Word Embedding using Word2Vec

Ramaseshan Ramachandran

Ramaseshan

Word2Vec[1]

[1] Mikolov, Tomas, Kai Chen, Gregory S. Corrado and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space." International Conference on Learning Representations (2013)

GOAL

- Process each word in a Vocabulary of words to obtain a respective numeric representation of each word in the Vocabulary
- Reflect semantic similarities, Syntactic similarities, or both, between words they represent
- Uncover latent semantic relationships
- Map each of the plurality of words to a respective vector and output a single merged vector that is a combination of the respective vectors

CONTEXT WORDS AND CENTRAL WORD

- **Continuous Bag of Words (CBOW)** Models – A central word is surrounded by context words. Given the context words identify the central word
  - Wish you many more happy returns of the day
- **Skip Gram Model**- Given the central word, identify the surrounding words
  - Wish you many more happy returns of the day

Let $\mathbb{V}$ be the universal vocabulary and V be a finite set of words found from a corpus. Let us consider a finite set of words $V \in \mathbb{V}$ as vocabulary.
The aim is to find a d-dimensional **dense** vector representation $\vec{w_i}$ for $\forall w_i \in V$.
Then, $\forall w_i \in V$, the word embedding is represented as a matrix $\mathbf{W} \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the size of the vocabulary and every $w_i \in \mathbb{R}^d$.

CONTINUOUS BAG OF WORDS (CBOW)

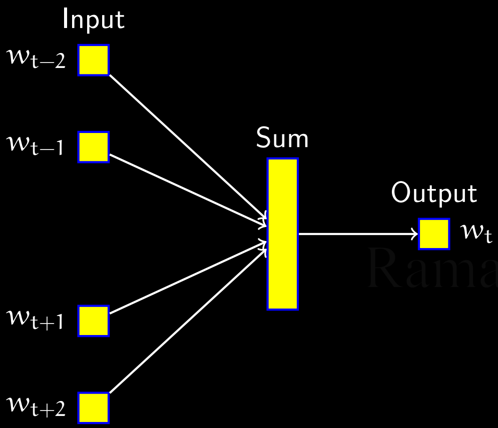


Figure: The CBOW architecture predicts the current word based on the context words of length n. Here the window size is 5

CBOW uses the sequence of words "Wish", "you", "a", "happy", "year" as a context and predicts or generates the central word "new"

- CBOW is used for learning the central word
- Maximize probability of word based on the word co-occurrences within a distance of n

SKIP GRAM MODEL

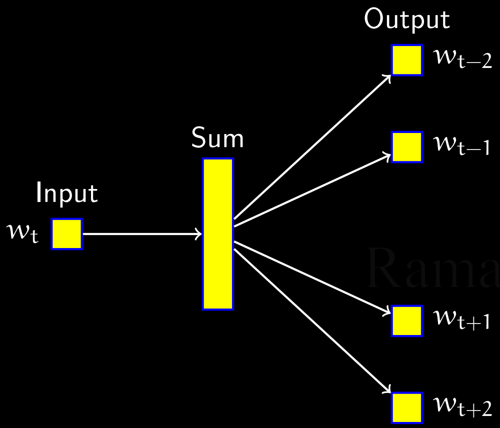


Figure: The SG architecture predicts the one context word at a time based on the center word. Here the window size is 5

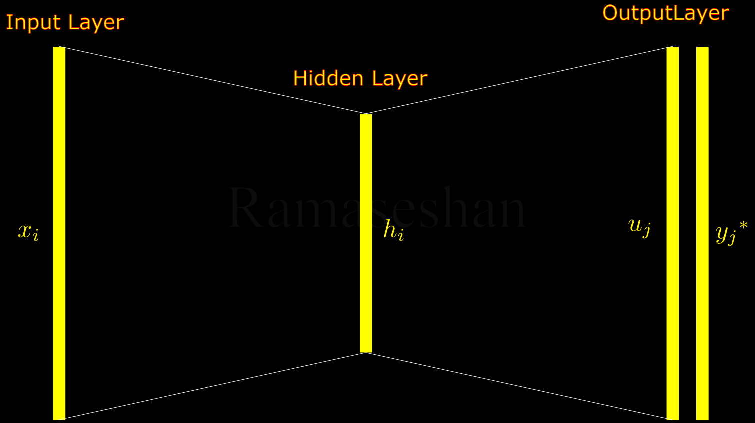
SG uses the central word "new" and predicts the context words "Wish", "you", "a", "happy", "year"

- SG is used to learn the context words given the central word
- Maximize probability of word based on the word co-occurrences within a distance of $[-n, +n]$ from the center word

SOURCE PREPARATION FOR TRAINING

Source Text	Training Samples
$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{many}}$ more happy returns of the day $\longrightarrow$	$(wish, you)$ $(wish, many)$
$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{more}}\boxed{\text{happy}}$ returns of the day $\longrightarrow$	$(you, Wish)$ $(you, more), (you, happy)$
$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}$ returns of the day $\longrightarrow$	$(many, Wish), (many, you)$ $(many, more), (many, happy)$
Wish $\boxed{\text{you}}\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}$ of the day $\longrightarrow$	$(more, many), (more, you)$ $(more, happy), (more, returns$
Wish you $\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}$ the day $\longrightarrow$	$(happy, many), (happy, more$ $(happy, returns), (happy, of)$
Wish you many $\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}\boxed{\text{the}}$ day $\longrightarrow$	$(returns, more), (returns, happ$ $(returns, of), (returns, the)$
Wish you many more $\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}\boxed{\text{the}}\boxed{\text{day}}$ $\longrightarrow$	$(of, happy), (of, returns)$ $(of, the), (of, day)$

ONE-WORD LEARNING



INPUT LAYER

$$t^{aback} = \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix} \dots t^{zoom} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 1 \\ 0 \end{pmatrix} t^{zucchini} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ \dots \\ 0 \\ 1 \end{pmatrix}$$

$$x_k = 1 \text{ and } x'_k = 0, \forall k' \neq k$$

HIDDEN LAYER

This neural network is fully connected. Input to the network is a one-hot vector. W is the N-dimensional vector representation of the word, v_w^{T}, presented as input

$$\mathrm{h} = \mathrm{W}^{\mathrm{T}}\mathrm{X} \tag{1}$$

Now $v_{w\mathrm{I}}$ of the matrix (W) is the vector representation of the input one-hot vector w_{I}.
From (1), h is a linear combination of input and weights.
In the same way. we get a score for u_{j}

$$u_{\mathrm{j}} = \mathrm{v}'^{\mathrm{T}}_{\mathrm{w_j}}\mathrm{h} = \mathrm{v}'^{\mathrm{T}}_{\mathrm{w_j}}\mathrm{v}_{\mathrm{w_I}} \tag{2}$$

where $v_{w_{\mathrm{I}}}$ is the vector representation of the input word w_{I} and $v'_{w_{\mathrm{j}}}$ is the j^{th} column of (W')

