CS7150.37338.202130 Day 7 Paper Questions

Xiaolin Zhao, Morgan Kohler, Aishwarya Vantipuli

TOTAL POINTS

6 / 6

QUESTION 1

1 Question 1 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts BN at test time for single image

- 0.5 pts BN for CNN

QUESTION 2

2 Question 2 1 / 1

✓ **- 0 pts Correct**

- 0.2 pts Interpretation

Good interpretation and challenge of the authors view.

QUESTION 3

3 Question 4 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts Click here to replace this description.

QUESTION 4

4 Question 5 1 / 1

✓ **- 0 pts Correct**

- 0.1 pts Interpretation

QUESTION 5

5 Question 6 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts Click here to replace this description.

QUESTION 6

6 Question 9 1 / 1

✓ **- 0 pts Correct**

CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 7 — Preparation Questions For Class
Due: Wednesday 2/10/2021 at 2:30pm via Gradescope

Names: [Morgan Kohler, Aishwarya Vantipuli, Xiaolin Zhao]

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written. Your answers should be as concise as possible. When asked to explain a figure, your response should have the following structure: provide context (state what experiment was being run / state what problem is being solved), state what has been plotted, remark on what we observe from the plots, and interpret the results.

Submit one document for your group and tag all group members. We recommend you use Overleaf for joint editing of this TeX document.

Directions: Read 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift' by Ioffe and Szegedy.

- Read entire paper.

Question 1. *In the context of CNNs, what is Batch Normalization? At test time, there may be only a single image passed into the network, and the variance of a quantity over only a single datapoint is undefined. How is this issue dealt with?*

Response:

Batch Normalization is an approach to reduce the internal covariate shift (which refers to the change in the distribution of network activations due to the change in layer input and parameters during training), and this is accomplished by introducing a normalization step that fixes the means and variances of layer inputs, thus the output distributions of a layer with respect to different batches are almost identical.

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

In the context of CNNs, individual feature maps are separately normalized. Mean and variance are taken over the combination of all training images. The expectation and variance are estimated at train time with only the minibatches (subsets).

For a single image at test time, the final result of the running average of mean and variance from training is used to normalize the layer outputs.

1 Question 1 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts BN at test time for single image

- 0.5 pts BN for CNN

Question 2. *Explain Figure 1.*

Response:

Context:

To verify the effects of internal covariate shift on training and the ability of Batch Normalization to combat it, a fully connected neural network with 3 hidden layer with and without Batch Normalization is trained separately on the MNIST dataset to classify the handwritten digits. The networks are trained for 50K steps, with 60 examples per mini-batch.

What is plotted:

Figure 1(a) describes, with training steps increased up to 50K, how the test accuracy changes for neural network trained with and without Batch Normalization.

In Figure 1(b) and 1(c), neural network trained without and with Batch Normalization respectively. The evolution of input distributions to a typical sigmoid activation from the last hidden layer over the course of training shown at 15, 50, 85 percentiles.

What we observe:

In figure 1(a) we observed that although both neural networks achieved almsot same test accuracy in the end, the batch-normalized network took much less time compared to the standard one, in other words, BN converges much faster with high test accuracy.

In figure 1(b) and 1(c) we observe that, without Batch Normalization, the distributions of all three percentiles change significantly over time both in their mean and variance. In contrast, with Batch Normalization, the distributions are much more stable over the whole training steps.

Interpretation:

Authors trying to prove Batch Normalization can help the network train faster and achieve higher accuracy by making the input distribution more stable and reduces the internal covariate shift. But the experiments shown in this paper are not sufficient to claim that BN reduces ICS to be true.

One reason could be because there's a hidden assumption for each layer to work well: input batches should be i.i.d, by applying batch normalization we assured this assumption to be true.

2 Question 2 1 / 1

✓ **- 0 pts Correct**

- 0.2 pts Interpretation

Good interpretation and challenge of the authors view.

Directions: Read 'How Does Batch Normalization Help Optimization?' by Santurkar et al.

- Read entire paper, except Sections 4-5.

Question 4. *Explain Figure 1.*

Response:

Context:

In order to explore the relation between Batch Norm and internal covariate shift (ICS), authors trained a standard VGG neural network on CIFAR-10 data with and without Batch Norm.

What is plotted:

Training and Testing accuracy of VGG network when trained in 15k steps with and without Batch Norm at different learning rates is plotted. Blue line indicates training/testing performance using Batch Norm with batch size of 128.Red line indicates same without batch normalization.

