

# Natural Language Processing; Word Representations



[YouTube Playlist](#)

**Maziar Raissi**

**Assistant Professor**

Department of Applied Mathematics

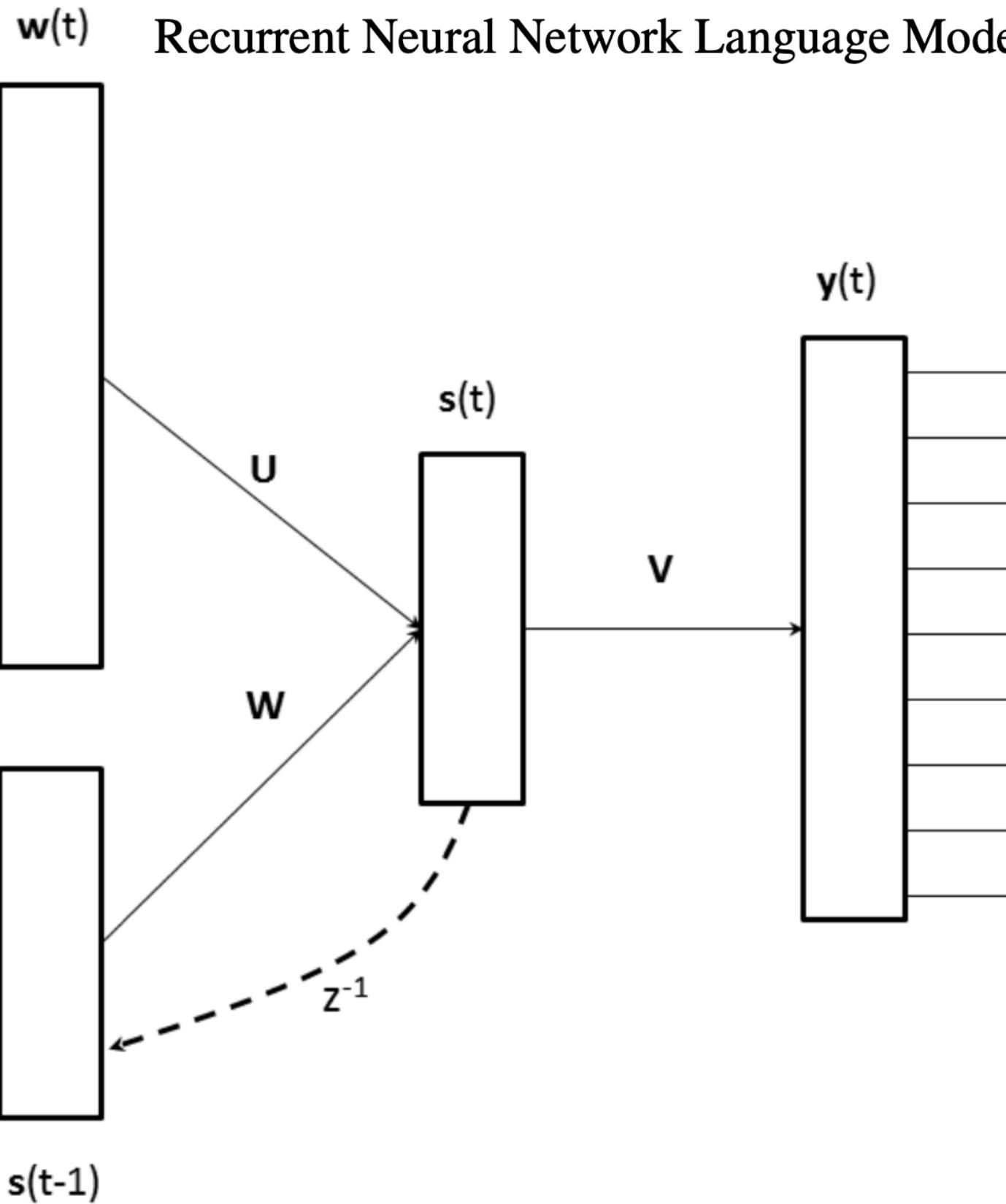
University of Colorado Boulder

[maziar.raissi@colorado.edu](mailto:maziar.raissi@colorado.edu)



Boulder

# Linguistic Regularities in Continuous Space Word Representations

[YouTube Playlist](#)

$w(t) \rightarrow$  input vector (input word at time  $t$  encoded using 1-of- $N$  coding)

$y(t) \rightarrow$  probability distribution over words

$s(t) \rightarrow$  hidden state maintaining a representation of the sentence history

$w(t)$  and  $y(t)$  have dimensionality of the vocabulary

$$s(t) = f(Uw(t) + Ws(t - 1))$$

$$y(t) = g(Vs(t))$$

$$f(z) = \frac{1}{1 + e^{-z}}, g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

Each column in  $U$  represents a word

## Measuring Linguistic Regularities (Syntactic & Semantic)

“ $a$  is to  $b$  as  $c$  is to ?”

