Data pre-processing and training technique

Liang Hong

2nd Feb, 2023

Outline

- torch.utils.data API
- Dataset split
- Overfitting/underfitting
- Data augmentation

torch.utils.data API

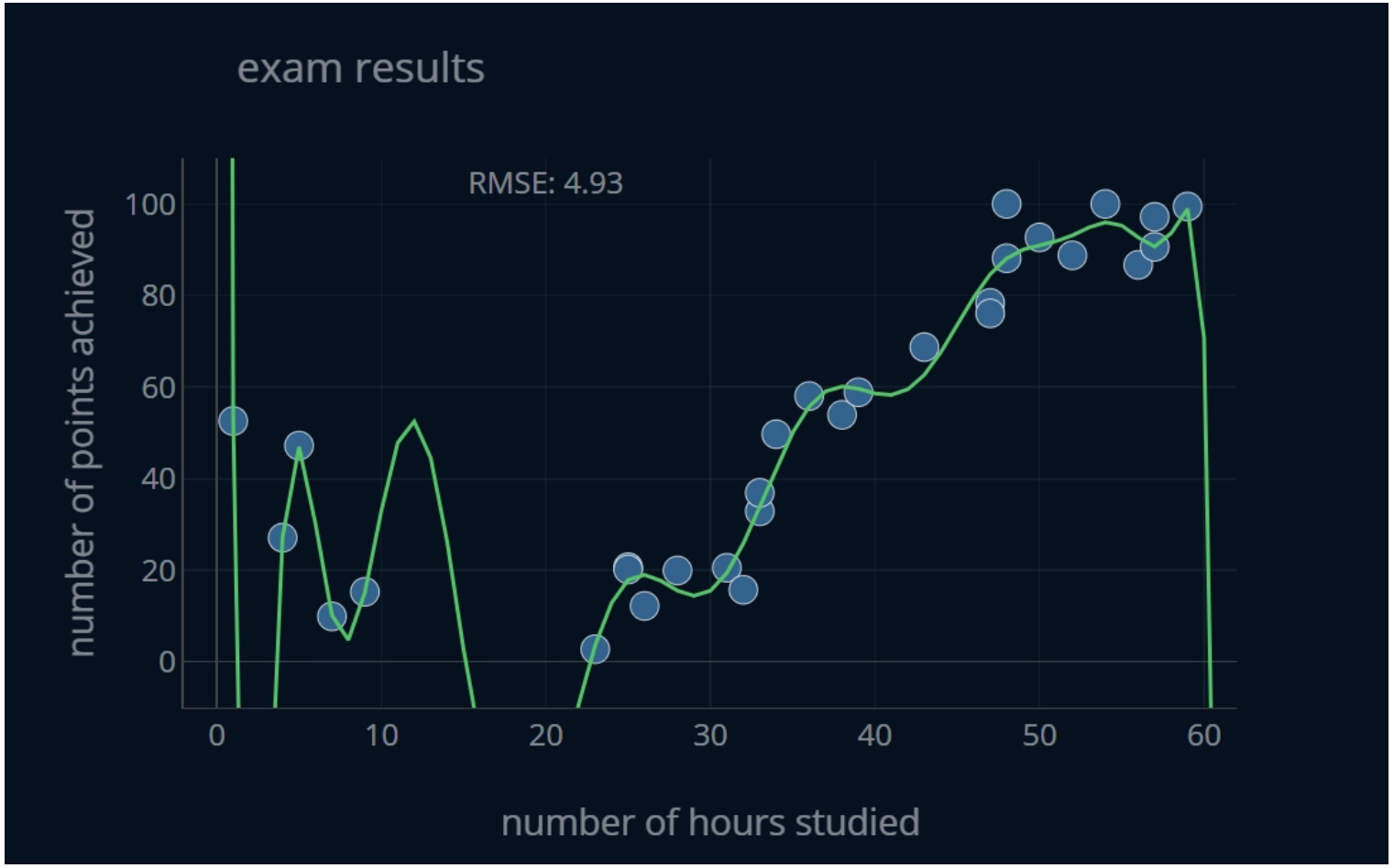
- torch.utils.data.Dataset
  - __getitem__() : pass in an index and return one data sample
  - __len__() : length of the dataset
- torch.utils.data.Sampler (not always needed)
  - __iter__() : iterate over the dataset
  - __len__() : len of iterator
- torch.utils.data.DataLoader
  - Dataset, sampler/batch size, shuffle
  - collate_fn : process batch data to tensor

