

---

[Supplementary Material]

## A ATTENTION

For the two inputs  $\mathbf{H}^1$  and  $\mathbf{H}^2$  of the attention module, we find the importance of  $\mathbf{h}^1$  and  $\mathbf{h}^2$  for each node. For example, for embedding  $\mathbf{h}^1$ , first we transform each embedding through a nonlinear transformation:

$$\hat{\mathbf{h}}_i^1 = \tanh(\mathbf{W} \cdot (\mathbf{h}_i^1)^T + \mathbf{b})$$

where  $\mathbf{W}$  is the weight matrix and  $\mathbf{b}$  is the bias. Then we use one shared attention vector  $\mathbf{q}$  to get the attention value  $w_i^1$ :

$$w_i^1 = \mathbf{q} \cdot \hat{\mathbf{h}}_i^1$$

Then the final attention coefficient can be calculated as:

$$\alpha_i^1 = \text{softmax}(w_i^1)$$

And larger  $\alpha_i^1$  implies the corresponding embedding is more important. Then the final output can be denoted as:

$$\mathbf{H} = \alpha^1 \cdot \mathbf{H}^1 + \alpha^2 \cdot \mathbf{H}^2$$