Machine Learning Engineer Course
Day 28

- Natural Language Processing -

DIVE INTO CODE

Thursday October 28, 2021

DIOP Mouhamed

Agenda

1. Check-in
2. Quick Review
3. Natural Language Processing
4. Sample code
5. To do by next class
6. Check-out

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What did you learn in the previous week?
(Anything is fine.)

Quick Review (ResNet and VGG)

Large-scale image recognition competition

- ResNet, VGG
- Transfer Learning
- The role of weights in transfer learning
- Learned weights

Introduction to Natural Language Processing

What is Natural Language Processing?

The languages that we humans use in our daily lives, such as Japanese, French, and Chinese, are called natural languages, and can be distinguished from programming languages, semiotics, and developed languages such as Esperanto (artificial languages).

Natural Language Processing (NLP) is a technology that allows computers to process such natural languages, and in Sprin21, we will consider the use of natural languages as input for machine learning.

What kind of machine learning model handles natural language?

Natural Language Processing

Machine learning models used in Natural Language Processing

Recurrent Neural Networks（RNN）

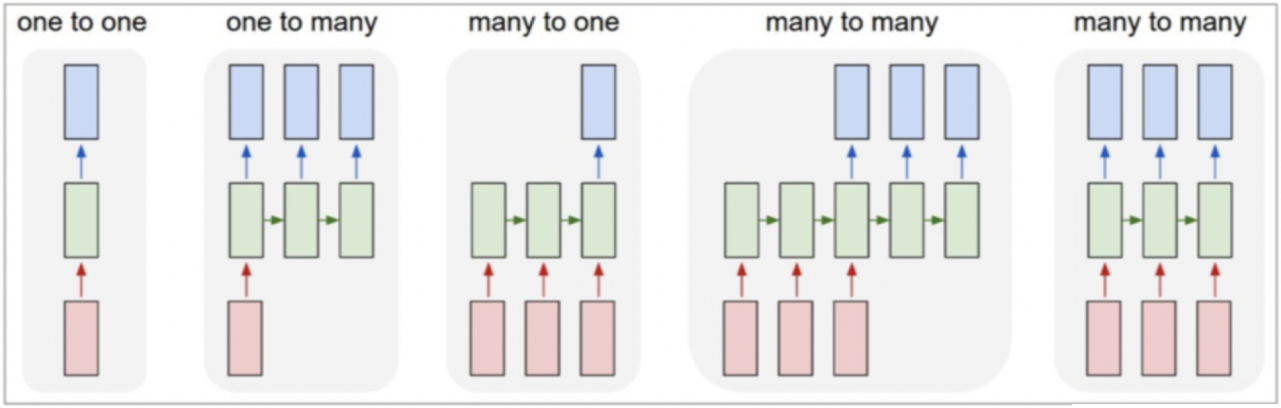
A type of deep learning model mainly used for sequence data analysis (time series data prediction). It is called a recursive neural network.

This RNN-based model has been applied to the following tasks.

Fields of Application:

Language modeling, neural machine translation, music generation, time series prediction, financial forecasting, etc.

The number of input and output vectors can be changed depending on the task.

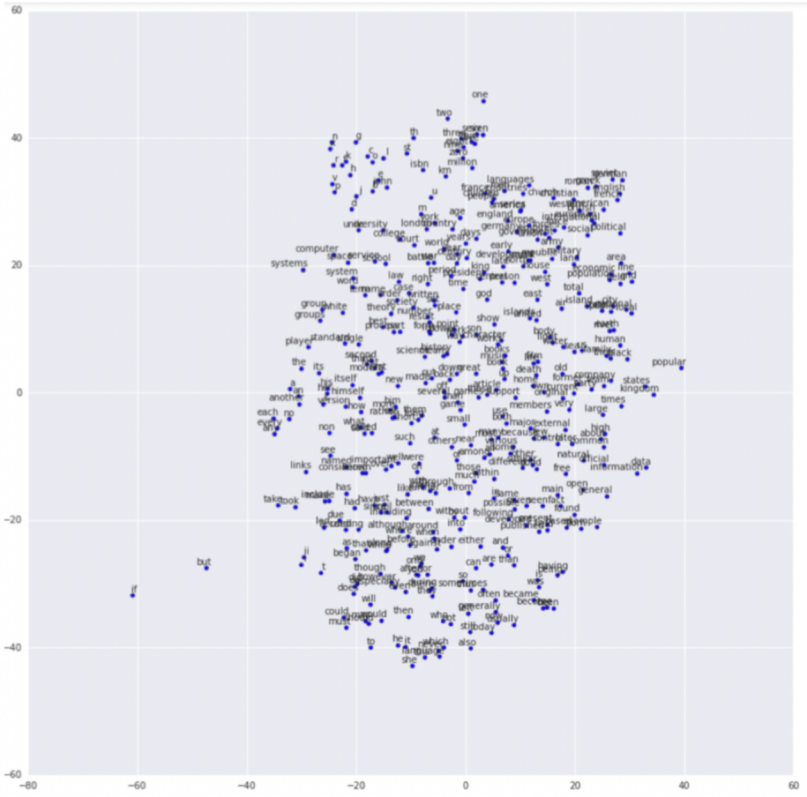


http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Natural Language Processing

What data should be used as input for natural language processing?

Vectorize string (symbolic) data to make it easier for machines to process. A vector representation, called Distributed Representation (or Word Embedding), is used for input data (sequential data) in language models.



What is distributed representation?

