UC San Diego

Neural Word Embedding as Implicit Matrix Factorization

By Omer Levy and Yoav Goldberg, Neural Information Processing Systems 2014

Presented by Ronald Yu

Background: SGNS (Word2Vec)

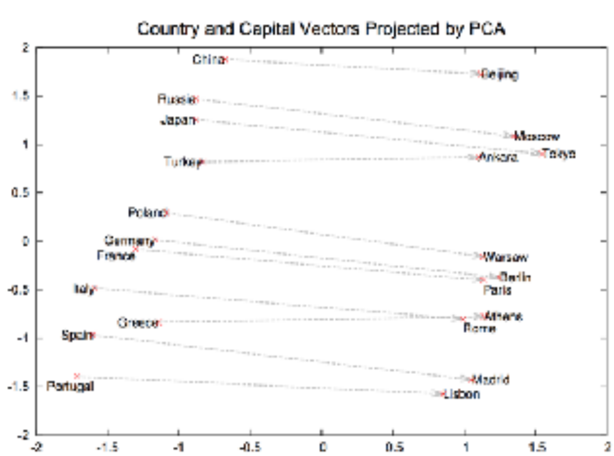
Maximize probability of word-context pairs appearing

$$P(D=1|w,c) = \sigma(\vec{w}\cdot\vec{c}) = \frac{1}{1+e^{-\vec{w}\cdot\vec{c}}}$$

Enhance algorithm by randomly sampling negative examples of word-context pairs

$$\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\mathbb{E}_{c_N\sim P_D}\left[\log\sigma(-\vec{w}\cdot\vec{c}_N)\right] \qquad P_D(c) = \frac{\#(c)}{|D|}$$

$$\ell = \sum_{w\in V_W}\sum_{c\in V_C}\#(w,c)\left(\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\mathbb{E}_{c_N\sim P_D}\left[\log\sigma(-\vec{w}\cdot\vec{c}_N)\right]\right)$$



SGNS As Implicit Matrix Factorization

- Goal of this paper: Formulate an explicit representation of what exactly SGNS is trying to optimize
- We have word matrix *W* and context matrix *C*, consider:

$$W \cdot C^{\top} = M \quad \text{of dimensions } |V_W| \times |V_C|$$

- Can we explicitly formulate *M* based on the loss function of SGNS?

SGNS As Implicit Matrix Factorization

For sufficiently large dimensionality, we have global objective

$$\ell = \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \left(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right] \right)$$

$$= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \left(\log \sigma(\vec{w} \cdot \vec{c}) \right) + \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \left(k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right] \right)$$

$$= \sum_{w \in V_W} \sum_{c \in V_C} \#(w,c) \left(\log \sigma(\vec{w} \cdot \vec{c}) \right) + \sum_{w \in V_W} \#(w) \left(k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right] \right)$$

$$\mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right] = \sum_{c_N \in V_C} \frac{\#(c_N)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c}_N)$$

This gives us a local objective for a specific word context pair (w,c):

$$\ell(w,c) = \#(w,c) \log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \#(w) \cdot \frac{\#(c)}{|D|} \log \sigma(-\vec{w} \cdot \vec{c})$$

SGNS As Implicit Matrix Factorization

$$\ell(w,c) = \#(w,c)\log\sigma(\vec{w}\cdot\vec{c}) + k\cdot\#(w)\cdot\frac{\#(c)}{|D|}\log\sigma(-\vec{w}\cdot\vec{c})$$

$$\frac{\partial\ell}{\partial x} = \#(w,c)\cdot\sigma(-x) - k\cdot\#(w)\cdot\frac{\#(c)}{|D|}\cdot\sigma(x)$$

Setting the derivative to zero, we obtain:

$$e^{2x} - \left(\frac{\#(w,c)}{k\cdot\#(w)\cdot\frac{\#(c)}{|D|}} - 1\right)e^{x} - \frac{\#(w,c)}{k\cdot\#(w)\cdot\frac{\#(c)}{|D|}} = 0$$

If we define $y = e^x$,

$$y = \frac{\#(w,c)}{k\cdot\#(w)\cdot\frac{\#(c)}{|D|}} = \frac{\#(w,c)\cdot|D|}{\#w\cdot\#(c)}\cdot\frac{1}{k}$$

$$\vec{w}\cdot\vec{c} = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)}\cdot\frac{1}{k}\right) = \log\left(\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)}\right) - \log k$$

