UNIVERSITY OF TARTU
1632

Self-Supervised Cell Discovery in Histopathology Images Using Vision Transformers

Project team: **Denys Kaliuzhnyi, Denys Krupovych, Pavel Chizhov**
Supervisor: **Mikhail Papkov**

Contacts:
denyskaliuzhnyi@gmail.com
dkrupovich99@gmail.com
chizhovpd@gmail.com

Institute of Computer Science

Repository:
github.com/DenysKaliuzhnyi/dino

Abstract

This poster presents the results of a project, which was done as a part of the Machine Translation course at the University of Tartu, Institute of Computer Science. The aim of the project was to apply a self-supervised method for image feature extraction to testis histopathology images dataset from East-Tallinn Central Hospital.

Introduction

Data labeling and annotation is extremely laborous and expensive for histopathology images: the amounts of data are large and new data is constantly added. Because of this, self-supervised methods application is of paramount importance in this field. A modern approach to image feature extraction through self-supervised pretraining called DINO [1] was recently applied to histopathology images from TCGA-BRCA breast cancer dataset [2].

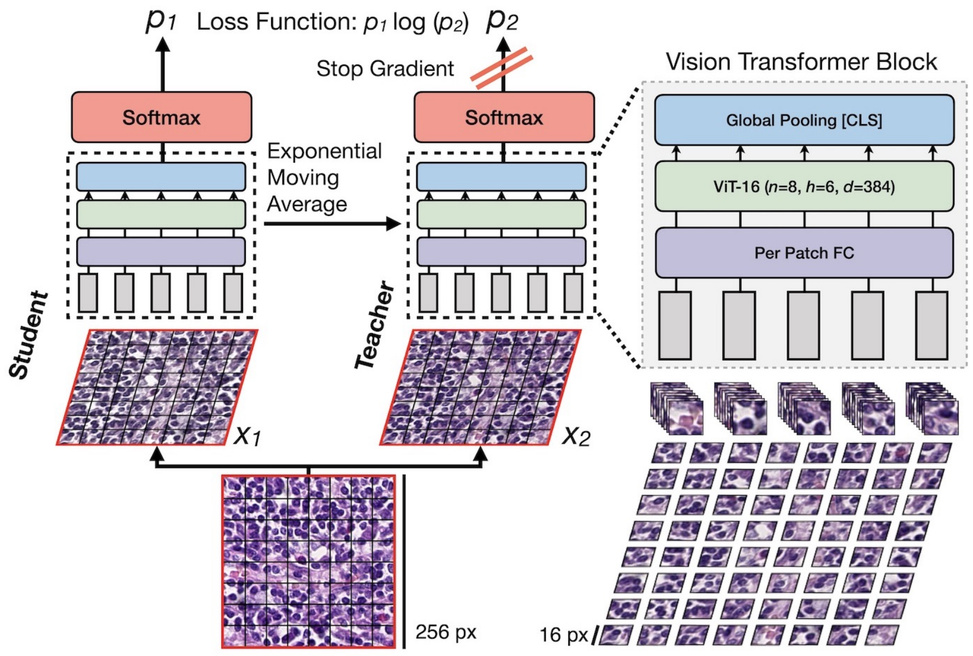

Figure 1 - DINO Architecture

The approach is to pass the augmented version of the image to the student and teacher networks in order to measure the similarity of the output features using cross-entropy loss. DINO uses a Vision Transformer (ViT) which splits the image into non-overlapping 16x16 patches and applies multi-head self-attention layer for feature extraction between the smaller patches [1].

Data

The data was presented as a collection of whole slide images (WSIs) of tissue fragments. A special cropping procedure was implemented to generate crops of size 256x256 pixels from WSIs.
A small set of images annotated with 5 cell types (Figure 2) was selected as a validation set.

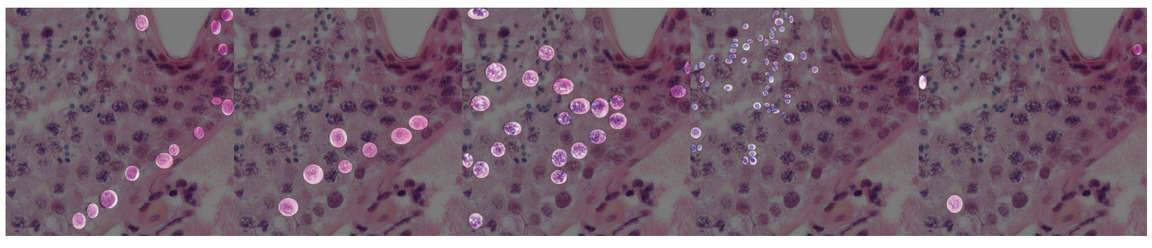

Figure 2 - Cell types: Spermatogonia, Sertoli, Primary spermatocyte, Spermatid, Garbage cell

Training

We decided to select the best checkpoint of the model from the histopathology article [2] as the pretrained model and fine-tune it on the generated dataset of 400000 crops.
Since DINO is a complex architecture with quite a lot of hyperparameters and the training process is very time-consuming, a considerable amount of time was spent on the hyperparameters search.
Figure 3 presents the difference in the training progress of the model with default hyperparameters, and the model with the best set of hyperparameters achieved in the project.

