00:00:00 intro: Lets reproduce GPT-2 (124M)

hi everyone so today we are going to be continuing our Zero to Hero series and in particular today we are going to reproduce the gpt2 model the 124 million version of it so when openi released gpt2 this was 2019 and they released it with this blog post on top of that they released this paper and on top of that they released this code on GitHub so open a/ gpt2 now when we talk about reproducing gpt2 we have to be careful because in particular in this video we're going to be reproducing the 124 million parameter model so the thing to realize is that there's always a miniseries when these are releases are made so there are the gpt2 miniseries made up of models at different sizes and usually the biggest model is called the gpt2 but basically the reason we do that is because you can put the model sizes on the x-axis of plots like

this and on the Y AIS you put a lot of uh Downstream metrics that you're interested in like translation summarization question answering and so on and you can chart out these scaling laws so basically as the model size increases you're getting better and better at Downstream metrics and so in particular for gpt2 if we scroll down in paper there are four models in the gpt2 miniseries starting at 124 million all the way up to 1558 million now the reason my numbers the way I say them disagree with this table is that this table is wrong if you actually go to the uh gpt2 uh GitHub repo they sort of say that um there was an error in how they added up the parameters but basically this is the 124 million parameter model Etc so the 124 million parameter had 12 layers in the Transformer and it had 768 channels in the Transformer 768 dimensions and I'm going to be assuming some familiarity with what these terms mean because I covered all of

this in my previous video let's build gpt2 uh let's build GPT from scratch so I covered that in the previous video in this playlist now if we do everything correctly and everything works out well by the end of this video we're going to see something like this where we're looking at the validation loss which basically um measures how good we are at predicting the next token in a sequence on some validation data that the model has not seen during training and we see that we go from doing that task not very well because we're initializing from scratch all the way to doing that task quite well um by the end of the training and hopefully we're going to beat the gpt2 uh 124 M model now previously when they were working on this this is already 5 years ago so this was probably a fairly complicated optimization at the time and the gpus and the compute was a lot smaller today you can reproduce this model in roughly an hour or probably less

even and it will cost you about 10 bucks if you want to do this on the cloud uh Cloud Compu a sort of computer that you can all rent and if you pay $10 for that computer you wait about an hour or less you can actually achieve a model that is as good as this model that open ey released and uh one more thing to mention is unlike many other models open ey did release the weights for gpt2 so those weights are all available in this repository but the gpt2 paper is not always as good with all of the details of training so in addition to the gpt2 paper we're going to be referencing the gpt3 paper which is a lot more Concrete in a lot of the hyp parameters and optimization settings and so on um and it's not a huge departure in the architecture from the GPT 2 uh version of the model so we're going to be referencing both gpt2 and gpt3 as we try to reproduce gpt2 124 M uh so let's go so the first thing I