Distinguished Engineering

Transformer 5/5
- Transformer, the dark knight

•••

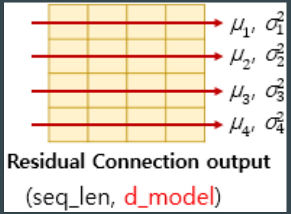
BW

Plan

- Prologue, seq2seq
- **Attention, please**
- Transformer, a new hope
- Transformer, revenge of the fallen
- Transformer, vision

Transformer

- Residual connection, Layer Normalization

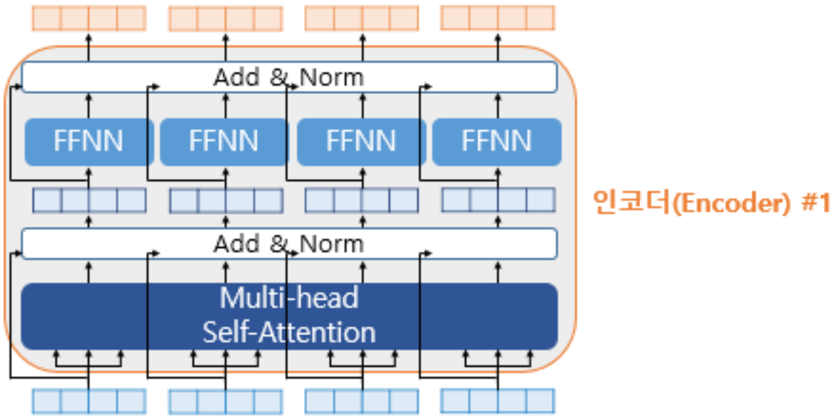


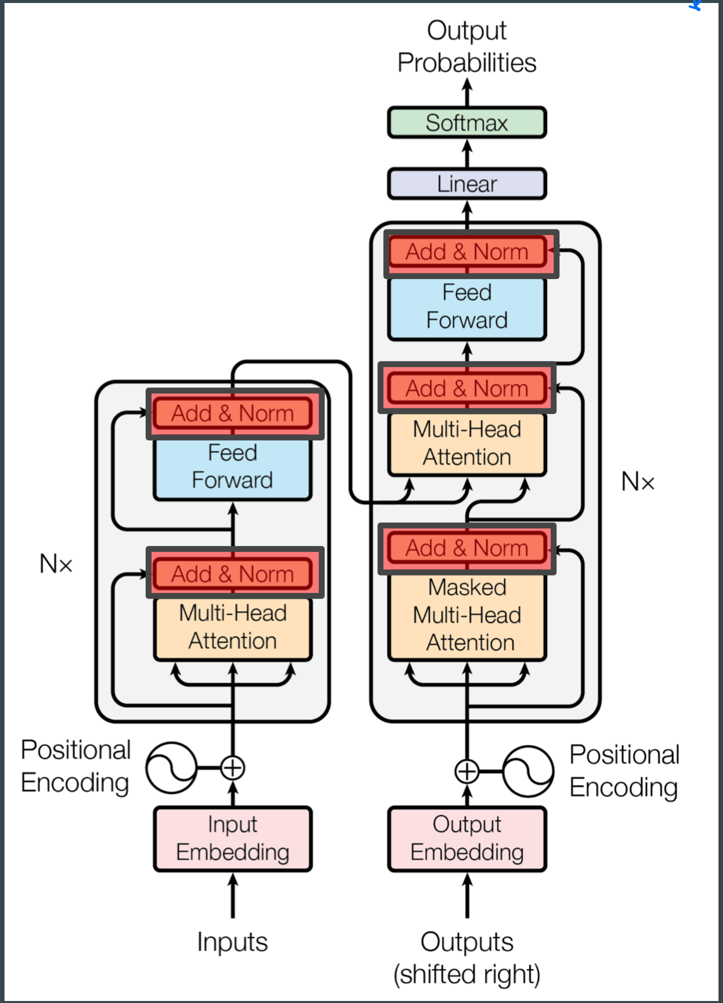
$$ln_i = LayerNorm(x_i)$$

$$\hat{x}_{i,k} = \frac{x_{i,k} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

