Notes on Dense Representations by Nils Reimers

compiled by D. Gueorguiev, 6/7/2025

Dense Representation Definition

Formally,

dimensional representation (embedding)

Find such function f with which the *semantically similar* text is close with respect to the Euclidean metric

What does *semantically similar* mean and what it depends on?

Universal text representations do not exist. Text representations depend on the context and task.

Example 1: *Nuclear energy is safe* *Nuclear energy is dangerous*

Example 2:

S1: Heidi Klum is pregnant

S2: Heidi Klum is a model S3: Naomi Campbell is pregnant

Applications:

a ) Clustering:

A graph showing a diagram

Description automatically generated with medium confidence

The blue points represent text corpora on the same topic, green are another set of text corpora on the same topic but different than that of the blue ones, and finally the orange dots represent text corpora on a third topic.

b ) Bitext Mining

A graph of a bitxt mining

Description automatically generated

we take a large corpus – say the English Wikipedia, encode the corpora in vector space (shown on the Figure above as blue points) then take the text corpora from the German Wikipedia, encode the corpora in the same vector space (shown as the red points) and find blue-red dot pairs which are closest to each other. This is popular use case for obtaining training data in large Machine Translation models.

c ) Search

A diagram of a search bar

Description automatically generated

Encode the set of documents in vector space and a query. Find the closest document to the supplied query.

d ) Multi-Modal Search

A screenshot of a white dog

Description automatically generated

We can map images, videos, tables, program code, etc to vector spaces and thus match text corpora to videos, images, tables, or program code segments.

e ) Zero-Shot Image Classification

A couple of dogs in the snow

Description automatically generated

Having a set of labels encoded in vector space we would like to perform image classification. For a given image we compute its embedding in the same vector space in which the labels were encoded (depicted in blue) and we find the closest blue point to the orange dot which represents the image encoding in vector space.

f ) Few-Shot Intent Classification

A diagram of a financial system

Description automatically generated with medium confidence

When you build a chatbot you usually build it to serve different use cases , that is – different intents. It could be used at bank so when a customer is asking for account balance or the customer wants to transfer money. In this situation we define two intents – one is to serve customer account balance and the other is to transfer customer money. We create a set of example sentences for each of these two intents as visualized on the Figure above.

We match the latest input from the customer with the closest datapoint in vector space. No need to train or fine-tune the model but just prepare well-chosen examples

In Summary:

Have some examples for every intent (checking balance, transferring money)

New utterance => find closest example => use intent

# References

[1] [Introduction to Dense Text Representations - Part 1, Nils Reimers, Jun 21, 2021](https://youtu.be/qmN1fJ7Fdmo?si=UDxDVSawWsnH56qJ)

[2] [Introduction to Dense Text Representations - Part 2, Nils Reimers, Jun 21, 2021](https://youtu.be/0RV-q0--NLs?si=8cktLBFigHlNZzi-)

[3] [Introduction to Dense Text Representation - Part 3, Nils Reimers, Jun 21, 2021](https://youtu.be/t4Gf4LruVZ4?si=C2fjB45Vsye0t97p)