0.1 Paper

Formula (1) is the the Skip-gram objective function (log probability):

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

T is the total number of words (size of corpus), c is the size of the training context (window size), w_t is the center word, w_{t+j} is the target word (from context).

Each (w_I, w_O) is a data sample (center word and target word).

Formula (2) is using softmax function to represent $p(w_O|w_I)$:

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^{\mathrm{T}} v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^{\mathrm{T}} v_{w_I})}$$

v_w and v'_w are the "input" and "output" embeddings of w, and W is the number of words in the vocabulary.

So the final objective function is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log\frac{\exp({v'_{w_{t+j}}}^{\mathrm{T}} v_{w_t})}{\sum_{w=1}^{W}\exp({v'_w}^{\mathrm{T}} v_{w_t})}$$

Formula (3) is about hierarchical softmax function.

Formula (4) is using the negative sampling to replace every log $p(w_O|w_I)$ term in the original objective function:

$$\log\, p(w_O|w_I) = \log\, \sigma({v'_{w_O}}^{\mathrm{T}} v_{w_I}) + \sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}[\log\, \sigma(-{v'_{w_i}}^{\mathrm{T}} v_{w_I})]$$

$P_n(w)$ is the noise distribution using logistic regression, where there are k negative samples for each data sample.

So the final objective function is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log\, \sigma({v'_{w_{t+j}}}^{\mathrm{T}} v_{w_t}) + \sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}[\log\, \sigma(-{v'_{w_i}}^{\mathrm{T}} v_{w_t})]$$

So I use the negative log probability from formula (4) as loss function of each data sample:

$$L(w_I, w_O) = -\log\,\sigma({v'_{w_O}}^{\mathrm{T}} v_{w_I}) - \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)}[\log\,\sigma(-{v'_{w_i}}^{\mathrm{T}} v_{w_I})]$$

And the loss function of whole dataset is

$$L = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} L(w_t, w_{t+j})$$

Maximize the objective function is equivalently to minimize the loss function. So the objective of learning algorithm is

$$\arg\min_{\theta} \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} L(w_t, w_{t+j})$$

where $\theta = \{v, v'\}$ (the input and output embeddings)

Using stochastic gradient descent:

- Initialize $\theta = \{v, v'\}$
- For N Iterations:
  - For each training sample (w_I, w_O)
    - $\Delta = -\nabla_\theta L(w_I, w_O)$ (the gradient)
    - $\theta = \theta + \alpha\Delta$ (α is the learning rate)

The partial derivative of $L(w_I, w_O)$ is

$$\Delta_{v_{w_I}} = -\frac{\partial L(w_I, w_O)}{\partial v_{w_I}} = [1 - \log\,\sigma({v'_{w_O}}^{\mathrm{T}} v_{w_I})]v'_{w_O} + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)}[-\log\,\sigma({v'_{w_i}}^{\mathrm{T}} v_{w_I}))]v'_{w_i}$$

$$\Delta_{v'_{w_O}} = -\frac{\partial L(w_I, w_O)}{\partial v'_{w_O}} = [1 - \log\,\sigma({v'_{w_O}}^{\mathrm{T}} v_{w_I})]v_{w_I}$$

$$\Delta_{v'_{w_i}} = -\frac{\partial L(w_I, w_O)}{\partial v'_{w_i}} = [-\log\,\sigma({v'_{w_i}}^{\mathrm{T}} v_{w_I})]v_{w_I}$$

Updating $\theta = \{v, v'\}$:

$$v_{w_I} = v_{w_I} + \alpha\Delta_{v_{w_I}}$$

$$v'_{w_O} = v'_{w_O} + \alpha\Delta_{v'_{w_O}}$$

$$v'_{w_i} = v'_{w_i} + \alpha\Delta_{v'_{w_i}}$$

0.2 Code

The paper's objective function is:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log\,\sigma({v'_{w_{t+j}}}^{\mathrm{T}}v_{w_t})+\sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}[\log\,\sigma(-{v'_{w_i}}^{\mathrm{T}}v_{w_t})]$$

The code's objective function is a little different:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,j\neq 0}\log\,\sigma({v'_{w_t}}^{\mathrm{T}}v_{w_{t+j}})+\sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}[\log\,\sigma(-{v'_{w_i}}^{\mathrm{T}}v_{w_{t+j}})]$$

So the relative loss function and gradient are a little different.

line 338 - line 360:
Initialization of syn0 , syn1 and syn1neg.
syn0 is v
syn1neg is v'
syn1 is used for hierarchical softmax

line 387 - line 405:
Building sentence.
Every 1000 words make up a sentence.

line 374 - line 386:
Updating learning rate every 10000 words (10 sentences)
Learning rate decreases linearly.

line 416:
"word": w_t

line 483 - line 531:
Skip-gram model (both negative samples and hierarchical softmax)

line 487:
"last word": w_{t+j}

line 489:
"l1": the index of w_{t+j} in syn0

line 508 - line 529:
Training Skip-gram model with negative sampling.

line 510:
target word as positive sample

line 514:
target word as negative samples

line 519:
"l2": the index w_t or w_i in syn1neg