深度學習系統與實現

LAB02 – Data augmentation

Dept. of Computer Science and
Information Engineering

National Chiao Tung University

Outline

- □ Background
- □ LAB 2-1, data augmentation
- □ LAB 2-2, data sampler
- □ LAB 2-3, practical experience
- □ Grading

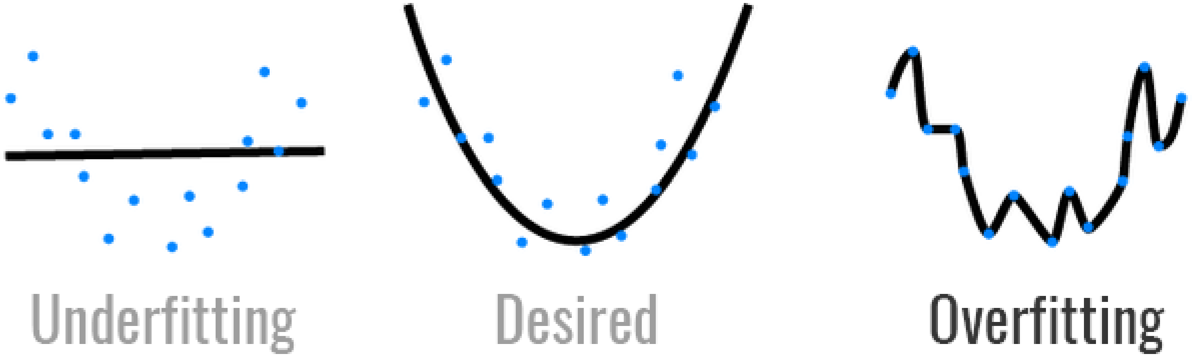
Data Augmentation

- Want the dataset to be diverse enough to include different positions, angles, lightings ...
- The best way to improve the performance of the deep learning model - Add more data
- Hard to gather more labeled data from the real world > augment existing datasets
- More robust !
  - e.g. use flip & rotate to cover different positions of the interestings

https://towardsdatascience.com/data-augmentation-and-images-7aca9bd0dbe8

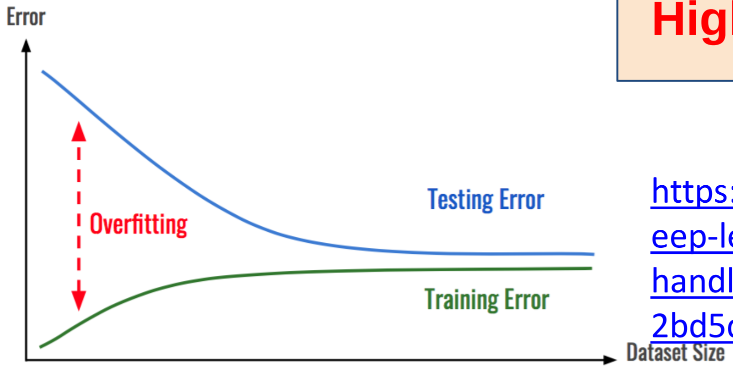
Overfitting problem

- Overfitting - model fits too well to the training set. Difficult to generalize to unseen data
- Solution - larger dataset, simpler model complexity, augmentation, regularization



High Bias

High Variance



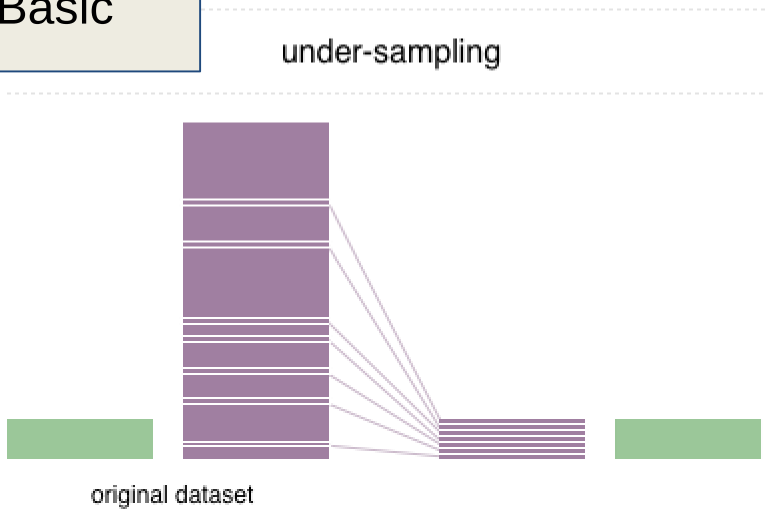
https://towardsdatascience.com/deep-learning-3-more-on-cnns-handling-overfitting-2bd5d99abe5d