OUTPUT LAYER

At the output layer, we apply the softmax to get the posterior distribution of the word(s). It is obtained by,

$$p(w_O|w_I) = y_{j^*} \tag{3}$$

where y_j is the output of the j^{th} unit in the output layer

$$y_* = \frac{\exp(u_j)}{\sum_{j'=1}^{V} \exp(u_{j'})} \tag{4}$$

$$= \frac{\exp(\mathrm{v}'^{T}_{\mathrm{w}_j} v_{wI})}{\sum_{j'=1}^{V} \exp(\mathrm{v}'^{T}_{\mathrm{w}_{j'}} v_{wI})} \tag{5}$$

where v_w, v'_w are the input vector (word vector) and output vector (feature vector) representations, of w_j and w'_j, respectively

UPDATE WEIGHTS - HIDDEN-OUTPUT LAYERS

The learning/training objective is to maximize (5) or minimize the error between the target and the computed value of the target which is $y_j^* - t_j$ and t_j is same as the input vector, in this case. We use cross-entropy as it provides us with a good measure of "error distance"

$$\max p(w_o|w_I) = \max(\log(y_j*)) - -\text{Maximize} \quad (6)$$
$$-E = u_j - \log(y_{j*}) - t_j - -\text{minimize} \quad (7)$$
$$= u_j * -\log \sum_{j'=1}^{V} \exp(u_j') - t_j \quad (8)$$

where E is the loss function.

- $\log p(x)$ is well scaled
- Selection of step size is easier
- With $p(x)$ multiplication may yield to near zero causing *underflow*
- For better optimization, $\log p(x)$ is considered (multiplication $\rightarrow$ addition)

UPDATE WEIGHTS (HO) - MINIMIZATION OF E

It is a special case of cross-entropy measurement between two probabilistic distributions y and t. To minimize E,take the partial derivative of E with respect to j^{th} unit of u_j

$$\frac{\partial E}{\partial u_j} = y_j - t_j = e_j \tag{9}$$

where e_j is the prediction error. Taking partial derivative with respect to the hidden-output weights, we get,

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w'_{ij}} = e_j \cdot h_i \tag{10}$$

Using the above equation (10),

$$w'^{\,new}_{ij} = w'^{\,old}_{ij} - \eta e_j \cdot h_i \text{ or} \tag{11}$$

$$\mathrm{v}^{(\mathrm{new})}_{\mathrm{w}_j} = \mathrm{v}'^{(\mathrm{old})}_{\mathrm{w}_j} - \eta e_j \cdot \mathbf{h} \quad \text{for } j = 1, 2, 3, \ldots, V \tag{12}$$

UPDATE INPUT TO HIDDEN WEIGHTS

Taking the derivative with respect to h_i, we get

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} e_j \cdot w'_{ij} = \mathrm{EH}_i \tag{13}$$

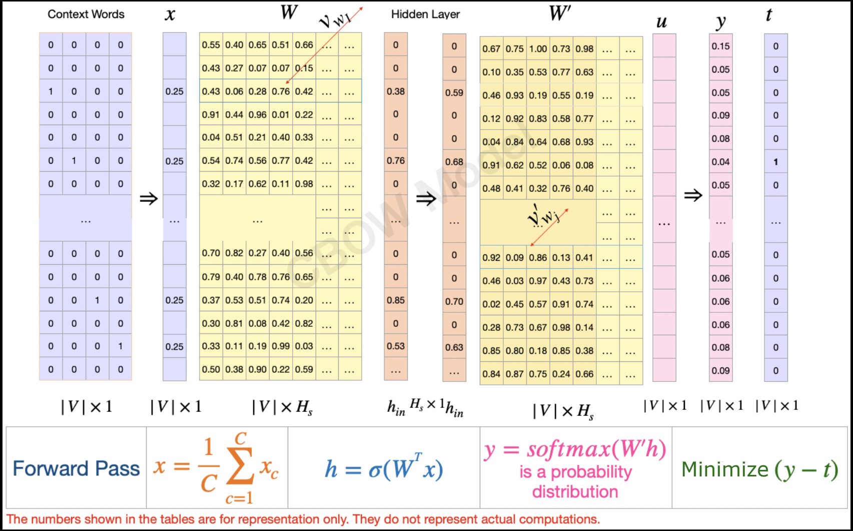
Taking the derivative with respect to w_{ki}, we get

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = \mathrm{EH}_i.x_k \tag{14}$$

Now the weights are updated using

$$v_{w_i}{}^{(new)} = v_{wi}{}^{(old)} - \eta \mathrm{EH}^T \tag{15}$$

CBOW IN MATRIX FORM



Context Words

0	0	0	0
0	0	0	0
1	0	0	0
0	0	0	0
0	0	0	0
0	1	0	0
0	0	0	0
…			
0	0	0	0
0	0	0	0
0	0	1	0
0	0	0	0
0	0	0	1
0	0	0	0

x

0.25
0.25
…
0.25
0.25

W

0.55	0.40	0.65	0.51	0.66	…	…
0.43	0.27	0.07	0.07	0.15	…	…
0.43	0.06	0.28	0.76	0.42	…	…
0.91	0.44	0.96	0.01	0.22	…	…
0.04	0.51	0.21	0.40	0.33	…	…
0.54	0.74	0.56	0.77	0.42	…	…
0.32	0.17	0.62	0.11	0.98	…	…
…					…	…
0.70	0.82	0.27	0.40	0.56	…	…
0.79	0.40	0.78	0.76	0.65	…	…
0.37	0.53	0.51	0.74	0.20	…	…
0.30	0.81	0.08	0.42	0.82	…	…
0.33	0.11	0.19	0.99	0.03	…	…
0.50	0.38	0.90	0.22	0.59	…	…

Hidden Layer

0	0
0	0
0.38	0.59
0	0
0	0
0.76	0.68
0	0
…	…
0	0
0	0
0.85	0.70
0	0
0.53	0.63
…	…

W'

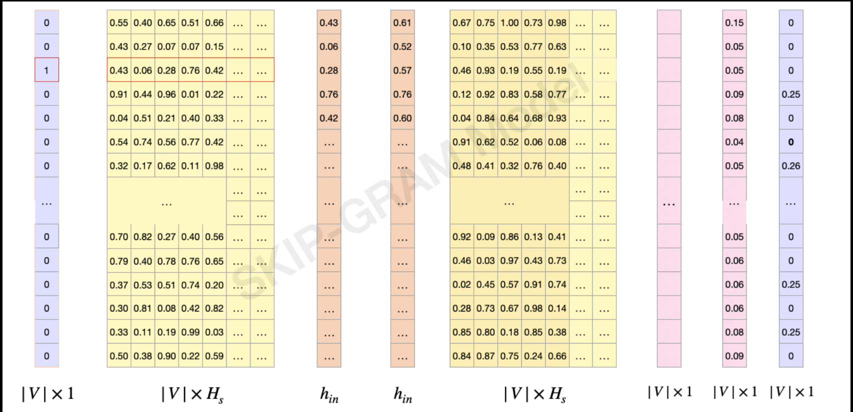
0.67	0.75	1.00	0.73	0.98	…	…
0.10	0.35	0.53	0.77	0.63	…	…
0.46	0.93	0.19	0.55	0.19	…	…
0.12	0.92	0.83	0.58	0.77	…	…
0.04	0.84	0.64	0.68	0.93	…	…
0.91	0.62	0.52	0.06	0.08	…	…
0.48	0.41	0.32	0.76	0.40	…	…
…					…	…
0.92	0.09	0.86	0.13	0.41	…	…
0.46	0.03	0.97	0.43	0.73	…	…
0.02	0.45	0.57	0.91	0.74	…	…
0.28	0.73	0.67	0.98	0.14	…	…
0.85	0.80	0.18	0.85	0.38	…	…
0.84	0.87	0.75	0.24	0.66	…	…

