



✓ Application of Deep Learning to Text and Images

Module 2, Lab 3: GloVe Word Vectors

This notebook supports the topics presented 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.

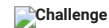
The term **Word vectors** refers to a family of related techniques, first gaining popularity via Word2Vec which associates an n -dimensional vector to every word in the target language.

- **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.



No coding is needed for an activity. You try to understand a concept, answer questions, or run a code cell.

Challenges are where you can practice your coding skills.

✓ Index

1. [GloVe Word Vectors](#)
2. [Cosine Similarity](#)

First, install the latest versions of the libraries.

```
# installing libraries
!pip install -U -q -r requirements.txt
```

```
↗ ERROR: Could not open requirements file: [Errno 2] No such file or directory: 'requirements.txt'
```

```
# install torchtext
!pip install torchtext
```

```
# installing libraries listed in requirements.txt
!pip install -U -q -r requirements.txt
```

```
#importing modules
from torchtext.vocab import GloVe
from sklearn.decomposition import PCA
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt
```

```
#modifying the url of glove data to be pulled from huggingface
GloVe.url['6B'] = 'https://huggingface.co/stanfordnlp/glove/resolve/main/glove.6B.zip'
```

```
%matplotlib inline
```

```
↗ Show hidden output
```

✓ 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](#) 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!



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

```
# Install PyTorch, ensuring CUDA compatibility if needed
# You might need to adjust the CUDA version (cu118, cu121, etc.) based on your system
!pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118

# Remove any existing torchtext installations
!pip uninstall -y torchtext

# Install a compatible version of torchtext (0.16.0 is known to work well with PyTorch 2.1)
!pip install torchtext==0.16.0

#install other needed packages
!pip install -U -q scikit-learn matplotlib

# Create a dummy requirements.txt file and add torchtext to it, this way it will remain if installed.
with open("requirements.txt", "w") as f:
    f.write("torchtext==0.16.0\n")
    f.write("scikit-learn\n")
    f.write("matplotlib\n")

# installing libraries listed in requirements.txt
!pip install -U -q -r requirements.txt

#importing modules
from torchtext.vocab import GloVe
from sklearn.decomposition import PCA
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt

#modifying the url of glove data to be pulled from huggingface
GloVe.url['6B'] = 'https://huggingface.co/stanfordnlp/glove/resolve/main/glove.6B.zip'

%matplotlib inline

# Load the model. You can change dim to 50, 100, 300
glove = GloVe(name="6B", dim=50)

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

# 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")

simCompare("actor", "pen", "film")
simCompare("cat", "dog", "sea")

#generate word embeddings for the words "computer" and "human" using pre-trained GloVe embedding.
print(f"computer -> {glove['computer']}\n")
print(f"human -> {glove['human']}\n")

#Determine if "car" is closer to "truck" than "bike".
simCompare("car", "truck", "bike")
```

 [Show hidden output](#)

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

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

cat -> tensor([ 0.4528, -0.5011, -0.5371, -0.0157,  0.2219,  0.5460, -0.6730, -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.

```
##### CODE HERE #####
```

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

✓ Cosine Similarity

You learned about cosine similarity in class, now let's look at an example. Use the [cosine_similarity\(\)](#) 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.

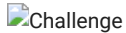
```
# define the similarity between two words
def similarity(w1, w2):
    return cosine_similarity([glove[w1].tolist()], [glove[w2].tolist()])

# 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")

simCompare("actor", "pen", "film")
simCompare("cat", "dog", "sea")
```

```
↔ 'actor' is closer to 'film' than 'pen'
'cat' is closer to 'dog' than 'sea'
```

Try it Yourself!



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

```
##### CODE HERE #####
```

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

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.