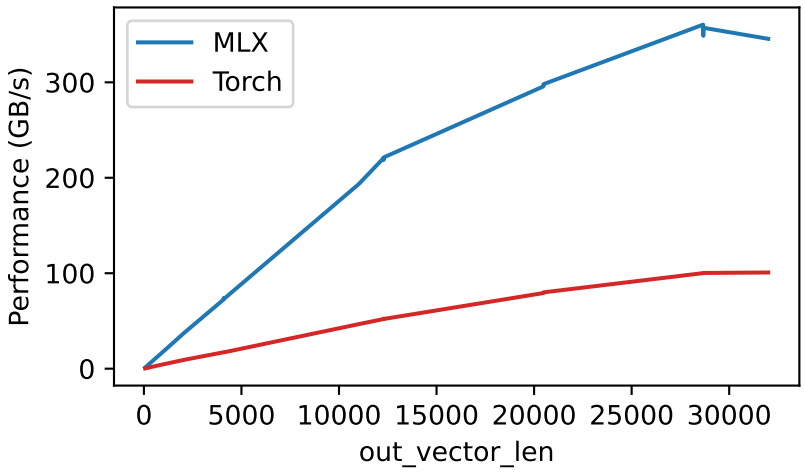
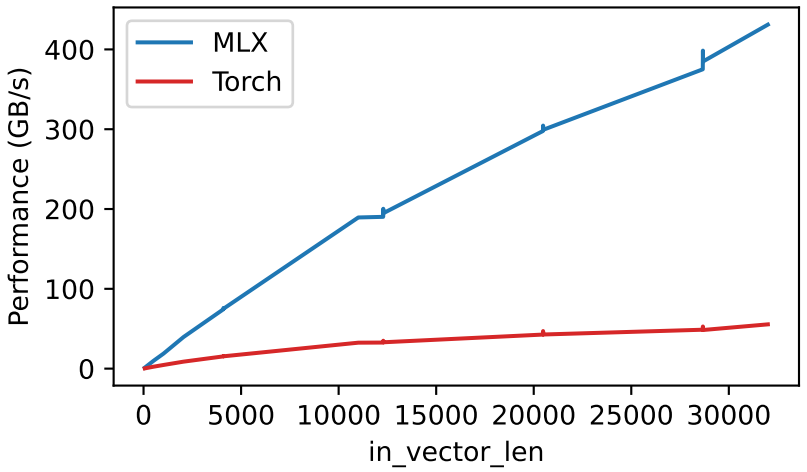
Apple M2 Ultra: float16 gemv

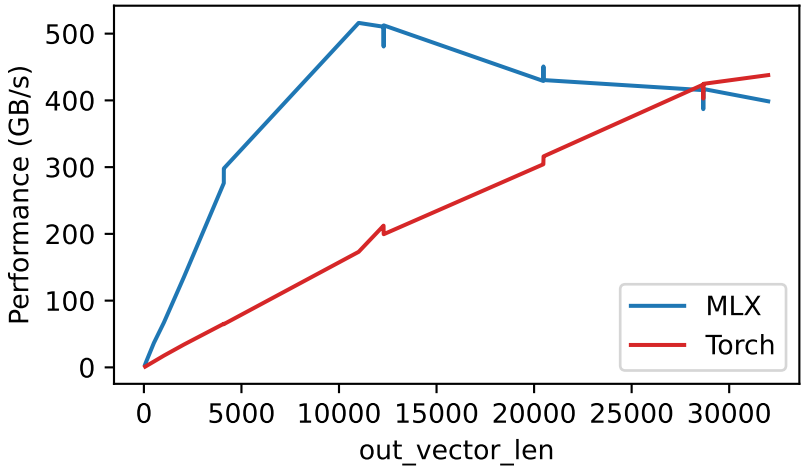
gemv ([out_vec_len, 128] X [128, 1]) | float16



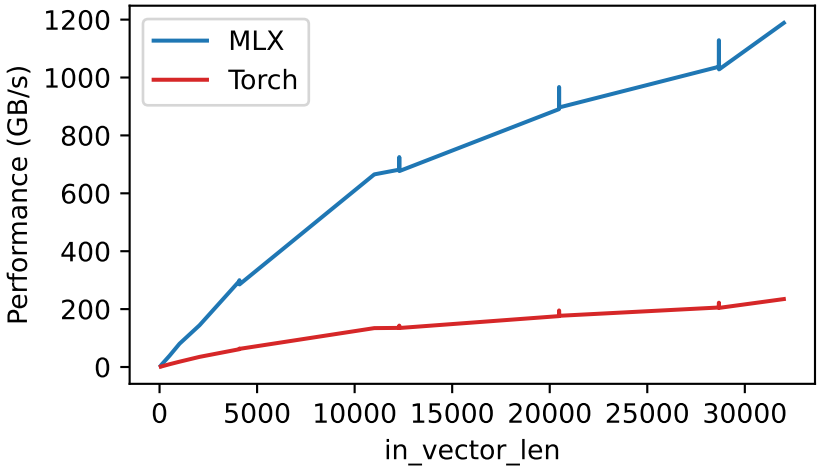
([128, in_vec_len] X [in_vec_len, 1])



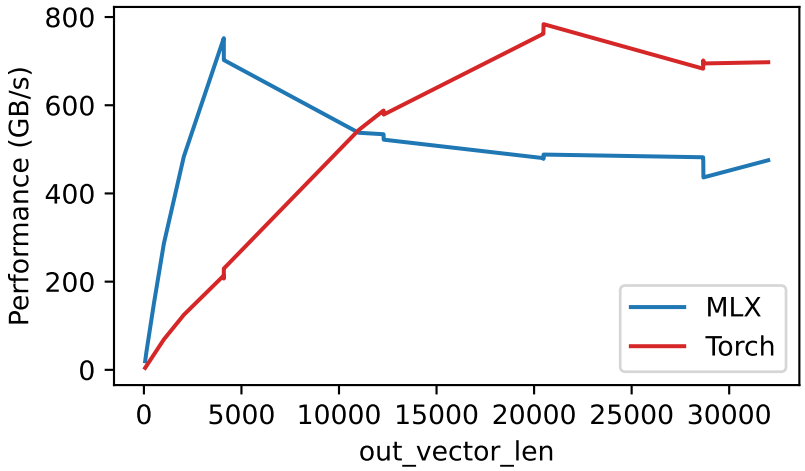
gemv ([out_vec_len, 512] X [512, 1]) | float16



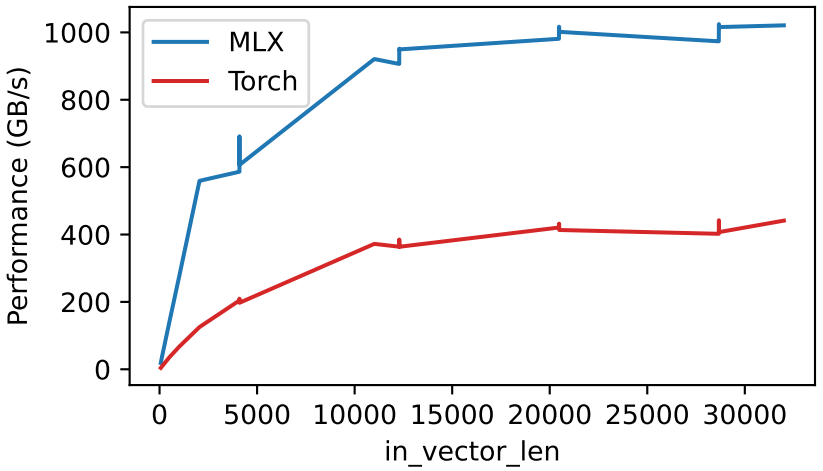
([512, in_vec_len] X [in_vec_len, 1])



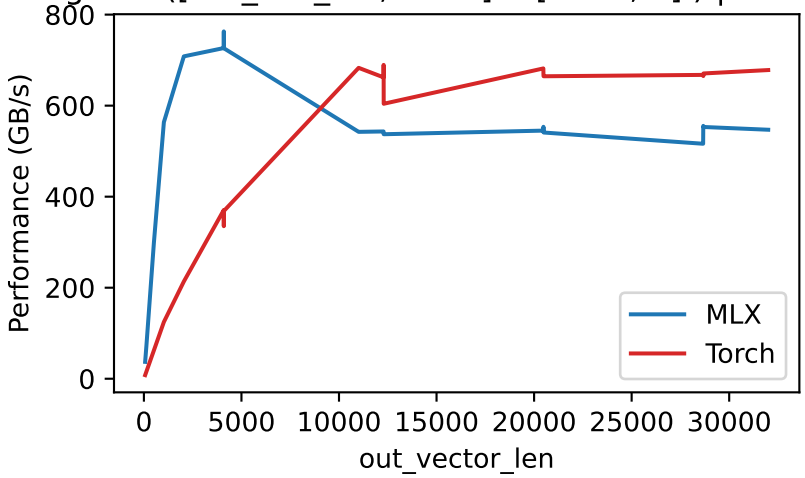
gemv ([out_vec_len, 2048] X [2048, 1]) | float16



([2048, in_vec_len] X [in_vec_len, 1])



gemv ([out_vec_len, 4096] X [4096, 1]) | float16



([4096, in_vec_len] X [in_vec_len, 1])