Attention Blocks

1. Tokenization (word → vector)

Embeds a token *w* into a higher dimensional embedding vector ***E***. This vector has yet to hold any contextual information.

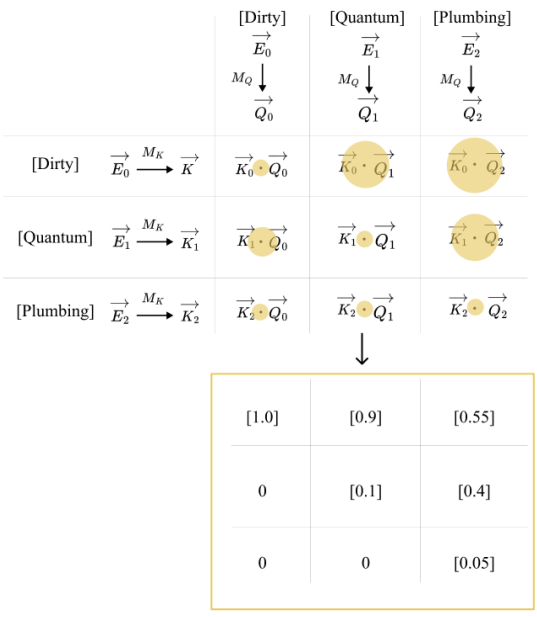
2. Attention Mechanism

A multi-headed attention block comprises many layered attention mechanisms (called single 'heads'). Each attention head adds some context to ***E***. Here is how it works:

Every head of attention has three matrices that contain learned weights. A query matrix M_Q , a key matrix M_K , and a value matrix M_V. The first matrix can be thought of as a question (ie. what is an adjective that precedes this noun?). The second matrix can be thought of as the answer (ie. does this word fit as an adjective for a noun?). The final matrix takes embedded vector ***E***s and translates them to a higher dimensional context space. These three matrices make up the parameters of the model that are populated during training on a large input sample that maps the context space.

Every ***E*** is multiplied by M_Q and M_K to generate a Query vector ***Q*** and a Key vector ***K***. A key matches a query well if they point in relatively the same direction in our context space, thus we dot product them.

Here is a one attention head for our adjective-noun example on the phrase: "Dirty Quantum Plumbing."



[1.0]	[0.9]	[0.55]
0	[0.1]	[0.4]
0	0	[0.05]

(Circles represent magnitudes, Dirty and Quantum are good adjective matches for the noun plumbing)

These values are then normalized to become column-stochastic. In addition, you never want later words in a sequence to influence previous words, as it defeats the purpose of trying to predict the next token. So you force this weights matrix to be upper triangular.

(I just made up some logical entries)

Now, To actually update each ***E*** with this weight matrix, you multiply the value matrix M_V by each word vector to obtain a vector ***V*** that represents a word's contextual direction in the higher dimensional context space. Then, multiply every ***V*** by its corresponding weight (influence) in the weight matrix.

Higher weighted words will therefore influence more. The final step is to sum up the column and update ***E.***

You do this for every column but here is an example of just "plumbing." We can see how dirty and quantum, which are adjectives of this type of plumbing, will greatly provide context to its embedding.

		[Dirty] $\vec{E}_0$	[Quantum] $\vec{E}_1$	[Plumbing] $\vec{E}_2$		
[Dirty]	$\vec{E}_0 \xrightarrow{M_V} \vec{V}_0$	[1.0]	[0.9]	[0.55] * $\vec{V}_0$	=	$0.55\vec{V}_0$
[Quantum]	$\vec{E}_1 \xrightarrow{M_V} \vec{V}_1$	0	[0.1]	[0.4] * $\vec{V}_1$	=	$+ 0.4\vec{V}_1$
[Plumbing]	$\vec{E}_2 \xrightarrow{M_V} \vec{V}_2$	0	0	[0.05] * $\vec{V}_2$	=	$+ 0.05\vec{V}_2$
					=	$\Delta\vec{E}_2$

$$\vec{E}_2' = \vec{E}_2 + \Delta\vec{E}_2$$