Define operator attributes and add data-driven post-training sparsificaiton capabilities

Manuj Sabharwal, Ken Koyanagi

Intel Corporation

1. Operator attribute

Background

- ONNX is popular for its model portability and supports weights sparsity
  - However, it can be difficult to determine:
    - Whether the model is a sparse model
    - What the sparsity format structure is
- There is currently `SparseTensor`

```
message SparseTensor {
  // This field MUST NOT have the value of UNDEFINED
  // This field MUST have a valid TensorProto.DataType value
  // This field MUST be present for this version of the IR.
  int32 elem_type = 1;
  TensorShapeProto shape = 2;
}
```

Proposal Idea

- ONNX operators support a sparsity attribute which indicate the following
  - sparsity : format structure
    - *e.g., unstructured, COO, 2:4, BSR, etc.*
  - sparsity % : float
    - *e.g., 0.50*
- This helps the give useful insights when viewing the portable model.
- ONNX runtime execution providers can determine how to use the sparse tensors or whether to use them as sparse representations.

Goal: Improve sparsity visibility within the community

2. Data-driven post-training sparsification

Background

- There is no single “best” sparsity format that is agreed upon at this stage.
- In turn, this creates a barrier to entry to evaluate sparsity as an interest area.
- Currently, there is no turn-key solution to see how a model would benefit from exploring sparsity with ONNX.
  - E.g., are there disk space benefits with Model A?
  - There are other toolkits which offer a single-line API to generate a sparse model based on a threshold or percentile.
- Common feedback we’ve heard is that it’s difficult to spend resources without knowing the roofline or potential.

Proposal:

Request for Analysis

Explore simple low-barrier approaches for data-free and data-driven post-training sparsity.

** details for how low-effort (e.g., sensitivity analysis) or already available tool adoption can be discussed as part of the exploration.*

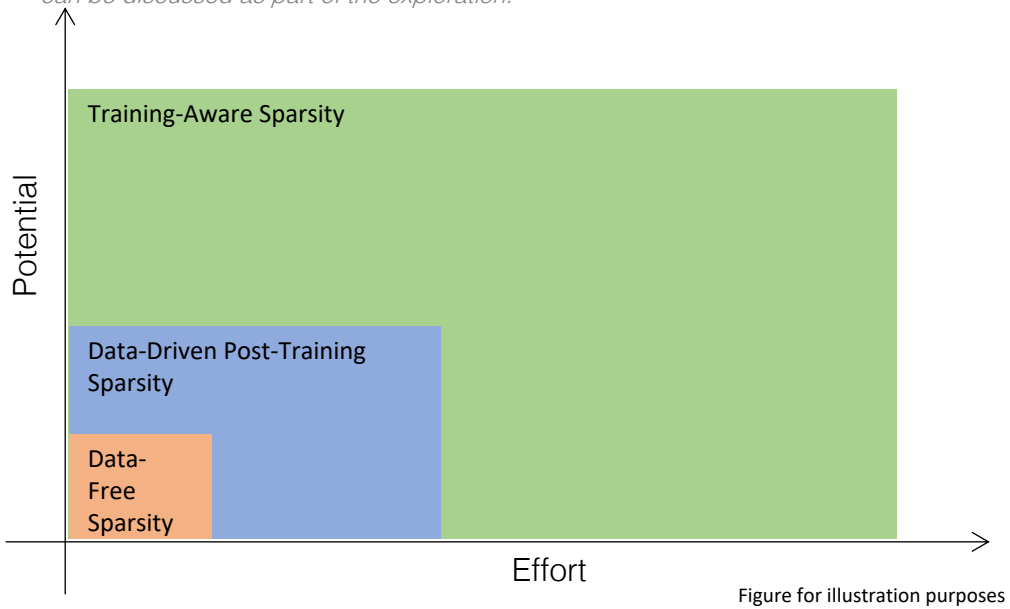

Figure for illustration purposes only.

Goal: Decrease the amount of time or effort needed to quickly see the potential of sparsity

Questions/Comments Welcome:

manuj.r.sabharwal@intel.com

ken.koyanagi@intel.com