Szóreprezentációk folytonos vektortérben

2014. szeptember 26.

Outline

Bengio et al 2013: A Neural Probabilistic Language Model

Mikolov et al 2014: Efficient Estimation of Word Representation in Vector Space

Pennington et al 2014: GloVe: Global Vectors for Word Representation

A dimenziók átka

- A hagyományos nyelvmodellek (pl. n-gram alapúak)
- általában 1-2 szónyi környezet
- nem veszik figyelembe a hasonlóságot

  “The cat is walking in the bedroom”
  “A dog was running in a room”

Elosztott reprezentáció

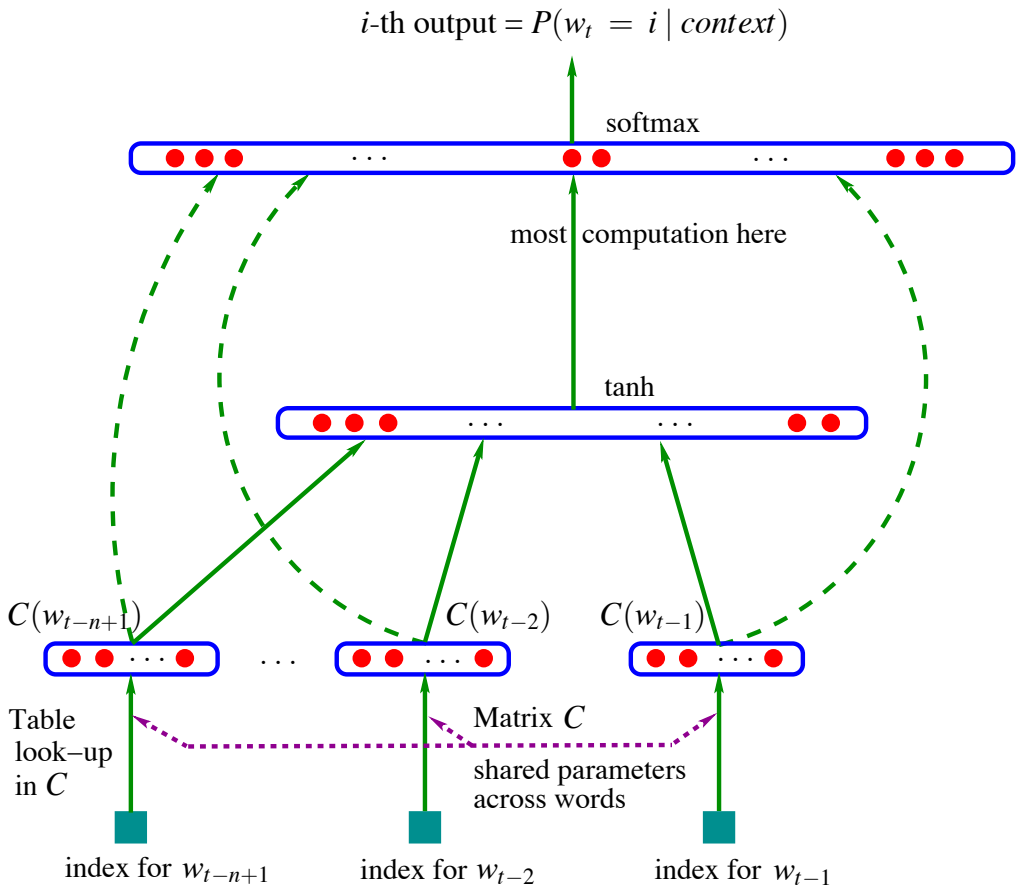


Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i-th word feature vector.

- Maximalizálni

$$L = \frac{1}{T} \sum_t \log f(w_t, w_{t-1}, \ldots, w_{t-n+1}; \theta) + R(\theta)$$

- Input

$$x = (C(w_{t-1}), C(w_{t-2}), \ldots, C(w_{t-n+1}))$$

  - $\dim(C) = |V| \times m$

- Output

$$y = b + Wx + U \tanh(d + Hx)$$

  - $\dim(b) = |V|$, $\dim(W) = |V| \times (n-1)m$, $\dim(U) = |V| \times h$, $\dim(d) = h$, $\dim(H) = h \times (n-1)m$

- Normalizálás (softmax)

$$f(w_t, w_{t-1}, \ldots, w_{t-n+1}) = \frac{e^y_{w_t}}{\sum_i e^y_i}$$

- Paraméterek száma

$$|V|(1 + nm + h) + h(1 + (n-1)m)$$

Eredmények

	n	c	h	m	direct	mix	train.	valid.	test.
MLP1	5		50	60	yes	no	182	284	268
MLP2	5		50	60	yes	yes		275	257
MLP3	5		0	60	yes	no	201	327	310
MLP4	5		0	60	yes	yes		286	272
MLP5	5		50	30	yes	no	209	296	279
MLP6	5		50	30	yes	yes		273	259
MLP7	3		50	30	yes	no	210	309	293
MLP8	3		50	30	yes	yes		284	270
MLP9	5		100	30	no	no	175	280	276
MLP10	5		100	30	no	yes		265	**252**
Del. Int.	3						31	352	336
Kneser-Ney back-off	3							334	323
Kneser-Ney back-off	4							332	321
Kneser-Ney back-off	5							332	321
class-based back-off	3	150						348	334
class-based back-off	3	200						354	340
class-based back-off	3	500						326	**312**
class-based back-off	3	1000						335	319
class-based back-off	3	2000						343	326
class-based back-off	4	500						327	312
class-based back-off	5	500						327	312

