机器学习复习2

4.2

**1. Try to give some examples of famous network structure.And why VGG, ResNet have many structures with different network layers ?**

**2. Why does ResNet (residual network) work?**

**2-2.There are two types of "skip connection block“ in ResNet named “identity block” and “convolutional block”. What’s the difference between them?**

**3. What does 1\*1 kernel convolution do?**

**3-2. What’s inception network’s motivation? Try to write down a “Inception module”.**

**4. What should we do when using a pre-train neural network to train a transfer learning classification task?**

1. There are many famous deep neural network models. Like LeNet-5, AlexNet, VGG, GoogLeNet, ResNet, Inception series, etc.

VGG and ResNet use multiple “blocks” with same padding convolution layer to construct their network models.

The output features of each “block” can stay the same dimension with input features, thus we can add different number of “blocks” in network and construct different model structures.

1. A very deep network is usually suffer from gradient vanish/explosion because all the layers need to learn the optimal function, which is hard to learn well.

Using short circuit connection, ResNet hopes each ‘block’ learns the residual function(the difference between the optimal function and the identity function), which is much easier to learn.

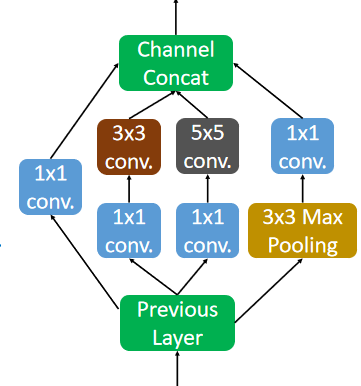
The short circuit connection allows the gradient to be directly back-propagated to earlier layers so ResNet can also avoid gradient vanish/explosion problem.

2-2. The “identity block” is the standard block used in ResNets, and corresponds to the case where the input activation has the same dimension as the output activation .

The "convolutional block“ is used when the input and output dimensions don't match up. The difference with the “identity block” is that there is a convolution layer in the shortcut path to resize the input to a correct dimension with output activation.

1. 1\*1 layer is also called “bottle neck” layer. It is equivalent to making a scale for the original feature map, and the scale size is trained by kernel parameters, which reduce the feature dimension (number of channels) and thus reduce the computation cost.

3-2. Inception network aims to increases the bandwidth of the network, increases the adaptability of the network to the input scale while reducing computational cost.



1. The number of predict classes of the target domain may be different with the source domain. Firstly we should redesign and train the softmax layer from the source domain to the target domain.

There is no need to train all network’s parameters since the front layers of pre-train network extract the generalized features, which are also useful in target domain. So we can just freeze the parameters of front layers and only train the behind layers of the pre-train network.

4.3

**3. What’s the Intersection Over Union(IoU) function?**

**4. Since some sliding windows (bounding boxes) may point out a same object because sliding stride is small, how to make sure the algorithm detects each object only once?**

**4-2. Try to write down a pseudo code to implement NMS algorithm.**

**5. Why should we use anchor boxes in object detection task?**

**6. YOLO algorithm is faster than Faster-RCNN algorithm but lower performance. Why?**

1. Intersection-over-Union (IoU), a concept used in target detection, is the Intersection rate of candidate bound and ground truth bound——the ratio of their Intersection to Union set. Usually if IoU score > 0.5 , that would be considered a good result.

