Active Learning using Deep Generative Models

Arpan Losalka [A0206973X]
Dipika Singhania [A0195129X]
Sun Xiaofei [A0198932M]

National University of Singapore, 2019-20, Semester II, CS6216

Active Learning

- Active Learning helps to develop label-efficient algorithms by **sampling the most representative queries to be labeled**

- **Importance: Reduce the high cost of acquiring labels**
    - Expert labelling, such as medical images, is the most prevailing application
    - Long labelling time per large scale data sample.

Related work

Selection of Space to Sample

- Data (images, video) itself
- Model Output Space
- Latent Space of Variational Autoencoder

How to cover all samples from space?

Samples must be diverse and cover entire space, not biased toward one region.

1. **Machine Learning Methods**
   a. Least Confidence
   b. Margin Entropy
   c. Geometry Based such as clustering

1. **Deep Learning Methods**
   a. Adversarial Discriminative Loss
   b. Ranking Loss with Adversarial Discriminative Loss

Related work

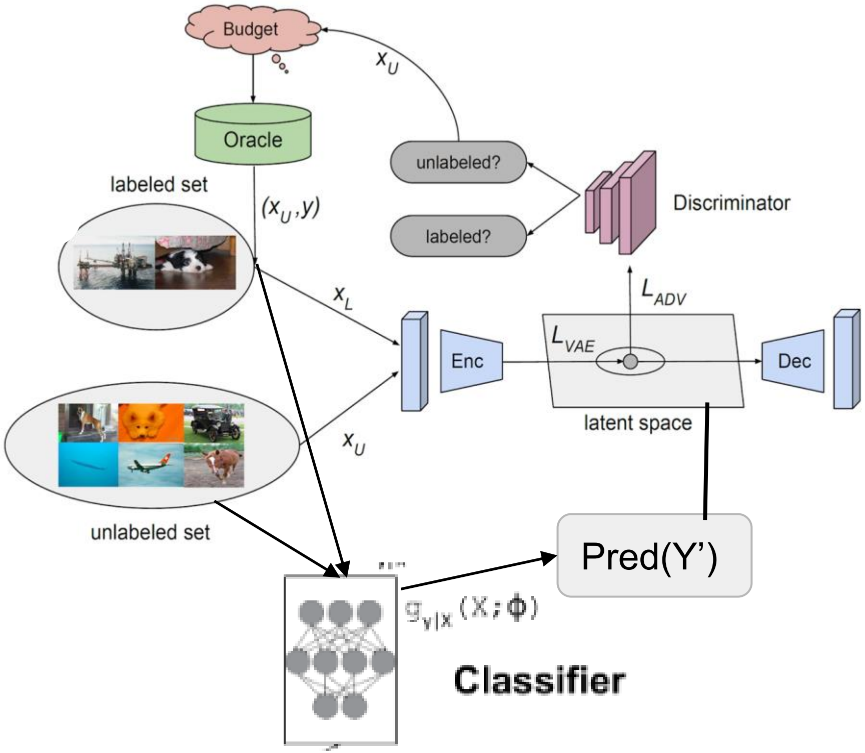
- **BALD-VAE: Generative Active Learning based on the Uncertainties of Both Labeled and Unlabeled Data** *- Lee, S. K., & Kim, J. H. (2019)*
    - Accounts for **uncertainty of labeled data** along with unlabeled data

- **Bayesian Generative Active Deep Learning** *- Tran, T., Do, T. T., Reid, I., & Carneiro, G. (2019)*
    - Combines **data augmentation** with **active learning**

- **Active Learning in VAE Latent Space** *- Tonnaer, L. M. A. (2017)*
    - Trains a VAE using available data, and defines various measures for ***informativeness, representativeness*** and ***diversity*** of query samples in the latent space of the VAE. Uses a linear combination of these measures to select queries for labeling.

- **Variational Adversarial Active Learning (VAAL)** *Sinha, S., Ebrahimi, S., & Darrell, T. (2019)*
    - Use adversarial discriminative loss in VAE latent Space

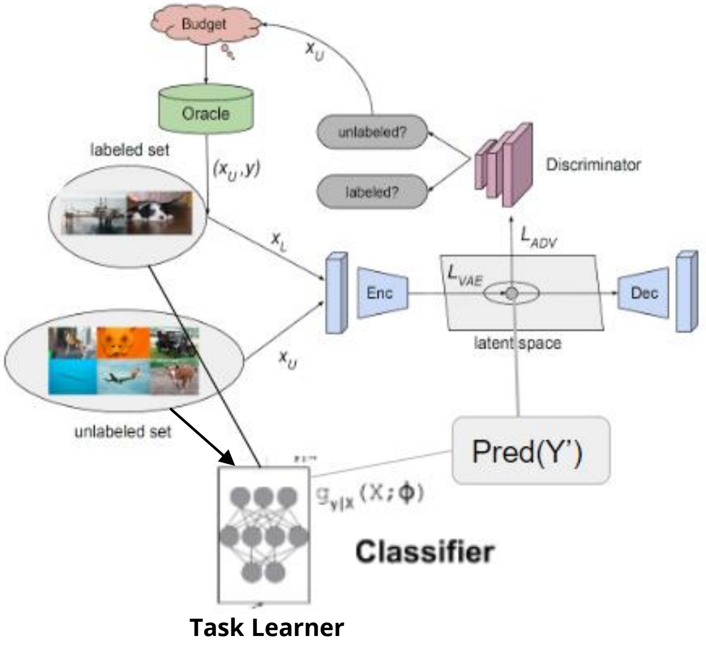
Proposed Model



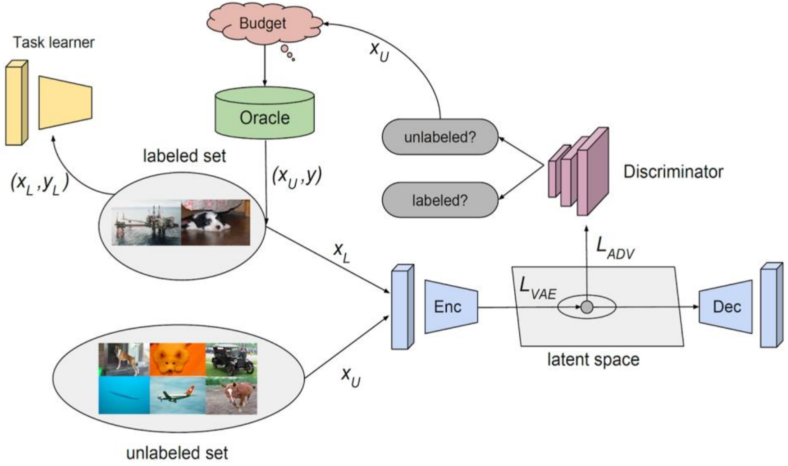
<u>M2 Model</u>: Semi-supervised learning with deep generative models (*Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014)*)

Proposed Model

M2-VAAL



VAAL

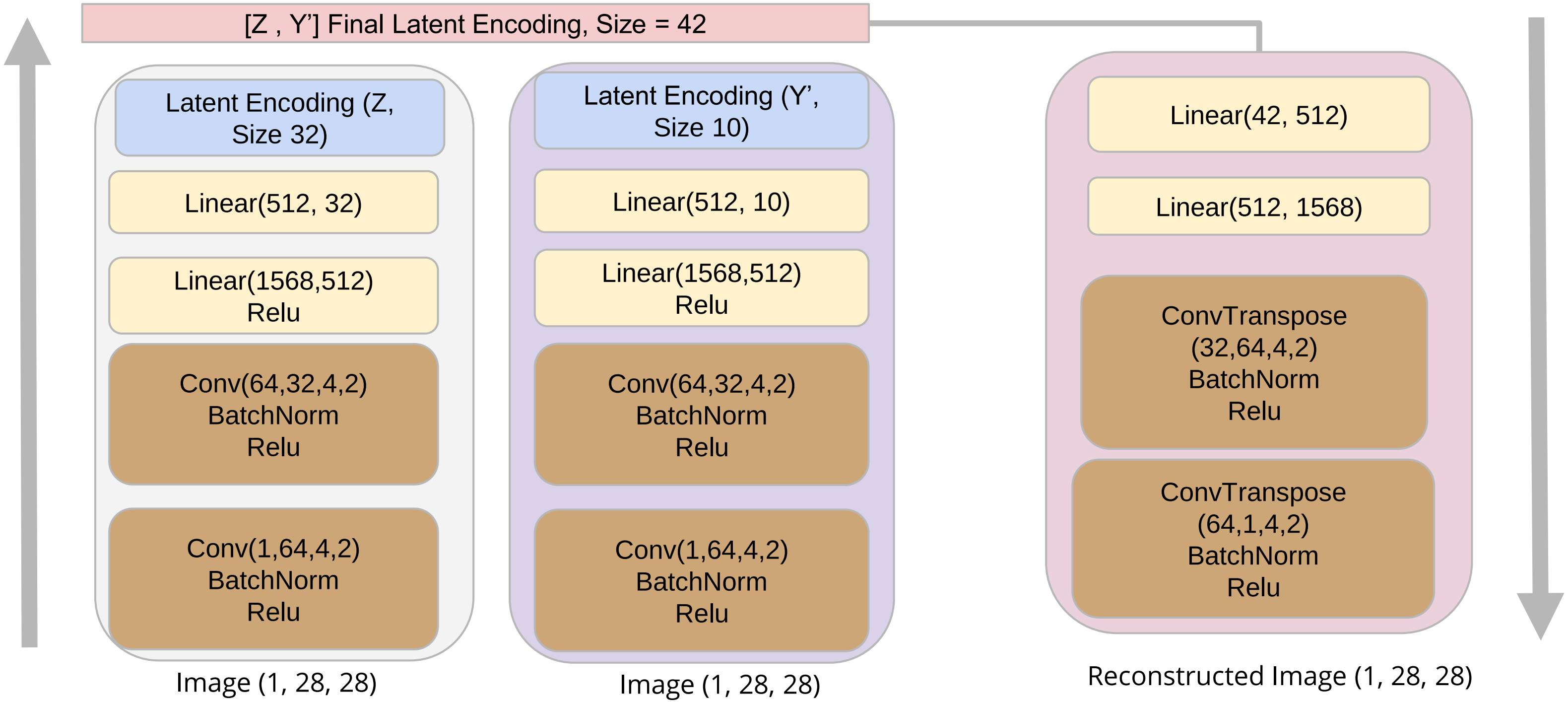


Components of our Architecture

- Variational autoencoder with classifier to learn a low dimensional latent space of the underlying distribution of data

- Adversarial Discriminator: Trained to map the latent space to binary label (1 or 0) to distinguish between labelled and unlabelled samples. Used for sampling labels which are most confidently classified as unlabelled.

Structure of VAE and Classifier Used



[Z , Y'] Final Latent Encoding, Size = 42

Latent Encoding (Z, Size 32)

Linear(512, 32)

Linear(1568,512)
Relu

Conv(64,32,4,2)
BatchNorm
Relu

Conv(1,64,4,2)
BatchNorm
Relu

Image (1, 28, 28)

VAE ENCODER

Latent Encoding (Y', Size 10)

Linear(512, 10)

Linear(1568,512)
Relu

Conv(64,32,4,2)
BatchNorm
Relu

Conv(1,64,4,2)
BatchNorm
Relu

Image (1, 28, 28)

CLASSIFIER

Linear(42, 512)

Linear(512, 1568)

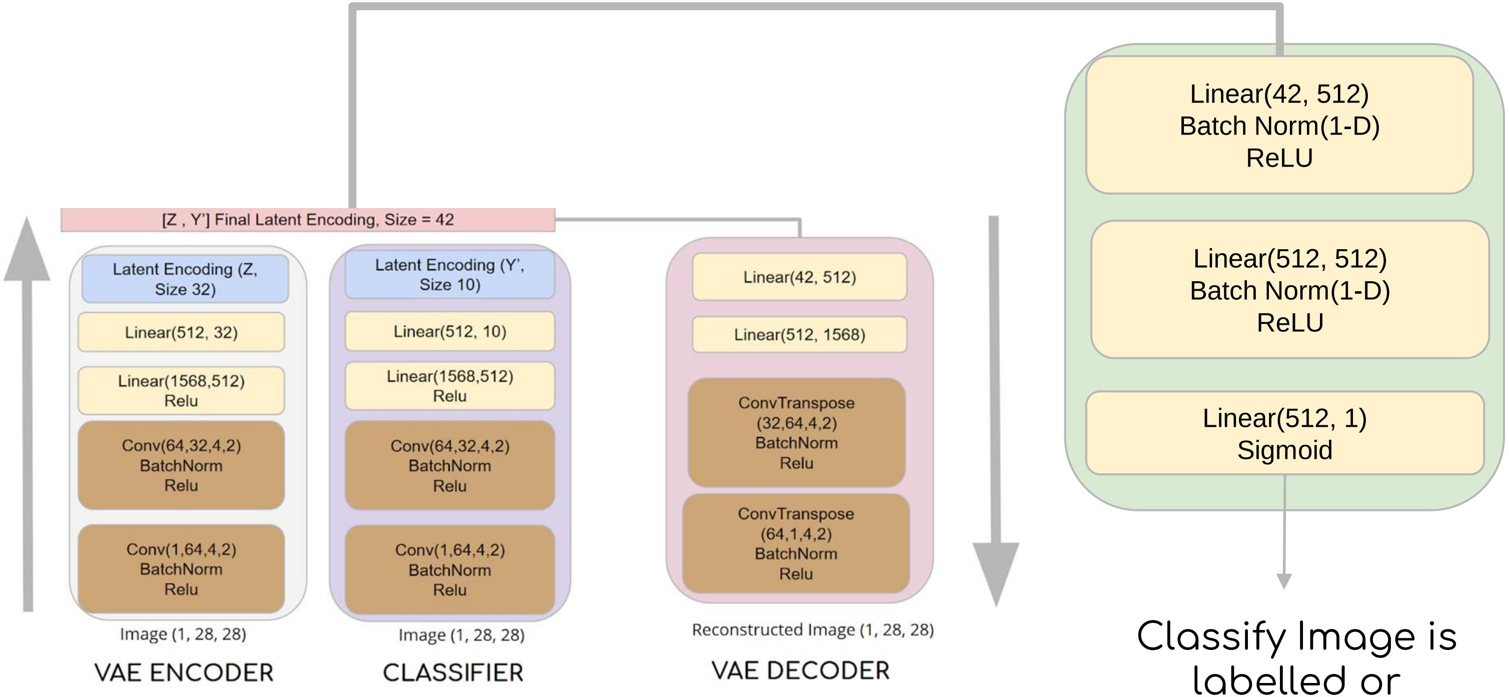
ConvTranspose
(32,64,4,2)
BatchNorm
Relu

ConvTranspose
(64,1,4,2)
BatchNorm
Relu

Reconstructed Image (1, 28, 28)

VAE DECODER

Structure of Discriminator Used


[Z , Y'] Final Latent Encoding, Size = 42
Latent Encoding (Z, Size 32)
Linear(512, 32)
Linear(1568,512) Relu
Conv(64,32,4,2) BatchNorm Relu
Conv(1,64,4,2) BatchNorm Relu
Image (1, 28, 28)
VAE ENCODER
Latent Encoding (Y', Size 10)
Linear(512, 10)
Linear(1568,512) Relu
Conv(64,32,4,2) BatchNorm Relu
Conv(1,64,4,2) BatchNorm Relu
Image (1, 28, 28)
CLASSIFIER
Linear(42, 512)
Linear(512, 1568)
ConvTranspose (32,64,4,2) BatchNorm Relu
ConvTranspose (64,1,4,2) BatchNorm Relu
Reconstructed Image (1, 28, 28)
VAE DECODER
Linear(42, 512) Batch Norm(1-D) ReLU
Linear(512, 512) Batch Norm(1-D) ReLU
Linear(512, 1) Sigmoid

Classify Image is labelled or unlabelled

Loss Used for Training the Model

- **VAE loss:** Minimizing the variational lower bound

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}^{trd} = \quad & \mathbb{E}[\log p_\theta(x_L|z_L)] - \beta \, \mathrm{D}_{\mathrm{KL}}(q_\phi(z_L|x_L)||p(z)) \\ & + \mathbb{E}[\log p_\theta(x_U|z_U)] - \beta \, \mathrm{D}_{\mathrm{KL}}(q_\phi(z_U|x_U)||p(z))\end{aligned}$$

- **Task Classifier Cross Entropy Loss:** Predict correct class of data

- **Adversarial Loss of Discriminator:** Minimize distance between L and U

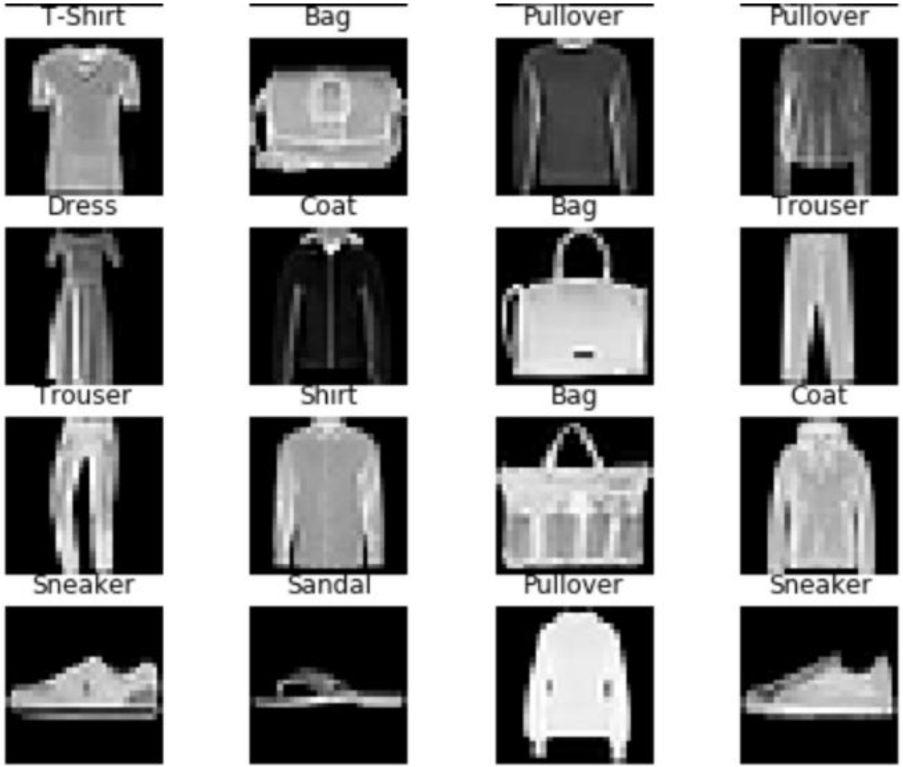
$$\mathcal{L}_{\mathrm{VAE}}^{adv} = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(D(q_\phi(z_U|x_U)))]$$

