1. NMT with RNNs

(g) The mask is used to mark the <pad> term and to set the corresponding value in e_t to -inf, which sets the corresponding value in α_t to zero (softmax). Otherwise, the decoder will use the information of <pad> term and maybe predict <pad> terms as outputs, which is not expected.

(i) 22.

(j) ① Dot product

- fast; no additional storage memory
- maybe too easy to get the information of s_t and h_i

② Multiplicative attention

- transition between ① and ③

③ Additive attention

- able to catch the information between value vectors and queries
- more storage memory; slow; more hyperparameters.