CS7150.37338.202130 Day 10 Paper Questions

Aishwarya Vantipuli, Harshita Ved, Aveek Choudhury

TOTAL POINTS

7 / 7

QUESTION 1

1 Question 1 **1 / 1**

✓ **- 0 pts** **Correct**

- 0.2 pts not clear where the weight decay term come from

QUESTION 2

2 Question 2 **1 / 1**

✓ **- 0 pts** **Correct**

- 0.2 pts Click here to replace this description.

QUESTION 3

3 Question 3 **1 / 1**

✓ **- 0 pts** **Correct**

QUESTION 4

4 Question 5 **1 / 1**

✓ **- 0 pts** **Correct**

QUESTION 5

5 Question 6 **1 / 1**

✓ **- 0 pts** **Correct**

QUESTION 6

6 Question 7 **1 / 1**

✓ **- 0 pts** **Correct**

QUESTION 7

7 Question 9 **1 / 1**

✓ **- 0 pts** **Correct**

- 0.2 pts Click here to replace this description.

CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 10 — Preparation Questions For Class
Due: Wednesday 2/24/2021 at 2:30pm via Gradescope

Names: Aishwarya Vantipuli, Aveek Choudhury, Harshita Ved

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 'ImageNet Classification with Deep Convolutional Neural Networks' (AlexNet).

- Read the entire paper

Question 1. *Explain all of the terms in the optimization algorithm presented in Section 5.*

Response:

Momentum - An exponentially weighted average of the prior updates to the weights can be included when the weights are updated. Here 0.9 momentum means the direction and speed at which the parameters move through parameter space.

Weight Decay - It is a regularization technique by adding a small penalty, usually the L2 norm of the weights to the loss function. Here 0.0005 (very low) value is used to avoid overfiting.

Learning Rate - The learning rate is a parameter that determines how much an updating step influences the current value of the weights. If it's large, correspondingly large modification of the weights w_i has to be made, not too large to overshoot local minima.

Gradient of Loss - It is the average of derivative of loss function with respect to weight over the batch D_i evaluated at w_i.

Authors used both momentum and weight decay on stochastic gradient descent to implement Learning rate schedule decay thus achieving optimal results.

SGD With Momentum

$$\Delta w_{i+1}(t+1) = -\epsilon \frac{\partial L}{\partial w_i} + \alpha \Delta v_i$$

SGD with Weight Decay

$$\Delta w_i(t+1) = -\epsilon \frac{\partial L}{\partial w_i} - \lambda \eta w_i$$

1 Question 1 1 / 1

✓ **- 0 pts** **Correct**

- 0.2 pts not clear where the weight decay term come from

Combined

$$\Delta w_i(t+1) = -\epsilon \frac{\partial L}{\partial w_i} - \lambda \eta w_i + \alpha \Delta v_i$$

Question 2. *The AlexNet paper used a learning rate schedule where the learning rate was lowered when validation error stopped improving. Why is it reasonable to have a schedule where learning rate decreases? Why wait until validation error stops improving (as opposed to imposing a specific schedule based on epoch number)?*

Response:

In order to achieve faster convergence, prevent oscillations and getting stuck in undesirable local minima the learning rate is often varied during training either in accordance to a learning rate schedule or by using an adaptive learning rate.Learning Rate Schedules seek to adjust the learning rate by reducing learning rate when there is no further improvement during training process thus preventing gradients stuck in local minima.

Reasons why authors didn't use step decay (epoch based) because in step decay, optimal values depends on dataset. Also in order to maintain large learning rate as long as possible it is a good idea to decrease learning rate when performance saturates.

2 Question 2 1 / 1

✓ **- 0 pts** **Correct**

- 0.2 pts Click here to replace this description.

Question 3. *Explain Figure 1.*

Response:

Context:

2 variants of a 4-layer neural network - one with ReLU and the other with tanh neurons were trained using the CIFAR-10 dataset for image classification in order to evaluate the performance benefits by using neurons with ReLU non-linearity. Learning rates for each network was chosen independently and no regularization was employed.