- **Objective function of the Discriminator:** Classify between L and U

$$\mathcal{L}_D = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(1 - D(q_\phi(z_U|x_U)))]$$

Dataset

- Fashion MNIST
- 55K training images
- 5K validation images
- 10K test images



Active Learning Algorithm

1. We start with initial labelled set of samples (X_L, Y_L), unlabelled samples (X_U)
2. We train the VAE, Classifier and Discriminator for 20 epochs.
3. We report test accuracy with the current Classifier model
4. Sample new set of labelled samples.
    - We choose set of samples which the discriminator predicts as unlabelled with highest confidence
5. Continue with step 2-4.

Initial Set of Results On Fashion MNIST

1000 images of data is: 84.91

2000 images of data is: 85.84

3000 images of data is: 86.58

4000 images of data is: 87.33

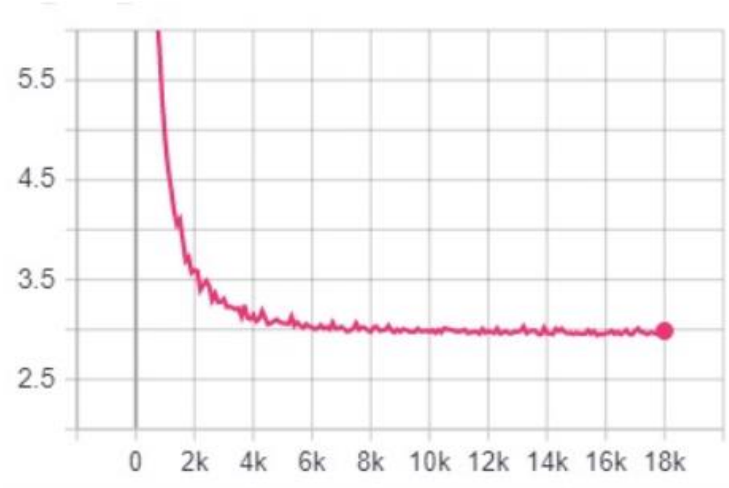
5000 images of data is: 88.66

6000 images of data is: 88.79

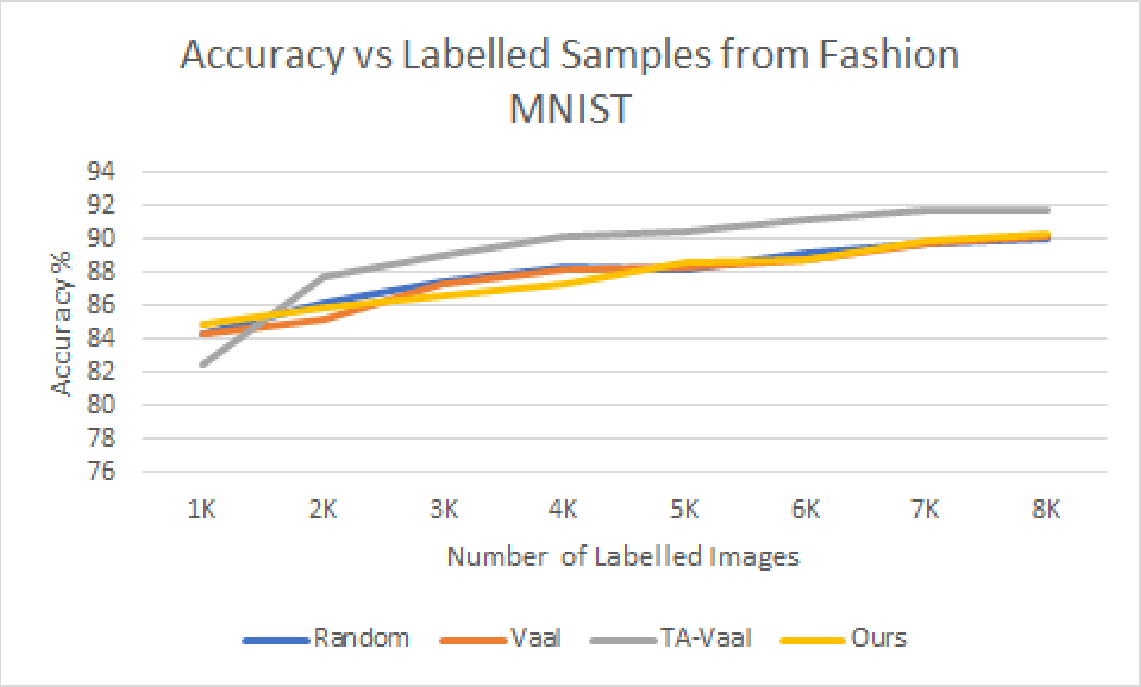
7000 images of data is: 89.94

8000 images of data is: 90.37

Decrease in Classifier Task Loss Across Iterations



Comparison to TA-VAAL, VAAL and Random Selection



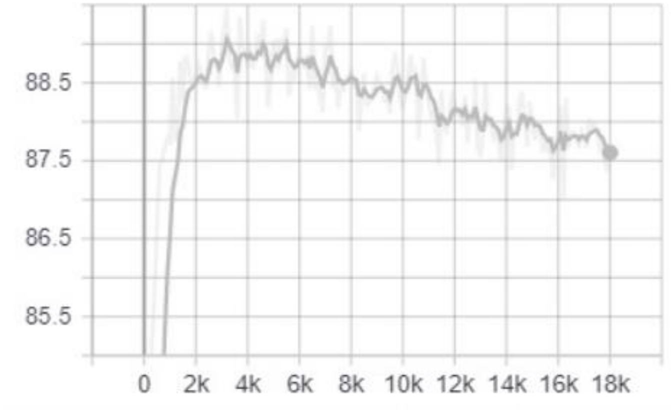
Task-aware Variational Adversarial Active Learning (TA-VAAL) *Kim, Kwanyoung, Dongwon Park, Kwang In Kim, and Se Young Chun (2020)*

