Project 1

Natural Language Processing

Project Description

Generate Word Embeddings and retrieve outputs of each layer with Keras based on the Classification task

Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation.

It is a distributed representation for the text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging natural language processing problems.

We will use the IMDb dataset to learn word embeddings as we train our dataset. This dataset contains 25,000 movie reviews from IMDB, labeled with a sentiment (positive or negative).

Dataset

The Dataset of 25,000 movie reviews from IMDB, labeled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes (integers). For convenience, the words are indexed by their frequency in the dataset, meaning the for that has index 1 is the most frequent word. Use the first 20 words from each review to speed up training, using a max vocab size of 10,000.

As a convention, "0" does not stand for a specific word, but instead is used to encode any unknown word.

Overview

1. Import test and train data
2. Import the labels (train and test)
3. Get the word index and then Create a key-value pair for word and word_id (12.5 points)
4. Build a Sequential Model using Keras for the Sentiment Classification task (10 points)
5. Report the Accuracy of the model (5 points)
6. Retrieve the output of each layer in Keras for a given single test sample from the trained model you built (2.5 points)

Project submissions and Evaluation Criteria

While we encourage peer collaboration and contribution, plagiarism, copying the code from other sources or peers will defeat the purpose of coming to this program. We expect the highest order of ethical behavior.

Submit the code on Olympus.

Submit the project milestone in a Jupiter notebook and submit it on Olympus for evaluation.

Project Support

You can clarify your queries by dropping a mail to Olympus

Happy Learning!