Lab 1(Lab Week 3): CNN on Cifar-10

Lab Objective:

In this lab, you will be asked to build popular network architecture (*Network In Network*, NIN) [1], and train it on Cifar-10 dataset. Moreover, you need to use data augmentation and Dropout [2] during training.

Turn in:

1. Experiment Report (3/21(二))
2. Demo date (3/21(二))

Requirements:

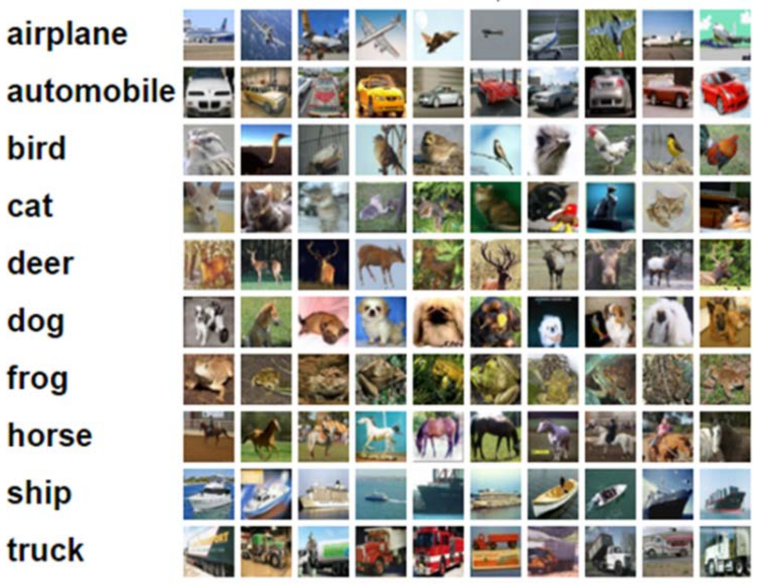
- Implement "**Network In Network**" (NIN) [1] convolutional architecture
- Implement data augmentation: translation and horizontal flipping
- Use "**Dropout**" [2] in NIN
- Train NIN+Dropout with/without data augmentation

Environment:

- Cifar-10 dataset

The CIFAR-10 dataset consists of 60000 32×32 color images (RGB) in **10** classes, with 6000 images per class. There are 50000 training images and 10000 test images.

Download: https://www.cs.toronto.edu/~kriz/cifar.html



Sample Code

There are many cifar-10 sample codes:

git clone https://github.com/tensorflow/models.git

Higher-level tensorflow API tflearn:

pip install git+https://github.com/tflearn/tflearn.git

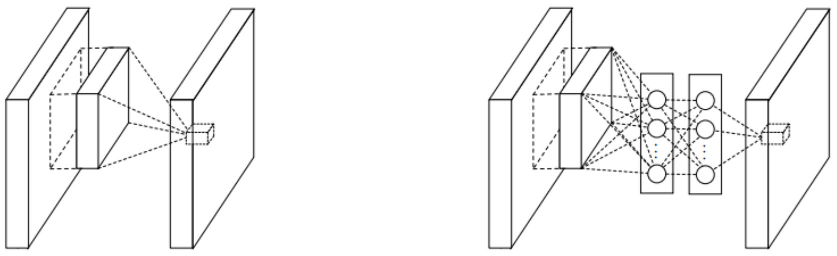
Or you can install other higher-level API, such as Keras.