Variational Adversarial Active Learning (VAAL) *Sinha, S., Ebrahimi, S., & Darrell, T. (2019)*

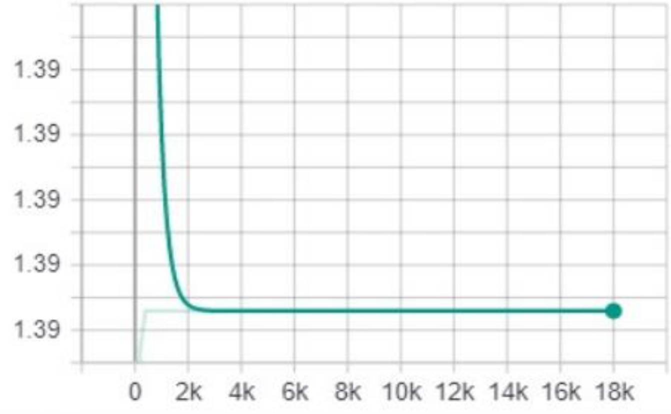
Why Discriminator of VAAL/ours performs poorly than random sampling?

Discriminator loss is constant at high accuracy

Test Accuracy across iterations



Discriminator Loss across iterations



- Discriminator selects samples at random
- At high accuracy, discriminator is unable to distinguish latent space of labelled and unlabelled data.

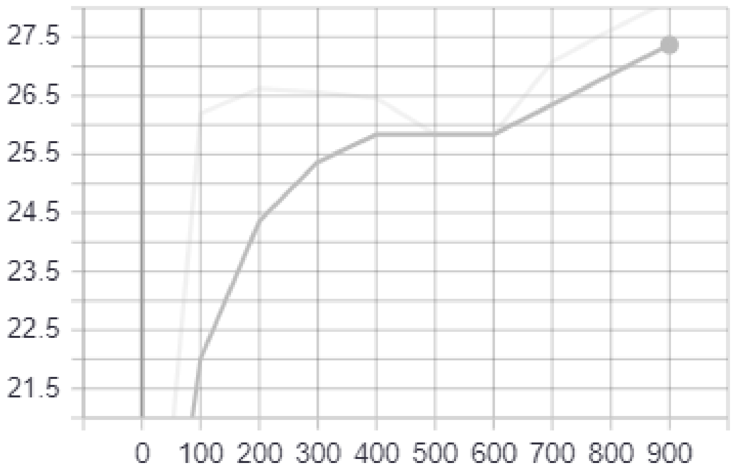
Starting with less labelled samples

- 500 samples containing only 3 classes out of 10 classes
  - {2: Pull Over, 4: Coat, 7: Sneaker}
- Sample very few new samples 64 samples of labelled data
- Stop when discriminator unable to differentiate between labelled and unlabelled data

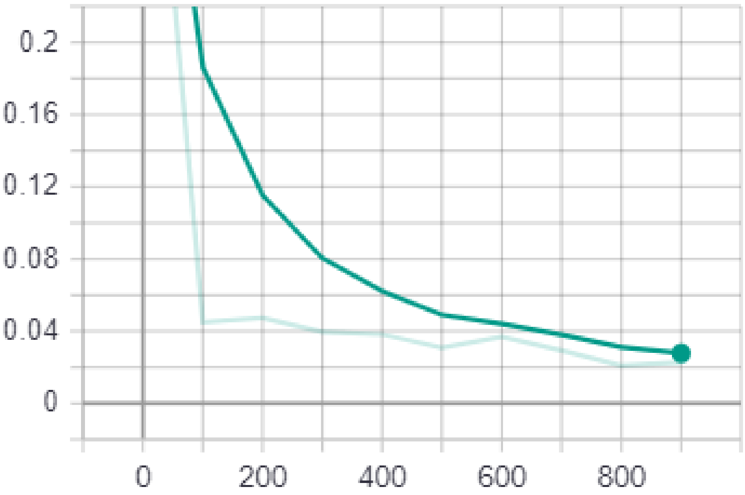
```
500  images of data is: 27.25
564  images of data is: 28.34
628  images of data is: 30.29
692  images of data is: 40.42
756  images of data is: 36.39
820  images of data is: 36.61
884  images of data is: 39.97
948  images of data is: 42.84
1012 images of data is: 48.52
1076 images of data is: 45.46
1140 images of data is: 52.14
1204 images of data is: 50.59
1268 images of data is: 50.82
1332 images of data is: 49.86
1396 images of data is: 52.44
```