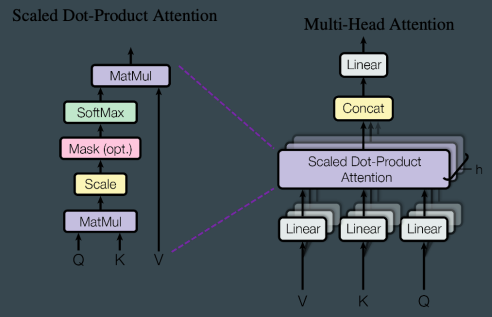
$$ln_i = \gamma\hat{x}_i + \beta = LayerNorm(x_i)$$

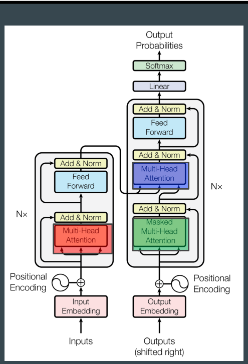
γ	1	1	1	1
β	0	0	0	0



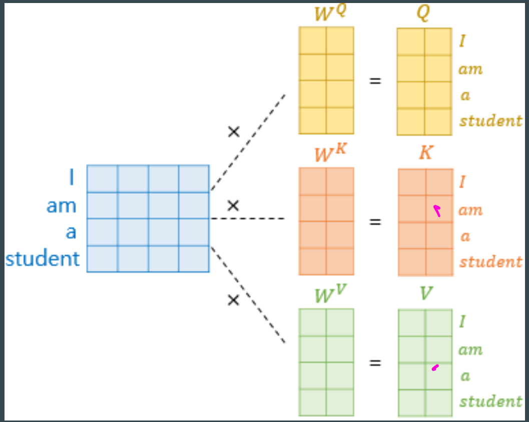


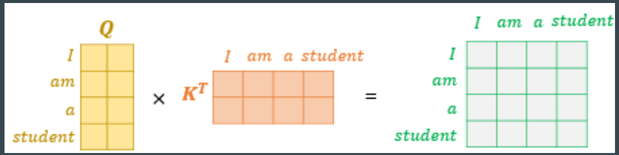
Transformer

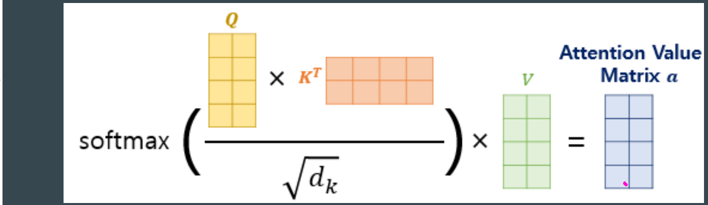




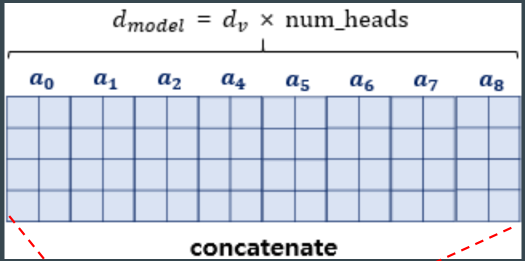
- Multi-head Attention

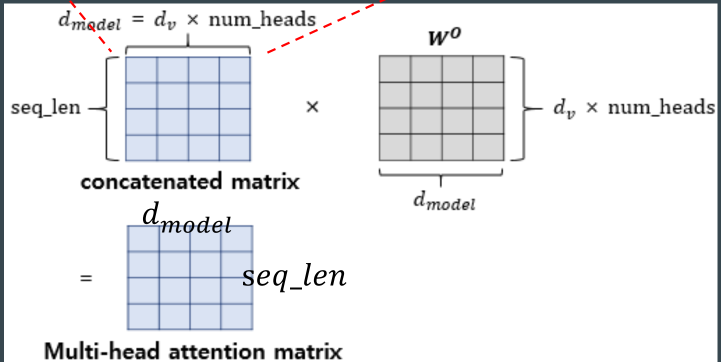




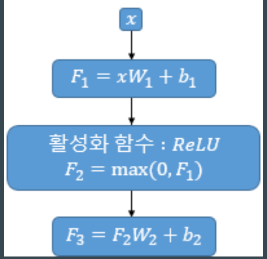


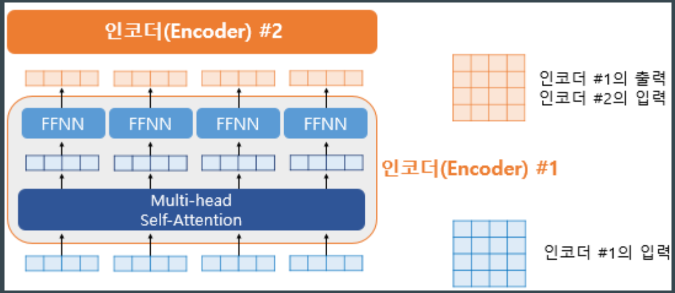
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



