Comparison of baseline model vs. finetuned model

1. Performance metrics

Generally, the finetuned model performed better than the baseline model on the cv-valid-dev dataset (n = 4,076). The word error rate (WER) and character error rate (CER) showed a difference of -0.0177 and -0.00817 respectively as seen in Table 1.

Table 1: Performance metrics of baseline model vs finetuned model on cv-valid-dev dataset

	WER (3 s.f.)	**CER (3 s.f.)**
Baseline model	0.108	0.0454
Finetuned model	0.0906	0.0372
Change	-0.0177	-0.00817

This slight improvement in model performance is expected, however improvements in the finetuning process could still be made to improve performance further.

2. Suggestions for improvements

Firstly, the finetuning dataset was done on subsampled dataset (n = 5,000 out of 380,377) due to time and resource constraints.

- ➢ Suggestion: Gradually increasing dataset size (e.g., 10k, 20k, 50k) incrementally while observing performance gains at each step. Hence, a learning curve can be made to determine if there is a performance plateau point vs. dataset size.

The hyperparameters chosen were also arbitrary in nature with the focus being the number of epochs to see the change in performance vs. epoch. Hence, a more systematic approach could be taken.

- ➢ Suggestion: Perform hyperparameter tuning using tools like Optuna to search for better learning rate, batch size, warm-up steps and weight decay. Furthermore, number of epochs could be increased as well to observe the performance plateau point vs. epoch number.

Next, different datasets could be used. An example would be a dataset with differing difficulty score (curriculum learning). This will allow the model to capture general patterns before tackling difficult samples.

- ➢ Suggestion: One way to go about this would be to augment the data to introduce defects. For e.g., adding background noise, shifting the pitch, stretching the length of the audio, random cropping and mixing of samples. This would also help with expanding the dataset and the generalization of the model.

Lastly, relying on a single train/validation split (e.g., 70-30) could lead to high variance in performance evaluation, especially on smaller datasets.

- ➢ Suggestion: Implement k-fold cross-validation to reduce performance variance across random splits this will in turn allow for a better estimation of model performance during finetuning.

Done by: Nigel Wee

Words: 345