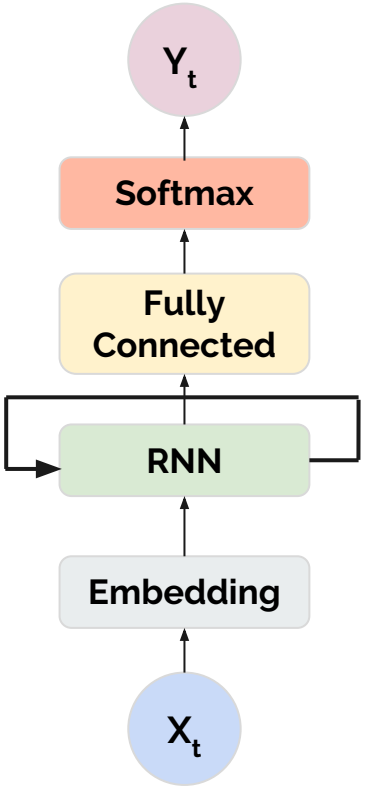
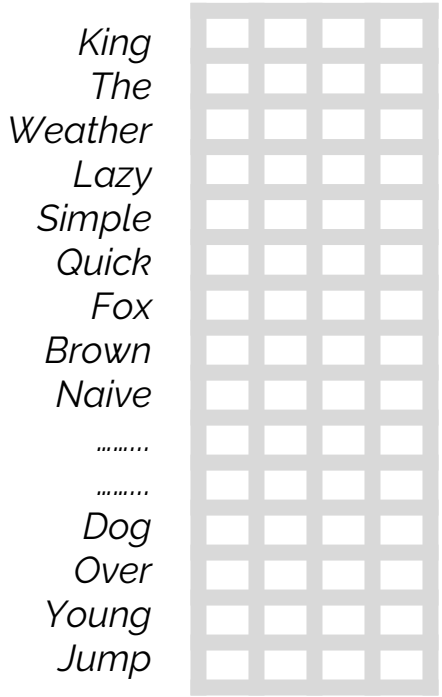
Introducing Sparse Tensor in Apache MXNet

Haibin Lin, @eric-haibin-lin

Motivation - Language Model

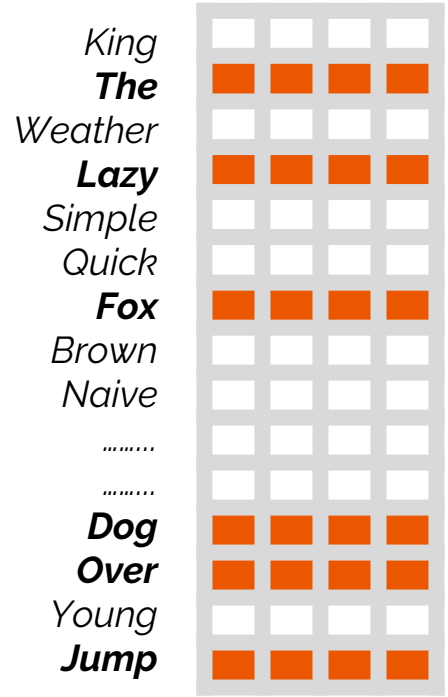


Gradients for Embeddings are Sparse



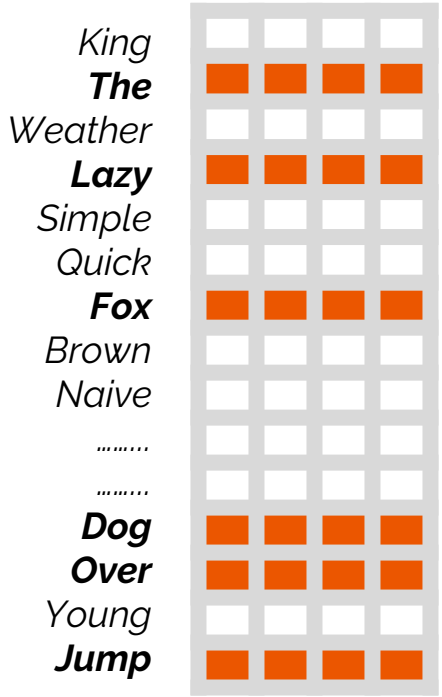
"The fox jumps over the lazy dog"

