Application of Deep Learning to Text and Images

Module 2, Lab 3: GloVe Word Vectors

This notebook supports the topics presented on on the **Word Embeddings** lecture.

In this lab you will learn how to use word embeddings. Word embeddings, or word vectors, are a way of representing words as numeric vectors in a high-dimensional space. These embeddings capture the meaning of the words, the relationships between them, and can be used as inputs to machine learning models for a variety of natural language processing tasks.

The term **Word vectors** refers to a family of related techniques, first gaining popularity via `Word2Vec` which associates an n-dimensional vector to every word in the target language.

- **Note**: Normally n is in the range of 50 to 500. In this lab, you will set it to 50

You will learn:

- What GloVe word vectors are
- How to load GloVe word vectors
- How to use GloVe to produce word vectors
- What cosine Similarity is
- How to use cosine similarity to compare words

You will be presented with two kinds of exercises throughout the notebook: activities and challenges.

Index

First, install the latest versions of the libraries.

In [9]:
```
# installing libraries
!pip install -U -q -r requirements.txt
```

In [10]:
```
from torchtext.vocab import GloVe

#from torchtext.vocab import GloVe
GloVe.url['6B'] = 'https://huggingface.co/stanfordnlp/glove/resolve/main/glo

from sklearn.decomposition import PCA
from sklearn.metrics.pairwise import cosine_similarity

%matplotlib inline
import matplotlib.pyplot as plt
```

GloVe Word Vectors

You learned about **Word2Vec** and **FastText** as word embedding techniques. Now you will use a set of pre-trained word embeddings. Pre-trained embeddings are created by someone else who took the time and computational power to train. This reduces your cost by not having to train the model yourself. One popular word embedding is **GloVe** embeddings. GloVe is a variation of a Word2Vec model. To learn more about GloVe, read the Project GloVe (https://nlp.stanford.edu/projects/glove/) website.

In this exercise, you will discover relationships between word vectors using the GloVe embeddings.

You can easily import GloVe embeddings from the Torchtext library. Here, you will get vectors with 50 dimensions.

The `name` parameter refers to the particular pre-trained model that should be loaded:

- Wikipedia 2014 + Gigaword 5
  - 6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download: `"6B"`
  - This is the model that you will load.
- Common Crawl
  - 42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download: `"42B"`
- Common Crawl
  - 840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download: `"840B"`
- Etc
  - See documentation in Stanford link above

Try it Yourself!

Activity

Run the cell below to load the GloVe embedding model and select the dimension.

In [11]:
```
# Load the model. You can change dim to 50, 100, 300
glove = GloVe(name="6B", dim=50)
```

Now that the data is loaded, you can access it and print example word embeddings.

In [12]:
```
print(f"cat -> {glove['cat']}\n")
```

```
cat -> tensor([ 0.4528, -0.5011, -0.5371, -0.0157,  0.2219,  0.5460, -0.67
30, -0.6891,
         0.6349, -0.1973,  0.3368,  0.7735,  0.9009,  0.3849,  0.3837,  0.
2657,
        -0.0806,  0.6109, -1.2894, -0.2231, -0.6158,  0.2170,  0.3561,  0.
4450,
         0.6089, -1.1633, -1.1579,  0.3612,  0.1047, -0.7832,  1.4352,  0.
1863,
        -0.2611,  0.8328, -0.2312,  0.3248,  0.1449, -0.4455,  0.3350, -0.
9595,
        -0.0975,  0.4814, -0.4335,  0.6945,  0.9104, -0.2817,  0.4164, -1.
2609,
         0.7128,  0.2378])

cat -> tensor([ 0.4528, -0.5011, -0.5371, -0.0157,  0.2219,  0.5460, -0.67
30, -0.6891,
         0.6349, -0.1973,  0.3368,  0.7735,  0.9009,  0.3849,  0.3837,  0.
2657,
        -0.0806,  0.6109, -1.2894, -0.2231, -0.6158,  0.2170,  0.3561,  0.
4450,
         0.6089, -1.1633, -1.1579,  0.3612,  0.1047, -0.7832,  1.4352,  0.
1863,
        -0.2611,  0.8328, -0.2312,  0.3248,  0.1449, -0.4455,  0.3350, -0.
9595,
        -0.0975,  0.4814, -0.4335,  0.6945,  0.9104, -0.2817,  0.4164, -1.
2609,
         0.7128,  0.2378])
```

What do these numbers mean?

You might notice that the tensor has 50 values in it. This is related to the dimension flag (`dim=50`) you set when you loaded the GloVe model. You can generate word embeddings for several words and use them to determine how closely related words are. This is a task that machine learning is really good at.

Try it Yourself!

Challenge

In the code block below, generate word embeddings for the words "computer" and "human" using pre-trained GloVe embedding.

