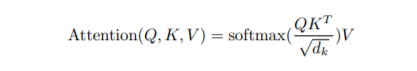
**SELF ATTENTION**

Self-Attention mechanism is primarily employed in NLP tasks but very recently has been adopted by computer vision field to develop Avant-Garde range of models that outperform old methods.

Let’s say we have an input image that can be turned into a sequence of pixels. So, an 8x8 image now becomes a linear sequence of dimension 1x64. Each image has **Queries, Keys** and **Value** vector of size 1x64. These vectors are basically the pixel sequence with different weights associated with it. Next we calculate dot product of **Queries** and **Keys** and divide by root of dimension of the ‘**Keys’** vector. The result is passed through SoftMax to get an attention score. The division is done to get more stable gradients and SoftMax is used for normalization. Formula given below:



So, if we have 4 images, the **Query**, **Key**, **Value** would be matrices of size 4x64 each. In practice, that’s how it works. After softmax, these matrices would produce multiple scores creating an attention map. Next, the softmax scores are multiplied to each **value** vector and finally all summed up to give the output of the attention layer.