Discriminator works with Lower Accuracy

Test Accuracy across iterations



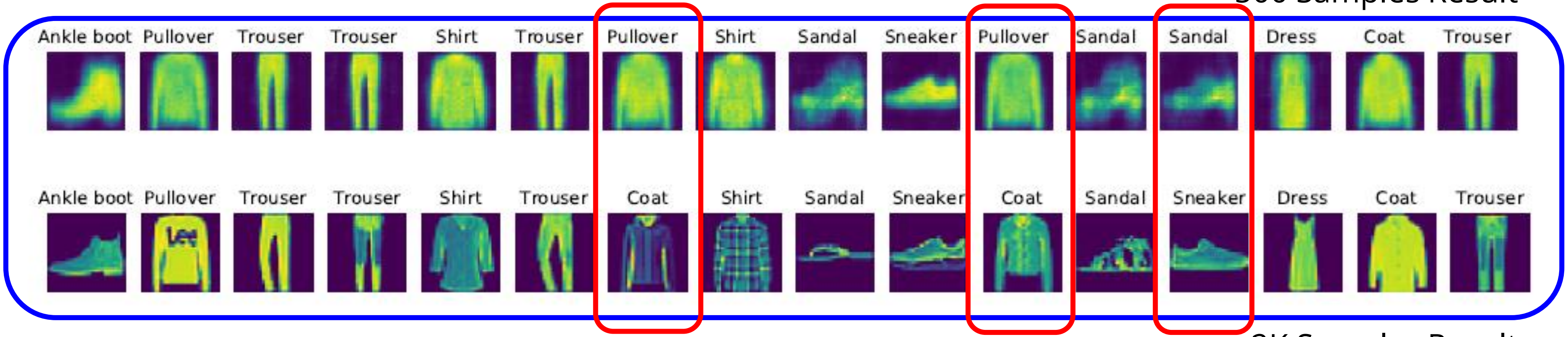
Discriminator Loss across iterations



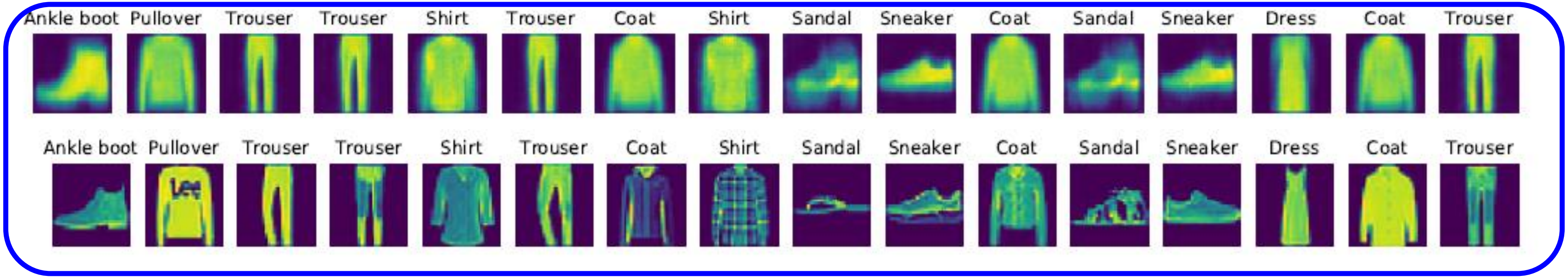
- Discriminator able to distinguish latent space of labelled and unlabelled data with lower test accuracy

Qualitative visualization of VAE performance

500 Samples Result


Ankle boot Pullover Trouser Trouser Shirt Trouser Pullover Shirt Sandal Sneaker Pullover Sandal Sandal Dress Coat Trouser
Ankle boot Pullover Trouser Trouser Shirt Trouser Coat Shirt Sandal Sneaker Coat Sandal Sneaker Dress Coat Trouser

2K Samples Result


Ankle boot Pullover Trouser Trouser Shirt Trouser Coat Shirt Sandal Sneaker Coat Sandal Sneaker Dress Coat Trouser
Ankle boot Pullover Trouser Trouser Shirt Trouser Coat Shirt Sandal Sneaker Coat Sandal Sneaker Dress Coat Trouser

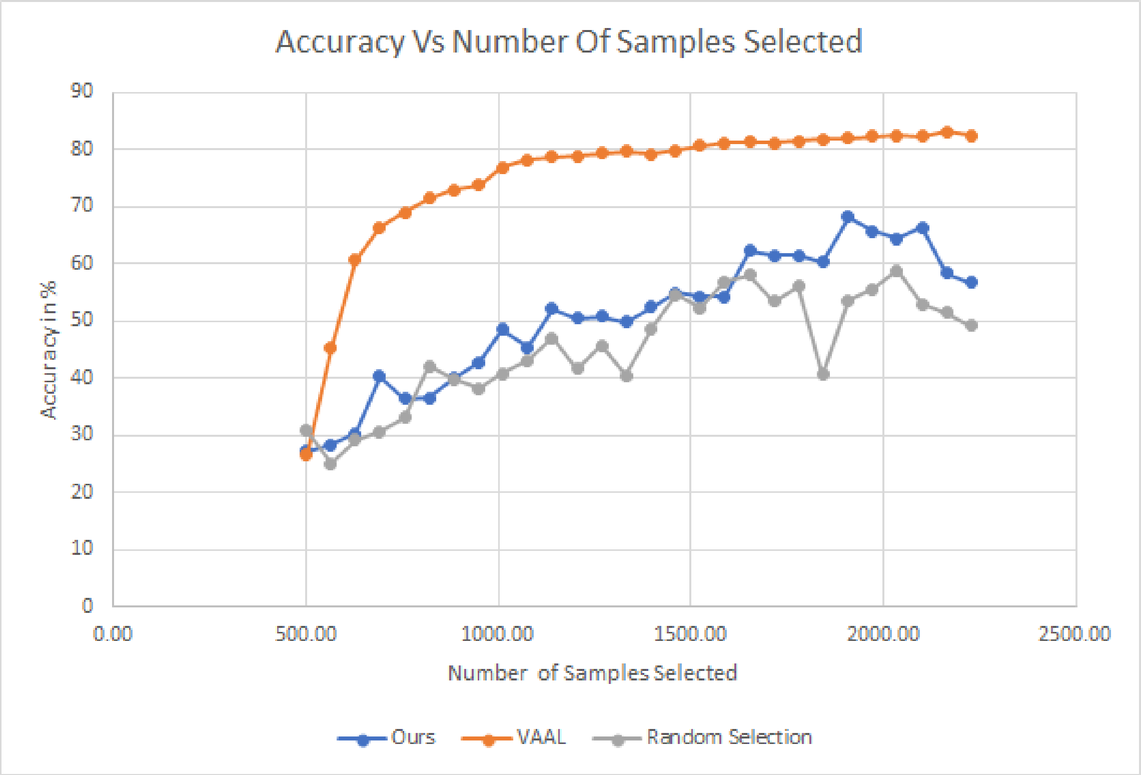
Discriminator Selection of Samples

1. Initial Budget: 500 Samples
   - [PullOver:166, Coat:167, Sneaker:167]

1. Next 64 Samples Budget
   - [T-Shirt:2, PullOver:4, Dress:11, Coat:7, Sandal:8, Skirt:13, Sneaker:5, Bag:10, Ankle-Boot:4]

1. Next 64 Samples Budget
   - [T-Shirt:1, PullOver:19, Dress:1, Coat:5, Sandal:14, Skirt:5, Bag:1, Ankle-Boot:18]

Less labelled samples: Comparison With Random and VAAL



- VAAL (VAE only) performed well with adversarial loss
- Our (M2-Model) degrades with adversarial loss