Also, distributions at layer 3 and layer 11 random inputs over training are plotted with(Blue) and without Batch Norm(Red) at 0.1 learning rate.

What we observe:

Fig (a) and Fig (b) shows a dramatic improvement in terms of optimization and generalization when trained using BatchNorm. In Fig (c) there is not much of a difference in distributional stability in networks with/without Batch Norm.

Interpretation:

The results reiterate the performance gains of BatchNorm stated in the original paper while further analyzing the underlying assumption of it's success. The plot of the layer distributions through time is more visually accessible than the distribution plots put forth in the original paper. Here we can see that the reduction in internal covariate shift, in this experiment, is perhaps not as drastic as stated in the original paper and assumed by the community. Visually the graphs look very similar with the standard network even having a more compact distribution than the standard + BatchNorm network. This result alone does not entirely disprove the original reasoning but this thinking is further bolstered later in the paper.

3 Question 4 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.

Question 5. *Explain Figure 2.*

Response:

Context:

In order to evaluate BatchNorm's success is due to controlling ICS, authors trained a new model by injecting random noise into a BatchNorm model using i.i.d. noise sampled from non-zero mean and non unit variance distribution. This model was compared with the other two models being VGG without BatchNorm (standard) and with BatchNorm (standard + BatchNorm).

What is plotted:

Training Accuracy over time is plotted as well as distributions of activations over time at various depths. This is for all the 3 models: Standard(red), with Batch Norm(blue), with Noisy Batch Norm(pink)

What we observe:

The performance difference between models with BatchNorm layersm and "noisy" BatchNorm layers is almost non-existent. Also, both these networks perform much better than standard networks. Moreover, the "noisy" BatchNorm network has qualitatively less stable distributions than even the standard, non-BatchNorm network, yet it still performs better in terms of training.

Interpretation:

This experiment intentionally added increased ICS and still observed an increase in performance relative to the standard network. The observation that the noisy BatchNorm model is almost identical to the performance of the regular BatchNorm network tells us that the performance gain due to BatchNorm likely does not stem from increased stability of layer input distributions. That even the noisy BatchNorm performs better than the standard network tells us that there is more to the success than reduction of ICS. Furthermore the visualization of layer distributions seems to indicate that the lowest ICS model is the standard model.

4 Question 5 1 / 1

✓ **- 0 pts Correct**

- 0.1 pts Interpretation

Question 6. *Explain Figure 3.*

Response:

Context:

The authors further test effect of BatchNorm on ICS by training a VGG network with and without BatchNorm and comparing various statistics directly related to ICS.

What is plotted:

Extent of ICS with and without BatchNorm layers for VGG and DLN are plotted. To isolate the effect of non-linearities as well as gradient stochasticity, they performed this analysis on (25-layer) deep linear networks (DLN).By calculating change in gradient(weights) at single input x, covariance shift is measured.

Training accuracy for VGG and Training Loss for DLN are plotted on left. For each layer cosine angle (ideally 1) and L2-difference of the gradients (ideally 0) before and after updates to the preceding layers is measured.

What we observe:

We observe that networks with BatchNorm often exhibit an increase in ICS.This is particularly striking in the case of DLN. In fact, in this case, the standard network experiences almost no ICS for the entirety of training.

Interpretation:

At low learning rates it is clear that BatchNorm has worse internal covariate shift than the standard network. While at higher learning rates, the covariate shift is roughly equal to that of the standard network. Yet we see in both the accuracy and loss graphs that the BatchNorm network outperforms the standard network for both small and large learning rates.

5 Question 6 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Click here to replace this description.

Question 9. *What is the point of having an explanation for why BatchNorm works?*

Response:

The authors of "How Does BatchNorm Help Optimization" hope that their work will shed light on the "underlying complexities of neural network training" and thus "inform further algorithmic progress in this area". So in other words, without the correct explanation it is unclear what direction future research should pursue. With only the ICS explanation, future research would likely focus on further reduction of the ICS and would likely not make significant progress. With a correct explanation like the smoothening of the loss landscape, then the research would likely focus on further increasing Lipschitzness. This direction would be the most likely to yield good results as there is more evidence to support this explanation. So without data-backed explanations, advancements in the field are essentially trial and error and up to random-chance. This is clearly unsatisfactory for efficient research in any field.

6 Question 9 1 / 1

✓ - 0 pts Correct