image R meeting

20 November 2024

- Ilaria Billato (Ph.D. Student), University of Padova
  - ilaria.billato@phd.unipd.it https://github.com/billila

TCGA images

- 11765 **diagnostic** images
- 33 tumor types
- saved in .svs format

~ 25 days to download all the images

~ 11 TB

TCIA images

- **diagnostic** images and also **radiomics**
- saved in different format .tiff, .png, .svs)
- TCIA Histopathology Custom Dataset Builder

~ ?

TCGA & TCIA

TCIA and TCGA provides different set of H&E images.

imageTCGA shiny app

imageTCGA

`imageTCGA` is an R package designed to provide an interactive Shiny application for exploring the TCGA Diagnostic Image Database. This application allows users to filter and visualize clinical data, geographic distribution, and other relevant statistics related to TCGA diagnostic images.

https://github.com/billila/imageTCGA

imageTCGA R package

Data Loading Options:

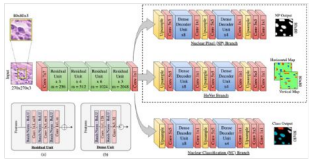
1. Matrix format
2. Spatial data format
3. Zarr format
4. S4 objects
5. Bumpy matrix
6. anndata

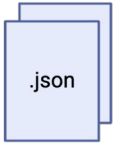
Functions:

1. Download TCGA Images
2. Extract Features (HoverNet, Prov-GigaPath)

HoVer-Net

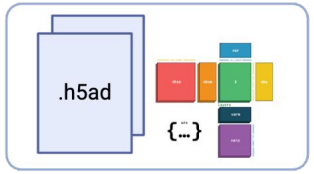
11765 H&E images

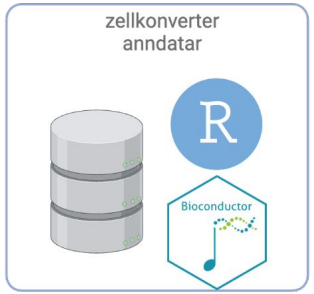




Feature Extraction

1. Input data - H&E TCGA
2. Hover-Net
3. Output Hover-Net .json
4. Feature Extraction
5. Outupt .h5ad
6. Import .h5ad in R and create a R package repository





how obtain json file

- run run_infer.py

```
python run_infer.py \
--gpu='0,1' \
--nr_types=6 \
--type_info_path=type_info.json \
--batch_size=64 \
--model_mode=fast \
--model_path=hovernet_fast_pannuke_type_tf2pytorch.tar \
--nr_inference_workers=6 \
--nr_post_proc_workers=6 \
wsi \
--input_dir=/home/exouser/hover_net \
--output_dir=/home/exouser/hover_net/output \
--save_thumb \
--save_mask
```

- ~ 2 days to analyze one image on my computer (still running)
- GPU on Anvil https://github.com/vqdang/hover_net/issues/286
- HoVer-Next ?

how obtain json file

HoVer-NeXt Inference

HoVer-NeXt is a fast and efficient nuclei segmentation and classification pipeline.

Supported are a variety of data formats, including all OpenSlide supported datatypes, `.npy` numpy array dumps, and common image formats such as JPEG and PNG. If you are having trouble with using this repository, please create an issue and we will be happy to help!

paper: https://openreview.net/pdf?id=3vmB43oqIO

json file -> Prostate Cancer by Mohamed

```
data                                    Large list (2 elements,  187 MB)
    $ mag: int 40
    $ nuc:List of 117935
     ..$ 1     :List of 5
     .. ..$ bbox     : int [1:2, 1:2] 15892 15912 64768 64804
     .. ..$ centroid : num [1:2] 15633 65054
     .. ..$ contour  : int [1:41, 1:2] 15622 15621 15617 15616 15616 15619 15620 15624 15625 15628 ...
     .. ..$ type_prob: num 1
     .. ..$ type     : int 3
     ..$ 2     :List of 5
     .. ..$ bbox     : int [1:2, 1:2] 15898 15925 64617 64644
     .. ..$ centroid : num [1:2] 15480 65062
     .. ..$ contour  : int [1:44, 1:2] 15467 15466 15466 15465 15466 15466 15471 15471 15473 15473 ...
     .. ..$ type_prob: num 1
     .. ..$ type     : int 3
     ..$ 3     :List of 5
     .. ..$ bbox     : int [1:2, 1:2] 15917 15927 64744 64774
     .. ..$ centroid : num [1:2] 15608 65073
     .. ..$ contour  : int [1:26, 1:2] 15592 15592 15593 15594 15594 15596 15597 15598 15603 15604 ...
     .. ..$ type_prob: num 1
     .. ..$ type     : int 3
```

