Search for questions, people, and topics

Ask New Question

Sign In

Convolutional Neural Networks Artificial Neural Networks Deep Learning Machine Learning

How does batch normalization behave differently at training time and test time?

5 Answers

Debiprasad Ghosh, Learned Machine learning for structural health monitoring
Answered May 31, 2016 · Upvoted by Carson Wu, M.S. Computer Vision & Deep Learning, Tongji University (2019)

Training will always be done with a mini-batch size. But as long as precomputed normalization parameters (means or variances) are used along with other trained network parameters, testing or inference can be performed with of any size.

Let me to explain what was answered by one of the author in [1] without expecting any Upvote from Quora users.

Consider you have many tagged picture of animals, you want to create an animal classifier, such that if you give any other picture of animal, your classifier can predict the class of the animal successfully.

Usually, to get confidence on your classifier, you have to divide your picture set into two subsets, one is for training of the network and other is for testing of the network.

Without using Batch Normalization

If you train your network without using batch normalization technique, either you can train your network with a single picture at a time or you can consider a mini-batch, where more than one number of training pictures will be used for single up-gradation of weights. Also, training in batches over one example at a time improves quality of gradient and parallelization efficiency.

Your trained network will classify any single picture during testing or inference. There is no necessity for testing with mini-batch. If your network successfully classifies your test samples, you are confident about your network for any animal picture to classify.

Using Batch Normalization

Alternatively, if you are interested to include batch normalization [2] in your classifier, during training you have to consider a mini-batch of pictures at a time. This requirement of mini-batch training is for learning of two extra parameters for every dimension for every hidden layers along with the weights associated with the network.

Necessity of these extra parameters is to get optimum scaling and shifting of activation outputs over zero means and unit variances towards elimination of

Still have a question? Ask your own!

What is your question? Ask

Related Questions

- Is batch normalization an absolute must for GAN training success?
- How does a person choose the best size of mini-batch in the test when the model is using batch normalization?
- Andrew Ng: How can people ultilize Deep Learning in an interactive manner other than the normal training-testing-deploying model?
- Will batch-normalization make difference when training with batch-size=1?
- Is batch normalization still effective in batch-sizes of 8-16 (in CNNs)? Could “moving average” help?
- Why does batch normalization help?
- What is the effect of batch normalization with a batch size of 1 in a deep neural network, in terms of performance and speed of training proce...
- When using batch normalization, where in the layer stack do you put it?
- How do I apply Batch Normalization to the convolutional layer of a CNN?
- Why does batch normalization enable higher learning rate?

+ Ask New Question

ɔnd sub-operation (scaling and shifting)
vith the other network parameter (weights)
rmalization parameter involved in first
change as the contents in every mini-

only on input and different trained
d shift and scale parameters. So,
identified during training, which will be

which will be used during inference, you
ated from many equal size mini-batch
ıg, you can track the moving averages of
ı be used for checking of accuracy of the
l normalization parameters.

eters are simply a linear transformation
be merged with respective trained scaling
ɔf the network.

your network with a mini-batch of
ıt should be the batch size during test.

f/1502.03167v...

- *Batch size during testing should not matter if you use pre-computed means and variances based on the activations on the training set.*
- *Another potential is to compute the mean and variation values based on your test-set activation distribution, but still not batch-wise. Also it should be helpful to combat domain transfer issues.*

[1] Christian Szegedy's answer to How does a person choose the best size of mini-batch in the test when the model is using batch normalization?

[2] Debiprasad Ghosh's answer to Why does batch normalization help?

9.6k Views · 15 Upvotes

Ask New Question Sign In

ɪh-sizes of 8-16 (in CNNs)? Could "moving

Related Questions

Is batch normalization an absolute must for GAN training success?

How does a person choose the best size of mini-batch in the test when the model is using batch normalization?

Andrew Ng: How can people ultilize Deep Learning in an interactive manner other than the normal training-testing-deploying model?

Will batch-normalization make difference when training with batch-size=1?

Is batch normalization still effective in batch-sizes of 8-16 (in CNNs)? Could "moving average" help?

Why does batch normalization help?

What is the effect of batch normalization with a batch size of 1 in a deep neural network, in terms of performance and speed of training proce...

When using batch normalization, where in the layer stack do you put it?

How do I apply Batch Normalization to the convolutional layer of a CNN?

Why does batch normalization enable higher learning rate?

+ Ask New Questio

ation: Accelerating Deep Network Training

ɪg/pdf/1502.03167v...

ɪen Batch Normalization (BN) calculates

t.

