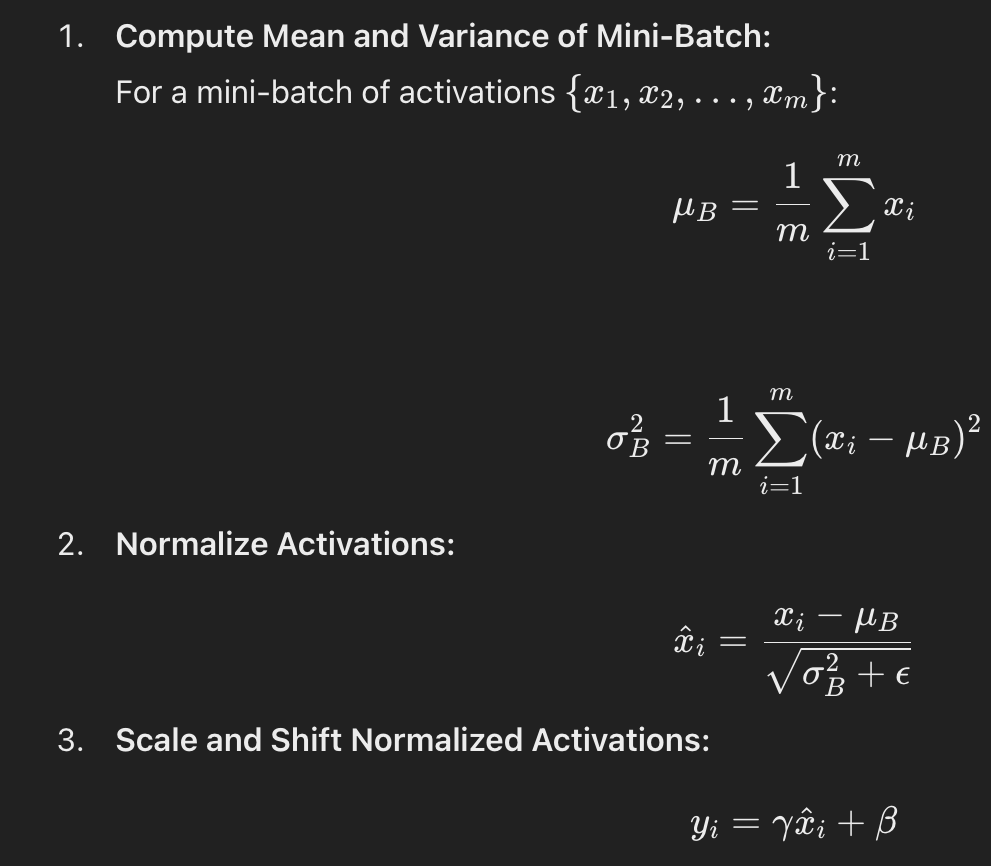
**REVIEW QUESTIONS**

**1.BATCH NORMALIZATION**

It stabilizes the training process by making the input to each layer maintain a constant distribution, which creates higher learning rates and faster convergence.

Steps:

γ-adjusts the normalized output

β**-**Shift parameter that shifts the normalized output

Above 2 are learnable parameters .

Running mean-keeps the track of mean in mini-batches

Running variance-keeps track on variance

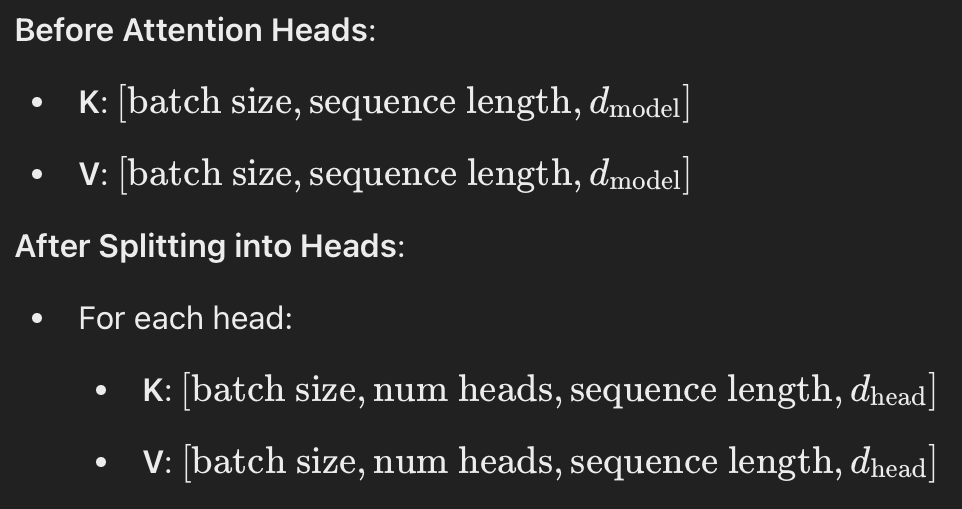
Hyperparamters-

Momentum -determines the update rate of mean and variance.

