

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

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

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \quad \ell = 1 \dots L \quad (2)$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad \ell = 1 \dots L \quad (3)$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \quad (4)$$

Standard **qkv** self-attention (SA, Vaswani et al. (2017)) is a popular building block for neural architectures. For each element in an input sequence  $\mathbf{z} \in \mathbb{R}^{N \times D}$ , we compute a weighted sum over all values  $\mathbf{v}$  in the sequence. The attention weights  $A_{ij}$  are based on the pairwise similarity between two elements of the sequence and their respective query  $\mathbf{q}^i$  and key  $\mathbf{k}^j$  representations.

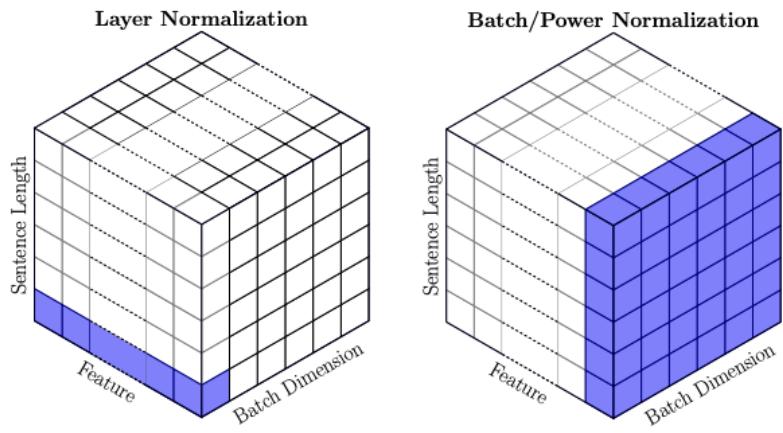
$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z} \mathbf{U}_{qkv} \quad \mathbf{U}_{qkv} \in \mathbb{R}^{D \times 3D_h}, \quad (5)$$

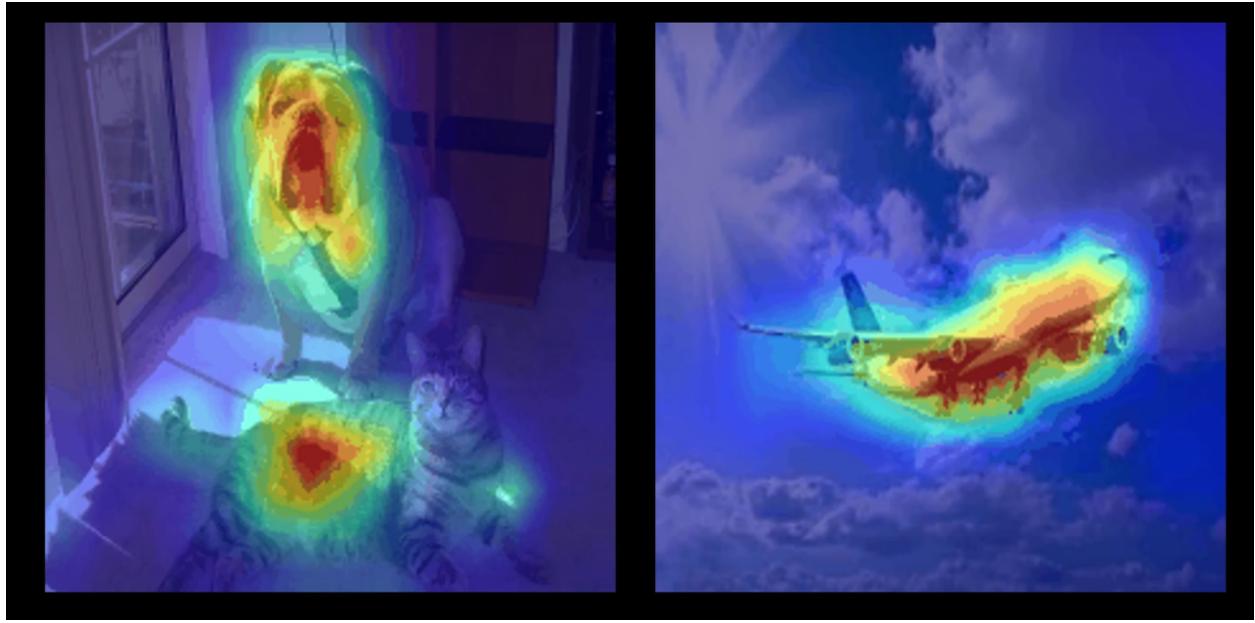
$$A = \text{softmax} \left( \mathbf{q} \mathbf{k}^\top / \sqrt{D_h} \right) \quad A \in \mathbb{R}^{N \times N}, \quad (6)$$

$$\text{SA}(\mathbf{z}) = \mathbf{A} \mathbf{v}. \quad (7)$$

Multihead self-attention (MSA) is an extension of SA in which we run  $k$  self-attention operations, called “heads”, in parallel, and project their concatenated outputs. To keep compute and number of parameters constant when changing  $k$ ,  $D_h$  (Eq. 5) is typically set to  $D/k$ .

$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(z); \text{SA}_2(z); \dots; \text{SA}_k(z)] \mathbf{U}_{msa} \quad \mathbf{U}_{msa} \in \mathbb{R}^{k \cdot D_h \times D} \quad (8)$$





	Strong inductive bias (translation invariance)		No strong inductive bias
	Not too data-hungry		Data-hungry
	Hierarchical structure (receptive field)		Global structure (attention)
CNNs		ViT	