In summary,

$$M_{ij}^{\text{SGNS}} = W_i\cdot C_j = \vec{w}_i\cdot\vec{c}_j = PMI(w_i,c_j) - \log k$$

Pointwise Mutual Information (PMI)

- Definition: $PMI(x,y) = \log\frac{P(x,y)}{P(x)P(y)}$
- We can estimate PMI as

$$PMI(w,c) = \log\frac{\#(w,c)\cdot|D|}{\#(w)\cdot\#(c)}$$

- Problems with PMI:
  - Computationally expensive
  - Ill defined:

$$PMI(w,c) = \log 0 = -\infty.$$

- Solution: Positive Pointwise Mutual Information

$$PPMI(w,c) = \max\left(PMI\left(w,c\right),0\right)$$

SGNS as Matrix Factorization

- We can represent more negative samples with shifted PPMI:

$$SPPMI_k(w, c) = \max\left(PMI\left(w, c\right) - \log k, 0\right)$$

- Obtain low-dimensionality representations with SVD:

$$W^{\text{SVD}_{1/2}} = U_d \cdot \sqrt{\Sigma_d} \qquad C^{\text{SVD}_{1/2}} = V_d \cdot \sqrt{\Sigma_d}$$

Matrix Factorization vs SGNS

- Advantages:
  - No hyper-parameters
  - works on larger datasets
- Disadvantages:
  - SGNS better favors reducing the loss of frequent word-context pairs
  - SGNS distinguishes between observed and negative samples
  - SGNS does not require a sparse matrix

Comparison of SGNS and Matrix Factorization

Method	PMI$-\log k$	SPPMI	SVD			SGNS		
			$d=100$	$d=500$	$d=1000$	$d=100$	$d=500$	$d=1000$
$k=1$	0%	0.00009%	26.1%	25.2%	24.2%	31.4%	29.4%	7.40%
$k=5$	0%	0.00004%	95.8%	95.1%	94.9%	39.3%	36.0%	7.13%
$k=15$	0%	0.00002%	266%	266%	265%	7.80%	6.37%	5.97%

Table 1: Percentage of deviation from the optimal objective value (lower values are better). See 5.1 for details.

Optimal Value: $\vec{w} \cdot \vec{c} = \text{PMI}(w,c) - \log k.$

SPPMI: $\vec{w} \cdot \vec{c} = \max(\text{PMI}(w,c) - \log k, 0)$

Other evaluation benchmarks

- Word analogy

$$\arg\max_{b^* \in V_W \setminus \{a^*, b, a\}} \cos(b^*, a^*) \cdot \cos(b^*, b) / (\cos(b^*, a) + \varepsilon)$$

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

More Results

- Summary of this table:
  - SGNS does better when k is larger
  - SGNS does much better on word analogy tasks
  - Otherwise competitive

WS353 (WordSim) [13]			MEN (WordSim) [4]			Mixed Analogies [20]			Synt. Analogies [22]		
Representation		Corr.	Representation		Corr.	Representation		Acc.	Representation		Acc.
SVD	(k=5)	0.691	SVD	(k=1)	0.735	SPPMI	(k=1)	0.655	SGNS	(k=15)	0.627
SPPMI	(k=15)	0.687	SVD	(k=5)	0.734	SPPMI	(k=5)	0.644	SGNS	(k=5)	0.619
SPPMI	(k=5)	0.670	SPPMI	(k=5)	0.721	SGNS	(k=15)	0.619	SGNS	(k=1)	0.59
SGNS	(k=15)	0.666	SPPMI	(k=15)	0.719	SGNS	(k=5)	0.616	SPPMI	(k=5)	0.466
SVD	(k=15)	0.661	SGNS	(k=15)	0.716	SPPMI	(k=15)	0.571	SVD	(k=1)	0.448
SVD	(k=1)	0.652	SGNS	(k=5)	0.708	SVD	(k=1)	0.567	SPPMI	(k=1)	0.445
SGNS	(k=5)	0.644	SVD	(k=15)	0.694	SGNS	(k=1)	0.540	SPPMI	(k=15)	0.353
SGNS	(k=1)	0.633	SGNS	(k=1)	0.690	SVD	(k=5)	0.472	SVD	(k=5)	0.337
SPPMI	(k=1)	0.605	SPPMI	(k=1)	0.688	SVD	(k=15)	0.341	SVD	(k=15)	0.208