CITY
UNIVERSITY OF LONDON
EST 1894

Academic excellence for
business and the professions

Unsupervised Anomaly Detection
CitAI Seminar 17/03/2021

Sergio Naval-Marimont[1], Giacomo Tarroni[1,2]

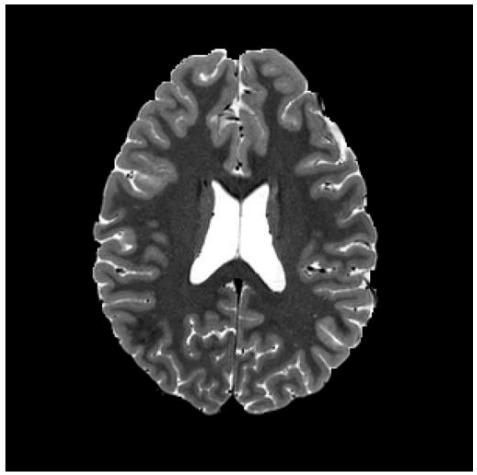
[1]MSc Data Science, City, University of London
[2]BioMedIA, Department of Computing, Imperial College London

Today's session…

- Introduction to unsupervised anomaly detection
- Deep unsupervised models
- Anomaly detection methods
- Some results

Introduction: Anomaly Detection in Medical Images

- Multiple deep learning methods proposed to localize anomalies in medical images.
- Fully supervised methods achieve high accuracies…
- …however:
    1. Rely on large and diverse annotated datasets for training.
    2. Specific to the anomalies annotated.

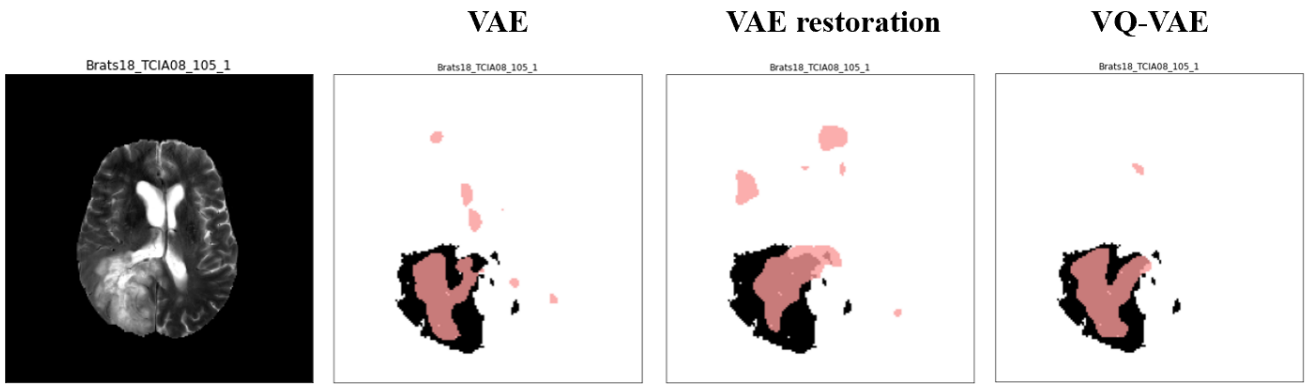


Normal or
Abnormal?

Unsupervised Anomaly Detection

- **Unsupervised Anomaly detection** does not require annotated datasets.
- Multiple approaches, generically based on generative models and two steps:
  1. Model the distribution of normal samples p(**X**): VAEs, GANs, AR…
  2. Method to identify / localize anomalies by comparing a test sample with the learnt distribution of normal samples
- Datasets are expensive to create and in principle, unsupervised methods are not limited to the annotated anomalies. Also, in terms of data available we can expect:

Healthy images without labels >> images with annotated anomalies



Models: Variational Auto-Encoders (VAE)

Objective: estimate p(***X***). Assumption, generative model:

$$\boldsymbol{z} \rightarrow \boldsymbol{X}$$

- We observe ***X*** (e.g. CT Scan, MRI,...), **z** is unobserved latent (low dimensional manifold where observed images sit). In VAEs ***z*** assumed to ***z*** ~ N(0,I)
- $p(\boldsymbol{X}) = \int P(\boldsymbol{X} \mid \boldsymbol{z})\, p(\boldsymbol{z})\, dz$. If ***z*** is continuous and $P(\boldsymbol{X} \mid \boldsymbol{z})$ is complicated likelihood (e.g. neural networks), this becomes easily intractable. Large datasets also makes sampling solutions (i.e. Montecarlo) not viable.
- In VAE we introduce a recognition model $Q(\boldsymbol{z} \mid \boldsymbol{X})$ [generates (μ, σ) for ***z***~N(μ, σ)] and derive the evidence lower bound (ELBO):

$$\log P(\boldsymbol{X}) - KL[Q(\boldsymbol{z}|\boldsymbol{X}), P(\boldsymbol{z}|\boldsymbol{X})] = E_{z\sim Q(\boldsymbol{z})}[\log P(\boldsymbol{X}|\boldsymbol{z})] - KL[Q(\boldsymbol{z}|\boldsymbol{X}), P(\boldsymbol{z})]$$

Evidence Lower Bound

Reconstruction given z

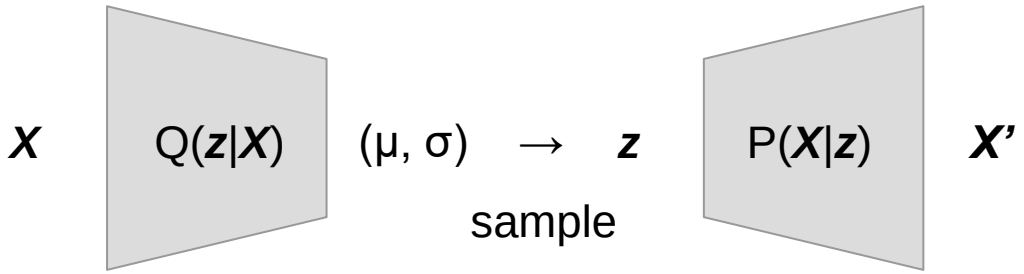
KL between recognition and prior for z

Auto-Encoding Variational Bayes Diederik P Kingma, Max Welling https://arxiv.org/abs/1312.6114

Models: Variational Auto-Encoders (VAE)

- **Training procedure**:
  1. Sample z ~ Q(***z***|***X***)
  2. Reconstruct P(***X***|***z***)
  3. Backprop to improve parameters of parametrized P and Q
- Problem: How do we calculate gradients through sampling step?:

  Reparametrization trick: express z~N(μ, σ) as z~μ + σε, with ε~N(0, I)

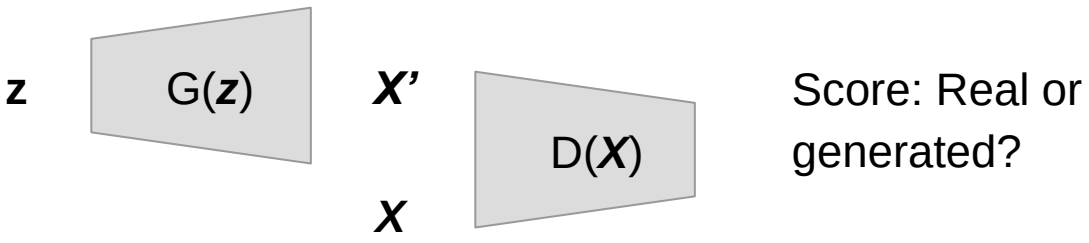


Often, likelihood P(X|z) assumed N(μ, σ), with σ = I (e.g. we just care about the mean, intensity of a pixel)

Auto-Encoding Variational Bayes Diederik P Kingma, Max Welling https://arxiv.org/abs/1312.6114

Models: Generative Adversarial Networks (GAN)

- We don't model directly P(X) but learn to sample from X: we learn P(***X***) implicitly.
- Game between two networks:
  - **Generator**: G(***z***) to learn to generate samples ***X'***. Generator is being seeded from a prior distribution (e.g. N(0,I))
  - **Discriminator**: D(***X***) to give a P(***X'***) of sample ***X'*** belonging to the real distribution (or in more recent GANs, just realness score). The discriminator provides information (gradients) to the generator on how to change its parameters so samples generated look real (and discriminator cannot differentiate between the real and generated samples).



