Advanced Deep Learning
Winter 2023/24
Vincent Christlein, Katharina Breininger

Sheet 4, starting from Feb 22th, 2023, due March 26th, 2023, 16:00

Topic 4: Self-Supervised Learning Challenge

This competition investigates the performance of large-scale retrieval of historical document fragments based on writer recognition. The analysis of historic fragments is a difficult challenge commonly solved by trained humanists. **Your task is to implement a suitable automatic framework able to relieve your colleagues from the humanities from this tedious task.**

For this challenge, you can work in teams with up to three people!

To simulate fragments, we extracted random text patches from historical document images. The goal is to find similar patches of the same page or manuscript. The document images are provided by several institutions and different genres (manuscripts, letters, charters). Examples can be found in Fig. 1.

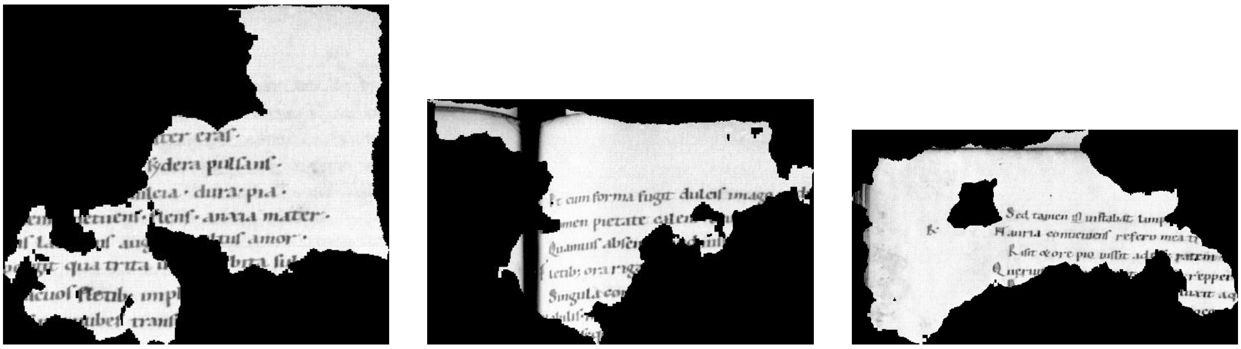

Figure 1: Example fragments.

1. Dataset

The full dataset is available below: `https://doi.org/10.5281/zenodo.3893807` and will be provided on the HPC cluster in `/home/janus/iwb6-datasets/FRAGMENTS`

It contains a training and a test set with the following image naming-convention: WID_PID_FID.jpg , where WID=writer id, PID: page id, FID= fragment id.

- Train set: contains ≈ 100 000 fragments using the Historical-IR19 as base dataset. They should all contain some text, however some fragments are quite small.
- Test set: contains about 20 000 additional fragments

For more information please see [1]. The dataset was created by Mathias Seuret. The generation code is publicly available below: https://github.com/seuretm/diamond-square-fragmentation

Important: For our run of the challenge, we will provide a **new** small private test set for deciding upon the ranking. You can of course use the current test set to compare your results with the participants of the original challenge [1]. We recommend to use the test dataset as an independent validation/test set during your model development and to not include it in your training set (of course you can add it and fine-tune/retrain your model for the final challenge testing). Furthermore, avoid “overoptimisation” on this test set.

2. Task

The task consists of finding all fragments corresponding to (a) the same page (b) the same writer using a document fragment as query. This means, your task is to create vector embeddings for all test fragments.

3. Evaluation

The evaluation will be done using a leave-one-image-out cross-validation approach. This means that every image of the test set will be used as query for which the other test images are going to be ranked according to their similarity with the query. The competition will be evaluated in two ways using mean average precision (mAP):

- On a writer-level, i. e., finding fragments of the same writer.
- On a page-level, i. e., finding fragments of the same page.

The final ranking will be the average rank between both tasks.

We will use the provided `eval_map.py` for the mAP computation and evaluation of the competition. You can also use it to evaluate your results, just call it with your $T \times T$ distance matrix, which can, for example, be computed by means of cosine distances:

```
dists = 1.0 - encs.dot(encs.T)
```

where `encs` would be the $T \times D$ matrix of your embeddings (each row corresponds to one embedding of a fragment and D denotes its dimensionality), s. `main.py` for an example.

4. Submission

One week before the deadline, we will upload the test dataset for which you have to compute the embeddings and the respective distance matrix. Submission steps:

1. Please compute the distance matrix, save it as `dm.pkl.gz` (s. `main.py`) and upload it to FAUBox, gdrive, or similar and send us the link.

2. Upload your code to StudOn.

5. Bonus points

You have to improve upon the baseline to be eligible for the bonus points.

Have fun and let us know if you encounter any issues!

References

[1] Mathias Seuret, Anguelos Nicolaou, Dominique Stutzmann, Andreas Maier, and Vincent Christlein. Icfhr 2020 competition on image retrieval for historical handwritten fragments. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 216–221, Sep. 2020.