## Positional encoding
https://towardsdatascience.com/master-positional-encoding-part-i-63c05d90a0c3
https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

http://nlp.seas.harvard.edu/2018/04/01/attention.html#background

Attention

- an attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and outputs are all vectors
- 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
- one possible compatibility function is the 'scaled dot-product attention'
  - involves finding a dot product between a query vector and a key vector to get a weight for the corresponding value vector
- another example is 'additive attention'
  - this uses a feed forward network with a single hidden layer
  - Q: to do what exactly?

Positional encoding
What you want from a good position encoding
1. Deterministic: given a position, the encoding representing that position should always be the same.
2. Unique: no two positions should share an encoding
3. Distance metric: encoding(p2) - encoding(p1) should always yield the same answer, independent of the sequence data being looked at
4. Generalization: a model should be able to generate/learn position encodings for positions it has not encountered before

- should the position encoding be a property of the model?
- or something in the input?
- or...?