Generative Adversarial Networks Ian J. Goodfellow, Jean Pouget-Abadie, et. al. https://arxiv.org/abs/1406.2661

Models: Generative Adversarial Networks (GAN)

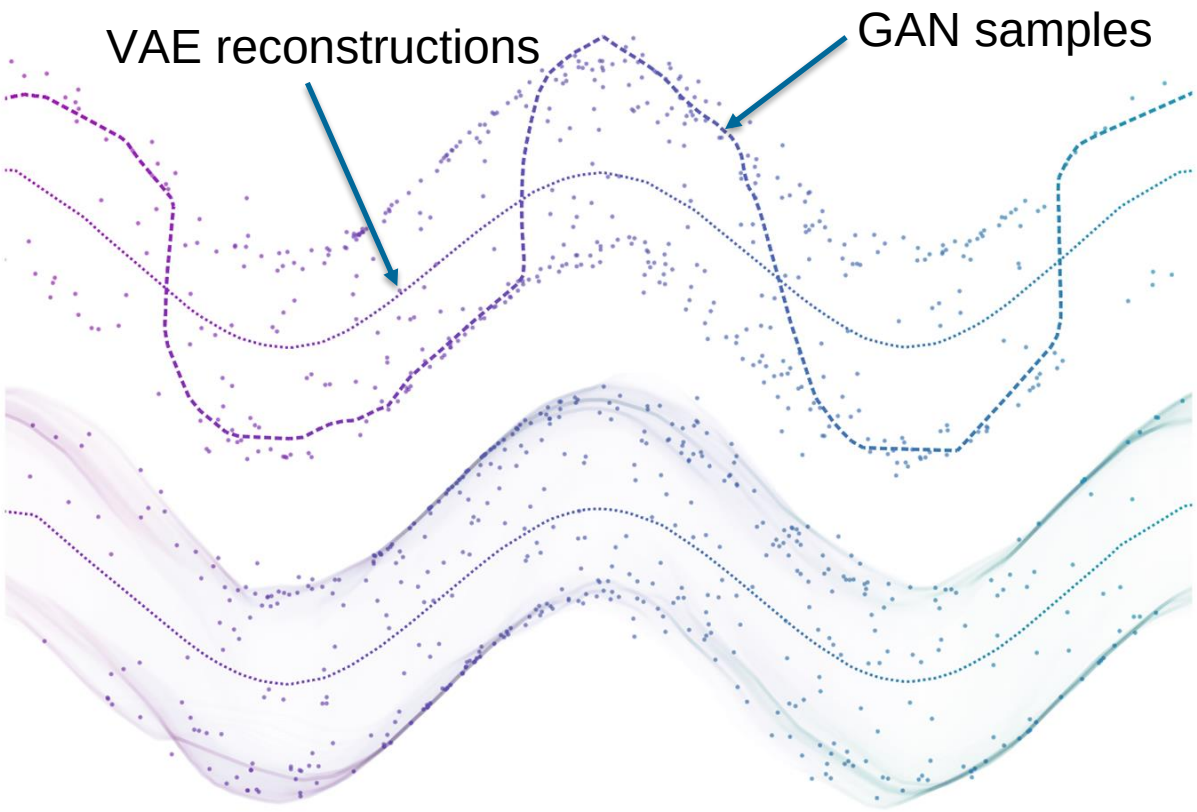


Figure 6: Illustration of a toy example with two-dimensional data and a one-dimensional latent space. Points: data, dotted line: manifold of reconstructions from VAE, dashed line/density: manifold of reconstruction with our model. Color encodes the position in the one-dimensional latent space. Top: with a deterministic generator of the form $G_{\theta_g}(z)$. Bottom: with a probabilistic generator of the form $G_{\theta_g}(z, \xi)$. (best seen with zoom and color)

- VAE learn the mean of the distribution (reconstructions tend to look blurry!)
- GAN learns to generate samples, no incentive to cover all the distribution! (*mode-collapse*)
- Implications for anomaly detection?

AVAE: Adversarial Variational Auto Encoder. Antoine Plumerault, Hervé Le Borgne, Céline Hudelo https://arxiv.org/abs/2012.11551

Models: Many more options to explore!

- Most of literature focus on VAEs and GANs, but many other ways to model P(X):
  - VAE and GAN variations (e.g. VQ-VAE)
  - Auto-Regressive
  - Flow models
  - Auto-Decoder

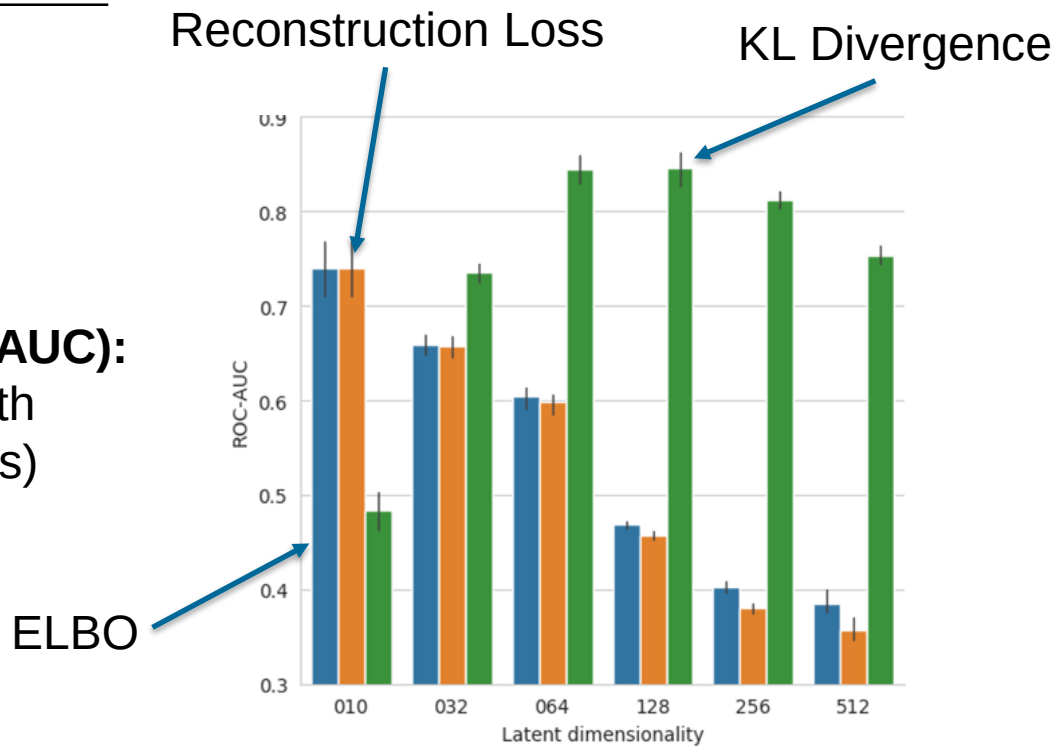
Methods: Intro

- We have a model that learns P(***X***), how do we know if a new sample is anomalous, or more interesting, where is anomalous?
- Sample and pixel-wise **Anomaly Scores** (AS)
  - **Reconstruction error** (*vanilla*)- Assumes that a model trained on normal images will not be able to reconstruct anomalies. L1 or L2 distances between reconstruction and original images.
  - In VAE, KL Divergence (pixel-wise AS not obvious)
  - **Restoration approaches** - Use models to turn test images into normal images (restore action). Then compare original vs restored. Generally, modify a test image to increase P(***X***)

Methods: Reconstruction AS with VAE

- Experiments on Brain MR images, predicting Gliomas:
  1. Reconstruction Loss performance decrease with latent size
  2. Given a model with sufficient capacity, KL Divergence term is a better sample-wise AS!

