Generating Histopathologic Images With Variational Autoencoders

Lynn Zheng
ruolinzheng@uchicago.edu

Presentation Outline

1. Motivation & Project Objective
2. Background Research & Reference
3. Dataset, Methodology, Technologies
4. Results
5. Challenges & Lessons Learned

Motivation & Project Objective

- Data imbalance in histopathologic image datasets
  - Kaggle Histopathologic Cancer Image Dataset: 100k tumor out of 220k training images
- Train generative models for data augmentation
  - Class-conditional GANs and VAEs
- Quantitatively evaluate the quality of generated data
  - Unlike MNIST or faces, we cannot say that generated data "look realistic"

Background Research

Generative models applied to histopathologic images

- Tschuchnig, M. E., Oostingh, G. J., & Gadermayr, M. (2020). Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. arXiv preprint arXiv:2004.14936.
- Raza, K., & Singh, N. K. A tour of unsupervised deep learning for medical image analysis. arXiv 2018. arXiv preprint arXiv:1812.07715.

Background Research

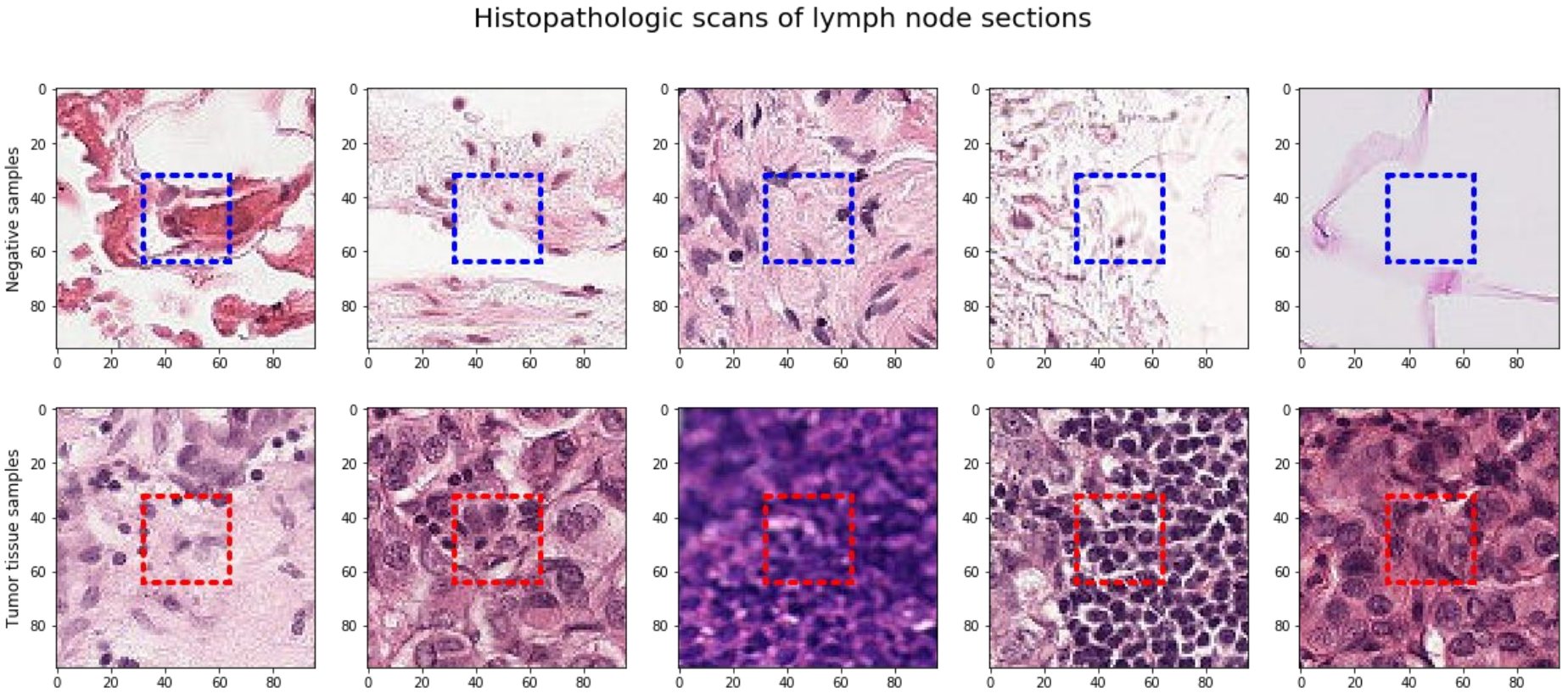
GANs and VAEs in more general domains

- Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. In Advances in Neural Information Processing Systems (pp. 14866-14876).
  - Quantitatively evaluate the performance of generative models, Vector-Quantized VAEs, GANs

Dataset

Kaggle Histopathologic Cancer Detection Dataset

- https://www.kaggle.com/c/histopathologic-cancer-detection/data
- Identify metastatic tissue in histopathologic scans of lymph node sections
- 220k, 96x96, RGB images
- "A positive label indicates that the center 32x32px region of a patch contains at least one pixel of tumor tissue"



source: https://www.kaggle.com/qitvision/a-complete-ml-pipeline-fast-ai

Dataset

160k training and 60k testing

- For training generators, use all 160k training images
- Generate 160k images using the generator
- For training classifiers, hold out 25% of the 160k images as validation
- Select the model checkpoint with the highest validation AUC as the best

Methodology

- Train generative models on histopathologic images
- Evaluate the quality of generated data by:
  - Training a binary (normal vs. tumor) classifier on generated data
  - Testing the classifier on real test data
  - If the generated data is similar to real data, the classifier should perform well
  - Baseline classifier trained on real training data
- Metric: AUC instead of accuracy since the dataset is unbalanced

