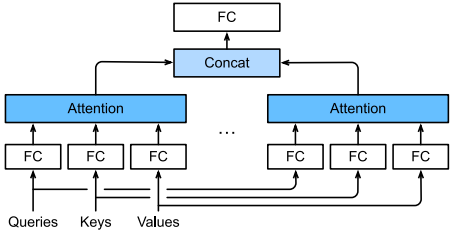
10.5. Multihead Attention

Lecture based on "Dive into Deep Learning" **http://D2L.AI** (Zhang et al., 2020)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Multi-Head Attention



- Queries, keys, and values are projected with h different linear layers.
- h projected queries, keys, and values are fed into attention pooling in parallel.
- h attention pooling outputs are concatenated and transformed with another learned linear projection to produce the final output.
- Each of h attention pooling outputs is a *head*.

Given the same set of queries, keys, and values we may want our model to combine different behaviors of the same attention mechanism, such as capturing dependencies of various ranges (e.g., shorter-range vs. longer-range) within a sequence.
Allow attention to jointly use different representation subspaces of queries, keys, and values by transforming with h independently learned linear projections.

Mathematical Definition

Given a query $\mathbf{q} \in \mathbb{R}^{d_q}$, a key $\mathbf{k} \in \mathbb{R}^{d_k}$, and a value $\mathbf{v} \in \mathbb{R}^{d_v}$, each attention head $\mathbf{h}_i$ $(i = 1, \ldots, h)$ is computed as

$$\mathbf{h}_i = f(\mathbf{W}_i^{(q)}\mathbf{q}, \mathbf{W}_i^{(k)}\mathbf{k}, \mathbf{W}_i^{(v)}\mathbf{v}) \in \mathbb{R}^{p_v},$$

where learnable parameters $\mathbf{W}_i^{(q)} \in \mathbb{R}^{p_q \times d_q}$, $\mathbf{W}_i^{(k)} \in \mathbb{R}^{p_k \times d_k}$ and $\mathbf{W}_i^{(v)} \in \mathbb{R}^{p_v \times d_v}$, and f is attention pooling, such as additive attention and scaled dot-product attention.

The multi-head attention output is a linear transformation of the concatenation of h heads:

$$\mathbf{W}_o \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_h \end{bmatrix} \in \mathbb{R}^{p_o},$$

where $\mathbf{W}_o \in \mathbb{R}^{p_o \times hp_v}$ are learnable parameters.

Based on this design, each head may attend to different parts of the input.
More sophisticated functions than the simple weighted average can be expressed.

Summary

Multi-head attention combines knowledge of the same attention pooling via different representation subspaces of queries, keys, and values. To compute multiple heads of multi-head attention in parallel, proper tensor manipulation is needed.