$$\begin{aligned} \boldsymbol{H}^w_{L-F} &= \text{Encoder}_w(\boldsymbol{H}^w_0) \\ \boldsymbol{H}^v_{L-F} &= \text{Encoder}_v(\boldsymbol{H}^v_0) \\ \boldsymbol{H}^w_L &= \text{Encoder}_s(\boldsymbol{H}^w_{L-F}) \\ \boldsymbol{H}^v_L &= \text{Encoder}_s(\boldsymbol{H}^v_{L-F}) \\ \boldsymbol{H}^w_l &= \left[\boldsymbol{w}_l^{[\texttt{T\_CLS}]}, \boldsymbol{w}_l^1, ..., \boldsymbol{w}_l^M, \boldsymbol{w}_l^{[\texttt{T\_SEP}]}\right] \\ \boldsymbol{H}^v_l &= \left[\boldsymbol{v}_l^{[\texttt{I\_CLS}]}, \boldsymbol{v}_l^1, ..., \boldsymbol{v}_l^N\right] \end{aligned} \tag{1}$$

- with $l \in \{1, ..., L - F, ..., L\}$
- we define $\boldsymbol{H}^w_L$ as the final output of the student model for the caption, and $\boldsymbol{H}^v_L$ as the final output of the student model for the image, with $\boldsymbol{H}^w_L \in \mathbb{R}^{(M+2)\times D}$ and $\boldsymbol{H}^v_L \in \mathbb{R}^{(N+1)\times D}$

Image-Text Matching with Feature Fusion