Exploration of ML Architectures, from Text Embeddings to Original Text

Path 3: The application path

Halvor

Introduction and Functionality of Text Embeddings

- Operates by representing text in a "semantic" high-dimensional vector space
- Text encoding into semantic vector spaces
- Organizes texts/words/sentences with similar themes close to each other in space
- Measured with cosine similarity to quantify semantic similarity
- Provides an understanding of contextual and thematic connections in texts
- Can be component in Language Model (LLM) assistants Retrieval Augmented Generation (RAG)
- And others

Exploration of ML Architectures

- *Objective: Predict original text based on vectors produced by the "decoder" from the text-embedding model*

- Initial use of simple ML architectures
- Later evaluation of more advanced architectures (Potential Project 2?)
- Inspired by the approach presented in https://arxiv.org/pdf/2310.06816.pdf , but with an approach accessible without the advanced complexity
- Explore the effectiveness of various models in predicting original text based on text-embedding vectors using a "decoder"

Unsupervised classification of text

- *Objective: Classify text using different unsupervised ML-methods and text embedding as input*

- Dataset containing longer texts, transformed using open-source text embedding models
- None or few labeled text (7%)
- Reduce dimensionality and cluster

Process

1. The dataset

2. Model Training:
   - Set up a simple NN
   - Train and evaluate

3. Take it from there ☺

Don’t hesitate to stop me if you find any of this interesting