**AML**: Questions before the test
week 1 introduction

- what are we estimating in the supervised scenario vs generative scenario (6) -
- formulate the cross entropy loss (19) *
- formulate SGD with momentum (21) *
- what's the difference between empirical risk and generalization (23) -
- give 3 'Real World Hypothesis Class' requirements (28) *
- why are linear function good for some problems but insufficient for others (29) -
- what are NNs good, ok and poor at according to Yadid (30) *
- In what sense does Classical ML theory not generalize well to deep networks (31) *
- give 2 regularization methods and their formulation (33) *
- survey the 5 steps of supervised ML pipeline (35) *

**Architectures**

- formulate the fc layers with Relus (45) -
- why not using only Fully-connected (FC) layers (with compression) (51) -
- how do we increase context with CNNs -
- describe 'depth-wise' convolutions and their advantages vs regular ones (61) -
- explain self attention in detail, and why are CNNs a private case of self attention (63-65) -
- explain the context of each token in the CNN vs Transformer scenarios -
- describe the problem with very deep NNs and the solution (69) -
- explain Batch/Layer normalization and why are they needed (73) *
- summarize the Two ways of reaching global context with few params (77) -
- Should you try to find your own architecture? -
- How do you use transformers for images?
- When would we use a transformer and when would we use CNNs in terms of data size and why? -

**Autoregressive models:**

- What are the 3 tasks generative models try to do? (93) *
- formulate the joint probability of N variables using bayes decomposition (102) *
- formulate the loss of the autoregressive model in general (103) *
- formulate the loss in simple terms (106) *
- Why is it hard to implement on images? *
- How do we sample from an autoregressive model? (109-111) *
  - What are the problems with the 'most likely' approach?
- How do you do point estimation on an autoregressive model? -
- formulate the perplexity measurement. (113)
- Example: Linear Language Model
  - what is the problem of a linear language model (114) (order)
  - What is good? (115) (words representations)
- Formulate the conditional scenario (119) *
- What are the ways to use it on continuous data? (121) (parametrization, discretisation) *
- How does the 'parti' model by google works? *

- What are the 3 major limitations of Autoregressive models? (133) *

**neural scaling laws:** (125)

- Where would you invest? (data, compute, bigger batch sizes) (130) *
- describe power law analysis and the GPT predicted accuracy (128) *
- Does the architecture of the transformer matter? (129) *

**variational inference**

- What is the posterior distribution $p_{\theta}(x|z)$ (141) -
- how do we compute the likelihood $p_{\theta}(x)$ given the latent variable $z$? (142) *
- Explain and formulate importance sampling vs monte carlo method (144-148)
- how do you compute $p(x)$ with monte carlo sampling?
  What's the problem with it? (149)
- Formulate the variational inference loss from importance sampling:
  $log(p(x)) \geq E_{z\sim q} log(p(x|z)) - KL(q||p)$:
- What is the best proposal distribution? (156)
- What loss are we optimizing and what parameters are we looking for in the latent optimization scenario of VI (157)
- What is the reparameterization trick and how does it help to propagate the gradients? (158)
- How is the VAE in the amortized version optimized in practice (the formal implementation vae) (161)
- Why are the results of VAE blurry? (164, 171)
- how do we measure $p(x)$ (point estimation) using the VAE formulation (formula and practice)? (165)
- Why don’t we use the same method for optimizing G,S? (we wanted the mean out of the log so we used jensen inequality)
- What does the conditional VAE formulation look like? (167-168)
  - formulate $p(x, c)$
  - after choosing the $q$ of the importance sampling as $q = q(z|x, c)$
  - formulate the ELBO inequality
  - formulate the loss

**Diffusion models**

- formulate the diffusion process SDE -> x(t+dt) = ? (176)
- formulate dx if we take the step size to 0
- what is $dW_t$ and $E[dW_t]$
- Can we invert a noise to a **specific** image? (180)
- why not just invert the sign of the dx in the SDE? (180)
- What is the difference between VP and VE processes? and Why to use VP (optimization considerations)
- write the Fokker-Planck equation and explain it in general terms (181)
- What is a score function? (184-187)
- How do we compute the estimation of the score function? (187 - 190)

