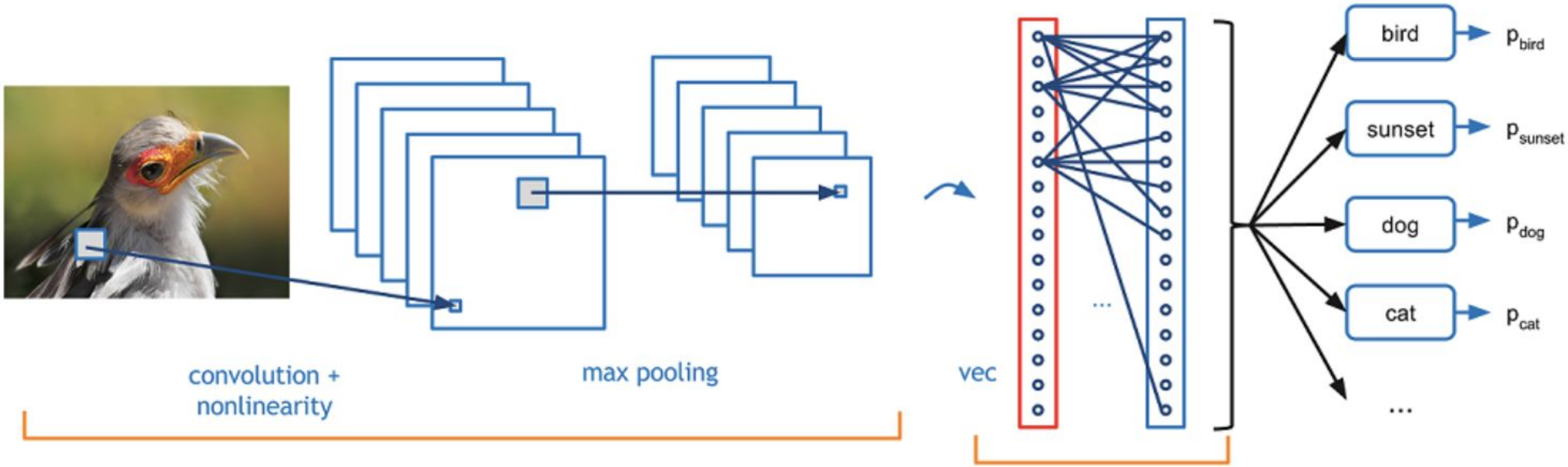
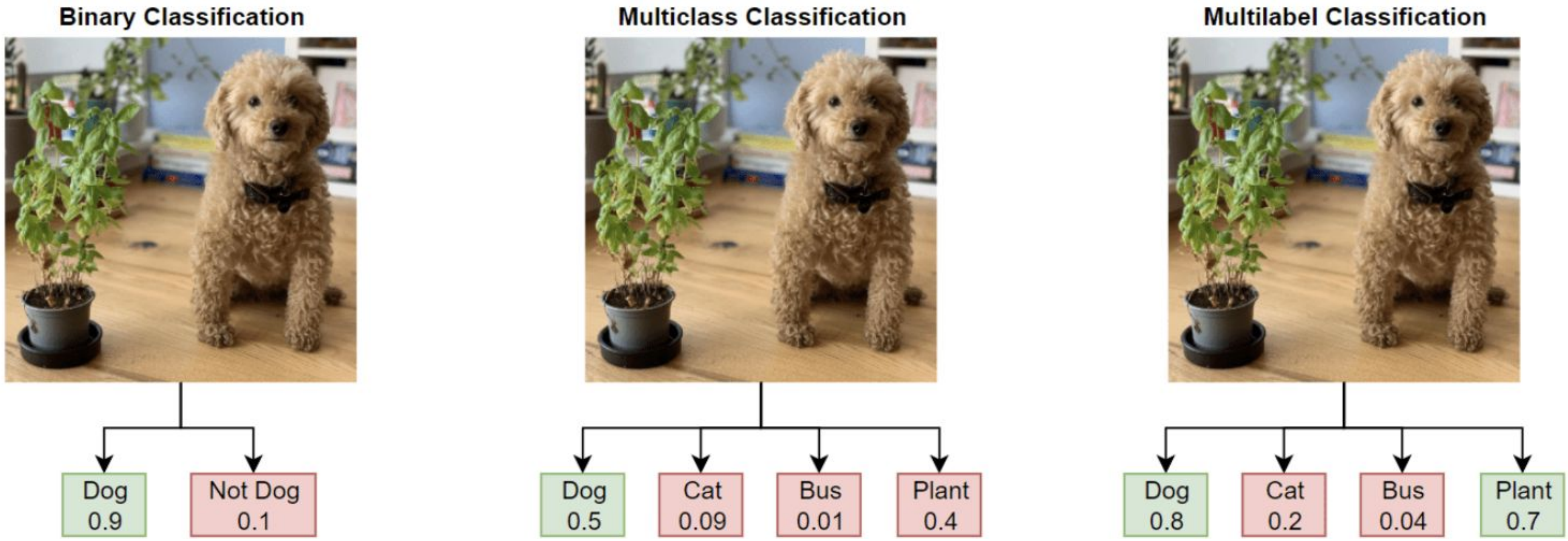
Image Classification

Vishal Reddy Mandadi

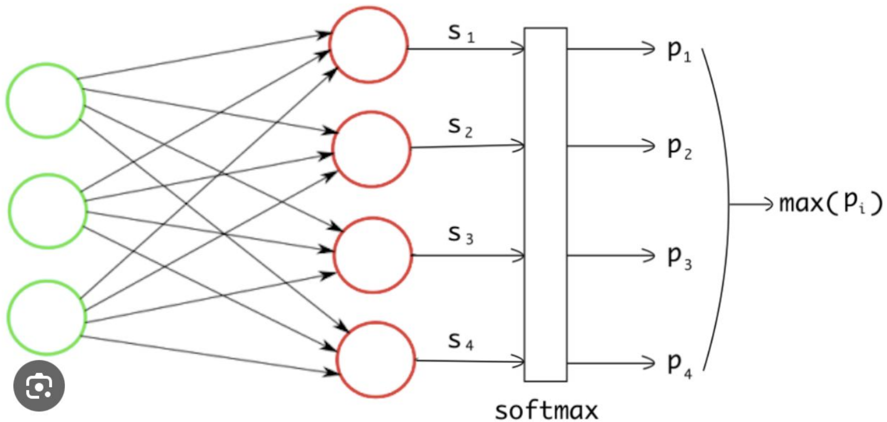
Image Classification



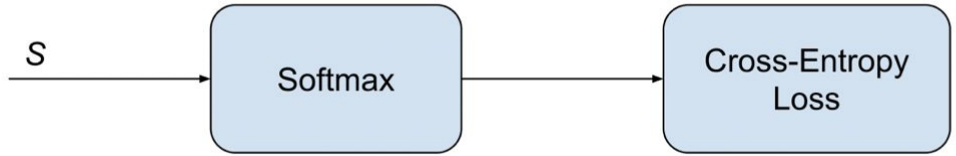
Binary vs Multi-class vs Multi-label Classification



Categorical Cross Entropy Loss



$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$



$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$

Evolution of Image Classification

LeNet (1998)

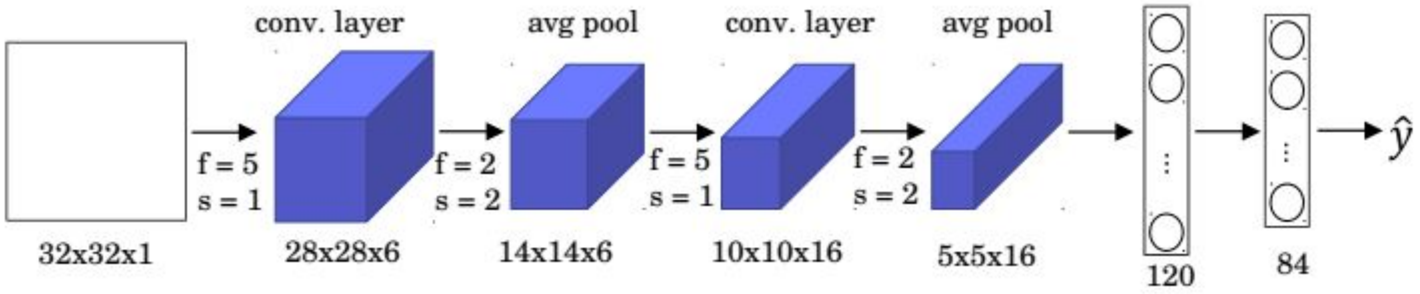


Figure 3: LeNet-5 neural network. Around 60k parameters.

AlexNet (2012)

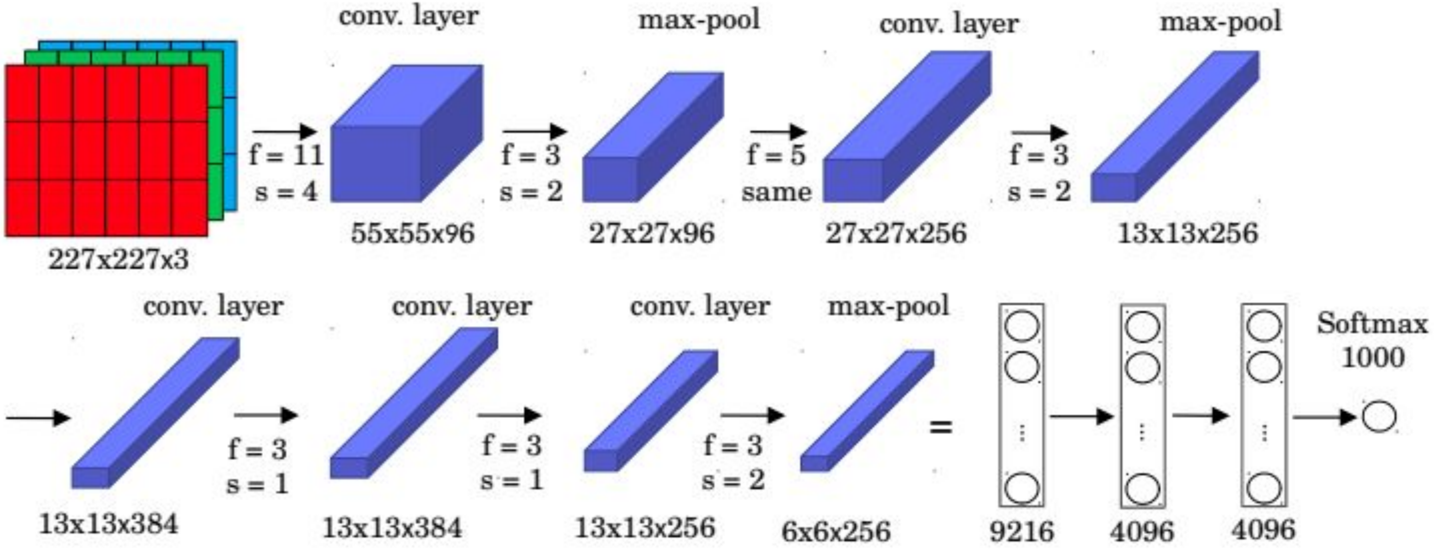


Figure 4: AlexNet neural network. Around 60 million parameters.

VGG-16 and VGG-19 (2014)

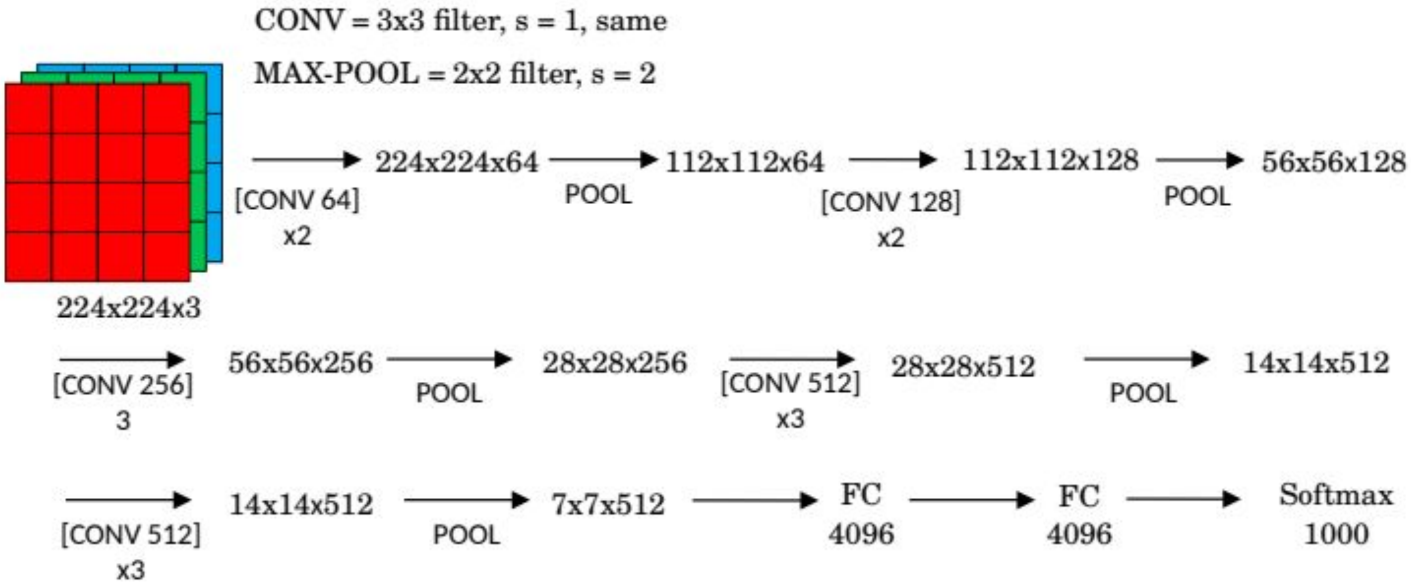


Figure 5: VGG-16. Around 138 million parameters.