Reason

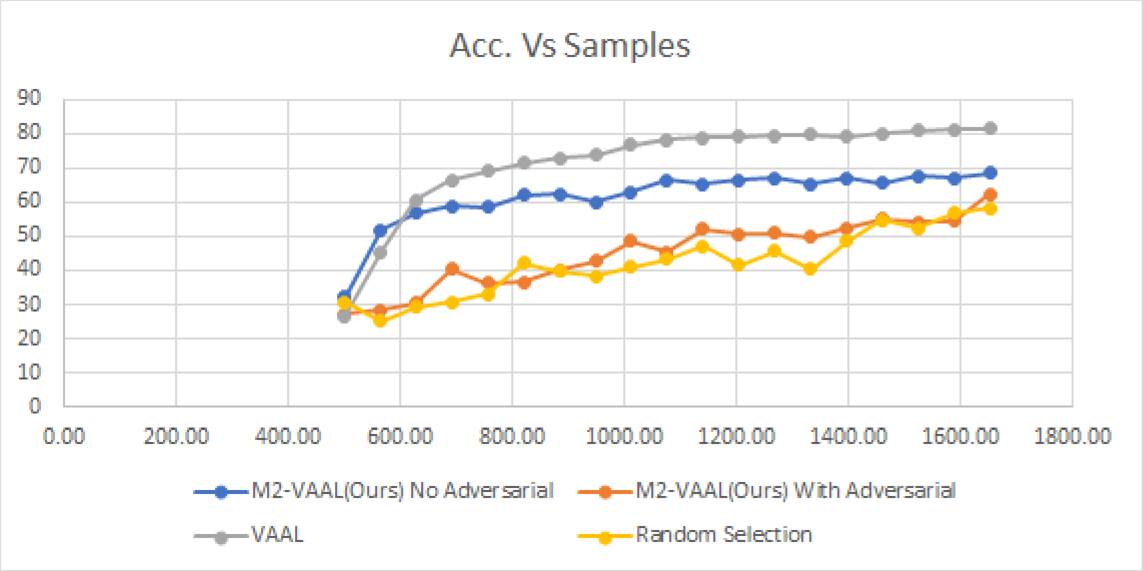
- Adversarial loss does orthogonal work to Task Classifier

$$\mathcal{L}_{\text{VAE}}^{adv} = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(D(q_\phi(z_U|x_U)))]$$

- It tries to make the Latent Space of Labelled Data equal to Latent Space of Unlabelled Data
- Our latent space consists of [Z, Y’], adv tries to equalize Y’ probabilities, thereby degrading performance of the classifier
- M2 model cannot be used with Adversarial Discriminative Loss

Using only the Classifier part of Discriminator

$$\mathcal{L}_D = -\mathbb{E}[\log(D(q_\phi(z_L|x_L)))] - \mathbb{E}[\log(1 - D(q_\phi(z_U|x_U)))]$$



- Accuracy without Adversarial Loss increased
- But, sample selection became less diverse, preventing it to reach maximum accuracy
- Discriminator model not suitable for sampling in M2 model

Increased Model Capacity

- We use a larger and more hyperparameter-tuned model with ~3M parameters to try and push the accuracy values.
- Dataset: Fashion-MNIST
- Training with 1000 labeled samples: *~85%* accuracy
- Training with 50 labeled samples: *~68%* accuracy
- We experiment in the *low labeled data size* regime to be able to observe more pronounced effect of *active* learning.

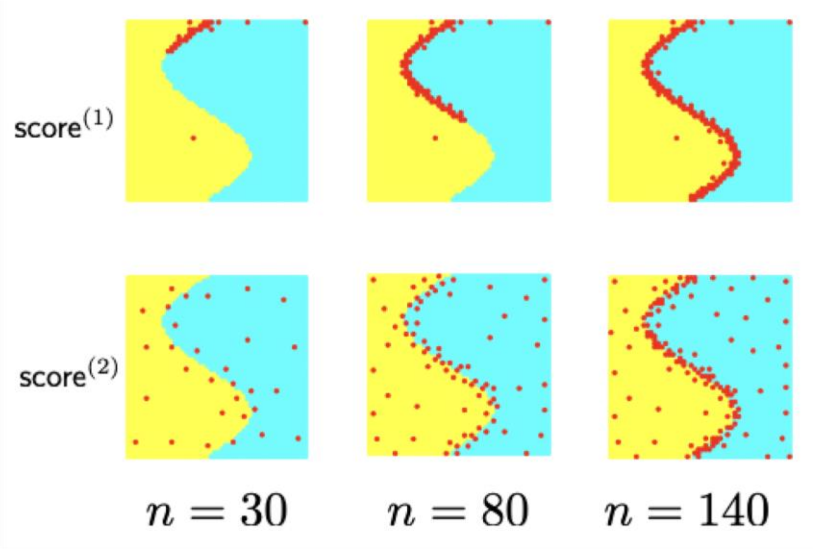
Datasets

- MNIST: Model accuracy very high
  - >95% with just 100 labeled samples
  - Not ideal for showing improvement using active learning
- CIFAR10 / CIFAR100: Input image size increases
  - Larger models required (Encoder + Decoder + Classifier)
  - Training time much longer
- Fashion-MNIST
  - Image dimensions same as MNIST
  - Accuracy not as high as MNIST
  - Good middle ground!

Acquisition Function

- On Fashion-MNIST dataset, we first look at ***norm-based losses***.

$$\text{score}^{(1)}(u) = \|f^u(x)\|\,, \qquad \text{score}^{(2)}(u) = \|f^u(x) - f(x)\|$$



Maximin Active Learning with Data-Dependent Norms - *Karzand, M., & Nowak, R. D. (2019)*

Acquisition Function

- The norm-based acquisition function requires computation of the norm of neural network weights
  - To understand which sample leads to maximum deviation in the weight norm
  - Notion of complexity of the sample w.r.t. the current model
- Extremely inefficient
  - for large scale datasets
  - for large models
- Need to look at other possibilities..!

Need for diversity

- Selecting a **single sample** for labeling at every iteration is inefficient.

- Need to select a **batch of samples** at every iteration.

- For each batch, need to select samples that are ***diverse***.

