Adversarial Active Learning for Deep Networks: a Margin Based Approach

Melanie Ducoffe and Frederic Precioso

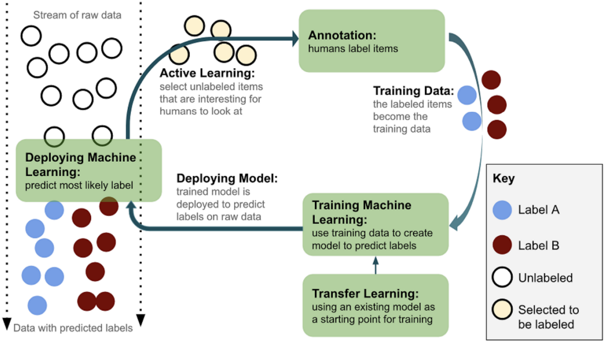
Enrico Maria Bachiorrini

Université Côte d'Azur

March 23rd 2021

UNIVERSITÉ CÔTE D'AZUR

What is active learning



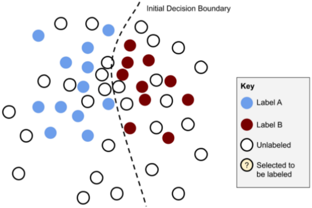
Challenges of active learning

- Minimize the number of annotations queried to the oracle.

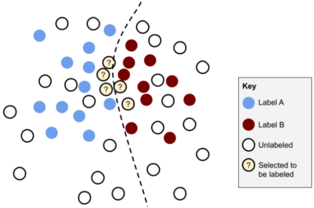
Margin based active learning

Only queries samples that are close to the decision boundary.

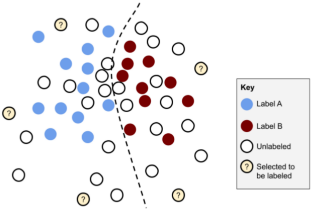
How to find the decision boundary?

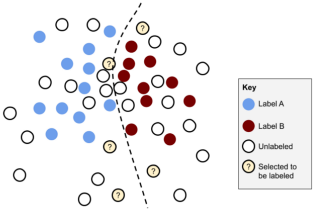


The boundary from a Machine Learning model, that would predict Label A to the left and Label B to the right.



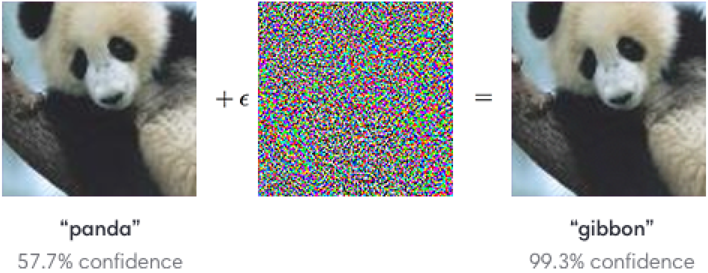
Uncertainty Sampling: selecting unlabeled items near the decision boundary.





Adversarial examples

Instances with small, intentional feature perturbations designed to cause the model to make a false prediction.



Algorithm 1 ***DFAL***: DeepFool Active Learning

Require: $\mathcal{L}$ set of initial labeled training examples
Require: $\mathcal{U}$ set of initial unlabeled training examples
Require: $\mathcal{H}$ set of hyper-parameters to train the network
Require: K the number of candidates
Require: n_{query} the number of data to query
Require: p: the L_p norm used ($p = 2$)
Require: N: the number of data to label