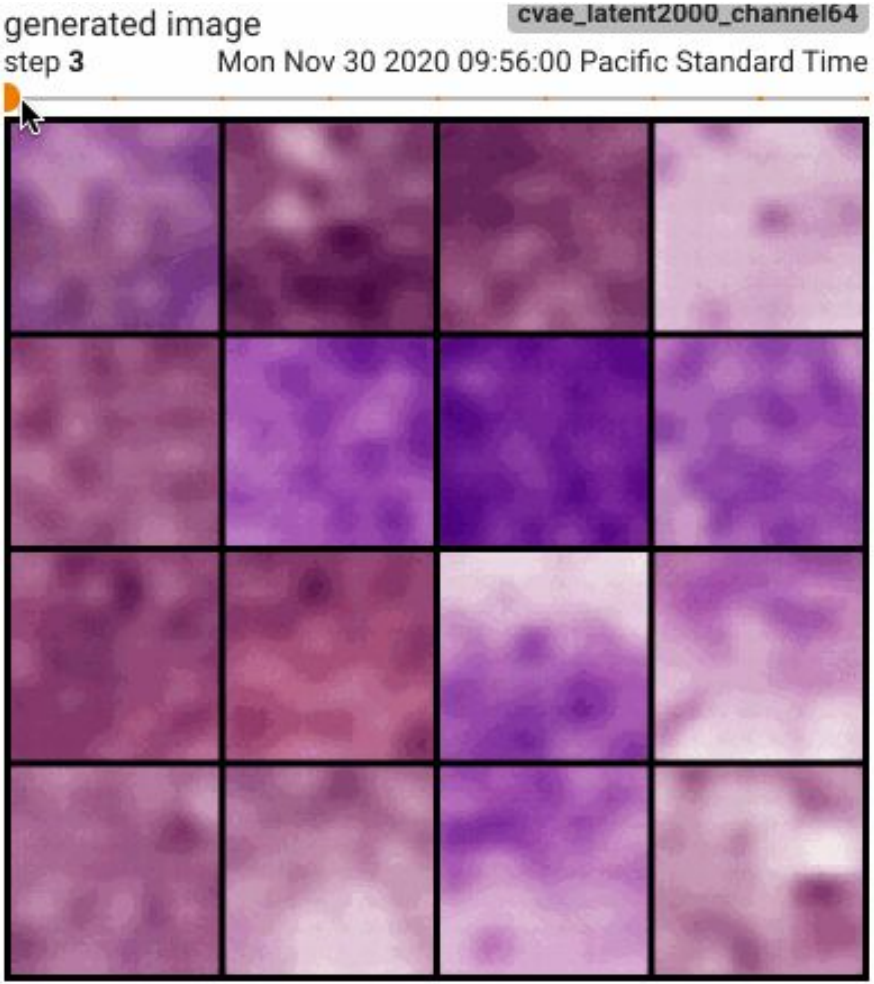
Technologies

- Implemented a vanilla convolutional classifier used on both real data and generated data
- Implemented class-conditional VAEs and GANs in PyTorch
- Models
  - VAE:
    - different latent dimensions, 100, 500, 2000
    - 4 convolution layers
    - number of channels: 16, 64
  - GAN: binary cross-entropy loss, Wasserstein loss
- Training time: Classifiers: 6 GPU hours; Generators: 2 - 6 GPU hours

Results