u	y	t
	0.15	0
	0.05	0
	0.05	0
	0.09	0
	0.08	0
	0.04	1
	0.05	0
…	…	…
	0.05	0
	0.06	0
	0.06	0
	0.06	0
	0.08	0
	0.09	0

Forward Pass	$x = \frac{1}{C}\sum_{c=1}^{C} x_c$	$h = \sigma(W^T x)$	$y = softmax(W'h)$ is a probability distribution	Minimize $(y - t)$

The numbers shown in the tables are for representation only. They do not represent actual computations.

SG IN MATRIX FORM



$\lvert V\rvert \times 1$	$\lvert V\rvert \times H_s$							h_{in}	h_{in}	$\lvert V\rvert \times H_s$							$\lvert V\rvert \times 1$	$\lvert V\rvert \times 1$	$\lvert V\rvert \times 1$
0	0.55	0.40	0.65	0.51	0.66	…	…	0.43	0.61	0.67	0.75	1.00	0.73	0.98	…	…		0.15	0
0	0.43	0.27	0.07	0.07	0.15	…	…	0.06	0.52	0.10	0.35	0.53	0.77	0.63	…	…		0.05	0
1	0.43	0.06	0.28	0.76	0.42	…	…	0.28	0.57	0.46	0.93	0.19	0.55	0.19	…	…		0.05	0
0	0.91	0.44	0.96	0.01	0.22	…	…	0.76	0.76	0.12	0.92	0.83	0.58	0.77	…	…		0.09	0.25
0	0.04	0.51	0.21	0.40	0.33	…	…	0.42	0.60	0.04	0.84	0.64	0.68	0.93	…	…		0.08	0
0	0.54	0.74	0.56	0.77	0.42	…	…	…	…	0.91	0.62	0.52	0.06	0.08	…	…		0.04	0
0	0.32	0.17	0.62	0.11	0.98	…	…	…	…	0.48	0.41	0.32	0.76	0.40	…	…		0.05	0.26
…			…			…	…	…	…			…			…	…	…	…	…
0	0.70	0.82	0.27	0.40	0.56	…	…	…	…	0.92	0.09	0.86	0.13	0.41	…	…		0.05	0
0	0.79	0.40	0.78	0.76	0.65	…	…	…	…	0.46	0.03	0.97	0.43	0.73	…	…		0.06	0
0	0.37	0.53	0.51	0.74	0.20	…	…	…	…	0.02	0.45	0.57	0.91	0.74	…	…		0.06	0.25
0	0.30	0.81	0.08	0.42	0.82	…	…	…	…	0.28	0.73	0.67	0.98	0.14	…	…		0.06	0
0	0.33	0.11	0.19	0.99	0.03	…	…	…	…	0.85	0.80	0.18	0.85	0.38	…	…		0.08	0.25
0	0.50	0.38	0.90	0.22	0.59	…	…	…	…	0.84	0.87	0.75	0.24	0.66	…	…		0.09	0

$X, W^e, W^c, h,$ and $\hat{Y}$ represent the input, embedding matrix, context matrix and output vector, respectively. h is the copy of the vector corresponding to the index of the input word in the vocabulary for skip-gram model. For CBOW model, it represents a vector whose elements are obtained using the the average linear sum of the vectors corresponding to the context words. x is represented by

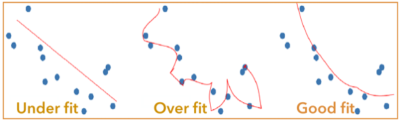
$$x = \frac{1}{C}(x_1 + x_2 + \ldots + x_n), \quad (1)$$

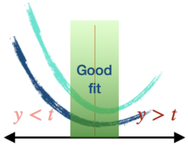
where n represents the number of context words. For skip-gram, $n = 1$ and for CBOW model $n > 1$.

UNDER FIT, OVER FIT AND GOOD FIT

During the forward pass, h_{in} is obtained using the linear combination the initial weight matrix and the one-hot vector of the corresponding context word (skip-gram model) or corresponding context words (CBOW model). Note that the index of the word in the vocabulary is represented by the position of 1 in the OHV. During the computations of Y, the propagated properties of the word vector, h, further passes on the properties of the word vector to the predicted output, Y, through the W^c.

$$\underset{|V|\times 1}{\overset{y}{\begin{bmatrix} 0.15 \\ 0.05 \\ 0.05 \\ 0.09 \\ 0.08 \\ 0.04 \\ 0.05 \\ \cdots \\ 0.05 \\ 0.06 \\ 0.06 \\ 0.06 \\ 0.08 \\ 0.09 \end{bmatrix}}} - \underset{|V|\times 1}{\overset{t}{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}} = \underset{|V|\times 1}{\overset{e}{\begin{bmatrix} 0.15 \\ 0.05 \\ 0.05 \\ 0.09 \\ 0.08 \\ -0.96 \\ 0.05 \\ \\ 0.05 \\ 0.06 \\ 0.06 \\ 0.06 \\ 0.08 \\ 0.09 \end{bmatrix}}}$$





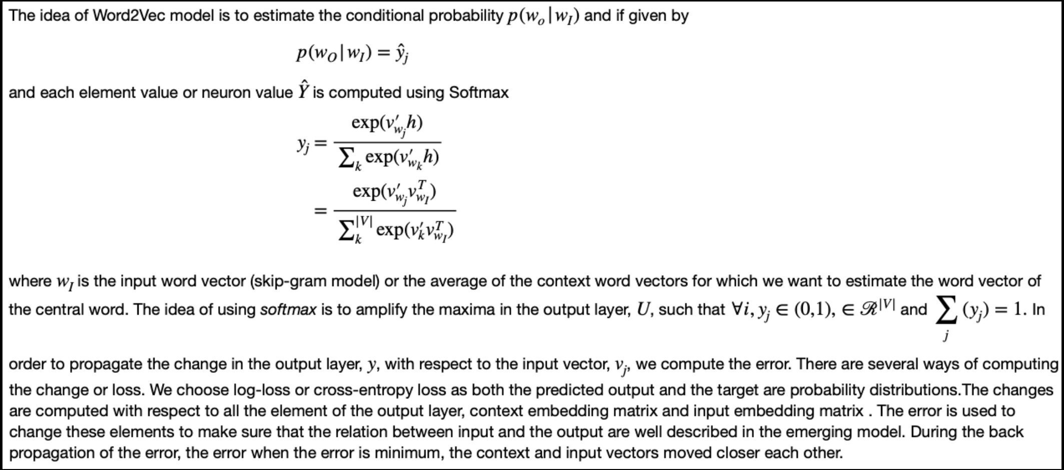
If $y - t$ is lower, then the predicted word is not close to the target word. In other words, the context vectors which are supposed to be similar to the target vector needs to be adjusted into to minimise the differences.

