Document Retrieval

Divyakumar Prajapati

The problem

Problem Statement

Task was to build search engine type system(for documents related to specific field) In which by searching for a query or suggestion it will return some k relevant document.

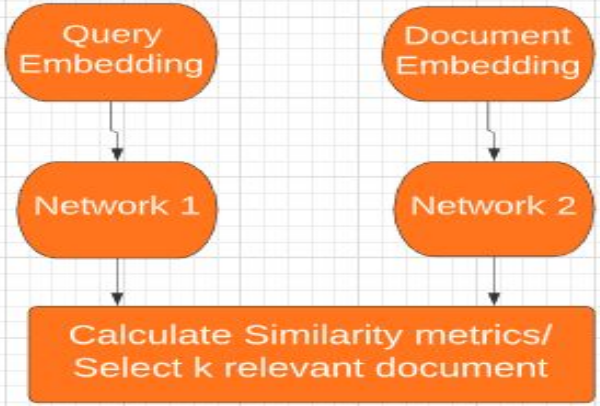
What is document retrieval ?

Document retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual.

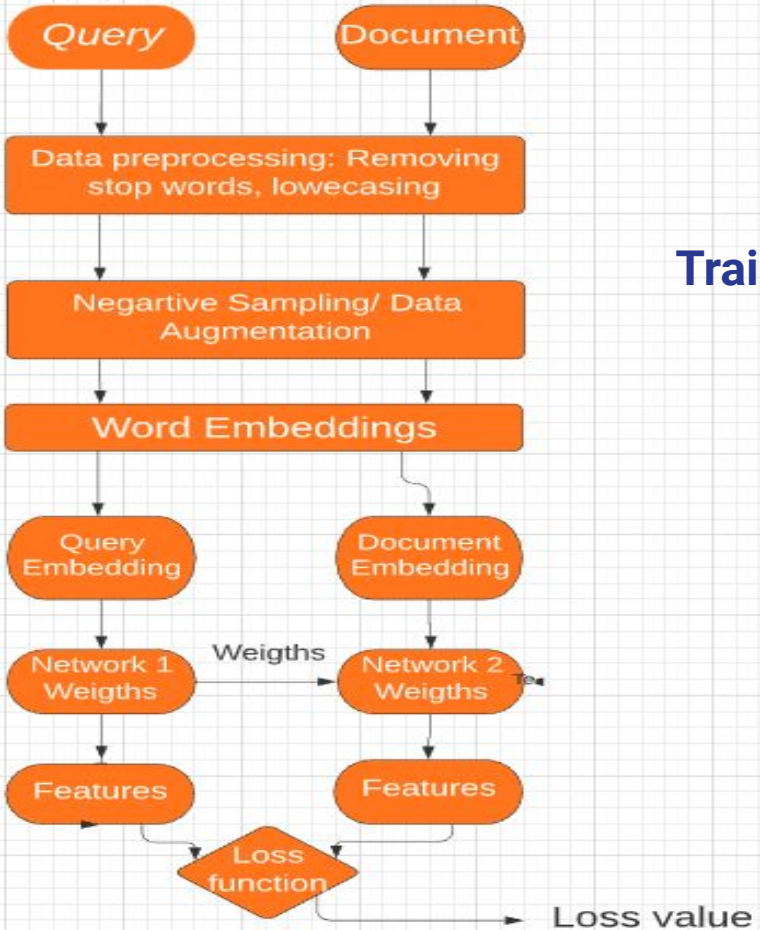
Table of Context

Architecture

Prediction



Training



Negative Sampling

Since data we got is like there is a question and document, so we have only positive samples for training our model. And model train on this data can lead bias. So, it was required to add negative samples as well to our data so that our contrastive loss model can learn to differentiate positive and negative samples. It may be easier for a model to distinguish among documents from different categories than those within the same category.
We use refined negative sampling for this.
In refined negative sampling, the model is presented with questions within the same category and subcategory (local sampling) as well as other categories (distance sampling) with a view to training a model capable of generating embeddings that are appropriately clustered not only across different categories, but also within each category.
We sample from answers that belong to the same category, from answers that belong to the same subcategory and from answers that do not belong to the category of the question. We choose these proportions heuristically with 50% from inside the category (in particular, 20% from the subcategory and 30% from the entire category) and the remaining 50% from outside the category.

Text Augmentation

More data we have, better performance we can achieve. However, it is very too luxury to annotate large amount of training data. Therefore, proper data augmentation is useful to boost up your model performance. And to teach our model to encode sentence with same context but with different structure.
We used Bert pre trained models for data augmentation.

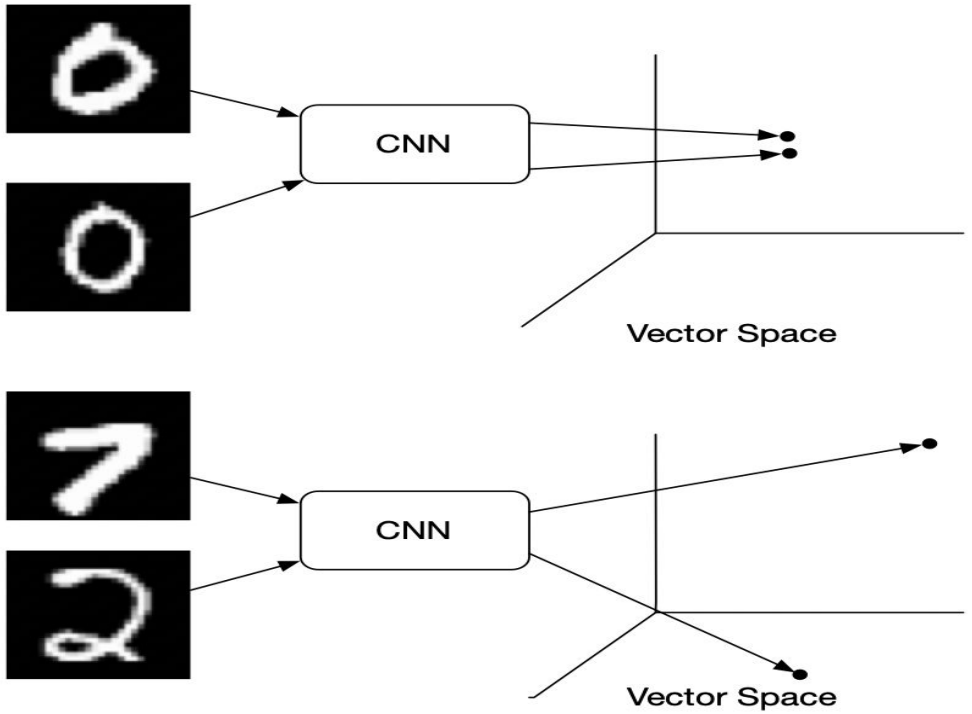
Feature Construction

It was very important to choose a encoder(for encoding a text which was going to pass to siamese network) which does not only take word occurring in account but also context.
For this we choose fastext to encode our data.
What is Fastext ?

fastText is a library for learning of word embeddings and text classification created by Facebook's AI Research (FAIR) lab. The model allows one to create an unsupervised learning or supervised learning algorithm for obtaining vector representations for words. Facebook makes available pretrained models for 294 languages. fastText uses a neural network for word embedding.

Siamese Network

When training a siamese network, 2 or more inputs are encoded and the output features are compared. This comparison can be done in a number of ways. Some of the comparisons are triplet loss, pseudo labeling with cross-entropy loss, and contrastive loss.
So what siamese does is shown in figure