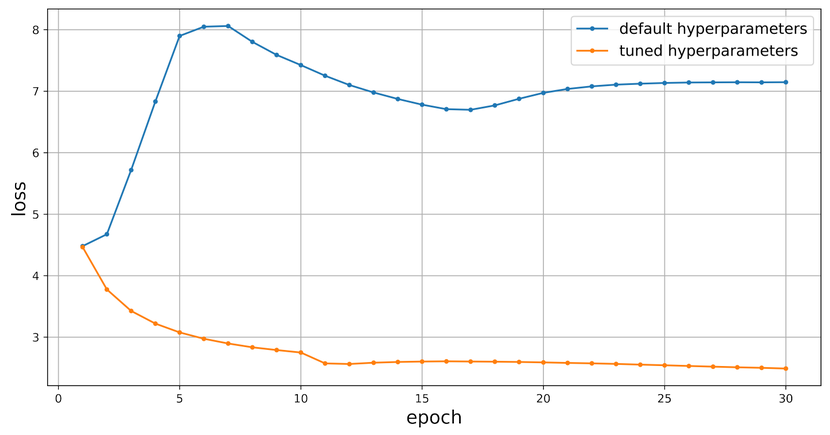

Figure 3 - Training loss history

The winning set of hyperparameters is mostly described by these changes:

- disabled gradient clipping
- increased teacher momentum
- 10 epochs with frozen last layer
- 0 warmup teacher temp epochs

In Figure 4 there are examples of images and 6 attention maps extracted from the best model's ViT.

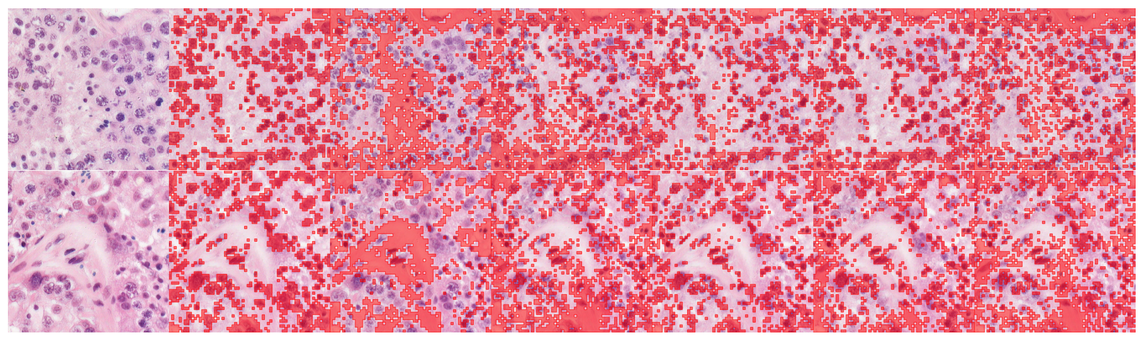

Figure 4 - Original image and attention maps 0-5

Evaluation

In order to see the correspondence between attention heads and cell types, we evaluated recall from two binary images - attention map and annotation mask. The recall metric was chosen because not all the cells on the images are annotated.

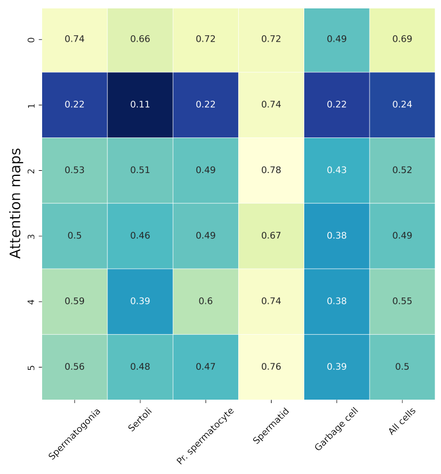

Figure 5 - Recall heatmap

In Figure 5 there is a heatmap depicting recall values for each attention head and each cell type. We can conclude, for example, that head 1 avoided cells and selected empty space instead, head 0 selected bigger cells (first three types). Every head has a good recall on Spermatid cells, this may be because they are small. Yet heads 4-5 seem to be focused on these exact cells. The same inference can be made from the visualizations in Figure 4.

Conclusion

After the thorough choice of hyperparameters the fine-tuned DINO model showed promising results, proving applicability of such a model to the task of cell detection in histopathology images.

References

[1] Mathilde Caron et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2021.

[2] Richard J Chen and Rahul G Krishnan. "Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology". In: *Learning Meaningful Representations of Life, NeurIPS 2021* (2021).