The main disadvantage of all these attention models is that, because we are computing all the various different options, they require many computations for every step, causing a linear increase in computation costs and massive use of memory.

(F) Transformer - attention only:

Transformer is a new model, hot off the presses from Google. In this model the network relies exclusively on attention units to learn the relationships between input and output, with no use of RNNs whatsoever. The Transformer's architecture allows far greater parallel computation, the effectiveness of which its authors illustrate by training their network for just 12 hours on 8 P100 GPUs to handle the WMT 2014 English-to-German translation task, and achieving the best results to date.

Transformer's attention mechanism is based on key-value pair attention, which resembles the NTM attention approach in that the key is the means of retrieval of the learned values, stored in memory. In a similar manner to NTM, the output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

Transformer's attention unit is Multi-Head Attention, which is made up by repeated use of a basic element called a "Scaled Dot-Product Attention". The input consists of queries and keys of dimension d_k and values of dimension d_v .

We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values.