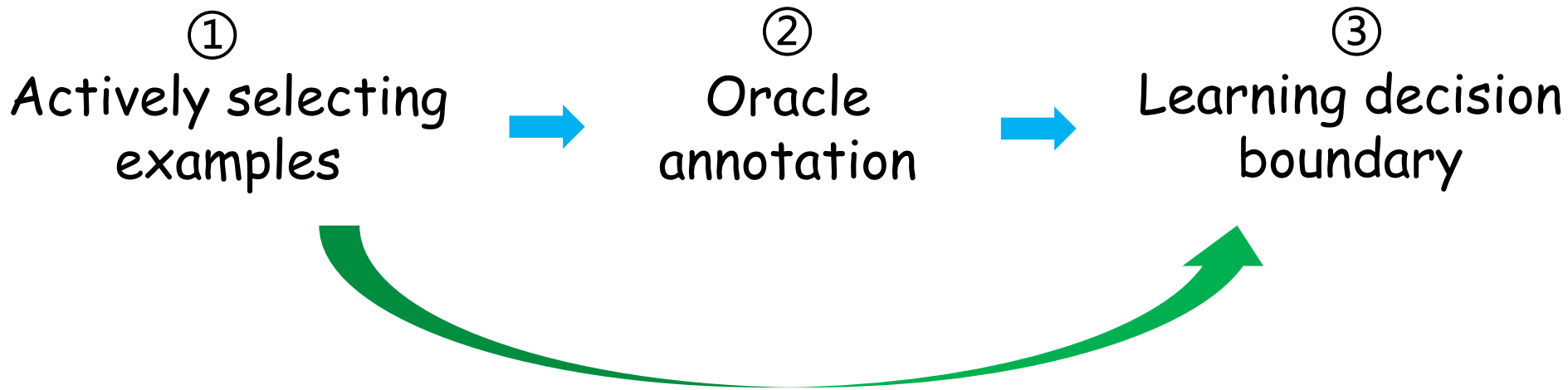
Active Decision Boundary Annotation with Deep Generative Models

Miriam Huijser
Aiir Innovations
Amsterdam, The Netherlands
https://aiir.nl/

Jan C. van Gemert
Delft University of Technology
Delft, The Netherlands
http://jvgemert.github.io/

Motivation

- Classical active learning



Can we skip step ② for a more efficient annotation ?

Active Decision Boundary Annotation

Problem formulation

N data samples $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ $y \in \{-1, 1\}$

Initial set $\mathcal{A}$ contains a handful of annotated images

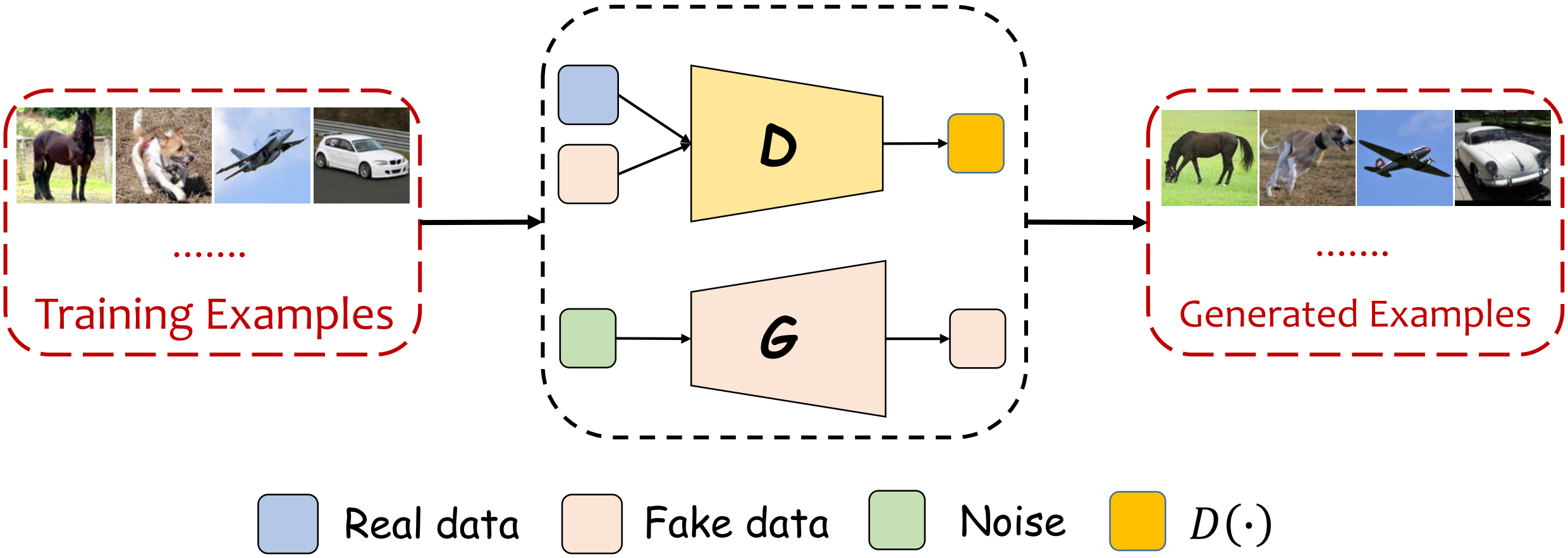
Each data sample x_i has a corresponding latent variable $\mathbf{z}_i \in \mathbb{R}^K$

Ω: The tightest hypersphere that contains all z_i

In each active learning iteration, we estimate a decision boundary $\hat{\theta}$.

The goal is to best approximate the real decision boundary θ while minimizing the number of iteration.