Lab Description:

- Network In Network (NIN)
    - Enhance model discriminability for local patches within the receptive field

    Traditional CNN: 3x3 conv + ReLU

    NIN: 3x3 conv + ReLU + 1x1 conv + ReLU + 1x1 conv +ReLU



(a) Linear convolution layer (b) Mlpconv layer

$$f_{i,j,k} = \max(w_k^T x_{i,j}, 0).$$

$$\begin{aligned} f^1_{i,j,k_1} &= \max({w^1_{k_1}}^T x_{i,j} + b_{k_1}, 0). \\ &\vdots \\ f^n_{i,j,k_n} &= \max({w^n_{k_n}}^T f^{n-1}_{i,j} + b_{k_n}, 0). \end{aligned}$$

- Full NIN architecture used in Cifar-10

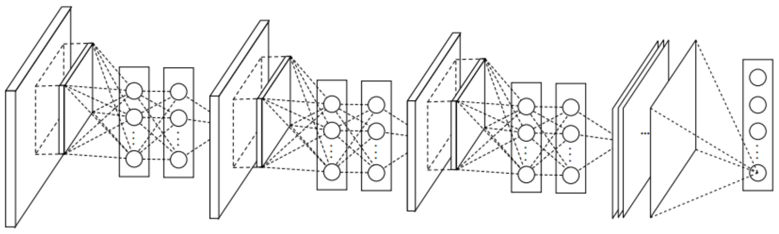
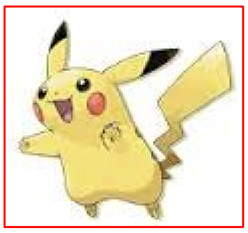


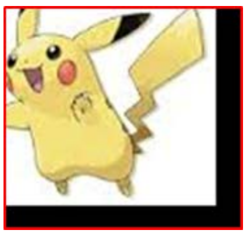
Figure 2: The overall structure of Network In Network. In this paper the NINs include the stacking of three mlpconv layers and one global average pooling layer.

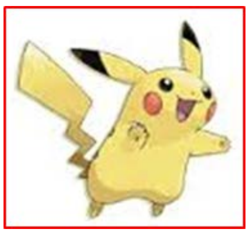
- Architecture Details:

Conv1	filter size = 5x5, # of filter =192, pad = 2, stride = 1	Act.=ReLU
mlp 1	filter size = 1x1, # of filter =160, pad = 0, stride = 1	Act.=ReLU
mlp 2	filter size = 1x1, # of filter =96, pad = 0, stride = 1	Act.=ReLU
Pool 1	3x3 max pooling, stride = 2	
	Dropout 0.5	
Conv2	filter size = 5x5, # of filter =192, pad = 2, stride = 1	Act.=ReLU
mlp 2-1	filter size = 1x1, # of filter =192, pad = 0, stride = 1	Act.=ReLU
mlp 2-2	filter size = 1x1, # of filter =192, pad = 0, stride = 1	Act.=ReLU
Pool 2	3x3 max pooling, stride = 2	
	Dropout 0.5	
Conv3	filter size = 3x3, # of filter =192, pad = 1, stride = 1	Act.=ReLU
mlp 3-1	filter size = 1x1, # of filter =192, pad = 0, stride = 1	Act.=ReLU
mlp 3-2	filter size = 1x1, # of filter =10, pad = 0, stride = 1	Act.=ReLU
Global Pool	8x8 average pooling, stride =1	
	Softmax	

- Data augmentation: Translation and Horizontal flipping:

