Project 2: Semi-Supervised Learing

Lecturer: Gao Huang
TA: Yang Yue

June 2024

In this assignment, you will work with a remote sensing dataset known as EuroSAT [4]. Your task is to tackle the challenge of land cover image classification using contemporary deep neural networks. You are restricted to using only the provided dataset for each specific task. The incorporation of additional data sources or pre-trained models is prohibited. It is important to note that your model's performance will contribute to your overall score in this project, so please be careful with your experiments

As illustrated in Figure 1, the EuroSAT dataset is derived from Sentinel-2 satellite imagery and encompasses 13 spectral bands. It includes 27,000 labeled and geo-referenced images distributed across 10 distinct categories of land cover. These categories are: "Annual Crop", "Highway", "Permanent Crop", "Sea Lake", "Forest", "Industrial", "Residential", "Herbaceous Vegetation", "Pasture", and "River".

Figure 1: The EuroSAT dataset.

1 Task A: Fully Supervised Learning

In folder `src`, we have provided a basic start code for loading the data, training loop, and model definition. You should read them and understand them. You are free to change any part of the code, but you are not encouraged to use architectures other than the ResNet18[3] we provided.

Problem 1 (10pt): Use the default parameters provided in the code to train your model. Report the model's performance. Plot the accuracy and loss curve of the training and testing set.

Problem 2 (40pt): Enhance your training process. Your optimizations should include, but are not limited to, the following actions:

- Adjusting hyperparameters such as number of epochs, learning rate, batch size, weight decay, etc.
- Try different optimizer types, such as RMSprop and Adam.
- Use common techniques such as dropout, data augmentation, and learning rate schedules.
- Any additional strategies you found useful.

Maintain a detailed log of your experiments and analyze the impact of each modification introduced. Additionally, provide the accuracy and loss curves for your most effective configuration.

2 Task B: Semi-Supervised Learning

The task of *semi-supervised learning* aims to learn from a dataset that is partially labeled. Specifically, the whole dataset $\mathcal{D} = \{\mathcal{D}_L, \mathcal{D}_U\}$ comprises a small labeled subset $\mathcal{D}_L = \{x_i, y_i\}_{i=1}^{N_L}$ and a large scale unlabeled subset $\mathcal{D}_U = \{u_j\}_{j=1}^{N_U}$. We hope to learn from the labeled subset while making the best use of the unlabeled subset.

Problem 3 (10 pt): Train your model on the labeled subset $\mathcal{D}_L$ and evaluate its performance. You should also optimize your training procedure.

There are many techniques to deal with partially labeled datasets. We provide two basic formulations. You can choose **one of them** to finish your project.

1. Pre-train then fine-tune: First, pre-train the model with self-supervised learning methods on the entire dataset $\mathcal{D}$, then fine-tune on the labeled subset $\mathcal{D}_L$. (See SimCLRv2 [2])

2. Joint training: Introduce another loss term for the unlabeled images. The optimization problem can be formulated as

$$\min_{\theta} \sum_{(x,y)\in\mathcal{D}_L} \mathcal{L}_s(x, y, \theta) + \alpha \sum_{u\in\mathcal{D}_U} \mathcal{L}_u(u, \theta). \tag{1}$$

where $\mathcal{L}_s$ is the supervised loss, $\mathcal{L}_u$ is the unsupervised loss. One of the most popular forms of unsupervised loss is to "guess" the labels of $u \in \mathcal{D}_U$ based on the model's predictions (called *pseudo labels*). Existing works obtain pseudo labels by:

- Using a (sharpened) average of the predictions from multiple augmentations as pseudo labels (See MixMatch [1])
- Using the prediction from a weakly augmented image as the pseudo label for a strongly augmented one (See FixMatch [5])
- ...

Problem 4 (40 pt): Use the above techniques or anything else you like to solve the semi-supervised learning problem on the Eurostat dataset. There's **no requirements** for what techniques you must adopt. Answer the following questions:

- What techniques have you implemented? Provide their mathematical formulations.
- Offer intuitive explanations for why these techniques are beneficial.
- What are the impacts of these techniques on the model performance? What accuracy does your model finally reach?

Bonus (20 pt): You will get extra points if you meet at least one of the following criteria:

- Conduct extensive experiments to analyze the effects of your methods, including sensitivity to method-specific hyperparameters.
- Your model reaches very strong performance.
- Develop and implement a novel method that differs from existing approaches.

3 Dataset

We randomly split the dataset into two parts, in which the first 70% is used for training while the remaining 30% is for testing. For task A, all the training data is labeled and used to train the network. For task B, the testing set is the same as task A. The labeled training subset has 25 samples per class, and the rest of the images in the training set are unlabeled.

Run the bash script `download_data.sh`[1], and the dataset will automatically be downloaded and unzipped. The dataset directory is structured as follows:

- Task A training set: `"Task_A/train/{AnnualCrop,...,River}/xxx.jpg"`.
- Task A testing set: `"Task_A/val/{AnnualCrop,...,River}/xxx.jpg"`.
- Task B labeled training subset: `"Task_B/train_labeled/{AnnualCrop,...,River}/xxx.jpg"`.
- Task B unlabeled training subset: `"Task_B/train_unlabeled/xxx.jpg"`.
- Task B testing set: `"Task_B/val/{AnnualCrop,...,River}/xxx.jpg"`.

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[1] If you failed to run the script, please copy its contents and paste it into the terminal.

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019.

[5] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.