- given a very noisy image, and a denoiser. Given the fact that a lot of images could result in this image given an added noise,  what is this denoiser going to estimate?
- How do we train a denoiser?
- write the process of DDIM sampling loop (pseudo code) (191)
- what is the intuition behind the ELBO log probability estimation of a sample (192)
- Why would we want to estimate p(x) for an image?
- How do we change the formulation (DDIM) for a conditional scenario? (195)
- formulate the denoiser in the Classifier-Free Guidance scenario. what does ω do?
- How does MagVIT work? (203)

**GANs**

- Does it have point estimation?
- what is Integral Probability Metrics (210-211)
- What is good in this idea and what is bad? (212)
- What is the alternative of IPM in GANs? (214-215)
- Why don't we optimize G(z) vs a trained discriminator d(x)? explain using IPM
- formulate the GAN loss and explain why does it work
- what is easy with GANs, and impossible (219)
- why is JS divergence giving the optimal discriminator but is still unusable? (221) (p=Not(q))
- What is the wisherstain distance? And why is it better then other matrices (225)
- Why do we want to enforce Lipschitz-1 on the discriminator function? and how do we do it? (5 ways)  (234-235)
- What is the problem with the loss that enforces lipschitz-1 on d(x) for x's from the training set? (only enforces it in certain places in the images distribution)
- What is Spectral Normalization? (235)  (largest eigenvalue smaller than 1, divide in sigma)
- explain the Wasserstein GAN
- How can you do disentanglement in styleGan? (243)

**Representation learning**

- What are 5 good criteria for choosing representations (248-253)?
- Why is learning 'Informativeness on Semantic Attributes' easier than learning 'Meaningful Distance on Semantic Attributes'.
- What is the **general Distance Preservation** Schema (256)
- What are we learning in these classical methods? (a matrix of representations)
- explain a Simple Case: **Classical Dimensional Scaling** (257)
- Give the final solution formula of the representations in **cMDS** and how to get to it (258)?
- So why not do PCA (259)?
- Explain the algorithm and the loss of **laplacian eigenmaps** (261)?
- Why do we have to add the $cov(E) = I$ term (261) Why does it ensure that solution won't collapse?
- What are the issues with **Local Methods**?
    - if the manifold consists of 2 'pages' or just noise the representations losses meaning

    - $cov(E) = I$ doesn't necessary solves all problems: it can put most of the embeddings around 0 and pad some other embeddings in random places to enforce the constraint
- formulate **SNE** Stochastic Neighborhood Embeddings (263)
    - What does sigma represent?
    - Why is it not a symmetric matrix, and how do you formulate a **Symmetric SNE**? (264)
    - Explain **tSNE** and what is the difference with SNE (265)
- What are 4 issues with classical methods? (267)
- What are the 2 main differences between DL representation learning and the classical methods? (268)
- What are the 2 problems with classical methods that are solved with DL? (269)
- Why do we use augmentations for generating the neighbors in representation learning? and what are the problems with augmentations ? (272)
- Explain **Contrastive learning** and formulate the loss,for one augmentation (273)
    - What is the intuition for the loss term? (275)
- Explain the **classifier interpretation to contrastive learning**, and what do we do to save computations (274)
- formulate the **VICReg** loss (Ex2) (277)
- formulate **CCA,** and its loss in the **Barlow Twins** scenario.
    - what is the rationale of learning the $corr(f(x_i), f(x'_i)) = 1$
- Explain the **SIMSIAM** Idea (281)

- Explain **Multi-Modal Supervision** for contrastive learning, and the Clip ML and DL tricks (282-284)
- Explain **masked auto encoding** (285)
    - What do we estimate?
    - What does it learn?
    - Explain how '**Bert'** was trained? What was the motivation? (286)
    - What is the difference in the final use of Masked Autoencoders for images and text? (for images: a way to do transfer learning, vs representation learning)
    - What can be a reason that Masked Autoencoders fail to learn good representations? (don't have to learn anything semantic)
    - What is the implementation trick for Masked Autoencoders for images for saving memory and computation, and why can transformers do that?
- Explain how **BEIT** works, was it better than the masked AE?

