Self-Supervised Deep Metric Learning for ancient papyrus fragments retrieval

Summary

Papyrologists struggle to study the manuscripts of ancient literature to understand how human societies were organized, their cultural, administrative, societal and economical aspects of lives that time. The ancient papyrus documents found are usually fragments which may be different in size, color, shape, texture, and style of writing. The same looking documents may be fragments of different documents because of the conservation conditions and vice versa and it makes it very difficult to associate any new found fragment with any previously found fragment of the same document. The authors of this study have proposed a solution using self-supervised Deep Metric Learning to reconstruct the papyri fragments prior to their analysis by papyrologists.

The authors have experimented on two datasets; HisFrag databse, which comes from the 2020 ICFHR HisFragIR20 competition, it contains 17222 fragments as training set and 2732 as test set. It does not require any pre-processing as labels are provided as file names. The other dataset which is also big contribution by the authors is pre-processing of Michigan database. Michigan database contains 17029 images of different language documents. Out of 14890 Greek Papyri, authors choose 6190 color images after removing 515 negative and infrared images, and 8185 duplicate images. After further processing, the dataset is left with 1118 papyri composed of 4579 fragments, each composed of atleast two fragments.

Siamese Neural Networks are used to learn similarity score between two fragments. Absolute element-wise difference of the two output vectors is calculated – the embeddings - and to feed this difference into two dense layers of size 512 that are connected to a sigmoid neuron as the final output. This allows to optimize using the Binary Cross Entropy loss function. Patch base approach is used to train the model. Balance mini-batches are built during training to avoid over sampling or under sampling a single class. First, maximum possible non-overlapping patches of size 64x64 pixels are extracted of each image then patches are ranked based on their content; text or background. Patch score is calculated by dividing total number of pixels by number of pixels that constitute text plus the subtraction from 1 of total number of pixel divided by the number of pixels that constitutes background. At the end, 5 patches representing each fragment with highest score are chosen, a higher number would increase a lot of computations.

The models are evaluated on un-trained dataset. Similarity score is computed by taking average of the scores of each pair of patches between all the possible fragments. From these scores a 2D similarity matrix is built and also ground truth matrix, telling if the fragments belong to same papyrus. From these metrics, all the metrics used in Hisfrag competition for evaluation are calculated, ie Mean Average Precision, Top-1 Accuracy, Precision at 10 and 100 (pr@10 and pr@100). With these metrics both models are performing differently, VGG16 seems to work best with Hisfrag dataset and Resnet50 is with Michigan dataset. The best is 87% top-1 accuracy on VGG16 with Hisfrag dataset. A model trained on another dataset is being tested focused on domain-adaption to check the results. The results are better with fine tuning with 50 papyri and then 1000 papyri but still these are not the results which are expected and far from the baseline results.

Self-supervised way is introduced by the authors over VGG16 and Resnet50. Previously, in baseline, the real information was assumed available at the papyrus level for training.

A model based on self-supervised learning is introduced where no training data is required. A pretext task to determine if two patches are from the same fragment and target task to determine if two fragments are from the same papyrus. In pretext task, patches are labeled with the identifiers of individual fragments as ground truth is not available in this case, and this can also mislabel a small portion. Probability of mislabeled pairs is 0.005 which is calculated on 10000 papyri in Hisfrag dataset which makes total 295000 patches. As the probability is very low, it is assumed that it should not affect the training performance much. When self-supervised approach is compared to domain adaptation, the results are worse when training with the papyrus level information but it works very well with fine-tuned papyrus level information. It is empirically proved that self-supervised approach combined with pre-trained model on the Hisfrag dataset is much more useful rather than annotating 1000 papyri. Given that, sample images are quite degraded still self-supervised model gives better results.

The authors have empirically proved Self-supervised Deep Metric Learning approach is far better than other approaches in this scenario with top-1 accuracy of 0.73 on Hisfrag dataset. The proposed model is also experimented on in a use case where only unlabeled data is available to the papyrologists. The results also suggests the model will be useful to reconstruct new found fragments.

The authors have discussed that in future thorough experiments will be conducted on the patch selection process as it will have huge impact on the model.