OTH
Amberg-Weiden

Advanced Topics in Machine Learning

Winter Semester 2024/2025

Prof. Dr.-Ing. Christian Bergler | OTH Amberg-Weiden

OTH Amberg-Weiden

Topics of Today: Advanced Deep Learning Strategies – Part I

- Representation Learning
- Transfer Learning
- Distillation Learning
- Deep Metric Learning (Focus: Contrastive Learning)

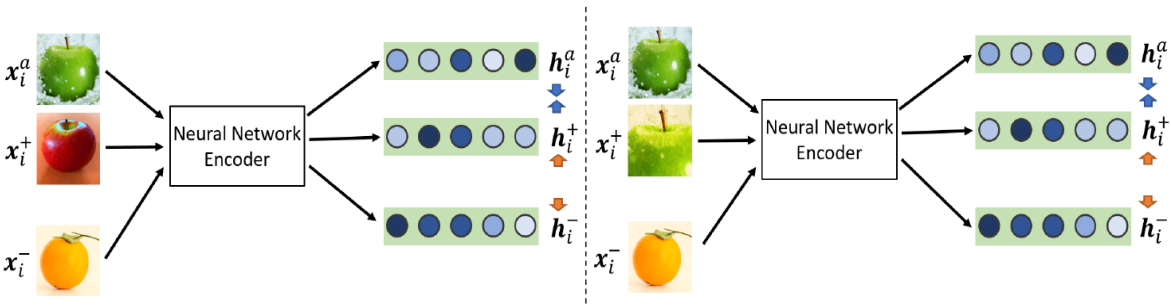
Topics of Today: Advanced Deep Learning Strategies – Part II

- Self-Supervised Learning
- Active Learning

Deep Learning Paradigms – Part I

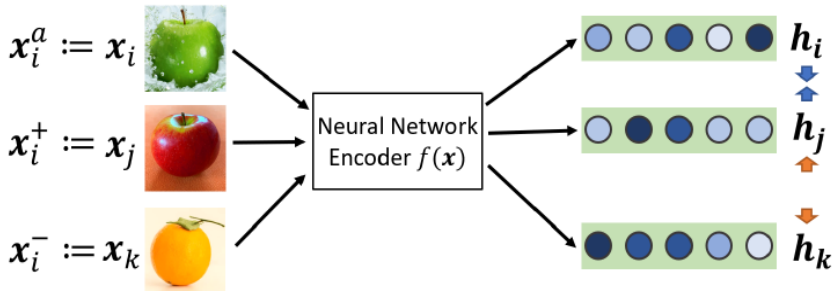
Contrastive Learning – Supervised VS. Unsupervised

From Supervised To Self-Supervised...



Source: Images from FAU Erlangen-Nuremberg, Pattern Recognition Lab, K.Breininger, V. Christlein, Advanced Deep Learning – Representation Learning

From Supervised To Self-Supervised...



$$L(x_i^a, x_i^+, x_i^-) = max(0, \epsilon + ||f(x_i^a) - f(x_i^+)||_2^2 - ||f(x_i^a) - f(x_i^-)||_2^2)$$

$$L(x_i^a, x_i^+, x_i^-) = max(0, \epsilon + ||h_i - h_i^+||_2^2 - ||h_i - h_i^-)||_2^2)$$

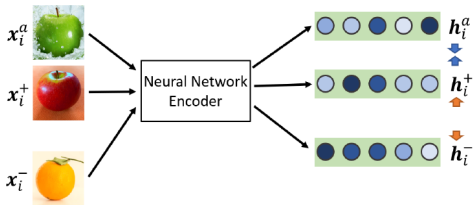
Source: Images from FAU Erlangen-Nuremberg, Pattern Recognition Lab, K.Breininger, V. Christlein, Advanced Deep Learning – Representation Learning

Deep Learning Paradigms – Part I

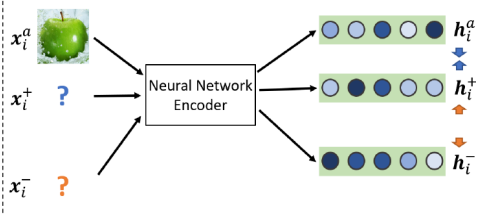
Contrastive Learning – Supervised VS. Unsupervised

From Supervised To Self-Supervised...

Supervised contrastive learning:



Unsupervised contrastive learning:



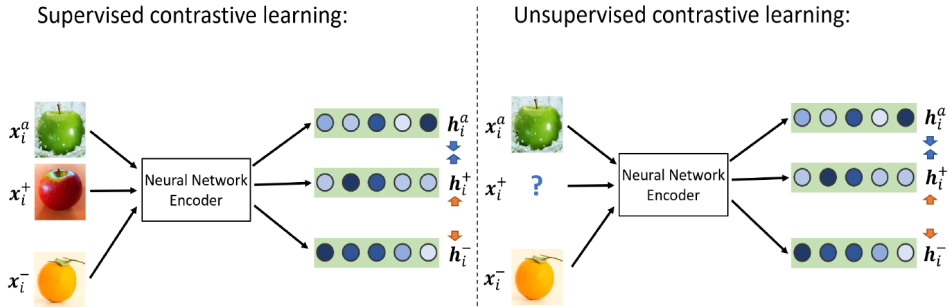
- How to find positive (x_i^+) and negative (x_i^-) samples in case the data is unlabeled?

Source: Images from FAU Erlangen-Nuremberg, Pattern Recognition Lab, K.Breininger, V. Christlein, Advanced Deep Learning – Representation Learning

Deep Learning Paradigms – Part I

Contrastive Learning – Supervised VS. Unsupervised

From Supervised To Self-Supervised...



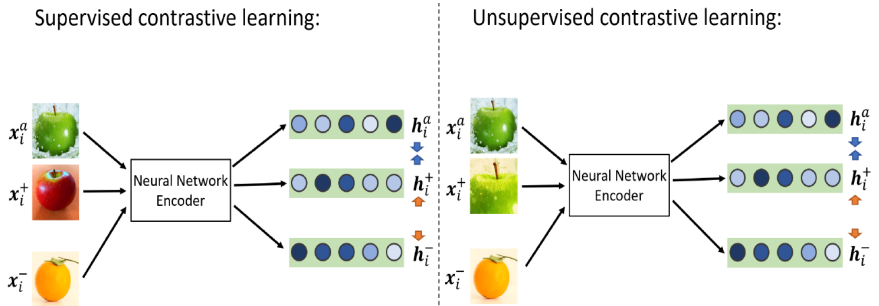
- Negative samples (x_i^-) are easy to identify! How about hard negatives?
- Use of (unsupervised) distance-based similarity measurements (clustering)!

Source: Images from FAU Erlangen-Nuremberg, Pattern Recognition Lab, K.Breininger, V. Christlein, Advanced Deep Learning – Representation Learning

Deep Learning Paradigms – Part I

Contrastive Learning – Supervised VS. Unsupervised

From Supervised To Self-Supervised...



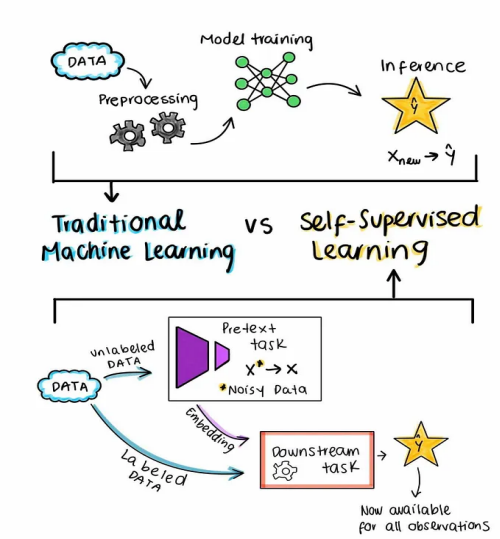
- How to find positive samples (x_i^+)? Are there any?
- Either: use of (unsupervised) distance-based similarity measurements (clustering)
- Or: Augmentation (use anchor x_i^a to create positive pair)!