| method | pros | cons |
|---|---|---|
| Contrastive | works good | ● for each iteration requires a pass on all images/ big batches |
| Barlow Twins | every sample is independent from other samples, doesn't need | |

| | | |
|---|---|---|
| | negative samples (doesn’t require big batches) | |
| SimSiam | doesn’t require big batches (negative samples) | isn’t very explainable, can collapse |
| VICReg | | ● VICreg requires tuning the variance threshold and the regularization weight parameters<br>● VICreg assumes that the data has a Gaussian distribution along each dimension |
| BERT | | ● not trained for autoregressive tasks |
| BEIT | | ● too complex, the tokenization proved to be unnecessary |
| MAE | ● good for pre training | ● bad for representation learning |

**Compositionality and Disentanglement**

- describe the Compositionality problem in representation learning (295)
    - we want to be able to separate between attributes in our representation (color and label of fruit)
- describe the biasing problem, and 2 ways to overcome it (296)

Disentanglement

- describe the disentanglement task (297)
- How does disentanglement entail compositionality and classification? (298-299)
- Formulate the Identifiability in the Linear Setting (301)
- In which scenario this problem is unidentifiable? (302)
- Explain the **ICA** method in general and what it tries to solve
- Explain **BetaVAE** for unsupervised disentanglement (304)
    - Why doesn’t it work?
        - What enforcing $cov(Z) = 1$ over the representations doest solve the disentanglement problem? (any rotation of the axis gives a $cov(Z) = 1$ but we don’t know that these are the features we want or a mixture of them)
- Explain the scenario of conditional disentanglement (not included in test)

**Anomaly detection**

- What is **Retrieval**? (315-317)
    - What is the Hubness problem? and what can be done to address it?
- What is the connection between anomaly detection and density estimation? (324)
    - How to choose a threshold?
- What is Out-of-Distribution Generalization, what is the assumption hidden inside it? (327)  (we decode the data from a highly dense representation of it so we need to have good idea what is it going to look like)
    - What is the problem with it? (328)
- Explain **Deep Nearest Neighbors** (331)
    - How does it approximate P(X)?
    - Why is it better than P(X) in sparse areas? (sparse and therefore noisy and bad estimation)
    - What is the problem with high dimensional representations?
    - Is it robust across datasets? (333)
- What is the **Deep Anomaly Segmentation** (SPADE) pipeline? (335)
- What are 4 desirable properties of anomaly detection methods? (339)
- How can we combine training data with external data? (340)
    - Explain the **PANDA**  pipeline example
        - How can they avoid collapse to a singular point? (342)
- If we had the best representations possible, can we solve AD? (354)
- Explain the Expressivity-Sensitivity Tradeoff of anomaly detection and the conclusion from it. (356-358)
- explain the **No Free Lunch** principle in anomaly detection (358)
- describe a way to learn **Representations Invariant to Nuisance Attributes** (360-363)
    - how can you train z not to reveal any information over d?

**clustering**

- Under what assumptions can we do Unsupervised classification using clustering?
- Explain **Hierarchical Clustering** (368)
  - What parameters can be tuned?
- Explain **kmeans,** what is the difference from Hierarchical Clustering (369)
- Explain **Gaussian Mixture Models:** (370-372)
  - formulate a latent variable with a small set of discrete values
  - What is the assumption over the data?
  - For each latent variable, what parameters do we need to find?
  - how do we find $P(z = i|x)$?
  - What is the loss we optimize?
  - How do we do unsupervised classification using GMM (show the formula)?
  - Where does the GMMs assumption fail? (bad k, gaussian assumption)
- Explain **Classification-Based Clustering (373)**
  - What are the 2 assumptions we use for this method? (373)
  - What are the first 2 losses used here? write their formulas (374-375)
  - What would happen if we choose only one of them?
  - Are the 2 of them enough? (no semantics)
  - Can we use it on the pixel domain? (answer: no semantics)
  - What does the final loss and the architecture look like? (376)
- Explain **Domain Generalization**
  - what is the solution now for that (a lot of data, variation, big models)
  - Explain **Domain Adaptation** (379-380)
    - Give an optional solution for that, write the loss and explain it
    - When will this fail?