Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

Sergey Ioffe
Christian Szegedy
Google, 1600 Amphitheatre Pkwy, Mountain View, CA 94043

Abstract

Training Deep Neural Networks is complicated by the fact that the distribution of each layer’s inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. We refer to this phenomenon as internal covariate shift, and address the problem by normalizing layer inputs. Our method draws its strength from making normalization a part of the model architecture and performing the normalization for each training mini-batch. Batch Normalization allows us to use much higher learning rates and be less careful about initialization, and in some cases eliminates the need for Dropout. Applied to a state-of-the-art image classification model, Batch Normalization achieves the same accuracy with 14 times fewer training steps, and beats the original model by a significant margin. Using an ensemble of batch-normalized networks, we improve upon the best published result on ImageNet classification: reaching 4.82% top-5 test error, exceeding the accuracy of human raters.

1. Introduction

Deep learning has dramatically advanced the state of the art in vision, speech, and many other areas. Stochastic gradient descent (SGD) has proved to be an effective way of training deep networks, and SGD variants such as momentum (Sutskever et al., 2013) and Adagrad (Duchi et al., 2011) have been used to achieve state of the art performance. SGD optimizes the parameters \( \Theta \) of the network, so as to minimize the loss

\[
\Theta = \arg \min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \ell(x_i, \Theta)
\]

where \( x_{1...N} \) is the training data set. With SGD, the training proceeds in steps, at each step considering a mini-batch \( x_{1...m} \) of size \( m \). Using mini-batches of examples, as opposed to one example at a time, is helpful in several ways. First, the gradient of the loss over a mini-batch \( \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \ell(x_i, \Theta)}{\partial \Theta} \) is an estimate of the gradient over the training set, whose quality improves as the batch size increases. Second, computation over a mini-batch can be more efficient than \( m \) computations for individual examples on modern computing platforms.

While stochastic gradient is simple and effective, it requires careful tuning of the model hyper-parameters, specifically the learning rate and the initial parameter values. The training is complicated by the fact that the inputs to each layer are affected by the parameters of all preceding layers – so that small changes to the network parameters amplify as the network becomes deeper.

The change in the distributions of layers’ inputs presents a problem because the layers need to continuously adapt to the new distribution. When the input distribution to a learning system changes, it is said to experience covariate shift (Shimodaira, 2000). This is typically handled via domain adaptation (Jiang, 2008). However, the notion of covariate shift can be extended beyond the learning system as a whole, to apply to its parts, such as a sub-network or a layer. Consider a network computing

\[
\ell = F_2(F_1(u, \Theta_1), \Theta_2)
\]

where \( F_1 \) and \( F_2 \) are arbitrary transformations, and the parameters \( \Theta_1, \Theta_2 \) are to be learned so as to minimize the loss \( \ell \). Learning \( \Theta_2 \) can be viewed as if the inputs \( x = F_1(u, \Theta_1) \) are fed into the sub-network

\[
\ell = F_2(x, \Theta_2).
\]
For example, a gradient descent step
\[ \Theta_2 \leftarrow \Theta_2 - \frac{\alpha}{m} \sum_{i=1}^{m} \frac{\partial F_2(x_i, \Theta_2)}{\partial \Theta_2} \]
(for mini-batch size \( m \) and learning rate \( \alpha \)) is exactly equivalent to that for a stand-alone network \( F_2 \) with input \( x \). Therefore, the input distribution properties that aid the network generalization – such as having the same distribution between the training and test data – apply to training the sub-network as well. As such it is advantageous for the distribution of \( x \) to remain fixed over time. Then, \( \Theta_2 \) does not need to readjust to compensate for the change in the distribution of \( x \).

Fixed distribution of inputs to a sub-network would have positive consequences for the layers outside the sub-network, as well. Consider a layer with a sigmoid activation function \( z = g(Wu + b) \) where \( u \) is the layer input, the weight matrix \( W \) and bias vector \( b \) are the layer parameters to be learned, and \( g(x) = \frac{1}{1 + \exp(-x)}. \) As \( |x| \) increases, \( g'(x) \) tends to zero. This means that for all dimensions of \( u = Wu + b \) except those with small absolute values, the gradient flowing down to \( u \) will vanish and the model will train slowly. However, since \( x \) is affected by \( W, b \) and the parameters of all the layers below, changes to those parameters during training will likely move many dimensions of \( x \) into the saturated regime of the nonlinearity and slow down the convergence. This effect is amplified as the network depth increases. In practice, the saturation problem and the resulting vanishing gradients are usually addressed by using Rectified Linear Units (Nair & Hinton, 2010) ReLU(\( x \)) = max(\( x, 0 \)), careful initialization (Bengio & Glorot, 2010; Saxe et al., 2013), and small learning rates. If, however, we could ensure that the distribution of nonlinearity inputs remains more stable as the network trains, then the optimizer would be less likely to get stuck in the saturated regime, and the training would accelerate.

We refer to the change in the distributions of internal nodes of a deep network, in the course of training, as Internal Covariate Shift. Eliminating it offers a promise of faster training. We propose a new mechanism, which we call Batch Normalization, that takes a step towards reducing internal covariate shift, and in doing so dramatically accelerates the training of deep neural nets. It accomplishes this via a normalization step that fixes the means and variances of layer inputs. Batch Normalization also has a beneficial effect on the gradient flow through the network, by reducing the dependence of gradients on the scale of the parameters or of their initial values. This allows us to use much higher learning rates without the risk of divergence. Furthermore, batch normalization regularizes the model and reduces the need for Dropout (Srivastava et al., 2014). Finally, Batch Normalization makes it possible to use saturating nonlinearities by preventing the network from getting stuck in the saturated modes.

In Sec. 4.2, we apply Batch Normalization to the best-performing ImageNet classification network, and show that we can match its performance using only 7% of the training steps, and can further exceed its accuracy by a substantial margin. Using an ensemble of such networks trained with Batch Normalization, we achieve the top-5 error rate that improves upon the best known results on ImageNet classification.

2. Towards Reducing Internal Covariate Shift

We define Internal Covariate Shift as the change in the distribution of network activations due to the change in network parameters during training. To improve the training, we seek to reduce the internal covariate shift. By fixing the distribution of the layer inputs \( x \) as the training progresses, we expect to improve the training speed. It has been long known (LeCun et al., 1998b; Wiesler & Ney, 2011) that the network training converges faster if its inputs are whitened – i.e., linearly transformed to have zero means and unit variances, and decorrelated. As each layer observes the inputs produced by the layers below, it would be advantageous to achieve the same whitening of the inputs of each layer. By whitening the inputs to each layer, we would take a step towards achieving the fixed distributions of inputs that would remove the ill effects of the internal covariate shift.

We could consider whitening activations at every training step or at some interval, either by modifying the network directly or by changing the parameters of the optimization algorithm to depend on the network activation values (Wiesler et al., 2014; Raiko et al., 2012; Povey et al., 2014; Desjardins & Kavukcuoglu). However, if these modifications are interspersed with the optimization steps, then the gradient descent step may attempt to update the parameters in a way that requires the normalization to be updated, which reduces the effect of the gradient step. For example, consider a layer with the input \( u \) that adds the learned bias \( b \), and normalizes the result by subtracting the mean of the activation computed over the training data: \( \hat{x} = x - E[x] \) where \( x = u + b, \mathcal{X} = \{x_1...N\} \) is the set of values of \( x \) over the training set, and \( E[x] = \frac{1}{N} \sum_{i=1}^{N} x_i \). If a gradient descent step ignores the dependence of \( E[x] \) on \( b \), then it will update \( b \leftarrow b + \Delta b \), where \( \Delta b \propto -\partial \ell/\partial \hat{x} \). Then \( u + (b + \Delta b) - E[u + (b + \Delta b)] = u + b - E[u + b] \). Thus, the combination of the update to \( b \) and subsequent change in normalization led to no change in the output of the layer nor, consequently, the loss. As the training continues, \( b \) will grow indefinitely while the loss remains fixed. This problem can get worse if the normalization not only centers but also scales the activations. We have observed this...
normalization speeds up convergence, even when the features are not decorrelated.

Note that simply normalizing each input of a layer may change what the layer can represent. For instance, normalizing the inputs of a sigmoid would constrain them to the linear regime of the nonlinearity. To address this, we make sure that the transformation inserted in the network can represent the identity transform. To accomplish this, we introduce, for each activation $x^{(k)}$, a pair of parameters $\gamma^{(k)}, \beta^{(k)}$, which scale and shift the normalized value:

$$ y^{(k)} = \gamma^{(k)} \bar{x}^{(k)} + \beta^{(k)}. $$

These parameters are learned along with the original model parameters, and restore the representation power of the network. Indeed, by setting $\gamma^{(k)} = \sqrt{\text{Var}[x^{(k)}]}$ and $\beta^{(k)} = E[x^{(k)}]$, we could recover the original activations, if that were the optimal thing to do.

In the batch setting where each training step is based on the entire training set, we would use the whole set to normalize activations. However, this is impractical when using stochastic optimization. Therefore, we make the second simplification: since we use mini-batches in stochastic gradient training, each mini-batch produces estimates of the mean and variance of each activation. This way, the statistics used for normalization can fully participate in the gradient backpropagation. Note that the use of mini-batches is enabled by computation of per-dimension variances rather than joint covariances; in the joint case, regularization would be required since the mini-batch size is likely to be smaller than the number of activations being whitened, resulting in singular covariance matrices.

Consider a mini-batch $B$ of size $m$. Since the normalization is applied to each activation independently, let us focus on a particular activation $x^{(k)}$ and omit $k$ for clarity. We have $m$ values of this activation in the mini-batch,

$$ B = \{x_{1\ldots m}\}. $$

Let the normalized values be $\tilde{x}_{1\ldots m}$, and their linear transformations be $y_{1\ldots m}$. We refer to the transform

$$ \text{BN}_{\gamma, \beta} : x_{1\ldots m} \to y_{1\ldots m} $$

as the Batch Normalizing Transform. We present the BN Transform in Algorithm 1. In the algorithm, $\epsilon$ is a constant added to the mini-batch variance for numerical stability.

The BN transform can be added to a network to manipulate any activation. In the notation $y = \text{BN}_{\gamma, \beta}(x)$, we indicate that the parameters $\gamma$ and $\beta$ are to be learned, but it should be noted that the BN transform does not independently process the activation in each training example. Rather, $\text{BN}_{\gamma, \beta}(x)$ depends both on the training example and the other examples in the mini-batch. The scaled and
shifted values \( y \) are passed to other network layers. The normalized activations \( \hat{x} \) are internal to our transformation, but their presence is crucial. The distributions of values of any \( \hat{x} \) has the expected value of 0 and the variance of 1, as long as the elements of each mini-batch are sampled from the same distribution, and if we neglect \( \epsilon \). This can be seen by observing that \( \sum_{i=1}^{m} \hat{x}_i = 0 \) and \( \frac{1}{m} \sum_{i=1}^{m} \hat{x}_i^2 = 1 \), and taking expectations. Each normalized activation \( \hat{z}^{(k)} \) can be viewed as an input to a sub-network composed of the linear transform \( y^{(k)} = \gamma^{(k)} \hat{z}^{(k)} + \beta^{(k)} \), followed by the other processing done by the original network. These sub-network inputs all have fixed means and variances, and although the joint distribution of these normalized \( \hat{z}^{(k)} \) can change over the course of training, we expect that the introduction of normalized inputs accelerates the training of the sub-network and, consequently, the network as a whole.

During training we need to backpropagate the gradient of loss \( \ell \) through this transformation, as well as compute the gradients with respect to the parameters of the BN transform. We use chain rule, as follows:

\[
\begin{align*}
\frac{\partial \ell}{\partial x_1} &= \frac{\partial \ell}{\partial y_1} \cdot \gamma \\
\frac{\partial \ell}{\partial y_1} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \cdot (x_i - \mu_B) \cdot \frac{-1}{\sqrt{\sigma_B^2 + \epsilon}} \\
\frac{\partial \ell}{\partial \mu_B} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial x_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \epsilon}} + \frac{\partial \ell}{\partial \sigma_B} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial \ell}{\partial \sigma_B} \cdot \frac{1}{m} \\
\frac{\partial \ell}{\partial \sigma_B^2} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \sigma_B} \cdot \hat{x}_i \\
\frac{\partial \ell}{\partial \gamma} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \\
\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \\
\end{align*}
\]

Thus, BN transform is a differentiable transformation that introduces normalized activations into the network. This ensures that as the model is training, layers can continue learning on input distributions that exhibit less internal covariate shift, thus accelerating the training. Furthermore, the learned affine transform applied to these normalized activations allows the BN transform to represent the identity transformation and preserves the network capacity.

### 3.1. Training and Inference with Batch-Normalized Networks

To Batch-Normalize a network, we specify a subset of activations and insert the BN transform for each of them, according to Alg. 1. Any layer that previously received \( x \) as the input, now receives \( \text{BN}(x) \). A model employing Batch Normalization can be trained using batch gradient descent, or Stochastic Gradient Descent with a mini-batch size \( m > 1 \), or with any of its variants such as Adagrad (Duchi et al., 2011). The normalization of activations that depends on the mini-batch allows efficient training, but is neither necessary nor desirable during inference; we want the output to depend only on the input, deterministically. For this, once the network has been trained, we use the normalization

\[
\hat{x} = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \epsilon}}
\]

using the population, rather than mini-batch, statistics. Neglecting \( \epsilon \), these normalized activations have the same mean 0 and variance 1 as during training. We use the unbiased variance estimate \( \text{Var}[x] = \frac{m}{m-1} \cdot E_B[\sigma_B^2] \), where the expectation is over training mini-batches of size \( m \) and \( \sigma_B^2 \) are their sample variances. Using moving averages instead, we can track the accuracy of a model as it trains. Since the means and variances are fixed during inference, the normalization is simply a linear transform applied to each activation. It may further be composed with the scaling by \( \gamma \) and shift by \( \beta \), to yield a single linear transform that replaces \( \text{BN}(x) \). Algorithm 2 summarizes the procedure for training batch-normalized networks.

### 3.2. Batch-Normalized Convolutional Networks

Batch Normalization can be applied to any set of activations in the network. Here, we focus on transforms that consist of an affine transformation followed by an element-wise nonlinearity:

\[
z = g(Wu + b)
\]

where \( W \) and \( b \) are learned parameters of the model, and \( g(\cdot) \) is the nonlinearity such as sigmoid or ReLU. This formulation covers both fully-connected and convolutional layers. We add the BN transform immediately before the nonlinearity, by normalizing \( x = W\hat{u} + \hat{b} \). We could also normalize the layer inputs \( u \), but since \( u \) is likely the output of another nonlinearity, the shape of its distribution is likely to change during training, and constraining its first and second moments would not eliminate the covariate shift. In contrast, \( W\hat{u} + \hat{b} \) is more likely to have a symmetric, non-sparse distribution, that is “more Gaus-
### 3.3. Batch Normalization enables higher learning rates

In traditional deep networks, too high a learning rate may result in the gradients that explode or vanish, as well as getting stuck in poor local minima. Batch Normalization helps address these issues. By normalizing activations throughout the network, it prevents small changes in layer parameters from amplifying as the data propagates through a deep network. For example, this enables the sigmoid nonlinearities to more easily stay in their non-saturated regimes, which is crucial for training deep sigmoid networks but has traditionally been hard to accomplish.

Batch Normalization also makes training more resilient to the parameter scale. Normally, large learning rates may increase the scale of layer parameters, which then amplify the gradient during backpropagation and lead to the model explosion. However, with Batch Normalization, backpropagation through a layer is unaffected by the scale of its parameters. Indeed, for a scalar $a$,

$$
\text{BN}(Wu) = \text{BN}((aW)u)
$$

and thus

$$
\frac{\partial \text{BN}(aWu)}{\partial a} = \frac{\partial \text{BN}(Wu)}{\partial a},
$$

so the scale does not affect the layer Jacobian nor, consequently, the gradient propagation. Moreover,

$$
\frac{\partial \text{BN}(aWu)}{\partial (aW)} = \frac{1}{a} \frac{\partial \text{BN}(Wu)}{\partial W},
$$

so larger weights lead to smaller gradients, and Batch Normalization will stabilize the parameter growth.

We further conjecture that Batch Normalization may lead the layer Jacobians to have singular values close to 1, which is known to be beneficial for training (Saxe et al., 2013). Consider two consecutive layers with normalized inputs, and the transformation between these normalized vectors: $\tilde{z} = F(\tilde{x})$. If we assume that $\tilde{x}$ and $\tilde{z}$ are Gaussian and uncorrelated, and that $F(\tilde{x}) \approx J\tilde{x}$ is a linear transformation for the given model parameters, then both $\tilde{x}$ and $\tilde{z}$ have unit covariances, and $I = \text{Cov}[\tilde{x}] = J\text{Cov}[\tilde{x}]J^T = JJ^T$. Thus, $J$ is orthogonal, which preserves the gradient magnitudes during backpropagation. Although the above assumptions are not true in reality, we expect Batch Normalization to help make gradient propagation better behaved. This remains an area of further study.

### 4. Experiments

#### 4.1. Activations over time

To verify the effects of internal covariate shift on training, and the ability of Batch Normalization to combat it, we considered the problem of predicting the digit class on the MNIST dataset (LeCun et al., 1998a). We used a very simple network, with a 28x28 binary image as input, and 3 fully-connected hidden layers with 100 activations each. Each hidden layer computes $y = g(Wu + b)$ with sigmoid nonlinearity, and the weights $W$ initialized to small random Gaussian values. The last hidden layer is followed
by a fully-connected layer with 10 activations (one per class) and cross-entropy loss. We trained the network for 50000 steps, with 60 examples per mini-batch. We added Batch Normalization to each hidden layer of the network, as in Sec. 3.1. We were interested in the comparison between the baseline and batch-normalized networks, rather than achieving the state of the art performance on MNIST (which the described architecture does not).

Figure 1(a) shows the fraction of correct predictions by the two networks on held-out test data, as training progresses. The batch-normalized network enjoys the higher test accuracy. To investigate why, we studied inputs to the sigmoid, in the original network $N$ and batch-normalized network $N_{BN}$ (Alg. 2) over the course of training. In Fig. 1(b,c) we show, for one typical activation from the last hidden layer of each network, how its distribution evolves. The distributions in the original network change significantly over time, both in their mean and the variance, which complicates the training of the subsequent layers. In contrast, the distributions in the batch-normalized network are much more stable as training progresses, which aids the training.

### 4.2. ImageNet classification

We applied Batch Normalization to a new variant of the Inception network (Szegedy et al., 2014), trained on the ImageNet classification task (Russakovsky et al., 2014). The network has a large number of convolutional and pooling layers, with a softmax layer to predict the image class, out of 1000 possibilities. Convolutional layers use ReLU as the nonlinearity. The main difference to the network described in (Szegedy et al., 2014) is that the $5 \times 5$ convolutional layers are replaced by two consecutive layers of $3 \times 3$ convolutions with up to 128 filters. The network contains $13.6 \cdot 10^6$ parameters, and, other than the top softmax layer, has no fully-connected layers. We refer to this model as Inception in the rest of the text. The training was performed on a large-scale, distributed architecture (Dean et al., 2012), using 5 concurrent steps on each of 10 model replicas, using asynchronous SGD with momentum (Sutskever et al., 2013), with the mini-batch size of 32. All networks are evaluated as training progresses by computing the validation accuracy @1, i.e. the probability of predicting the correct label out of 1000 possibilities, on a held-out set, using a single crop per image.

In our experiments, we evaluated several modifications of Inception with Batch Normalization. In all cases, Batch Normalization was applied to the input of each nonlinearity, in a convolutional way, as described in section 3.2, while keeping the rest of the architecture constant.

#### 4.2.1. Accelerating BN Networks

Simply adding Batch Normalization to a network does not take full advantage of our method. To do so, we applied the following modifications:

- **Increase learning rate.** In a batch-normalized model, we have been able to achieve a training speedup from higher learning rates, with no ill side effects (Sec. 3.3).
- **Remove Dropout.** We have found that removing Dropout from BN-Inception allows the network to achieve higher validation accuracy. We conjecture that Batch Normalization provides similar regularization benefits as Dropout, since the activations observed for a training example are affected by the random selection of examples in the same mini-batch.
- **Shuffle training examples more thoroughly.** We enabled within-shard shuffling of the training data, which prevents the same examples from always appearing in a mini-batch together. This led to about 1% improvement in the validation accuracy, which is consistent with the view of Batch Normalization as a regularizer: the randomization inherent in our method should be most beneficial when it affects an example differently each time it is seen.
- **Reduce the $L_2$ weight regularization.** While in Inception an $L_2$ loss on the model parameters controls overfitting, in modified BN-Inception the weight of this loss is reduced by a factor of 5. We find that this improves the accuracy on the held-out validation data.
- **Accelerate the learning rate decay.** In training Inception, learning rate was decayed exponentially. Because our network trains faster than Inception, we lower the learning rate 6 times faster.
- **Remove Local Response Normalization** While Inception and other networks (Srivastava et al., 2014) benefit from it, we found that with Batch Normalization it is not necessary.
- **Reduce the photometric distortions.** Because batch-normalized networks train faster and observe each training example fewer times, we let the trainer focus on more “real” images by distorting them less.
4.2.2. Single-Network Classification

We evaluated the following networks, all trained on the LSVRC2012 training data, and tested on the validation data:

**Inception**: the network described at the beginning of Section 4.2, trained with the initial learning rate of 0.0015.

**BN-Baseline**: Same as Inception with Batch Normalization before each nonlinearity.

**BN-x5**: Inception with Batch Normalization and the modifications in Sec. 4.2.1. The initial learning rate was increased by a factor of 5, to 0.0075. The same learning rate increase with original Inception caused the model parameters to reach machine infinity.

**BN-x30**: Like BN-x5, but with the initial learning rate 0.045 (30 times that of Inception).

**BN-x5-Sigmoid**: Like BN-x5, but with sigmoid nonlinearity \( g(t) = \frac{1}{1+\exp(-t)} \) instead of ReLU. We also attempted to train the original Inception with sigmoid, but the model remained at the accuracy equivalent to chance.

In Figure 2, we show the validation accuracy of the networks, as a function of the number of training steps. Inception reached the accuracy of 72.2% after 31 · 10^6 training steps. The Figure 3 shows, for each network, the number of training steps required to reach the maximum accuracy reached by the network.

![Figure 2. Single crop validation accuracy of Inception and its batch-normalized variants, vs. the number of training steps.](image)

In Figure 2, we show the validation accuracy of the networks, as a function of the number of training steps. Inception achieved the accuracy of 72.2% after 31 · 10^6 training steps. The Figure 3 shows, for each network, the number of training steps required to reach the maximum accuracy reached by the network.

<table>
<thead>
<tr>
<th>Model</th>
<th>Steps to 72.2%</th>
<th>Max accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception</td>
<td>31.0 · 10^6</td>
<td>72.2%</td>
</tr>
<tr>
<td>BN-Baseline</td>
<td>13.3 · 10^6</td>
<td>72.7%</td>
</tr>
<tr>
<td>BN-x5</td>
<td>2.1 · 10^6</td>
<td>73.0%</td>
</tr>
<tr>
<td>BN-x30</td>
<td>2.7 · 10^6</td>
<td>74.8%</td>
</tr>
<tr>
<td>BN-x5-Sigmoid</td>
<td></td>
<td>69.8%</td>
</tr>
</tbody>
</table>

**BN-x5-Sigmoid** achieves the accuracy of 69.8%. Without Batch Normalization, Inception with sigmoid never achieves better than 1/1000 accuracy.

4.2.3. Ensemble Classification

The current reported best results on the ImageNet Large Scale Visual Recognition Competition are reached by the Deep Image ensemble of traditional models (Wu et al., 2015) and the ensemble model of (He et al., 2015). The latter reports the top-5 error of 4.94%, as evaluated by the ILSVRC test server. Here we report a test error of 4.82% on test server. This improves upon the previous best result, and exceeds the estimated accuracy of human raters according to (Russakovsky et al., 2014).

For our ensemble, we used 6 networks. Each was based on BN-x30, modified via some of the following: increased initial weights in the convolutional layers; using Dropout (with the Dropout probability of 5% or 10%, vs. 40% for the original Inception); and using non-convolutional Batch Normalization with last hidden layers of the model. Each network achieved its maximum accuracy after about 6 · 10^6 training steps. The ensemble prediction was based on the arithmetic average of class probabilities predicted by the constituent networks. The details of ensemble and multicroc inference are similar to (Szegedy et al., 2014).

We demonstrate in Fig. 4 that batch normalization allows us to set new state-of-the-art on the ImageNet classification challenge benchmarks.
5. Conclusion

We have presented a novel mechanism for dramatically accelerating the training of deep networks. It is based on the premise that covariate shift, which is known to complicate the training of machine learning systems, also applies to sub-networks and layers, and removing it from internal activations of the network may aid in training. Our proposed method draws its power from normalizing activations, and from incorporating this normalization in the network architecture itself. This ensures that the normalization is appropriately handled by any optimization method that is being used to train the network. To enable stochastic optimization methods commonly used in deep network training, we perform the normalization for each mini-batch, and backpropagate the gradients through the normalization parameters. Batch Normalization adds only two extra parameters per activation, and in doing so preserves the representation ability of the network. We presented an algorithm for constructing, training, and performing inference with batch-normalized networks. The resulting networks can be trained with saturating nonlinearities, are more tolerant to increased training rates, and often do not require Dropout for regularization.

Merely adding Batch Normalization to a state-of-the-art image classification model yields a substantial speedup in training. By further increasing the learning rates, removing Dropout, and applying other modifications afforded by Batch Normalization, we reach the previous state of the art with only a small fraction of training steps – and then beat the state of the art in single-network image classification. Furthermore, by combining multiple models trained with Batch Normalization, we perform better than the best known system on ImageNet, by a significant margin.

Our method bears similarity to the standardization layer of (Gülçehre & Bengio, 2013), though the two address different goals. Batch Normalization seeks a stable distribution of activation values throughout training, and normalizes the inputs of a nonlinearity since that is where matching the moments is more likely to stabilize the distribution. On the contrary, the standardization layer is applied to the output of the nonlinearity, which results in sparser activations. We have not observed the nonlinearity inputs to be sparse, neither with nor without Batch Normalization. Other notable differences of Batch Normalization include the learned scale and shift that allow the BN transform to represent identity, handling of convolutional layers, and deterministic inference that does not depend on the minibatch.

In this work, we have not explored the full range of possibilities that Batch Normalization potentially enables. Our future work includes applications of our method to Recurrent Neural Networks (Pascanu et al., 2013), where the internal covariate shift and the vanishing or exploding gradients may be especially severe, and which would allow us to more thoroughly test the hypothesis that normalization improves gradient propagation (Sec. 3.3). More study is needed of the regularization properties of Batch Normalization, which we believe to be responsible for the improvements we have observed when Dropout is removed from BN-Inception. We plan to investigate whether Batch Normalization can help with domain adaptation, in its traditional sense – i.e. whether the normalization performed by the network would allow it to more easily generalize to new data distributions, perhaps with just a recomputation of the population means and variances (Alg. 2). Finally, we believe that further theoretical analysis of the algorithm would allow still more improvements and applications.

Acknowledgments

We thank Vincent Vanhoucke and Jay Yagnik for help and discussions, and the reviewers for insightful comments.
References


Dean, Jeffrey, Corrado, Greg S., Monga, Rajat, Chen, Kai, Devin, Matthieu, Le, Quoc V., Mao, Mark Z., Ranzato, Marc’ Aurelio, Senior, Andrew, Tucker, Paul, Yang, Ke, and Ng, Andrew Y. Large scale distributed deep networks. In NIPS, 2012.


Jiang, Jing. A literature survey on domain adaptation of statistical classifiers, 2008.


