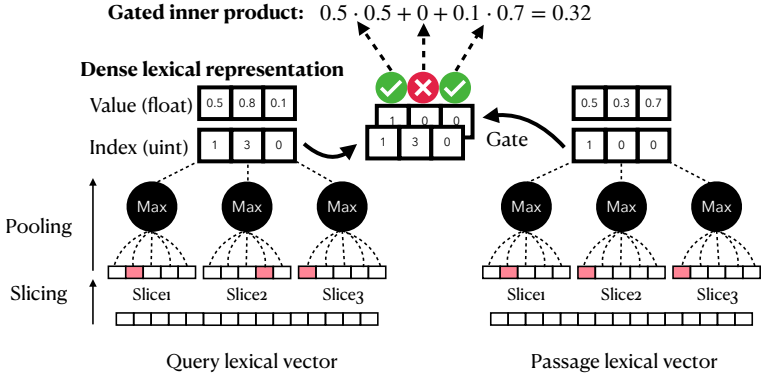

Fig. 1. Illustration of densified lexical representations (DLRs) and gated inner product (GIP). Both query and passage lexical representations are fixed-width vectors where the number of dimensions is equal to the vocabulary size. Our approach first groups these high-dimensional vectors into M slices, each with N dimensions (e.g., $M = 3$, $N = 5$ here). For each slice, the maximum value is selected. These values from the original vector (Value) and their positions in each slice (Index) are recorded separately. When computing the gated inner product between two vectors, only the dimensions with the same index are considered.