DeepVariant training using Parabricks

Table of contents

DeepVariant is a data analysis pipeline employing a deep neural network to identify genetic variants from next-generation DNA sequencing (NGS) data. While DeepVariant is exceptionally precise for various NGS data, there might be users keen on crafting tailored deep learning models meticulously suited for highly specific data.

The DeepVariant training pipeline has three major steps:

1. Run make_examples in “training” mode on the training and validation data sets,
2. Shuffle each set of examples and generate a data configuration file for each, and
3. Run model_train and model_eval.

Parabricks currently contains a GPU accelerated version of the first two steps.

Run make_examples in training mode

The "make_examples" step processes the input data, producing output suitable for use in subsequent steps. The output produced will include a label field.

Beginning with version 1.4.0, DeepVariant introduced an additional parameter in their WGS configuration through the `--channels "insert_size"` flag.

Depending on the nature of your data, you may wish to adjust the flags for the make_examples step, potentially leading to varying formats for the output examples. Please see the DeepVariant documentation for details regarding these options.

make_examples Quick Start

This code runs the "make_examples" step, combining the reference, BAM, VCF and BED files into a format suitable for use by the shuffle, model_train and model_eval steps.

```
# This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/deepvariant_train:4.2.0-1 \ pbrun make_examples \ --ref /workdir/${REFERENCE_FILE} \ --reads /workdir/${INPUT_BAM} \ --truth-variants /workdir/${TRUTH_VCF} \ --confident-regions /workdir/${TRUTH_BED} \ --examples /outputdir/${TFRECORD_FILE} \ --disable-use-window-selector-model \ --channel-insert-size
```

Compatible make_examples Baseline Command

```
( seq 0 $((N_SHARDS-1)) |\ parallel --halt 2 --line-buffer \ sudo docker run --volume <INPUT_DIR>:/workdir --volume <OUTPUT_DIR>:/outputdir \ google/deepvariant:"${BIN_VERSION"} \ /opt/deepvariant/bin/make_examples \ --mode training \ --ref "/workdir/${REF}" \ --reads "/workdir/${INPUT_BAM}" \ --examples "/outputdir/validation_set.with_label.tfrecord@${N_SHARDS}.gz" \ --truth_variants "/workdir/${TRUTH_VCF" \ --confident_regions "/workdir/${TRUTH_BED}" \ --task {} \ --channels "insert_size" )
```

make_examples Reference

Run `deepvariant make_examples` in training mode to create tensorflow.Examples.

make_examples Input/Output file options

--ref REF

Genome reference to use. Must have an associated FAI index as well. Supports text or gzipped references. Should match the reference used to align the BAM file provided to --reads. (default: None)

Option is required.
--reads READS

Aligned, sorted, indexed BAM file containing the reads we want to call. Should be aligned to a reference genome compatible with --ref. (default: None)

Option is required.

--interval-file INTERVAL_FILE

Path to a BED file (.bed) for selective access. This option can be used multiple times. (default: None)

--confident-regions CONFIDENT_REGIONS

Regions that we are confident are hom-ref or a variant in BED format. Contig names must match those of the reference genome. (default: None)

Option is required.

--truth-variants TRUTH_VARIANTS

Tabix-indexed VCF file containing the truth variant calls for this labels which we use to label our examples. (default: None)

Option is required.

--examples EXAMPLES

Path to write tf.Example protos in TFRecord format. (default: None)

Option is required.

--proposed-variants PROPOSED_VARIANTS

Path of the vcf.gz file, which has proposed variants for the make examples stage. (default: None)

make_examples Tool Options:

--num-cpu-threads-per-stream NUM_CPU_THREADS_PER_STREAM

Number of CPU threads to use per stream. (default: 6)

--num-zipper-threads NUM_ZIPPER_THREADS

Number of threads for compression and writing output files. (default: 4)

--num-streams-per-gpu NUM_STREAMS_PER_GPU

Number of streams to use per GPU. (default: 2)

--disable-use-window-selector-model

Change the window selector model from Allele Count Linear to Variant Reads. This option will increase the accuracy and runtime. (default: None)

--gvcf

Generate variant calls in .gvcf Format. (default: None)

--norealign-reads

Do not locally realign reads before calling variants. Reads longer than 500 bp are never realigned. (default: None)

--sort-by-haplotypes

Reads are sorted by haplotypes (using HP tag). (default: None)

--keep-duplicates

Keep reads that are duplicate. (default: None)

--vsc-min-count-snps VSC_MIN_COUNT_SNPS

SNP alleles occurring at least this many times in the AlleleCount will be advanced as candidates. (default: 2)

--vsc-min-count-indels VSC_MIN_COUNT_INDELS

Indel alleles occurring at least this many times in the AlleleCount will be advanced as candidates. (default: 2)

--vsc-min-fraction-snps VSC_MIN_FRACTION_SNPS

SNP alleles occurring at least this fraction of all counts in the AlleleCount will be advanced as candidates. (default: 0.12)