Inception Net (GoogLeNet) (2014)

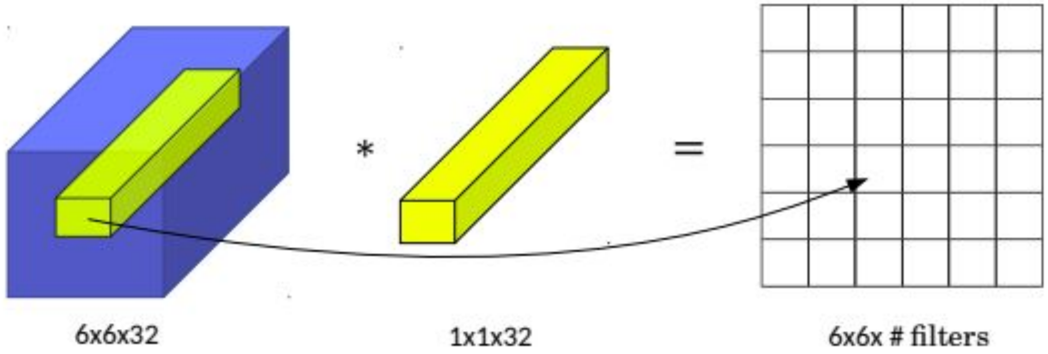


Figure 11: 1×1 convolution. The filter has size $1 \times 1 \times 32$ elements (weights). The number of filters correspond to the number of channels of the output.

Inception Net

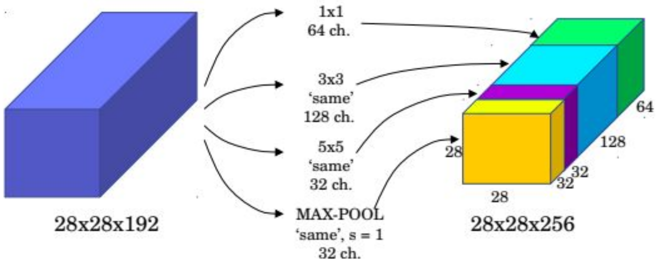


Figure 13: Inception module with 1×1, 3×3, 5×5 convolutional layers, and max-pooling.

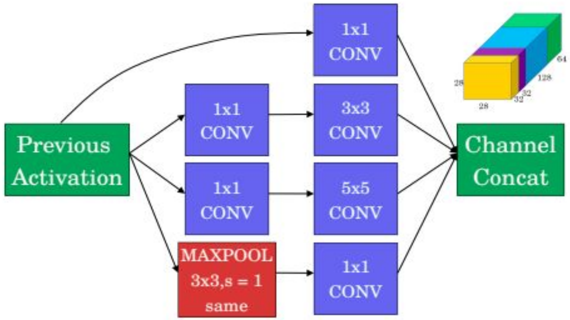


Figure 14: Inception module with 1×1, 3×3, 5×5 convolutional layers, and max-pooling with intermediate 1×1 convolutions.

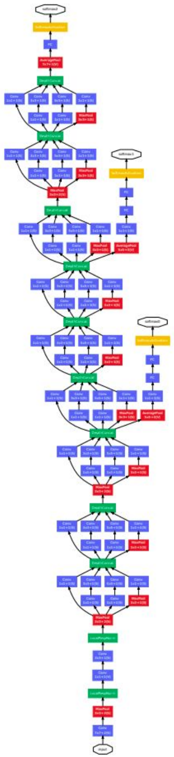


Figure 15: GoogLeNet network with all the bells and whistles [7].

ResNets (2015)

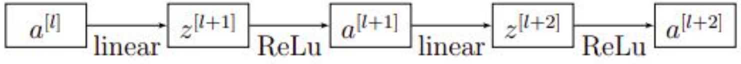


Figure 6: Plain network structure for layers l to $l+2$.

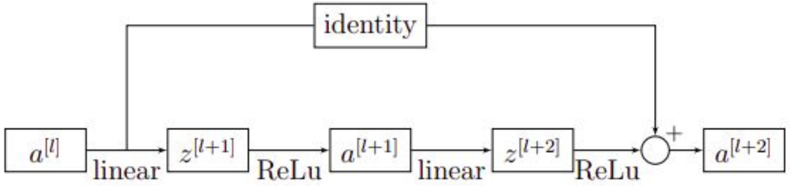


Figure 7: Residual network structure for layers l to $l+2$.