Gradients for Embeddings are Sparse

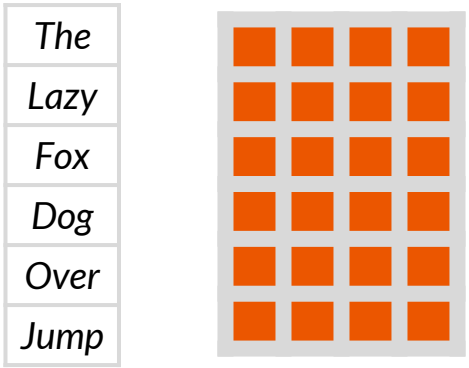


"The fox jumps over the lazy dog"

Gradients for Embeddings are Sparse



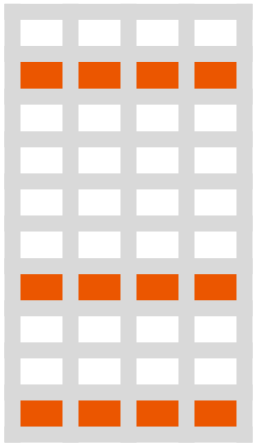
"The fox jumps over the lazy dog"



indices **data**

MXNet Sparse Tensor Types

- **Sparse Gradient (e.g. word embeddings in NLP)**
- Sparse Data (e.g. recommender systems, social networks, computational ads)

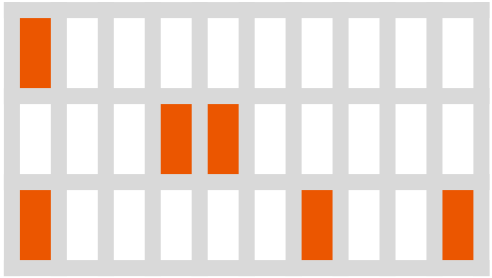


Row Sparse format

mx.nd.sparse.RowSparseArray

MXNet Sparse Tensor Types

- Sparse Gradient (e.g. word embeddings in NLP)
- **Sparse Data (e.g. recommender systems, social networks, computational ads)**



Compressed Sparse Row (CSR) format

mx.nd.sparse.CSRNDArray

MXNet Sparse Feature

- **Sparse Data Formats**
- Sparse Operators
- Multi-GPU & Multi-Machine Communication

Loading Sparse Data

```
>>> import mxnet as mx
>>> import scipy.sparse as spsp
>>> sp_matrix = spsp.rand(2, 3, density=0.3, dtype='float32')
>>> print(sp_matrix.toarray())
[[ 0.          0.          0.89449775 ]
 [ 0.          0.57767099  0.         ]]
```

Loading Sparse Data

```
>>> import mxnet as mx
>>> import scipy.sparse as spsp
>>> sp_matrix = spsp.rand(2, 3, density=0.3, dtype='float32')
>>> print(sp_matrix.toarray())
[[ 0.         0.         0.89449775 ]
 [ 0.         0.57767099 0.         ]]

>>> x = mx.ndarray.sparse.csr_matrix(sp_matrix, ctx=mx.cpu())
>>> x
<CSRNDArray 2x3 @cpu(0)>
```

MXNet Sparse Feature

- Sparse Data Formats
- **Sparse Operators**
- Multi-GPU & Multi-Machine Communication