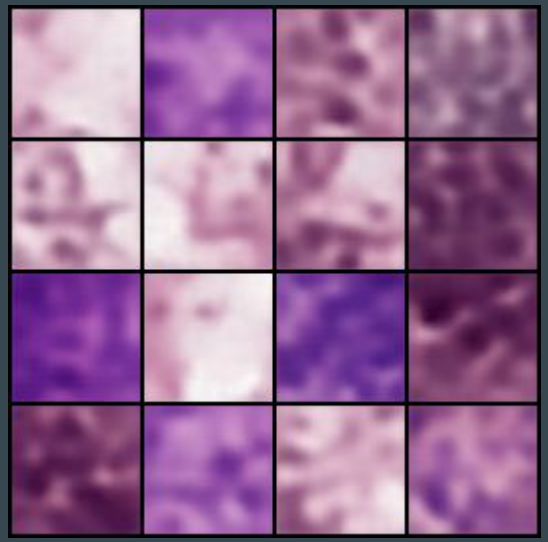
Images generated by a VAE with 2000 latent dimensions and 64 convolutional channels.

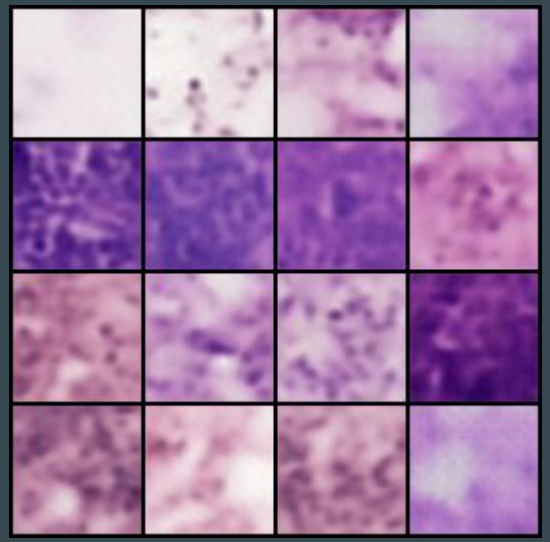
The top two rows are negative/normal samples and the bottom two rows are positive/tumor samples

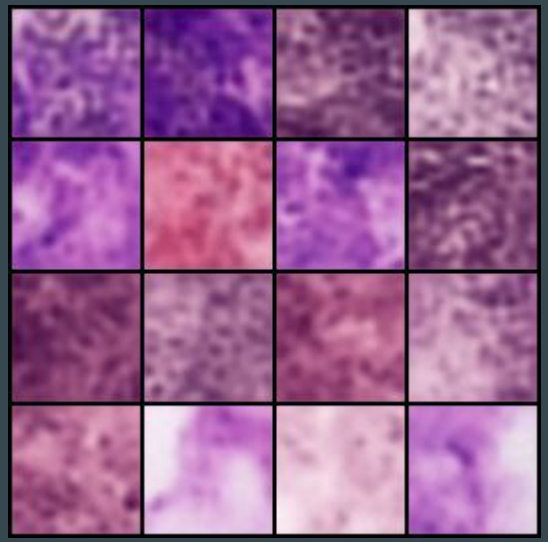


Results

VAE-generated images grow less blurry with a larger latent dimension and more convolutional channels; Left to right: 100, 500, 2000 latent dimensions