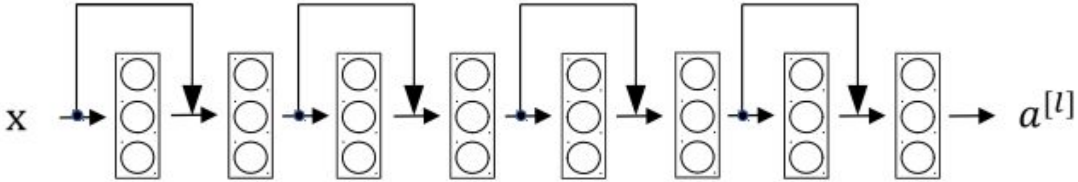
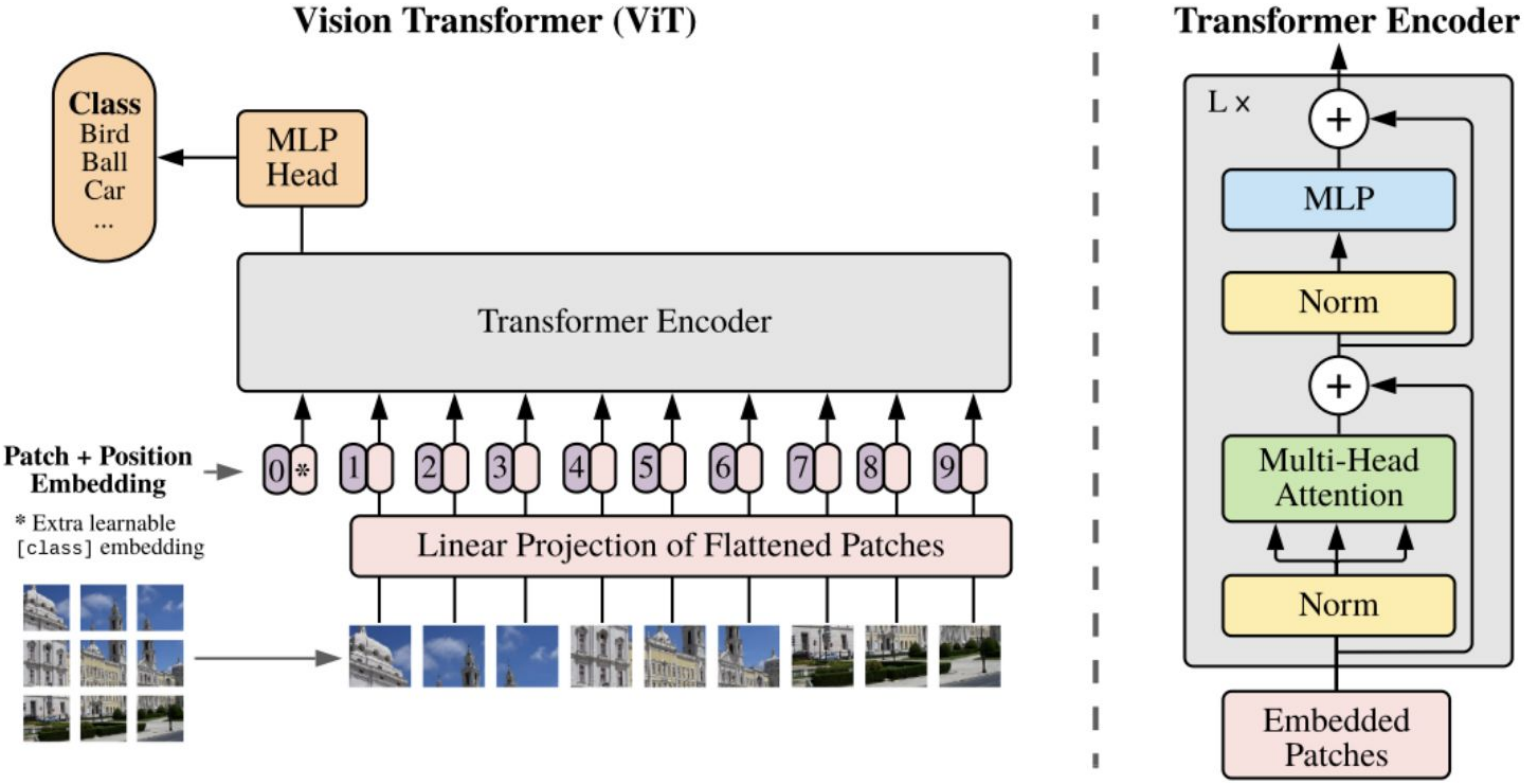


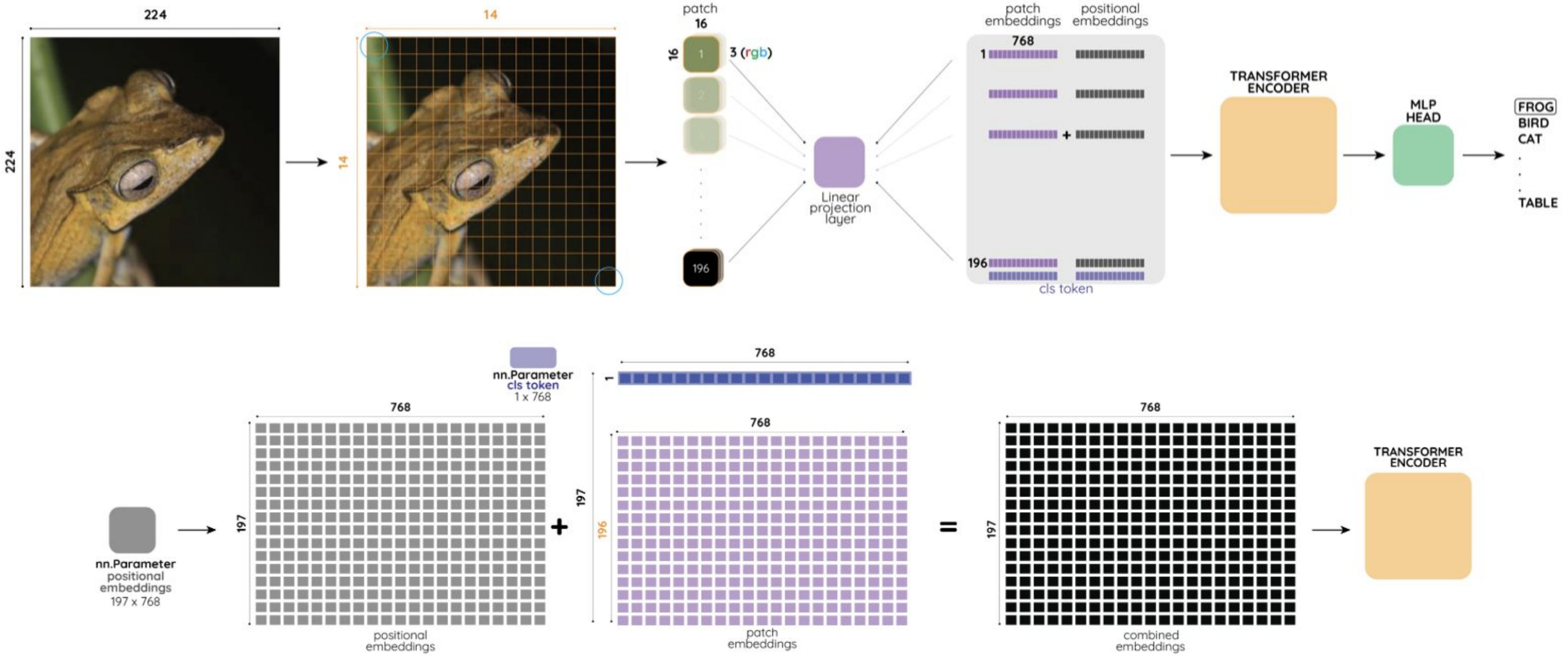
Figure 8: Residual network structure for layers l to $l+2$.