ɪiance base on mini-batch. You can see

$\mathcal{B} = \{x_{1...m}\}$;

// mini-batch mean

ɪini-batch variance

// normalize

// scale and shift

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

In test: While mean and variance were calculated using the population, rather than mini-batch, statistics.

normalization

$$\widehat{x} = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}}$$

using the population, rather than mini-batch, statistics. Ne-

The actual implement have a little difference with the original paper. For details implement, here is a tutorial Implementing Batch Normalization in Tensorflow from R2RT .

In training phase:

Step 1, the model will calculate batch_mean and batch_var base on the input batch

```
batch_mean, batch_var = tf.nn.moments(x,[0])
```

Step 2, pop_mean and pop_var will be update base on batch_mean, batch_var with decay. In this tutorial decay will be 0.999

```
decay = 0.999
```

Still have a question? Ask your owı

Ask

storm
_var + epsilon)

)p_var, skip step 1 and step 2 in training
pop_var directly:

+ epsilon)

Petersburg State University

Szegedy) it appears to be better (for
ne exact procedure for N=1 during
ınce over the entire image (normalization is
1). This is also known as Instance
Normalization and is shown to do better for Style Transfer as well as in practice for classification/detection networks, so appears to be better across the board.

https://arxiv.org/pdf/1607.08022...

It's also simpler/lends to better parallelization than BN and is closer in spirit to the good old Divisive Normalization technique from the 90s :p To speedup the process you can sparsely subsample the image in CUDA using coalesced reads (warp size=32) and striding the offset every, say 32*8 linear patches in W (can cross over to the next height line). This can rather significantly increase your inference accuracy, intuitively due to with standard process means/variances are naturally diverging as activations propagate down through the network. (planning to submit a paper to arxiv about it along with a fast CUDA kernel implementation). In tensorflow it's achieved by using is_training=True however cuDNN batch normalization is quite slow for some reason. The above CUDA kernel will substantially reduce this time. Another idea I had (that I haven't yet tested) is to insert batch norm only every 3-rd or 4-th layer, thus minimizing the divergence of mean/var over depth and using precomputed/static means/variances for the rest (which are "free" because they can be baked into the network). This would further reduce batch norm overhead while producing results close to "optimal", ie identical procedure during training and inference.

ou
N
ads

Ask New Question   Sign In

smatch between N>1 during training and
ı norm by batch (by that I mean
producing N*C values, of course it's no
it's 100% consistent between
tuitively becomes just per-channel
tional invariance property.

ıe learning

ı layers see the whole batch, compute the
rameters.

see only one test data point at a time, hence
; a whole batch is infeasible (and is
variance estimated during training. By the
ould give a good estimation.

mputer Vision and Deep Learning

.n` and `variance` statistics. In training, it
ıile in testing, they are obtained from the

Related Questions

Is batch normalization an absolute must for GAN training success?

How does a person choose the best size of mini-batch in the test when the model is using batch normalization?

Andrew Ng: How can people ultilize Deep Learning in an interactive manner other than the normal training-testing-deploying model?

Will batch-normalization make difference when training with batch-size=1?

Is batch normalization still effective in batch-sizes of 8-16 (in CNNs)? Could "moving average" help?

Why does batch normalization help?

What is the effect of batch normalization with a batch size of 1 in a deep neural network, in terms of performance and speed of training proce...

When using batch normalization, where in the layer stack do you put it?

How do I apply Batch Normalization to the convolutional layer of a CNN?

Why does batch normalization enable higher learning rate?

+ Ask New Questio

Related Questions

Why does batch normalization help?

What is the effect of batch normalization with a batch size of 1 in a deep neural network, in terms of performance and speed of training proce...

When using batch normalization, where in the layer stack do you put it?

How do I apply Batch Normalization to the convolutional layer of a CNN?

Why does batch normalization enable higher learning rate?

Does Residual Learning work without batch normalization?

What is the difference between dropout and batch normalization?

Can a dropout of neural network training stagnate when combined with batch normalization?

How does one use moving averages to track the accuracy of a model as it trains in Batch Normalization?

Is batch normalization over hyped for training deep neural networks?

+ Ask New Question

Still have a question? Ask your owı

Ask

Ask New Question Sign In

Related Questions

Is batch normalization an absolute must for GAN training success?

How does a person choose the best size of mini-batch in the test when the model is using batch normalization?

Andrew Ng: How can people ultilize Deep Learning in an interactive manner other than the normal training-testing-deploying model?

Will batch-normalization make difference when training with batch-size=1?

Is batch normalization still effective in batch-sizes of 8-16 (in CNNs)? Could “moving average” help?

Why does batch normalization help?

What is the effect of batch normalization with a batch size of 1 in a deep neural network, in terms of performance and speed of training proce...

When using batch normalization, where in the layer stack do you put it?

How do I apply Batch Normalization to the convolutional layer of a CNN?

Why does batch normalization enable higher learning rate?

+ Ask New Questio

Still have a question? Ask your owı

Ask