--vsc-min-fraction-indels VSC_MIN_FRACTION_INDELS

Indel alleles occurring at least this fraction of all counts in the AlleleCount will be advanced as candidates. (default: None)

--min-mapping-quality MIN_MAPPING_QUALITY

By default, reads with any mapping quality are kept. Setting this field to a positive integer i will only keep reads that have a MAPQ >= i. Note this only applies to aligned reads. (default: 5)

--min-base-quality MIN_BASE_QUALITY

Minimum base quality. This option enforces a minimum base quality score for alternate alleles. Alternate alleles will only be considered if all bases in the allele have a quality greater than min_base_quality. (default: 10)

--mode MODE

Value can be one of [shortread, pacbio, ont]. By default, it is shortread. If mode is set to pacbio, the following defaults are used: --norealign-reads, --alt-aligned-pileup diff_channels, --vsc-min-fraction-indels 0.12. If mode is set to ont, the following defaults are used: -norealign-reads, --variant-caller VCF_CANDIDATE_IMPORTER. (default: shortread)

--alt-aligned-pileup ALT_ALIGNED_PILEUP

Value can be one of [none, diff_channels]. Include alignments of reads against each candidate alternate allele in the pileup image. (default: None)

--variant-caller VARIANT_CALLER

Value can be one of [VERY_SENSITIVE_CALLER, VCF_CANDIDATE_IMPORTER]. The caller to use to make examples. If you use VCF_CANDIDATE_IMPORTER, it implies force calling. Default is VERY_SENSITIVE_CALLER. (default: None)

--add-hp-channel

Add another channel to represent HP tags per read. (default: None)

--parse-sam-aux-fields

Auxiliary fields of the BAM/CRAM records are parsed. If either --sort-by-haplotypes or --add-hp-channel is set, then this option must also be set. (default: None)

--use-wes-model

If passed, the WES model file will be used. Only used in shortread mode. (default: None)

--run-partition

Divide the whole genome into multiple partitions and run multiple processes at the same time, each on one partition. (default: None)

--gpu-num-per-partition GPU_NUM_PER_PARTITION

Number of GPUs to use per partition. (default: None)

--include-med-dp

If True, include MED (default: None)

--normalize-reads

If True, allele counter left align INDELs for each read. (default: None)

--pileup-image-width PILEUP_IMAGE_WIDTH

Pileup image width. Only change this if you know your model supports this width. (default: 221)

--channel-insert-size

If True, add insert_size channel into pileup image. By default, this parameter is true in WGS and WES mode. (default: None)

--no-channel-insert-size

If True, don't add insert_size channel into the pileup image. (default: None)

--max-read-size-512

Allow deepvariant to run on reads of size 512bp. The default size is 320 bp. (default: None)

--prealign-helper-thread

Use an extra thread for the pre-align step. This parameter is more useful when --max-reads-size-512 is set. (default: None)

--max-reads-per-partition MAX_READS_PER_PARTITION

The maximum number of reads per partition that are considered before following processing such as sampling and realignment. (default: 1500)

--partition-size PARTITION_SIZE

The maximum number of basepairs allowed in a region before splitting it into multiple smaller subregions. (default: 1000)

--track-ref-reads

If True, allele counter keeps track of reads supporting ref. By default, allele counter keeps a simple count of the number of reads supporting ref. (default: None)

--phase-reads

Calculate phases and add HP tag to all reads automatically. (default: None)

--dbg-min-base-quality DBG_MIN_BASE_QUALITY

Minimum base quality in a k-mer sequence to consider. (default: 15)

--ws-min-windows-distance WS_MIN_WINDOWS_DISTANCE

Minimum distance between candidate windows for local assembly (default: 80)

--channel-gc-content

If True, add gc (default: None)

--channel-hmer-deletion-quality

If True, add hmer deletion quality channel into pileup image (default: None)

--channel-hmer-insertion-quality

If True, add hmer insertion quality channel into pileup image (default: None)

--channel-non-hmer-insertion-quality

If True, add non-hmer insertion quality channel into pileup image (default: None)

--skip-bq-channel

If True, ignore base quality channel. (default: None)

--aux-fields-to-keep AUX_FIELDS_TO_KEEP

Comma-delimited list of auxiliary BAM fields to keep. Values can be [HP, tp, t0] (default: HP)

--vsc-min-fraction-hmer-indels VSC_MIN_FRACTION_HMER_INDELS

Hmer Indel alleles occurring at least this be advanced as candidates. Use this threshold if hmer and non-hmer indels should be treated differently (Ultima reads)Default will use the same threshold for hmer and non-hmer indels, as defined in vsc_min_fraction_indels. (default: None)

--vsc-turn-on-non-hmer-ins-proxy-support

Add read-support from soft-clipped reads and other non-hmer insertion alleles,to the most frequent non-hmer insertion allele. (default: None)

--consider-strand-bias

If True, expect SB field in calls and write it to vcf (default: None)