In [14]:

```
############## CODE HERE ###############
import numpy as np

def load_glove_embeddings(glove_model):
    embeddings = {}
    for i, word in enumerate(glove_model.itos):
        embeddings[word] = glove_model.vectors[i].numpy()
    return embeddings

embeddings_index = load_glove_embeddings(glove)

# Retrieve embeddings for "computer" and "human"
computer_vector = embeddings_index.get("computer")
human_vector = embeddings_index.get("human")

print("Embedding for 'computer':", computer_vector)
print("Embedding for 'human':", human_vector)

############## END OF CODE #################
```

```
Embedding for 'computer': [ 0.079084 -0.81504   1.7901    0.91653   0.1079
7  -0.55628  -0.84427
 -1.4951    0.13418   0.63627   0.35146   0.25813  -0.55029   0.51056
  0.37409   0.12092  -1.6166    0.83653   0.14202  -0.52348   0.73453
  0.12207  -0.49079   0.32533   0.45306  -1.585    -0.63848  -1.0053
  0.10454  -0.42984   3.181    -0.62187   0.16819  -1.0139    0.064058
  0.57844  -0.4556    0.73783   0.37203  -0.57722   0.66441   0.055129
  0.037891  1.3275    0.30991   0.50697   1.2357    0.1274   -0.11434
  0.20709 ]
Embedding for 'human': [ 0.61854   0.11915  -0.46786   0.31368   1.0334
0.95964   0.87803
 -1.0346    1.6322    0.29347   0.80844  -0.058903  0.021251  0.40986
  0.54443  -0.33311   0.53712  -0.35823   0.29374   0.090151 -0.92049
  0.69386   0.39098  -0.64392   0.77831  -1.7215   -0.48393  -0.50327
 -0.22508   0.099192  3.2095   -0.31554  -0.71754  -1.6752   -1.3537
  0.15195   0.054557 -0.1633   -0.027993  0.3917   -0.55007  -0.079205
  0.63389   0.51446   0.70124   0.27638  -0.53445   0.064808 -0.21974
 -0.52048 ]
Embedding for 'computer': [ 0.079084 -0.81504   1.7901    0.91653   0.1079
7  -0.55628  -0.84427
 -1.4951    0.13418   0.63627   0.35146   0.25813  -0.55029   0.51056
  0.37409   0.12092  -1.6166    0.83653   0.14202  -0.52348   0.73453
  0.12207  -0.49079   0.32533   0.45306  -1.585    -0.63848  -1.0053
  0.10454  -0.42984   3.181    -0.62187   0.16819  -1.0139    0.064058
  0.57844  -0.4556    0.73783   0.37203  -0.57722   0.66441   0.055129
  0.037891  1.3275    0.30991   0.50697   1.2357    0.1274   -0.11434
  0.20709 ]
Embedding for 'human': [ 0.61854   0.11915  -0.46786   0.31368   1.0334
0.95964   0.87803
 -1.0346    1.6322    0.29347   0.80844  -0.058903  0.021251  0.40986
  0.54443  -0.33311   0.53712  -0.35823   0.29374   0.090151 -0.92049
  0.69386   0.39098  -0.64392   0.77831  -1.7215   -0.48393  -0.50327
 -0.22508   0.099192  3.2095   -0.31554  -0.71754  -1.6752   -1.3537
  0.15195   0.054557 -0.1633   -0.027993  0.3917   -0.55007  -0.079205
  0.63389   0.51446   0.70124   0.27638  -0.53445   0.064808 -0.21974
 -0.52048 ]
```

Cosine Similarity

You learned about cosine similarity in class, now let's look at an example. Use the cosine_similarity() (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html) function from scikit-learn to easily calculate cosine similarity between word vectors.

Try it Yourself!

Activity

Run the cell below to calculate cosine similarity between word vectors.

In [17]:
```
# define the similarity between two words
def similarity(w1, w2):
    return cosine_similarity([glove[w1].tolist()], [glove[w2].tolist()])


# Say if w1 is closer to w2 than w3
def simCompare(w1, w2, w3):
    s1 = similarity(w1, w2)
    s2 = similarity(w1, w3)
    if s1 > s2:
        print(f"'{w1}'\tis closer to\t'{w2}'\tthan\t'{w3}'\n")
    else:
        print(f"'{w1}'\tis closer to\t'{w3}'\tthan\t'{w2}'\n")
```

In [18]:
```
simCompare("actor", "pen", "film")
simCompare("cat", "dog", "sea")
```

```
'actor'	is closer to	'film'	than	'pen'

'cat'	is closer to	'dog'	than	'sea'

'actor'	is closer to	'film'	than	'pen'

'cat'	is closer to	'dog'	than	'sea'
```

Try it Yourself!

Challenge

Write code to determine if "car" is closer to "truck" than "bike".

In [22]:
```
############### CODE HERE ################

# Retrieve embeddings for words
car_vector = embeddings_index.get("car")
truck_vector = embeddings_index.get("truck")
bike_vector = embeddings_index.get("bike")

# Compute cosine similarities
similarity_car_truck = cosine_similarity([car_vector], [truck_vector])[0][0]
similarity_car_bike = cosine_similarity([car_vector], [bike_vector])[0][0]

# Determine which is closer
if similarity_car_truck > similarity_car_bike:
    print("'Car' is closer to 'Truck' than 'Bike'.")
else:
    print("'Car' is closer to 'Bike' than 'Truck'.")

 ############### END OF CODE ##################
```

```
'Car' is closer to 'Truck' than 'Bike'.
'Car' is closer to 'Truck' than 'Bike'.
```

Conclusion

You have now seen how to use word embeddings and determine relationships between word vectors using the GloVe embeddings.

Next Lab: Word Embeddings

In the next lab of this module you will learn how to build a recurrent neural network (RNN) with PyTorch. It will also show you how to implement a simple RNN-based model for natural language processing.