AS sample-wise performance (ROC-AUC):
(Brain MR Images with Gliomas as anomalies)

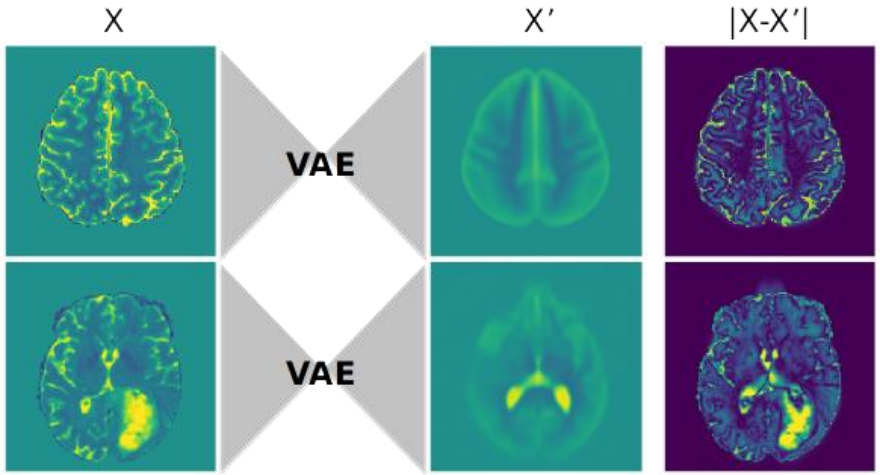


Similar conclusions are described in 'Unsupervised Anomaly Localization using Variational Auto-Encoders' Zimmerer, D., Isensee, F., et al. (2019)

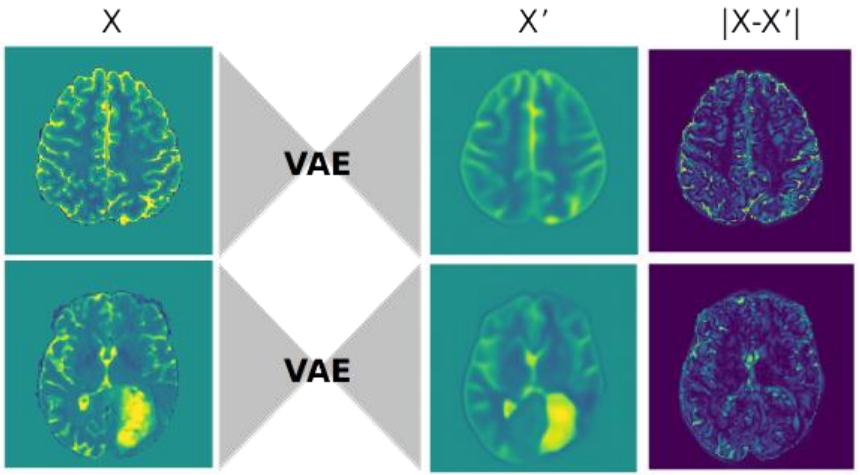
Methods: Reconstruction AS with VAE

- Why reconstruction does not work as Anomaly Score?
  - If the latent space is small, reconstructions are blurry (means!)
  - If the latent space is large, the VAE will be able to reconstruct anomalies.

- Model with 10d latent space -



- Model with 128d latent space -



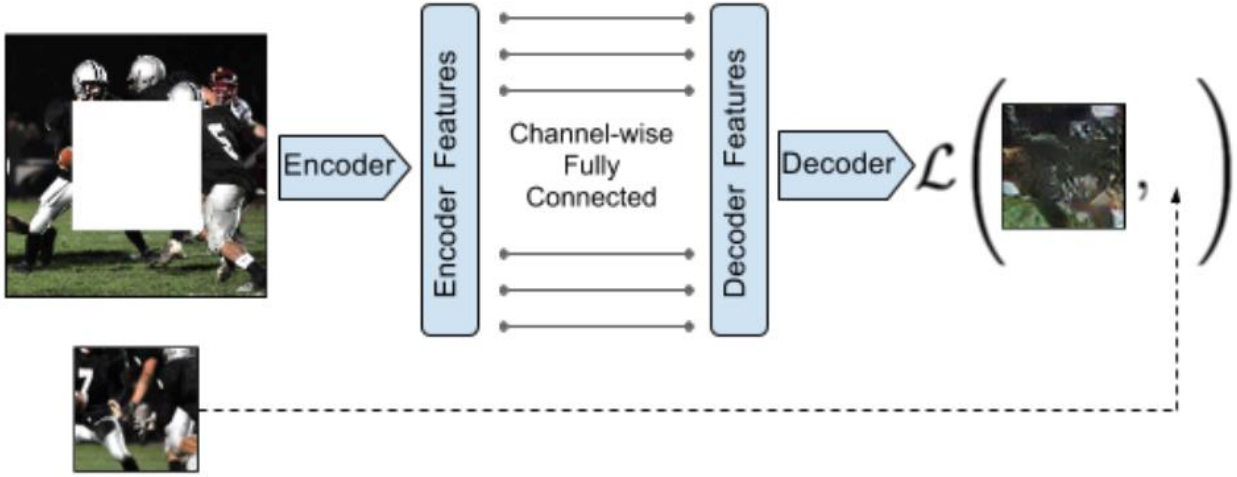
Methods: Alternative AS

- Expressive VAEs reconstructing anomalies are an issue, multiple strategies explored:
  - Use adversarial loss to check if reconstructions are still in normal distribution: VAE-GAN[1,2]
  - Force the model to learn structure: add Context-Encoding tasks[3]
  - Use KL divergence component assigned to pixels: KL-Grad[4]
  - Restoration: Use learnt distribution to increase P(***X***) [5]
  - Estimate density of the latent space a posteriori with a GM or AR model (my MSc thesis)

1 - Baur, C. et al. (2018) 'Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images'
2 - Chen, X. and Konukoglu, E. (2018) 'Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders'.
3 - Zimmerer, D., Kohl, S., et al. (2019) 'Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection -- Short Paper'.
4 - Zimmerer, D., Isensee, F., et al. (2019) 'Unsupervised Anomaly Localization using Variational Auto-Encoders'.
5 - Chen, X. et al. (2020) 'Unsupervised Lesion Detection via Image Restoration with a Normative Prior'.

Methods: Context Encoding (self-supervised learning)

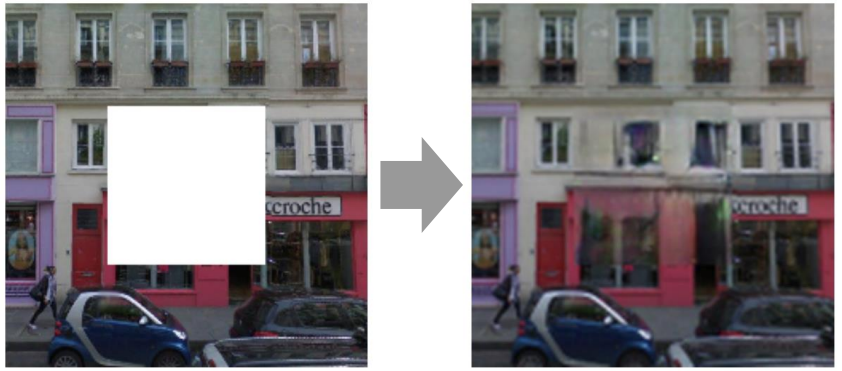
- High level idea: Add a task so the model needs to learn the distribution, not just encode-decode: Predict occluded images sections



$$\mathcal{L}_{rec}(x) = \|\hat{M} \odot (x - F((1 - \hat{M}) \odot x))\|_2^2$$

$$\mathcal{L}_{adv} = \max_D \quad \mathbb{E}_{x \in \mathcal{X}}[\log(D(x)) + \log(1 - D(F((1 - \hat{M}) \odot x)))]$$

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}$$



Pathak et al 2016

Methods: Context Encoding

- Context-Encoding improves anomaly detection performance[1]:

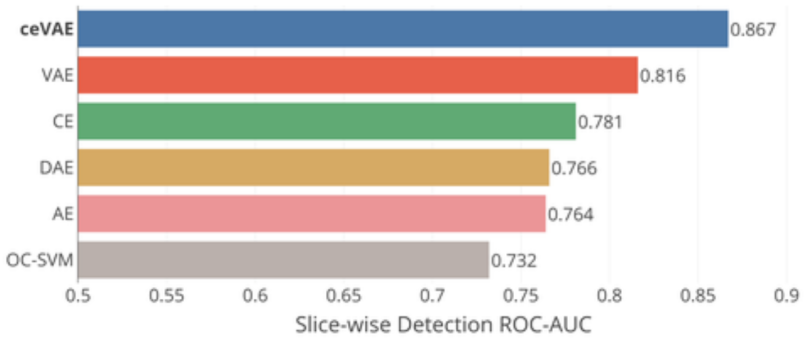


Fig. 2: Comparison of slice-wise anomaly detection performance of different models on the BraTS-2017 dataset.