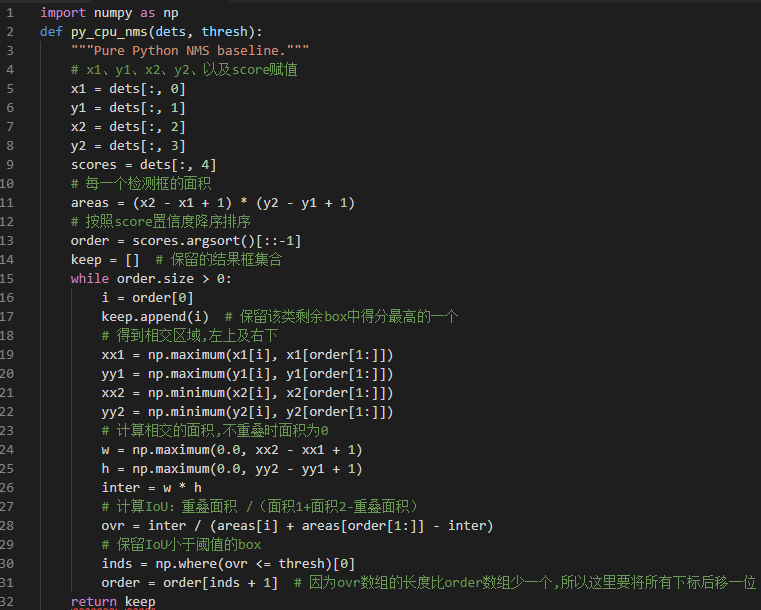


1. Non-max suppression algorithm is used to choose the best bounding box of each object.

Firstly it choose the highest probability bounding box for an object, then it suppressed bounding boxes which have high IoU score with the chosen box.

The algorithm would repeat several times until all bounding boxes are chosen or suppressed.

4-2.



1. Anchor boxes algorithm can avoid the problem that each grid cell can only detect one object.

Using anchor boxes, each object in training image is assigned to grid cell that contains object’s midpoint and anchor box for the grid cell with highest IoU. Thus each grid cell can detect multiple (number of anchor boxes) objects.

Anchor boxes also help achieve better result by specializing some outputs to special object. Such as person(tall and narrow object) and car(short and wide object).

1. YOLO makes predictions based on global information about the image.

Compared with Faster-RCNN, YOLO can reduce the error rate of false detection (the detection of background as an object) by more than half , but it’s performance like poor positional accuracy , poor detection of small objects and dense objects.

Faster-RCNN is a two steps method, where it firstly propose some regions that have an object with high probability and then classify it.

It reduce the class imbalance problem (the number of windows with not object is much larger than the number of windows with object.) but leads to higher computation cost.

4.4

**2. How can we deal with the problem, where we just have a few images in database in face recognition task?**

**2.2Try to write down the formula for “triplet loss”.**

**3. Is the one/few-shot learning problem belongs to transfer learning problem? Why?**

**4. How to define a loss function to train a style transfer network?**

**4.2Try to write down the formula for “content similarity” and “style similarity”.**

1. This problem is called “one/few shot learning”.

We can use triplet loss to train a siamese network. Triplet loss is defined by triples of image, the distance between positive and anchor samples should be smaller enough (with a margin) than the distance between negative and anchor samples.

Since there are too many negative samples, we should choose the “hard” negative samples (small distance with anchors) in training progress.

2-2.



The “margin”α is used to avoid the network learns all images equal to zero.

The max function is used to avoid the network just learns a small part of the training sets, where it just pull the distance of some samples very large and ignore the others.

1. Yes.

One/few shot learning means that only one/a few labeled samples per target class can be used in training phase. It’s a small sample learning problem.

Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

Small sample learning belongs to transfer learning, we can use a pre-train classification model to build a siamese network.

1. The loss function should has two parts to evaluate the content similarity and style similarity respectively.

Content similarity can be calculated as L2-distance between features of the content image and generated one. The features to compute content loss should be chose as the output of network’s middle layer.

Style similarity is defined as F-norm between correlation matrix of the content image and generated one. Correlation matrix means the activations across channels, images with same style should have similar correlation matrix.

4-2. Content similarity for layer l (choose a middle layer of network):



Style similarity:







And the total loss function:



5.1

**1.Why does standard DNN can NOT work well in time series problem?**

**2.Try to write down the forward propagation formula (or update equations) of basic RNN architecture in step t.**

**3.What problem have we encountered if all weights and activations in RNN are taking on the value of NaN during training?**

**4.What’s the difference about “simplified GRU” and “full GRU”?**

**5.How many gates in LSTM? Try to explain their role in LSTM respectively.**

1. Different from general tasks, sequence problem’s inputs and outputs can be different lengths in different examples, for example language translation, each sequence may have different number of words.

The same word in different positions of sequence may express the same meaning, but a standard DNN can NOT learns that because it does NOT share features learned across different positions of sequence.