Generative Adversarial Nets



Active Decision Boundary Annotation

Method

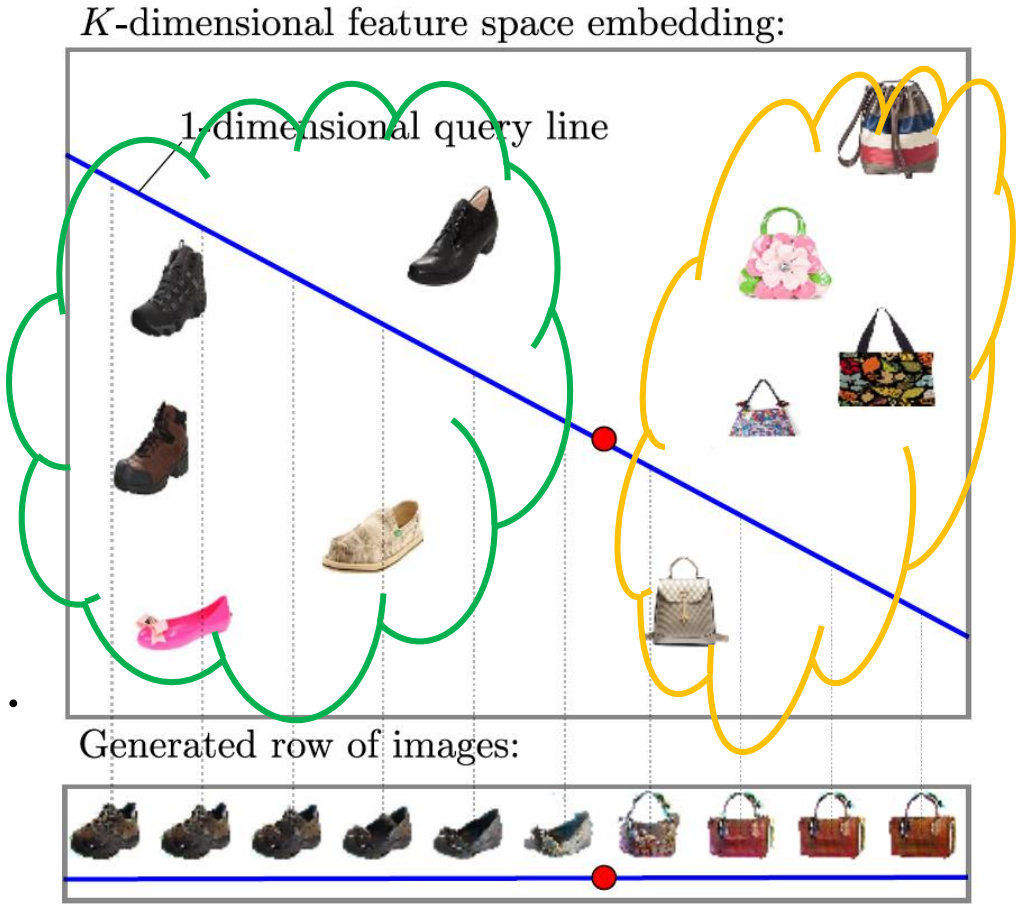
Utilize the power of deep generative models for active learning.

I. Use all unlabeled images to learn a K-dimensional embedding.

II. Select a 1-d query line and employ a GAN to generate visual samples along this line.

III. Annotate the point between two samples of different classes as the decision boundary and use it to improve the classification model.

Repeat step II and step III



Active Decision Boundary Annotation

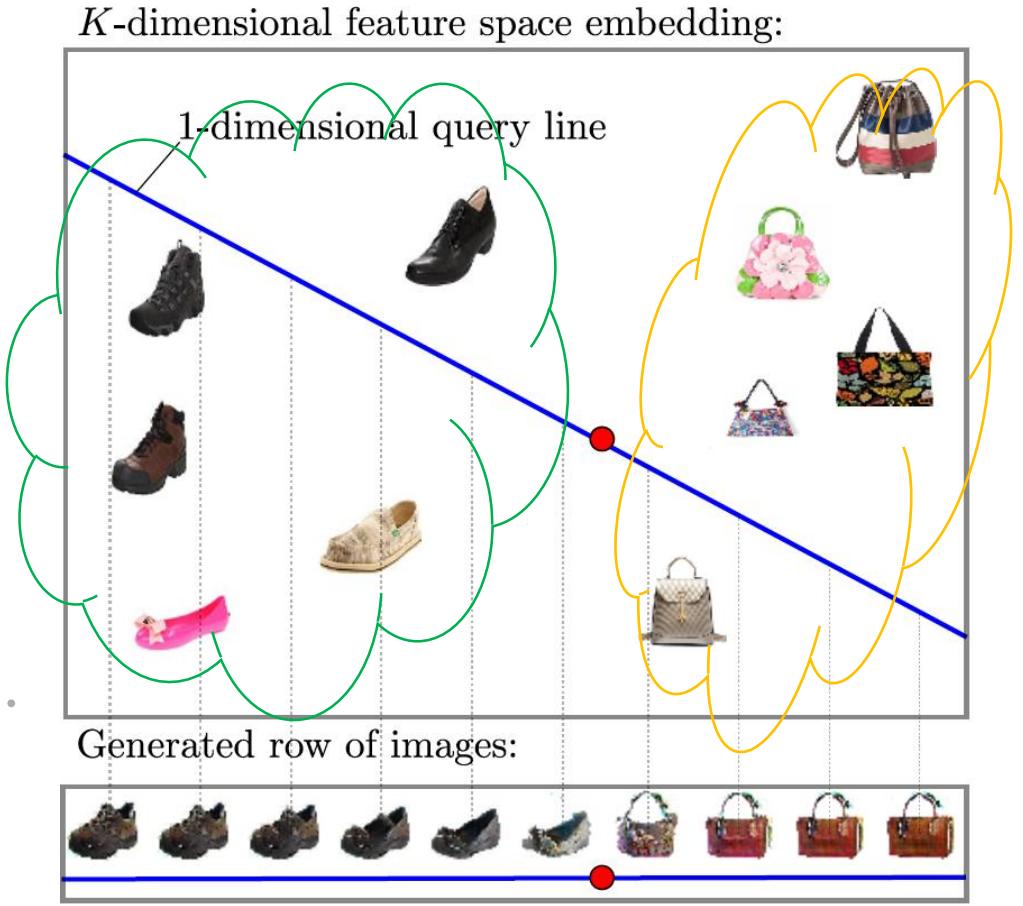
Method

Utilize the power of deep generative models for active learning.