Source: Images from FAU Erlangen-Nuremberg, Pattern Recognition Lab, K.Breininger, V. Christlein, Advanced Deep Learning – Representation Learning

Deep Learning Paradigms – Part I

Self-Supervised Learning

- Definition: form of unsupervised learning, where the mode learns patterns from the data itself without the need of labeled samples → Possibility to automatically generate large-scale data corpora

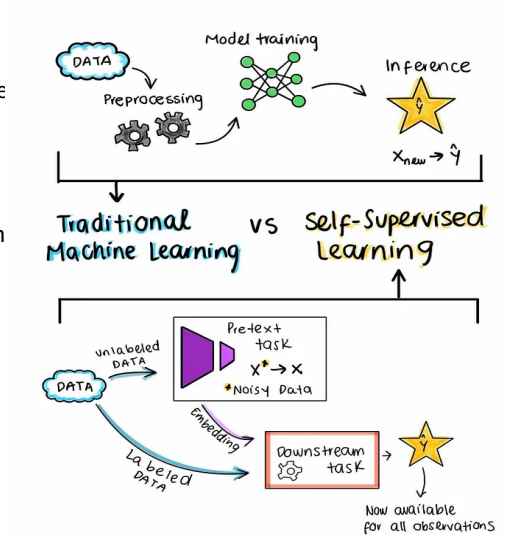


Source: Image from https://engineering.rappi.com/self-supervised-learning-making-the-most-of-large-scale-unlabelled-data-3931a41c84dc

Deep Learning Paradigms – Part I

Self-Supervised Learning

- Definition: form of unsupervised learning, where the mode learns patterns from the data itself without the need of labeled samples → Possibility to automatically generate large-scale data corpora
- Motivation: supervised learning is costly & time-consumin SSL aims to reduce the need by creating training pairs (pseudo labels) between input & ground truth

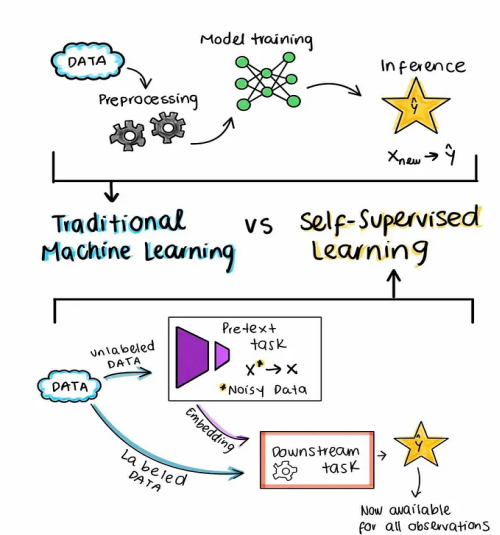


Source: Image from https://engineering.rappi.com/self-supervised-learning-making-the-most-of-large-scale-unlabelled-data-3931a41c84dc

Deep Learning Paradigms – Part I

Self-Supervised Learning

- Definition: form of unsupervised learning, where the mode learns patterns from the data itself without the need of labeled samples → Possibility to automatically generate large-scale data corpora
- Motivation: supervised learning is costly & time-consumin SSL aims to reduce the need by creating training pairs (pseudo labels) between input & ground truth
- Core Concepts: pretext task, contrastive learning, masked modeling, clustering-based, generative

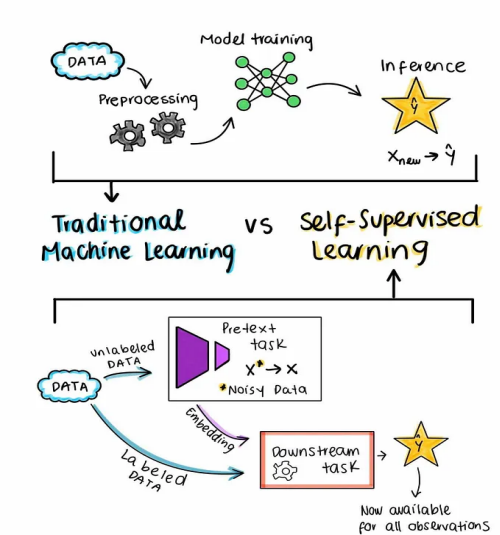


Source: Image from https://engineering.rappi.com/self-supervised-learning-making-the-most-of-large-scale-unlabelled-data-3931a41c84dc

Deep Learning Paradigms – Part I

Self-Supervised Learning

- Definition: form of unsupervised learning, where the mode learns patterns from the data itself without the need of labeled samples → Possibility to automatically generate large-scale data corpora
- Motivation: supervised learning is costly & time-consumin SSL aims to reduce the need by creating training pairs (pseudo labels) between input & ground truth
- Core Concepts: pretext task, contrastive learning, masked modeling, clustering-based, generative
- SSL Review/Survey Studies Concepts:
  - Jie Gui, A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends
  - Utku Ozbulak, Know Your Self-supervised Learning: A Survey Image-based Generative and Discriminative Training

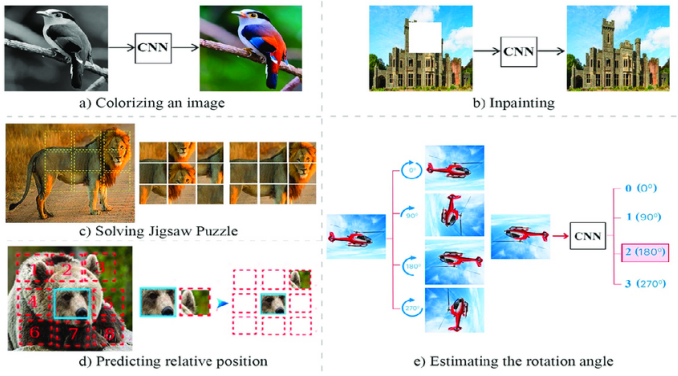


Source: Image from https://engineering.rappi.com/self-supervised-learning-making-the-most-of-large-scale-unlabelled-data-3931a41c84dc

Deep Learning Paradigms – Part I

Self-Supervised Learning

Pretext Task



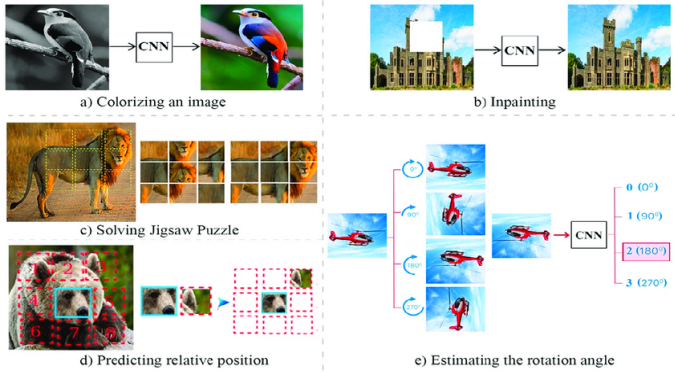
- Generate pseudo-labels from the data (images) itself

Source: Image from Saleh Ali Albelwi, "Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging"

Deep Learning Paradigms – Part I

Self-Supervised Learning

Pretext Task

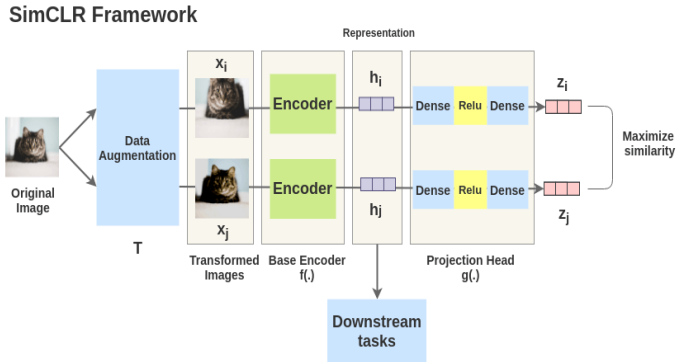


- Generate pseudo-labels from the data (images) itself
- Solving pretext tasks allows the model to extract useful latent representations that later improve the downstream tasks (colorization, rotation prediction, inpainting, puzzle-based, relative positioning, ...)

Source: Image from Saleh Ali Albelwi, "Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging"

OTH Amberg-Weiden

Contrastive Learning

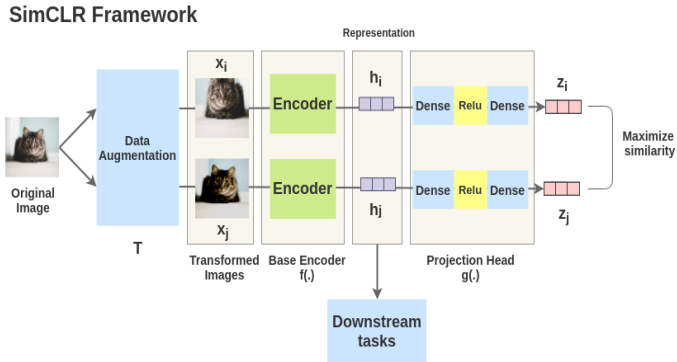


- Ting Chen et al., A simple framework for contrastive learning of visual representations (SimCLR), here: GitHub-SimCLR, in addition to visual explanation

Source: Image from Saleh Ali Albelwi, "Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging"