What is plotted:

The plot shows number of iterations taken by CNN using ReLU (solid line) and tanh (dashed line) to reach 0.25 training error.

What we observe:

CNN with ReLU converges six times faster than with tanh

Interpretation:

The observations leads us into inferring that the usage of neurons with non-linearity as ReLU enables faster convergence. Considering the size of neural network in this experiment, it would be difficult to train large networks and on big data sets using saturating functions such as tanh.

3 Question 3 1 / 1

✓ - 0 pts Correct

Question 5. *Explain the different data augmentation strategies used in the AlexNet paper. What do each of these strategies accomplish?*

Response: **First Augmentation Strategy:** Generating image translations and horizontal reflections. First, extracting random 224 × 224 patches (and their horizontal reflections) from the 256×256 images and then training network on these extracted patches. At test time, the network makes a prediction by extracting five 224 × 224 patches (the four corner patches and the center patch) as well as their horizontal reflections (hence ten patches in all), and averaging the predictions made by the network's soft-max layer on the ten patches.

This increases the size of our training set by a factor of 2048, though the resulting training examples are, of course, highly interdependent. Without this scheme, network suffers from substantial over-fitting, which would have forced us to use much smaller networks. This first strategy improves error rate by making the model position and reflection invariant.

Second Augmentation Strategy: Altering the intensities of the RGB channels in training images. Specifically, we perform PCA on the set of RGB pixel values throughout the ImageNet training set by eigenvalues, eigenvectors decomposition for each image. Then adding these eigenvectors with magnitude of their eigenvalue times some random number drawn from Gaussian distribution with 0 mean and 0.1 standard deviation.

The second strategy improves error rate by making the model invariant to changes in the intensity and color of the illumination. This scheme reduces the top-1 error rate by over 1%.

4 Question 5 1 / 1

✓ - 0 pts Correct

Directions: Read 'Deep Residual Learning for Image Recognition' (ResNets).

- Read Section 1, 3, 4.1

Question 6. *The ResNet paper reports 3.57% error on the ILSVRC. Some people would claim this performance is superhuman. Look up the rate of error achieved by humans. Why is the human error rate not 0%? (After all, wasn't it labelled by humans?) Do you think it is fair to say that this net can achieve superhuman performance at image classification?*

Response:

Humans have achieved an estimated 5.1 percent error rate. "To our knowledge, our result is the first to surpass human-level performance…on this visual recognition challenge," the researchers wrote.(**Source**)

The human error rate is not 0 because humans have no trouble distinguishing between a sheep and a cow. But computers are not perfect with these simple tasks. However, when it comes to distinguishing between different breeds of sheep, this is where computers outperform humans. The computer can be trained to look at the detail, texture, shape and context of the image and see distinctions that can't be observed by humans. Thus, a human like us, who does not have deep knowledge about species of animal, can barely distinguish between them. Thus humans are bound to make errors on some prediction, hence not 0. Also, the data is collected by different humans. search engine results are themselves tagged by other humans who are domain experts on those classes.

ResNets can be considered to be superhuman, where they can outperform an average human being at classification. However it remains to be seen if humans are trained with the entire dataset, whether they can be 100% accurate or not.

5 Question 6 1 / 1

✓ - 0 pts Correct

Question 7. *Explain the right column of Figure 3. Include the meaning of the text in each of the boxes, what the solid arrows mean, what the dashed arrows mean, what "pool, /2" and "avg pool" mean.*

Response:

Context:

A 34-layer residual network was trained on the ImageNet dataset for classification task.

What is plotted:

The architecture of the 34-layer residual network is plotted to show the configuration of the layers and the connections.

What we observe:

The network layers are coded using 6 colored boxes - orange (1 layer), purple (6 layers), green (8 layers), red (12 layers), blue (6 layers) and white (1 layer). Shortcut connections are represented every 2 convolution layers using arrows, solid arrows indicating identity shortcuts (when input and output dimensions are same) whereas the dotted arrows indicating increase in dimensions (different input and output dimensions).