I. Use all unlabeled images to learn a K-dimensional embedding.

II. Select a 1-d query line and employ a GAN to generate visual samples along this line.

III. Annotate the point between two samples of different classes as the decision boundary and use it to improve the classification model.



Active Decision Boundary Annotation

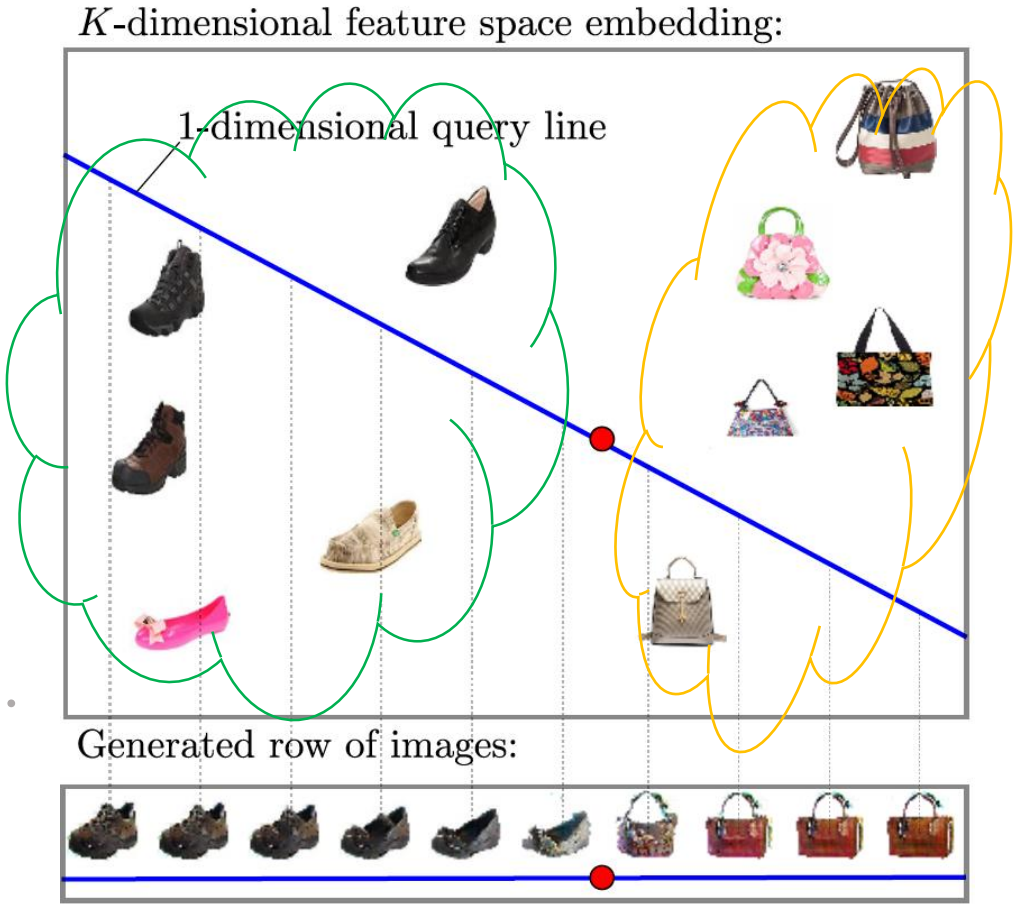
Method

Utilize the power of deep generative models for active learning.

I. Use all unlabeled images to learn a K-dimensional embedding.

II. Select a 1-d query line and employ a GAN to generate visual samples along this line.

III. Annotate the point between two samples of different classes as the decision boundary and use it to improve the classification model.



Implement Detail

Constructing the query line

- Uncertainty sampling

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} 1 - P_{\hat{\theta}}(\hat{y}|\mathbf{z})$$

Principle: Generated images along the query line q should undergo a class-change.

Ensure q is perpendicular to the current estimated decision boundary $\hat{\theta}$

Solution: $\mathbf{q}(t) = \mathbf{z}^p + (\mathbf{z}^* - \mathbf{z}^p)t$

where $\mathbf{z}^p = \mathbf{z}^* - \frac{(\hat{\mathbf{w}}^\intercal \mathbf{z}^* + \hat{b})}{\hat{\mathbf{w}}^\intercal \hat{\mathbf{w}}} \hat{\mathbf{w}}$