OTH Amberg-Weiden

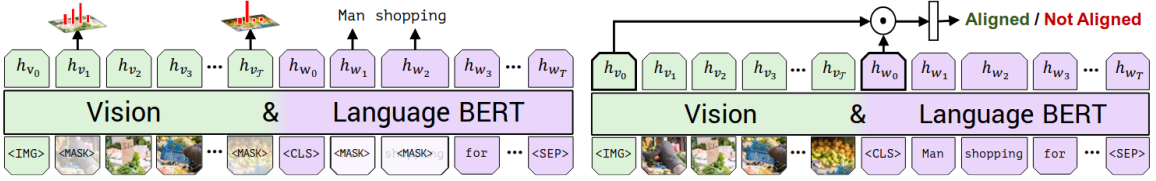
Contrastive Learning



- Ting Chen et al., A simple framework for contrastive learning of visual representations (SimCLR), here: GitHub-SimCLR, in addition to visual explanation
- Data augmentation, encoder, projection head, contrastive loss (normalized temperature-scaled cross-entropy loss (NT-Xent))

Source: Image from Saleh Ali Albelwi, "Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging"

Masked-Based Learning

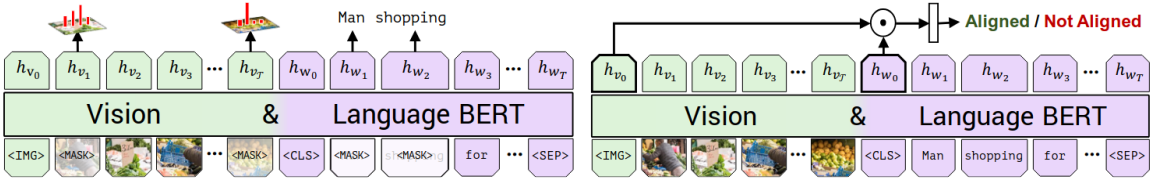


(a) Masked multi-modal learning (b) Multi-modal alignment prediction

- Predicts a part of the data based on another, using dependencies within data (vision- and/or text-based masking – example: ViLBERT (Vision-and-Language BERT)

Source: Jiasen Lu et al.: "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks"

Masked-Based Learning

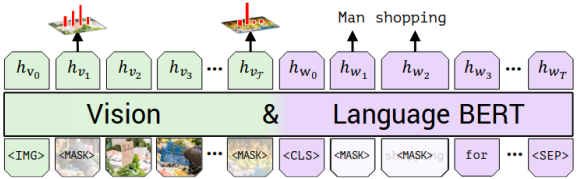


(a) Masked multi-modal learning (b) Multi-modal alignment prediction

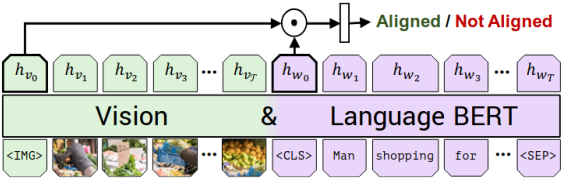
- Predicts a part of the data based on another, using dependencies within data (vision- and/or text-based masking – example: ViLBERT (Vision-and-Language BERT)
- Semantic understanding needs to be learned by the model in order to be able to reconstruct missing contextual content, irrespective of the data

Source: Jiasen Lu et al.: "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks"

Masked-Based Learning



(a) Masked multi-modal learning

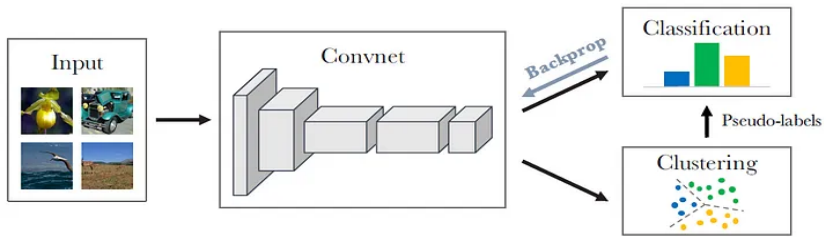


(b) Multi-modal alignment prediction

- Predicts a part of the data based on another, using dependencies within data (vision- and/or text-based masking – example: ViLBERT (Vision-and-Language BERT)
- Semantic understanding needs to be learned by the model in order to be able to reconstruct missing contextual content, irrespective of the data
- Randomly showing the model different, but not the entire, input (trade-off: masking (out) content w.o. distorting the actual signal too much)

Source: Jiasen Lu et al.: "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks"

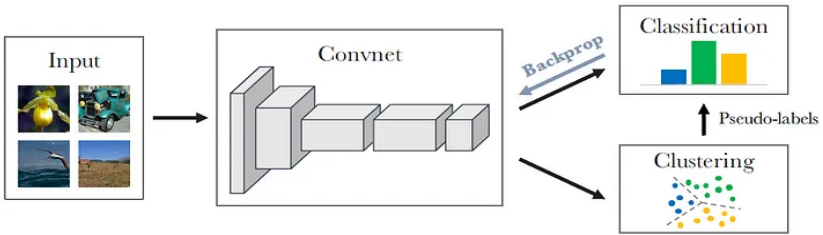
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Generate useful feature representations by grouping similar samples together

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

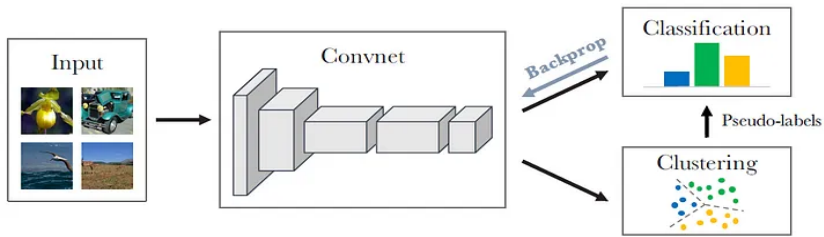
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Generate useful feature representations by grouping similar samples together
- General Approach: clustering the data (e.g., using K-means), assigning pseudo-labels to each cluster content, supervised training via cluster as class assignments

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

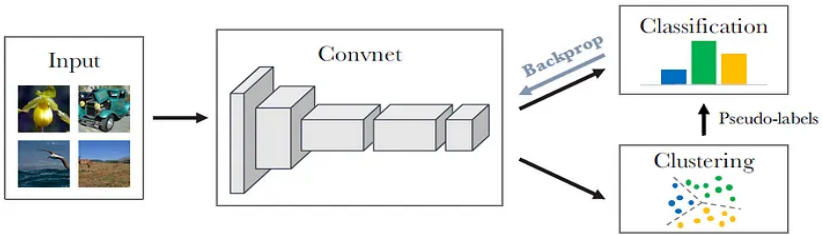
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Generate useful feature representations by grouping similar samples together
- General Approach: clustering the data (e.g., using K-means), assigning pseudo-labels to each cluster content, supervised training via cluster as class assignments
- Criterion: $\min_{\theta,\phi} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(g_\phi(f_\theta(x_n), y_n)$, with $\mathcal{L}$ as the negative log-softmax function

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

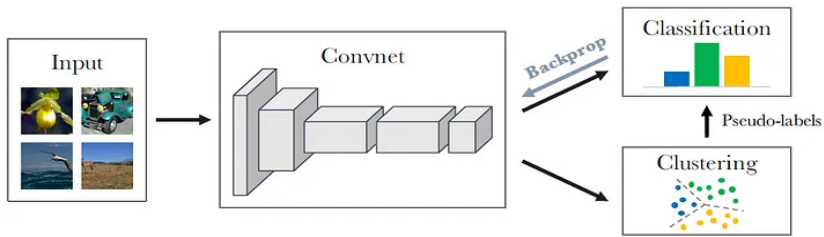
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Generate useful feature representations by grouping similar samples together
- General Approach: clustering the data (e.g., using K-means), assigning pseudo-labels to each cluster content, supervised training via cluster as class assignments
- Criterion: $\min_{\theta,\phi} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(g_\phi(f_\theta(x_n), y_n)$, with $\mathcal{L}$ as the negative log-softmax function
- Feedback Loop – better features lead to more coherent clusters, and better clusters guide the model to refine its features in the next iteration

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

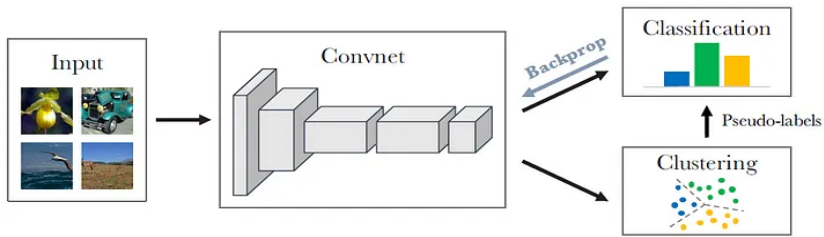
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Traditional clustering techniques (e.g. k-means) use the latent feature vectors, produced by $f_\theta(x_n)$ & clusters them into k (fixed) distinct groups using a geometric criterion