--p-error P_ERROR

Basecalling error for reference confidence model. (default: 0.001)

--channel-ins-size

If true, add another channel to represent size of insertions. (good for flow-based sequencing) (default: None)

--max-ins-size MAX_INS_SIZE

Max insertion size for ins_size_channel, larger insertions will look like max (have max intensity) (default: 10)

--disable-group-variants

If using vcf_candidate_importer and multi-allelic sites are split across multiple lines in VCF, set to True so that variants are not grouped when transforming CallVariantsOutput to Variants. (default: None)

-L INTERVAL, --interval INTERVAL

Interval within which to call the variants from the BAM/CRAM file. Overlapping intervals will be combined. Interval files should be passed using the --interval-file option. This option can be used multiple times (e.g. "-L chr1 -L chr2:10000 -L chr3:20000+ -L chr4:10000-20000"). (default: None)

Common options:

--logfile LOGFILE

Path to the log file. If not specified, messages will only be written to the standard error output. (default: None)

--tmp-dir TMP_DIR

Full path to the directory where temporary files will be stored.

--with-petagene-dir WITH_PETAGENE_DIR

Full path to the PetaGene installation directory. By default, this should have been installed at /opt/petagene. Use of this option also requires that the PetaLink library has been preloaded by setting the LD_PRELOAD environment variable. Optionally set the PETASUITE_REFPATH and PGCLOUD_CREDPATH environment variables that are used for data and credentials (default: None)

--keep-tmp

Do not delete the directory storing temporary files after completion.

--no-seccomp-override

Do not override seccomp options for docker (default: None).

--version

View compatible software versions.

GPU options:

--num-gpus NUM_GPUS

Number of GPUs to use for a run. GPUs 0..(NUM_GPUS-1) will be used.

Run shuffle

The shuffling of TensorFlow example data is a crucial stage in model training. In the DeepVariant training process the examples are globally shuffled as part of the preprocessing step.

This script shuffles TensorFlow records locally and in-memory.

shuffle Quick Start

> *# This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR.* docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \
> nvcr.io/nvidia/clara/deepvariant_train:4.2.0-1 \ pbrun shuffle \ --input-pattern-list /workdir/validation_set.with_label.tfrecord-?????-of-00016.gz \ --output-pattern-prefix /outputdir/validation_set.with_label.shuffled \ --output-dataset-config-pbtxt /outputdir/validation_set.dataset_config.pbtxt \ --output-dataset-name HG001 \ --direct-num-workers 16

Compatible shuffle Baseline Command

```
python3 shuffle_tfrecords_lowmem.py \ --
input_pattern_list="${INPUT_DIR}/validation_set.with_label.tfrecord=?????-of-
00016.gz" \ --
output_pattern_prefix="${OUTPUT_DIR}/validation_set.with_label.shuffled" \ --
output_dataset_config="${OUTPUT_DIR}/validation_set.dataset_config.pbtxt" \ --
putput_dataset_name="HG001" \ --direct_num_workders=16 \ --step=1
```

shuffle Reference

Shuffle examples globally.

Shuffle Input/Output file options

--output-dataset-config-pbtxt OUTPUT_DATASET_CONFIG_PBTXT

Human-readable version of DeepVariantDatasetConfig. (default: None)

Option is required.

--input-pattern-list INPUT_PATTERN_LIST [INPUT_PATTERN_LIST ...]

TFRecord filename pattern. (default: None)

Option is required.

Shuffle Tool Options:

--output-pattern-prefix OUTPUT_PATTERN_PREFIX

Filename pattern for the output TFRecords. (default: None)

Option is required.

--output-dataset-name OUTPUT_DATASET_NAME

Option is required.

--direct-num-workers DIRECT_NUM_WORKERS

Number of writer threads (default: 1)

Common options:

--logfile LOGFILE

Path to the log file. If not specified, messages will only be written to the standard error output. (default: None)

--tmp-dir TMP_DIR

Full path to the directory where temporary files will be stored.

--with-petagene-dir WITH_PETAGENE_DIR

Full path to the PetaGene installation directory. By default, this should have been installed at /opt/petagene. Use of this option also requires that the PetaLink library has been preloaded by setting the LD_PRELOAD environment variable. Optionally set the PETASUITE_REFPATH and PGCLOUD_CREDPATH environment variables that are used for data and credentials (default: None)

--keep-tmp

Do not delete the directory storing temporary files after completion.

--no-seccomp-override

Do not override seccomp options for docker (default: None).

--version

View compatible software versions.

GPU options:

--num-gpus NUM_GPUS

Number of GPUs to use for a run. GPUs 0..(NUM_GPUS-1) will be used.

Run model_train and model_eval

We provide a Jupyter Notebook with a more detailed example of re-training DeepVariant 1.5 using Parabricks and additional instructions on the *model_train* and *model_eval* steps.

See also the DeepVariant training documentation, and the original Shuffle program.