Advanced Deep Learning
Winter 2024/25
Vincent Christlein, Soroosh Tayebi Arasteh

Sheet 4, starting from Feb 12th, 2024, due April 2nd, 2024, 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This competition investigates the performance of large-scale retrieval of historical document fragments based on writer recognition.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To simulate fragments, we extracted random text patches from historical document images.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 writer of a 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 at `https://doi.org/10.5281/zenodo.3893807` and will soon also be provided on the HPC cluster below `/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 $\approx$ 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 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 using mean average precision (mAP) @ rank 50:

- Only on a writer-level, i. e., finding fragments of the same writer.

You can use the provided `eval_map.py` for the mAP computation (or use directly the kaggle leaderboard) 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.

Note that the code computes the full mean average precision of the full square distance matrix. In contrast, in kaggle, we will only use the first 50 entries for computing the mean average precision, s. kaggle for more details.

4. Submission

You have to submit a Nx51 matrix, where the first column corresponds to the query filenames and then the 50 closest filenames to it, s. kaggle for more details.

1. Submit via kaggle.
2. Upload your code to StudOn – please ensure that you mention your team name in your code header.

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.