Sparse Matrix-Dense Vector Multiplication

```
>>> w = mx.ndarray.random.uniform(shape=(3, 1))
>>> w
[[ 0.54881352]
 [ 0.59284461]
 [ 0.85794562]]
<NDArray 3x1 @cpu(0)>
>>> y = mx.ndarray.sparse.dot(x, w)
>>> y
[[ 0.75519383]
 [ 0.34246913]]
<NDArray 2x1 @cpu(0)>
```

Sparse Matrix-Dense Vector Multiplication

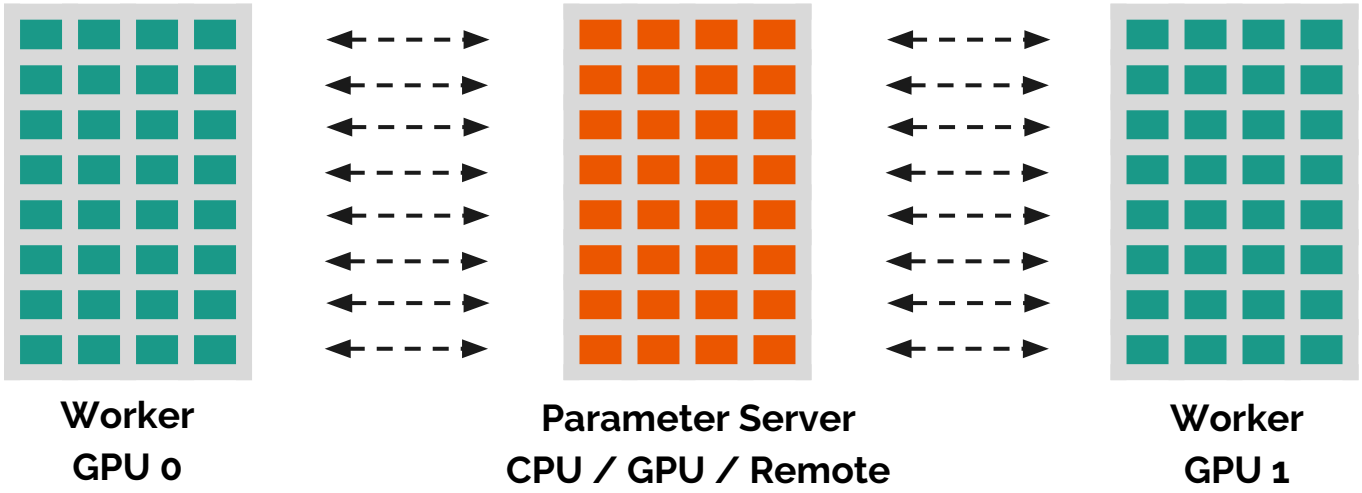
- KDD Cup 2010 dataset on r4.8xlarge (CPU)