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

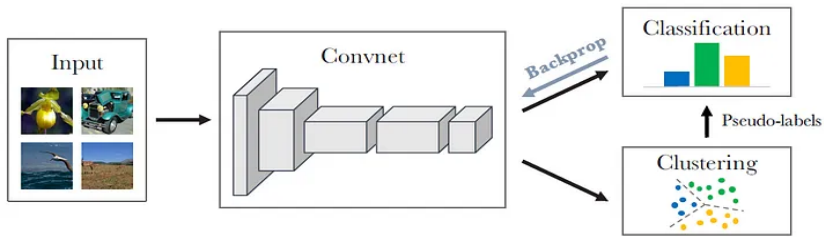
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Traditional clustering techniques (e.g. k-means) use the latent feature vectors, produced by $f_\theta(x_n)$ & clusters them into k (fixed) distinct groups using a geometric criterion
- Jointly learns a $d \times k$ centroid matrix C (d=embedding size, k=number of cluster centers) and the respective cluster assignments y_n

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

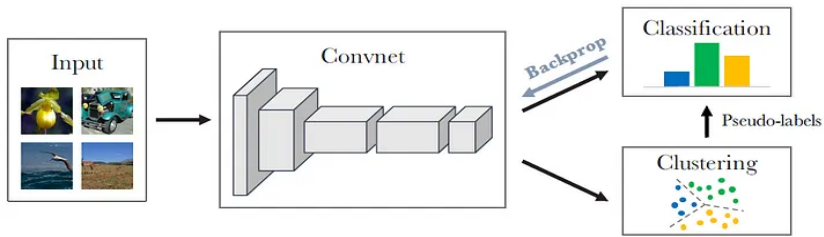
Clustering-Based Learning (Deep Clustering, Different Architectural Designs)



- Traditional clustering techniques (e.g. k-means) use the latent feature vectors, produced by $f_\theta(x_n)$ & clusters them into k (fixed) distinct groups using a geometric criterion
- Jointly learns a $d \times k$ centroid matrix C (d=embedding size, k=number of cluster centers) and the respective cluster assignments y_n
- $\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} min_{y_n \in \{0,1\}^k} ||f_\theta(x_n) - Cy_n||_2^2$, with $y_n^T 1_k = 1$ (one cluster per sample)

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

Clustering-Based Learning (Deep Clustering, Different Architectural Designs)

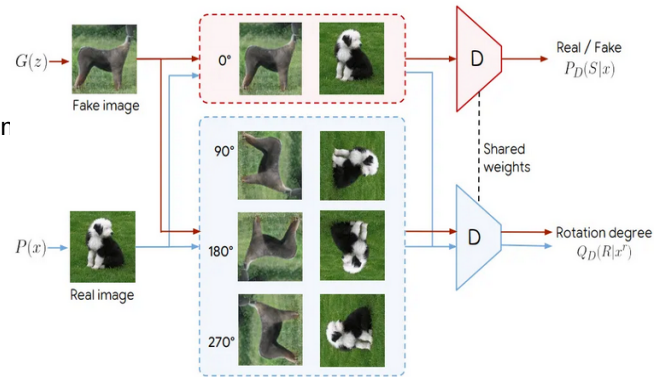


- Traditional clustering techniques (e.g. k-means) use the latent feature vectors, produced by $f_\theta(x_n)$ & clusters them into k (fixed) distinct groups using a geometric criterion
- Jointly learns a $d \times k$ centroid matrix C (d=embedding size, k=number of cluster centers) and the respective cluster assignments y_n
- $\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} min_{y_n \in \{0,1\}^k} ||f_\theta(x_n) - Cy_n||_2^2$, with $y_n^T 1_k = 1$ (one cluster per sample)
- Assign each sample to the nearest cluster, update centroids to minimize overall distance

Source: Image from Mathilde Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features"

Generative-Based Learning

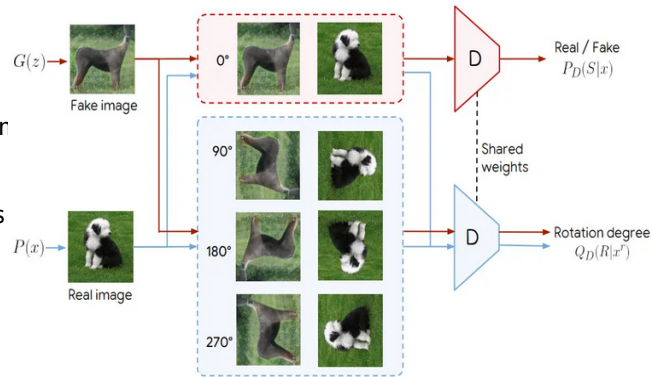
- Idea: generative SSL–approaches train the model to generate or reconstruct parts of the given input data, allowing the model to capture meaningful paradigms (e.g. pretext task – inpaintir rotation, and other augmentations)



Source: Image from Ting Chen et al., "Self-Supervised GANs via Auxiliary Rotation Loss"

Generative-Based Learning

- Idea: generative SSL–approaches train the model to generate or reconstruct parts of the given input data, allowing the model to capture meaningful paradigms (e.g. pretext task – inpaintir rotation, and other augmentations)
- Architectures: traditional autoencoders (AE), variational autoencoders (VAE), and generative adversarial networks (GANs) → in general unsupervised, conditioning generation w.r.t. partial (augmented) data samples



Source: Image from Ting Chen et al., "Self-Supervised GANs via Auxiliary Rotation Loss"

Challenges and Applications

- Data augmentation: appropriate augmentations can be challenging, especially outside of vision tasks

Challenges and Applications

- Data augmentation: appropriate augmentations can be challenging, especially outside of vision tasks
- Negative sampling in contrastive learning: efficient and diverse negative samples are essential for contrastive SSL (appropriate pairs/triplets)

Challenges and Applications

- Data augmentation: appropriate augmentations can be challenging, especially outside of vision tasks
- Negative sampling in contrastive learning: efficient and diverse negative samples are essential for contrastive SSL (appropriate pairs/triplets)
- Representation collapse: in non-contrastive methods, there is often a lack of diversity with respect to the entire representation space, finally being very similar

Challenges and Applications

- Data augmentation: appropriate augmentations can be challenging, especially outside of vision tasks
- Negative sampling in contrastive learning: efficient and diverse negative samples are essential for contrastive SSL (appropriate pairs/triplets)
- Representation collapse: in non-contrastive methods, there is often a lack of diversity with respect to the entire representation space, finally being very similar
- Transferability across domains: SSL representations may not always generalize well to different domains without a careful design

Challenges and Applications

- Data augmentation: appropriate augmentations can be challenging, especially outside of vision tasks
- Negative sampling in contrastive learning: efficient and diverse negative samples are essential for contrastive SSL (appropriate pairs/triplets)
- Representation collapse: in non-contrastive methods, there is often a lack of diversity with respect to the entire representation space, finally being very similar
- Transferability across domains: SSL representations may not always generalize well to different domains without a careful design
- Various applications: computer vision (image classification, object detection, image segmentation), NLP (LLMs, such as Generative pre-trained transformer – GPT), Speech and Audio Processing (speech recognition, sound event classification), Reinforcement Learning, and others

What is it about?

- Human-based data labeling is labor intensive, costly (human resources, time), and error-prone

Source: Image from https://makeameme.org/meme/after-finishing-labelling

What is it about?

- Human-based data labeling is labor intensive, costly (human resources, time), and error-prone
- In active learning the model selectively queries the most informative data samples to label

Source: Image from https://makeameme.org/meme/after-finishing-labelling

OTH Amberg-Weiden

What is it about?

- Human-based data labeling is labor intensive, costly (human resources, time), and error-prone
- In active learning the model selectively queries the most informative data samples to label
- Helps to prioritize labeling efforts, especially in cases where annotating each example requires significant resources

Source: Image from https://makeameme.org/meme/after-finishing-labelling

OTH Amberg-Weiden

What is it about?

