GloVe on Spark

Alex Adamson

Global Vectors for Word Representation (Pennington et. al, 2015)

- Form word representations by factorizing matrix of context-weighted cooccurrence counts:

$$J = \sum_{i,j=1}^{|C|} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

$$f(x) = \begin{cases} (x/x_{max})^\alpha & x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$

Distributing GloVe

- Main operations:
  - Full matrix multiplies
    - Dot-products of word vectors, context word vectors
    - Applying chain rule: $W \leftarrow W - \eta \cdot (f(X) \circ (W^T\tilde{W} + B + \tilde{B} - \log X) \cdot \tilde{W})$
    - Communication not terrible if matrices are partitioned the same (same cols/rows per block)
  - Elementwise multiply/divide, add/subtract, application of Double => Double:
    - Used heavily in computing cost/gradients:

$$\frac{\partial J}{\partial (W^T\tilde{W} + B + \tilde{B} - \log X)} \propto f(X) \circ (W^T\tilde{W} + B + \tilde{B} - \log X)$$

    - Can be communication free if both are block matrices, same size, same cols/rows per block
- Keys:
  - Avoiding conversions between matrix types
  - Applying elementwise multiplication by f(X) as early as possible to make cost/partial gradient matrix as sparse as possible before other operations (empty blocks => no multiplication)