**四、 Single Headed Attention RNN (SHA-RNN)**

* 單頭注意力 (Single Headed Attention)

Transformer 模型建立於無序基礎，並只透過注意力機制完成訓練，但每層網路都有數十個注意力頭 (Attention Heads)，運算過程中因無法得知有效頭數而耗費多餘的運算資源。相較而言，SHA-RNN的注意力機制只保留一個頭完成向量之注意力乘積與得分。

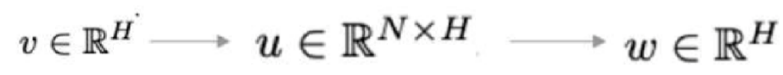
* 結構變更

主要以 Transformer 之自注意力機制為基礎修改，四層結構中，每層先做LSTM，進行層標準化 (Layer Normalization) 後連接注意力機制，實際上為8層網路。

結構變更後與 Transformer不同之處在於本模型只對Q做全連接層，並以sigmoid產生Q, K。然而，過程中亦發現 LSTM 輸出須經過全連接層轉換為Q,K,V，因此本研究改進之核心部分在於作者提出改造後的前饋層(Boom Layer)。

* Boom Layer

為了減少運算量，基於Transformer前饋層改造後的 Boom Layer將原本前饋層透過激活函數GeLU轉換成 N 倍向量，分成 N 等分後加總將維度轉換回原維度。



<< 維度轉換示意圖