- Human-based data labeling is labor intensive, costly (human resources, time), and error-prone
- In active learning the model selectively queries the most informative data samples to label
- Helps to prioritize labeling efforts, especially in cases where annotating each example requires significant resources
- Idea: Achieve similar or better performance with fewer labeled examples using an intelligent way of selecting those samples, thus the model can learn more efficiently & accurately, while reducing the amount of required labeled data

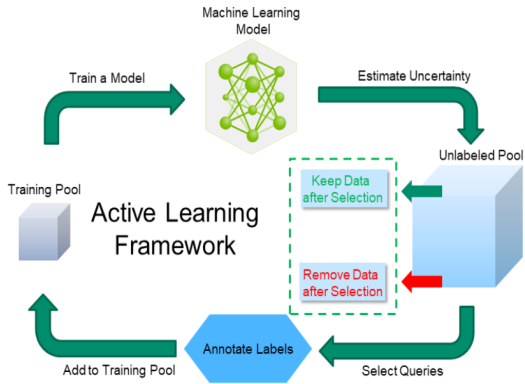
Source: Image from https://makeameme.org/meme/after-finishing-labelling

Deep Learning Paradigms – Part I

Active Learning

Process: starts with a small set of labeled data, selects data points for labeling in an iterative & intelligent way, minimizing the cost of annotating

1. Train a model using the initial small (labeled) training pool



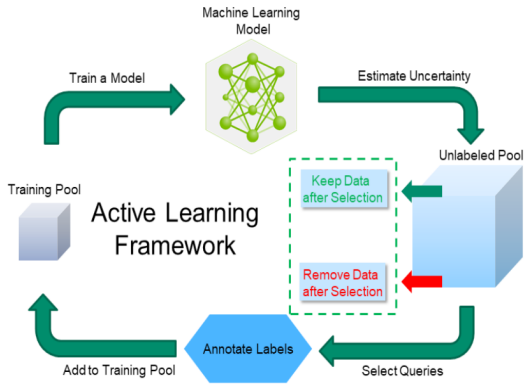
Source: Image from https://viso.ai/deep-learning/active-learning/

Deep Learning Paradigms – Part I

Active Learning

Process: starts with a small set of labeled data, selects data points for labeling in an iterative & intelligent way, minimizing the cost of annotating

1. Train a model using the initial small (labeled) training pool
2. Prediction of new unlabeled data samples



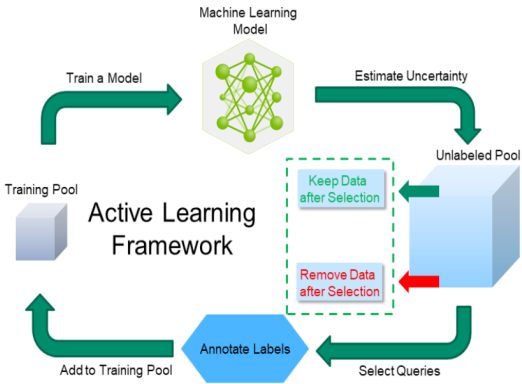
Source: Image from https://viso.ai/deep-learning/active-learning/

Deep Learning Paradigms – Part I

Active Learning

Process: starts with a small set of labeled data, selects data points for labeling in an iterative & intelligent way, minimizing the cost of annotating

1. Train a model using the initial small (labeled) training pool
2. Prediction of new unlabeled data samples
3. Uncertainty estimation: different query strategies (e.g. classical-, uncertainty-, pool-, entropy-, stream-based sampling) to select the most informative samples for labeling



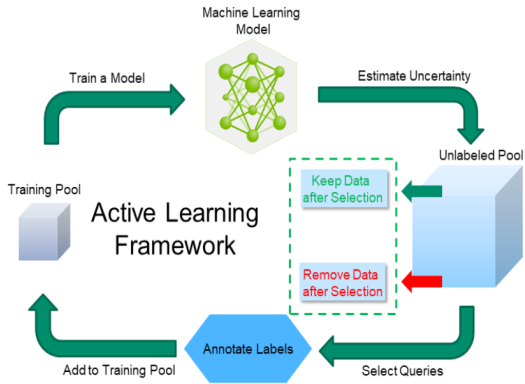
Source: Image from https://viso.ai/deep-learning/active-learning/

Deep Learning Paradigms – Part I

Active Learning

Process: starts with a small set of labeled data, selects data points for labeling in an iterative & intelligent way, minimizing the cost of annotating

1. Train a model using the initial small (labeled) training pool
2. Prediction of new unlabeled data samples
3. Uncertainty estimation: different query strategies (e.g. classical-, uncertainty-, pool-, entropy-, stream-based sampling) to select the most informative samples for labeling
4. Human-based annotation of the machine-selected data samples



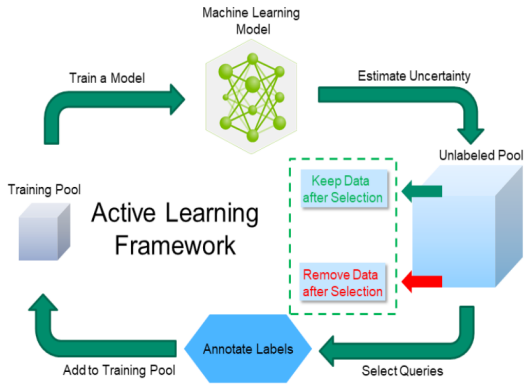
Source: Image from https://viso.ai/deep-learning/active-learning/

Deep Learning Paradigms – Part I

Active Learning

Process: starts with a small set of labeled data, selects data points for labeling in an iterative & intelligent way, minimizing the cost of annotating

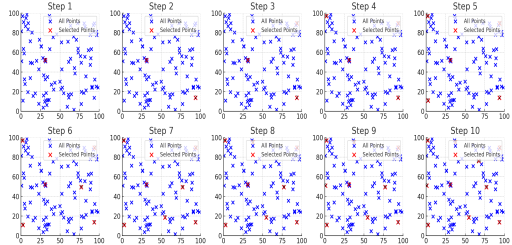
1. Train a model using the initial small (labeled) training pool
2. Prediction of new unlabeled data samples
3. Uncertainty estimation: different query strategies (e.g. classical-, uncertainty-, pool-, entropy-, stream-based sampling) to select the most informative samples for labeling
4. Human-based annotation of the machine-selected data samples
5. Add new data to the training pool & retrain

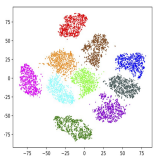


Source: Image from https://viso.ai/deep-learning/active-learning/

Query Strategy – Classical

- The following techniques do not rely on any model predictions, it considers sample distances only





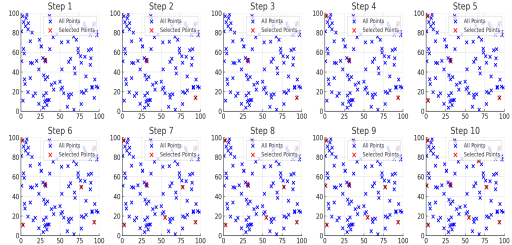
Source: Image from https://en.wikipedia.org/wiki/Farthest-first_traversal# & Yazhou Ren et al., "Deep Density-based Image Clustering"

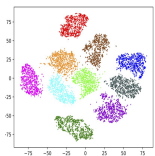
Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Classical

- The following techniques do not rely on any model predictions, it considers sample distances only
- Random Sampling: random, no strategy behind!





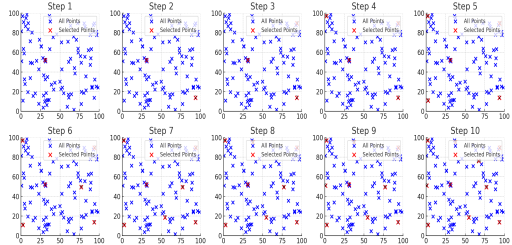
Source: Image from https://en.wikipedia.org/wiki/Farthest-first_traversal# & Yazhou Ren et al., "Deep Density-based Image Clustering"

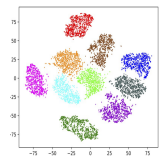
Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Classical

- The following techniques do not rely on any model predictions, it considers sample distances only
- Random Sampling: random, no strategy behind!
- Farthest-First Traversal: always picking the point farthest from the selected set of points (each new point is as different as possible from the previous points) → Greedy algorithm!



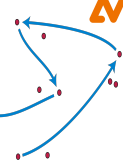


Source: Image from https://en.wikipedia.org/wiki/Farthest-first_traversal# & Yazhou Ren et al., "Deep Density-based Image Clustering"

Deep Learning Paradigms – Part I

Active Learning

first graph selects random and second graph selcts fardest away

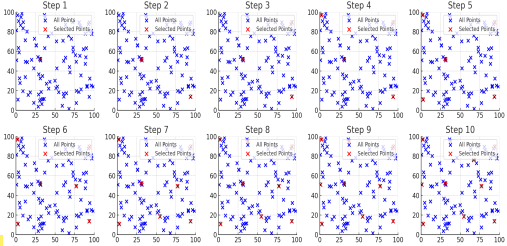


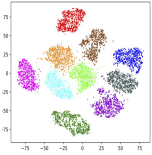
Query Strategy – Classical

- The following techniques do not rely on any model predictions, it considers sample distances only
- Random Sampling: random, no strategy behind!

not a good idea

- Farthest-First Traversal: always picking the point farthest from the selected set of points (each new point is as different as possible from the previous points) → Greedy algorithm!
- K-medoids clustering: a variant of k-means, taking not the mean point as cluster center, but an actual data point in the cluster (medoid = the most centrally located object, with minimum sum of distances to other points)



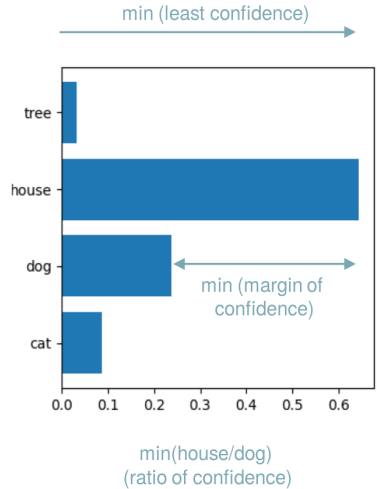


Source: Image from https://en.wikipedia.org/wiki/Farthest-first_traversal# & Yazhou Ren et al., "Deep Density-based Image Clustering"

