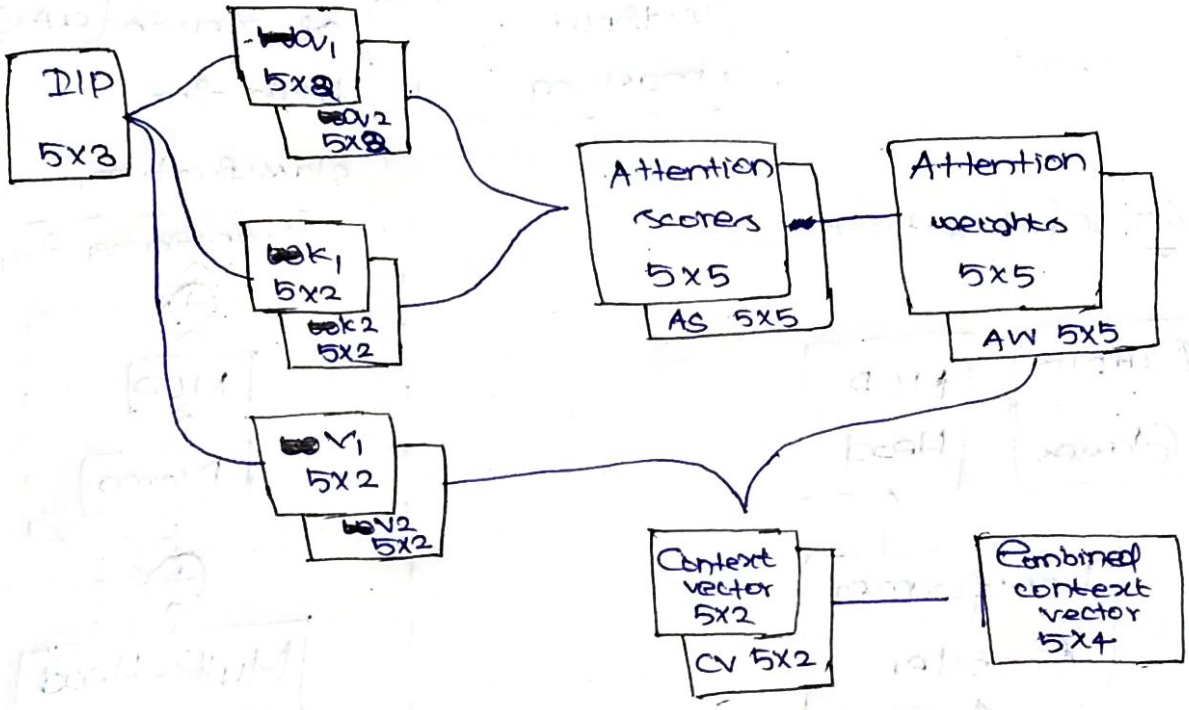
Lec-6

Introduction to Multi-Head Attention



* Instead of single attention head, there will be multiple attention head run in parallel.
* Each attention head will have,
  - → its own Q, K, V weights matrices.
  - → learn different patterns

* The outputs are concatinated along the feature dimension. Each token has 4 dimensional context vector instead of 2

* GPT-2 small has 12 transformer block and 12 attention head per block.

* GPT-3 has 96 transformer block and 96 attention head per block.

—— x ——