UNIVERSITÄT KOBLENZ · LANDAU

Presentation
on

Paragraph Segmentation Using Semi-Supervised Deep Clustering Network

Under the guidance of Prof. Zeyd Boukhers

Presented by

Vidyasagar Aithal Radhakrishna
Soujanya Basangari
Toralben Arvindbhai Davara
Varuni Gururaja
Sai Manasa Tanniru

Agenda

3. Pre-Processing

Description about the data preparation before sending it to our model.

4. Methodolgy

Description about our model, training phase and evaluation of trained model.

2. Data Generation

A short description of methodology involved in generating a large data tp train our model.

5. Results and Evaluation

Discussion on the results achieved by our model on the different set of test data.

1. Introduction

A brief introduction about our application. Problem statement? How can it be tackled? Design of our model to solve the problem?

6. User Interface

A small demonstration of our website application.

Introduction

- PDF is one of the most popular and powerful electronic document formats
- Extracting information from the PDF document is a tricky job
- A huge number of applications are available online to extract the content. But the variety is so confusing and there is no clear winner
- Deep Learning techniques are booming nowadays and could be used to solve numerous problems
- Our application is developed on semi-supervised model combined with clustering technique
- End result is the clustered data, which will be used to get the segmented paragraphs in an oredered manner

Data Generation

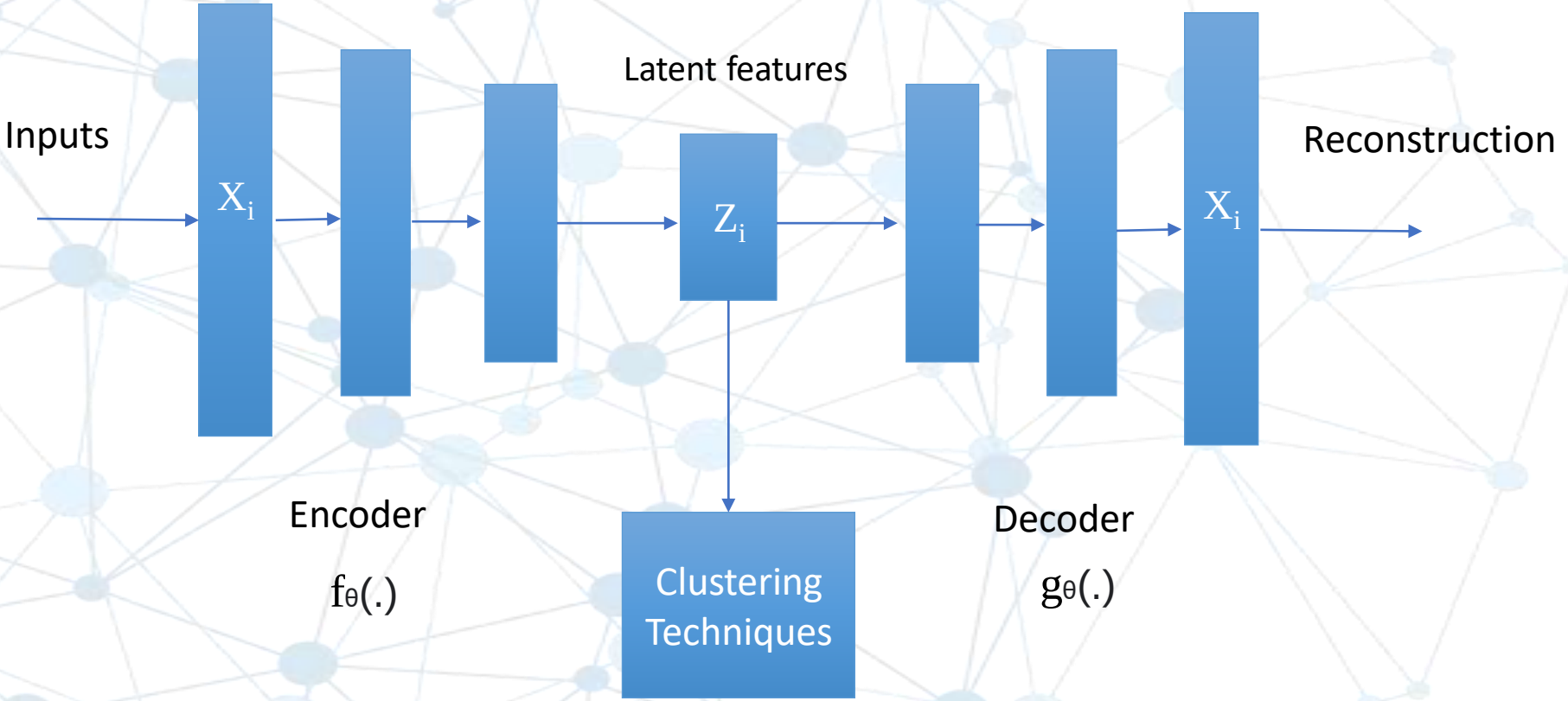
- Metadata records from ArXiv On Kaggle2 (1.7 million articles in JSON format)
- 1000 Pdfs with 200 paragraphs, headers, and footnotes each
- CSV files are generated parallelly, containing each lines of pdfs with respective labels
- Scraped webpages to create large set of pdfs (Unlabelled data)

  - Starting line of paragraph as ***1***
  - Rest lines of paragraph as ***2***
  - Header and footer as ***3***

Pre-Processing

- Feature extraction is done using PyMuPDF
- Along with the each line of text the layout information of the line are also derived using the PyMuPDF.
- From the extracted features, we also derive a few more features by the method of transormation.
- All together around 26 handcrafted features are extracted.
- A few of the examples are as follows: Ishorizontaltab, isenddot, isstartnumber, etc.

