00:31:00 implementing the forward pass to get logits

model okay now before we can actually generate from this model we have to be able to forward it we didn't actually write that code yet so here's the forward function so the input to the forward is going to be our indices our tokens uh token indices and they are always of shape B BYT and so we have batch dimension of B and then we have the time dimension of up to T and the T can't be more than the block size the block size is is the maximum sequence length so B BYT indices arranged is sort of like a two-dimensional layout and remember that basically every single row of this is of size up to uh block size and this is T tokens that are in a sequence and then we have B independent sequences stacked up in a batch so that this is efficient now here we are forwarding the position embeddings and the token embeddings and this code should

be very recognizable from the previous lecture so um we basically use uh a range which is kind of like a version of range but for pytorch uh and we're iterating from Z to T and creating this uh positions uh sort of uh indices um and then we are making sure that they're in the same device as idx because we're not going to be training on only CPU that's going to be too inefficient we want to be training on GPU and that's going to come in in a bit uh then we have the position embeddings and the token embeddings and the addition operation of those two now notice that the position embed are going to be identical for every single row of uh of input and so there's broadcasting hidden inside this plus where we have to create an additional Dimension here and then these two add up because the same position embeddings apply at every single row of our example stacked up in a batch then we forward the Transformer

blocks and finally the last layer norm and the LM head so what comes out after forward is the logits and if the input was B BYT indices then at every single B by T we will calculate the uh logits for what token comes next in the sequence so what is the token B t+1 the one on the right of this token and B app size here is the number of possible tokens and so therefore this is the tensor that we're going to obtain and these low jits are just a softmax away from becoming probabilities so this is the forward pass of the network and now we can get load and so we're going to be able to generate from the model imminently okay so now we're going to