If $y - t$ is higher, then the predicted word moved further away from the target, again the context vectors are further away from the expected result.

If $y - t$ is very small, then we are close to achieving the desired result. The learning becomes slow, the weight updates do not yield any significant change In such case, learning is complete and the context vectors are now closer to the target vector.

Since the target and the estimated values are probability distributions, it is a good idea to use the cross-entropy loss for the propagation of error. In other words, the movement of the word vector towards the expected one is proportional to the error.

OUTPUT GENERATION



The idea of Word2Vec model is to estimate the conditional probability $p(w_o | w_I)$ and if given by

$$p(w_O | w_I) = \hat{y}_j$$

and each element value or neuron value $\hat{Y}$ is computed using Softmax

$$y_j = \frac{\exp(v'_{w_j} h)}{\sum_k \exp(v'_{w_k} h)}$$
$$= \frac{\exp(v'_{w_j} v^T_{w_I})}{\sum_k^{|V|} \exp(v'_k v^T_{w_I})}$$

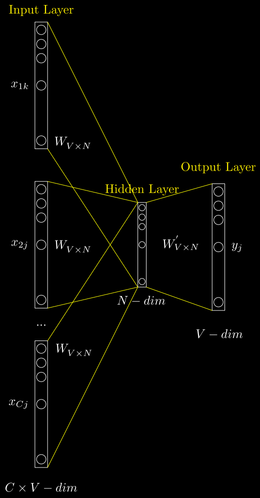
where w_I is the input word vector (skip-gram model) or the average of the context word vectors for which we want to estimate the word vector of the central word. The idea of using *softmax* is to amplify the maxima in the output layer, U, such that $\forall i, y_j \in (0,1), \in \mathscr{R}^{|V|}$ and $\sum_j (y_j) = 1$. In order to propagate the change in the output layer, y, with respect to the input vector, v_j, we compute the error. There are several ways of computing the change or loss. We choose log-loss or cross-entropy loss as both the predicted output and the target are probability distributions.The changes are computed with respect to all the element of the output layer, context embedding matrix and input embedding matrix . The error is used to change these elements to make sure that the relation between input and the output are well described in the emerging model. During the back propagation of the error, the error when the error is minimum, the context and input vectors moved closer each other.

SOME INSIGHTS ON OUTPUT-HIDDEN-INPUT LAYER WEIGHT UPDATES

- The prediction error E propagates the weighted sum of all words in the vocabulary to every output vector v'_j
- The change in the input vector is defined by the output vector which in turn is updated due to the prediction error
- The model parameters accumulate the changes until the system reaches a state of equilibrium
- The rows in the Input-Hidden layer (v_j) stores the features of the words in the vocabulary V

CBOW MODEL FOR MULTIPLE WORDS

- C is the number of context words
- V is the size of the vocabulary
- h_i receives average of the vectors of the input context words
- Output vector v'_{wj} is the column vector in the W' representing relationship between the context words and the target word
- Softmax is used for the output layer probability distribution for the target word



The CBOW Model

INPUT-HIDDEN WEIGHT VECTORS AND LOSS FUNCTION

The hidden units receive values from the linear combination of the context vectors and the weights

$$u_j = \mathrm{v}'^{T}_{\mathrm{w}_j}\mathrm{h} = \mathrm{v}'^{T}_{\mathrm{w}_j}\mathrm{v}_{\mathrm{w}_I} \tag{16}$$

$$\begin{aligned} \mathrm{h} &= \frac{1}{C}\mathrm{W}^{T}(x_1 + x_2 + x_3 + \cdots + x_C) \\ &= \frac{1}{C}(\mathrm{v}_{w1} + \mathrm{v}_{w2} + \mathrm{v}_{w3} + \cdots + \mathrm{v}_{wC}) \end{aligned} \tag{17}$$

The equation for v'_j can be borrowed from (2) and E is

$$\begin{aligned} E &= -\log p(w_O | w_{I,1}, w_{I,2}, w_{I,3}, \ldots, w_{I,C}) \\ &= -v'_{wO}\cdot \mathrm{h} + \log \sum_{j'=1}^{V} \exp\left(\mathrm{v}'_{\mathrm{w}_j}{}^{T}\cdot \mathrm{h}\right) \end{aligned} \tag{18}$$

UPDATE INPUT AND OUTPUT VECTORS

There is no change in the hidden-output weights[3] (19)as the computations remain the same. The new $\mathrm{v}_{\mathrm{w}_{\mathrm{I},c}}^{(\mathrm{new})}$ is written as

$$\mathrm{v}_{\mathrm{w}_{\mathrm{I},c}}^{(\mathrm{new})} = \mathrm{v}_{\mathrm{w}_{\mathrm{I},c}}^{(\mathrm{old})} - \frac{1}{C} \cdot \eta \mathrm{EH}^{T}, \text{for} \quad j = 1, 2, 3, \ldots C \tag{20}$$

where η is the learning rate.

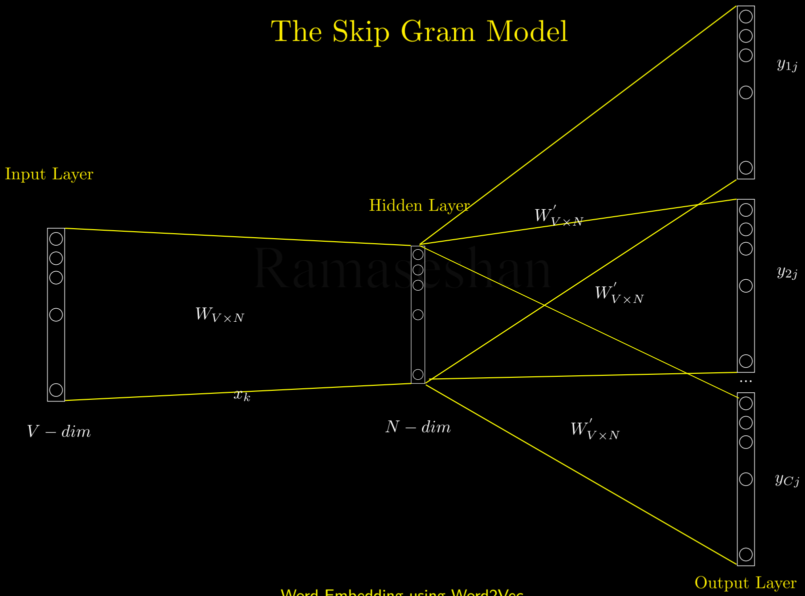
3

$$\mathrm{v}_{\mathrm{w}_j}^{(\mathrm{new})} = \mathrm{v}'^{(\mathrm{old})}_{\mathrm{w}_j} - \eta e_j \cdot \mathrm{h} \quad \text{for } j = 1, 2, 3, \ldots, V \tag{19}$$

WHAT DOES IT LEARN?

