Vision Transformers

John Chiasson[1] *and Ruthvik Vaila*[2]
Boise State University[1] *and LambWeston Holdings Inc.*[2]

Data: CIFAR $b \times 3 \times 32 \times 32$

same_conv_layer_stack
`n_conv_layers` (default is 1)
Stack of `n_conv_layers` convolution layers with each layer of the form
input channels: 3
output channels: 3
padding: 2
kernal: $3 \times 5 \times 5$

conv_proj_layer
Embed the image where $e = 512$ is the embedding size.
To do this go from 3 32×32 channels to 512 8×8 channels.
That is, apply 512 kernels of size $3 \times 4 \times 4$ with a **stride of 4**
to convert $b \times 3 \times 32 \times 32$ channels/maps to $b \times 512 \times 8 \times 8$ channels/maps.

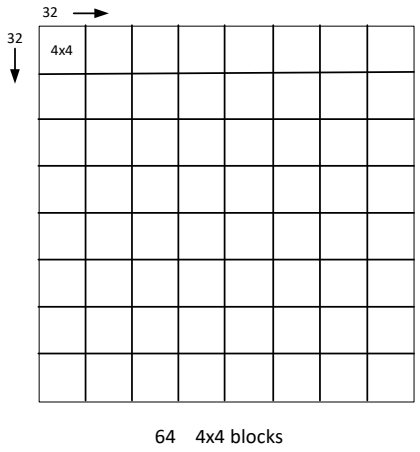

Figure 1: Each image is now represented as a "sentence" of $8 \times 8 = 64$ words with each word in $\mathbb{R}^{512}$.

Flatten height & width and rearrange
Flatten $b \times 512 \times 8 \times 8$ channels/maps to $b \times 512 \times 64$ channels/maps.
Rearrange $b \times 512 \times 64$ channels/maps to $b \times 64 \times 512$ channels/maps.

For each image include class token of size $b \times 1 \times 512$.
(During training only this part is sent to the classifier.)
Concatenate the class token $b \times 1 \times 512$ with $b \times 64 \times 512$ so that the batch images have shape

$$b \times 65 \times 512.$$

That is, a "sentence" contains 65 words and each word is embedded (represented) as a vector in $\mathbb{R}^{512}$.

Position Encoding
Make 65 position tokens with have size 65×512 to encode the posiion of the 65 "words" of the image.
All 65 position tokens are intialized to the same random 65×512 tensor.
Add the position tokens to the word embeddings (done in the code by broadcasting).

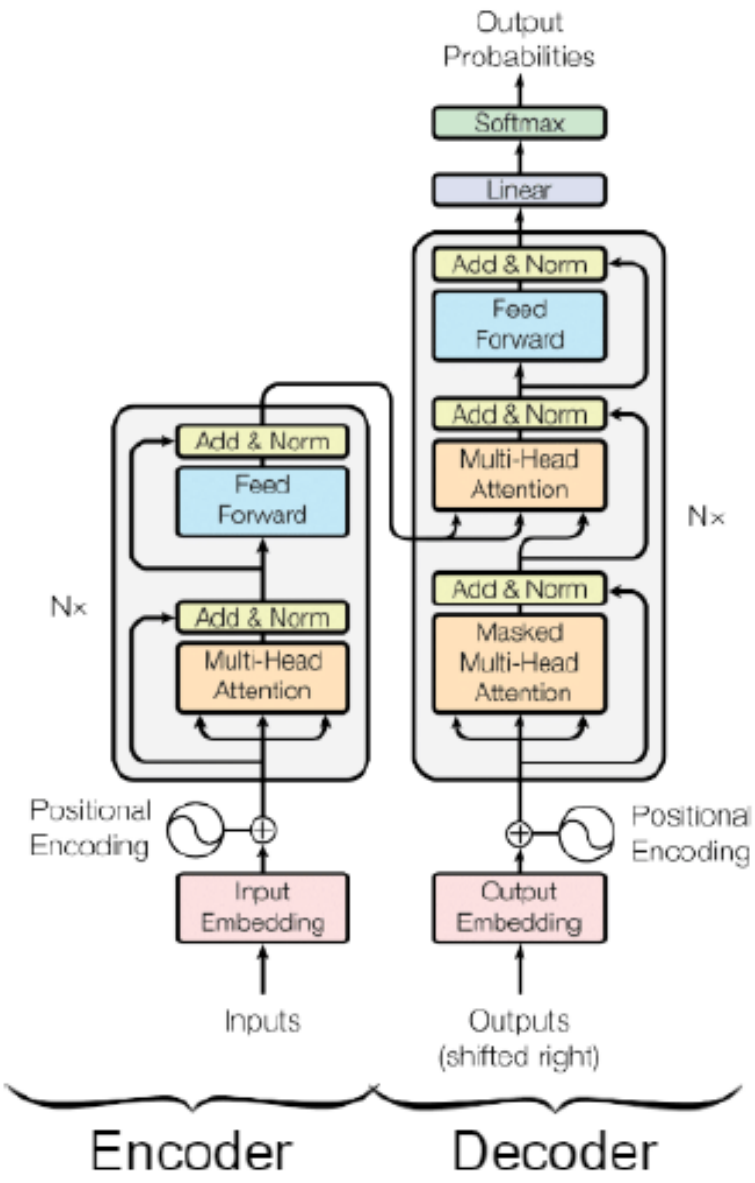

Figure 2: From *Attention Is All You Need* at *https://arxiv.org/abs/1706.03762*