“family is to families as car is to cars”

“closing is to shirt as dish is to bowl”

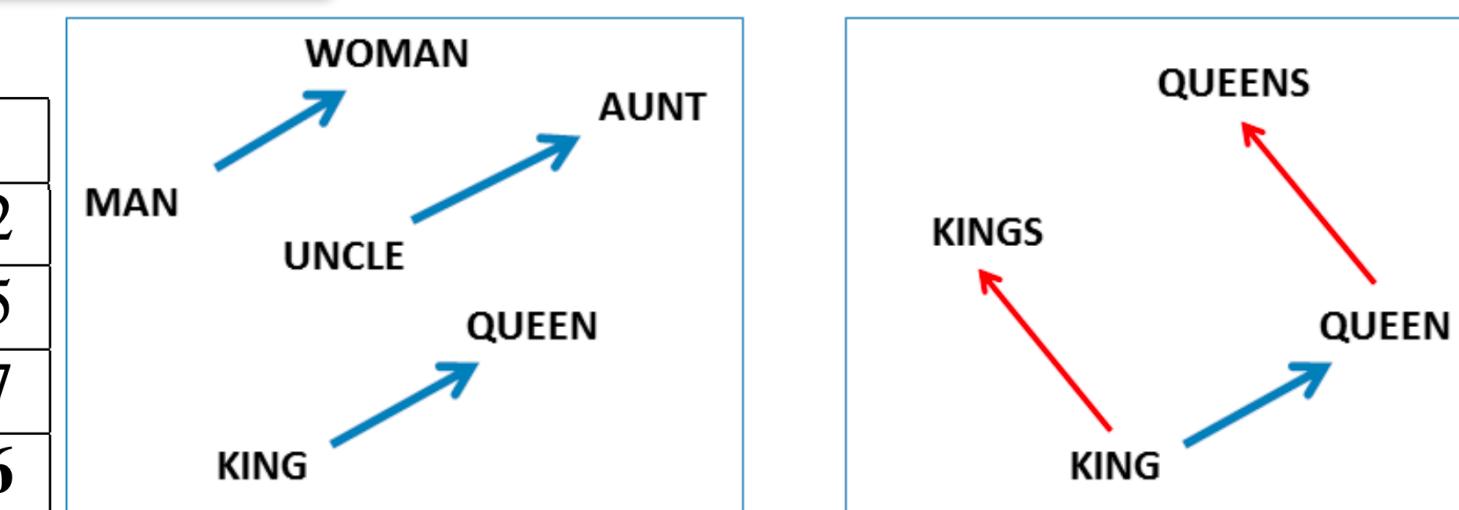
## The Vector Offset Method

$$y = x_b - x_a + x_c$$

$$d = \arg \max_w \frac{x_w y}{\|x_w\| \|y\|} = \cos(x_w, y)$$

## Syntactic Regularities

Method	Adjectives	Nouns	Verbs	All
RNN-80	9.3	5.2	30.4	16.2
RNN-320	18.2	19.0	45.0	28.5
RNN-640	21.0	25.2	54.8	34.7
<b>RNN-1600</b>	<b>23.9</b>	<b>29.2</b>	<b>62.2</b>	<b>39.6</b>



## semantic regularities

Method	Spearman's $\rho$	MaxDiff Acc.
LSA-640	0.149	0.364
RNN-80	0.211	0.389
RNN-320	0.259	0.408
RNN-640	0.270	0.416
<b>RNN-1600</b>	<b>0.275</b>	<b>0.418</b>

SemEval 2012 task of measuring relation similarity

Category	Relation	Patterns Tested	# Questions	Example
Adjectives	Base/Comparative	JJ/JJR, JJR/JJ	1000	good:better rough:---
Adjectives	Base/Superlative	JJ/JJS, JJS/JJ	1000	good:best rough:---
Adjectives	Comparative/Superlative	JJS/JJR, JJR/JJS	1000	better:best rougher:---
Nouns	Singular/Plural	NN/NNS, NNS/NN	1000	year:years law:---
Nouns	Non-posessive/Possessive	NN/NN_POS, NN_POS/NN	1000	city:city's bank:---
Verbs	Base/Past	VB/VBD, VBD/VB	1000	see:saw return:---
Verbs	Base/3rd Person Singular Present	VB/VBZ, VBZ/VB	1000	see:sees return:---
Verbs	Past/3rd Person Singular Present	VBD/VBZ, VBZ/VBD	1000	saw:sees returned:---