 # init the training set
 $k = 0$
 $\mathcal{L}_0 = \mathcal{L}$
 $\mathcal{U}_0 = \mathcal{U}$
 while k<N **do**
  # Train the network $\mathcal{A}_k$ given the current labeled training set
  $\mathcal{A}_k$=training($\mathcal{H}, \mathcal{L}_k$)
  # Select randomly a pool of data $\mathcal{S}_k$ of size K
  $\mathcal{S}_k \subseteq \mathcal{U}_k; \quad | \mathcal{S}_k | = K$
  for $x_i \in \mathcal{S}_k$ **do**
   #compute adversarial attacks with L_p norms
   $r_i \leftarrow DeepFool(x_i, \mathcal{A}_k; p)$
  end for
  # query the labels of the n_{query}-th samples $\mathcal{Q}_k$ owing the smallest L_p norm perturbation
  $index_k \leftarrow argsort(< r_i, r_i >_p | \ i = 1..K)$
  $\mathcal{Q}_k \leftarrow \{x_j \mid j \in index_k[0 : n_{query}]\} \cup \{x_j + r_j \mid j \in index_k[0 : n_{query}]\}$
  $\mathcal{L}_{k+1} \leftarrow \mathcal{L}_k \cup \mathcal{Q}_k$
  $\mathcal{U}_{k+1} \leftarrow \mathcal{U}_k \setminus \mathcal{Q}_k$
 end while

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	**82.77**	96.23	**97.71**	98.02	–
BALD	51.88	91.96	93.69	94.24	–
CEAL	71.81	94.81	96.77	97.33	–
CORE-SET	78.86	**96.52**	97.53	**98.03**	–
EGL	58.44	73.86	78.57	78.57	–
uncertainty	57.96	92.52	94.84	96.41	–
RANDOM	77.56	92.83	94.63	95.31	**99.04**

(a) *MNIST* (LeNet5)

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	**84.28**	**96.90**	**97.98**	**98.59**	–
BALD	53.73	91.47	94.32	94.32	–
CEAL	50.87	90.69	90.69	90.69	–
CORE-SET	78.80	96.68	97.46	97.88	–
EGL	37.92	91.84	93.99	93.99	–
uncertainty	45.57	88.36	94.27	94.60	–
RANDOM	69.79	91.96	94.05	94.46	**98.98**

(b) *MNIST* (VGG8)

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	94.62	98.50	98.98	99.10	–
BALD	93.10	97.95	97.95	97.95	–
CEAL	84.65	98.50	99.00	99.12	–
CORE-SET	92.50	**98.75**	**99.07**	**99.25**	–
EGL	75.07	95.47	95.47	95.47	–
uncertainty	**95.78**	98.35	98.85	98.98	–
RANDOM	95.50	98.07	98.07	98.07	**99.70**

(c) *Shoe-Bag* (LeNet5)

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	**87.73**	98.53	**99.30**	99.50	–
BALD	86.78	95.35	97.83	97.83	–
CEAL	84.20	98.78	99.25	**99.52**	–
CORE-SET	0.50	**99.12**	99.12	99.12	–
EGL	0.50	97.28	97.28	97.28	–
uncertainty	83.75	83.75	83.75	83.75	–
RANDOM	86.78	95.83	97.08	97.08	**99.50**

(d) *Shoe-Bag* (VGG8)

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	**82.56**	**89.63**	**90.72**	**91.09**	–
BALD	72.65	87.18	88.34	88.45	–
CEAL	70.46	87.04	88.31	89.39	–
CORE-SET	79.58	88.93	90.54	90.53	–
EGL	57.48	64.05	64.05	69.85	–
uncertainty	69.24	86.89	88.54	89.09	–
RANDOM	78.09	87.03	88.98	89.42	**95.46**

(e) *Quick-Draw* (LeNet5)

	Accuracy (%)				
# annotations	100	500	800	1000	All
DFAL	**84.23**	**91.52**	**93.16**	**93.91**	–
BALD	82.00	89.94	91.92	92.87	–
CEAL	64.45	79.66	85.73	88.65	–
CORE-SET	66.71	89.93	92.28	92.62	–
EGL	63.12	86.80	90.06	90.06	–
uncertainty	52.77	88.05	89.31	91.03	–
RANDOM	78.28	88.13	89.71	89.94	**96.75**

(f) *Quick-Draw* (VGG8)