Implement Detail

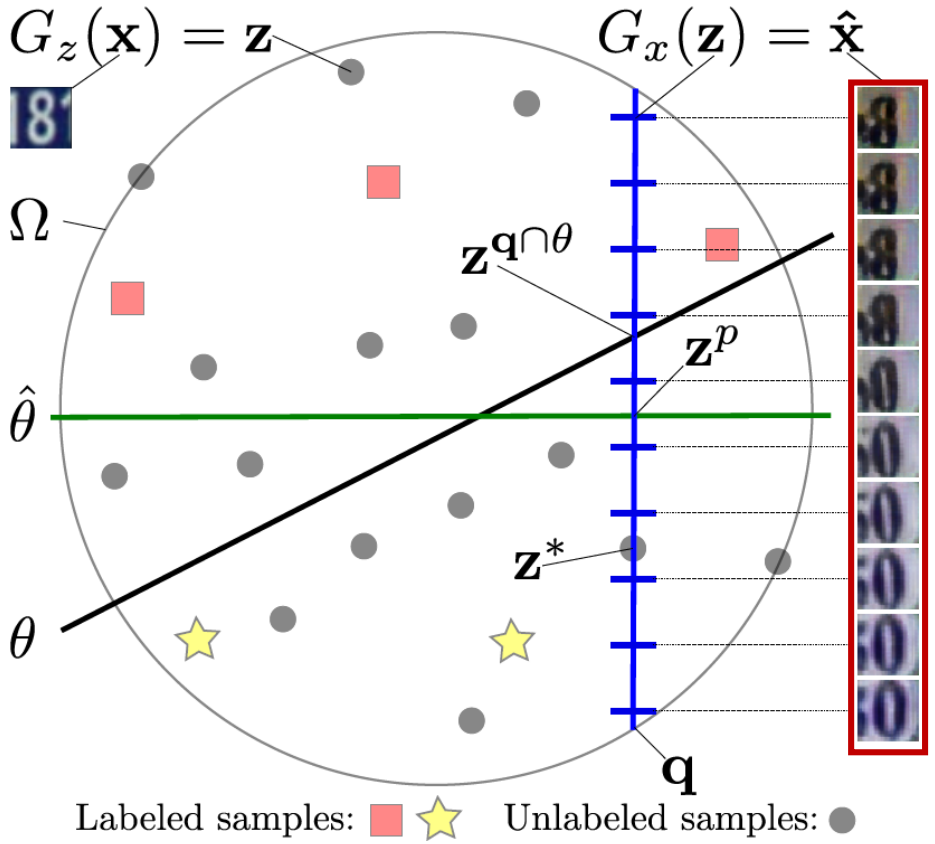
Construct the image row

Decide which image samples along the query line q to show to oracle.

I. Restrict the size of q lie within the tightest hypersphere Ω

$$\mathcal{H} = \{\mathbf{z} \in \mathbb{R}^K : ||\mathbf{z} - \bar{\mathbf{z}}|| \leq r\} \text{ where } r = \max_{i=0}^{N} ||\mathbf{z_i} - \bar{\mathbf{z}}||$$

II. Uniformly sample s latent samples from $\mathbf{q} \cap \mathcal{H}$ and decode the images as $G(\mathbf{z}) = \hat{\mathbf{x}}$.



Active Decision Boundary Annotation

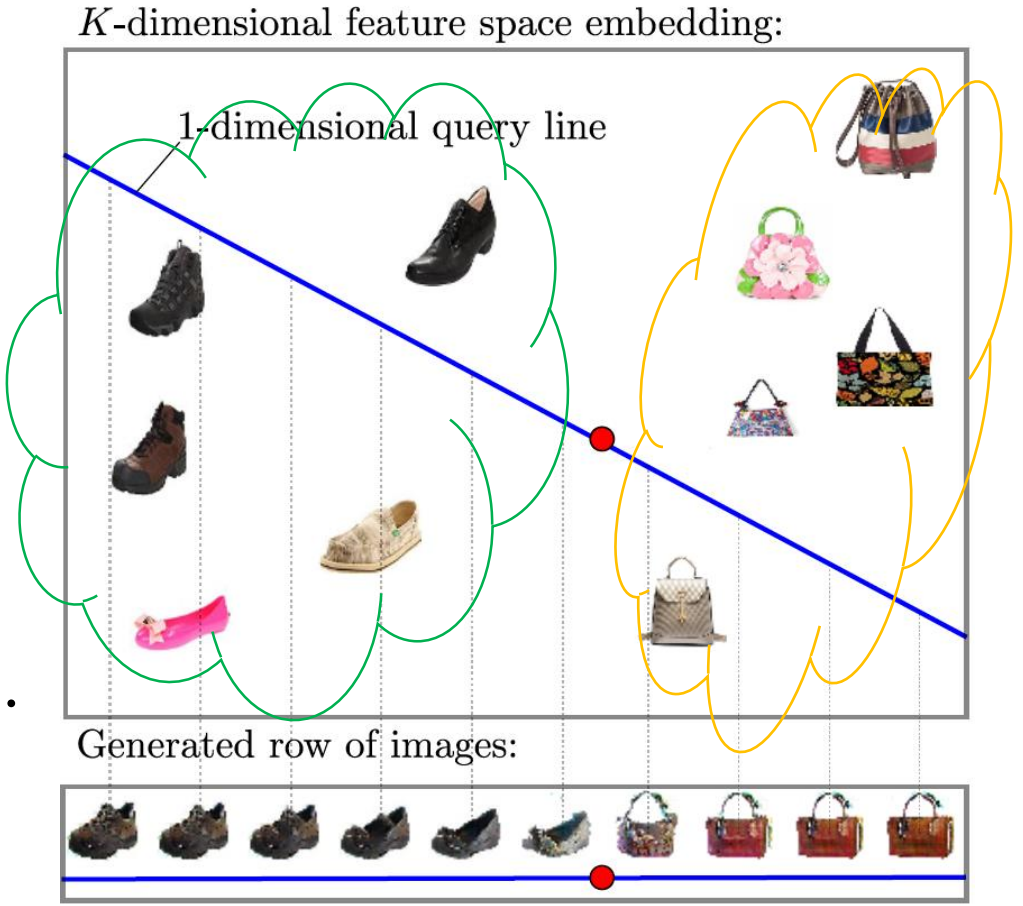
Method

Utilize the power of deep generative models for active learning.

I. Use all unlabeled images to learn a K-dimensional embedding.

II. Select a 1-d query line and employ a GAN to generate visual samples along this line.

III. Annotate the point between two samples of different classes as the decision boundary and use it to improve the classification model.