Boulder

# Distributed Representations of Words and Phrases and their Compositionality



[YouTube Playlist](#)

## Skip-gram Model

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

$$p(w_O | w_I) = \frac{\exp(v'_{w_O}^\top v_{w_I})}{\sum_{w=1}^W \exp(v'_{w_i}^\top v_{w_I})}$$

$w_1, w_2, \dots, w_T \rightarrow$  training words

$c \rightarrow$  size of the training context

$w_I = w_t, w_O = w_{t+j}$

$W \rightarrow$  vocabulary size

$v_w \rightarrow$  input vector representation of  $w$

$v'_w \rightarrow$  output vector representation of  $w$

$W \approx 10^5 - 10^7$  terms

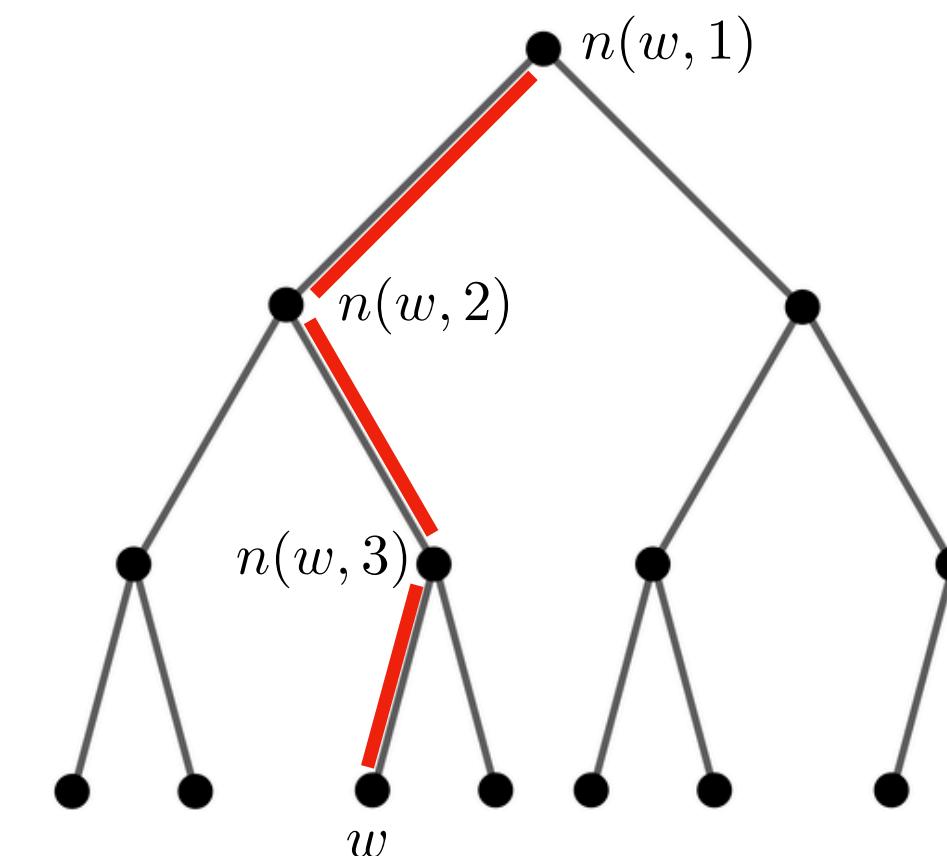
**Learning Phrases**      “New York Times”

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}$$

$\delta \rightarrow$  discounting coefficient

## Hierarchical Softmax

$\log_2(W)$  evaluations, rather than  $W$



$n(w, j) \rightarrow$   $j$ th node on the path from the root to  $w$

$L(w) \rightarrow$  length of this path

$n(w, 1) \rightarrow$  root

$n(w, L(w)) = w$

$\text{ch}(n) \rightarrow$  an arbitrary fixed child (e.g., left) of a node  $n$

$[\![\text{TRUE}]\!] = 1, [\![\text{FALSE}]\!] = -1$

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma([\![n(w, j+1) = \text{ch}(n(w, j))]\!] \cdot v'_{n(w,j)}^\top v_{w_I})$$

$$p(w|w_I) = \sigma(v'_{n(w,1)}^\top v_{w_I}) \sigma(-v'_{n(w,2)}^\top v_{w_I}) \sigma(v'_{n(w,3)}^\top v_{w_I})$$

binary Huffman tree: assign short codes to the frequent words

## Negative Sampling

Noise Contrastive Estimation (NCE): a good model should be able to differentiate data from noise by means of logistic regression

$$\log p(w_O | w_I) \approx \log \sigma(v'_{w_O}^\top v_{w_I}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} \underbrace{[\log \sigma(-v'_{w_i}^\top v_{w_I})]}_{\text{noise distribution}}$$