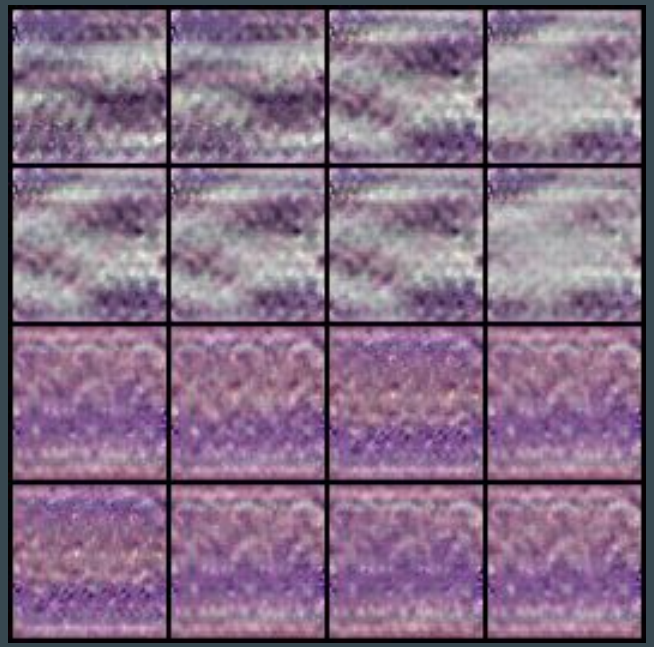


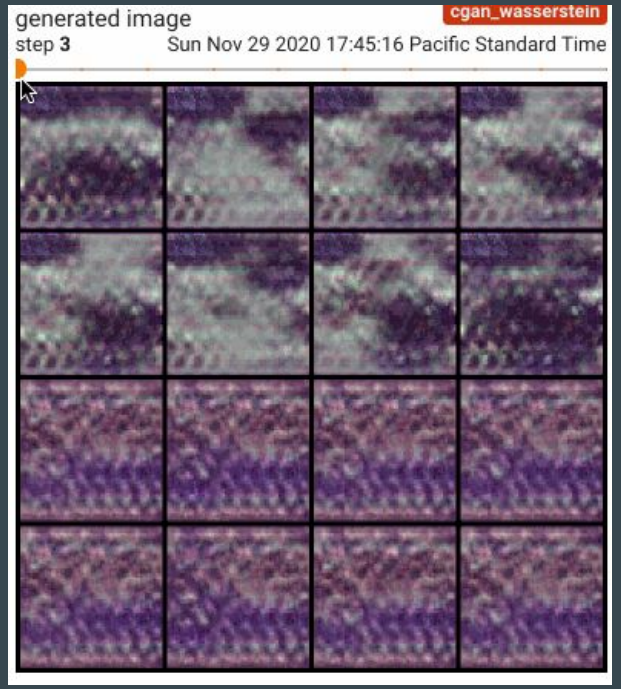




Results

GANs suffered from mode collapse despite various attempts





Results

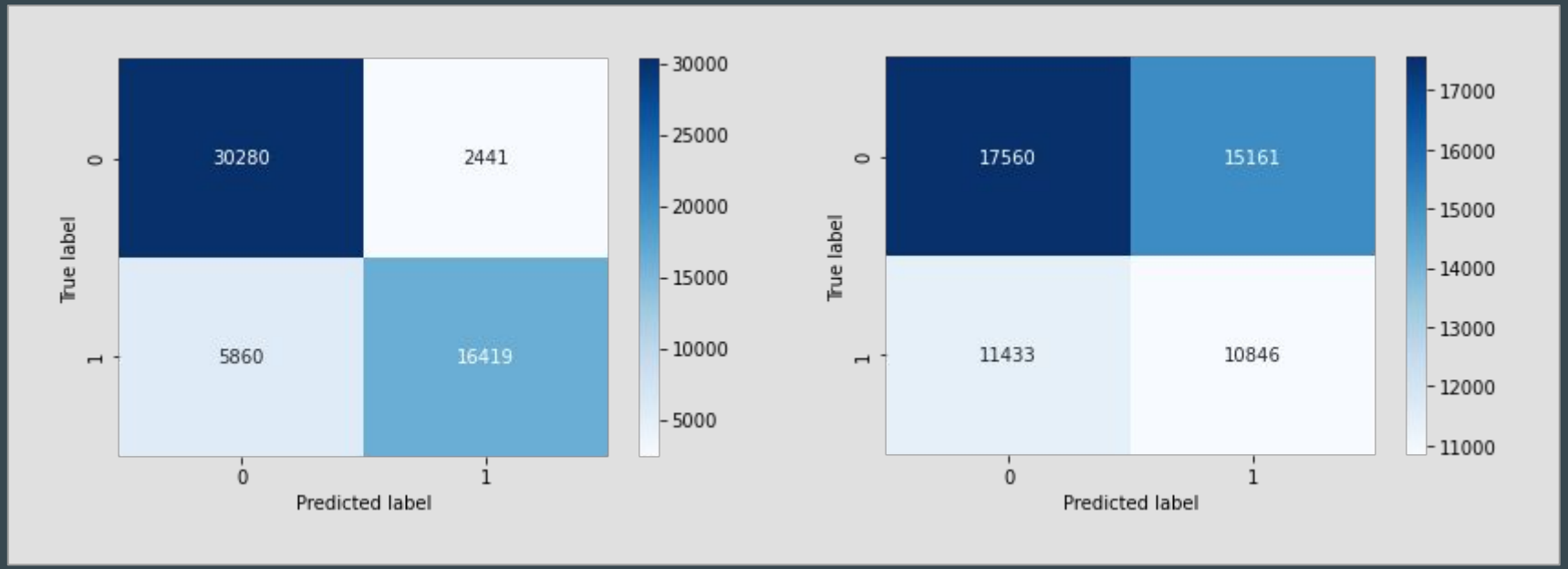
Test loss, accuracy, and AUC of classifiers trained on real vs. generated data

Model	Latent dimensions	Channels	Test loss	Test accuracy	Test AUC
Baseline	-	-	0.3576	0.8491	0.9225
VAE	100	16	40.67	0.5531	0.4328
	500	16	26.81	0.5496	0.4814
	2000	16	**4.317**	0.5165	**0.5065**
	2000	64	9.875	0.5608	0.4940

Table 1: Classifer test accuracy and AUC

Results

Confusion matrices using a 0.5 cutoff: baseline (left), "best" model 2000-latent-dim (right) seems to have erred a lot on false positives



Challenges

- Mode collapse in GANs
  - Modify the training scheduling of the generator and the discriminator
  - Train discriminator more often than generator
  - Pre-train the discriminator
- Low-resolution images from the VAEs
  - This is expected as VAEs learn the "mean" of the data distribution
  - Impossible to achieve pixel-level precisions, hence generated images are too blurry to be useful as histopathologic images
  - Recall Kaggle's dataset description: "at least one pixel of tumor tissue"

Challenges

Computing resource limitation

- Time: insufficient to fine-tune hyperparameters in the model architecture (latent dimensions, etc.) or for training (learning rate, optimizers, etc.)
- Space: challenging to train large models
  - GPU ran out of memory for a 4000-latent-dimension VAE with 256 convolutional channels; roughly a 7 GB model

Lessons Learned & Conclusion

- Generative models that perform well on other datasets might fail to translate into a different domain like histopathologic images
  - I trained VAEs and GANs with the same architecture on MNIST and Celeba, and achieved visually-sound results
  - Unlike digits or human faces that are highly structured, histopathologic images, especially pixels of tumor tissues, are highly variable
- Possible Improvements
  - Architectural improvement for VAEs (VQ-VAE) and GANs
  - Stabilize GAN convergence

Q & A