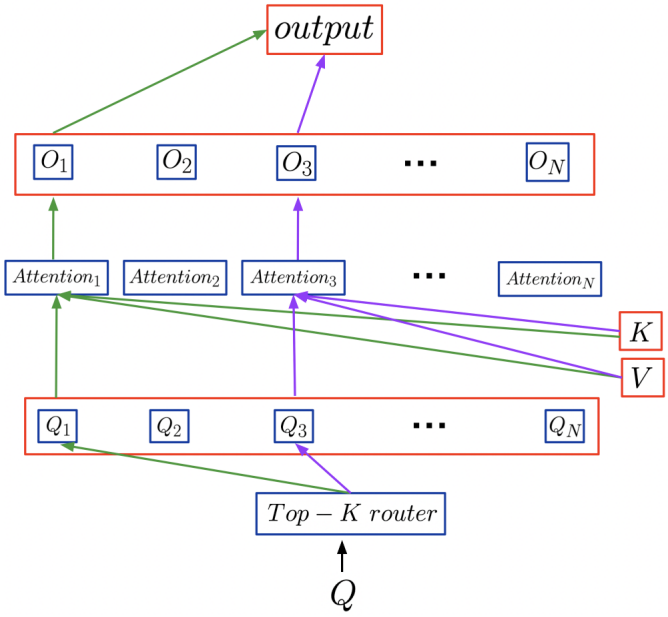
Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Mixture of Experts - Attention Heads



The routing probability for each expert is:

$$\boxed{P_{i,t} = Softmax_i(Q_t \cdot W_g)}$$

We can define the next subset:

$$\boxed{G(Q_t) = TopK(P_{i,t}, K)}$$

We define the weights:

$$\boxed{w_{i,t} = \frac{P_{i,t}}{\sum_{j \in G(Q_t)} P_{j,t}}}$$

We define the output:

$$\boxed{O_t = \sum_{j \in G(Q_t)} O_{j,t} \cdot w_{j,t}}$$