(Note that author did not include adversarial loss as proposed in Pathak et al 2016)

1 - Zimmerer, D., Kohl, S., et al. (2019) 'Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection -- Short Paper'.

Methods: Restoration

- **Restoration**: Differs from a reconstruction because it includes an action to modify test image so it becomes in-distribution. Anomalies can be localized using residuals test - restoration
  - With **VAEs**[1] - Modify pixels with backpropagation so we lower a loss composed of VAE ELBO + Total Variation norm (keeps image consistent):

$$\hat{X} = \text{argmax}_X[-\lambda \|X - Y\|_{TV} + \text{ELBO}(X)]$$

The above expression is used to iterate until convergence, being:

- λ - weighting of the data consistency term
- Y - test image
- X - current restoration (X = Y at t=0)
- $\hat{X}$ - new restoration

1 - Chen, X. et al. (2020) 'Unsupervised Lesion Detection via Image Restoration with a Normative Prior'.

Methods: Restoration

Table 1: Summarized AUC and DSC for GMVAE(TV) and baseline methods. FPR and FNR are calculated from T_{ls} at DSC_AUC. DSC1, DSC5, DSC10 are calculated from T_{ls}s at $l_{FPR} = 0.01, 0.05, 0.10$. For GMVAE(TV), $\lambda = 1.8$. *na:not available*.

Methods	DSC_AUC	AUC	FPR	FNR	DSC1	DSC5	DSC10
VAE(TV) (ours)	0.34±0.18	0.80	**0.11**	0.40	**0.34±0.20**	0.36±0.27	0.40±0.24
GMVAE(TV) (ours)	**0.37±0.18**	**0.83**	0.12	**0.34**	0.22±0.21	**0.46±0.23**	**0.43±0.20**
VAE-256	*na*	0.67	0.26	0.43	*na*	*na*	*na*
VAE-128	0.22±0.14	0.69	0.21	0.46	0.09±0.06	0.19±0.15	0.26±0.17
AAE-128	0.23±0.13	0.70	0.25	0.43	0.03±0.03	0.18±0.14	0.23±0.15
AnoGAN	0.19±0.10	0.65	0.33	0.37	0.02±0.02	0.10±0.06	0.19±0.13

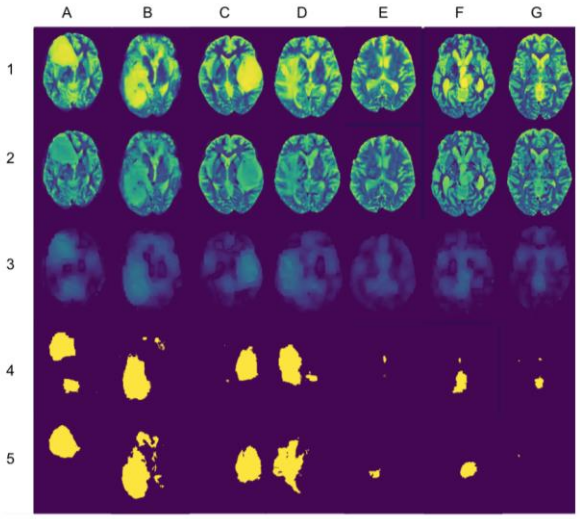
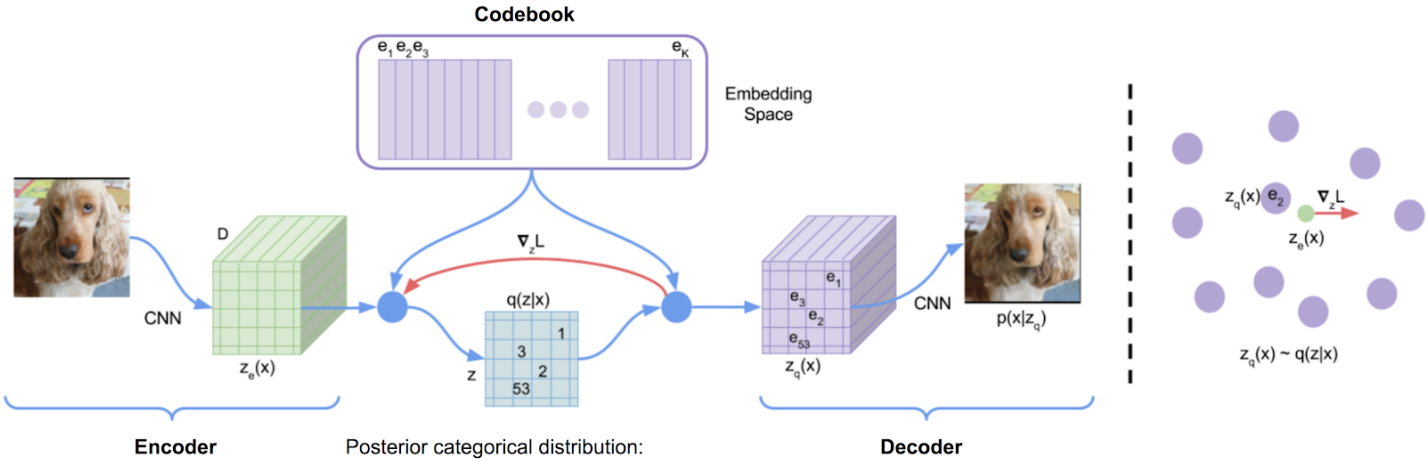


Figure 1: Segmentation by GMVAE(TV) at DSC5. Top to bottom: images with lesions, restored images, residual images, predicted segmentations, groundtruth segmentations.

1 - Chen, X. et al. (2020) 'Unsupervised Lesion Detection via Image Restoration with a Normative Prior'.

Model: Vector Quantised - VAE

Vector-Quantised VAE from original paper [1]



$$q(z = e_k|x) = \begin{cases} 1 \text{ if } k = \arg\min_i \|z_e(x) - e_i\|_2 \\ 0 \text{ otherwise} \end{cases}$$

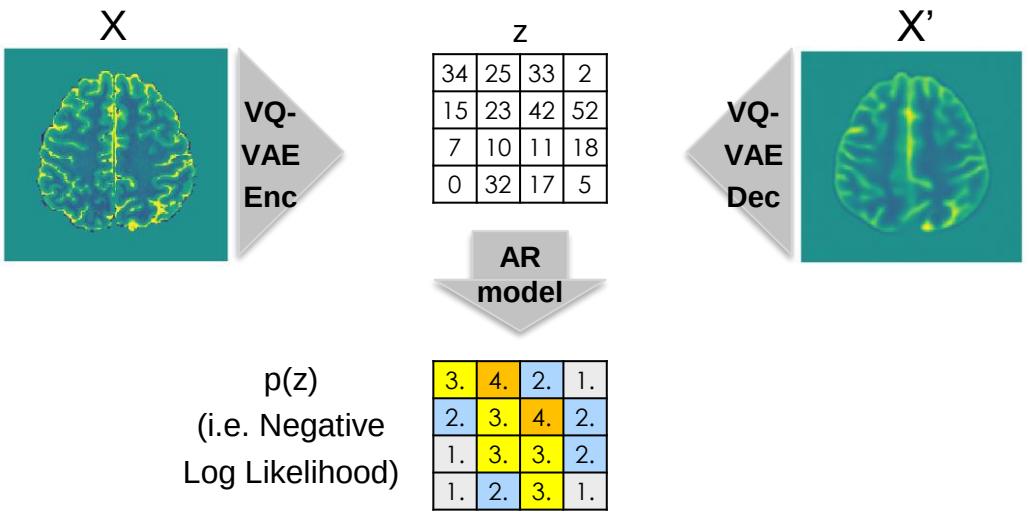
$$L = \log p(x|\, z_q(x)) + \ \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\ \|z_e(x) - \text{sg}[e]\|_2^2$$