- Distributed representation of words as vectors
- The learned vectors explicitly encode many linguistic regularities and patterns
- The learning should produce a similar word vectors for those words that appeared in similar context. How do we find out?
- Comparing the word vectors for similarity? Cosine similarity?
- Has the learned word vectors address stemming? run, running, ran as similar?
  - He runs half-marathon
  - He ran half-marathon
  - He is running half-marathon
- How about car, cars, automobile?
- How about awesome, fantastic, great?

SKIP-GRAM MODEL



SKIP-GRAM MODEL - FORWARD AND BACK PROPAGATION UPDATE OF WEIGHTS

$$\text{Hidden layer } \mathbf{h} = W^{T}x \tag{21}$$

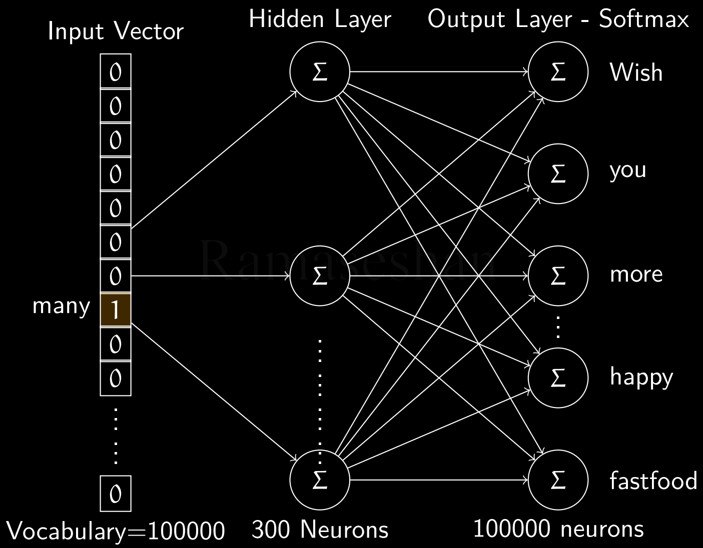
$$\text{Output neuron } u_c, j = \mathbf{v}'T_{w_j}.\mathbf{h}, \text{for } c = 1, 2, ...C \tag{22}$$

$$\text{cross entropy } E = -\sum_{c=1}^{C} u_j{}^*{}_c + C \cdot \log \sum_{j'=1}^{|v|} \exp(u_j') \tag{23}$$

$$\text{Context weights } w_{ij}'^{\text{new}} = w_{ij}'^{\text{old}} - \eta \cdot EI_j \cdot h_i \tag{24}$$

$$\text{input weights } w_{ij}^{\text{new}} = w_{ij}^{\text{old}} - \eta E_j \cdot h_i \tag{25}$$

NEURAL NETWORK ARCHITECTURE - A SAMPLE



A PYTHON IMPLEMENTATION FOR WORD EMBEDDING - 1/5

Initialization

```

    def setup_corpus(self,corpus_dir='/home/ramaseshan/Dropbox/NLPClass/2019/SmallCorpus/'):
        self.corpus = PlaintextCorpusReader(corpus_dir, '.*')

    def init_model_parameters(self,context_window_size=5,word_embedding_size=70,epochs=400,eta=0.01):
        self.context_window_size = context_window_size
        self.word_embedding_size = word_embedding_size
        self.epochs = epochs
        self.eta = eta

    def initialize_weights(self):
        self.embedding_weights = np.random.uniform(-0.9, 0.9, (self.vocabulary_size, self.word_embedding_size)) #input weights
        self.context_weights = np.random.uniform(-0.9, 0.9, (self.word_embedding_size, self.vocabulary_size)) #input weights
```

A PYTHON IMPLEMENTATION FOR WORD EMBEDDING - 2/5

Forward pass

$$H = W^T X$$

$$U = W'^T H = W'^T . W^T X$$

```python
    def forward_pass(self,X):
        H = np.dot(self.embedding_weights.T, X)
        U = np.dot(self.context_weights.T, H)
        y_hat = self.softmax(U)
        return y_hat, H, U
```

A PYTHON IMPLEMENTATION FOR WORD EMBEDDING - 3/5

Back propagation

$$w'^{\,new}_{ij} = w'^{\,old}_{ij} - \eta e_j \cdot h_i \text{ or}$$

$$v^{(new)}_{w_j} = v'^{(old)}_{w_j} - \eta e_j \cdot h \quad \text{for } j = 1, 2, 3, \ldots, V$$

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = EH_i . x_k$$

```
    def back_propagation(self,X,H,E):
        delta_context_weights = np.outer(H, E)
        delta_embedding_weights = np.outer(X, np.dot(self.context_weights, E.T
    ))

        # Change the weights using the back propagation values
        self.context_weights = self.context_weights - (self.eta *
    delta_context_weights)
        self.embedding_weights = self.embedding_weights - (self.eta *
    delta_embedding_weights)
        pass
```

A PYTHON IMPLEMENTATION FOR WORD EMBEDDING - 4/5

Training

$$E = -v'_{wO} \cdot h + \log \sum_{j'=1}^{V} \exp\left({v'_{w_j}}^{T} \cdot h\right) \tag{26}$$

```
    def train(self):
        for i in range(0, self.epochs):
            for target_word, context_words in np.array(self.training_samples):
                #for all the words
                y_hat, H, U = self.forward_pass(target_word)
                # compute error for all context words
                EI = np.sum([np.subtract(y_hat, word) for word in
                                context_words],axis=0)
                # back propagation to adjust weights
                self.back_propagation(target_word, H,EI)
                #Compute the error
                self.error[i] = -np.sum([U[word.index(1)]
                                        for word in context_words]) + \
                                          len(context_words) * \
                                          np.log(np.sum(np.exp(U)))
```

A PYTHON IMPLEMENTATION FOR WORD EMBEDDING - 5/5

Word vector for ***deep*** and similar words

$$\begin{bmatrix} -2.01970447 & 0.68963328 & 0.35593417 & 0.64125108 & \dots & 0.91503001 \end{bmatrix}$$

Word	Similarity
deep	1.0
heard	0.767841548247
depth	0.706466540662
well	0.684150968491
sound	0.662830677002
peso	0.507131975602
hit	0.464345901325
after	0.458074275823
water	0.424398813383

SOURCE PREPARATION FOR TRAINING

Source Text

$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{many}}$ more happy returns of the day$\longrightarrow$

