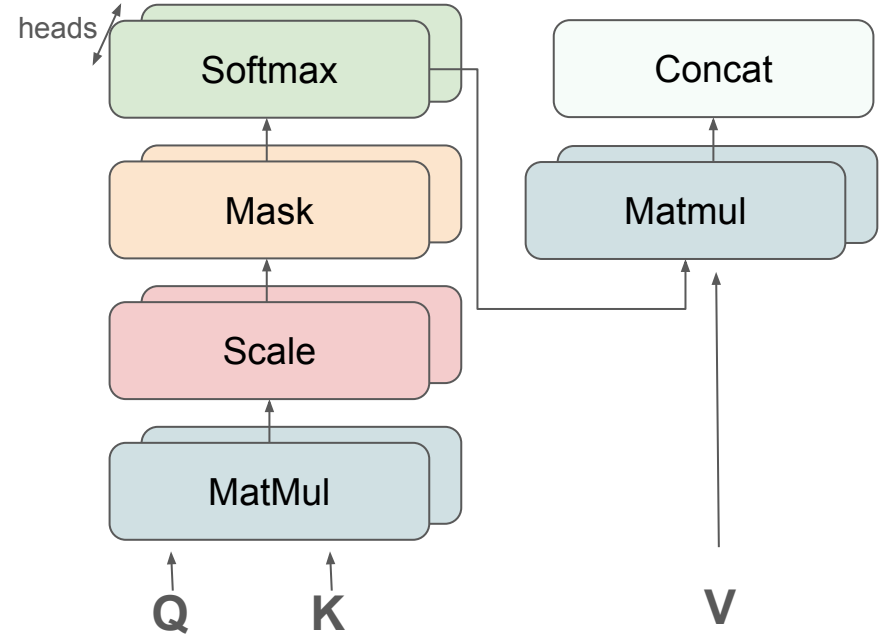
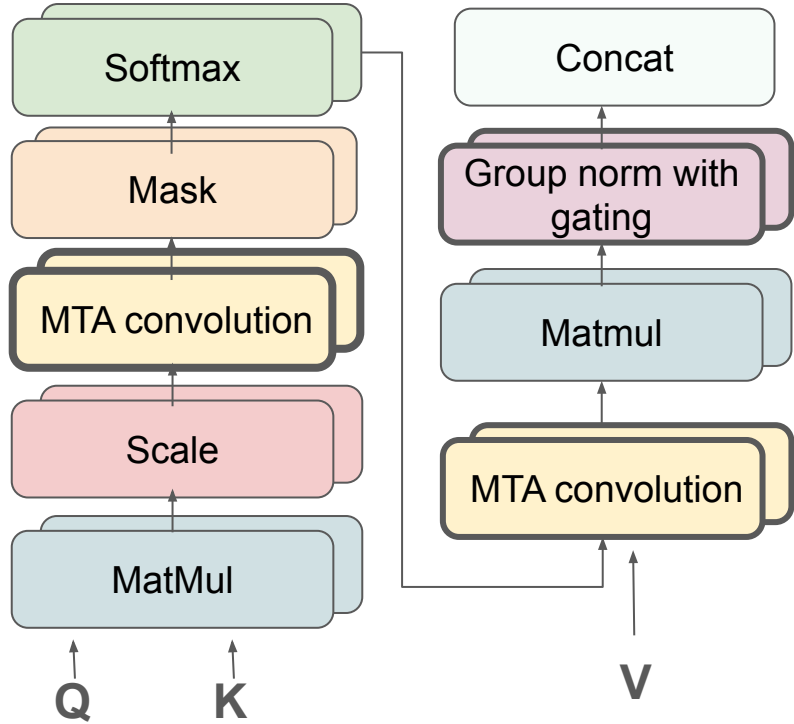


Standard multi-head attention



Multi-token attention