- Latent space is a 2d matrix of discrete variables
- Model is built around a dictionary that maps to an embedding space
- Vector Quantisation: no gradient for argmin, approximated with *straight-through* estimator
- MOOD Challenge implementation:
  - Images pre-processed to 160x160 2d axial slices
  - Encoder-Decoders with 4 ResNet blocks at each resolution
  - Latent space:
    - Brain: 20x20, 128 categories
    - Abdominal: 10x10, 128 categories

1 - van den Oord, A., Vinyals, O. and Kavukcuoglu, K. (2017) 'Neural Discrete Representation Learning'.

Model: Auto Regressive model for prior of VQ-VAE

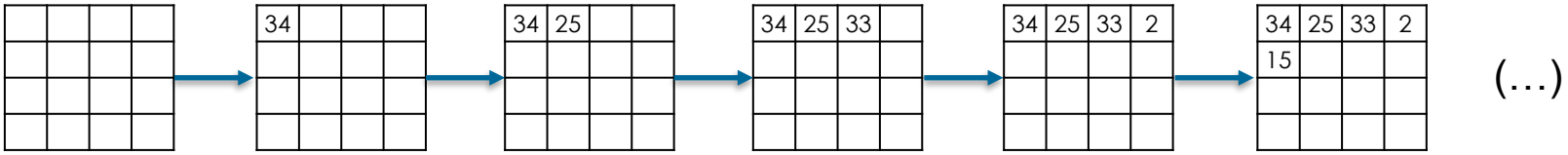
VQ-VAE is paired with an Auto-Regressive (AR) model to learn its prior.



In an AR model joint probability is modelled using factorization:

$$p(x) = \prod_{i}^{N} p(x_i | x_1, \ldots, x_{i-1})$$

AR models are generative, they allow sampling iteratively:



Method: Sample-wise score with VQ-VAE

Assumptions:

- VQ-VAE can reconstruct abnormal regions, however….
- …localized abnormal regions translate into latent codes with low probability assigned by the prior model
- The AR model tent to be very confident in background areas and thus have higher confidence in images with a higher foreground / background ratio

Sample-wise score:

- Define a threshold λ of log-likelihood of latent codes that are highly unlikely by looking at the distribution in the normal holdout dataset
- Score: Sum of negative log-likelihood, considering only codes above threshold:

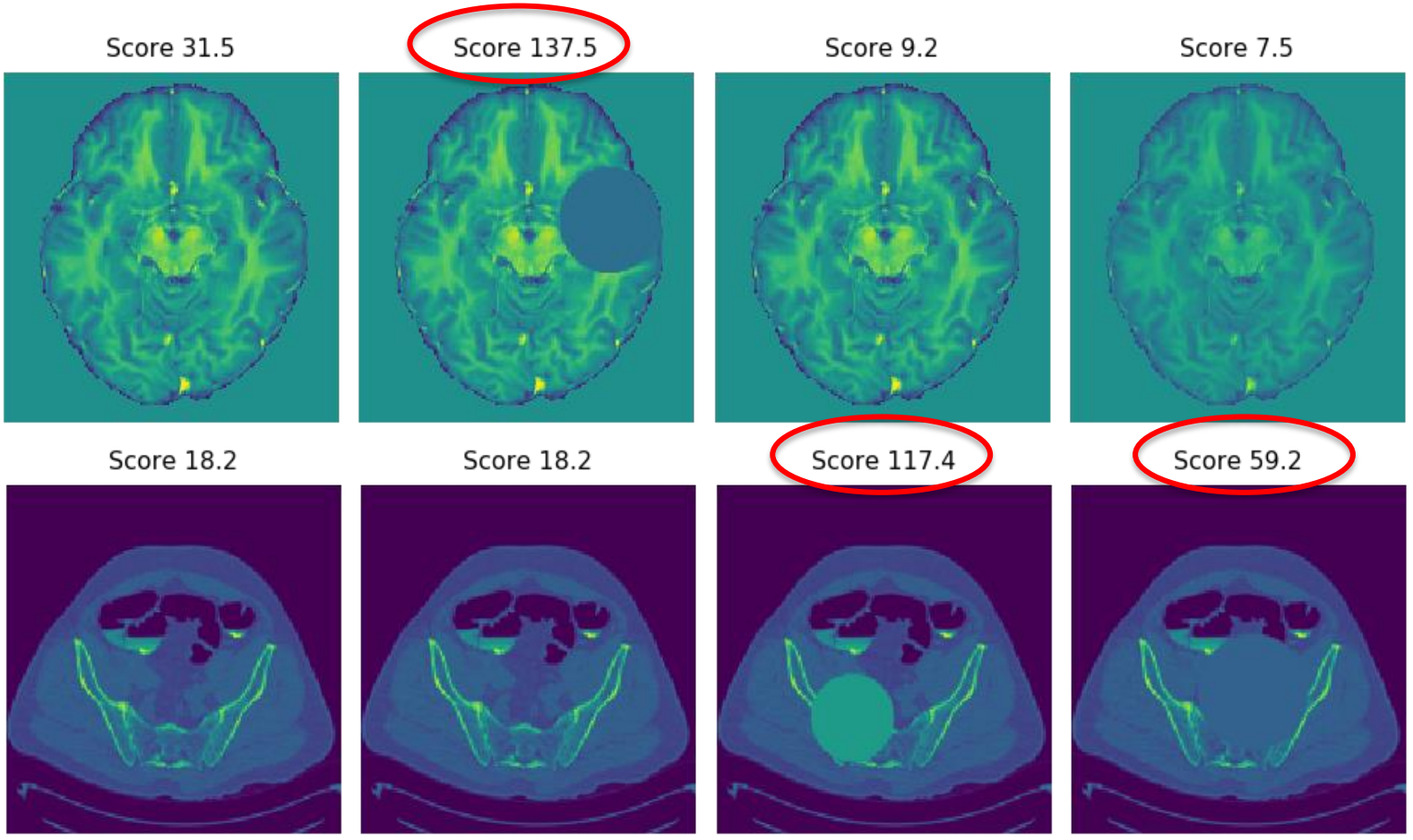
$$Score_{sample} = \sum_{i}^{N} \xi(p(x_i))$$

$$\xi(z) = \begin{cases} -\log(z) \text{ if } -\log(z) > \lambda \\ 0 \text{ otherwise} \end{cases}$$

- MOOD Challenge implementation: $\lambda = 7$. 32 axial slices per volume are processed, score is the sum over the slices

Methods: Sample-wise score with VQ-VAE

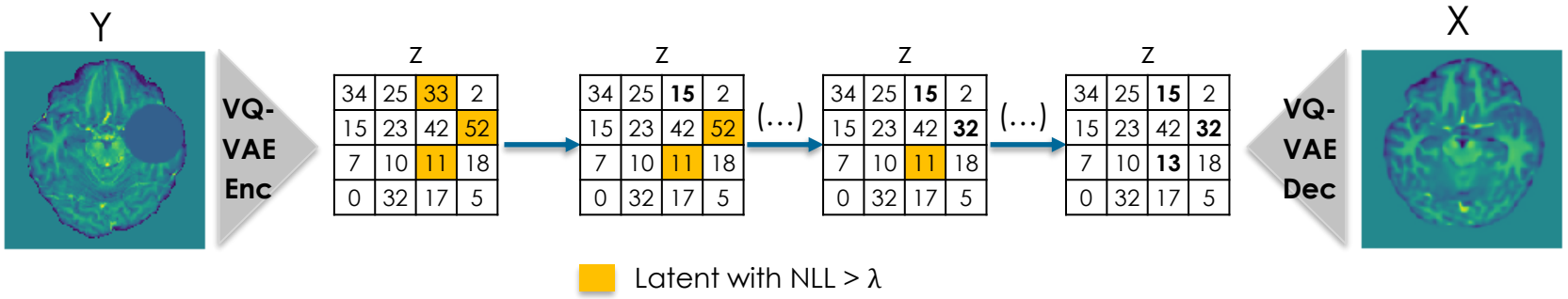
Toy slice example score:



Methods: Latent Restoration with VQ-VAEs

Pixel-wise score:

- Replace high loss latent codes with samples from AR (i.e. latent variables with NLL > λ).