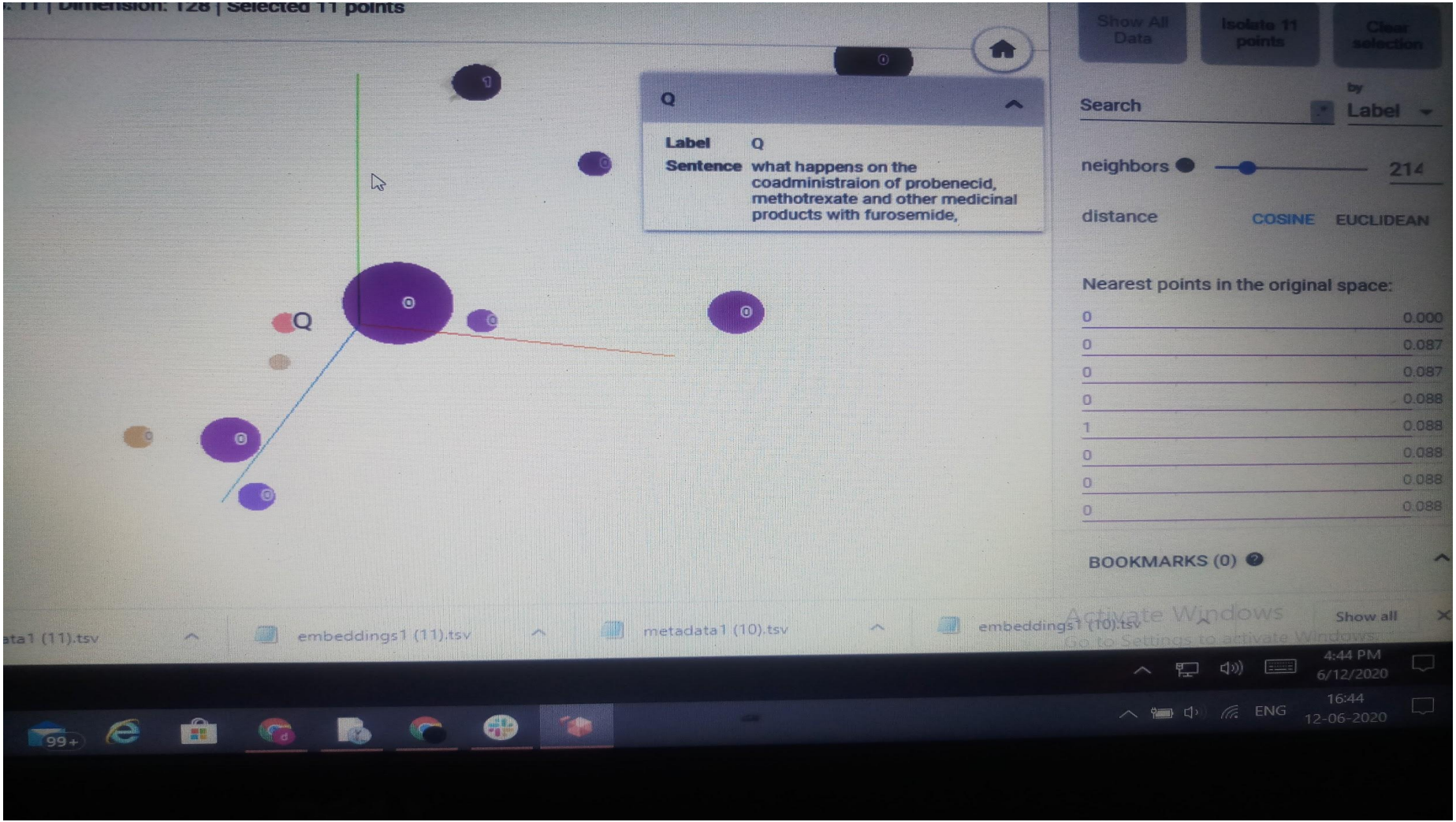
Why Contrastive loss instead of cross entropy loss?

Since we have the class labels for MNIST inputs we could use a regular network and Categorical Cross-Entropy loss. The problem is when we don't have nicely labeled data, as is usually the case. There is much more unlabeled data than labeled data available in the world. This is where contrastive loss comes in.

Contrastive loss takes the output of the network for a positive example and calculates its distance to an example of the same class and contrasts that with the distance to negative examples. Said another way, the loss is low if positive samples are encoded to similar (closer) representations and negative examples are encoded to different (farther) representations.

Evaluation

So to evaluate we took a question and if it returns a right document in k related documents than we give it 1 and otherwise 0 taking average of all tests can give us accuracy of our model.


Show All Data
Isolate 11 points
Clear selection
Search
by
Label
neighbors
214
distance
COSINE
EUCLIDEAN
Q
Label Q
Sentence what happens on the coadministraion of probenecid, methotrexate and other medicinal products with furosemide,
Nearest points in the original space:
0 0.000
0 0.087
0 0.087
0 0.088
1 0.088
0 0.088
0 0.088
0 0.088
BOOKMARKS (0)
embeddings1 (11).tsv
metadata1 (10).tsv
embeddings1 (10).tsv
Show all
Activate Windows
Go to Settings to activate Windows.
4:44 PM
6/12/2020
ENG
16:44
12-06-2020