Progress Report

Daniel Biro

February 21, 2021

Contents

1 Experiments - workflow version: 1.3.0; experiments revision: YYY **2**
1.1 Methods 2
1.2 Results 3

1 Experiments - workflow version: 1.3.0; experiments revision: YYY

1.1 Methods

This update includes two bigger changes to the project, the first being the tweaked data generation process and the second being the saving of intermediate models to produce graphs of the evolution of test set statistics over the course of the training.

The data generation was augmented according to what was discussed in the last meeting. The usage of the purity parameter was corrected such that purity % of reads are taken from the tumor sample at each type (unmutated, germline, somatic) and 1-purity % are taken from the normal sample, to build the separate features for the two samples. Gap counts are also considered now, bringing the features of the classification data up to 6 from the previous 4.

Now the germline mutations are considered to be the ones in the tumor VCF that are also present in the normal VCF, previously it was the other way around. This resulted in a smaller dataset since previously all normal VCF entries were designated as germline variants, but now only the tumor VCF entries are considered and split conditioned on being in the normal VCF as well (I will create a figure to make the change more clear in the thesis itself).

Dataset size changed in the following way for the training set:

- Germline variant: 5322 $\rightarrow$ 2464
- Somatic variant: 1872 $\rightarrow$ 1859
- Normal (unmutated, sum of the previous two): 7194 $\rightarrow$ 4323
- Total training set size: 12950 $\rightarrow$ 7782
- Total validation set size: 1438 $\rightarrow$ 864

And for the test set:

- Germline variant: 5264 $\rightarrow$ 2809
- Somatic variant: 2269 $\rightarrow$ 1691
- Normal (unmutated, sum of the previous two): 7533 $\rightarrow$ 4500
- Total test set size: 15066 $\rightarrow$ 9000

As can be seen, most of the loss comes in the germline variants category, probably due to what I mentioned in the paragraph above. Note that there was also been a change in the VCF of the normal sample since the last time the data was generated, which could also explain some of this.

For these germline variants now the reads from the tumor sample are copied to the normal sample with some noise added. We were discussing that this might make things more realistic and weaken the previously strong signal, but as presented in the results, it actually made the signal cleaner and stronger.

The other change was that now the models can be saved at every X epoch, controlled by the save_frequency parameter. I have been using 50, so over 2000 epochs of training there are 40 saves in total. For each saves the models of each run are evalated and graphs are drawn based on their performance, similarly as for the training and validation sets.

All new changes are up on GitHub, the documentation will be updated later.

1.2 Results

All of the models for both the classification and genotyping task have been re-run on the new data, as previously 10 times for 2000 epochs each.

Since the dataset size decreased, the running time of the models decreased as well. The new running times for 1000 epochs are roughly:

- GRU (and other basic models: 0:25 $\rightarrow$ 0:12
- Bidirectional GRU: 0:52 $\rightarrow$ 0:31
- Perceptron: 0:06 $\rightarrow$ 0:06
- Transformer 2:32 $\rightarrow$ 1:36
- Genotyping GRU (and other basic models: 0:20 $\rightarrow$ 0:12
- Genotyping Bidirectional GRU: 0:50 $\rightarrow$ 0:32
- Genotyping Perceptron: 0:06 $\rightarrow$ 0:06
- Genotyping Transformer 3:30 $\rightarrow$ 2:05

The tables and figures mentioned here will be included as separate files in the folder that this report arrives in to avoid clutter and the time intensive fiddling with latex tables.

First the classification task results are presented.

Table 1 showcases the comparison between the model results between training on the old and the new data across all metrics averaged for the three classes. Tables 2-4 show the training, validation and test set results for each of the three classes (germline variant, somatic variant, normal (unmutated)) separately. The gaps in the old data results are due to lack of time to train those models (the best peforming ones from the previous 1000 epoch long training experiments were prioritised).

While the preformance deterioriated for the simple GRU model, it has improved substantially for the perceptron, bidirectional with dropout and the Transformer, by roughly 3% on all datasets. This brings model performance near perfection on the training and validation set for many models and still very good on the test set, demonstrating some generalisation capability.

Based on test set performance a new best model emerged, the 1 encoder layer Transformer. While it performed relatively bad in the validation set, it topped the test set performance of the perceptron (best on training and validation both as in previous experiments) by roughly 1%.

All graphs for the 3 best performing models (Transformer, Bidirectional GRU with 5% dropout, Perceptron) are in the folders Figure 1-3.

For the genotyping task, the old and new results are displayed in table 5. The gaps again mean that those models were not considered due to lack of time. In the new results the only gap is the simple bidirectional GRU, which is currently being trained.

For this task the new data generation led to mixed results. While for GRU there was a slight increase in most metrics accross all datasets, the perceptron did substantially worse. The Transformer generally got worse everywhere as well but it still peformed better than the perceptron (2nd best model) on the test set and now it peforms better on the training and validation set as well.

Generally these models seems to overfit to the training data, as it can be observed from their graphs as well.

The graphs for Transformer, Bidirectional GRU with 5% dropout, Perceptron are in the folders Figure 4-6.