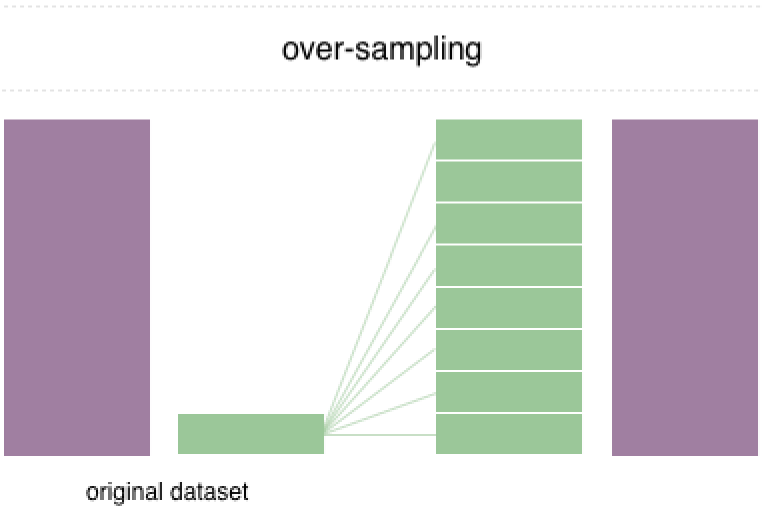
Imbalanced Dataset

- Take detection of diseases for example
  - Diseased patients are rare
  - More normal samples than disease ones
- "Resampling" is adopted to solve this problem
  - majority class -> remove sample
  - minority class -> add new sample (augmentation)

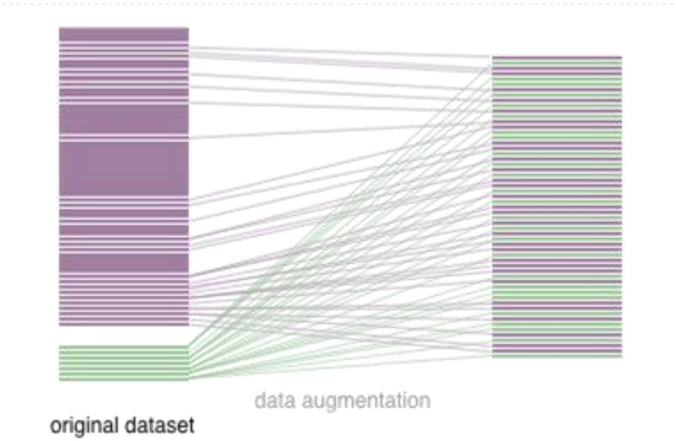
Data sampling

Basic





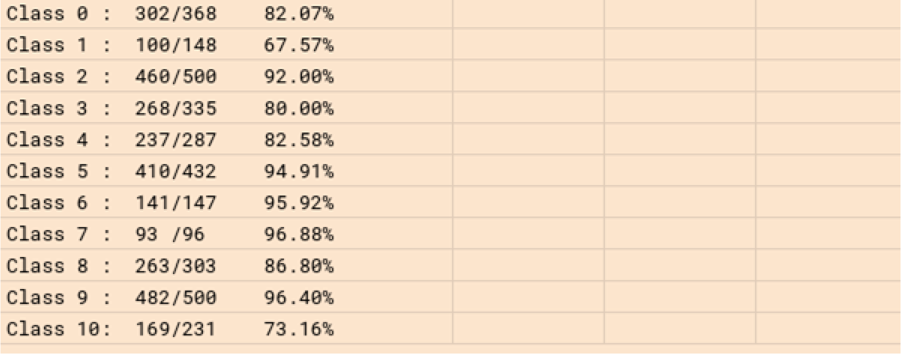
Advanced



Rebalance the class distribution !!