$P_n(w) \propto U(w)^{3/4}$  unigram distribution

## Countering the imbalance between rare and frequent words

each word  $w_i$  in the training set is discarded with probability

$t = 10^{-5} \rightarrow$  threshold,  $f \rightarrow$  frequency

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

**Evaluation** syntactic & semantic analogies

$$\text{“Paris”} = \arg \min_w d(\text{vec(“Berlin”)} - \text{vec(“Germany”)} + \text{vec(“France”)}, \text{vec}(w))$$



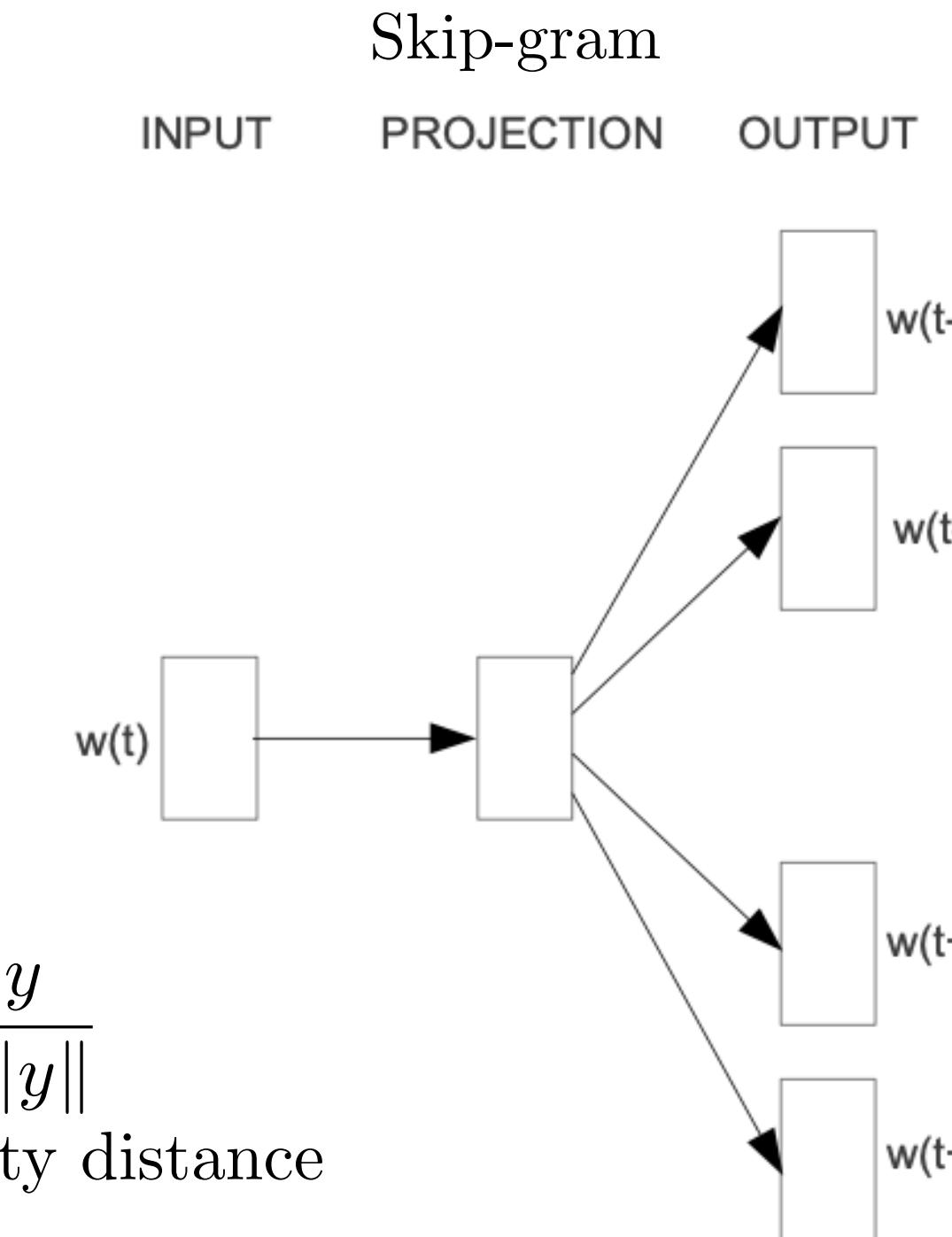
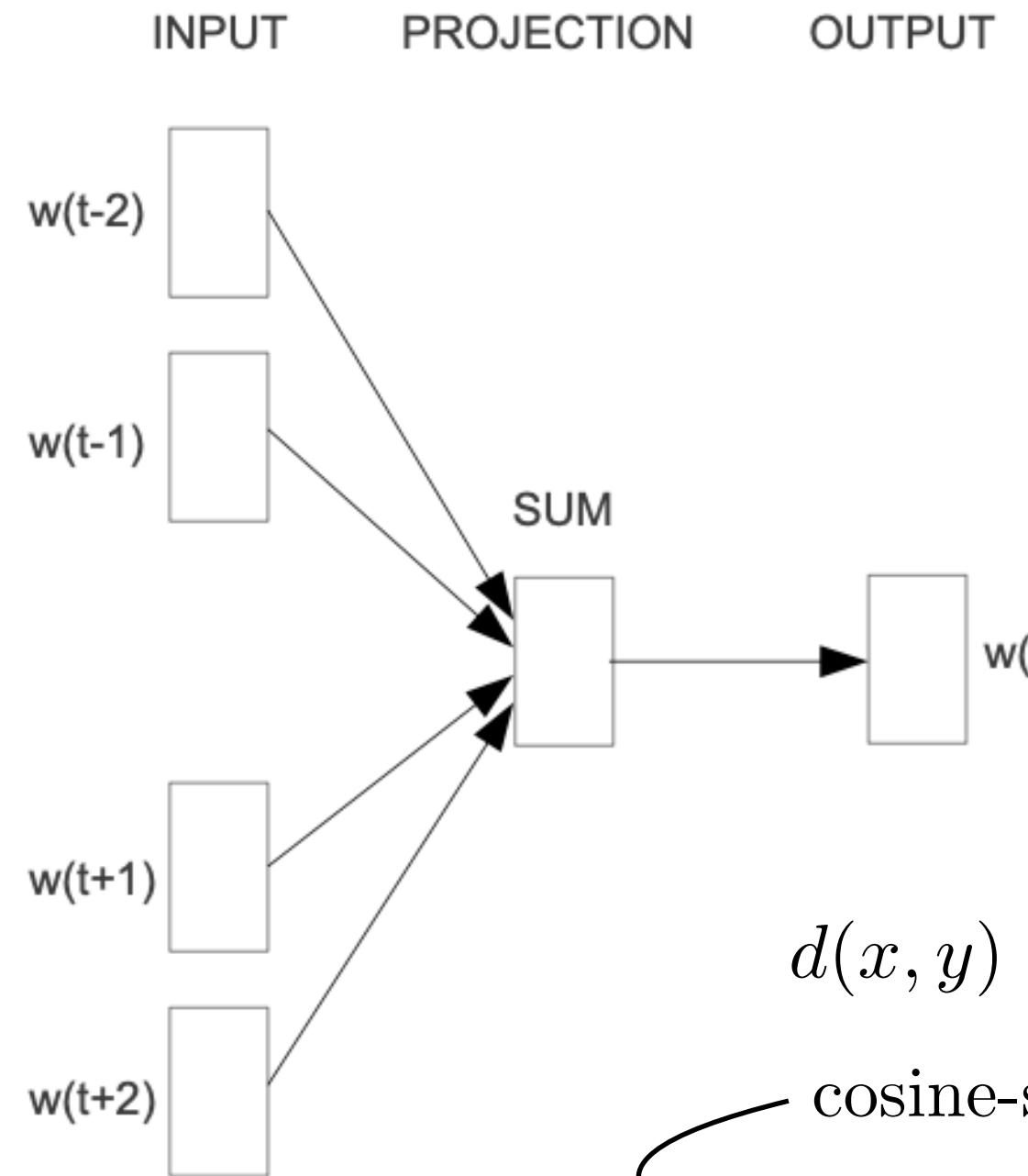
Boulder