Outline

Bengio et al 2013: A Neural Probabilistic Language Model

Mikolov et al 2014: Efficient Estimation of Word Representation in Vector Space

Pennington et al 2014: GloVe: Global Vectors for Word Representation

- Cél: jó minőségű szóvektorok tanulása
- multiple degrees of similarity - szintakitikai és szemantikai

$$v(\text{"}King\text{"}) - v(\text{"}Man\text{"}) + v(\text{"}Woman\text{"}) \sim v(\text{"}Queen\text{"})$$

Időköltségek

- Feedforward Neural Net Language Model (NNLM)

$$Q = n \times m + n \times m \times h + h \times |V|$$

bináris fa hierarchiába rendezve a szótárat $\log_2(|V|)$

- Recurrent NNLM

$$Q = h \times h + h \times |V|$$

Nincs projekciós réteg, $m = h$

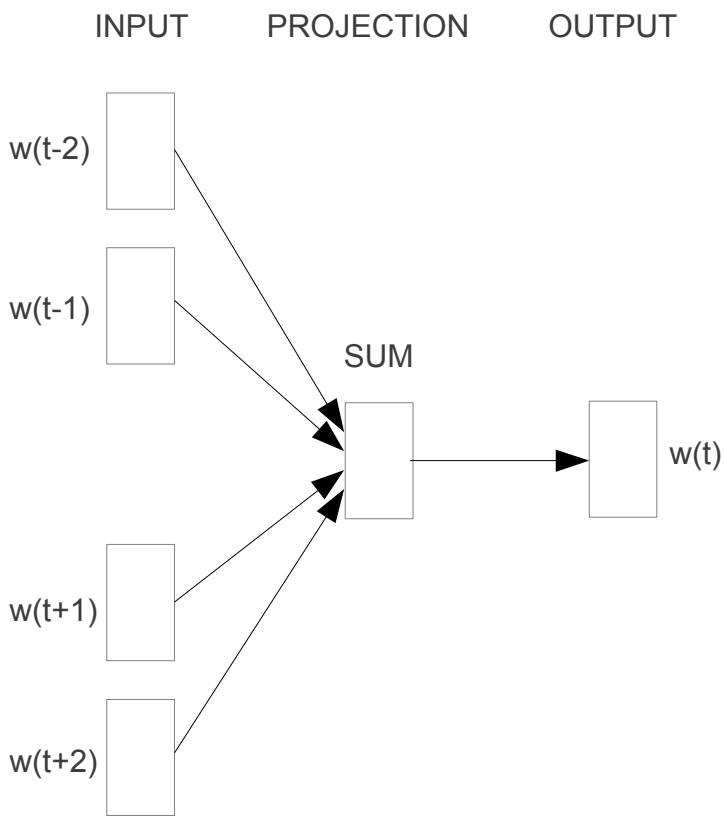
Egyszerűbb modellek szóvektortanulásra

- Continuous Bag-of-Words
  Szóvektorok átlagát veszi a környezetből, nincs rejtett réteg.

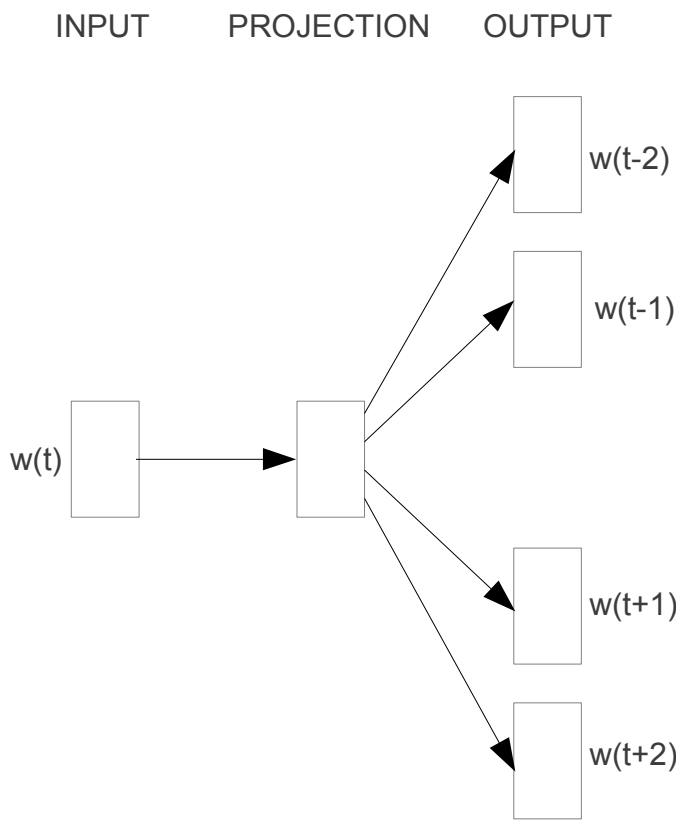
$$Q = n \times m + m \times \log_2(|V|)$$