Constraints of LAB 2

- Continue the previous lab, try to solve the problems about dataset imbalance



```
Class 0 :  302/368    82.07%
Class 1 :  100/148    67.57%
Class 2 :  460/500    92.00%
Class 3 :  268/335    80.00%
Class 4 :  237/287    82.58%
Class 5 :  410/432    94.91%
Class 6 :  141/147    95.92%
Class 7 :  93 /96     96.88%
Class 8 :  263/303    86.80%
Class 9 :  482/500    96.40%
Class 10:  169/231    73.16%
```

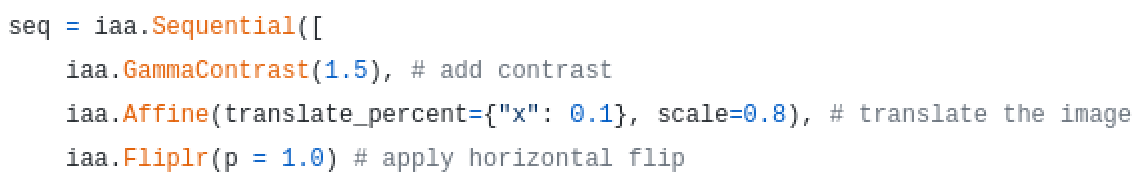
- The baseline accuracy of skewed-food11 need to be better than 80% (recommend: ResNet18, MobileNetV2)
- Cannot add or delete any images of the skewed-food11 dataset in your file system

LAB 2-1
Data Augmentation (1)

- DL frameworks usually have built-in data augmentation functions, but lack some critical features (e.g. noising)
- There are some popular image augmentation python packages(imgaug, Augmentor) designed specifically for deep learning

LAB 2-1
Data Augmentation (2)

- Take imgaug as an example
  - Over 60 image augmentation techniques
    - gaussian noise
    - blurring
    - hue/saturation changes
  - Support augmentation with segmentation masks, bounding boxes, key points
  - Augmentation pipelines



```
seq = iaa.Sequential([
    iaa.GammaContrast(1.5), # add contrast
    iaa.Affine(translate_percent={"x": 0.1}, scale=0.8), # translate the image
    iaa.Fliplr(p = 1.0) # apply horizontal flip
```

LAB 2-1
Data Augmentation (3)

- PyTorch Integration (20%)
  - try to migrate imgaug/Augmentor to torchvision.transforms
  - Visualize samples of Food11 w/ the effects by the migration from imgaug/Augmentor to torchvision.transforms
- Balanced augmentation (30%)
  - Try to implement the "augmentaion" function in the customized Dataset (ex: Food11Dataset in food11_dataset.py)
  - Balanced resample the training set with the above transform functions (balance distribution)

LAB 2-2
Data Sampler (1)

- Type of Pytorch Datasets
  - Map-style datasets
    - implements the __getitem__() and __len__() protocols
    - maintain a map from indices/keys to data samples
    - e.g. Food11Dataset in food11_dataset.py
  - Iterable-style datasets
    - implements the __iter__() protocol
    - iter(dataset), return a stream of data from a database
    - suitable for dynamic batch size
- All the works in LAB 2 must be done through Map-style datasets