Dataset split

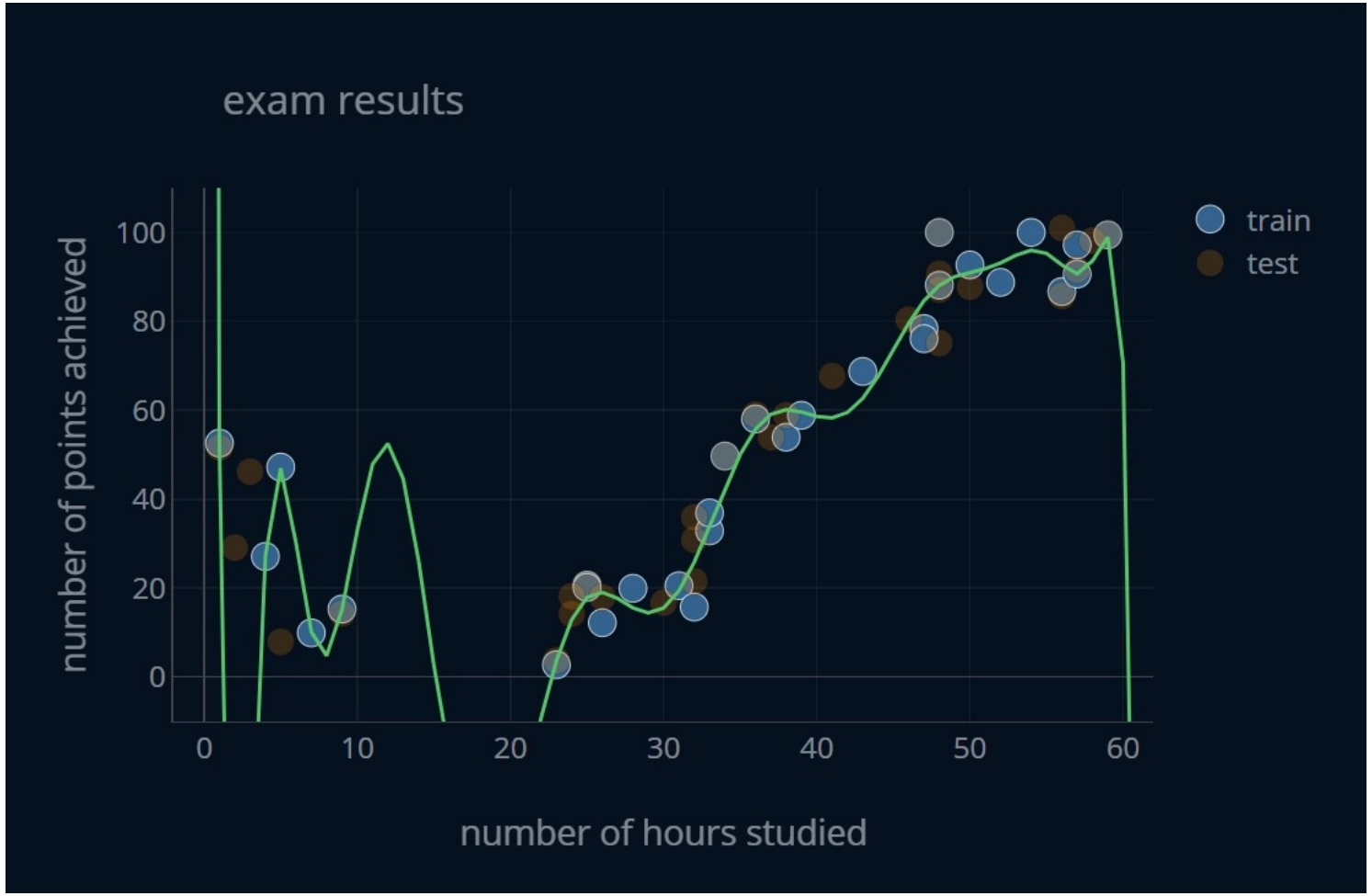
- Normally, train:eval:test = 7:2:1 (by Zhihua Zhou)
- Find whether there are duplicates separated in different set, or very similar ones (e.g. fig in train and augmented fig in test)
- Cross validation for limited data

- -> why we need to split dataset? Generalizability!