Implement Detail

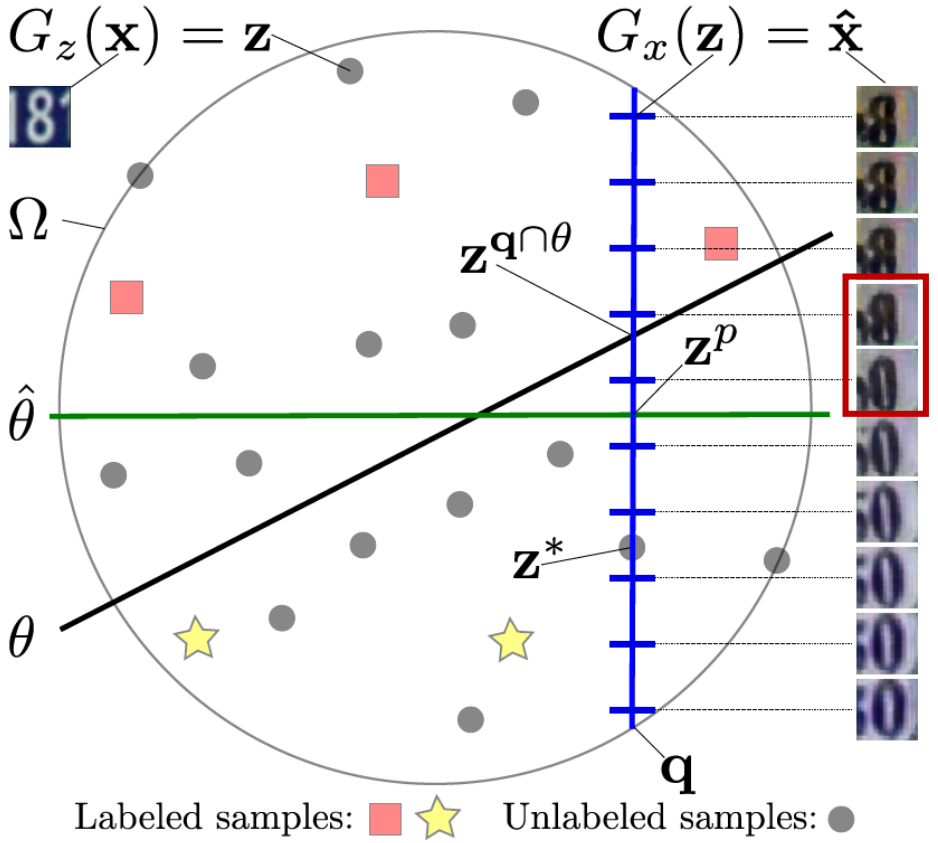
Annotating the decision boundary

- ✓ Human oracle
- ✓ oracle-classifier (trained on ground-truth labels)

After annotating

$\mathcal{B}$: stores the decision boundary annotations $\mathbf{z}^{\mathbf{q}\cap\theta} = \mathbf{q}\cap\theta$

$\mathcal{A}$: stores the pair $(\mathbf{z}^*, y)$



Implement Detail

Model optimization using boundary annotation

Update $\hat{\theta}$ using both labeled sample and decision boundary annotations.

- For labeled samples in $\mathcal{A}$

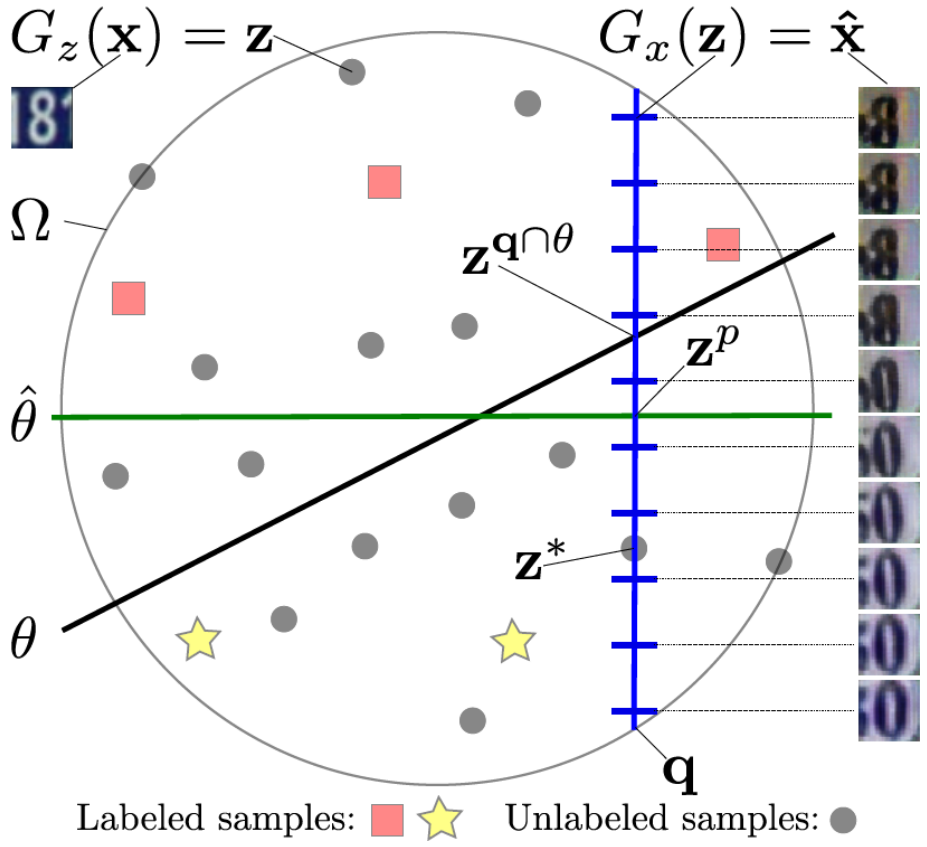
$$\mathcal{L}_{\text{class}} = \frac{1}{|\mathcal{A}|} \sum_{(\mathbf{z},y)\in\mathcal{A}} \max\left(0, 1 - y(\hat{\mathbf{w}}^{\intercal}\mathbf{z} + \hat{b})\right)$$

- For decision boundary annotations in $\mathcal{B}$

$$\mathcal{L}_{\text{regress}} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{z}\in\mathcal{B}} \left(\hat{\mathbf{w}}^{\intercal}\mathbf{z} + \hat{b}\right)^2$$