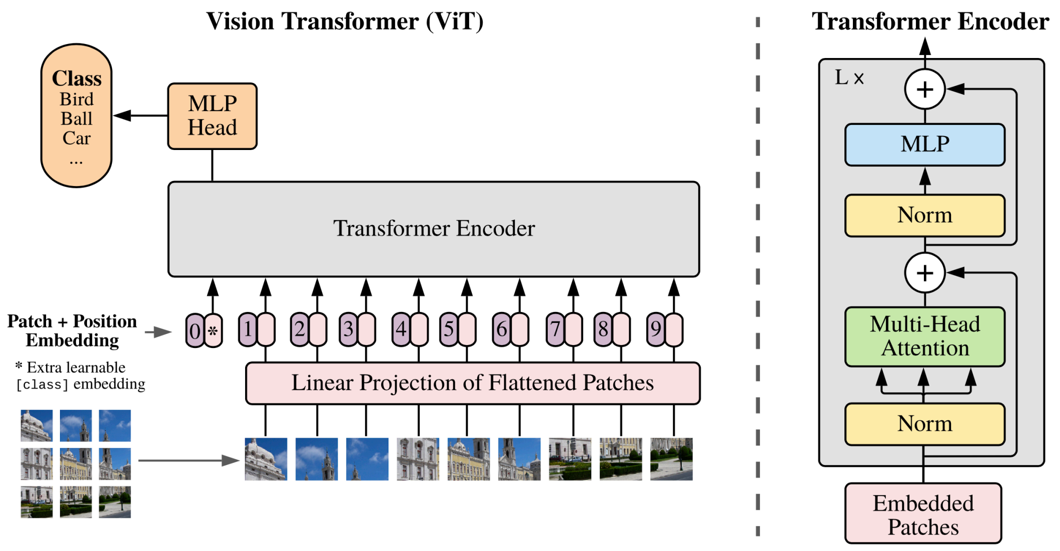


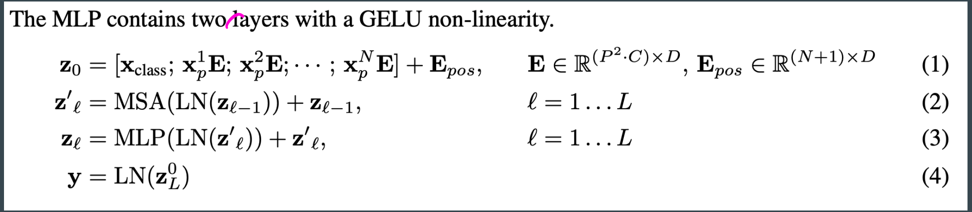
- FFNN





An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR2021





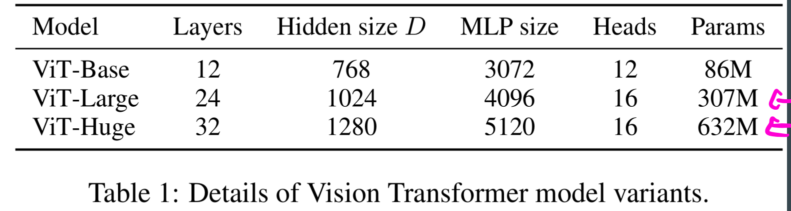
The MLP contains two layers with a GELU non-linearity.

$$\mathbf{z}_0 = [\mathbf{x}_\text{class}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C)\times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{1}$$

$$\mathbf{z'}_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0) \tag{4}$$



Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

ViT

- +
  - Simple (almost) architecture & well-proved performance
  - Scalability
  - Less Training Time
  - Excellent performance
- -
  - Less inductive bias —> requires more data!!
    - Less data, less performance

?

- Order
- Embedding
- Activation
  - GELU vs ReLU

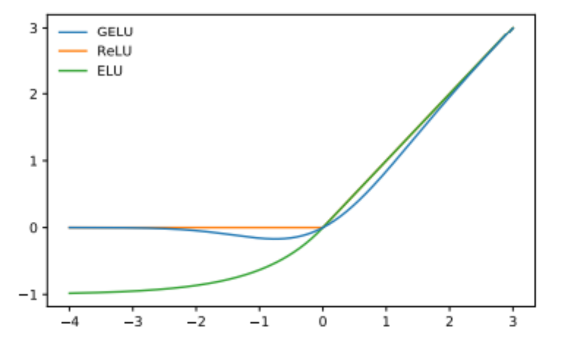
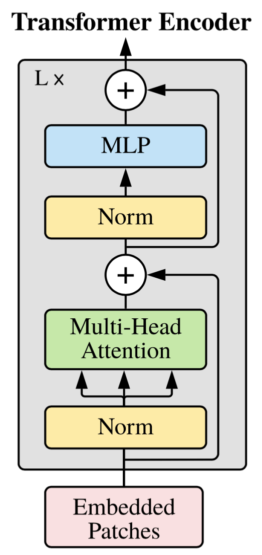
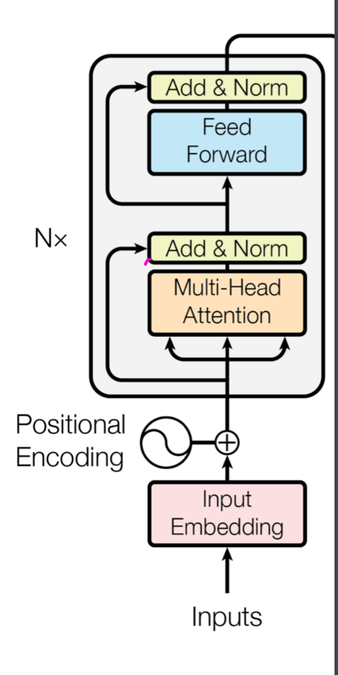