Overfitting

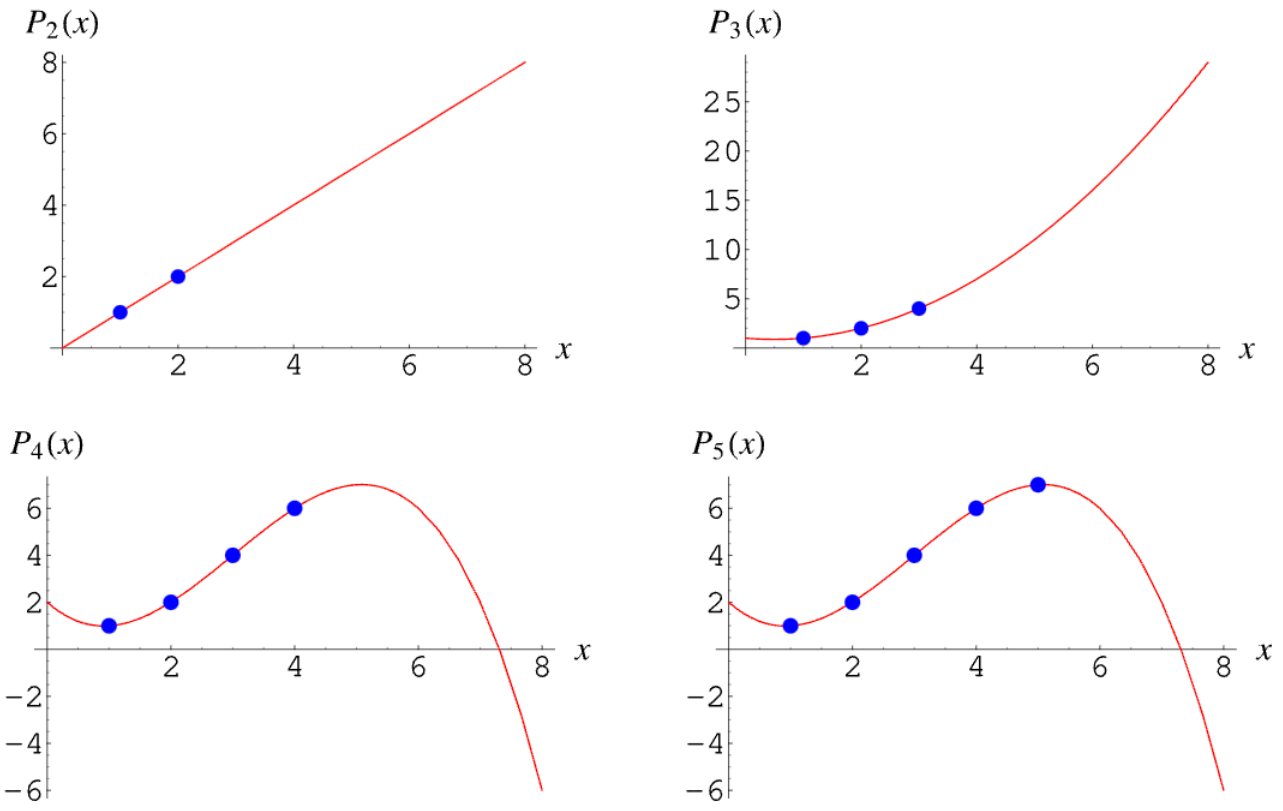


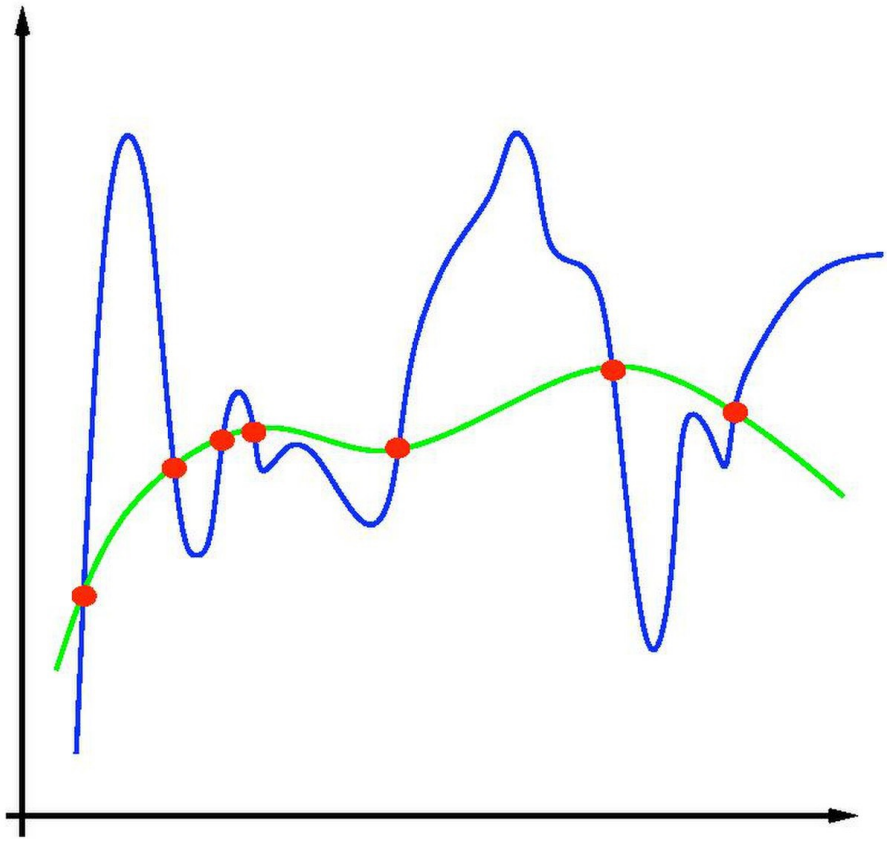
Overfitting



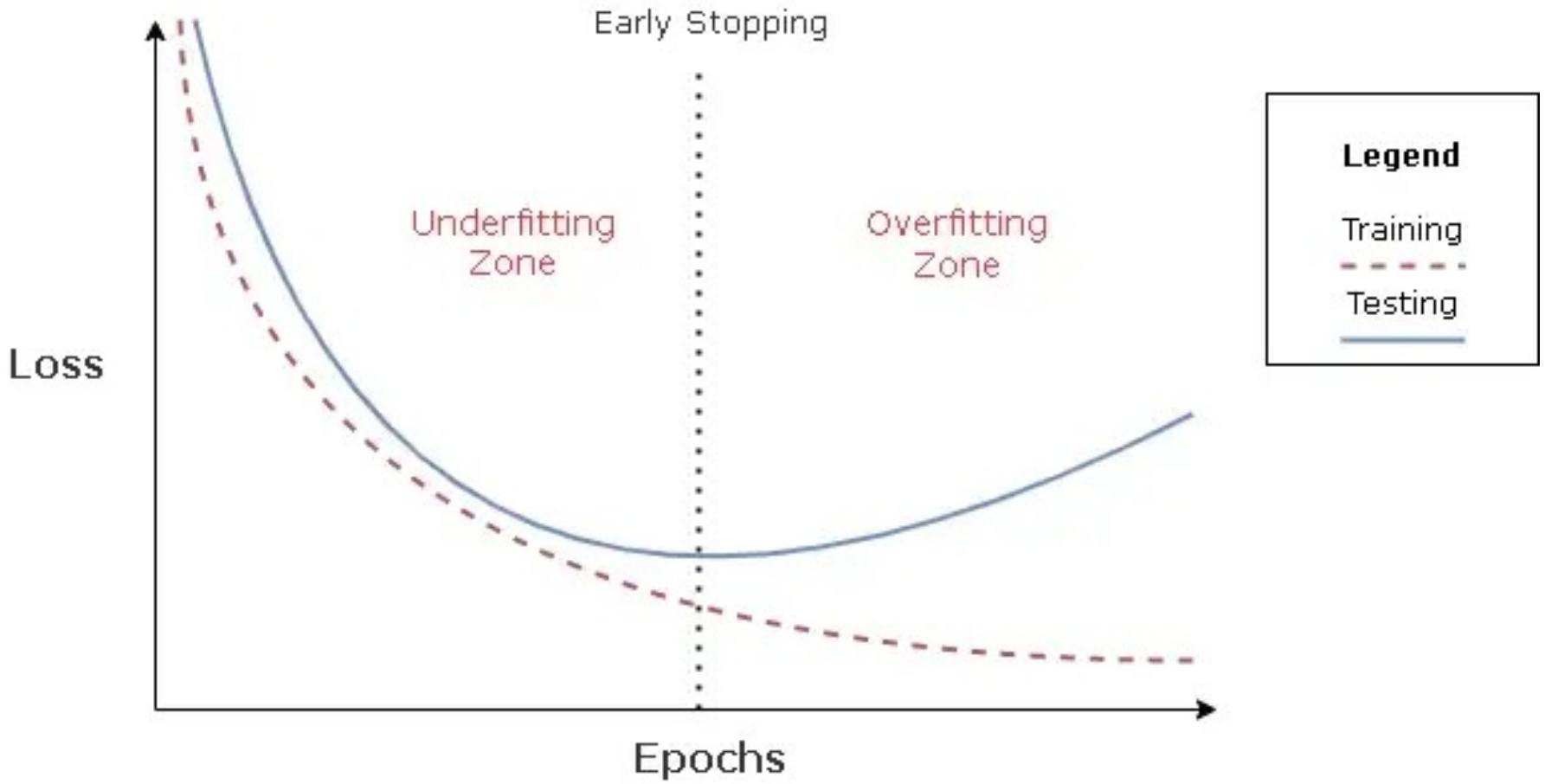
Overfitting

Example: lagrange interpolation





Overfitting



Overfitting

- When to stop?
  - Metric stop improving
  - The observed metric has not improved more than a given minimum change to be considered an improvement
  - If you set a learning rate decay, you may also combine the value of lr
- Other ways to deal with overfitting?
  - Use more data

