Word Embeddings
GloVe

國立臺灣大學 資訊工程學系
陳縕儂 助理教授
http://vivianchen.idv.tw

Comparison

- Count-based
  - LSA, HAL (Lund & Burgess), COALS (Rohde et al), Hellinger-PCA (Lebret & Collobert)
  - Pros
    - ✓ Fast training
    - ✓ Efficient usage of statistics
  - Cons
    - ✓ Primarily used to capture word similarity
    - ✓ Disproportionate importance given to large counts

- Direct prediction
  - NNLM, HLBL, RNN, Skipgram/CBOW, (Bengio et al; Collobert & Weston; Huang et al; Mnih & Hinton; Mikolov et al; Mnih & Kavukcuoglu)
  - Pros
    - ✓ Generate improved performance on other tasks
    - ✓ Capture complex patterns beyond word similarity
  - Cons
    - ✓ Benefits mainly from large corpus
    - ✓ Inefficient usage of statistics

Combining the benefits from both worlds → GloVe

GloVe

- Idea: ratio of co-occurrence probability can encode meaning
- P_{ij} is the probability that word w_j appears in the context of word w_i

$$P_{ij} = P(w_j \mid w_i) = X_{ij}/X_i$$

- Relationship between the words w_i and w_j

	x = solid	x = gas	x = water	x = random
$P(x \mid \text{ice})$	large	small	large	small
$P(x \mid \text{stream})$	small	large	large	small
$\frac{P(x \mid \text{ice})}{P(x \mid \text{stream})}$	large	small	~ 1	~ 1

Pennington et al., "GloVe: Global Vectors for Word Representation," in EMNLP, 2014.

GloVe

- The relationship of w_i and w_j approximates the ratio of their co-occurrence probabilities with various w_k

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

$$F((v_{w_i} - v_{w_j})^T v'_{\tilde{w}_k}) = \frac{P_{ik}}{P_{jk}} \qquad F(\cdot) = \exp(\cdot)$$

$$v_{w_i} \cdot v'_{\tilde{w}_k} = v_{w_i}^T v'_{\tilde{w}_k} = \log P(w_k \mid w_i)$$

Pennington et al., "GloVe: Global Vectors for Word Representation," in EMNLP, 2014.

GloVe

$$v_{w_i} \cdot v'_{\tilde{w}_j} = v_{w_i}^T v'_{\tilde{w}_j} = \log P(w_j \mid w_i)$$

$$= \log P_{ij} = \log(X_{ij}) - \log(X_i)$$

$$P_{ij} = X_{ij}/X_i$$

$$v_{w_i}^T v'_{\tilde{w}_j} + b_i + \tilde{b}_j = \log(X_{ij})$$

$$C(\theta) = \sum_{i,j=1}^{V} f(P_{ij})(v_{w_i} \cdot v'_{\tilde{w}_j} - \log P_{ij})^2$$

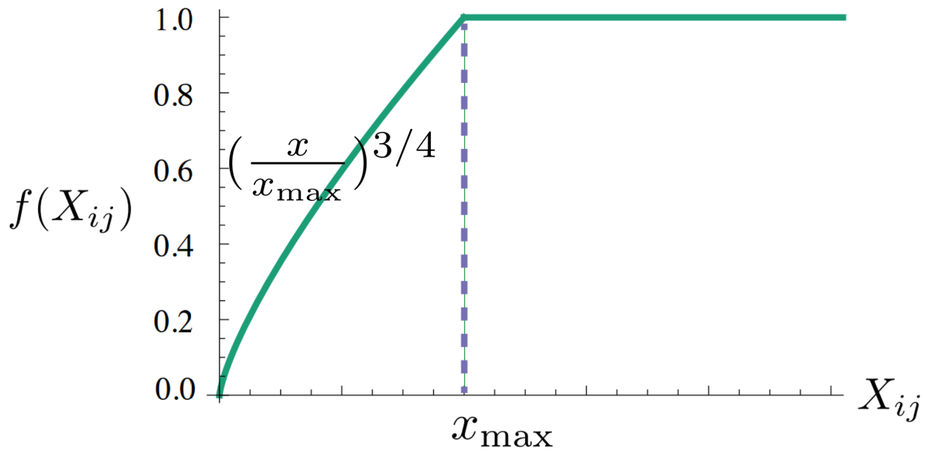
$$C(\theta) = \sum_{i,j=1}^{V} f(X_{ij})(v_{w_i}^T v'_{\tilde{w}_j} + b_i + \tilde{b}_j - \log X_{ij})^2$$

Pennington et al., "GloVe: Global Vectors for Word Representation," in EMNLP, 2014.

GloVe – Weighted Least Squares Regression Model

$$C(\theta) = \sum_{i,j=1}^{V} f(X_{ij})(v_{w_i}^T v'_{\tilde{w}_j} + b_i + \tilde{b}_j - \log X_{ij})^2$$

- Weighting function should obey
  - $f(0) = 0$
  - $f(x)$ should be non-decreasing so that *rare co-occurrences* are not overweighted
  - $f(x)$ should be relatively small for large values of x, so that *frequent co-occurrences* are not overweighted



fast training, scalable, good performance even with small corpus, and small vectors

Pennington et al., "GloVe: Global Vectors for Word Representation," in EMNLP, 2014.