- Usual acquisition functions tend to ignore diversity!
  - For e.g., for the least confidence based sample selection, similar samples will tend to have similar prediction confidence. Hence, the method will end up selecting similar samples in a batch, leading to unnecessary querying of labels.

Proposed methods

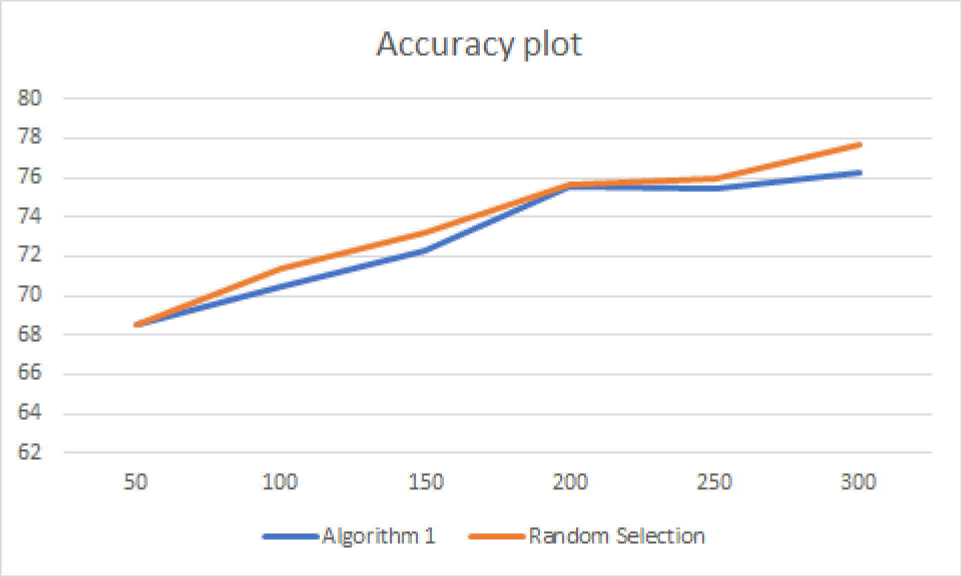
- **Algorithm 1:**

  For each class, for each unlabeled sample classified to that class, do until batch_size/num_classes samples are selected:

  1. Find the labeled data points which are correctly classified to the same class as an unlabeled sample (call them anchor points).
  2. Compute the sum of distances of the unlabeled sample to the anchor points in the latent space of the semi-supervised M2 model.
  3. Choose the unlabeled sample with the maximum distance, include in the set of anchor points, and repeat.

- Promotes diversity among samples by
  - Selecting samples from each class
  - Maximizing distances from *known* samples in a well-structured latent space

Algorithm 1



Problem

Ends up selecting samples from all classes, even if some of the classes need to be prioritized. Also, prediction confidence not taken into account.

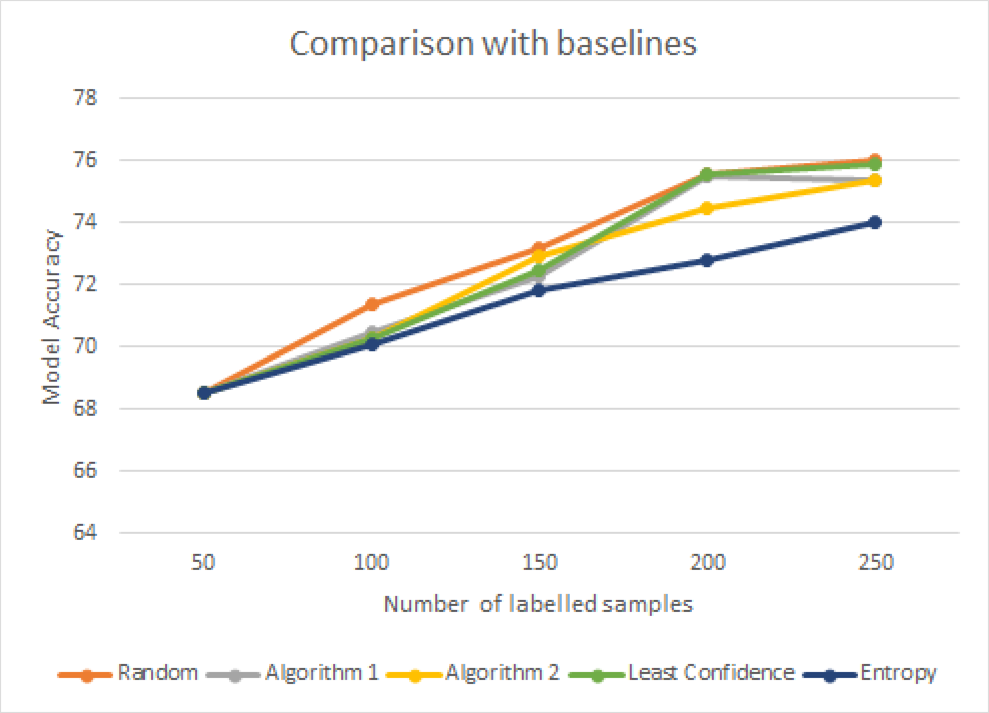
Proposed methods

- **Algorithm 2:**
  For each unlabeled sample, do:
  - Find the labeled data points which are correctly classified to the same class as an unlabeled sample.
  - Compute the *average sum of distances* of the unlabeled sample to these labeled samples in the latent space of the M2 model.
  - Choose the unlabeled samples with the minimum prediction confidence, weighted by normalised distance.

- Promotes diversity among samples by
  - Maximizing distances from *known* samples in a well-structured latent space.
  - Does not force sample selection from every class.

- However, accuracy values are comparable to the previous algorithm.

Comparison with Baselines



<u>Baselines</u>

1. Least confidence based
2. Maximum entropy based

Takeaways

- Recently proposed advanced acquisition functions do not necessarily apply in our case.
  - due to inefficiency when used in conjunction with deep neural networks, or
  - due to inherent properties of the semi-supervised VAE model.
- Proposed algorithms show similar performance as random selection.
  - *Silver lining*: Proposed methods tend to select diverse and informative samples, unlike random sampling.
  - Based on relation with the confusion matrices observed during training.

Thank you