ViT (2020)

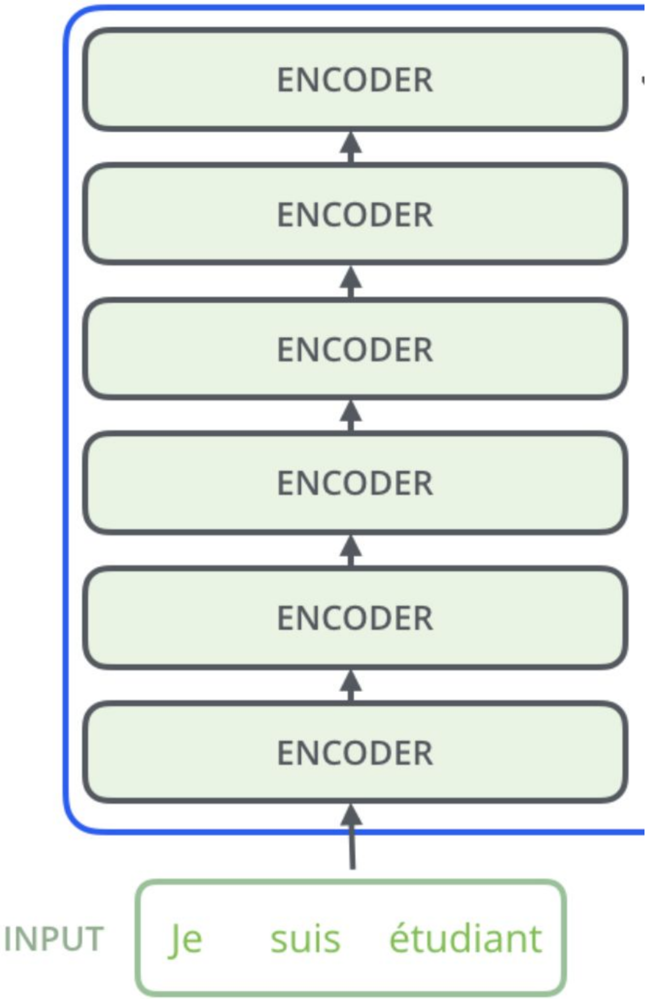
Overall Architecture (ViT)



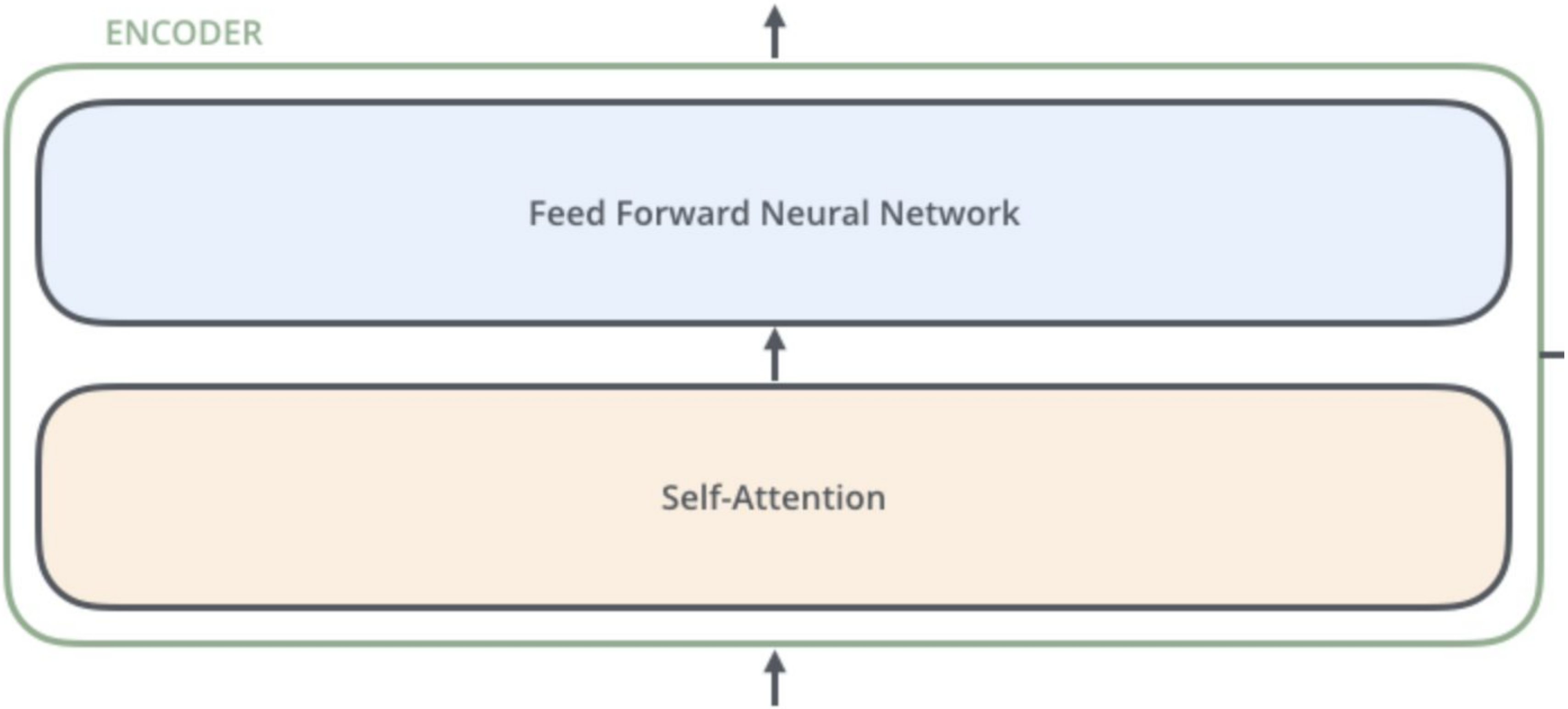
ViTs


224
224
14
14
patch
16
16
3 (rgb)
1
2
196
Linear projection layer
patch embeddings
positional embeddings
768
1
+
196
cls token
TRANSFORMER ENCODER
MLP HEAD
FROG
BIRD
CAT
TABLE
nn.Parameter
cls token
1 x 768
768
1
768
197
768
196
768
197
nn.Parameter
positional embeddings
197 x 768
197
+
=
positional embeddings
patch embeddings
combined embeddings
TRANSFORMER ENCODER