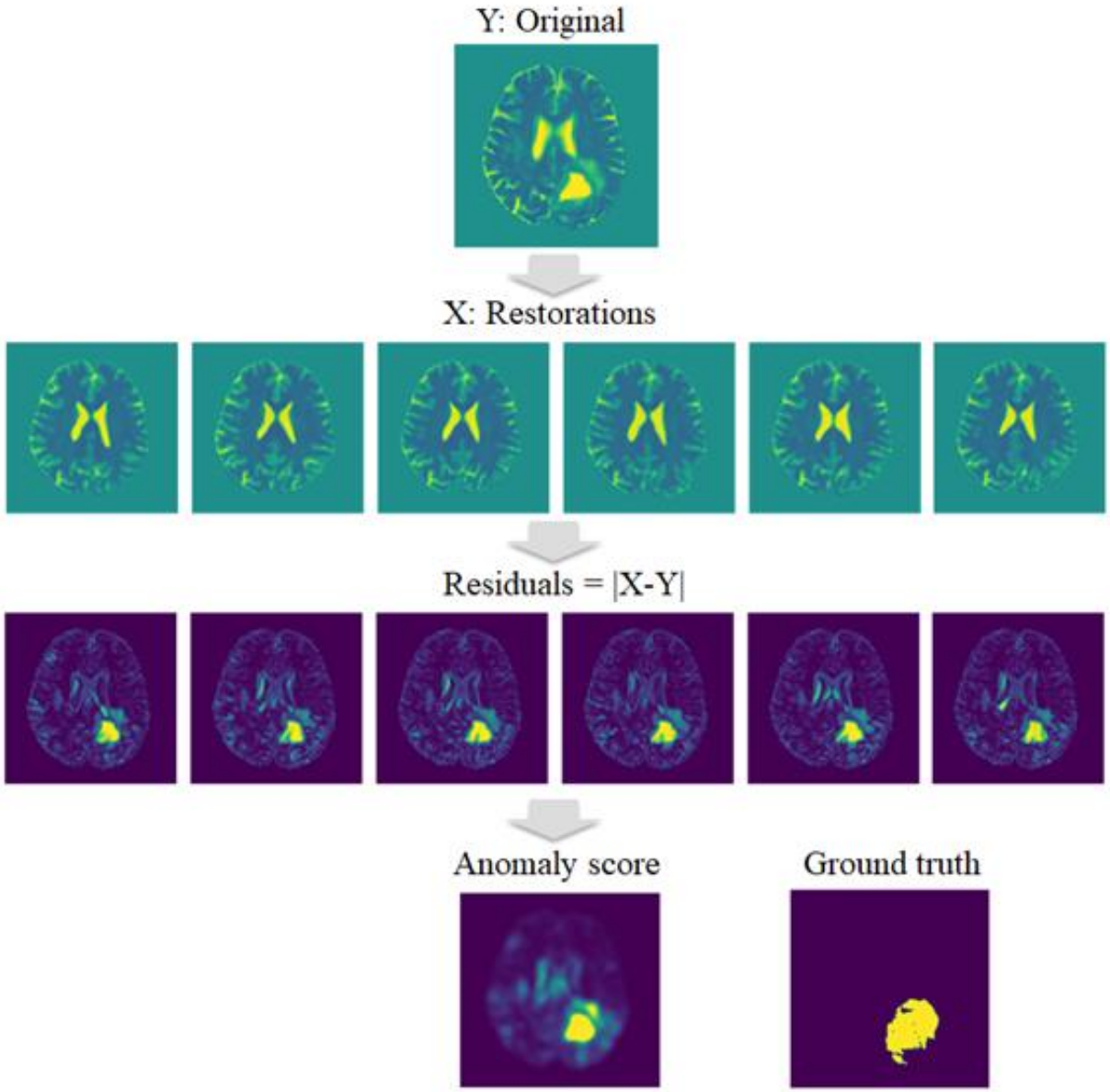
- Decode *restored* latents to generate image restorations *X.*
- To reduce variance, generate multiple restorations (j ∈ 1,2,...,S) for each image
- Pixel-score, residual $\widehat{D}$, is then:

$$Score_{pixel} = \sum_{j}^{S} \varpi_j \left| \mathrm{Y} - X_j \right|$$

$$\varpi_j = \mathrm{softmax}(k \,/ \sum_{i}^{P} \left| \mathrm{Y}^i - X_j^i \right|)$$

- Some restorations drift too much from original image. ϖ_j is introduced to remove weight from restorations that have lost consistency.
- Implementation: λ = 5, S = 15, *k = 3 (softmax temperature)*
- Post-processing: MinPooling + AvgPooling

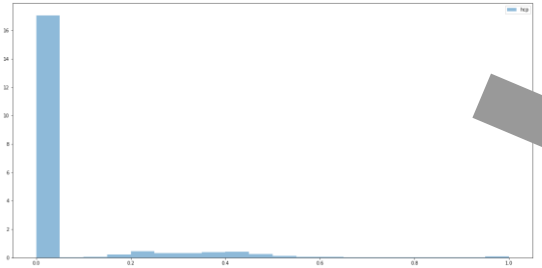
Methods: Latent Restoration with VQ-VAEs



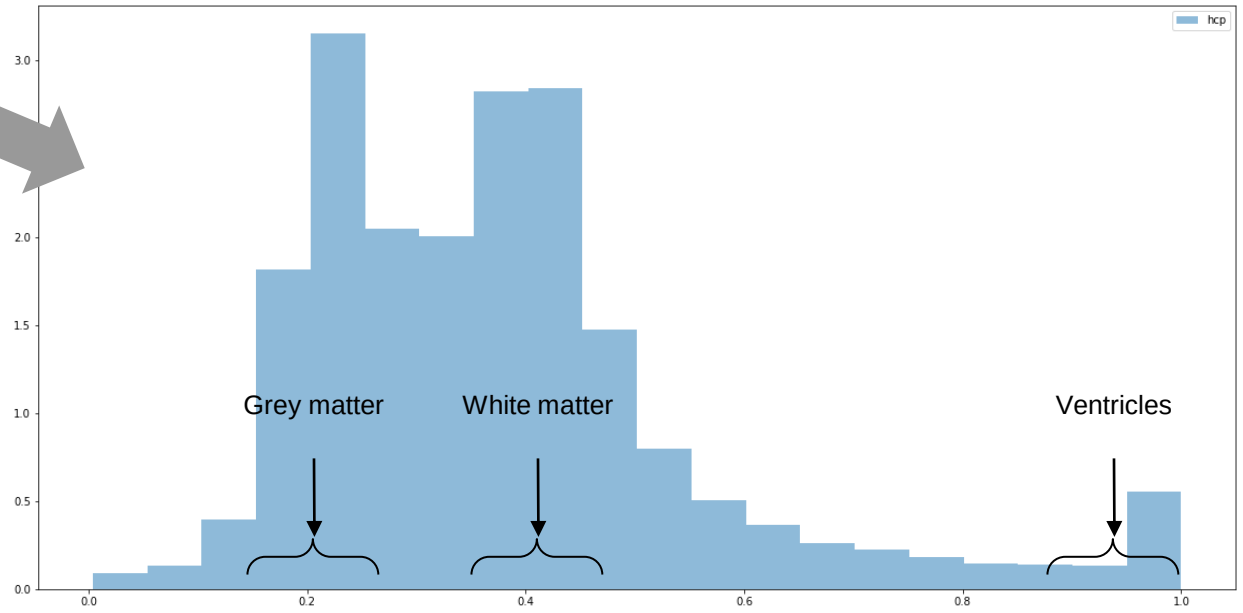
Methods: Other ideas with VAEs

Additionally, are L1 or L2 a good measure to localize anomalies in medical images?

