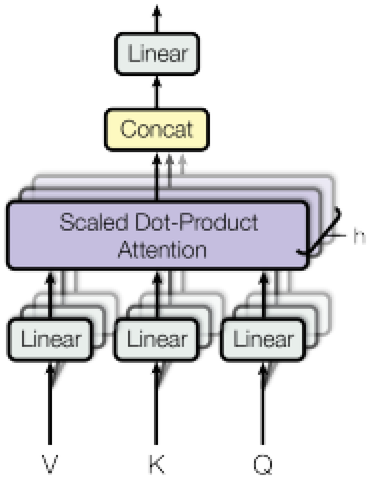
Computer Vision Task

Preprocessing the Data:
The training images paths and text prompts were captured from the JSON file.
Rather than training our model on several instances,I have used **Multi head attention** mechanism to encode multiple relationships and nuances for each word.

Pre training:
It consists of three components:
1.**Query**:What we are looking for?
2.**Key**:What we have?
3.**Value**:What we want to output?

Multi-Head Attention



To allow the parallel computation of multiple heads,transposition methods are used

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

The pretrained weights are imported and loaded .
The phrase cut images were transformed for visualization..

Again I made use of Hugging face library for most part of my code.