A vector that correlates "contextual similarity" with "semantic similarity".

Ideas:

"It is based on the Distributional Hypothesis [1], which states that "if two words are used in similar contexts (i.e., a county of words occurring around the word), then they are assumed to be similar.

[1] https://www.tandfonline.com/doi/abs/10.1080/00437956.1954.11659520

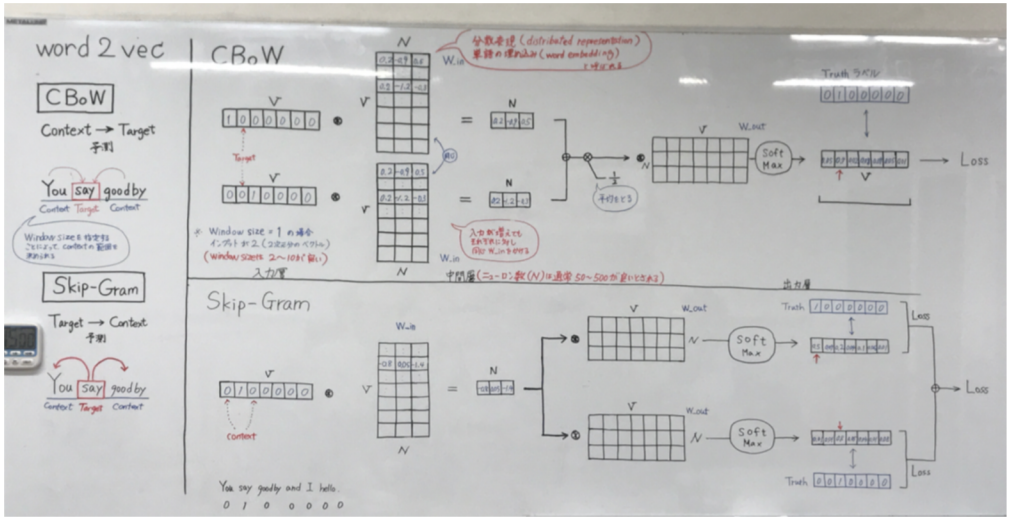
What kind of vector is it?

Example: Generated from Word2Vec

Distributed representations are generated in the training of neural networks (Word2Vec). Tens of thousands of words are replaced by a vector of distributed representations of arbitrary dimensions.

The distributed representation is embedded as the smallest unit in a vector space where words with similar contexts occupy close spatial positions (fixed length vectors of 50-500 dimensions).

The individual dimensions of a word vector usually have no inherent meaning. They represent the positions and distances between vectors by an overall pattern.



Local representation

Represent words as lexical indices

Distributed vs. local representations:

For distributed representations, we can vectorize them using a method called BOW (Bag of Words), which is a classical method (unlike Word2Vec, it is not a neural network). This is called a local representation.

In this method, words are encoded in a one-hot representation by considering them as IDs.

Such one-hot vectors can be viewed as words existing independently of each other in a multidimensional space where each word occupies one dimension and is independent of other dimensions (no mapping to other dimensions).

	a	bad	film	good	is	movie	this	very
0	0	0	0	1	1	1	1	1
1	1	0	1	1	1	0	1	0
2	0	2	0	0	0	0	0	3

BoW TF-IDF

Problems with One-Hot Vectors by BOW:

①Very common words appear in many sentences ("the," "and," "or," etc.).

②Sentences in which the same word is repeated are not suitable as input data.

These problems can be solved in TF - IDF (an advanced method of BoW).

In TF - IDF, the total number of words in a sentence (TF) is weighted by their frequency of occurrence in all sentences (df(t)). The higher the value of the frequency, the smaller the inverse document frequency (IDF), and the lower the value of the word.

	a	bad	film	good	is	movie	this	very
0	0	0	0	1	1	1	1	1
1	1	0	1	1	1	0	1	0
2	0	2	0	0	0	0	0	3

BoW

	a	bad	film	good	is	movie	this	very
0	0.00000	0.00000	0.00000	0.417796	0.417796	0.549351	0.417796	0.417796
1	0.51742	0.00000	0.51742	0.393511	0.393511	0.000000	0.393511	0.000000
2	0.00000	0.65918	0.00000	0.000000	0.000000	0.000000	0.000000	0.751985

TF-IDF

Problems with Distributed Representation

On the other hand, distributed representation also has its problems, such as the loss of language-specific syntactic structures.

Sample code

How to solve problems
“Natural_Language_Processing”

[Problem 1] Scratch implementation of BoW
[Problem 2] Calculation of TF-IDF
[Problem 3] Learning with TF-IDF
[Problem 4] Scratch implementation of TF-IDF
[Problem 5] Preprocessing the corpus
[Problem 6] Learning Word2Vec

Sprint 21 – NLP

Explanation about this Sprint is given but please try it on your own first.

Sprint 21 – NLP

Please work on your own after class and submit your assignments on DIVER.

Sprint 21 – NLP

A Sample Code of this Sprint is given but please try it on your own.

Sprint 21 – NLP

Please work on your own after class and submit your assignments on DIVER.

ToDo by next class

Next class will be Zoom : Thursday November 4th, 2021 19:30 ~ 20:30

🚨ToDo: Recurrent Neural Network
https://diver.diveintocode.jp/curriculums/1982

Check-out

3 minutes Please post the following point to Zoom chat.

Q. Current feelings and reflections
(joy, anger, sorrow, anticipation, nervousness, etc.)

Thank You For Your Attention