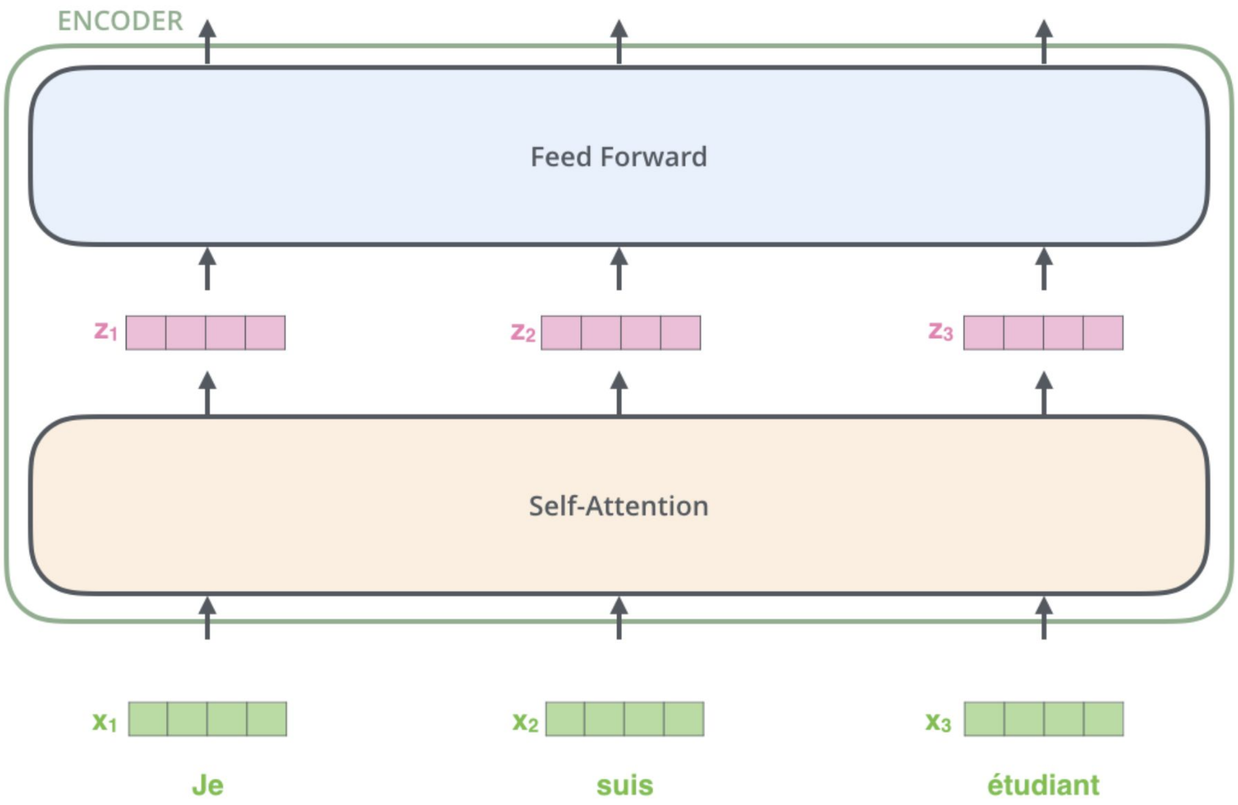
Transformer Encoder



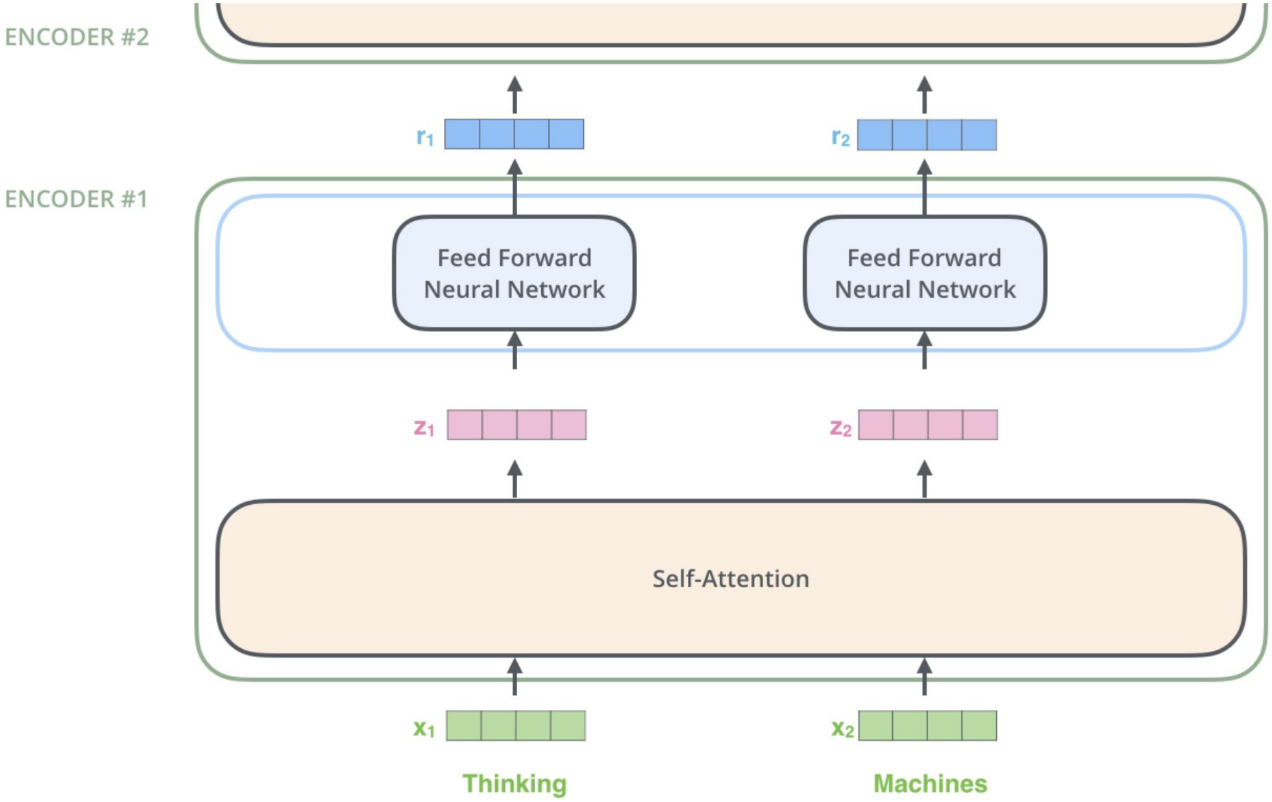
Transformer Encoder



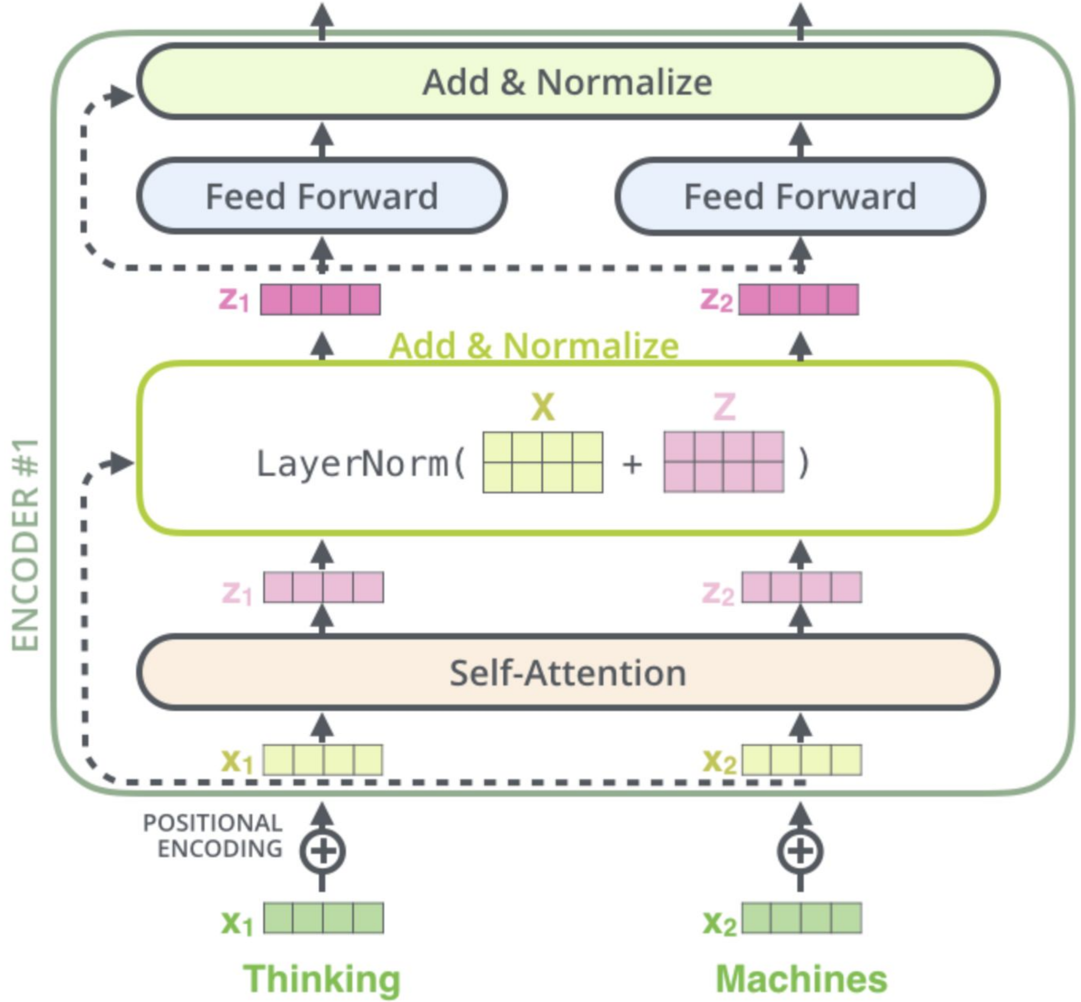
Transformer Encoder



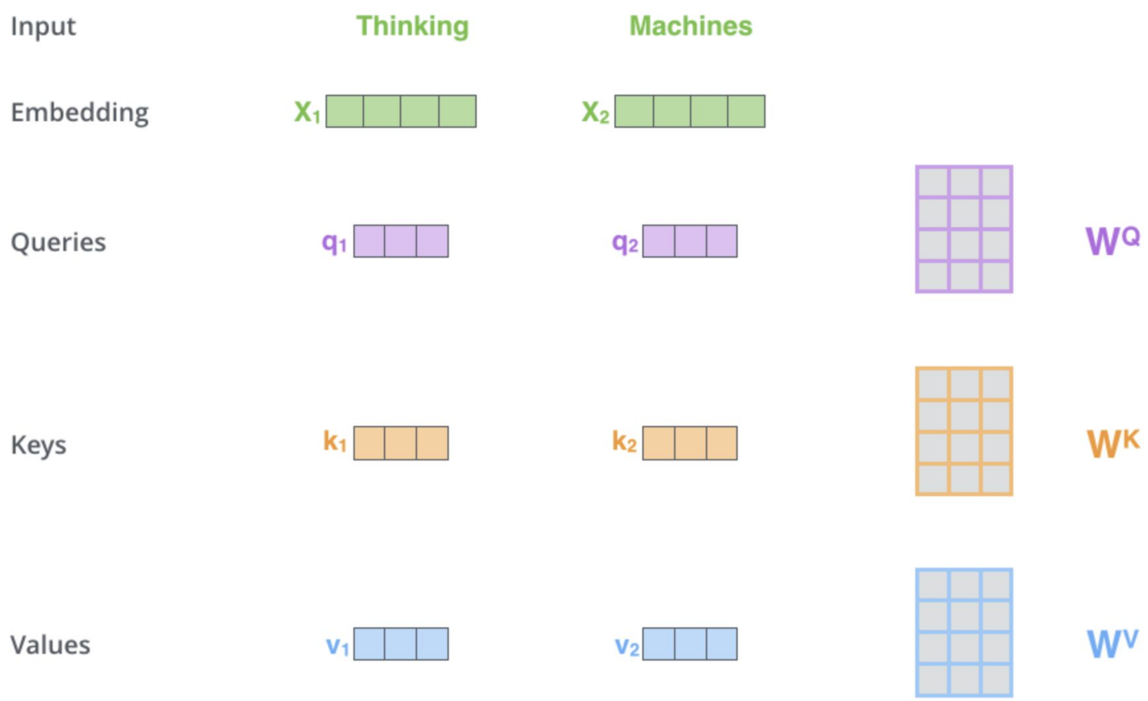
Transformer Encoder



Transformer Encoder

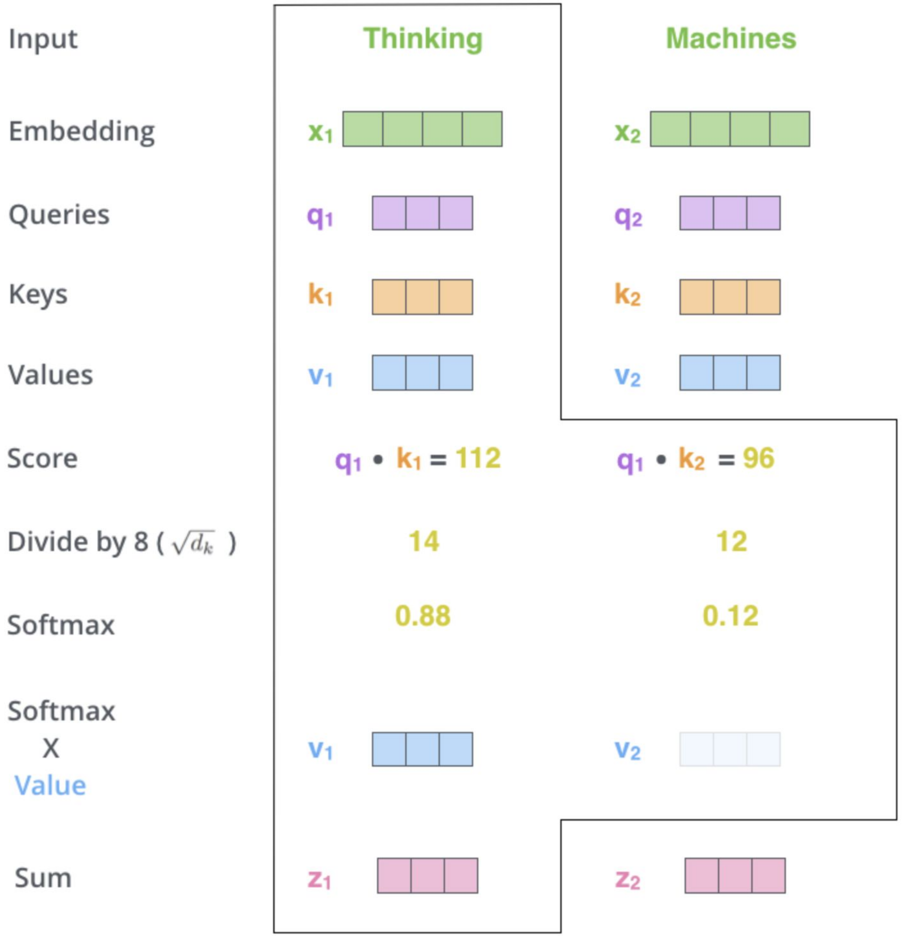


Transformer Encoder

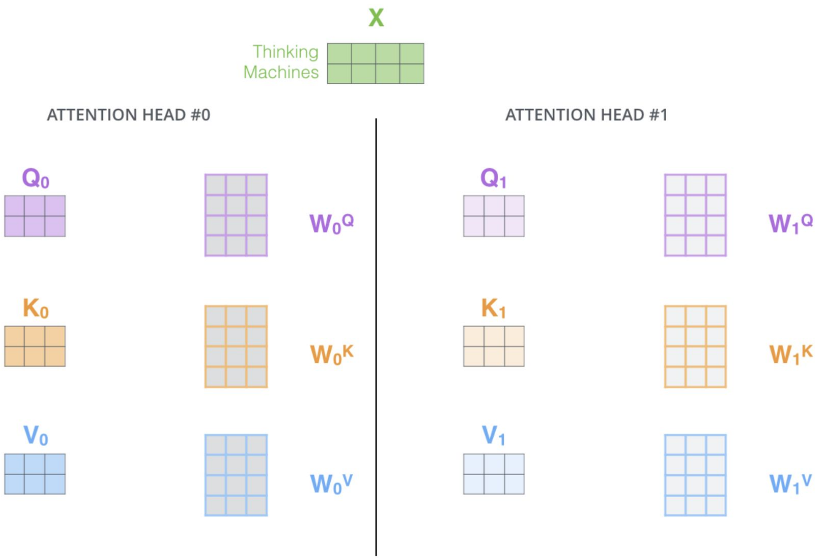