$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{more}}\boxed{\text{happy}}$ returns of the day$\longrightarrow$

$\boxed{\text{Wish}}\boxed{\text{you}}\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}$ returns of the day$\longrightarrow$

Wish $\boxed{\text{you}}\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}$ of the day$\longrightarrow$

Wish you $\boxed{\text{many}}\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}$ the day$\longrightarrow$

Wish you many $\boxed{\text{more}}\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}\boxed{\text{the}}$ day$\longrightarrow$

Wish you many more $\boxed{\text{happy}}\boxed{\text{returns}}\boxed{\text{of}}\boxed{\text{the}}\boxed{\text{day}}$ $\longrightarrow$

Training Samples

$(wish, you)$

$(wish, many)$

$(you, Wish)$

$(you, more), (you, happy)$

$(many, Wish), (many, you)$

$(many, more), (many, happy)$

$(more, many), (more, you)$

$(more, happy), (more, returns$

$(happy, many), (happy, more$

$(happy, returns), (happy, of)$

$(returns, more), (returns, happ$

$(returns, of), (returns, the)$

$(of, happy), (of, returns)$

$(of, the), (of, day)$

SUB-SAMPLING

- The words (of, the) in the pairs $(of, happy), (returns, the)$ do not give much information about the words happy and returns, respectively. Similarly, some pairs reappear with the order of the words switched.
- Some words could also be randomly removed from the based on the frequencies
- Words with less frequency or infrequent words appearing as context words could be discarded as they may not provide contextual information to the central word

SUB-SAMPLING IN WORD2VEC.C-GOOGLE

Here is the code for sub-sampling used by word2vec.c that randomly removes a word from the sample using an algorithm similar to Linear Congruential generator (LCG) algorithm.

```c
      if (word == 0) break;
      // The subsampling randomly discards frequent words while keeping the
  ranking same
      if (sample > 0) {
        real ran = (sqrt(vocab[word].cn / (sample * train_words)) + 1) * (
  sample * train_words) / vocab[word].cn;
        next_random = next_random * (unsigned long long)25214903917 + 11;
        if (ran < (next_random & 0xFFFF) / (real)65536) continue;
      }
```

$$\text{let}\quad f(x) = \frac{vocab[word].cn}{train\_words} \quad \text{and} \quad ran = \left(\sqrt{f(x)}+1\right) \times \frac{1}{f(x)} \tag{27}$$

where $vocab[word].cn$ is the count of the word $word$ and train_words represents all all the training words.Then, the probability of keeping the word is decided based on the generated random value $random$. If $ran < random$ keep it, else discard the word

TROUBLE WITH THE SIZE OF THE NETWORK

- All weights $(output \rightarrow hidden)$ and $(hidden \rightarrow input)$ are adjusted by taking a training sample so that the prediction cycle minimizes the loss function
- This amounts to updating all the weights in the neural network - amounts to several million weights for a network which has input neurons, $|V| = 1M$, and hidden unit size as 300
- In addition, we should consider the several million training samples pairs

NEGATIVE SAMPLING

- The size of the network is proportional to the size of the vocabulary V. For every training cycle of input, the every weight in the network needs to be updated
- For every training cycle, Softmax function computes the sum of the output neuron values
- Cost of updating all the weights in the fully connected network is very high
- Is it possible to change only a small percentage of the weights?
- Select a small number of *negative* words
- While updating the weights, these samples output zero while the positive sample(s) will retain its value
- During the backpropagation, the weights related to the negative and positive words are changed and the rest will remain untouched for the current update
- This reduces drastically the computation $\smile$

SELECTING A NEGATIVE SAMPLE

$$P(w_i) = \frac{f(w_i)}{\sum_{j=0}^{n} f(w_j)} \tag{28}$$

$$P(w_i) = \frac{f(w_i)^{\frac{3}{4}}}{\sum_{j=0}^{n} f(w_j)^{\frac{3}{4}}} \tag{29}$$

It is important to choose more frequent words.This equation decreases the probability of choosing the less frequent words. One way to implement is to create a unigram table filled with the words according to the probability. The frequently occurring words would be repeated several times according to their frequency thereby increasing the probability of choosing the frequent words for the ***negative*** samples

NEGATIVE SAMPLING - AN EXAMPLE

- Modify a small percentage of the weights, rather than all of them
- Consider chalk and blackboard pair for one-word learning one-word learning.
  - We want blackboard to have one $(= 1)$ at the output layer
  - Traditionally, we want rest of the words to have zero $(= 0)$
  - with Negative sampling, we pick a random set of words. Let us say we pick up words coffee, car, elephant, tower as negative samples. Instead of making every word zero, we turn the negative values to be equal to zero and rest unchanged

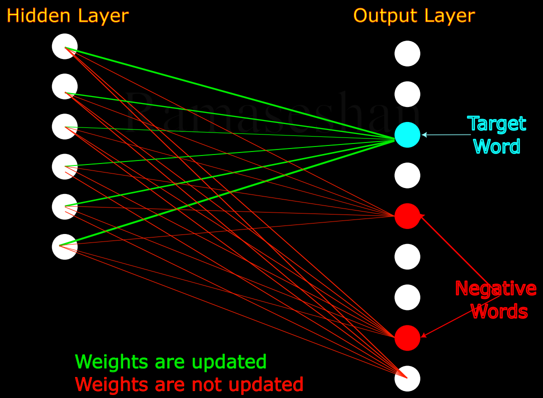
Logistic regression is used loss function used in when Negative Sampling is used:

$$E = -\left(\log\left(\sigma(z_{\text{pos}}\right) + \sum_{i=1}^{k} \log\left(1 - \sigma(z_{\text{neg}_i}\right)\right) \tag{30}$$

where z_{pos} is the input for the positive word, aiming for a value close to 1 and z_{neg_i} (for $i = 1$ to k) represents the inputs for each of the k negative words, aiming for values close to 0.

ADVANTAGES OF NEGATIVE SAMPLING

- **Faster training time** - especially for large datasets. It allows scaling of the
- **Better accuracy** Improves accuracy by avoiding over fitting
- **Optimized utilization of memory**



SOFTMAX

To deal with classification with multiple classes, softmax is very useful.If there are k classes in the data set, this activation function fits the classes in the range [0,1] by calculating the probability. This is best suited for the finding the activation value of the neurons in the output layer.It is a normalized exponential function

$$P(C_k|x_j) = \frac{e^{a_j}}{\sum_k e^{a_k}}, where \quad k = 1, K \tag{31}$$

HIERARCHICAL SOFTMAX

- Has a flat hierarchy with a probability value for every output node of depth $= 1$
- Normalized over the probabilities of all $|V|$ words
- Error correction happens for every output$\rightarrow$hidden units
- Huge costs if the vocabulary size $|V|$ is of the order of several thousands
- Decompose the flat hierarchy into a binary tree
- Form a hierarchical description of a word as a sequence of $O(log_2|V|)$ decisions and there by reducing the computing complexity of Softmax - $O(|V|) \rightarrow O(log_2(|V|))$
- Lay the words in a tree-based hierarchy - words as leaves
- Binary tree with $|V|-1$ nodes for left (0) and right(1) traversal
- Every leave represents the probability of the word

