

Chapter 5

1. Why do we first resize to a large size on the CPU, and then to a smaller size on the GPU?

It's a data augmentation strategy called presizing. `aug_transforms` applies various data transforms, and starting from the target image dimensions can introduce artifacts such as empty zones or degrade data, etc due to not having enough data to work off of. Hence, `aug_transforms` have to be done from a larger dimension with the target dimension as the resulting image of the operation.

However, since `aug_transforms` is applied on the GPU, it expects the images to be of the same dimension so that they can be collated into a tensor. Therefore, the images are pre-resized on the CPU into relatively larger dimensions.

2. If you are not familiar with regular expressions, find a regular expression tutorial, and some problem sets, and complete them. Have a look on the book website for suggestions.

3. What are the two ways in which data is most commonly provided, for most deep learning datasets?

- Individual files representing items of data, such as text documents or images, possibly organized into folders or with filenames representing information about those items
- A table of data (e.g., in CSV format) in which each row is an item and may include filenames providing connections between the data in the table and data in other formats, such as text documents and images

4. Look up the documentation for L and try using a few of the new methods is that it adds.

5. Look up the documentation for the Python pathlib module and try using a few methods of the Path class.

6. Give two examples of ways that image transformations can degrade the quality of the data.

- Rotating an image by 45 degrees fills corner regions of the new bounds with emptiness
- Many rotation and zooming operations will require interpolating to create pixels. These interpolated pixels are derived from the original image data but are still of lower quality.

7. What method does fastai provide to view the data in a DataLoader?

`DataLoaders.show_batch`

8. What method does fastai provide to help you debug a DataBlock?

`DataLoaders.summary`

9. Should you hold off on training a model until you have thoroughly cleaned your data?

It is best to create a baseline model as soon as possible. The model will make the process of cleaning the data easier by highlighting outliers.

10. What are the two pieces that are combined into cross entropy loss in PyTorch?

Cross entropy loss is the loss created by calculating the negative log likelihood of the softmax of the predictions.

11. What are the two properties of activations that softmax ensures? Why is this important?

- The softmax of activations add up to 1.

- It amplifies small changes in activations. If one of the activations is slightly larger than the other, the softmax will amplify the difference.

This is important because softmax makes it easy to select one class over the others, making it ideal for training a classifier when each item has a definite label.

12. When might you want your activations to not have these two properties?

For multi-label classifications or classification problems where each item can have none or greater than 1 labels.

13. Calculate the “exp” and “softmax” columns of Fig 5-3 yourself.

```
import math
activations = [0.02, -2.49, 1.25]

exp_activations = [math.exp(act) for act in activations]
exp_activations
OUTPUT: [1.0202013400267558, 0.08290996657517266, 3.4903429574618414]

sm_activations = [act/sum(exp_activations) for act in exp_activations]
sm_activations
OUTPUT: [0.22209894370956398, 0.01804959009253818, 0.7598514661978979]

sum(sm_activations)
OUTPUT: 1.0
```

14. Why can't we use torch.where to create a loss function for datasets where our label can have more than two categories?

Because `torch.where` can only be used to select between two possibilities, and multi-class classification can have multiple possibilities which requires indexing.

15. What is the value of $\log(-2)$? Why?

Not defined. Because

```
x = e**y
log(x) = y
```

Since `e**y` is always positive, `x` is always positive. Therefore, $\log(-2)$ is not defined.

16. What are two good rules of thumb for picking a learning rate from the learning rate finder?

- One order of magnitude less than where the minimum loss is achieved
- The point where the slope of the loss is maximum before the minimum loss is achieved

17. What two steps does the `fine_tune` method do?

- Trains the head of the model (as a result of `.freeze()`) for `freeze_epochs` (usually 1) with the given learning rate `base_lr`.
- Unfreezes (as a result of `.unfreeze()`) and trains the whole model for `epochs` with learning rate varying from `base_lr/200` for inner layers to `base_lr/2` for the outermost layer.

18. In Jupyter notebook, how do you get the source code for a method or function?

Use `??` before or after the function

19. What are discriminative learning rates?

Discriminative learning rates refers to the training trick of using different learning rates for different layers of the model. This is commonly used in transfer learning. The idea is that when you train a pretrained model, you don't want to drastically change the earlier layers as it contains information regarding simple features like edges and shapes. But later layers may be changed a little more as it may contain information regarding facial feature or other object features that may not be relevant to your task. Therefore, the earlier layers have a lower learning rate and the later layers have higher learning rates.

20. How is a Python slice object interpreted when passed as a learning rate to fastai?

The first value of the slice object is the learning rate for the earliest layer, while the second value is the learning rate for the last layer. The layers in between will have learning rates that are multiplicatively equidistant throughout that range.

21. Why is early stopping a poor choice when using one cycle training?

Before the days of 1cycle training it was very common to save the model at the end of each epoch, and then select whichever model had the best accuracy out of all of the models saved in each epoch. This is known as early stopping. However, this is very unlikely to give you the best answer, because those epochs in the middle occur before the learning rate has had a chance to reach the small values, where it can really find the best result. Therefore, if you find that you have overfit, what you should actually do is retrain your model from scratch, and this time select a total number of epochs based on where your previous best results were found.

22. What is the difference between resnet 50 and resnet101

The number 50 and 101 refer to the number of layers in the models. Therefore, ResNet101 is a larger model with more layers versus ResNet50.

23. What does to_fp16 do?

This enables mixed-precision training, in which less precise numbers are used in order to speed up training in modern GPUs.