# Efficient Estimation of Word Representations in Vector Space



YouTube Video

CBOW: Continuous Bag of Words



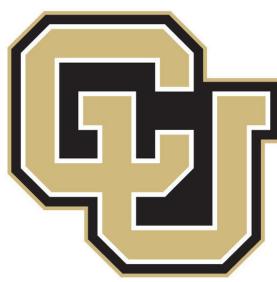
Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	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

“smallest” =  $\arg \min_w d(\text{vec}(\text{"biggest"}) - \text{vec}(\text{"big"}) + \text{vec}(\text{"small"}), \text{vec}(w))$

“Paris” =  $\arg \min_w d(\text{vec}(\text{"Berlin"}) - \text{vec}(\text{"Germany"}) + \text{vec}(\text{"France"}), \text{vec}(w))$

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

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



Boulder



[YouTube Playlist](#)

# GloVe: Global Vectors for Word Representation

1) Global matrix factorization methods

2) Local context window methods

$X \rightarrow$  matrix of word-word co-occurrence counts

$X_{ij} \rightarrow$  # times word  $j$  occurs in the context of word  $i$

$\log(X_{ij}) = w_i^T \tilde{w}_j + b_i + \tilde{b}_j \rightarrow$  assumption

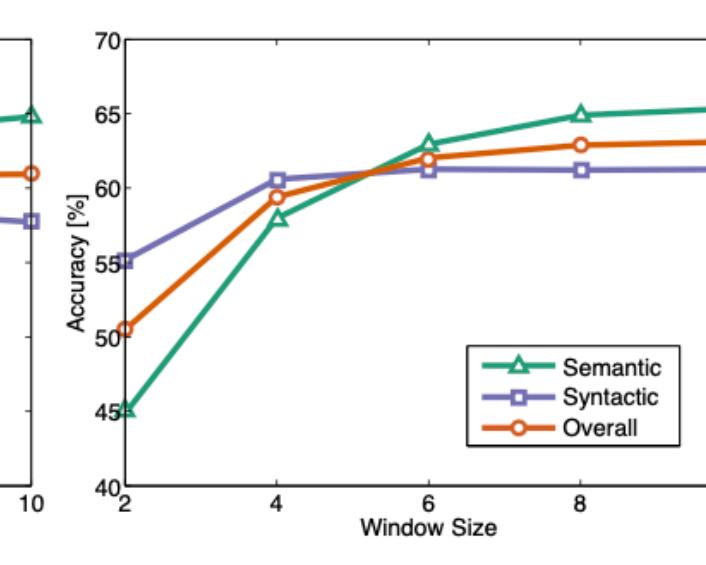
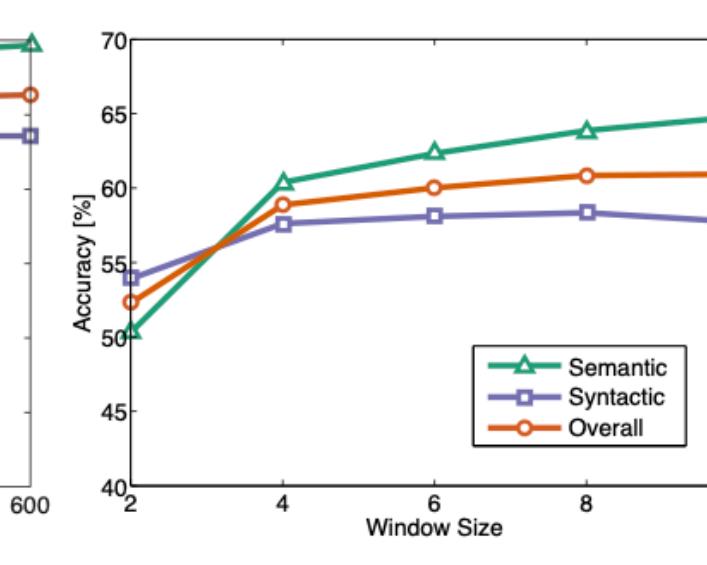
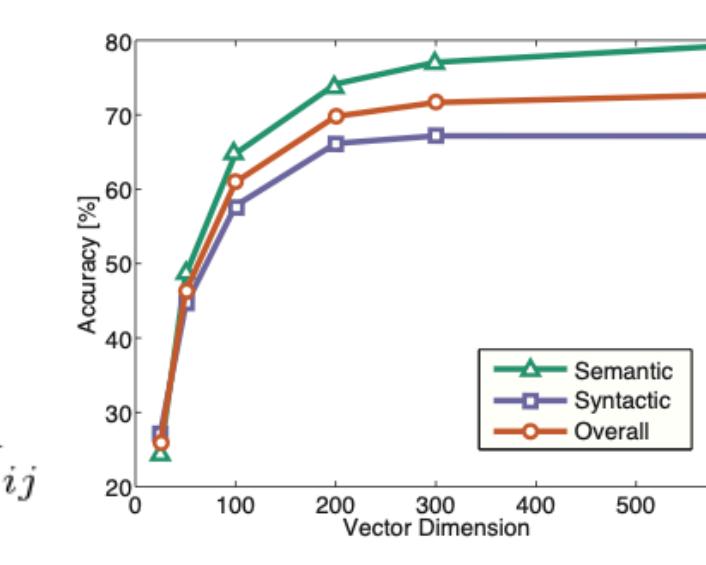
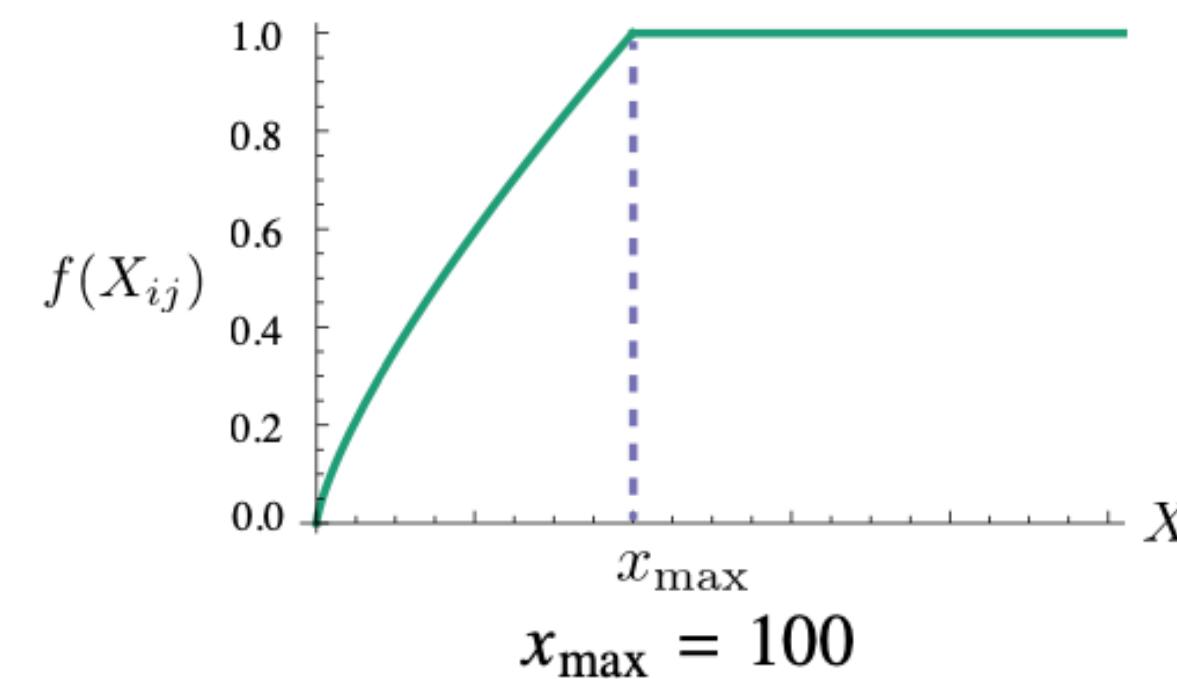
$w_i \in \mathbb{R}^d \rightarrow$  word vector

$\tilde{w}_j \in \mathbb{R}^d \rightarrow$  context word vector

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

weighted least squares

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases} \quad \alpha = 3/4$$



Relationship to skip-gram

$$Q_{ij} = \frac{\exp(w_i^T \tilde{w}_j)}{\sum_{k=1}^V \exp(w_i^T \tilde{w}_k)} \rightarrow \text{probability that word } j \text{ appears in the context of word } i$$

$$J = - \sum_{\substack{i \in \text{corpus} \\ j \in \text{context}(i)}} \log Q_{ij}$$