- Histogram of MRI -



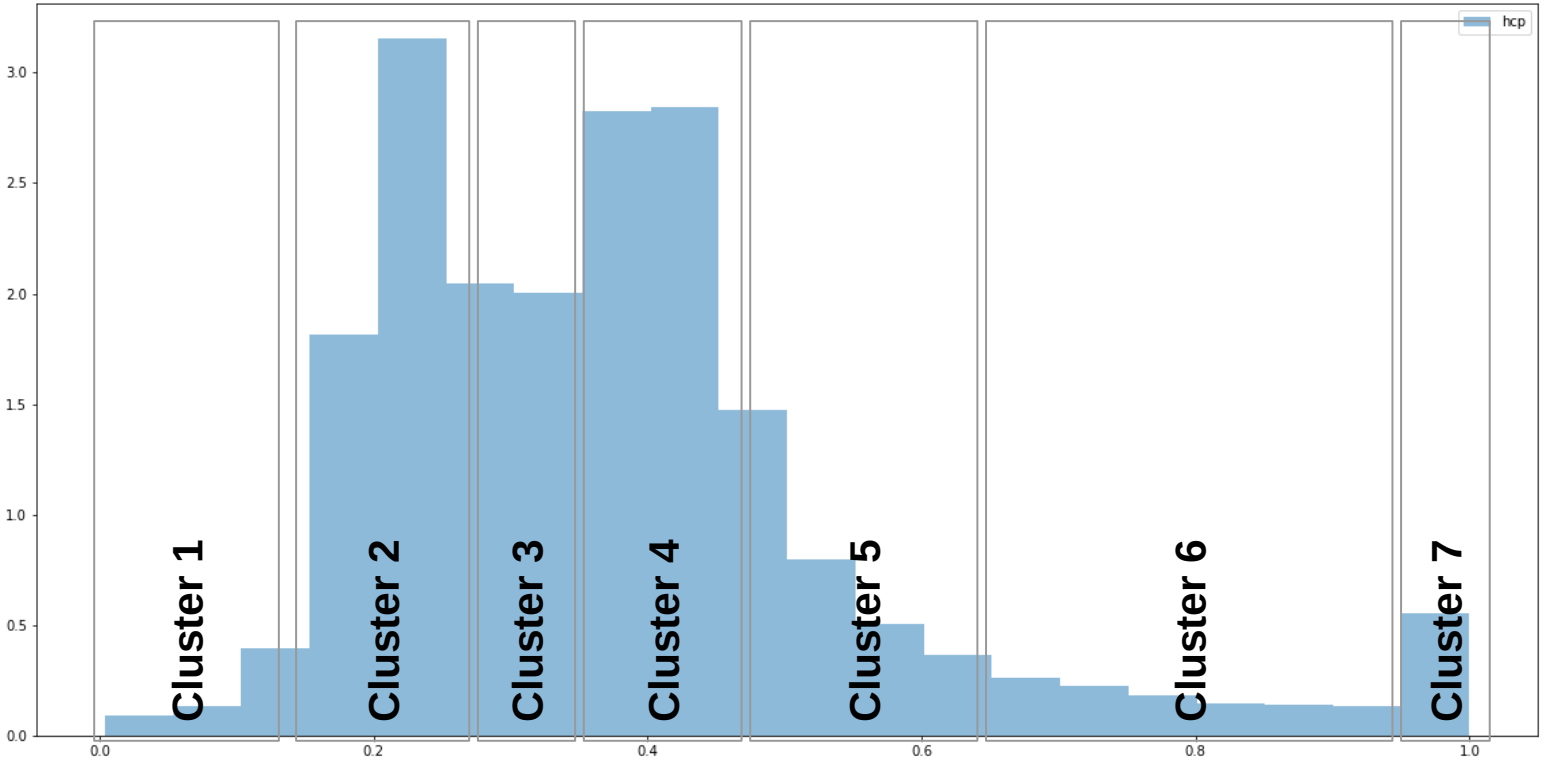
- Histogram of MRI (without background pixels) -



The distance in pixel intensities might not be relevant, some errors are penalized much more than others (i.e. intensities that are further away from the mean)

Methods: Other ideas with VAEs

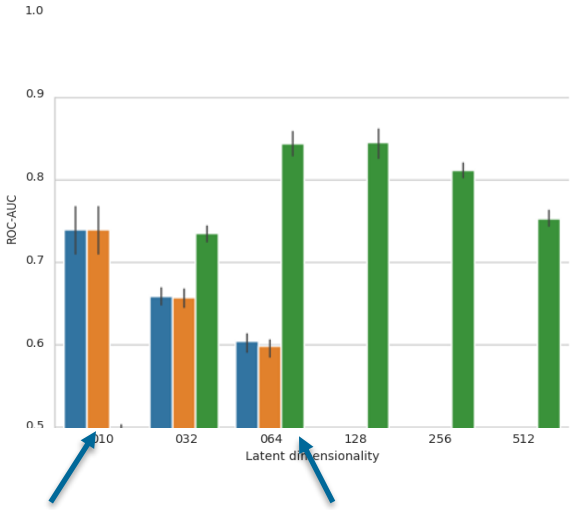
- Idea: Avoid distance in pixel intensity and focus on predicting the tissue type.
  - ...but we don't have tissue types….
    - Approximate tissue types using clusters of intensities (kMeans). Encode pixels to clusters of intensities: Regression → Classification
      - Cross-entropy as reconstruction loss and AS: Errors between tissues become symmetric



Methods: Other ideas with VAEs

- The categorical image encoding makes a big difference in anomaly detection performance!
- KL-Divergence is still a better Sample-wise Anomaly Score:

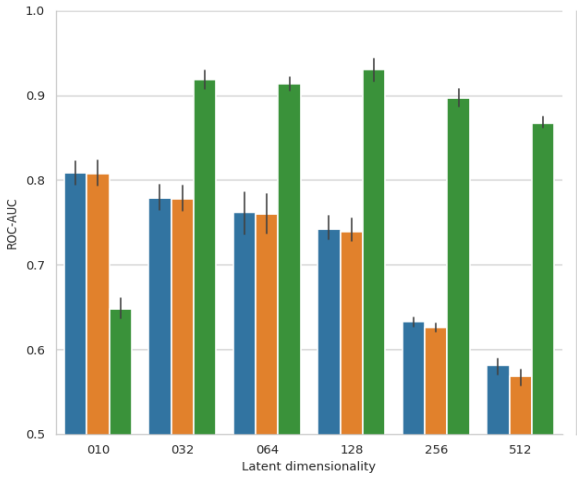
L1 Reconstruction Loss



Reconstruction Loss

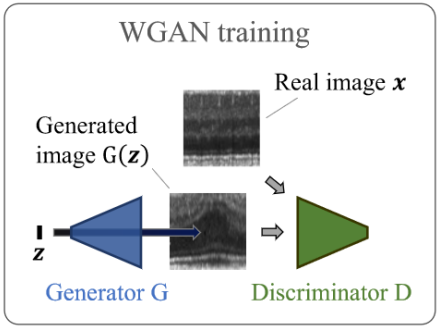
KL Divergence

Cross-Entropy Reconstruction Loss

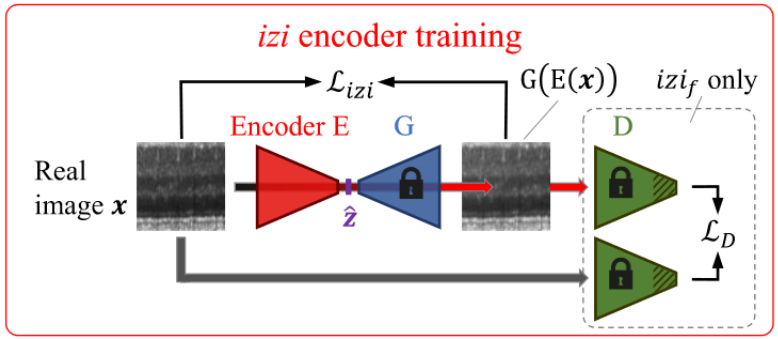


Methods: GAN restorations

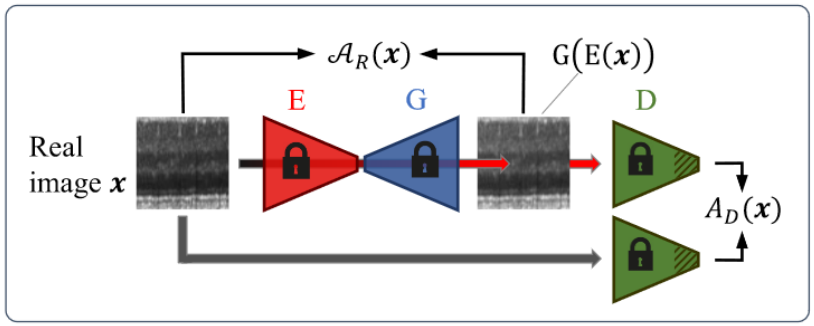
- **AnoGAN & f-AnoGAN**[1] - Find the equivalent healthy image from the GAN. Either by using backprop in latent space (original AnoGAN) or by using an encoder trained after GAN (fast-AnoGAN)



1) Train WGAN on healthy images



2) Train an Encoder to retrieve ***z***



3) At inference time, AS is defined:

$$A(\boldsymbol{X}) = A_r(\boldsymbol{X}) + \lambda\, A_d(\boldsymbol{X})$$

Being…
A_r the reconstruction loss
A_d the discriminator loss

1 - Schlegl, T. (2019) 'f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks'

Results

- Baur, C. et al. (2021) 'Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study':

  ~ 1st Restoration GMVAE, 2nd Context Encoding / AnoGAN

- Medical Out-of-Distribution Challenge @ MICCAI 2020
  - **Sample-wise**: 1st Foreign Patch Interpolation (supervised on synthetic anomalies), 2nd VQ-VAE.
  - **Pixel-wise:** 1st and 2nd supervised on synthetic anomalies, 3rd VQ-VAE shared. (VQ-VAE did very poorly on the Abdominal dataset!)
- MSc. Thesis (brain MRI dataset):

	Average Precision	ROC-AUC	DSC	FPR @95 Recall
VAE MSE reconstruction	0.270	0.942	0.339	0.242
VAE Cross-entropy reconstruction	0.354	0.949	0.415	0.219
VAE KL Grad	0.240	0.857	0.343	0.653
VAE Restoration	0.314	0.924	0.353	0.306
VQ-VAE L1 reconstruction	**0.463**	**0.977**	**0.528**	**0.105**
VQ-VAE Cross-entropy	0.414	0.972	0.466	0.122

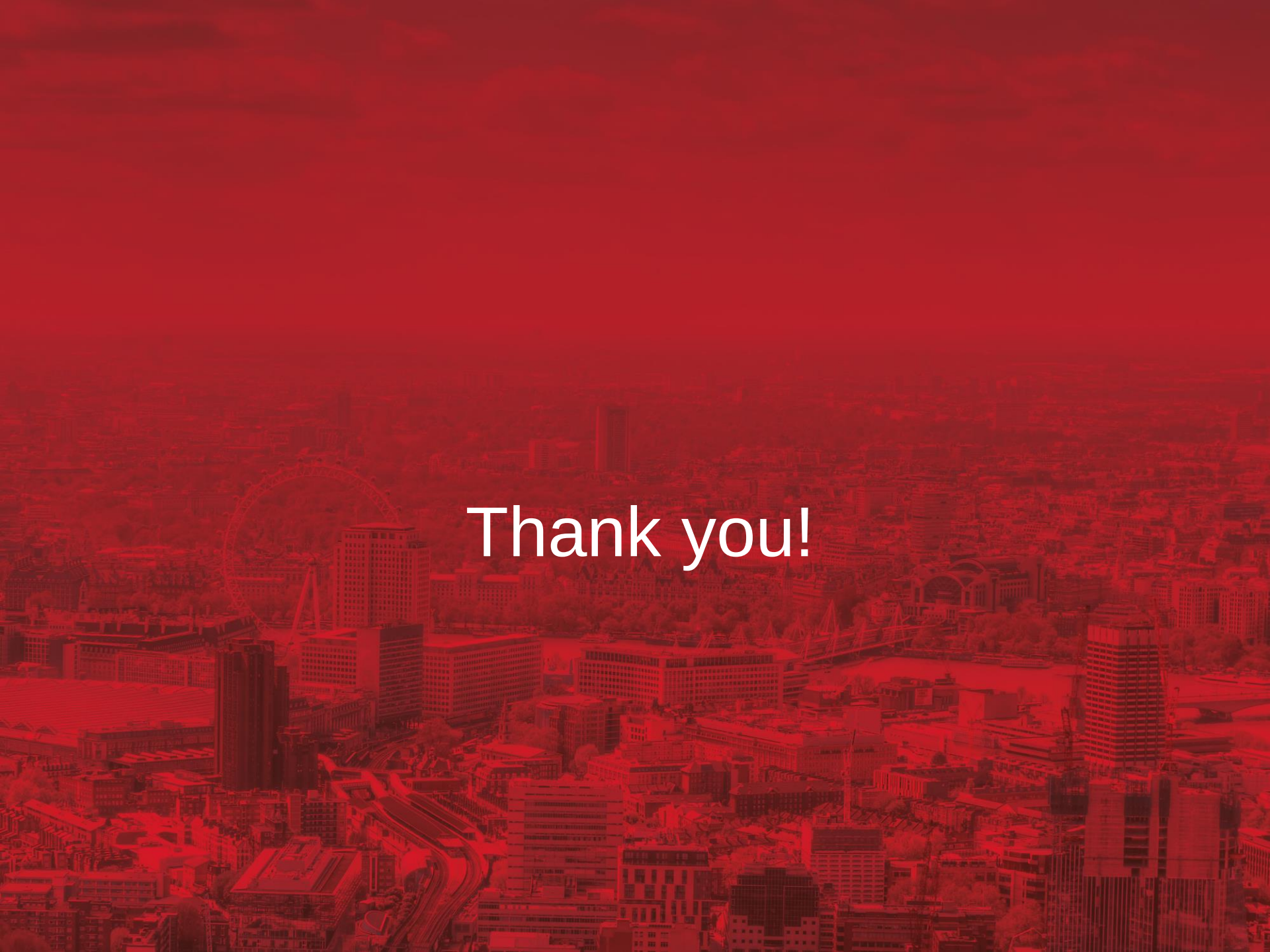
Table 4. Brain test set results for pixel-wise anomaly detection. VQ-VAE with 20x20 latent space. DSC coefficient reported is the maximum for the precision-recall curve

Discussion & Conclusions

- Unsupervised results have improved in recent year but still poor (when compared with supervised learning). Good research opportunity!
- Supervised learning on synthetic anomalies achieves significantly better results
- We have models that are able to learn the distribution of normal images and generate good samples …. However, open research question is how to leverage these to identify anomalies
- *Healthy images without labels >> Images with annotated anomalies*

  …however, in datasets publicly available, it is the opposite (segmentation challenges,...)

Q&A


Thank you!