The architecture starts with a 7x7 kernel - 64 filters - stride 2 layer (orange) which down-samples the input to 112x112x64, followed by a "pool,/2" block which indicates a max pool with stride 2 - further reducing the size of output to 56x56x64. The following layers follow the pattern of 3x3 covolutions with feature maps in the dimension of [64, 128, 256, 512], with shortcut connections (3) bypassing every 2 convolution layers.

The output from "pool,/2" is passed onto the blue layers having 3x3 kernel and 64 filters. Since the dimensions remain same, solid arrows used to show identity shortcuts ever 2 layers. Post the purple layers, the first green layer has a stride of 2 to reduce the size to 28x28x128. The shortcut connection here needs dimension change - identity mapping with extra padding to match dimensions (done by 1x1 convolutions) and hence shown using dotted arrow. Each of the following 7 green layer indicates 128 filters of 3x3 kernel, with solid shortcut arrows. Followed by the green layers is the set of 12 red-boxed layers which indicate 256 filters of 3x3 kernel, having shortcut connections every 2 layers (solid except the first). The first layer of the set of red boxes has a stride of 2 to reduce the output size to 14x14x256, and the shortcut is dotted line to indicate change in dimension. The set of 6 blue-boxed layers following the same pattern as red boxes, where the first layer has a stride of 2 to reduce the size to 7x7x512 and dotted shortcut arrow, and the rest of the layers having 512 filters of 3x3 kernel with solid shortcuts every 2 convolution layers.

The above layers are followed by a global average pooling layer indicated by "avg pool" and a 1000-way fully connected layer with softmax to map to the 1000 classes. The global average pooling layer can be seen as averaging all feature dimensions given by h*w*d to 1*1*d.

Interpretation:

When the shortcut connections are made for performing identity mappings, the network has no added parameters and comparable to the complexity of the non-residual nets.

6 Question 7 1 / 1

✓ - 0 pts Correct

Question 9. *Explain Figure 4 of the ResNet paper. Make sure to explain why there are two sudden steep drops in error % in both plots.*

Response:

Context:

18 and 34-layers neural network in their plain and residual architecture variants were trained on the ImageNet 2012 classification dataset in order to compare their training and validation errors. The dataset comprised of 1000 classes and the network was trained on 1.28 million training images, evaluated on 50K validation images along with testing on 100k test images.

What is plotted:

The training and validation errors during the training procedure are plotted as a function of iterations. The left plot shows the comparison for 18 and 34-layers plain network whereas the right plot shows the comparison for the residual network variants. The training error is denoted by the thin curve and the validation error by the bold curve.

What we observe:

On the left, i.e. plain architecture variant, we observe that the 34-layer network has a higher training error throughout the training procedure. The same phenomena is observed in the case of validation error as well, the 18-layer network having a lower error rate than the 34-layer network.

On the right, i.e. residual architecture variant, we observe the reverse phenomena where the 34-layer network performs better than the 18-layer network both in terms of training and validation errors. The 34-layer network exhibits considerably lower training error than the 18-layer network.

From both plots, we observe that the performance of the 18-layer plain and residual networks is comparable in terms of error rate. Also, we observe two sudden steep drop in error rate for all variants in both the plots.

Interpretation:

The 34-layer ResNet exhibits considerably lower training error and higher generalizability to the validation data indicating accuracy gains from increased depth and overcoming the degradation problem, unlike the plain architecture, also verifying the effectiveness of residual learning on extremely deep systems. ResNet can be seen as easing the optimization by providing faster convergence at the early stage.

The 2 sudden steep drops in error % in both the plots can be attributed to the learning rate schedule and the point of time during the training procedure when the rates were decreased to aid the training of the network. Since no clear indication is provided in the paper in terms of the type of schedule, it is difficult to interpret if the schedule is iteration or validation-error plateau-ing rate based. From the plot, we see that coincidentally, the drop occurs at the same iteration count for all the methods. Also, we can observe that the validation error start plateau-ing just before the drop occurs.

7 Question 9 1 / 1

✓ **- 0 pts** **Correct**

- 0.2 pts Click here to replace this description.