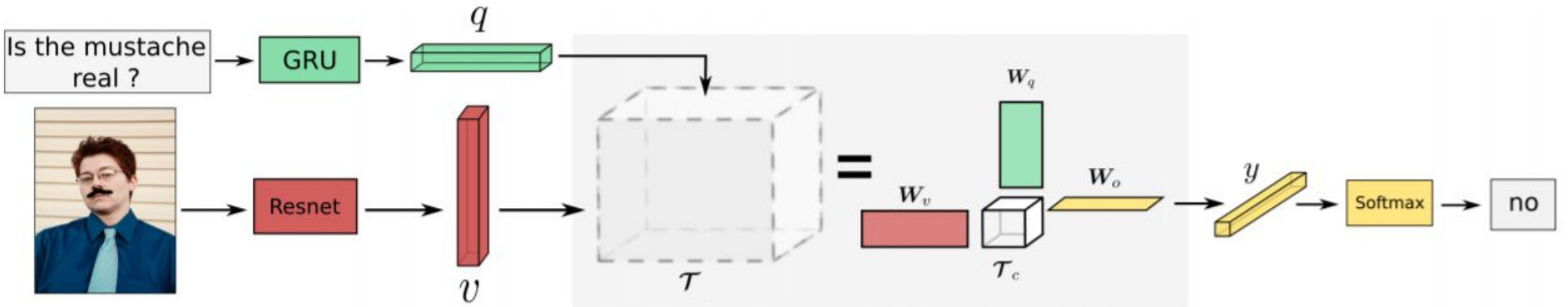

Figure 2: MUTAN fusion scheme for global Visual QA. The prediction is modeled as a bilinear interaction between visual and linguistic features, parametrized by the tensor $\mathcal{T}$. In MUTAN, we factorise the tensor $\mathcal{T}$ using a Tucker decomposition, resulting in an architecture with three intra-modal matrices $\boldsymbol{W}_q$, $\boldsymbol{W}_v$ and $\boldsymbol{W}_o$, and a smaller tensor $\mathcal{T}_c$. The complexity of $\mathcal{T}_c$ is controlled *via* a structured sparsity constraint on the slice matrices of the tensor.