Assignment 5: Implement the Continuous Bag of Words (CBOW) Model

Name: Zahir Mulla

In [9]:
```
# Rollno: B512012
# class: BE-IT(B)
```

In [1]:
```
#importing libraries
from keras.preprocessing import text
from keras.preprocessing import sequence
from keras.utils import pad_sequences
from keras.utils import to_categorical
import numpy as np
import pandas as pd
```

In [2]:
```
#taking random sentences as data
data = """Deep learning (also known as deep structured learning) is part of a broad
Deep-learning architectures such as deep neural networks, deep belief networks, dee
"""
dl_data = data.split()
```

In [3]:
```
#tokenization
tokenizer = text.Tokenizer()
tokenizer.fit_on_texts(dl_data)
word2id = tokenizer.word_index

word2id['PAD'] = 0
id2word = {v:k for k, v in word2id.items()}
wids = [[word2id[w] for w in text.text_to_word_sequence(doc)] for doc in dl_data]

vocab_size = len(word2id)
embed_size = 100
window_size = 2

print('Vocabulary Size:', vocab_size)
print('Vocabulary Sample:', list(word2id.items())[:10])
```

```
Vocabulary Size: 75
Vocabulary Sample: [('learning', 1), ('deep', 2), ('networks', 3), ('neural', 4),
('and', 5), ('as', 6), ('of', 7), ('machine', 8), ('supervised', 9), ('have', 10)]
```

In [4]:
```
#generating (context word, target/label word) pairs
def generate_context_word_pairs(corpus, window_size, vocab_size):
    context_length = window_size*2
    for words in corpus:
        sentence_length = len(words)
        for index, word in enumerate(words):
            context_words = []
            label_word   = []
            start = index - window_size
            end = index + window_size + 1

            context_words.append([words[i]
                                 for i in range(start, end)
                                 if 0 <= i < sentence_length
                                 and i != index])
            label_word.append(word)

            x = pad_sequences(context_words, maxlen=context_length)
            y = to_categorical(label_word, vocab_size)
            yield (x, y)

i = 0
for x, y in generate_context_word_pairs(corpus=wids, window_size=window_size, vocab
    if 0 not in x[0]:
        # print('Context (X):', [id2word[w] for w in x[0]], '-> Target (Y):', id2wo

        if i == 10:
            break
        i += 1
```

In [5]:

```
#model building
import keras.backend as K
from keras.models import Sequential
from keras.layers import Dense, Embedding, Lambda

cbow = Sequential()
cbow.add(Embedding(input_dim=vocab_size, output_dim=embed_size, input_length=windov
cbow.add(Lambda(lambda x: K.mean(x, axis=1), output_shape=(embed_size,)))
cbow.add(Dense(vocab_size, activation='softmax'))
cbow.compile(loss='categorical_crossentropy', optimizer='rmsprop')

print(cbow.summary())

# from IPython.display import SVG
# from keras.utils.vis_utils import model_to_dot

# SVG(model_to_dot(cbow, show_shapes=True, show_layer_names=False, rankdir='TB').cr
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 4, 100)            7500

 lambda (Lambda)             (None, 100)               0

 dense (Dense)               (None, 75)                7575

=================================================================
Total params: 15075 (58.89 KB)
Trainable params: 15075 (58.89 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
None
```

In [6]:

```
for epoch in range(1, 6):
    loss = 0.
    i = 0
    for x, y in generate_context_word_pairs(corpus=wids, window_size=window_size, v
        i += 1
        loss += cbow.train_on_batch(x, y)
        if i % 100000 == 0:
            print('Processed {} (context, word) pairs'.format(i))

    print('Epoch:', epoch, '\tLoss:', loss)
    print()
```

```
Epoch: 1        Loss: 433.4012360572815

Epoch: 2        Loss: 429.2743980884552

Epoch: 3        Loss: 426.0699987411499

Epoch: 4        Loss: 422.92839217185974

Epoch: 5        Loss: 420.40380811691284
```

In [7]:

```
weights = cbow.get_weights()[0]
weights = weights[1:]
print(weights.shape)

pd.DataFrame(weights, index=list(id2word.values())[1:]).head()
```

```
(74, 100)
```

Out[7]:

	0	1	2	3	4	5	6	7
deep	-0.032821	-0.007050	0.045066	0.006504	-0.018352	0.000622	-0.019529	0.005322
networks	0.050687	-0.057524	0.065664	0.030807	-0.039223	0.051308	-0.023726	-0.046275
neural	-0.031908	-0.033468	0.037874	-0.014744	-0.033155	-0.035152	0.019606	-0.021598
and	-0.030528	-0.023409	-0.010739	0.033079	-0.024328	-0.030563	-0.049606	0.003453
as	0.038382	-0.020246	-0.032325	0.033337	-0.021431	0.031506	-0.025571	0.029777

5 rows × 100 columns

In [8]:

```
from sklearn.metrics.pairwise import euclidean_distances

distance_matrix = euclidean_distances(weights)
print(distance_matrix.shape)

similar_words = {search_term: [id2word[idx] for idx in distance_matrix[word2id[sear
                   for search_term in ['deep']}

similar_words
```

```
(74, 74)
```

Out[8]:

```
{'deep': ['bioinformatics', 'artificial', 'applied', 'human', 'unsupervised']}
```

In [ ]: