RETHINKING DEEP ACTIVE LEARNING: USING UNLABELED DATA AT MODEL TRAINING

Oriane Siméoni
Mateusz Budnik
Yannis Avrithis
Guillaume Gravier

INRIA, IRISA, Univ Rennes, CNRS
Training DL model for the classification task

- Requires large annotated datasets
- Annotation done by humans
- Long and **fastidious** process
LEARNING WITH LESS SUPERVISION

- class 1
- class 2
- class 3

Supervised
LEARNING WITH LESS SUPERVISION
LEARNING WITH LESS SUPERVISION

- class 1
- class 2
- class 3
- unlabeled

Supervised  Semi-supervised  Unsupervised
LEARNING WITH LESS SUPERVISION

Supervised  Semi-supervised  Active  Unsupervised

•

•

•

•
Active Learning

Machine learning model

Annotator

L

U
Active Learning

Machine learning model

Select a budget $b$ of images

Annotator
Active Learning

Machine learning model

Select a budget \( b \) of images
Active Learning

Machine learning model

L

U

Select a budget $b$ of images

Annotator
Active Learning

Machine learning model

L

Select a budget $b$ of images

U

Annotator
Active Learning

One cycle

Select a budget $b$ of images

- Relevant before Deep Learning
- Not studied much in the context of Deep Learning
What is the best solution?

Baselines

- **Random**
  Selects uniformly random images.

---

What is the best solution?

Baselines

- **Random**
  Selects uniformly random images.

- **Geometry**
  Selects most distant image to its nearest labeled or previously acquired examples.

---

What is the best solution?

Baselines

- **Random**
  Selects uniformly random images.

- **Geometry\(^1\)**
  Selects most distant image to its nearest labeled or previously acquired examples.

- **Uncertainty**
  Selects most uncertain images: highest entropy of the classifier output probabilities.

---


What is the best solution?

Baselines

- **Random**
  Selects uniformly random images.

- **Geometry**
  Selects most distant image to its nearest labeled or previously acquired examples.

- **Uncertainty**
  Selects most uncertain images: highest entropy of the classifier output probabilities.

- **CEAL**
  Uses unlabeled data.

---

What is the best solution?

Baselines

- **Random**
  Selects uniformly random images.

- **Geometry**
  Selects most distant image to its nearest labeled or previously acquired examples.

- **Uncertainty**
  Selects most uncertain images: highest entropy of the classifier output probabilities.

- **CEAL**
  Uses unlabeled data.

---


**Experimental Details**

- **Network**
  - 13-layer convolutional network
  - Model trained from scratch

- **Training**
  - Very dependent on the data → 5 repetitions

- **Metrics**
  - Average accuracy and standard deviation

- **Datasets**
  - MNIST (10 cls, 60000 imgs)
  - SVHN (10 cls, 73257 imgs)
  - CIFAR-10 (10 cls, 50000 imgs)
  - CIFAR-100 (100 cls, 50000 imgs)

---


**What is the best solution?**

No clear **winner**.
THE IDEA

What if we could

- Improve results **with no additional supervision**
- Use unlabeled data during training
USING MORE UNLABELED DATA

Machine learning model

L
Annotator

U
Selection function
 USING MORE UNLABELED DATA

Unsupervised pre-training → Machine learning model → Annotation → Selection function

L → U

Semi-supervised method
Using more unlabeled data

Unsupervised pre-training → Machine learning model

Semi-supervised method → Selection function

L → Annotator → U
Improving the model using unlabeled data

Unsupervised pre-training

Following Deep Cluster\textsuperscript{8} to pre-train CNN
- Assign classes to data given closest centroids
- Train the network
- Re-assign classes

INTEGRATING INFORMATION FROM UNLABELED DATA

CIFAR-10 ($b = 100$)
Random
CoreSet
Uncertainty

CEAL

zero.osf
one.osf
two.osf
three.osf
four.osf

three.osf/zero.osf
four.osf/zero.osf
five.osf/zero.osf
six.osf/zero.osf

cycle

average accuracy

CIFAR-10 (b = 100)

Benefits
- performed only once at the beginning of the process
- can bring up to 6% improvement
IMPROVING ACTIVE LEARNING CYCLES

- Use unlabeled data in each cycle
- Adding semi-supervised learning

- Iterative label propagation following Iscen et al\(^9\).
  - Construct a reciprocal $k$-nn graph on data features
  - Label propagation
  - Train classifier using pseudo-labels

---

ADDING SEMI-SUPERVISION

![Graph showing average accuracy over cycles for CIFAR-10 (b = 100)]
Adding semi-supervision

Benefits

- Results improved by up to 15% from baselines
- Taking advantage of the whole dataset
- Suits better deep learning models
ADDING SEMI-SUPERVISION

![Graphs showing average accuracy over cycles for different datasets and values of parameter b.](image)

- **CIFAR-10 (b = 100)**
  - Average accuracy increases over cycles.
  - Different line styles and colors represent different conditions or models.

- **CIFAR-10 (b = 1000)**
  - Similar trend as CIFAR-10 (b = 100).

- **SVHN (b = 100)**
  - Average accuracy increases over cycles.

- **CIFAR-100 (b = 1000)**
  - Average accuracy increases over cycles.

The graphs illustrate the impact of semi-supervision on model performance across different datasets and parameter settings.
Conclusions

Take home message

- Active learning benefits from using **unlabeled** data
- We obtain **better** models requiring **less labeled** data
- **Random** selection of images is best with small budgets
- The selection method **does not** appear to **matter**

Contributions

- First results mixing active learning and unlabeled methods in the context of Deep Learning
- Proposition to rethink **Deep Active Learning**
  - using a scenario integrating unlabeled data
  - to always compare to Random with small budgets