Rethinking deep active learning: Using unlabeled data at model training

Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, Guillaume Gravier
Inria, IRISA, CNRS, Univ Rennes

Motivation

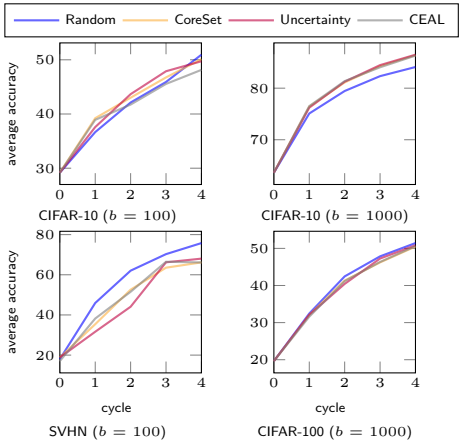
Training a model costs **a lot** of human labeling effort

- Known possible solution to reduce costs: **Active Learning**
  - Selecting most informative images to be labeled
  - Was efficient for methods **before deep learning** – one image at a time
- However,
  - Deep learning models need more images for training – is AL still relevant?
  - Unlabeled images are used **only for acquisition**
  - Why not taking advantage of the **unlabeled** images?

What is the best AL method?

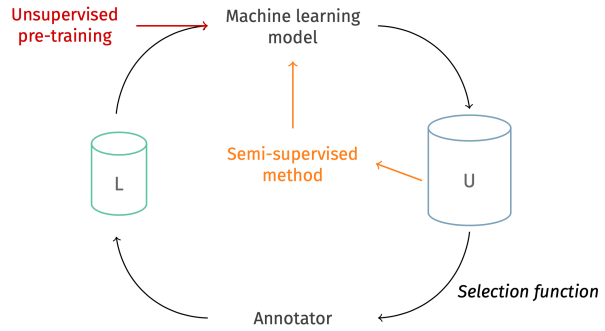
Baselines

- Random
  Selects uniformly random images.
- Geometry [2, 4]
  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 [5]
  Uses unlabeled data.



No clear winner

Using more unlabeled data

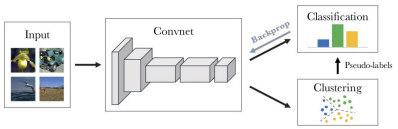


Our idea:
Adding **unlabeled** data to the active learning pipeline:

- Unsupervised pre-trained model performed once
- Using pseudo-labeled images at training, taking advantage of the whole dataset

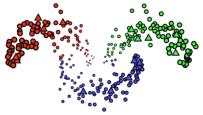
Integrating information from unlabeled data

- Improving the model using unlabeled data
- **Unsupervised pre-training**
- Following Deep Cluster [1] to pre-train CNN
  - Assign classes to data given closest centroids
  - Train the network
  - Re-assign classes

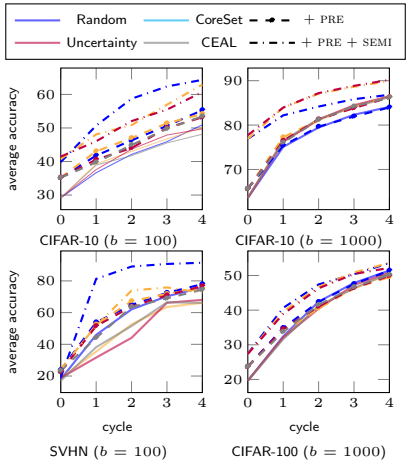


Improving active learning cycles

- Use unlabeled data in each cycle
- Adding **semi-supervised learning**
- Iterative label propagation following [3]
  - Construct a reciprocal k-nn graph on data features
  - Label propagation
  - Train classifier using pseudo-labels



Results



Adding **unsupervised pre-training**

- Training performed only once at the beginning of the process
- Brings **up to 6%** improvement

Adding **semi-supervised learning**

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

Conclusions

- 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

References:

[1] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. *arXiv preprint arXiv:1807.05520*, 2018.

[2] Y. Geifman and R. El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

[3] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label propagation for deep semi-supervised learning. In *CVPR*, 2019.

[4] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv*, 2018.

[5] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Trans. CSVT*, 27(12):2591–2600, 2017.

Contact: oriane.simeoni@inria.fr