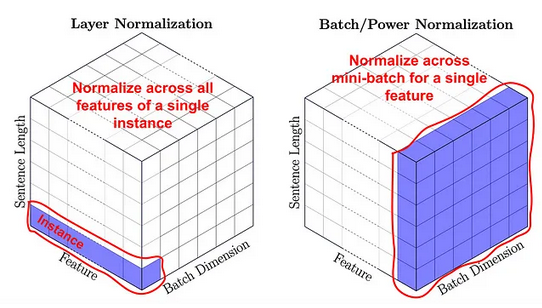
**Layer normalization**

======================================================================

Layer Normalization is a technique used in Transformers to stabilize and accelerate training by normalizing the inputs across the features for each token independently.

Layer Normalization (LayerNorm) normalizes the values within each individual token (or data point), rather than across the batch (like BatchNorm does).

--------------------------------------------------------------------------------------------------------------



--------------------------------------------------------------------------------------------------------------

**Why is it used in Transformers?**

Transformers don't process sequences sequentially (like RNNs), and they often deal with variable-length inputs. BatchNorm is sensitive to batch sizes and order, so **LayerNorm is preferred** because:

* It works the same regardless of batch size.
* It stabilizes the training dynamics.
* It improves convergence.

--------------------------------------------------------------------------------------------------------------