Multiplying x1 by the WQ weight matrix produces q1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

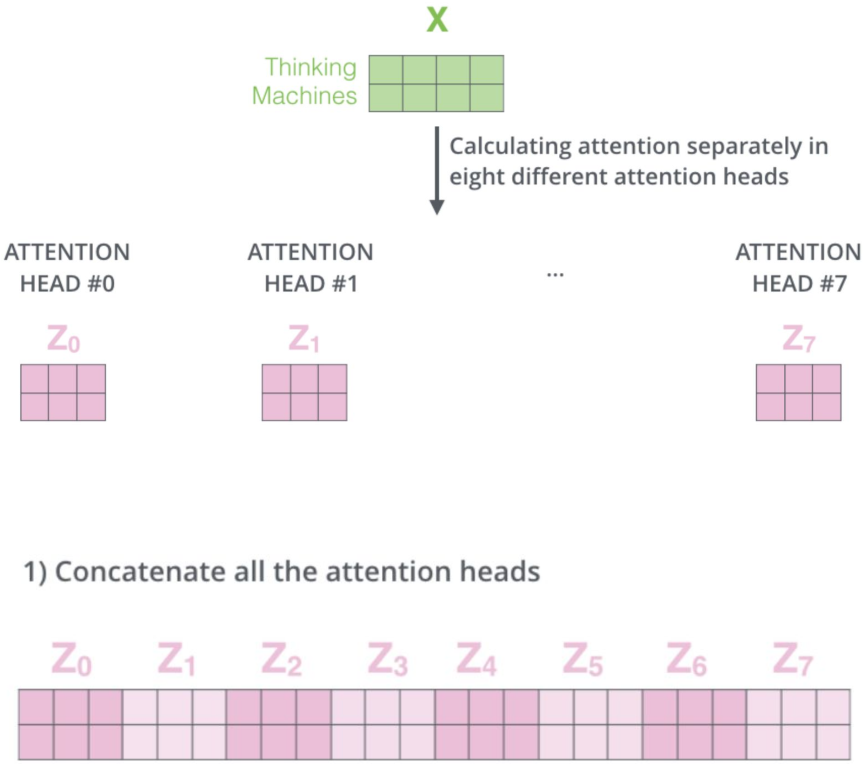
Transformer Encoder



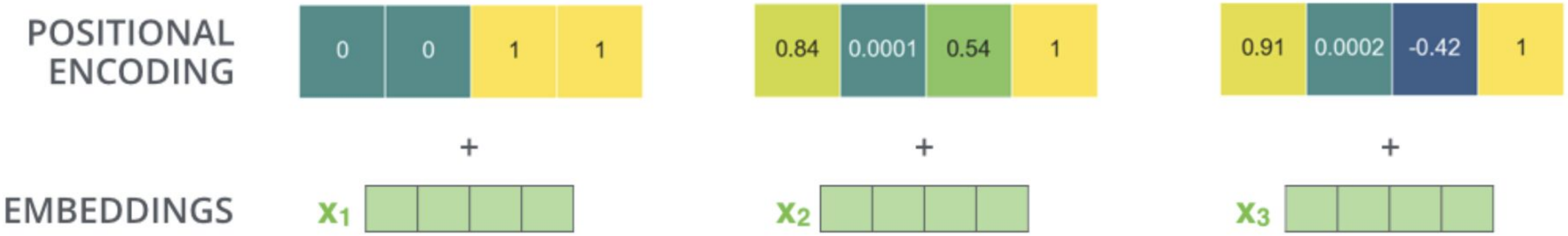
Multi-head Attention



With multi-headed attention, we maintain separate Q/K/V weight matrices for each head resulting in different Q/K/V matrices. As we did before, we multiply X by the WQ/WK/WV matrices to produce Q/K/V matrices.



Positional Encodings



Resources

1. Blog on Attention - Attention? Attention! | Lil'Log
2. A wonderful blog on Transformers - The Illustrated Transformer – Jay Alammar
3. Blog on Vision Transformers - Vision Transformer
4. Original Paper on Transformers in NLP (2017) - [1706.03762] Attention Is All You Need
5. Original Paper on Vision Transformers(2020) - [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

The End