OTH Amberg-Weiden

Query Strategy – Uncertainty

- Idea: model queries data points for which it has the highest uncertainty in its predictions

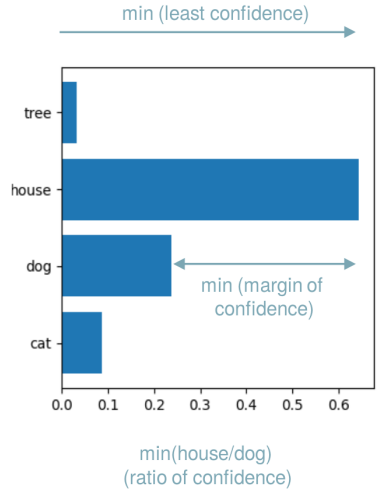


Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

OTH Amberg-Weiden

Query Strategy – Uncertainty

- Idea: model queries data points for which it has the highest uncertainty in its predictions
- Least Confidence Sampling: selects all the samples across the data pool, where the model's highest target class predictions is lowest, indicating uncertainty $\rightarrow \min_{samples} \ max_c \ y_c$



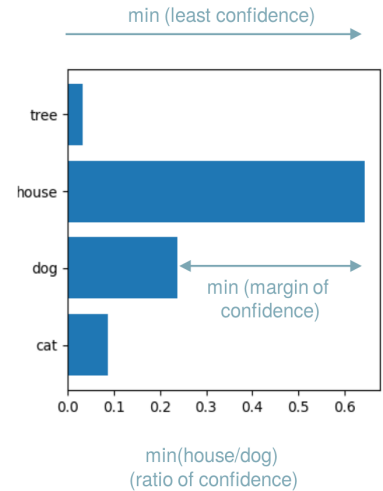
Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Uncertainty

- Idea: model queries data points for which it has the highest uncertainty in its predictions
- Least Confidence Sampling: selects all the samples across the data pool, where the model's highest target class predictions is lowest, indicating uncertainty $\rightarrow \min_{samples} max_c\ y_c$
- Margin Sampling: samples where the difference between the highest and second-highest prediction probability is smallest (ambiguity!) $\rightarrow \min_{samples} (max_c\ y_c - max_{c_{max}}\ y_c)$



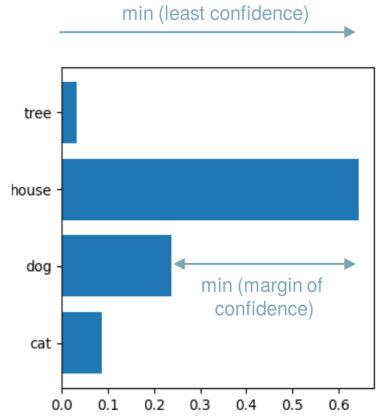
Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Uncertainty

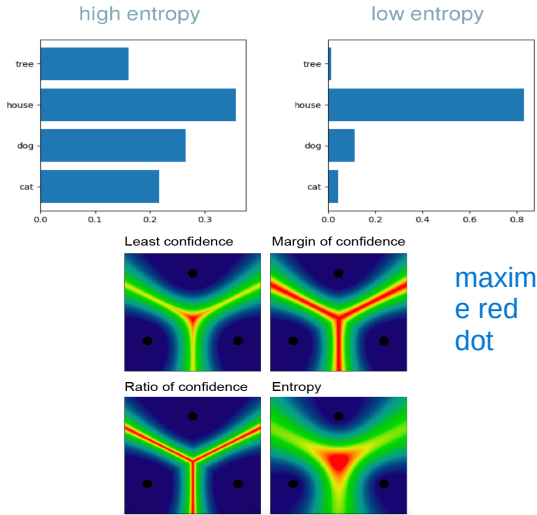
- Idea: model queries data points for which it has the highest uncertainty in its predictions
- Least Confidence Sampling: selects all the samples across the data pool, where the model's highest target class predictions is lowest, indicating uncertainty $\rightarrow \min_{samples} max_c\ y_c$
- Margin Sampling: samples where the difference between the highest and second-highest prediction probability is smallest (ambiguity!) $\rightarrow$ $\min_{samples}\ (max_c\ y_c - max_{c_{max}}\ y_c)$
- Ratio-based Sampling: smallest ratio between two top predictions $\rightarrow \min_{samples}\ (\frac{max_c\ y_c}{max_{c_{max}}\ y_c})$



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Query Strategy – Uncertainty

- Entropy-based Samling: Entropy $\mathbb{H}$ measures the overall uncertainty of the prediction, considering all the class probabilities
- Higher entropy $\mathbb{H}$ indicates greater uncertainty across multiple classes
- Entropy $\mathbb{H}$ for a prediction with n classes: $\max_{samples} \mathbb{H} = \max_{samples} (-\sum y_c \, log(y_c))$
- Provides a more global view of uncertainty



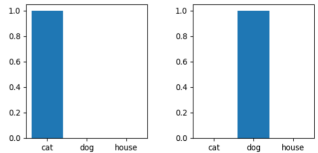
Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML
Source: Image from https://livebook.manning.com/book/human-in-the-loop-machine-learning/chapter-3/231

Query Strategy – Entropy

- Selection strategies for ensembles? $\rightarrow$ Choosing data points where ensemble members have the highest disagreement! Strategies: Highest Disagreement & Maximum Mutual Information (Bayesian Active Learning by Disagreement – BALD)

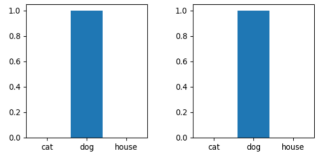
High Disagreement

$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ small



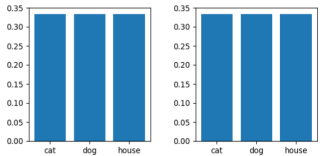
No Disagreement

$\mathbb{H}(\mathbb{E}(y))$ small
$\mathbb{E}(\mathbb{H}(y))$ small



Data Disagreement

$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ large



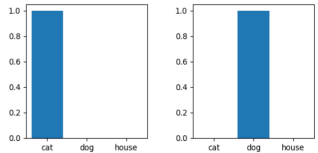
Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

OTH Amberg-Weiden

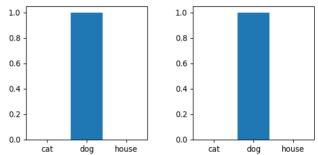
Query Strategy – Entropy

- Selection strategies for ensembles? $\rightarrow$ Choosing data points where ensemble members have the highest disagreement! Strategies: Highest Disagreement & Maximum Mutual Information (Bayesian Active Learning by Disagreement – BALD)
- Highest Disagreement: select samples where ensemble members (different models, similar data) have the most varied predictions ($\mathbb{H}(\mathbb{E}(y))$)

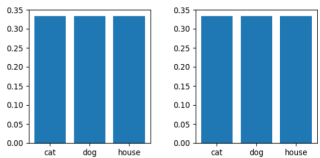
High Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ small



No Disagreement
$\mathbb{H}(\mathbb{E}(y))$ small
$\mathbb{E}(\mathbb{H}(y))$ small



Data Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ large



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

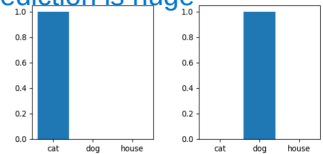
OTH Amberg-Weiden

entropy of average prediction is huge

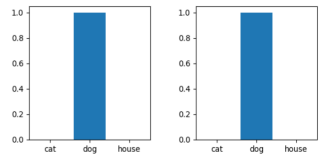
Query Strategy – Entropy

- Selection strategies for ensembles? $\rightarrow$ Choosing data points where ensemble members have the highest disagreement! Strategies: Highest Disagreement & Maximum Mutual Information (Bayesian Active Learning by Disagreement – BALD)
- Highest Disagreement: select samples where ensemble members (different models, similar data) have the most varied predictions ($\mathbb{H}(\mathbb{E}(y))$)
- Maximum Mutual Information: select samples maximizing the mutual information between the model predictions, quantifying uncertainty about the model's knowledge itself ($\mathbb{H}(\mathbb{E}(y)$ and $\mathbb{E}(\mathbb{H}(y))$)

High Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ small

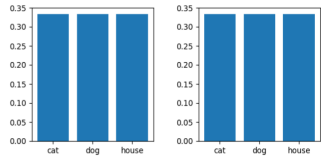


No Disagreement
$\mathbb{H}(\mathbb{E}(y))$ small
$\mathbb{E}(\mathbb{H}(y))$ small



analyse your model if we get this

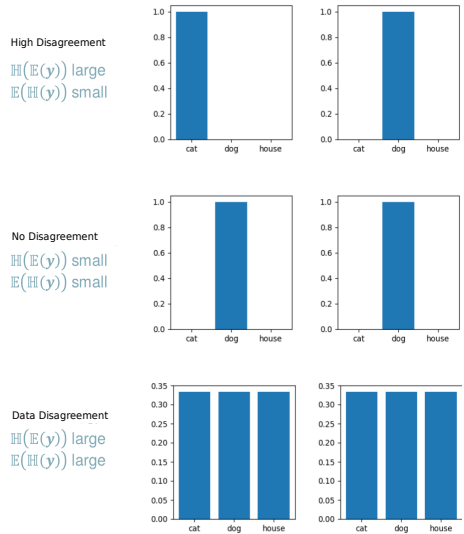
Data Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ large



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Query Strategy – Entropy

- Maximum Mutual Information:
  $\max_{samples} \mathbb{I} = \max_{samples} \mathbb{H}(\mathbb{E}(y)) - \mathbb{E}(\mathbb{H}(y))$
  - $\mathbb{I}$ as mutual information (epistemic uncertainty)
  - $\mathbb{H}(\mathbb{E}(y)$ = entropy of the average prediction (high if model's aggregated predictions spread across classes)
  - $\mathbb{E}(\mathbb{H}(y))$ = average entropy of individual predictions (how confident each model is about its own prediction, low value, low internal uncertainty)

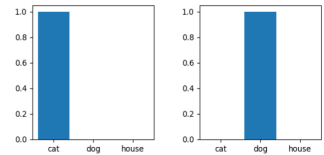


Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

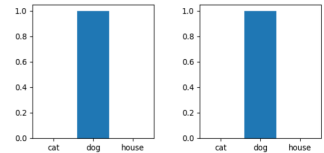
Query Strategy – Entropy

- Maximum Mutual Information:
  $\max_{samples} \mathbb{I} = \max_{samples} \mathbb{H}(\mathbb{E}(y)) - \mathbb{E}(\mathbb{H}(y))$
  - $\mathbb{I}$ as mutual information (epistemic uncertainty)
  - $\mathbb{H}(\mathbb{E}(y)$ = entropy of the average prediction (high if model's aggregated predictions spread across classes)
  - $\mathbb{E}(\mathbb{H}(y))$ = average entropy of individual predictions (how confident each model is about its own prediction, low value, low internal uncertainty)
- Variation in predictions from multiple forward passes with dropout enabled, which simulates an ensemble by generating diverse outputs for the same input

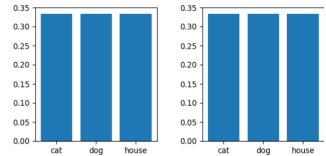
High Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ small



No Disagreement
$\mathbb{H}(\mathbb{E}(y))$ small
$\mathbb{E}(\mathbb{H}(y))$ small



Data Disagreement
$\mathbb{H}(\mathbb{E}(y))$ large
$\mathbb{E}(\mathbb{H}(y))$ large

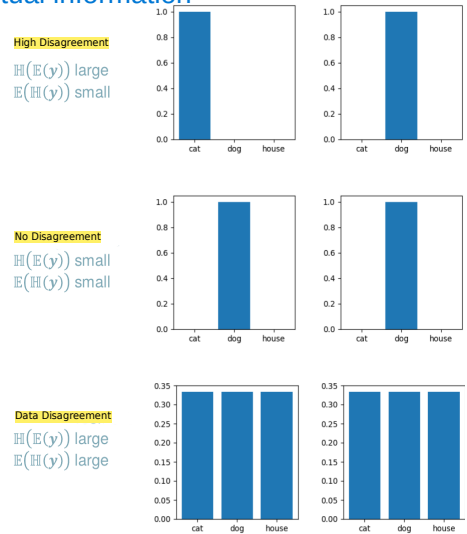


Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

goal to maximize the mutual information

Query Strategy – Entropy

- Maximum Mutual Information:
  $\max_{samples} \mathbb{I} = \max_{samples} \mathbb{H}(\mathbb{E}(y)) - \mathbb{E}(\mathbb{H}(y))$
  - $\mathbb{I}$ as mutual information (epistemic uncertainty)
  - $\mathbb{H}(\mathbb{E}(y)$ = entropy of the average prediction (high if model's aggregated predictions spread across classes)
  - $\mathbb{E}(\mathbb{H}(y))$ = average entropy of individual predictions (how confident each model is about its own prediction, low value, low internal uncertainty)
- Variation in predictions from multiple forward passes with dropout enabled, which simulates an ensemble by generating diverse outputs for the same input
- $\mathbb{H}(\mathbb{E}(y))$ is high, $\mathbb{E}(\mathbb{H}(y))$ is low $\rightarrow$ confident, but confident in different classes



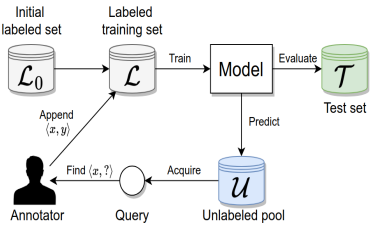
Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

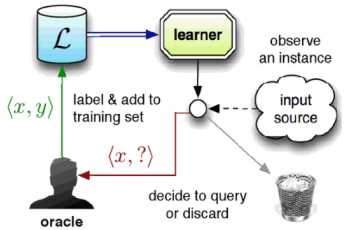
Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Pool & Stream

- Pool-based Sampling:
  - Model has access to a large pool of unlabeled data
  - Evaluation of all samples in the pool based on a selection criterion (uncertainty, highest disagreement)
  - Picks the most informative samples to query!
  - Use-Case: huge offline datasets



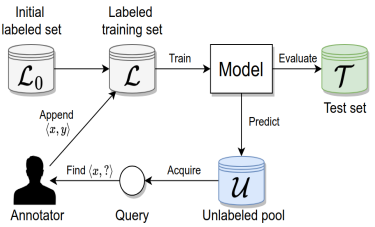


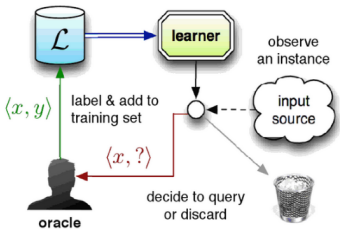
Deep Learning Paradigms – Part I

Active Learning

Query Strategy – Pool & Stream

- Pool-based Sampling:
  - Model has access to a large pool of unlabeled data
  - Evaluation of all samples in the pool based on a selection criterion (uncertainty, highest disagreement)
  - Picks the most informative samples to query!
  - Use-Case: huge offline datasets
- Stream-based Sampling:
  - Data samples arrive sequentially, 1-by-1, as a stream of unlabeled samples
  - The model assesses informativeness based on a selection criterion, in case of high uncertainty or disagreement, it is queried for labeling or discarded
  - Use-Case: real-time data feed (user activity log, many unseen data)





OTH Amberg-Weiden

Batch-Selection

- Batch-wise labeling of samples is more convenient for humans than just labeling samples 1-by-1, with long waiting times in between (retraining of classifier)

OTH Amberg-Weiden

Batch-Selection

- Batch-wise labeling of samples is more convenient for humans than just labeling samples 1-by-1, with long waiting times in between (retraining of classifier)
- How to select sample batches?

OTH Amberg-Weiden

Batch-Selection

- Batch-wise labeling of samples is more convenient for humans than just labeling samples 1-by-1, with long waiting times in between (retraining of classifier)
- How to select sample batches?
- Highest uncertainty/disagreement $\rightarrow$ Probably worse than random sampling!

OTH Amberg-Weiden

Batch-Selection

- Batch-wise labeling of samples is more convenient for humans than just labeling samples 1-by-1, with long waiting times in between (retraining of classifier)
- How to select sample batches?
- Highest uncertainty/disagreement → Probably worse than random sampling!
- Highest joint mutual information (BatchBALD)
  - Instead of selecting one sample at a time w.r.t. a maximal mutual information, BatchBALD computes $\mathbb{I}$ for all possible combinations of samples in a batch
  - 1. multiple predictions per sample in a batch
  - 2. compute $\mathbb{I} = \max_{batch} \mathbb{H}(\mathbb{E}(y)) - \mathbb{E}(\mathbb{H}(y))$ (w.r.t. all samples in a batch) → Pick batch with max $\mathbb{I}$

Batch-Selection

- BatchBALD → maximizing mutual information to reduce epistemic uncertainty, DBAL → emphasize diversity in the batch, labeled samples cover a broad range of the data space (no redundant information!)

