HTX xData Test Q4

The pretrained model and finetuned model are evaluated via generating a transcription, then comparing their results. The word error rate (WER) of the pretrained and finetuned model are as shown below:

Model	WER
Pretrained	0.108
Finetuned	0.038

As seen from the WER, the finetuned model performs similarly to the pretrained model results when it is evaluated against the libspeech dataset, showing that the finetuned model has learnt the appropriate features in the common voice dataset.

The table below shows some instances in which the finetuned model results did not match with the ground truth, whereas the pretrained model matches:

Finetuned Text	Ground Truth Text
BUT THE ENGLISHMAN APPEARED NOT TO ATTAUCH ANY IMPORTANCE TO IT	BUT THE ENGLISHMAN APPEARED NOT TO ATTACH ANY IMPORTANCE TO IT
THE REFERENCES ARE GOOD IN THE VERY NARROW AREA OF HISTOGRAMH METHODS	THE REFERENCES ARE GOOD IN THE VERY NARROW AREA OF HISTOGRAM METHODS
AT WASN'T MUCH ANYWAY	IT WASN'T MUCH ANYWAY
A LITTLE LEMONGRASS SHOULD FRESTEN IT UP	A LITTLE LEMONGRASS SHOULD FRESHEN IT UP
THE THE BOY AWOKE AS THE SUN ROSE	THE BOY AWOKE AS THE SUN ROSE
THE BOY WATCHED AS HIS COMPANION WENT TO HIS HORSE AND WITHDREW A CIMITAR	THE BOY WATCHED AS HIS COMPANION WENT TO HIS HORSE AND WITHDREW A SCIMITAR
BUT SHE'S GONE	BUT SHE IS GONE
RED YOU WOULD HAVE TO HAVE BEEN BORN AN ARAB TO UNDERSTAND HE ANSWERED	YOU WOULD HAVE TO HAVE BEEN BORN AN ARAB TO UNDERSTAND HE ANSWERED
THE HORIZON WAS TINGED WITH RED AND SUDDENLY THE SIGN APPEARED	THE HORIZON WAS TINGED WITH RED AND SUDDENLY THE SUN APPEARED
GIME ME MY ROBE	GIVE ME MY ROBE

As we can see, it sometimes makes spelling mistakes, such as “ATTACH” is spelled as “ATTAUCH”, “FRESTEN” instead of "FRESHEN”, or “CIMITAR” instead of "SCIMITAR”. One possible way to remedy such errors could be to include more samples where similar words are included. Another way could be to add a spell-checking stage after the transcription is completed to correct for such errors. 84/4076 samples observe such errors.

The table below shows instances where the pretrained model results did not match with the ground truth, whereas the finetuned model matches:

Pretrained Text	Ground Truth Text
ARE YOU GOING TO TALK OR WANT YOU	ARE YOU GOING TO TALK OR AREN'T YOU
HE MOVED ABOUT INVISIBLE BUT EVERY ONE COULD HEAL HIM	HE MOVED ABOUT INVISIBLE BUT EVERYONE COULD HEAR HIM
WE COULD GET TO THE PYRAMIDS BY TO MORROW SAID THE OTHER TAKING THE MONEY	WE COULD GET TO THE PYRAMIDS BY TOMORROW SAID THE OTHER TAKING THE MONEY
I'M TRYING TO THINK OF SOMETHING BEFORE THOSE REPORTES GETBACK	I'M TRYING TO THINK OF SOMETHING BEFORE THOSE REPORTERS GET BACK
WE DON'T HAVE TO GIVE UP OUR CLAP	WE DON'T HAVE TO GIVE UP OUR CLUB
CAN I GET A WHAT WHAT	CAN I GET A WOOT WOOT
PEOPLE SAW ME COMING AND WELCOMED ME HE SOUGHED	PEOPLE SAW ME COMING AND WELCOMED ME HE THOUGHT
THE SIMMON BLUE THAT DAY AS IT HAD NEVER BLOWN BEFORE	THE SIMUM BLEW THAT DAY AS IT HAD NEVER BLOWN BEFORE
THE FOLLOWING NIGHT THE BOY APPEARED AT THE ALCHEMIS'S TENT WITH THE HORSE	THE FOLLOWING NIGHT THE BOY APPEARED AT THE ALCHEMIST'S TENT WITH A HORSE
SAYS FERACHE TO GO THE LINET	SAYS FOR US TO GO THE LIMIT

Like the finetuned model, the pretrained model makes spelling mistakes from the ground truth text, such as “TO MORROW” instead of "TOMORROW”, “REPORTES” instead of "REPORTERS”, or “SOUGHED” instead of “THOUGHT”. In the pretrained model case, it appears to have heard a different word from the ground truth, such as “WHAT” instead of “WOOT”, or “CLAP” instead of “CLUB”. Such errors may need more contextual knowledge to provide the correct transcription. 1145/4076 samples encountered such issues.

