Lecture 20: Layer Normalization

Why is it needed?

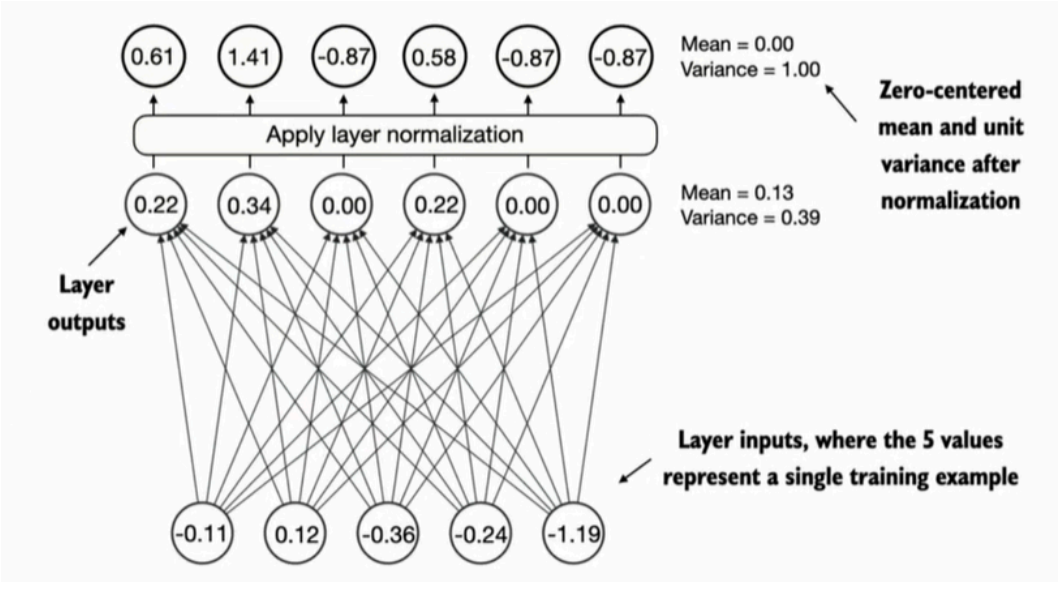
Training deep neural networks with many layers can be challenging due to problems like - vanishing gradient problem or exploding gradients problem. This leads to unstable training dynamics.

Layer normalization improves the stability and efficiency of neural network training.

Main Idea: Adjust outputs of neural networks to have mean zero

Layer normalization keeps gradients stable. It also prevents a problem like **internal covariate shift** (**Internal covariate shift** is the change in the distribution of intermediate layer activations during training due to parameter updates in previous layers.)

In GPT-2 and modern transformer architectures, layer normalization is typically applied before and after the multi head attention module and before the final output layer.



Layer and Batch normalization are two very different things. Layer normalization normalizes along the feature dimension, and independent of the batch size.