Word vectors

Vlad Shakhuro

MISIS UNIVERSITY

19 September 2025

Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

Word representations

Text is a sequence of symbols

Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

One-hot vectors

text

sequence

symbol

One-hot vectors

text

sequence

symbol

Problems:

- vector size is too large
- vectors don't capture *meaning*, since all vectors have same distances

Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

Denotational semantics

Meaning (Webster dictionary)

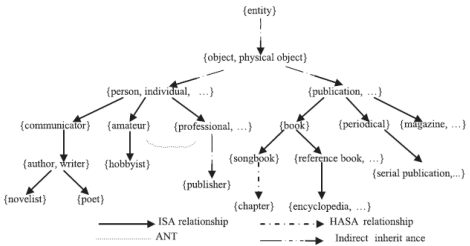
- the idea that is represented by a word, phrase, etc.
- the idea that a person wants to express by using words, signs, etc.
- the idea that is expressed in a work of writing, art, etc.

signifier (symbol) ⇔ signified (idea or thing)

tree ⇔ {🌳, 🎄, 🌴}

WordNet

A thesaurus containing lists of synonims and hypernims (“is a” relations)



Problems with WordNet

Problems with WordNet

- Requires human labour to create and adapt
- Missing nuances, synonims may be correct only in some contexts
- Missing new meaning of words, impossible to keep up-to-date
- Subjective, i.e. depends on human that collected thesaurus
- Can’t be used to accurately compute word similarity

Distributional semantics

Words which frequently appear in similar contexts have similar meaning. (Harris 1954, Firth 1957)

When a word **w** appears in a text, its context is a set of words that appear nearby (within a fixed-sized window). We have to encode information about contexts into word vectors.

…debt problems turning into **banking** crises as happened in 2009…
…Europe needs unified **banking** regulation to replace the hodgepodge…
…India has just given its **banking** system a shot in the arm…

Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

Idea

Learn word vectors by teaching them to predict contexts:

1. Take a huge text corpus
2. Go over the text with sliding window:
   2.1 For the central word, compute probabilities of context words
   2.2 Adjust the vectors to increase these probabilities

… Words are the **primary** building blocks of meaning …

Mikolov et al. Efficient Estimation of Word Representations in Vector Space. ICLR 2013

Idea

Learn word vectors by teaching them to predict contexts:

1. Take a huge text corpus
2. Go over the text with sliding window:
   2.1 For the central word, compute probabilities of context words
   2.2 Adjust the vectors to increase these probabilities

… Words are the primary **building** blocks of meaning …

Mikolov et al. Efficient Estimation of Word Representations in Vector Space. ICLR 2013

Objective function

word2vec learns parameters θ that maximize training data likelihood:

$$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \leqslant j \leqslant m, \\ j \neq 0}} P(w_{t+j} | w_t, \theta)$$

Objective function

word2vec learns parameters θ that maximize training data likelihood:

$$L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \leqslant j \leqslant m, \\ j \neq 0}} P(w_{t+j} \mid w_t, \theta)$$

To make optimization simpler, we use negative log-likelihood as loss function:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \leqslant j \leqslant m, \\ j \neq 0}} \log P(w_{t+j} \mid w_t, \theta)$$

How to compute $P(w_{t+j} \mid w_t, \theta)$?

For each word w we will have two vectors:

- v_w when w is a central word
- u_w when w is a context word
  (used only during training)

All v_w and u_w are trained parameters θ

$$P(o \mid c) = \frac{\exp\left(u_o^T v_c\right)}{\sum_{w \in W} \exp\left(u_w^T v_c\right)}$$

Training step

$$J(\theta) = -\frac{1}{T}\log L(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m \leqslant j \leqslant m, \\ j \neq 0}} \log P(w_{t+j} \mid w_t, \theta)$$

Now we can compute gradient of loss w.r.t. v_{primary} and all u_w and update parameters with gradient descent

Training step

… Words are the **primary** building blocks of meaning …

$$-\log P(\text{blocks} \mid \text{primary}) = -\log \frac{\exp\left(u_{blocks}^T v_{primary}\right)}{\sum_{w \in W} \exp\left(u_w^T v_{primary}\right)} =$$

$$= -\underbrace{u_{blocks}^T v_{primary}}_{increase} + \log \sum_{w \in W} \exp\left(\underbrace{u_w^T v_{primary}}_{decrease}\right)$$

Now we can compute gradient of loss w.r.t. v_{primary} and all u_w and update parameters with gradient descent

