BERT based Vanilla Models - Initial Experiments

This report details the findings from running multiple preliminary experiments on a simple model architecture that adds a classification head on top of Beto and mBERT foundational models.

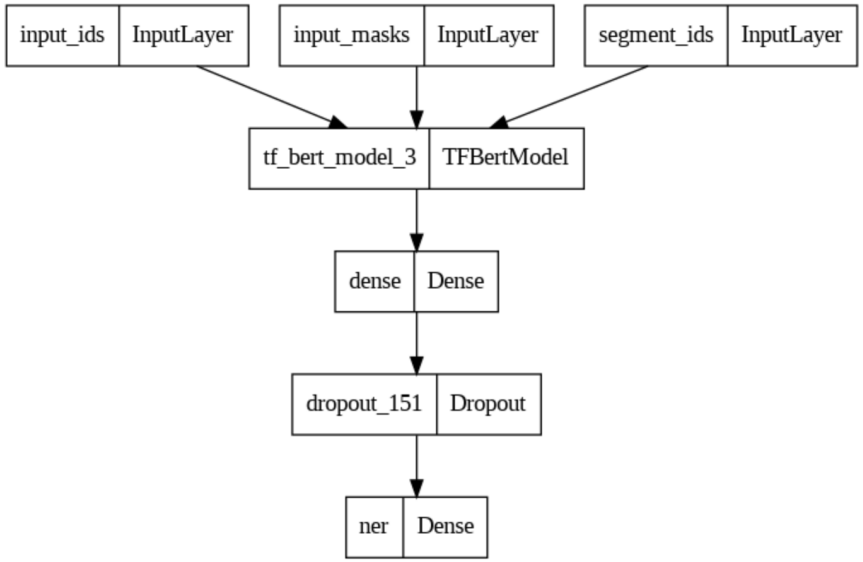
Ram Senthamarai

Results

Top model is Beto based model with all layers frozen while same with one 1 unfrozen layer comes close second.

- Test Accuracy with all layers frozen after 30 epochs: 0.882
- Validation Accuracy with all layers frozen after 30 epochs: 0.574
- Test Accuracy with 1 layer unfrozen after 30 epochs: 0.865
- Validation Accuracy with 1 layer unfrozen after 30 epochs: 0.699

Model Architecture



Classification layer added on top of BERT

Hyper Parameters

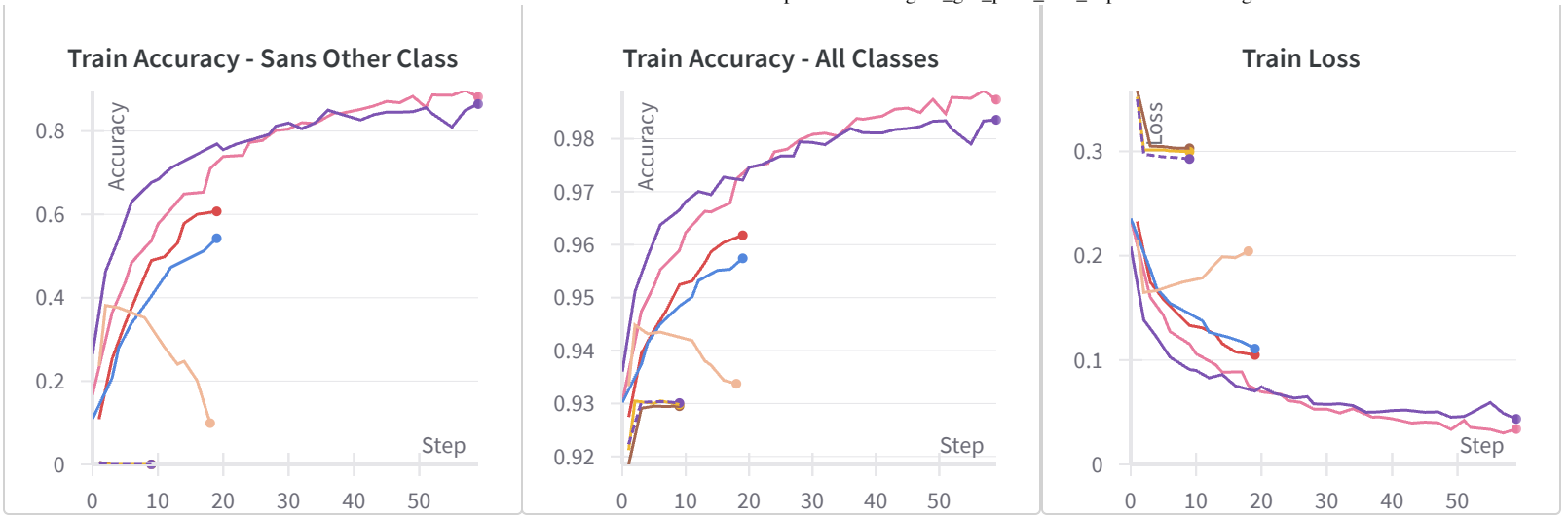
We ran multiple experiments with varying configurations:

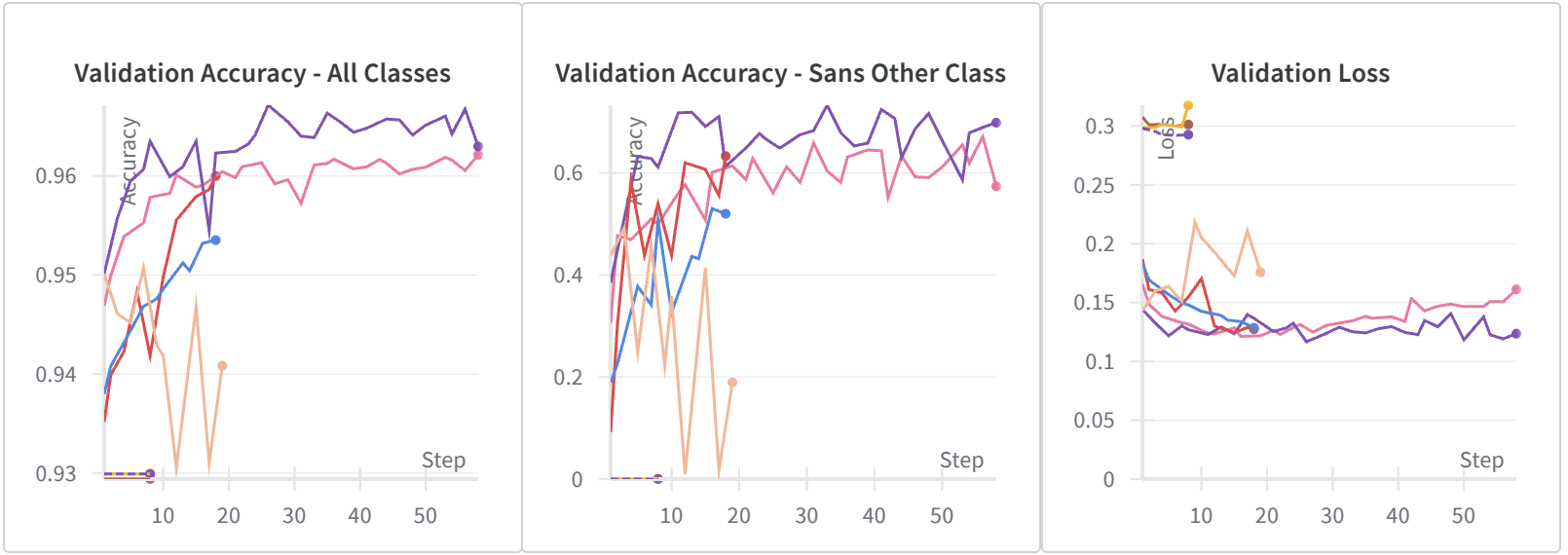
- Foundation model: mBERT or Beto
- Optimizer: Adam custom tweaked or Adam default
- Retrained Layers: None, All, Top 1, Top 2

Metrics

We address class imbalance (93% of tokens in training dataset belong to class "Other") by creating a custom evaluation metric and loss function that evaluates accuracy and cross-entropy loss after ignoring Other class tokens.

Findings





- Unfreezing more than 1 layer results in a very degraded model performance (attributable to lack of sufficient training data).

- Beto based model is slightly ahead of the mBert based model.
- Beto based model with all layers frozen seems to be slightly overfitting (a slight upward trend in validation loss) after 10 epochs or so. The trend is very small. Experimenting with some generalization techniques like adding a dropout layer might help.
- Looks like Beto based model with the top layer unfrozen is not overfitting as much as the other model
- Seems like both models can be trained for more epochs but the learning has slowed down. Tweaking optimizer parameters might help.

Next Steps

Given these findings we are going to focus on the Beto based model (with all layers frozen as well as 1 layer unfrozen) for next set of experiments:

- Add drop out layer to avoid overfitting
- Train for more epochs
- Hyper tuning of optimizer parameters
- Include F-Beta, Precision, Recall and Confusion Matrix in analysis.

Created with ❤️ on Weights & Biases.

https://wandb.ai/langbot/langbot_gec_plain_bert_experiments/reports/BERT-based-Vanilla-Models-Initial-Experiments--Vmlldzo2MDUwMTk2