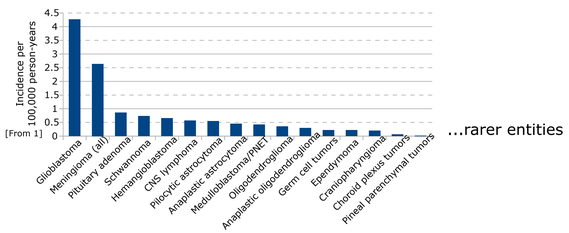
Creating Synthetic Glioma and Brain Tissue Histology

Marcello DiStasio[1,2] and David Meredith[1,3]

Departments of (1) Pathology, Brigham and Women's Hospital, (2) Pathology, Boston Children's Hospital, and (3) Oncologic Pathology, Dana Farber Cancer Institute, Boston MA 02115

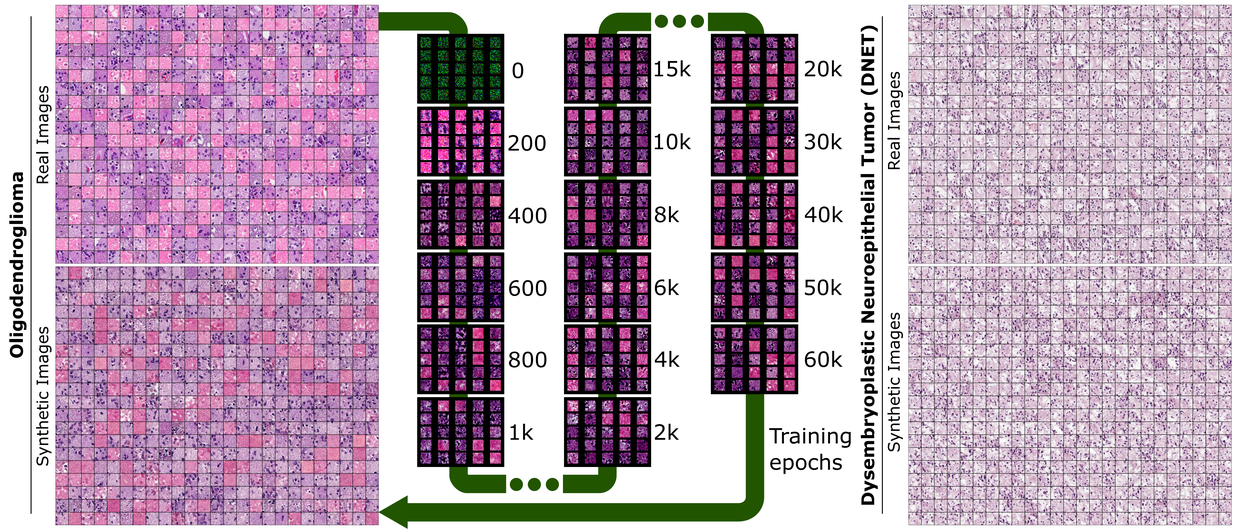
Introduction

The promise of digital image analysis for use in neuropathology is subject to constraints of automation and robustness of feature identification. Many advanced machine-learning algorithms have been applied to digitized slide images for purposes of segmentation and feature extraction, but these all rely on the availability of a large volume of high-quality training data to achieve high accuracy. Surgical neuropathology involves a wide variety of histologic appearances, including multiple morphologic variants of common entities (e.g. glioblastoma), and many entities which are encountered at low frequency (e.g. dysembryoplastic neuroepithelial tumor).



Training any robust machine-learning based classifier on raw neuropathology histologic images will suffer from the classic problem of unbalanced training data, in which a few histologic patterns are highly represented, and other histologic patterns are poorly represented. This biases training, creating algorithms with high sensitivity for the common patterns, and low sensitivity for rare patterns.

Results



Methods

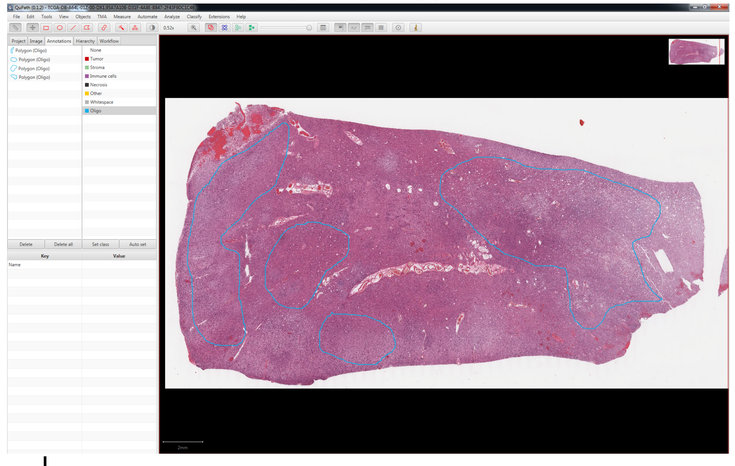
We present here a method for 'bootstrapping' both rare and common histologic features to arbitrarily high representation in a training data set by training a group of algorithms to generate novel, synthetic, realistic histologic images based on limited input (e.g. a single H&E stained slide). These algorithms, known as generative adversarial networks (GANs), once trained, can produce a vast number of novel images that share features with the training images. By training a GAN on a limited number (e.g. 1000) of image patches drawn from a single or few slide(s), a huge number of novel images can be generated (>1e419 64x64 pixel images; larger than the number of protons in the known universe).

Training Data Source

Whole Slide Images:
- »TCGA data portal
- »Slides scanned during clinical workflow

Known histologic diagnoses:
- •Glioblastoma, IDH-wildtype, W.H.O. grade IV
- •Oligodendroglioma, IDH-mutant and 1p/19q-codeleted, W.H.O. grade II
- •Dysembryoplastic Neuroepithelial Tumor, W.H.O. grade I

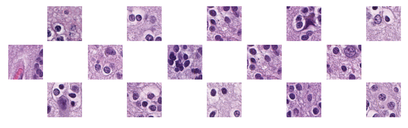
▾Manual selection of large tumor regions in QuPath [5]