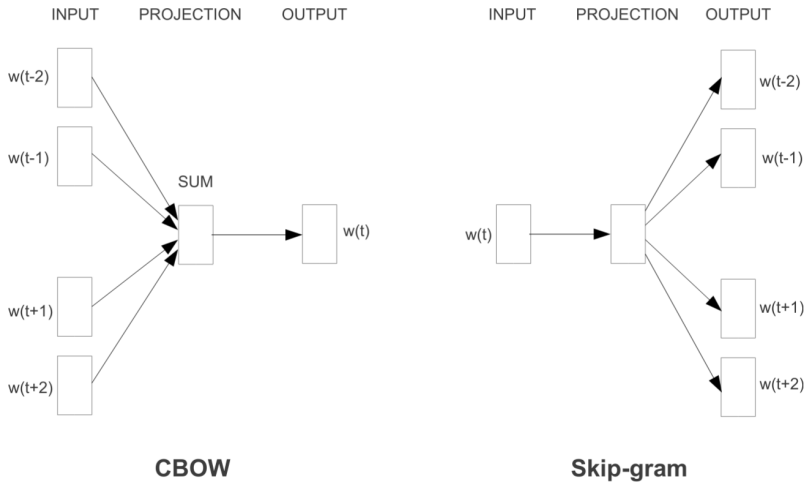
Faster training with negative sampling

$$-\log P(\text{blocks}\,|\,\text{primary}) = -\underbrace{u_{blocks}^{T} v_{primary}}_{increase} + \log \sum_{w \in W} \exp\left(\underbrace{u_{w}^{T} v_{primary}}_{decrease}\right)$$

Updating $|W| + 2$ vectors is too expensive, use small number (i.e. $K = 10$) of random negative samples

Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013

Skip-gram vs CBOW



Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

Directly counting word occurence

- I like deep learning
- I like flying
- I enjoy NLP

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

$$N(w, c)$$

Using SVD to obtain word and context vectors

GloVe: Global Vectors for Word Representation

$$J(\theta) = \sum_{w,c \in W} f(N(w,c)) \cdot (u_c^T \tilde{v}_w + b_c + \tilde{b}_w - \log N(w,c))^2$$

Weighting function f to:

- penalize rare events
- not to over-weight frequent events

Outline

1. Why do we need word representations?
2. One-hot encodings
3. What is meaning?
4. word2vec
5. Counting methods and GloVe
6. Evaluation of word vectors

Intrinsic and extrinsic evaluation

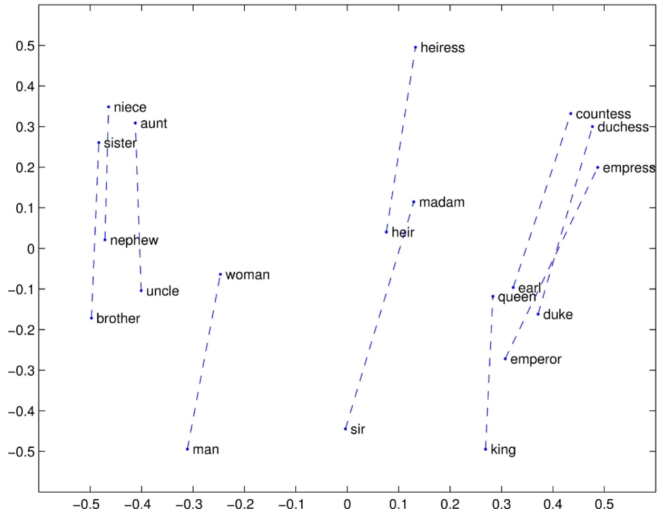
Intrinsic:

- Evaluation on a specific/intermediate subtask
- Fast to compute
- Helps to understand that system
- Not clear if it's helpful unless correlation to real task is established

Extrinsic:

- Evaluation on a real task
- Can take a long time to compute accuracy
- Unclear if the subsystem is the problem or its interaction or other subsystems, have to ablate

Intrinsic evaluation: vector linearity



Intrinsic evaluation: meaning similarity

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10
book	paper	7.46
computer	internet	7.58
plane	car	5.77
professor	doctor	6.62
stock	phone	1.62
stock	CD	1.31
stock	jaguar	0.92

Intrinsic evaluation: meaning similarity

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

Conclusion

We reviewed following topics:

- why do we need word representations
- denotational and distributional semantics
- word2vec approach
- counting methods and GloVe
- approaches to evaluate quality of obtained embeddings