- The final loss:

$$\mathcal{L} = \frac{1}{2}\mathcal{L}_{\text{class}} + \frac{1}{2}\mathcal{L}_{\text{regress}} + \lambda||\hat{\mathbf{w}}||^2$$



Experiment: configurations

- Data sets
  - MNIST, 50k for training and 10k for testing.
  - SVHN
  - Shoe-Bag
- Evaluation metric
  - Area Under the Learning Curve (AULC)

$$\mathrm{AULC} = \sum_{i=1}^{N} \frac{1}{2}(\mathrm{acc}_{i-1} + \mathrm{acc}_i)$$

Experiment: Evaluating various query strategies

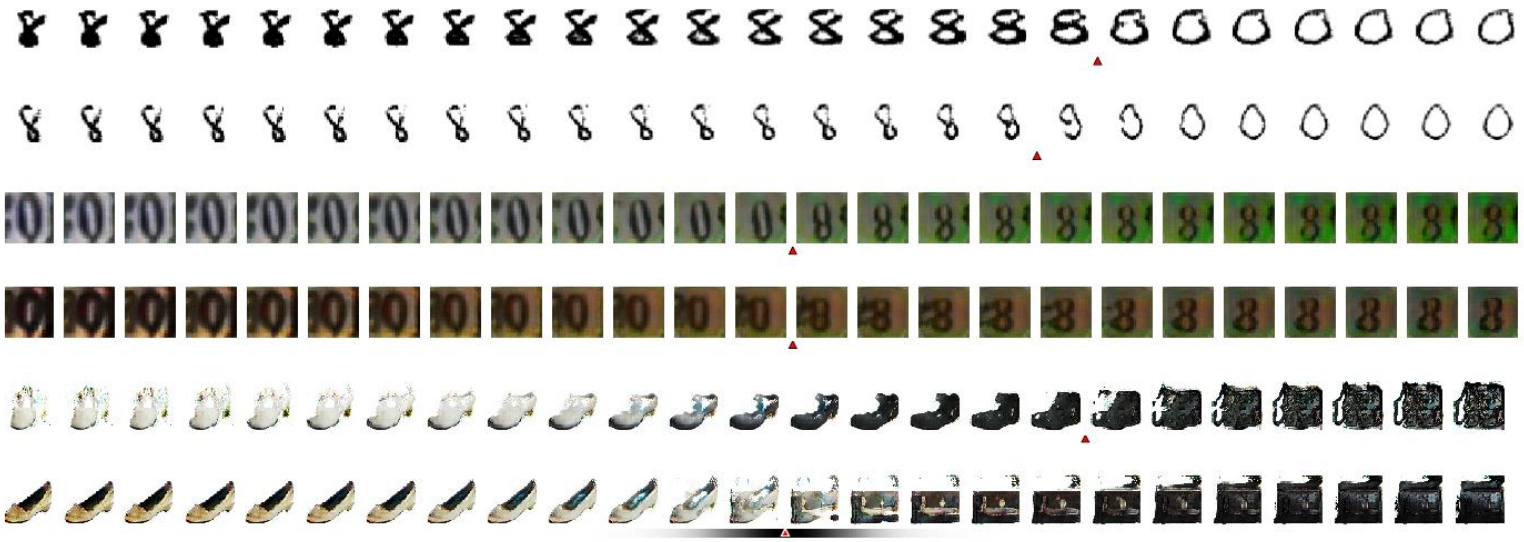
Four sample-based query strategies

- Random
- Uncertainty
- Uncertainty-dense
- K-cluster centroid

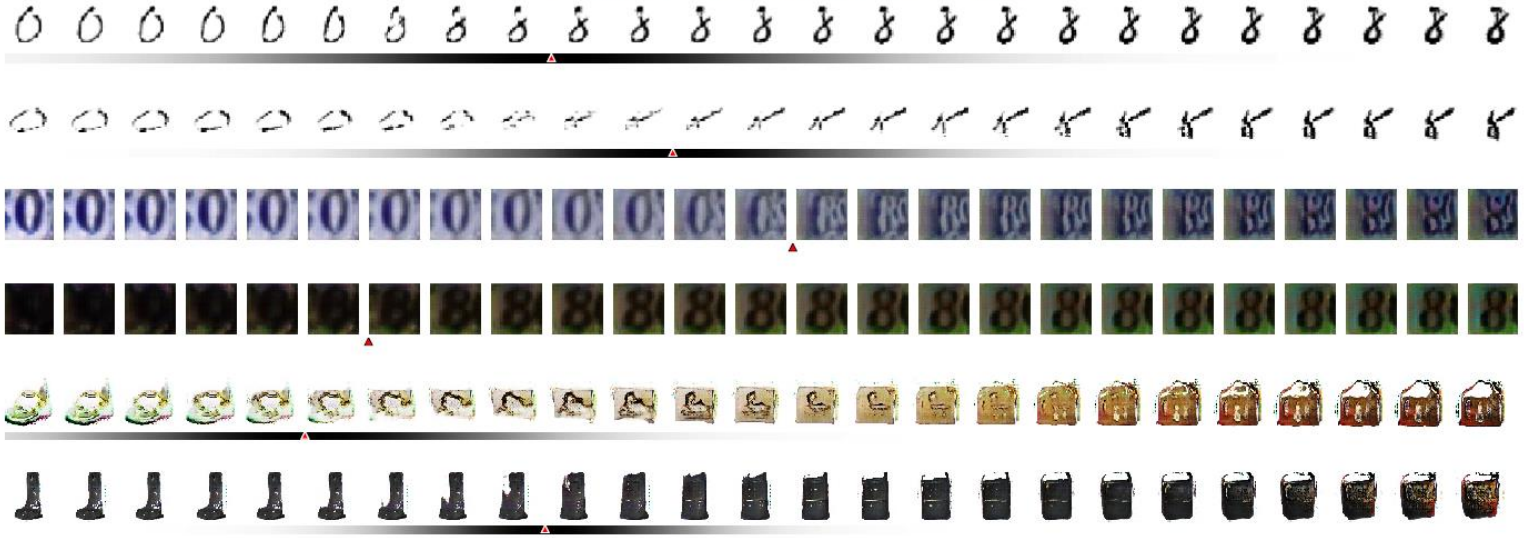
	MNIST 0 vs. 8		SVHN 0 vs. 8		Shoe-Bag	
Strategy	Sample	Boundary (ours)	Sample	Boundary (ours)	Sample	Boundary (ours)
Uncertainty	144.0 ± 0.5	$\mathbf{145.8 \pm 0.4}$	118.7 ± 1.3	$\mathbf{124.3 \pm 1.0}$	143.2 ± 0.6	$\mathbf{145.4 \pm 0.5}$
Uncertainty-dense	135.6 ± 10.5	$\mathbf{142.0 \pm 10.8}$	99.6 ± 5.8	$\mathbf{116.8 \pm 2.5}$	112.0 ± 6.6	$\mathbf{135.2 \pm 3.0}$
5 Cluster centroid	141.7 ± 0.4	$\mathbf{145.0 \pm 0.3}$	98.0 ± 4.9	$\mathbf{106.3 \pm 1.6}$	131.0 ± 1.6	$\mathbf{143.7 \pm 0.3}$
Random	142.2 ± 1.0	$\mathbf{145.1 \pm 0.5}$	116.2 ± 1.9	$\mathbf{124.7 \pm 1.1}$	140.5 ± 1.1	$\mathbf{145.0 \pm 0.4}$