OTH Amberg-Weiden

Batch-Selection

- BatchBALD → maximizing mutual information to reduce epistemic uncertainty, DBAL → emphasize diversity in the batch, labeled samples cover a broad range of the data space (no redundant information!)
- Balances uncertainty-based selection (samples where the model lacks confidence) with diversity-based selection (avoid redundancy within the batch)
  → Not only uncertain, but diverse, each sample provides unique insights to the model

Batch-Selection

- BatchBALD → maximizing mutual information to reduce epistemic uncertainty, DBAL → emphasize diversity in the batch, labeled samples cover a broad range of the data space (no redundant information!)
- Balances uncertainty-based selection (samples where the model lacks confidence) with diversity-based selection (avoid redundancy within the batch)
  → Not only uncertain, but diverse, each sample provides unique insights to the model
- Approach: uncertainty estimation, diversity criterion, batch selection

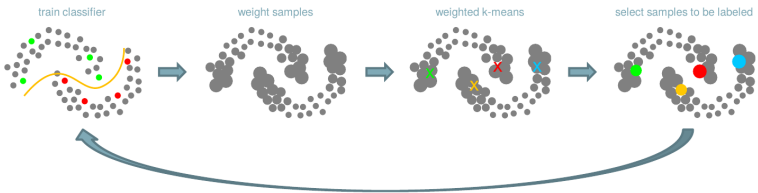
Batch-Selection

- BatchBALD → maximizing mutual information to reduce epistemic uncertainty, DBAL → emphasize diversity in the batch, labeled samples cover a broad range of the data space (no redundant information!)
- Balances uncertainty-based selection (samples where the model lacks confidence) with diversity-based selection (avoid redundancy within the batch)
  → Not only uncertain, but diverse, each sample provides unique insights to the model
- Approach: uncertainty estimation, diversity criterion, batch selection
- Distance between sample embeddings in the model's feature space (clustering technique k-means) or a distance metric (cosine or euclidean distance)

cluserting and euclidean distance can be different

Batch-Selection

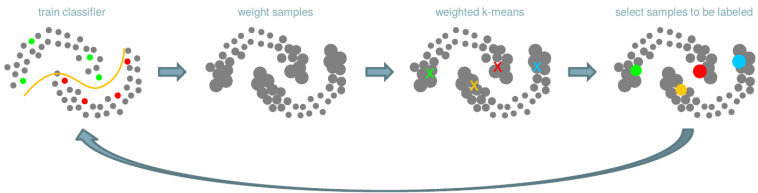
- DBAL ensures that selected samples are spread out capturing various input space regions



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Batch-Selection

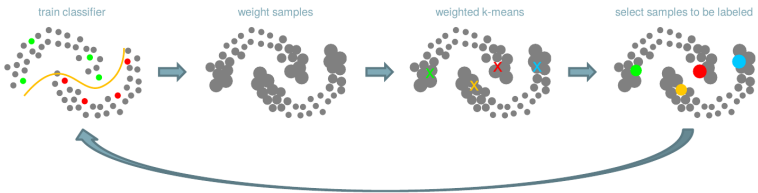
- DBAL ensures that selected samples are spread out capturing various input space regions
- Approach: 1. train classifier on labeled data, 2. weight samples based on uncertainty/disagreement, 3. run weighted k-means (for a batch of k samples to be labeled), 4. select samples closest to each of the k-centroids



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Batch-Selection

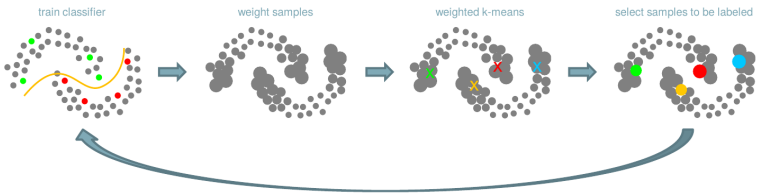
- DBAL ensures that selected samples are spread out capturing various input space regions
- Approach: 1. train classifier on labeled data, 2. weight samples based on uncertainty/disagreement, 3. run weighted k-means (for a batch of k samples to be labeled), 4. select samples closest to each of the k-centroids
- DBAL mini-batch that maximizes both objectives (uncertainty, diversity)



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Batch-Selection

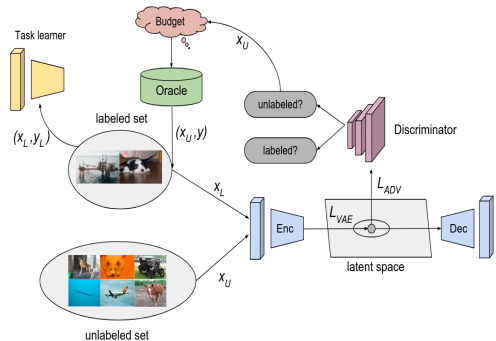
- DBAL ensures that selected samples are spread out capturing various input space regions
- Approach: 1. train classifier on labeled data, 2. weight samples based on uncertainty/disagreement, 3. run weighted k-means (for a batch of k samples to be labeled), 4. select samples closest to each of the k-centroids
- DBAL mini-batch that maximizes both objectives (uncertainty, diversity)
- High uncertainty is still prioritized, but within that high-uncertainty set, DBAL picks samples that are diverse from one another



Source: Image from OTH-AW, Electrical Engineering, Media and Computer Science, Thomas Nierhoff – Vorlesung Advanced Topics in ML

Variational Adversarial Active Learning (VAAL)

- Variational autoencoder and adversarial discriminator compete against each other to learn a latent representation to identify samples that are different from labeled data

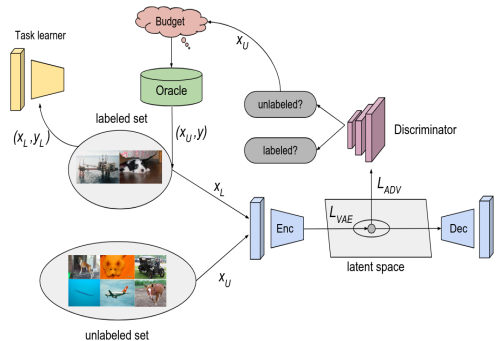


Source: Samarth Sinha et al., "Variational Adversarial Active Learning"

OTH Amberg-Weiden

Variational Adversarial Active Learning (VAAL)

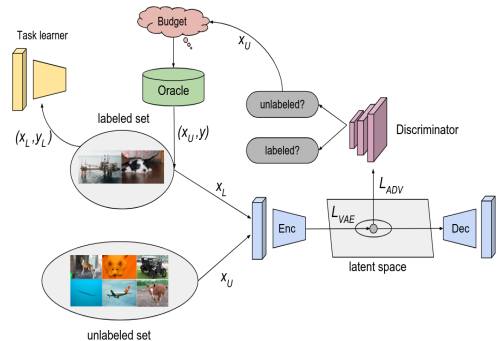
- Variational autoencoder and adversarial discriminator compete against each other to learn a latent representation to identify samples that are different from labeled data
- Discriminator uses latent space of VAE to determine if samples belong to labeled or unlabeled data
  $\rightarrow$ VAE tries to fool discriminator (all labeled!)



Source: Samarth Sinha et al., "Variational Adversarial Active Learning"

Variational Adversarial Active Learning (VAAL)

- Variational autoencoder and adversarial discriminator compete against each other to learn a latent representation to identify samples that are different from labeled data
- Discriminator uses latent space of VAE to determine if samples belong to labeled or unlabeled data
  $\rightarrow$ VAE tries to fool discriminator (all labeled!)
- Idea: VAAL uses a VAE and an adversarial discriminator to learn a latent representation, to identify samples, clearly different from labeled data



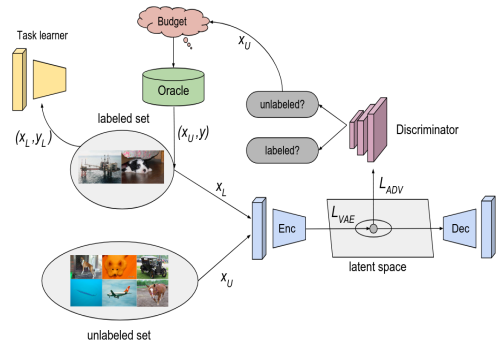
Source: Samarth Sinha et al., "Variational Adversarial Active Learning"

OTH Amberg-Weiden

Variational Adversarial Active Learning (VAAL)

- Variational autoencoder and adversarial discriminator compete against each other to learn a latent representation to identify samples that are different from labeled data
- Discriminator uses latent space of VAE to determine if samples belong to labeled or unlabeled data
  $\rightarrow$ VAE tries to fool discriminator (all labeled!)
- Idea: VAAL uses a VAE and an adversarial discriminator to learn a latent representation, to identify samples, clearly different from labeled data
- Key: unlabeled samples that the discriminator struggles, are likely to be informativ

unlabeld data is taken to human for annotation



vae tries to fool discriminator

Source: Samarth Sinha et al., "Variational Adversarial Active Learning"

Further Questions?

https://www.oth-aw.de/hochschule/ueber-uns/personen/bergler-christian/

Source: https://emekaboris.medium.com/the-intuition-behind-100-days-of-data-science-code-c98402cdc92c