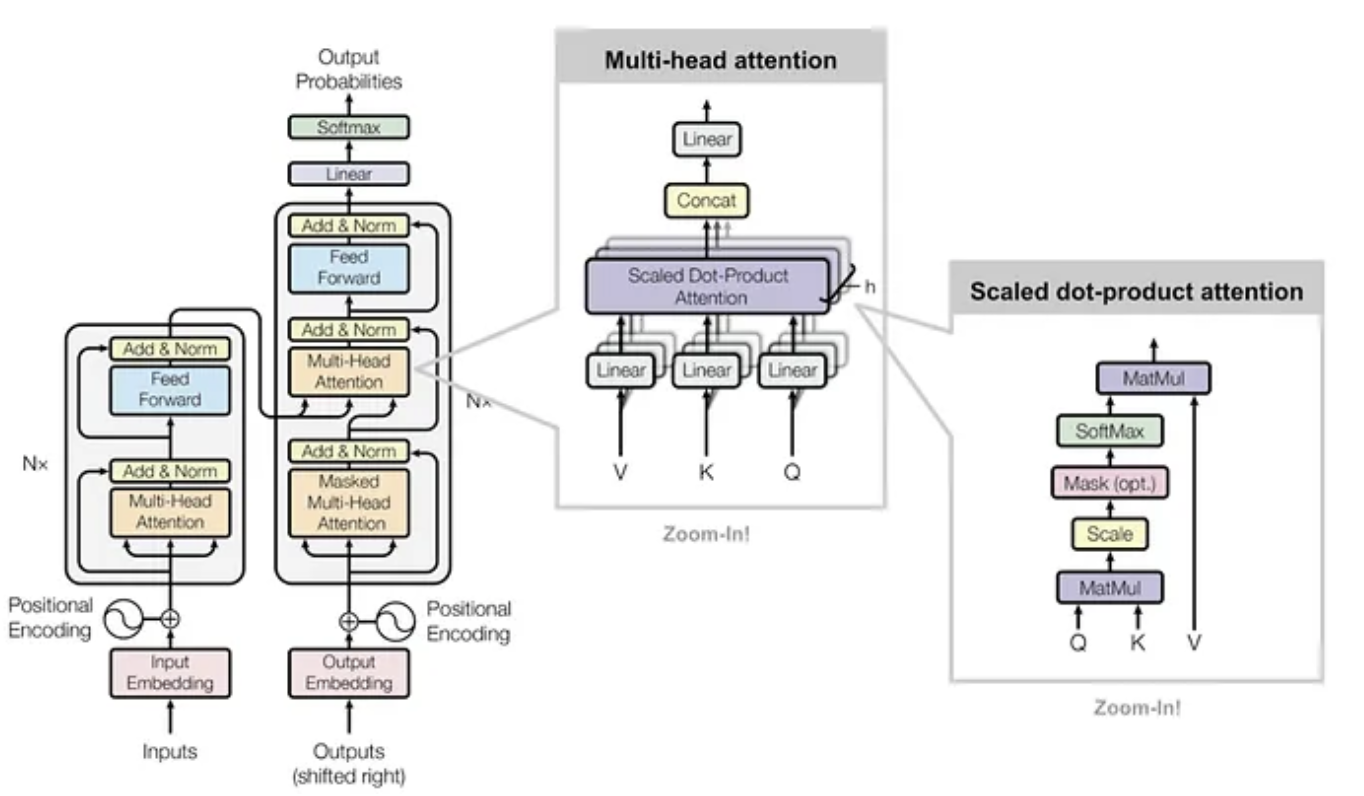
Epsilon- added to prevent division by 0.

**2.RNN used for timeline analysis**

The Bidirectional Recurrent Neural Network (BRNN) used when the sequence values depend not only on previous activation but also on future activation in the sequence. This makes BRNNs particularly useful for timeline analysis as they process data in both forward and backward directions. For ex, in speech recognition, understanding the current sound might require knowledge of both preceding and succeeding sounds in a sentence.The BRNN architecture adds an additional layer that processes information backward in time, enhancing the network's ability to capture context from the entire sequence.