Table 1: AULC results for four active learning query strategies. Results are on MNIST (classifying 0 and 8), SVHN (classifying 0 and 8) and Shoe-Bag after 150 queries, where the maximum possible AULC score is 150. The results are averaged over 15 repetitions. For each row, the significantly best result is shown in bold, where significance is measured with a paired t-test with $p < 0.05$. SVHN is the most difficult dataset. Uncertainty sampling is generally the best query strategy. Boundary annotation significantly outperforms sample annotations for all datasets for all query strategies.

Experiment: Evaluating generative model quality



(a) Query lines with high human consistency.



(b) Query lines with low human consistency.

	lines without change	samples deviation
MNIST 0 vs. 8	2	4
SVHN 0 vs. 8	1	1
Shoe-Bag	5	9

Table 2: Annotation consistency results averaged over 10 query line annotations from 10 human oracles. We show the number of lines marked as having no class change and the average deviation in number of images, rounded up, from the average annotation per line. Human consistency is worse for the non-uniform Shoe-Bag dataset. The more uniform datasets MNIST and SVHN have quite accurate human consistency.

Experiment: Evaluating annotation noise

Sampling noise (# images)	MNIST 0 vs. 8		SVHN 0 vs. 8		Shoe-Bag	
	Sample	Boundary (ours)	Sample	Boundary (ours)	Sample	Boundary (ours)
0	144.2 ± 0.5	$\mathbf{146.0 \pm 0.3}$	119.1 ± 1.5	$\mathbf{124.0 \pm 0.9}$	143.1 ± 0.6	$\mathbf{145.4 \pm 0.5}$
1	144.2 ± 0.5	$\mathbf{145.9 \pm 0.3}$	119.1 ± 1.5	$\mathbf{123.4 \pm 1.1}$	143.1 ± 0.6	$\mathbf{145.2 \pm 0.4}$
2	144.2 ± 0.5	$\mathbf{145.4 \pm 0.5}$	119.1 ± 1.5	$\mathbf{121.4 \pm 2.1}$	143.1 ± 0.6	$\mathbf{144.7 \pm 0.9}$
3	144.2 ± 0.5	$\mathbf{145.0 \pm 0.4}$	119.1 ± 1.5	$\mathbf{121.1 \pm 1.2}$	143.1 ± 0.6	$\mathbf{144.5 \pm 0.7}$
4	144.2 ± 0.5	144.2 ± 0.4	119.1 ± 1.5	119.1 ± 0.9	143.1 ± 0.6	$\mathbf{143.9 \pm 0.5}$
5	$\mathbf{144.2 \pm 0.5}$	143.6 ± 0.4	119.1 ± 1.5	113.6 ± 10.7	143.1 ± 0.6	143.0 ± 0.7

Table 3: AULC results for noisy boundary active learning with uncertainty sampling for MNIST (classifying 0 and 8), SVHN (classifying 0 and 8) and Handbags vs. Shoes after 150 queries (maximum possible score is 150). Each experiment is repeated 15 times. For each row, the significantly best result is shown in bold, where significance is measured with a paired t-test with $p < 0.05$. Noise has been added to the boundary annotation points; not to the image labels. Results worsen with more added noise, with the turning point of the significant better performance of Boundary around a sampling noise of 4 images for MNIST and SVHN, and 5 images for Shoe-Bag.

Experiment: Evaluating a human oracle

Experiment 4: Evaluating a human oracle

	MNIST 0 vs. 8		SVHN 0 vs. 8		Shoe-Bag	
Annotation	Sample	Boundary (ours)	Sample	Boundary (ours)	Sample	Boundary (ours)
Human oracle	8.5 ± 0.7	$\mathbf{8.8 \pm 0.3}$	5.7 ± 0.4	5.8 ± 0.4	8.1 ± 0.5	8.2 ± 0.4
SVM oracle	8.9 ± 0.3	$\mathbf{9.1 \pm 0.3}$	6.3 ± 0.4	6.4 ± 0.4	8.7 ± 0.3	8.8 ± 0.4

Table 4: AULC results for a human and a SVM oracle for sample-based active learning and our boundary active learning for MNIST (classifying 0 and 8), SVHN (classifying 0 and 8) and Shoe-Bag after 10 queries (maximum possible score is 10). The experiments are repeated 15 times and significant results per row are shown in bold for $p < 0.05$. Results always improve for boundary annotation, but these improvements are not significant for SVHN and Shoe-Bag.