	Dense Dot	Sparse Dot
Speed	121.38 ms	0.21 ms
Memory	~ 5 GB	~ 100 MB

MXNet Sparse Feature

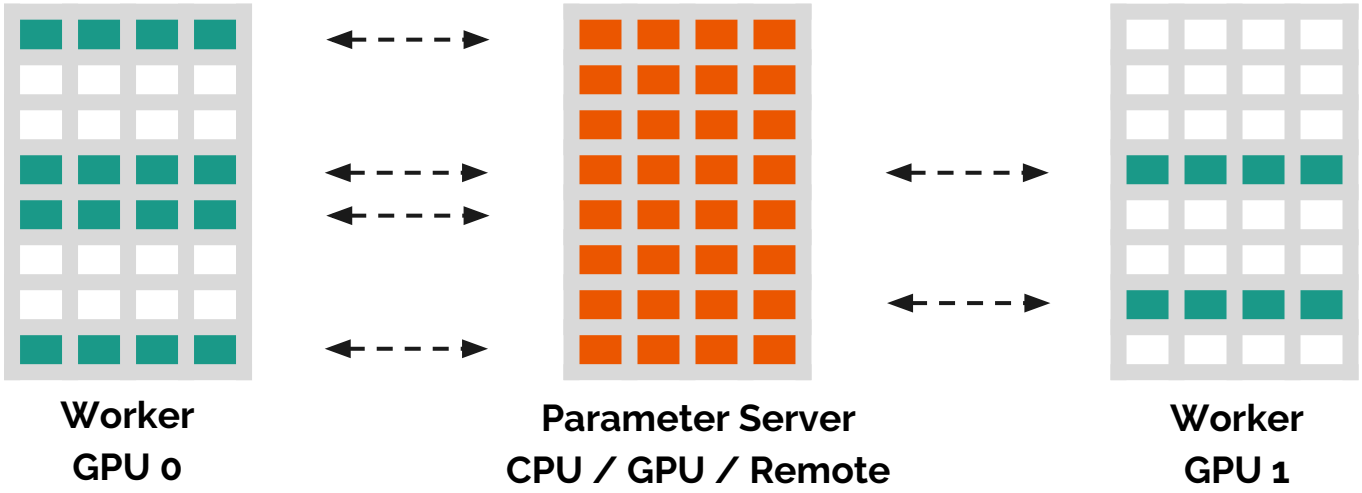
- Sparse Data Formats
- Sparse Operators
- **Multi-GPU & Multi-Machine Communication**

Multi-GPU & Multi-Machine Communication



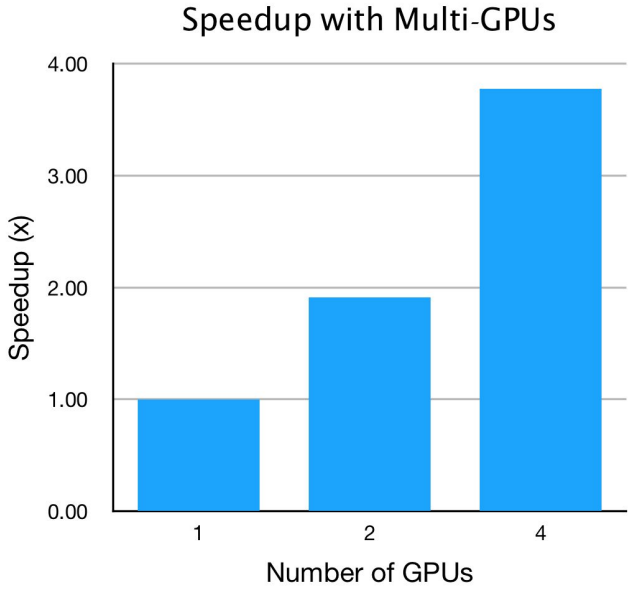
- Broadcasting the full model consumes lots of bandwidth

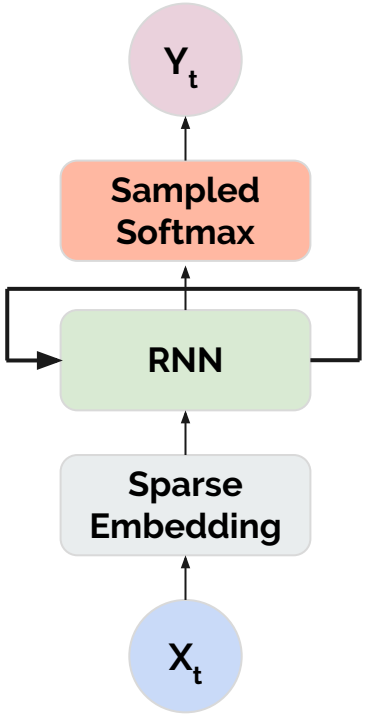
Multi-GPU & Multi-Machine Communication



- Broadcasting the partial model required for computation

Scalability Benchmarks





LSTM 2048-512 on Google Billion Words dataset, p2.4xlarge

Getting Started

Examples

Tutorials

Feature Request

https://github.com/apache/incubator-mxnet/**issues**