**Multi-head attention consists of four parts:**

* Linear layers and split into heads.
* Scaled dot-product attention.
* Concatenation of heads.
* Final linear layer.

Each multi-head attention block gets three inputs; Q (query), K (key), V (value). These are put through linear (Dense) layers and split up into multiple heads.

The scaled\_dot\_product\_attention defined above is applied to each head (broadcasted for efficiency). An appropriate mask must be used in the attention step. The attention output for each head is then concatenated (using tf.transpose, and tf.reshape) and put through a final Dense layer.

Instead of one single attention head, query, key, and value are split into multiple heads because it allows the model to jointly attend to information at different positions from different representational spaces. After the split, each head has a reduced dimensionality, so the total computation cost is the same as a single head attention with full dimensionality.

Multi-head attention gồm 4 phần:

Các lớp tuyến tính và chia các heads.

Scaled dot-product(tích vô hướng) attention

Mối liên hệ giữa cá đầu

Lớp tuyến tính cuối cùng

Mỗi khối multi-head nhận 3 inputs(Q (query), K (key), V (value)), chúng được đặt qua các lớp tuyến tính (Mật độ) và chia thành nhiều Heads.

Scaled dot-product attention được áp dụng cho mỗi head. Một mask thích hợp phải được sử dụng ở bước tập trung. Tỉ lệ mức độ tập trung của mỗi đầu được nối( sử dụng tf.transpose và tf.reshape) và đặt qua lớp dense cuối dùng.

Thay vì 1 head, query, key, và value chia thành nhiều heads vì nó cho phép các model tiếp cận thông tin tại các vị trí, không gian khá nhau. Sau khi phân tách, mỗi đầu sẽ giảm một chiều(dimensonality: chiều không gian), vì vậy chi phí tính toán bằng tính toán đơn head với toàn bộ chiều không gian