## Method 2

## Compressing the Linear Layer



The original layer

$$\boldsymbol{y}^{N_{out}} = \boldsymbol{W}^{N_{out} \times N_{in}} \boldsymbol{x}^{N_{in}} + \boldsymbol{b}^{N_{out}}$$

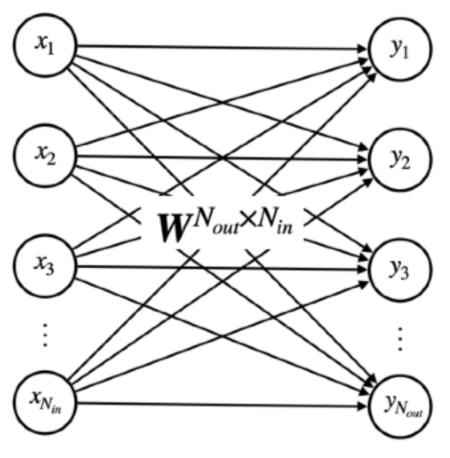
Tucker-2 approximation of weight matrix

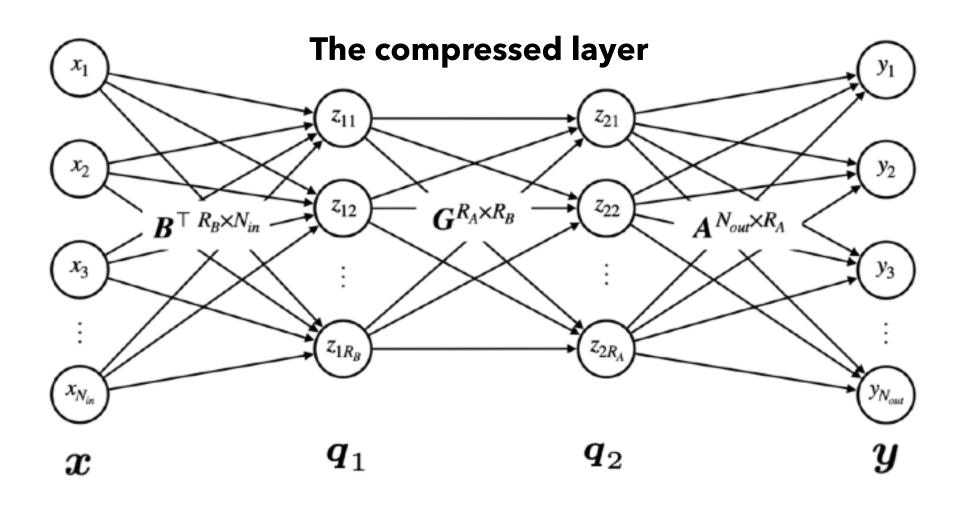
$$oldsymbol{W} pprox oldsymbol{G} imes_1 oldsymbol{A} imes_2 oldsymbol{B} = oldsymbol{A} oldsymbol{G} oldsymbol{B}^ op$$

The compressed layer

$$m{y}^{N_{ ext{out}}} pprox m{A}^{N_{ ext{out}} imes R_A} m{G}^{R_A imes N_{ ext{in}}} m{B}^{ op R_B imes N_{ ext{in}}} m{x}^{N_{ ext{in}}} m{x}^{N_{ ext{in}}} + m{b}^{N_{ ext{out}}}$$







## Method 2 rank selection Variational Bayesian Matrix Factorization



Acts as a good heuristic, but not guaranteed to yield the optimal ranks for the master problem

Given a matrix  $\boldsymbol{V}$  it is assumed:

$$oldsymbol{V}^{L imes M} = oldsymbol{U}^{L imes M} + oldsymbol{E}^{L imes M}$$

Goal is to find matrices **A** and **B** such that

$$\boldsymbol{U} = \boldsymbol{B}\boldsymbol{A}^\top$$

Probabilistic model of  $\boldsymbol{V}$  gives the posterior distributions of  $\boldsymbol{A}$  and  $\boldsymbol{B}$ 

Generally a non-convex problem, but analytical solution have been found