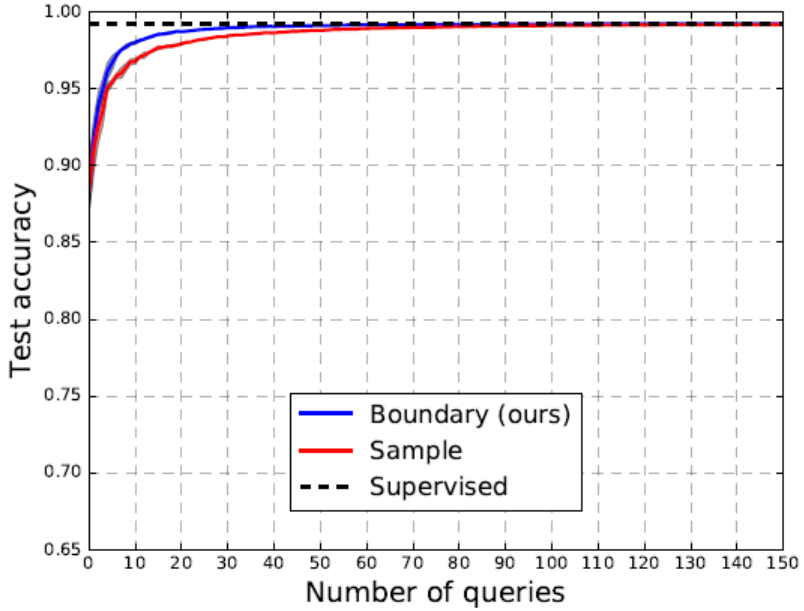
Experiment: Generalization over classes

Experiment 5: Full dataset evaluation

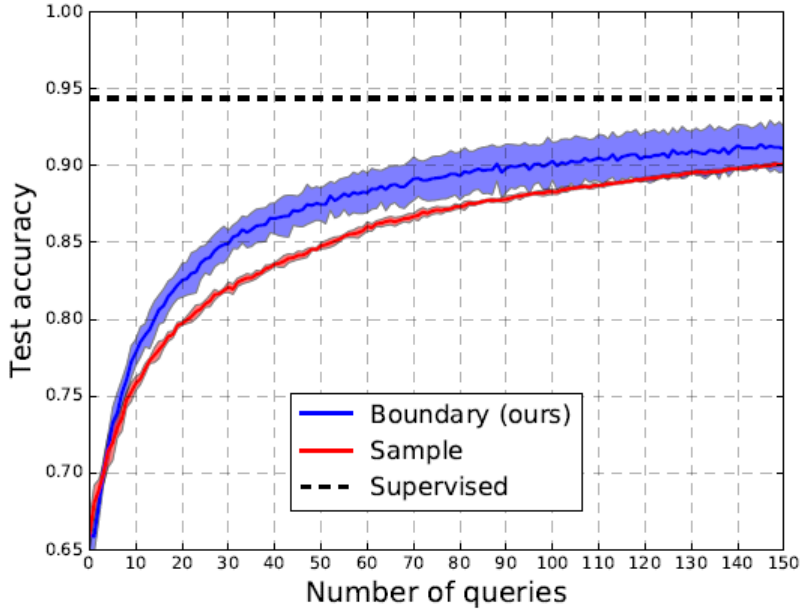
	Sample	Boundary (ours)
MNIST	147.8 ± 0.06	$\mathbf{148.3 \pm 0.04}$
SVHN	127.8 ± 0.2	$\mathbf{130.9 \pm 1.9}$
Shoe-Bag	143.2 ± 0.6	$\mathbf{145.4 \pm 0.5}$

Table 5: AULC results for sample-based active learning and boundary active learning for all datasets after 150 queries (maximum possible score is 150), averaged over all class pairs. The experiments are repeated 5 times and significant results are shown in bold. Significance is measured with a paired t-test with $p < 0.05$. For all datasets our method significantly improves over sample-based active learning.

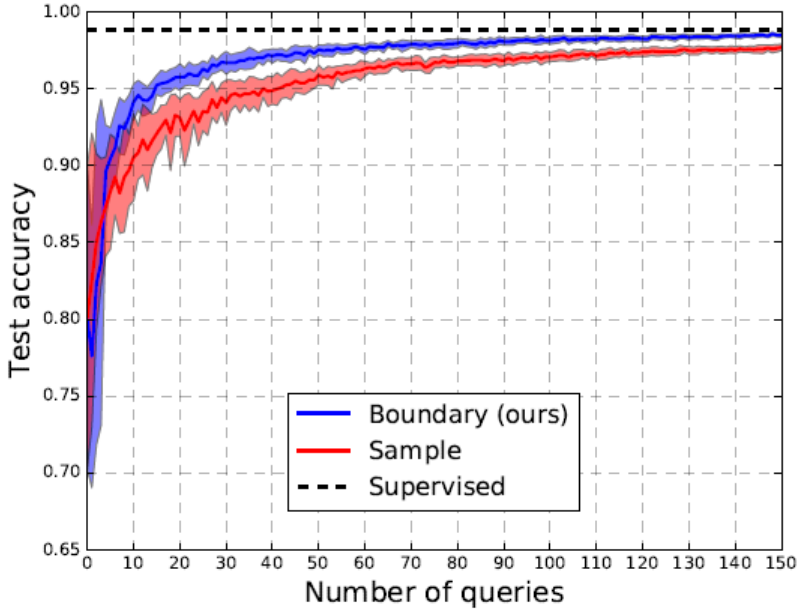
Experiment: Generalization over classes



(a) MNIST averaged over all classes.



(b) SVHN averaged over all classes.



(c) Shoe-Bag both classes.

Figure 4: Learning curves over all datasets and all class pairs using uncertainty sampling as query strategy. The experiments are repeated 5 times, standard deviations are indicated by line width. The fully supervised oracle-SVM is the upper bound. Our boundary method outperforms the sample-based method.

Thanks