json file

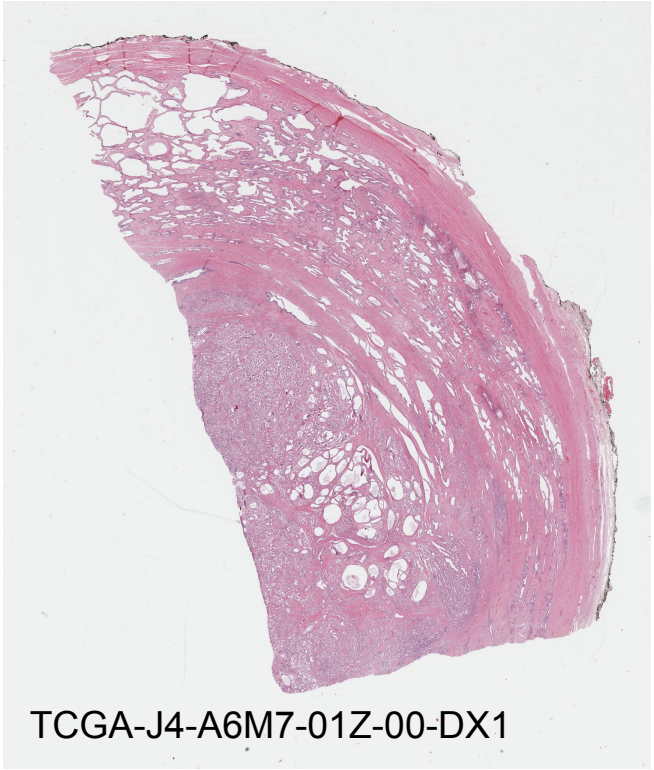
JSON structure explanation

- **mag**: This typically indicates the magnification level of the slide. In this case, it's set to 40.
- **nuc**: This is a dictionary containing the detected nuclei. Each key (like 1, 2, etc.) represents a unique nucleus detected in the image.

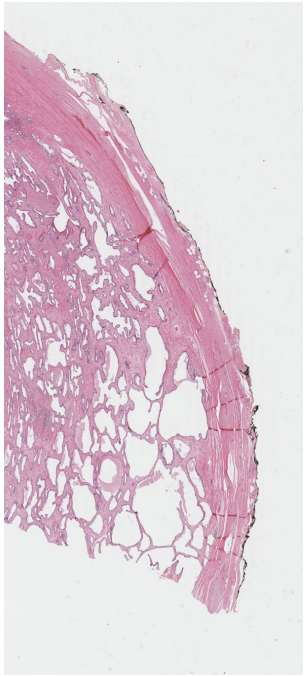
Each nucleus entry contains:

- **bbox**: The bounding box coordinates of the nucleus, represented as two points (top-left and bottom-right).
- **centroid**: The coordinates of the centroid of the nucleus, which is the center point calculated based on the contour.
- **contour**: A list of points that outline the shape of the nucleus. This is useful for visualizing the exact shape and boundaries of the detected nucleus.
- **type_prob**: The probability associated with the type of the nucleus, indicating the model's confidence in its classification.
- **type**: This typically represents the class label assigned to the nucleus (e.g., different types of cells or states).

json nuclear polygon by Lucio

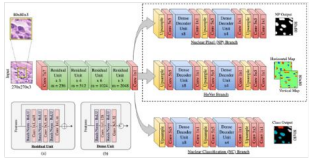


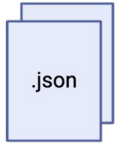
TCGA-J4-A6M7-01Z-00-DX1

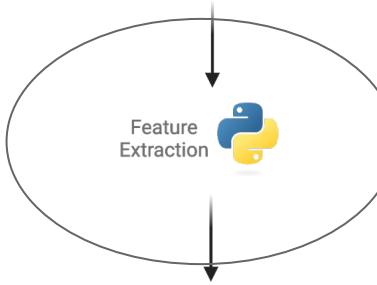


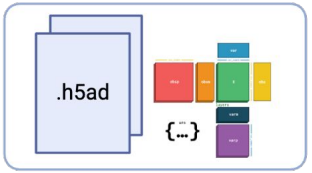
HoVer-Net

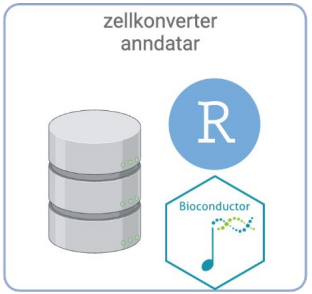
11765 H&E images











1. Input data - H&E TCGA
2. Hover-Net
3. Output Hover-Net .json
4. Feature Extraction
5. Outupt .h5ad
6. Import .h5ad in R and create a R package repository

feature extraction

python code from Mohamed
h5ad example from Lucio

.h5ad object

```
> library(anndata)
> anndata <-
anndata::read_h5ad("/home/ilaria/Docume
nts/cuny/h5ad/pca_adata_20x.h5ad")
> anndata
AnnData object with n_obs × n_vars =
55370851 × 1
    obs: 'type', 'slide_id'
```

```
> rhdf5::h5ls(h5ad_path)
           group       name       otype  dclass           dim
0              /          X H5I_DATASET   FLOAT 1 x 55370851
1              /     layers   H5I_GROUP
2              /        obs   H5I_GROUP
3           /obs     _index H5I_DATASET  STRING     55370851
4           /obs   slide_id   H5I_GROUP
5  /obs/slide_id categories H5I_DATASET  STRING          864
6  /obs/slide_id      codes H5I_DATASET INTEGER     55370851
7           /obs       type   H5I_GROUP
8      /obs/type categories H5I_DATASET  STRING            6
9      /obs/type      codes H5I_DATASET INTEGER     55370851
10             /       obsm   H5I_GROUP
11             /       obsp   H5I_GROUP
12             /        uns   H5I_GROUP
13             /        var   H5I_GROUP
14          /var     _index H5I_DATASET  STRING            1
15             /       varm   H5I_GROUP
16             /       varp   H5I_GROUP
```

Prov-GigaPath

Article | Open access | Published: 22 May 2024

A whole-slide foundation model for digital pathology from real-world data

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, … Hoifung Poon + Show authors

Nature **630**, 181–188 (2024) | Cite this article

- **1,384,860,229** 256 × 256 image **tiles** in **171,189** H&E-stained and immunohistochemistry pathology **slides**
- from biopsies and resections of 31 major tissue types in over **30,000 patient**