HIERARCHICAL SOFTMAX - 2

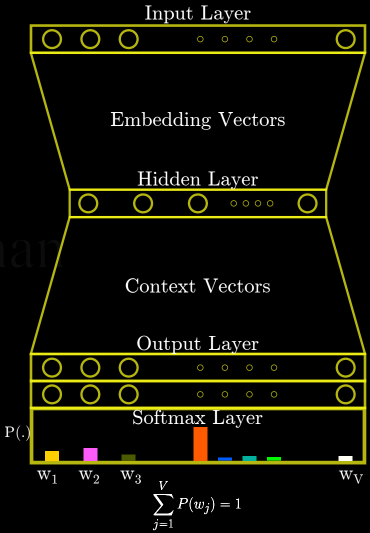
- Path length of a balanced Tree is $log_2(|V|)$. If the $|V| = 1 \quad million \quad words$, then the path length $= 19.9$ bits/word
- Constructing an Huffman encoded-tree would help frequent words to have short unique binary codes
- Learn to take these probabilistic decisions instead of directly predicting each word's probability [1]
- Every intermediate node denotes the relative probabilities of its child nodes
- The path to reach every leaf (word) is unique
- H-Softmax in many cases increases the prediction speed by more than 50X times

SOFTMAX

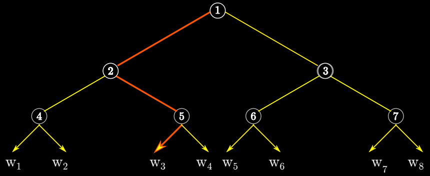
Softmax is a normalized exponential function

$$P(C_k|x_j) = \frac{e^{a_j}}{\Sigma_k e^{a_k}}, where \quad k = 1, K \quad (32)$$

- Has a flat hierarchy with a probability value for every output node of depth $= 1$
- Normalized over the probabilities of all $|V|$ words
- Error correction happens for every output$\rightarrow$hidden units
- Huge costs if the vocabulary size $|V|$ is of the order of several thousands

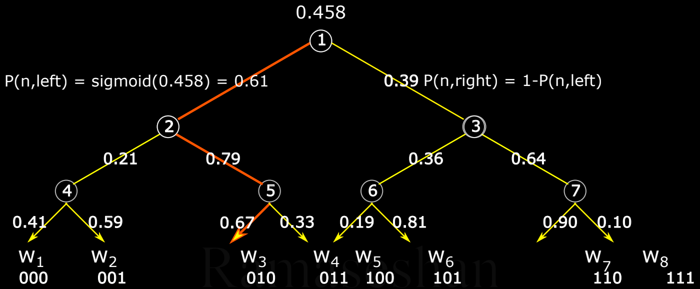


BALANCED BINARY TREE



- Move the words into a binary tree - depth depends on the vocabulary
- The path to any word in the vocabulary is known
- Traverse through the binary tree to reach any word
- At every step, make a binary decision to reach the word
- The length to reach any word in a balanced tree is $\log_2(|V|)$
- Words could be arranged using
  - random order
  - IS-A relationship
  - TF-IDF frequency

BALANCED BINARY TREE WITH 8 WORDS



$P(w_i) = \prod_{j \in N_L} P(n(w,j))$,where N_L is the list of nodes to reach the word and

$\mathbf{P}(\mathbf{W}) = \sum_{i=1}^{V} P(w_i) = 1$

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$P(w_i)$	0.05	0.08	0.32	0.16	0.04	0.11	0.22	0.02

Thus the hierarchical Softmax is a well defined multinomial distribution among all words

HIERARCHICAL SOFTMAX - ADVANTAGES

- Decomposes the flat hierarchy into a binary tree
- The path to reach every leaf (word) is unique
- Lays the words in a tree-based hierarchy - words as leaves
- Binary tree with $|V|-1$ nodes for left and right traversal
- Every intermediate node denotes the relative probabilities of its child nodes
- Every leaf represents the probability of the word

HIERARCHICAL SOFTMAX - ADVANTAGES

- Each node is indexed by a bit vector corresponding to the path from the root to the node
- Normalized values for the words are calculated without finding the probability for every word
- The entire vocabulary is partitioned into classes
- ANN learns to take these probabilistic decisions instead of directly predicting each word's probability[4]

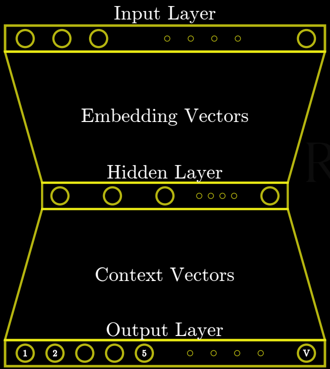
[4] Yoshua Bengio et al. "A Neural Probabilistic Model" In: J.Mach.Learn Res.3 (Mar.2003), pp.1137-1155. issn: 1532-4435

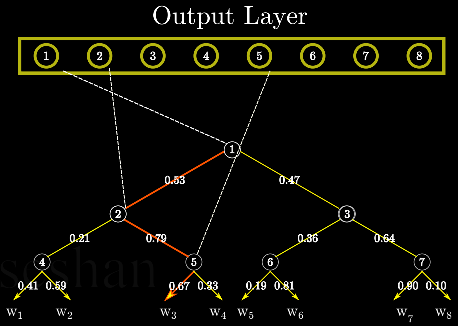
HIERARCHICAL SOFTMAX - ADVANTAGES

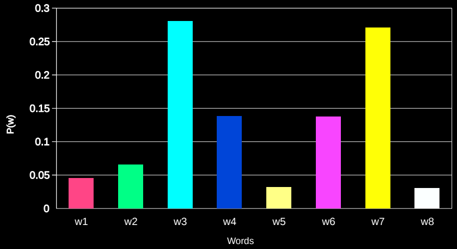
- Forms a hierarchical description of a word as a sequence of $O(\log_2|V|)$ decisions
- Reduces the computing complexity of Softmax - $O(|V|) \rightarrow O(\log_2(|V|))$
- Path length of a balanced Tree is $\log_2(|V|)$. If the $|V| = 1 \quad million \quad words$, then the path length $= 19.9$ bits/word
- A balanced binary tree should provide an exponential speed-up, on the order of $$\frac{|V|}{\log_2(|V|)}$$
- Constructing an Huffman encoded-tree would help frequent words to have short unique binary codes
- H-Softmax in many cases increases the prediction speed by more than 50X times

Note: Hierarchical softmax significantly reduces the computational complexity and it doesn't completely eliminate the need to compute probabilities for all words

WORD2VEC WITH HIERARCHICAL SOFTMAX







UPDATING WEIGHTS - 1/3

Let $L(w)$ be the number of nodes to traverse to the word from the root and $n(w,i)$ is the i^{th} node on this path and the associated vector in the context matrix is $v_{n(w,i)}$. $ch(n)$ is the child node [2][1][3][4]. Then the probability of word is

$$P(w|w_i) = \prod_{j=1}^{L(w)-1} \sigma([n(w,j+1) = ch(n(w,j))].v^T_{n(w,j)}h)$$

$$= \prod_{j=1}^{L(w)-1} \sigma([n(w,j+1) = ch(n(w,j))].v^T_{n(w,j)}v_{w_i}) \tag{33}$$

$$\text{where}[x] = \begin{cases} 1, & \text{if } x \text{ is true} \\ -1, & \text{otherwise} \end{cases} \text{ and } \sigma(.)\text{is the sigmoid function}$$

if the child node $ch(n(w,j))$ is left of the parent node, then the term $[n(w,j+1) = ch(n(w,j))]$ is 1, and equal to -1, if the path goes to the right.