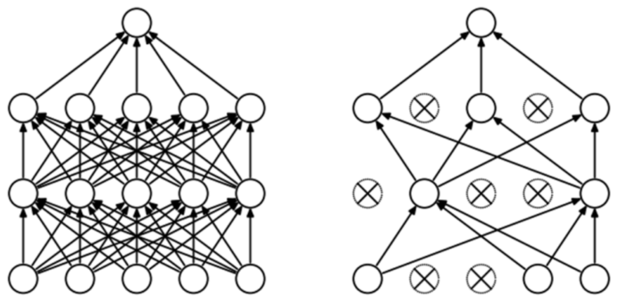

Original


Translation

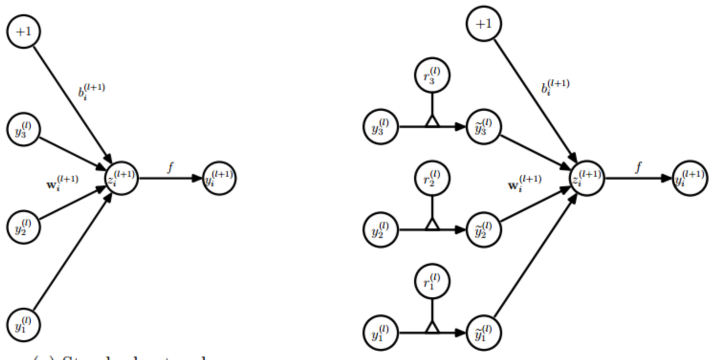

Horizontal flipping

- Dropout

Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from **co-adapting** too much. During training, dropout samples from an exponential number of different “thinned” networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has **smaller weights**. This significantly reduces overfitting and gives major improvements over other regularization methods.


(a) Standard Neural Net (b) After applying dropout.

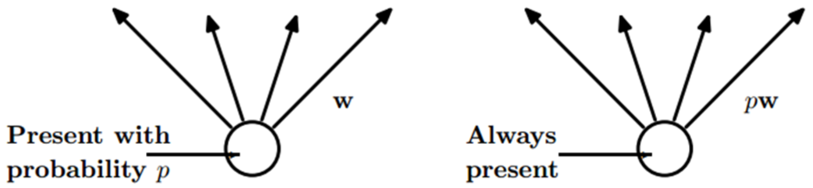
Network with Dropout



(a) Standard network (b) Dropout network

Figure 3: Comparison of the basic operations of a standard and dropout network.

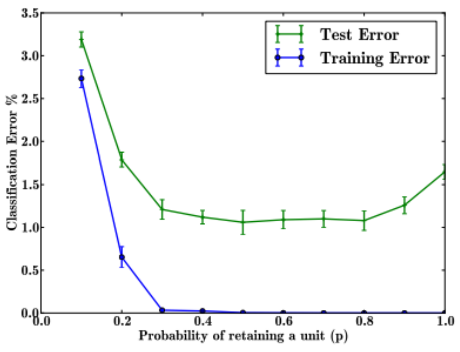
Dropout at testing



(a) At training time (b) At test time

The effect of Dropout rate

If the architecture is held constant, having a small p means very few units will turn on during training. It can be seen that this has led to *underfitting* since the training error is also high. We see that as p increases, the error goes down. It becomes flat when **$0.4 \leq p \leq 0.8$** and then increases as p becomes close to 1.



(a) Keeping n fixed.

Implementation Details:

- Training Hyperparameters:
  - Method: SGD with momentum
  - Mini-batch size: 128 (391 iterations for each epoch)
  - Total epochs: 164, momentum 0.9 (if you use momentum SGD)
  - Initial learning rate: 0.1, divide by 10 at 81, 122 epoch
  - Loss function: cross-entropy

- Data augmentation parameters:
  - Translation: Pad **4** zeros in each side and random cropping back to 32x32 size
  - Horizontal flipping: With probability **0.5**

Methodology:

- Estimated training time: ~1 hour with single titan X in Torch
- Estimated memory usage: ~1300 MB (without memory reduction)
- Estimated testing error rate: 12%~10% without data augmentation (10%~8% with)
- 9.47% with data augmentation in my implementation (8.81% in the paper)
- It may due to the preprocessing of input images (whitening & global normalization)

Extra Bonus:

- All convolutions NIN (remove pooling layers)
- Reproduce the experiment of Dropout rate in [2].

References:

[1] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[2] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

[3] https://www.tensorflow.org/tutorials/deep_cnn/

Report Spec:

1. Introduction
2. Experiment setup
   - The detail of your model
   - Report all your training hyper-parameters
3. Result
   - The comparison between with and without data augmentation
     - Final Test error
     - Training loss curve (you need to record training loss every epoch)
     - Test error curve (you need to record test error every epoch)
4. Discussion