- Continuous Skip-gram
  1 szó az input, megjósolja a környező szavakat.

$$Q = C \times (m + m \times \log_2(|V|)$$



CBOW



Skip-gram

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

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

Table 2: *Accuracy on subset of the Semantic-Syntactic Word Relationship test set, using word vectors from the CBOW architecture with limited vocabulary. Only questions containing words from the most frequent 30k words are used.*

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

Table 3: *Comparison of architectures using models trained on the same data, with 640-dimensional word vectors. The accuracies are reported on our Semantic-Syntactic Word Relationship test set, and on the syntactic relationship test set of [20]*

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

Table 4: *Comparison of publicly available word vectors on the Semantic-Syntactic Word Relationship test set, and word vectors from our models. Full vocabularies are used.*

Model	Vector Dimensionality	Training words	Accuracy [%]		
			Semantic	Syntactic	Total
Collobert-Weston NNLM	50	660M	9.3	12.3	11.0
Turian NNLM	50	37M	1.4	2.6	2.1
Turian NNLM	200	37M	1.4	2.2	1.8
Mnih NNLM	50	37M	1.8	9.1	5.8
Mnih NNLM	100	37M	3.3	13.2	8.8
Mikolov RNNLM	80	320M	4.9	18.4	12.7
Mikolov RNNLM	640	320M	8.6	36.5	24.6
Huang NNLM	50	990M	13.3	11.6	12.3
Our NNLM	20	6B	12.9	26.4	20.3
Our NNLM	50	6B	27.9	55.8	43.2
Our NNLM	100	6B	34.2	**64.5**	50.8
CBOW	300	783M	15.5	53.1	36.1
Skip-gram	300	783M	**50.0**	55.9	**53.3**

Table 5: *Comparison of models trained for three epochs on the same data and models trained for one epoch. Accuracy is reported on the full Semantic-Syntactic data set.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days]
			Semantic	Syntactic	Total	
3 epoch CBOW	300	783M	15.5	53.1	36.1	1
3 epoch Skip-gram	300	783M	50.0	55.9	53.3	3
1 epoch CBOW	300	783M	13.8	49.9	33.6	0.3
1 epoch CBOW	300	1.6B	16.1	52.6	36.1	0.6
1 epoch CBOW	600	783M	15.4	53.3	36.2	0.7
1 epoch Skip-gram	300	783M	45.6	52.2	49.2	1
1 epoch Skip-gram	300	1.6B	52.2	55.1	53.8	2
1 epoch Skip-gram	600	783M	56.7	54.5	55.5	2.5

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Outline

Bengio et al 2013: A Neural Probabilistic Language Model

Mikolov et al 2014: Efficient Estimation of Word Representation in Vector Space

Pennington et al 2014: GloVe: Global Vectors for Word Representation

- a szóanalógiás feladat: King - Queen = Man - Woman
- vektorok esetén szépen elegánsan vektoriális kivonás
- erre a feladatra fókuszálva dolgozzák ki a glove modellt

megfigyelés 1: i, j, k szavak, $P()$ valószínűségek: $\frac{P(k|i)}{P(k|j)} \gg 1$, ha k jelentése kapcsolatos az i, j szavakkal, de az i szóra jellemző, j-re nem; és közel van 1-hez, ha inadekvát. Tehát a $w_i - w_j$ szóvektorok különbsége a valószínűségük hányadosaival lehet arányos a modellben.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

megfigyelés 2: $w_i - w_j$ különbség és a w_k próbaszó (probe word) vektorokra szimmetrikus is lehetne a modell (mert a szó-kontextusszó viszony szimmetrikus)

megfigyelés 3: a valószínűségek skalárok, a szóvektorok vektorok, de a két vektor skaláris szorzatát kipróbálhatnánk

levezethető, hogy ilyen tulajdonságokkal rendelkezik ez az egyszerű modell:

$$w_i \cdot w_k + b = \log(X_{ik})$$

a modell betanításához (a w_i vektorok kiszámításához) a legkisebb négyzetek megfelelő (a glove-ban még egy súlyfüggvénnyel szorzunk, ami kiegyenlíti a szógyakoriságbeli nagyságrendi különbségeket)
a modell szerencsére a $|V|^2$-nél sokkal jobban méretezhető, mert az együttes előfordulások eloszlása nem egyenletes, hanem hatványtörvény szerinti (ritka mátrix; polinomiálisan kevesebb számítás szükséges)

Eredmények

világverő :)

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

Table 3: Spearman rank correlation on word similarity tasks. All vectors are 300-dimensional. The CBOW* vectors are from the `word2vec` website and differ in that they contain phrase vectors.

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Table 4: F1 score on NER task with 50d vectors. *Discrete* is the baseline without word vectors. We use publicly-available vectors for HPCA, HSMN, and CW. See text for details.

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**