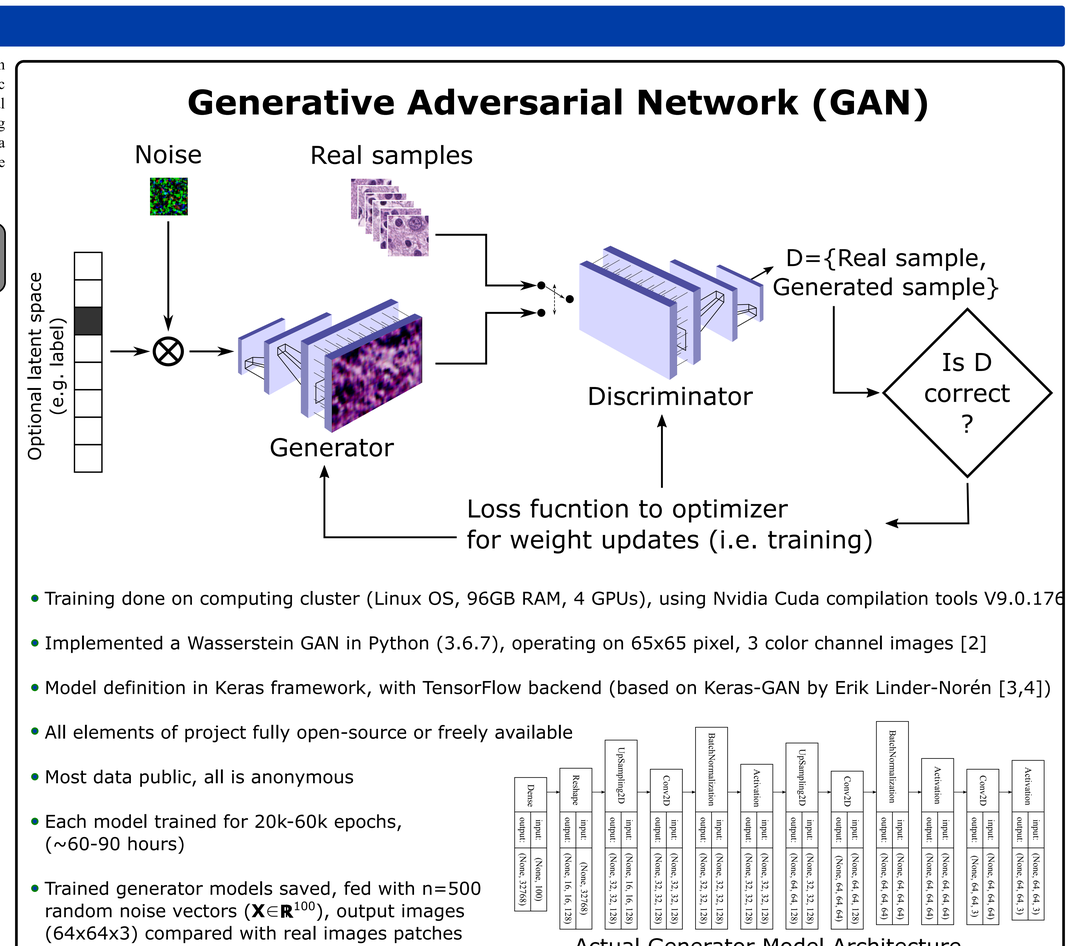


Custom script (Apache Groovy)

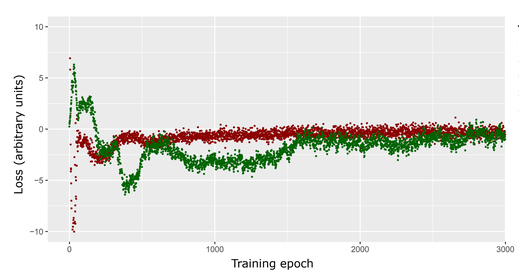
Exports labeled tile images from annotated regions in QuPath.

65x65 pixel patches (n = ~150,000 per whole slide image)





- Training done on computing cluster (Linux OS, 96GB RAM, 4 GPUs), using Nvidia Cuda compilation tools V9.0.176
- Implemented a Wasserstein GAN in Python (3.6.7), operating on 65x65 pixel, 3 color channel images [2]
- Model definition in Keras framework, with TensorFlow backend (based on Keras-GAN by Erik Linder-Norén [3,4])
- All elements of project fully open-source or freely available
- Most data public, all is anonymous
- Each model trained for 20k-60k epochs, (~60-90 hours)
- Trained generator models saved, fed with n=500 random noise vectors ($\mathbf{X} \in \mathbb{R}^{100}$), output images (64x64x3) compared with real images patches



◂Optimizer objective (loss) function results for training of the generator and discriminator on the oligodendroglioma training data set. Both networks achieve stability after about 1500 training epochs (~2-3 hours on 4 GPUs).

Code for exporting tiles and training GANs, as well as fully trained models are available at http://chelly.us/lab

Conclusions

- Using this method, we have created a set of models that generate synthetic image patches based on single slides of normal brain tissue, glioblastoma, oligodendroglioma (grade II), and other rarer brain tumor types.
- Synthetic images recapitulate morphologic and architectural features of particular tumors:
  - Neuropil texture
  - Nuclear contours
  - Chromatin density and texture
  - Cellular density
  - Spatial relationships between cells / nuclei
- Models are an anonymized, compressed representation of an individual tumor, or of a tumor type.
- These models can be easily deployed by machine learning researchers to enhance the training and testing of new algorithms for application to routine histologic images of neuropathologic entities.

References

[1] de Robles P, Feist KM, Frolkis AD, Pringsheim T, Atta C, St. Germaine-Smith C, Day L, Lam D, Jette N. The wordlwide incidence and prevalence of primary brain tumors: a systematic review and meta-analysis. Neuro-Oncology, 17(6), June 2015: 776-783

[2] Arjovsky M, Chintala S, Bottou L. 2017. Wasserstein GAN. arXiv:1701.07875

[3] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. 2017. Improved Training of Wasserstein GANs. arXiv:1704.00028; https://github.com/keras-team/keras-contrib/blob/master/examples/improved_wgan.py

[4] Keras-GAN. Erik Linder-Norén. https://github.com/eriklindernoren/Keras-GAN

[5] Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. Scientific Reports (2017). https://doi.org/10.1038/s41598-017-17204-5