If each word in sequence is presented as a one-hot vector, using whole sequence as the DNN’s input would also lead to huge number of parameters.



1. It’s a gradient explosion problem.

Because of the long-term dependence of back propagation, The gradient will be expressed as the product of multiple weights and the derivative of activation functions. In extreme cases, the value of the weight becomes so large that the result overflows (NaN value).

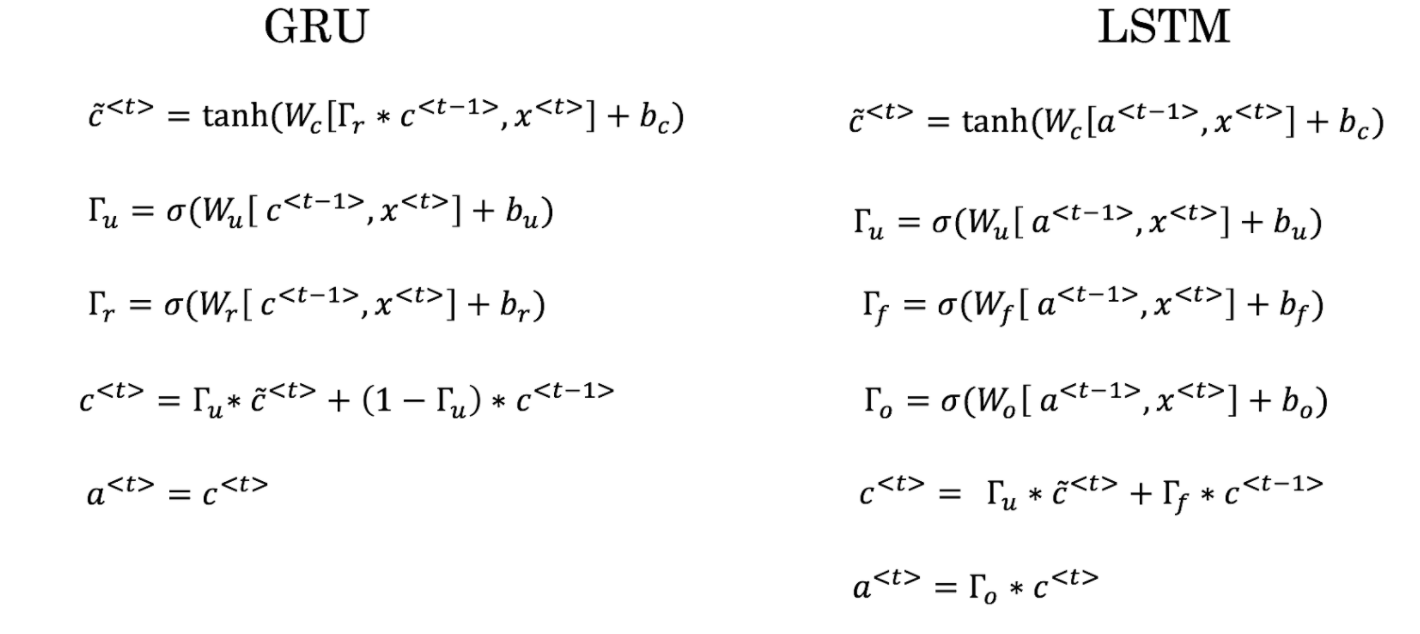
If the sample sequence is very long, it is easy to encounter the gradient explosion(or gradient vanish) problem in a standard RNN. As a solution, GRU and LSTM can solves this problem through the “Gate mechanism”.

1. Here is the update equations for “full GRU”.

For “Simplified GRU”, Andrew remove the Γr term(and set Γr = 1 to calculate  ).

Γr can be seen as a “relevance gate”, which determines how much information was forgotten in the past hidden layer outputs.

It helps capture short-term dependencies in time series data.



1. “Forget gate” Γf determines how much information was forgotten in the past hidden layer outputs.

“Update gate” Γu controls the degree of update of the hidden state  .

“Output gate” Γo determines what hidden layer value  to output.

 and  should both be output at a time step.Here is the update equations for LSTM.