UPDATING WEIGHTS - 2/3

Since the sum of the probabilities of at the node is 1, we can prove that

$$\sigma(v_n^T v_{w_i}) + \sigma(-v_n^T v_{w_i}) = 1, \text{ since } 1 - \sigma(x) = \sigma(-x) \tag{34}$$

Example

$$P(w|w_i) = \sigma(v^T_{n(w,j)} v_{w_i}).\sigma(-v^T_{n(w,j)} v_{w_i}).\sigma(v^T_{n(w,j)} v_{w_i})$$

To train the model, we need to minimize the negative log likelihood $-\log P(w|w_i)$

$$E = -\sum_{j=1}^{L(w)-1} \log \sigma([.]u_j'), \text{ where } u_j = v_j'.h \tag{35}$$

$$\tag{36}$$

where $t_j = 1$, if $[n(w, j+1) = ch(n(w,j))] = 1$ and $t_j = 0$ otherwise

UPDATING WEIGHTS - 3/3

$$\frac{\partial E}{\partial u_j} = \sigma(u_j - 1)[.] \tag{37}$$

$$= \begin{cases} \sigma(u_j) - 1 & \text{for } [.] = 1 \\ \sigma(u_j) & \text{for } [.] = -1 \end{cases} \tag{38}$$

$$= \sigma(u_j) - t_j \tag{39}$$

where $t_j = 1$, if $[n(w, j+1) = ch(n(w, j))] = 1$, else $t_j = 0$

$$\frac{\partial E}{\partial v'_j} = \sigma(v'_j.h) - t_j \tag{40}$$

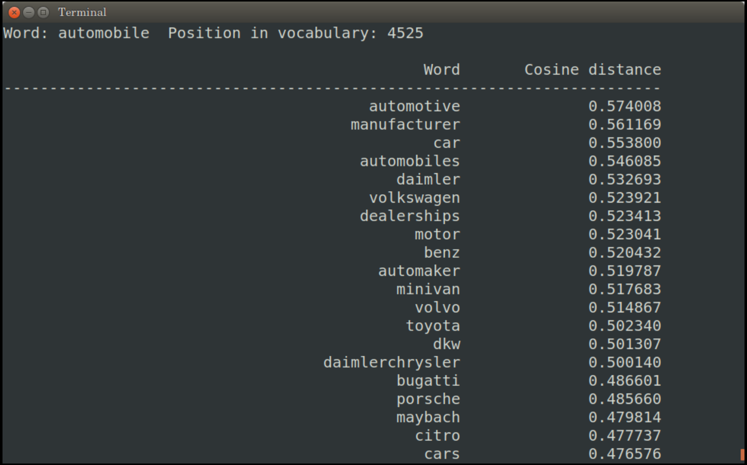
$$v'^{new}_j = v'^{old}_j - \eta(\sigma(v'_j.h) - t_j).h \tag{41}$$

$$\frac{\partial E}{\partial h} = \sum_{j=1}^{L(W)-1} \frac{\partial E}{\partial v'_j h} . \frac{\partial v'_j h}{\partial h} = EH \tag{42}$$

$$v^{new}_{w_i} = v^{old}_{w_i} - \eta . EH^T \tag{43}$$

WORD2VEC - RESULTS

The source code for word2vec is available at https://github.com/dav/word2vec. Word similarity for the word ***automobile***



```
Word: automobile  Position in vocabulary: 4525

                                              Word       Cosine distance
------------------------------------------------------------------------
                                        automotive              0.574008
                                      manufacturer              0.561169
                                               car              0.553800
                                       automobiles              0.546085
                                           daimler              0.532693
                                        volkswagen              0.523921
                                       dealerships              0.523413
                                             motor              0.523041
                                              benz              0.520432
                                         automaker              0.519787
                                           minivan              0.517683
                                             volvo              0.514867
                                            toyota              0.502340
                                               dkw              0.501307
                                   daimlerchrysler              0.500140
                                           bugatti              0.486601
                                          porsche               0.485660
                                           maybach              0.479814
                                             citro              0.477737
                                              cars              0.476576
```

BUILDING WORD2VEC FROM SCRATCH

1. **Data Preparation**
    - Preprocess the text corpus
    - Create a vocabulary of unique words.
    - Construct a reverse lookup to map words to their indices
2. **Generate Training Data**
    - Initialize a sliding window of size w
    - For each word in the corpus:
        - Extract the center word and context words
        - Add the pair to the training data
3. **Implement Skip-Gram Model**
    - Define the following layers:
        - *Input layer*: One-hot encoded center word
        - *Hidden layer*: Word embeddings
        - *Softmax layer:* predict context words
4. **Training**
    - Initialize the model parameters
    - For each epoch:
        - For every training data, ompute the loss
        - Update the model parameters
5. **Negative Sampling**
    - Use negative sampling speed up the training
        - Optimize the model to distinguish context words from negative words
6. **Training Loop**
    - Implement a training loop to iterate through the following steps:
        - Compute the gradients of the model parameters
        - Update the model parameters based on the gradients
7. **Save Word Embeddings**
    - Save the word embeddings to a file for future use
8. **Evaluation**
    - Evaluate the word embeddings using tasks like word similarity, word analogy, or text classification.
9. **Optimization and Performance**
    - Optimize your code for performance, as training Word2Vec can be computationally intensive
        - Use multi-threading or distributed computing if needed

LIMITATIONS

- Separate training is required for phrases
- Embeddings are learned based on a small local window surrounding words - good and bad share the almost the same embedding
- Does not address polysemy
- Does not use frequencies of term co-occurrences

REFERENCES

[1] Yoshua Bengio et al. "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3 (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944966.

[2] Jeffrey A. Dean Tomas Mikolov Kai ChenGregory S. Corrado. U.S. pat. US9037464B1. May 2015.

[3] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[4] Andriy Mnih and Geoffrey Hinton. "A Scalable Hierarchical Distributed Language Model". In: *Proceedings of the 21st International Conference on Neural Information Processing Systems*. NIPS'08. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008, pp. 1081–1088. ISBN: 978-1-6056-0-949-2. URL: