Batch Normalization

杜行健

翔英学院 2015 级

May 22, 2017

Overview

Covariate Shift

Given the same observation $X = x$, the conditional distributions of Y are the same in the source and target domains. However, the marginal distributuions of X may be different in the source and the target domains. Formally, we assume that $P_s(Y|X = x) = P_t(Y|X = x) forall x \in \chi$, but $P_s(X) \neq P$. This difference between the two domains is called *covariate shift*.

Internal Covariate Shift

In the case of deep networks, the input to each layer is affected by parameters in all the input layers. So even small changes to the network get amplified down the network. This leads to change in the input distribution to internal layers of the deep network and is known as internal covariate shift.
It is well established that networks converge faster if the inputs have been whitened (ie zero mean, unit variances) and are uncorrelated and internal covariate shift leads to just the opposite.

Vanishing Gradient

Saturating nonlinearities (like tanh or sigmoid) can not be used for deep networks as they tend to get stuck in the saturation region as the network grows deeper. Some ways around this are to use:

- Nonlinearities like ReLU which do not saturate
- Smaller learning rates
- Careful initializations

Vanishing Gradient

Saturating nonlinearities (like tanh or sigmoid) can not be used for deep networks as they tend to get stuck in the saturation region as the network grows deeper. Some ways around this are to use:

- Nonlinearities like ReLU which do not saturate
- Smaller learning rates
- Careful initializations

Vanishing Gradient

Saturating nonlinearities (like tanh or sigmoid) can not be used for deep networks as they tend to get stuck in the saturation region as the network grows deeper. Some ways around this are to use:

- Nonlinearities like ReLU which do not saturate
- Smaller learning rates
- Careful initializations

Standard Score

Let us say that the layer we want to normalize has d dimensions $X = (x_1, \cdot, x_d)$. Then, we can normalize the k^{th} dimension as follows:

$$\hat{x}^k = \frac{x^k - E[x^k]}{\sqrt{Var[x^k]}} \tag{1}$$

We also need to scale and shift the normalized values otherwise just normalizing a layer would limit the layer in terms of what it can represent. For example, if we normalize the inputs to a sigmoid function, then the output would be bound to the linear region only.
So the normalized input x^k is transformed to:

$$y^k = \lambda^k \hat{x} + \beta^k \tag{2}$$

where λ and β are parameters to be learned.

Moreover, just like we use mini-batch in Stochastic Gradient Descent (SGD), we can use mini-batch with normalization to estimate the mean and variance for each activation.
The transformation from x to y as described above is called Batch Normalizing Tranform. This BN transform is differentiable and ensures that as the model is training, the layers can learn on the input distributions that exhibit less internal covariate shift and can hence accelerate the training.
At training time, a subset of activations in specified and BN transform is applied to all of them.
During test time, the normalization is done using the population statistics instead of mini-batch statistics to ensure that the output deterministically depends on the input.

Advantages Of Batch Normalization

- Reduces internal covariant shift.
- Reduces the dependence of gradients on the scale of the parameters or their initial values.
- Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
- Allows use of saturating nonlinearities and higher learning rates.

Advantages Of Batch Normalization

- Reduces internal covariant shift.
- Reduces the dependence of gradients on the scale of the parameters or their initial values.
- Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
- Allows use of saturating nonlinearities and higher learning rates.

Advantages Of Batch Normalization

- Reduces internal covariant shift.
- Reduces the dependence of gradients on the scale of the parameters or their initial values.
- Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
- Allows use of saturating nonlinearities and higher learning rates.

Advantages Of Batch Normalization

- Reduces internal covariant shift.
- Reduces the dependence of gradients on the scale of the parameters or their initial values.
- Regularizes the model and reduces the need for dropout, photometric distortions, local response normalization and other regularization techniques.
- Allows use of saturating nonlinearities and higher learning rates.

Thanks for your attention!

Any questions?