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}(x/\sqrt{2})\right]$$



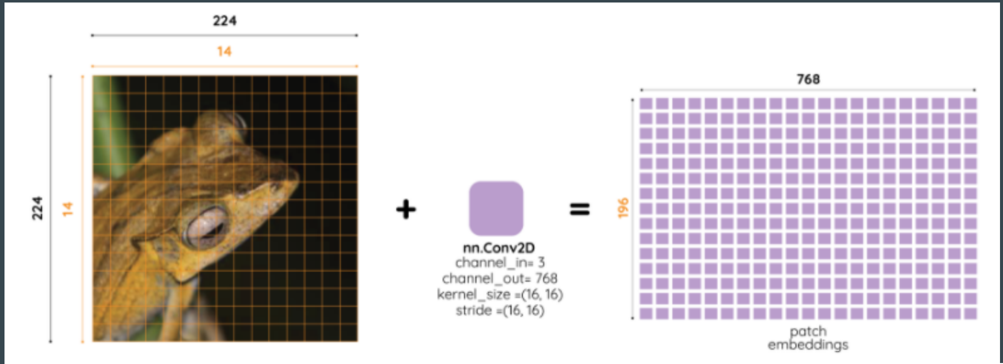
ViT

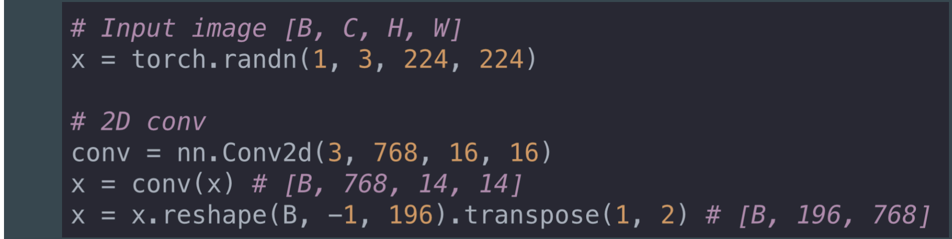


Transformer

ViT

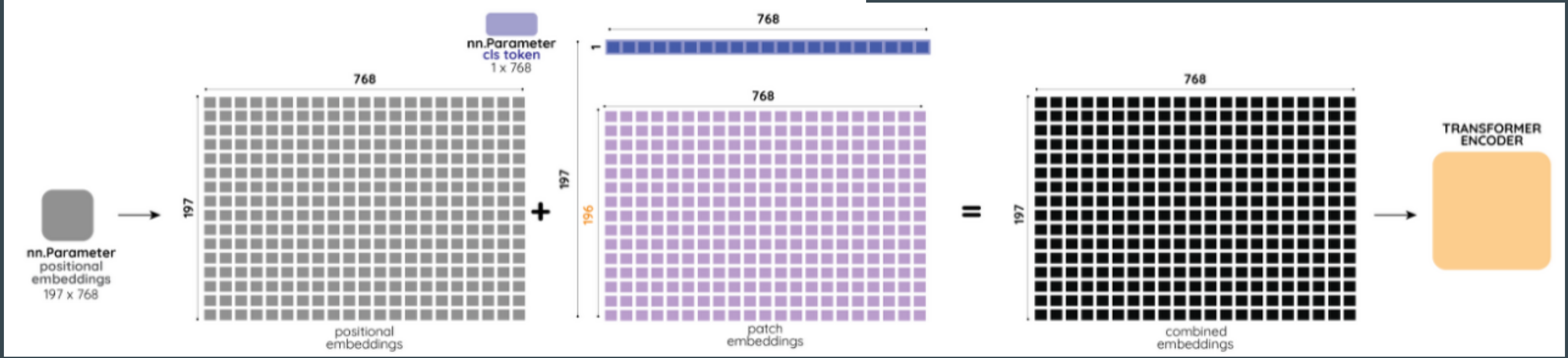
- Positional Embedding





```
# Input image [B, C, H, W]
x = torch.randn(1, 3, 224, 224)

# 2D conv
conv = nn.Conv2d(3, 768, 16, 16)
x = conv(x) # [B, 768, 14, 14]
x = x.reshape(B, -1, 196).transpose(1, 2) # [B, 196, 768]
```



Everything can be found in

https://wikidocs.net/book/2155

https://github.com/ukairia777/tensorflow-nlp-tutorial

https://hongl.tistory.com/232

https://dev-woong.tistory.com/38