The table below shows instances where both finetuned model and pretrained model did not match with ground truth:

Ground Truth Text	Finetuned Text	Pretrained Text
AT OTHER TIMES AT A CRUCIAL MOMENT I MAKE IT EASIER FOR THINGS TO HAPPEN	AT OTHER TIMES AT A CRUCIAL MOMENT YOU MAKE IT EASIER FOR THINGS TO HAPPEN	AT OTHER TIMES AT A CRUISING MOMENT YOU MAKE IT EASIER FOR THINGS TO HAPPEN
AND THEN THEY WANT THE PERSON TO CHANGE	AND THEN THEY WANTED THE PERSON TO CHANGE	AND THEN THEY WANTED THE PERSON TO CHANGE
THE BASKETBALL BOUNCED OFF HIS SHIELD OF TITANIUM	THE BASKETBALL BOUNCED ON THE SHIELS OF TITANIUM	THE BASKETBALL BOUNCED OF HIS SHIELD OF TITANIUM
GOTTA BE GENTLE TO SUIT ME	THET GOTTA BE GENTLE TO SEE ME	THET GOINTO BE GENTLE TO SIDME
SUPPOSE THERE WAS A SHANE YORK AND HE WALKED INTO THIS OFFICE	SUPPOSE THERE WAS A SHANEYOK AND HE WALKED INTO THE LITE OFFICE	SUPPOSE THERE WAS A SHAME YOG AND HE WALKED INTO POLISE OFFICE
I NEED YOU TO BE SPONTANEOUS HE ASKED ME OUT TO DIN DIN	I NEED YOU TO BE SPONTANEOUS HE HAS KAD ME OUT YOUR DIN DIN	I NEED YOU TO PEA SPONTENEOUS HE HAS CATTED ME OUT YOUR DINGDING
THE BOY LOOKED AROUND FOR THE OVENS AND OTHER APPARATUS USED IN ALCHEMY BUT SAW NONE	THE BOY LOOKED AROUND FOR THE OVENS AND OTHER APPARATUS USED IN ALCHEMY BUT SAW NOUNG	THE BOY LOOKED AROUND FOR THE OVENS AND OTHER APPARATUS USED IN ALCEMY TUT SONAN
THEY'RE FORMING CLUBS	THEY 'RE FORMING CLUBS	THEY ARE FORMING CLUBS
THIS MORNING I FOUND A CALCULATOR TAPED TO MY WII	THIS MORNING I FOUND A CALCULATOR TAPED TO MY WIIT	THIS MORNING I FOUND A CALCULATOR TAKE TO MY WEAT
ON TOP OF ALL THAT THE WEEDS KEEP GROWING AND THE GARBAGE HAS TO BE TAKEN OUT	ON TOP OF ALL THAT THE WIED'S KEEP GROWING AND THE GARBIT HAS TO BE TAKEN OUT	ON TOP OF ALL THAT THE WEEDS KEEP GROWING AND THE GABBIT HAS TO BE TAKEN OUT

From the table, the errors seen are commonly spelling mistakes of words that are less commonly used but do retain their phonetics. For example, the ground truth “BUT SAW NONE” is transcribed as “ BUT SAW NOUNG” for the finetuned model, and “TUT SONAN” in the pretrained model. 690/4076 samples have such error.

Overall, this ML model performs well on the Common Voice dataset, even without much finetuning. While there are some spelling mistakes, most of the transcribed text is largely similar to the ground truth. An increase in the number of training epochs would likely improve the accuracy of the finetuned model. As the finetuned model performs better than the pretrained model, it is likely that the finetuned model learnt from the accent, pace, pitch etc. variability of the common voice dataset. Some experiments to improve the accuracy may include increasing the size of the model, especially if the

variability of the dataset is large. Another experiment could be to replace the CTC decoder with the n-gram language model, which will help with alleviating the spelling mistakes observed in the output.