Data augmentation

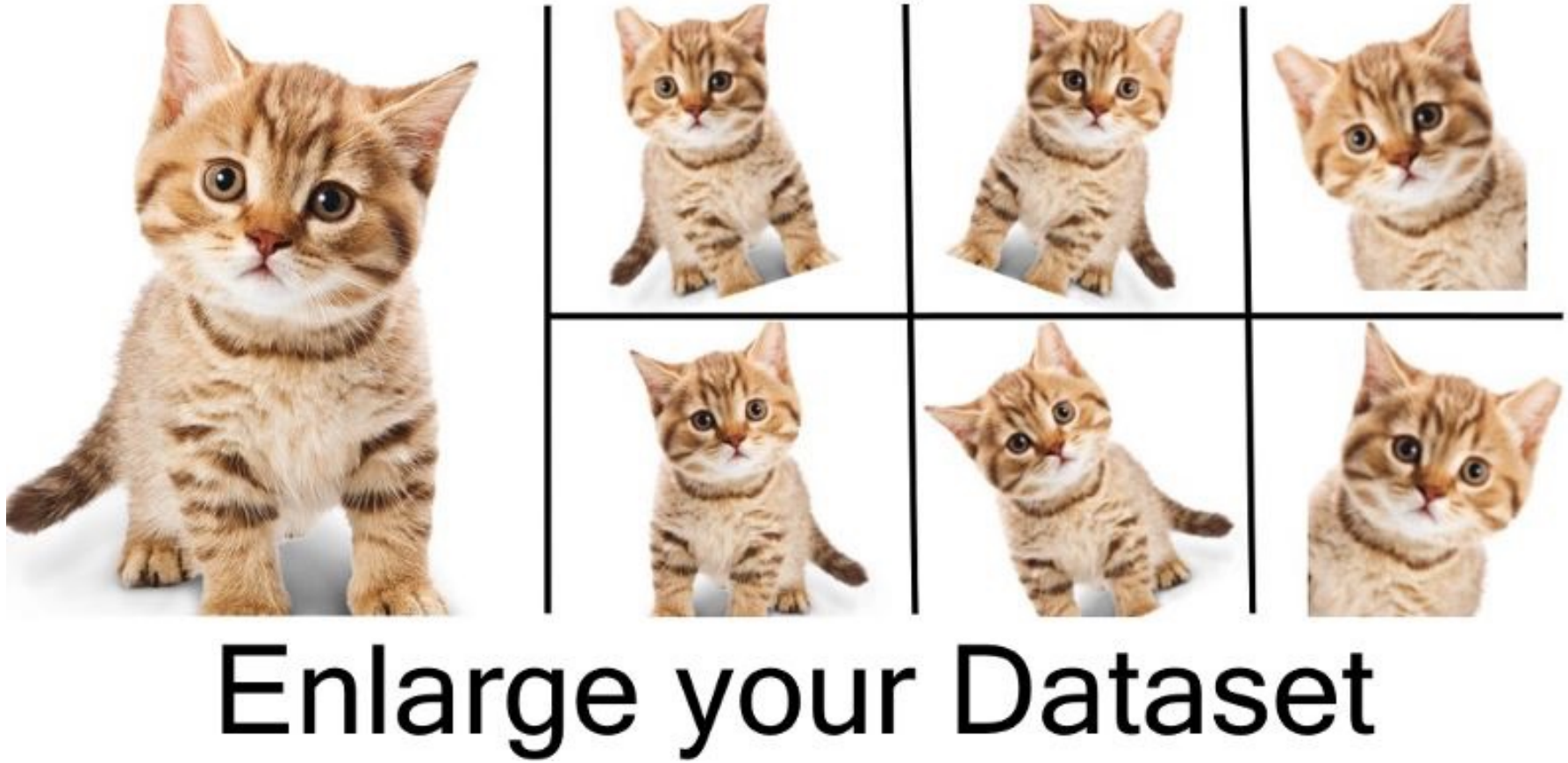


Image data augmentation

- Translation
- Rotation
- Flipping(do not apply to characters !)