Quantization Aware Training – AIMET TensorFlow

Quantization simulation and finetuning using the AIMET library.

The general procedure for quantization is to use AIMET's QuantizationSimModel to compute new encodings, then finetune the model.

```
from aimet_tensorflow.quantsim import QuantizationSimModel
```

# This script utilizes AIMET to perform Quantization aware training on a resnet50	
	# pretrained model with the ImageNet data set.This is intended as a working example
	# to show how AIMET APIs can be invoked.
	# Scenario parameters:
	AIMET quantization aware training using simulation model - QuantizationSimModel
	Quant Scheme: 'tf'
	rounding_mode: 'nearest'
	default_output_bw: 8, default_param_bw: 8
	Encoding computation using 5 batches of data
	Input shape: [1, 3, 224, 224]
	Learning rate: 0.001
	Decay Steps: 5

Evaluate

	Evaluate the specified session using the specified number of samples from the validation set.
	AIMET's QuantizationSimModel.compute_encodings() expects the function with this signature
	to its eval_callback parameter.
	:param sess: The sess graph to be evaluated.
	:param iterations: The number of batches of the dataset.

	:return: The accuracy for the sample with the maximum accuracy.

Train

Trains the session graph. The implementation provided here is just an example,	
	provide your own implementation if needed.
	:param sess: The sess graph to train.
	:param update_ops_name: list of name of update ops (mostly BatchNorms' moving averages).
	tf.GraphKeys.UPDATE_OPS collections is always used
	in addition to this list

create_quant_sim_model

Apply quantizer simulator on the original model and return its object.	
	:param sess: The sess with graph.
	:param start_op_names: The list of input op names of the sess.graph
	:param output_op_names: The list of output op names of the sess.graph
	:param use_cuda: If True then use a GPU for QuantizationSimModel
	:param parity_config_file: Config file for H/W parity
	:param evaluator: A callback function that is expected to run forward passes on a session
	:return: QuantizationSimModel object

perform_qat (Quantization Aware Training)

1. Instantiates Data Pipeline for evaluation and training	
2. Loads the pretrained resnet50 keras model	
3. Calculates floating point accuracy	

4. Quantization Sim Model	
4.1. Creates Quantization Sim model using AIMET QuantizationSimModel	
4.2. Calculates and logs the accuracy of quantizer sim model	
5. Quantization Aware Training	
5.1. Trains the quantization aware model	
5.2. Calculates and logs the accuracy of quantization Aware trained model	
5.3. Exports quantization aware model so it is ready to be run on-target	
	:param config: This argparse.Namespace config expects following parameters:
	tfrecord_dir: Path to a directory containing ImageNet TFRecords.
	This folder should contain files starting with:
	'train*': for training records and 'validation*': for validation records
	parity_config_file: An optional parity config file, used in Quantizer
	use_cuda: A boolean var to indicate to run the test on GPU.
	logdir: Path to a directory for logging.
	epochs: Number of epochs (type int) for training.
	learning_rate: A float type learning rate for model training
	decay_steps: A number used to adjust(decay) the learning rate after every decay_steps
	epochs in training.

References

https://github.com/quic/aimet/blob/develop/Examples/tensorflow/quantization/qat.py