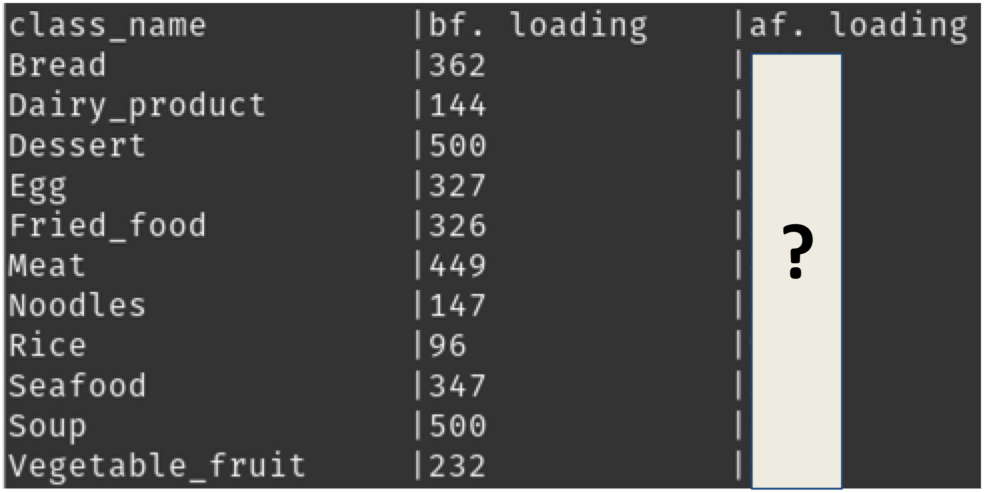
LAB 2-2
Data Sampler (2)

- For Map-style Datasets, torch.utils.data.Sampler can be used in data loading.
- Sampler object can yield the next index/key to fetch at each time
- Every Sampler subclass has to provide
  - __iter__() method, a way to iterate over indices
  - __len__() method, the length of the returned iterators

LAB 2-2
Data Sampler (3)

- use torch.utils.data.WeightedRandomSampler to do the balanced sampling (30%)
- compare the results between RandomSampler & WeightedRandomSampler

The difference between the amount of each class and the average need to be less than 10% after WeightedRandomSampler with specific weights



```
class_name           |bf. loading     |af. loading
Bread                |362             |
Dairy_product        |144             |
Dessert              |500             |
Egg                  |327             |
Fried_food           |326             |
Meat                 |449             |
Noodles              |147             |
Rice                 |96              |
Seafood              |347             |
Soup                 |500             |
Vegetable_fruit      |232             |
```

LAB 2-3
Practical experience

- Please combine the above methods & try to improve the average per-class accuracy (20%)
- Show the per-class accuracy & average per-class accuracy in your report
- How could you make improvement? Please explain in details in your report
- If it doesn't improve, please explain your experiment & analysis in your report

Grading

- LAB 2-1 (50%)
- LAB 2-2 (30%)
- LAB 2-3 (20%)
- Bonus (15%)
  - Reproduce the methods of other papers to solve the imbalance dataset problem (ex: Class-Balanced Loss Based on Effective Number of Samples, CVPR '19)
  - You should finish the above labs first

Total:
115

- Submission: source code + report (.ipynb is accepted)(E3)
  - zip format (ex: DLSR_lab2_{student id}.zip) 未依格式者,扣該lab成績5分
- Deadline : 2020/03/30,23:59 (Mon)(2 weeks)
- Demo : Date To Be Determined

Report Spec.

- EX:
  - Introduction - Problems & solutions
  - Experiment setup
  - Results
  - Analysis / Discussion
  - Others ...

Reference

- Pytorch – dataset:
  - https://pytorch.org/docs/stable/data.html#torch.utils.data.Dataset
- Augmentor:
  - https://github.com/mdbloice/Augmentor
- Imgaug Document & git-repo :
  - https://imgaug.readthedocs.io/en/latest/
  - https://github.com/aleju/imgaug
- How to use Imgaug:
  - https://colab.research.google.com/drive/109vu3F1LTzD1gdVV6cho9fKGx7lzbFll#scrollTo=rQ6DFPNvVD8s

Reference

- □ Overview of popular augmentation packages and PyTorch examples:
  - https://towardsdatascience.com/data-augmentation-for-deep-learning-4fe21d1a4eb9
- □ Imbalanced Dataset Sampler:
  - https://github.com/ufoym/imbalanced-dataset-sampler
- □ Pytorch - Sampler:
  - https://pytorch.org/docs/stable/data.html#torch.utils.data.Sampler
- □ Paper - A survey on Image Data Augmentation for Deep Learning