**3.What is the shape of K and V in transformers?**

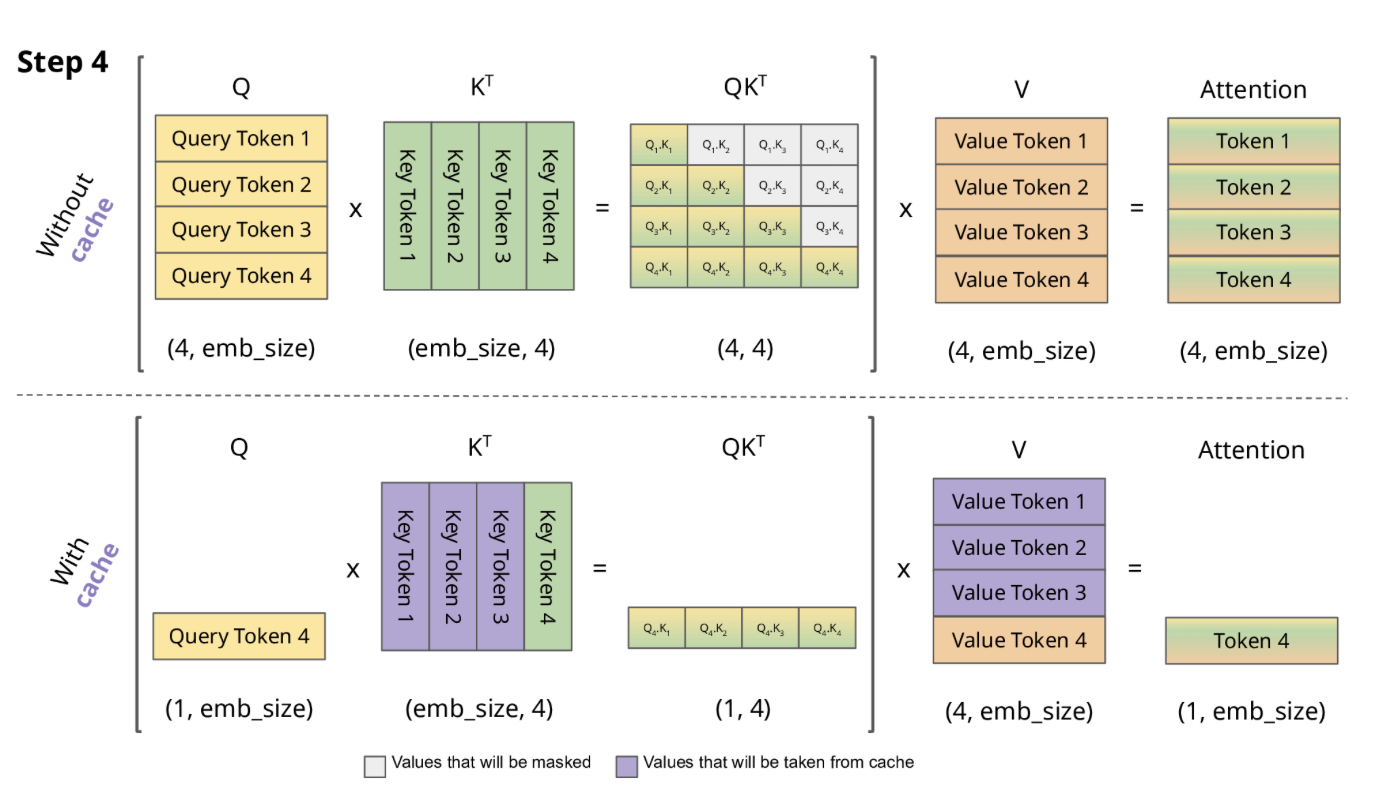
**4.Kv caching**

The Key and Value states are used for calculating the scaled dot-product attention.

KV caching occurs during multiple token generation steps and only happens in the decoder .

Since the decoder is traditional, at each generation step we are recalculating the same previous attention, even if we want to calculate the attention for the new token.

Here we use KV caching,

By caching previous keys and values , we focus on new activation.

**5.multi-class and multi-label**

\*multi-class

With multiclass classification, the model will select just one winning label.

Here each instance is assigned one label.

Used in MNIST digit recognition

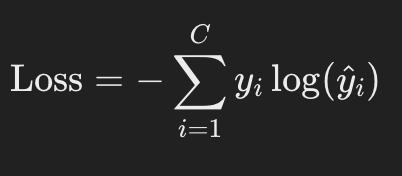
\*multi-label

With multilabel classification, the model has the option of outputting multiple labels if relevant.

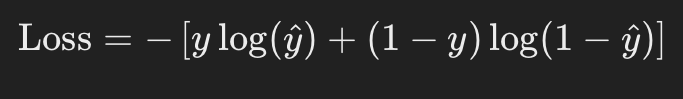
Here one instance can be assigned with multiple labels.

Used in tagging images

**5.loss function in softmax**

Cross-entropy loss

Gives the difference between predicted probability by softmax and true target labels.

Binary -entropy loss

Categorical cross-entropy loss is used when vectors are on hot encoded . Its functionality is similar to cross-entropy loss.