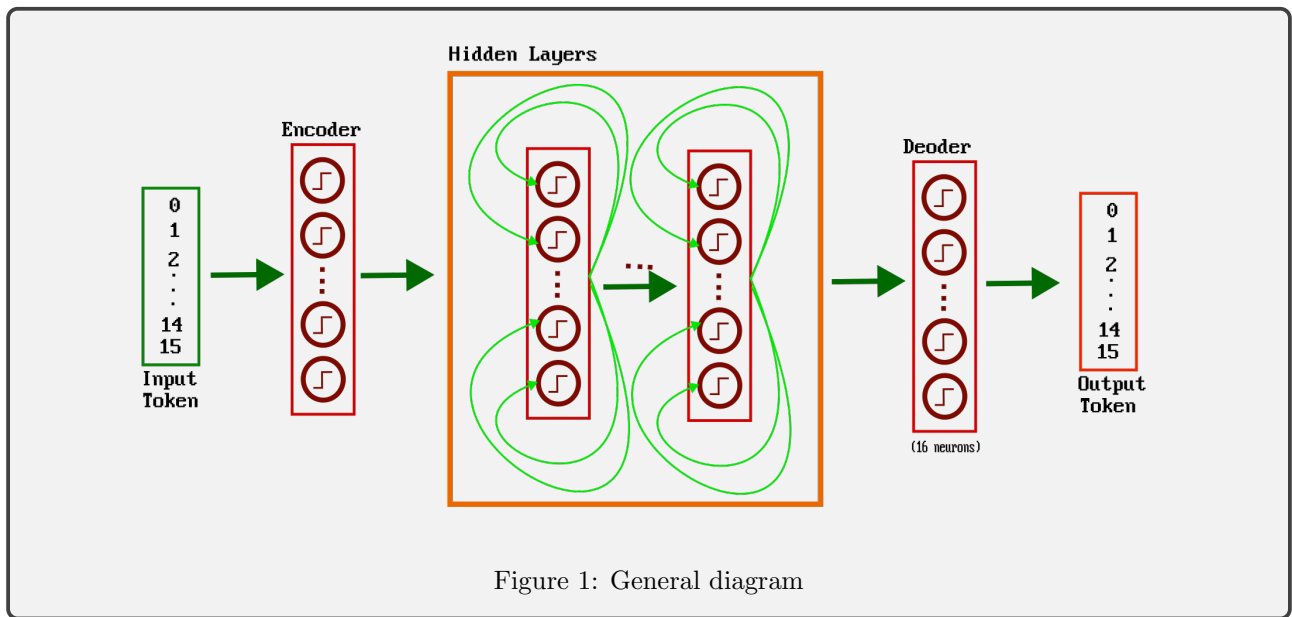
Abstract

In the past couple of years generative AI has received a lot of attention and is getting increased traction in various fields, being used for generating audio, video, image and text. However, current day AI solutions are centrally hosted and the generation is being done by transformers running distributed on a huge array of GPUs. This solution has many drawbacks, mainly the power consumption which comes from such a large set of GPUs, but it also is a problem when it comes to privacy. All the inference requests have to be forwarded to those server farms and it's not certain what the operator will do with the logs of the inference requests. Considering that OpenAI is an American company and as such subject to FISA, it makes it impossible for public administration offices like the European federal bureaucracies to use ChatGPT, despite its potential benefits and decrease in inefficiencies it would offer to government administration processes which are well known to be inherently inefficient by nature. Personal GPT addresses those problems by packing a Large Language Model onto a chip which can be attached to a computer over USB and can generate text without the need of an internet connection. All the data, confidential or not, stays within your own four walls.

1 Introduction

The Personal GPT architecture consists of three types of layers:

1. The Input layer, also known as the encoder
2. The Hidden layers, where all the parameters for the knowledge are being stored
3. The Output layer, also known as the decoder, which maps the positional vector provided by the hidden network back to a token.



1.1 Encoder Layer

The encoder layer is composed of a configurable amount of input perceptrons, which are wired in a recurrent neural network configuration. In this RNN configuration, the neuron does not only contain input synapses for the 16 input token bits (the token space of the GPT2 tokenizer) but also an input bit array containing the last encoded positional vector. This way, the positional encoding becomes a continuous time series.

1.2 Hidden Layer

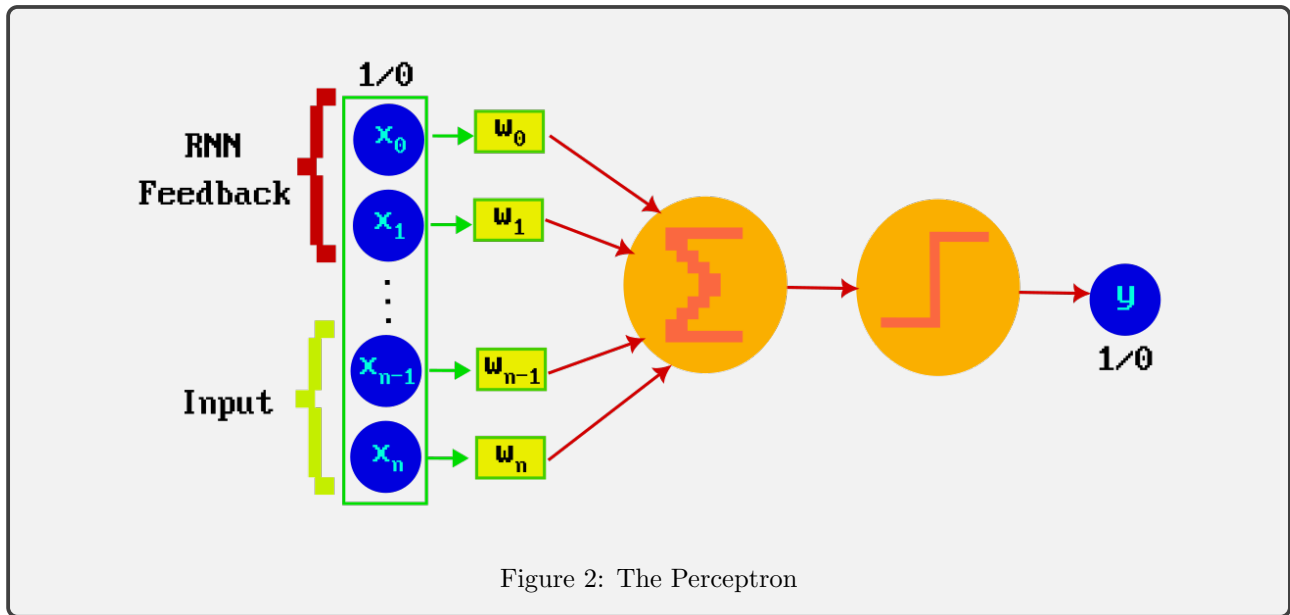
The hidden layer a multiple layers like the encoder layer being put into series, resulting in a neuron matrix. The width and the height of this matrix can be configured with parameters as well.

1.3 Decoder Layer

The decoder layer is the only non recurrent layer because it's only job is to map the output of the recurrent neuron matrix to a corresponding token.

2 The Perceptron

The perceptron which is being used in our design can be seen as a diagram below.



The inputs of the perceptron synapses are binary either one or zero, and the output of the perceptron is also binary either one or zero due to the activation function having been chosen to be the step function.

3 Continuous Positional Encoding