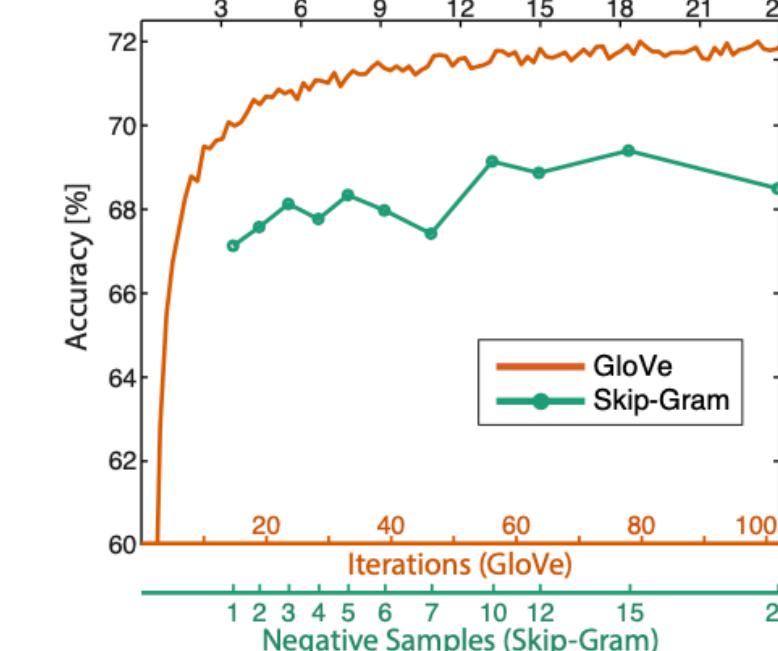
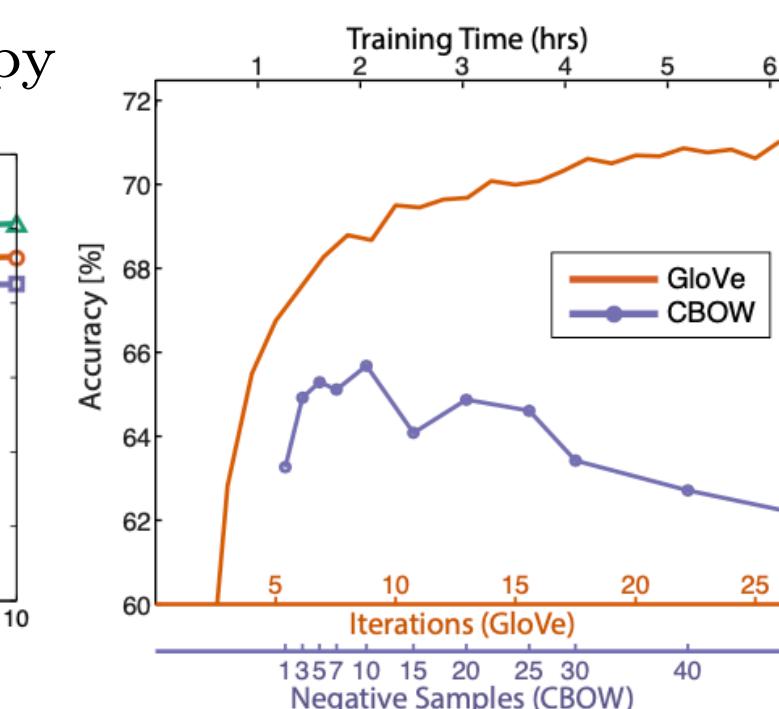
$$J = - \sum_{i=1}^V \sum_{j=1}^V X_{ij} \log Q_{ij} \rightarrow \text{group together those terms that have the same values for } i \text{ and } j$$

$$X_i = \sum_k X_{ik} \rightarrow \# \text{ times any word occurs in the context of word } i$$

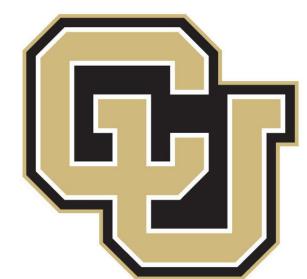
$$P_{ij} = \frac{X_{ij}}{X_i} \rightarrow \text{prob. that word } j \text{ occurs in the context of word } i$$

$$J = - \sum_{i=1}^V X_i \sum_{j=1}^V P_{ij} \log Q_{ij} = \sum_{i=1}^V X_i \underbrace{H(P_i, Q_i)}_{\text{cross-entropy}}$$

Model	Dim.	Size	Sem.	Syn.	Tot.
CBOW <sup>†</sup>	300	6B	63.6	<u>67.4</u>	65.7
SG <sup>†</sup>	300	6B	73.0	66.0	69.1
GloVe	300	6B	<u>77.4</u>	67.0	<u>71.7</u>







Boulder



# Questions?

[YouTube Playlist](#)

---