Methodology



Architecture of Clustering using autoencoder

Methodolgy

Training Process

Training process involves two phases in our application as follows:

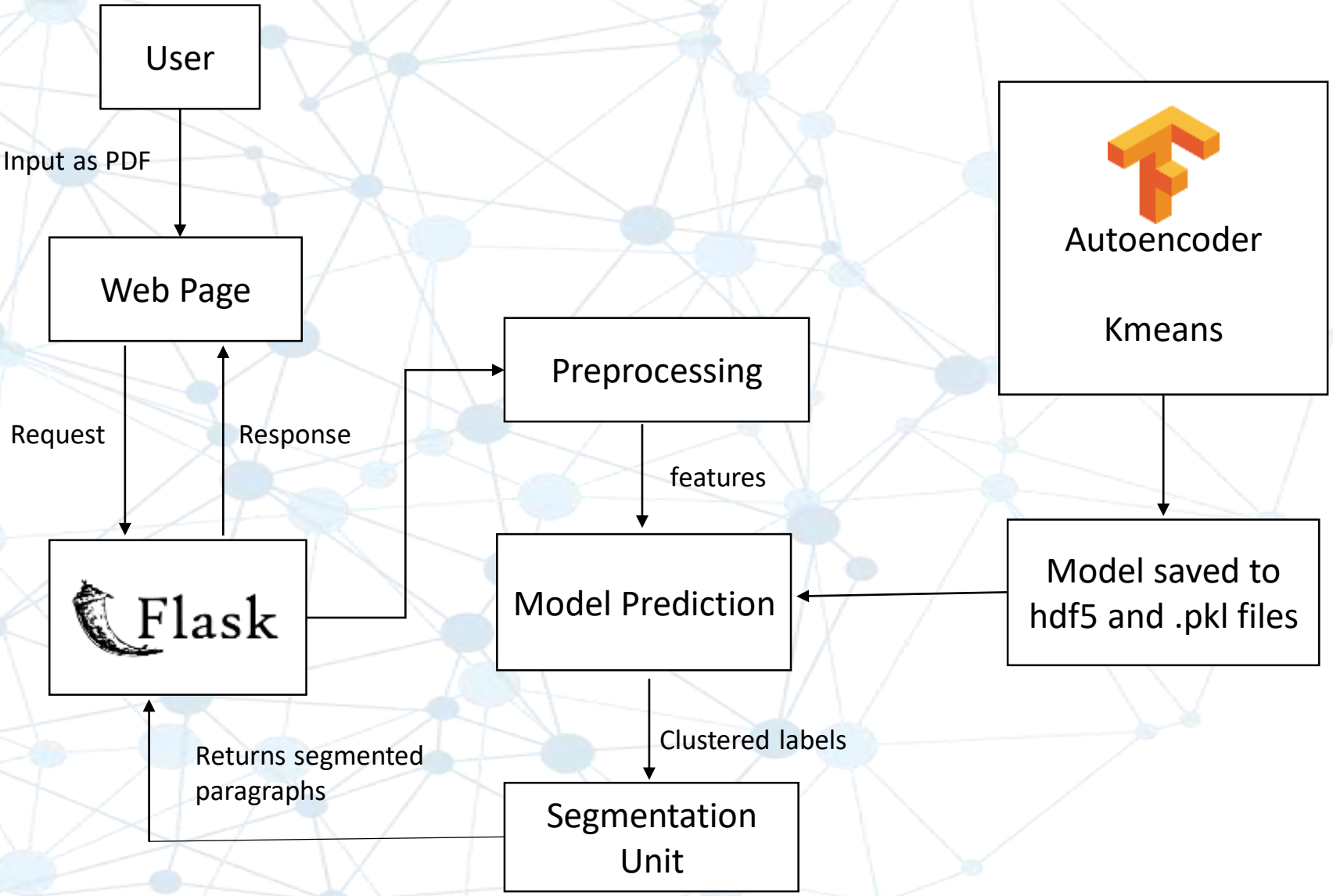
- **Phase 1 –** Pre-train model with labeled dataset
- **Phase 2**-Two step process
  - Step 1: Train model with both datasets without true labels(labeled and unlabeled)
  - Step 2: Train using only labeled dataset until the desired accuracy score is obtained

The accuracy achieved on our trained model is 0.99

Results & Evaluation

	Accuracy	Precision	Recall	F-score
Using Kmeans				
Set1	0.84	0.84	0.839	0.84
Set2	0.85	**0.85**	0.84	0.84
Set3	0.81	0.81	0.81	0.81
Using Gaussian mixture model				
Set1	0.83	0.835	0.839	0.835
Set2	0.84	**0.84**	0.84	0.83
Set3	0.78	0.78	0.78	0.79

User Interface



Thank You

For your Attention