Chapter 2

1. Where do text models currently have a major deficiency?

Deep learning is good at generating context-approproate text, such as replies to social media posts or imitating a particular author's style. However, deep learning is not good at generating *correct* responses There is no way, for instance, to combine a knowledge base of medical information with a deep learning model for generating medically correct natural language responses.

2. What are the possible negative societal implications of text generation models?

Context-appropriate, highly compelling responses on social media could be used at a massive scale to spread disinformation, create unrest, amd encourage conflict.

3. In situations where a model might make mistakes, and those mistakes could be harmful, what is a good alternative to automating a process?

Deep learning should be used not as an entirely automated process, but as part of a process in which the model and a human user interact closely. The predictions of the deep learning model, specially the cases that can have high-priority effects, could be reviewed by human experts for them to evaluate if the prediction is indeed correct and what the next steps should be. This can potentially make humans orders of magnitude more productive than they would be with entirely manual methods, and result in more accurate processes than using a human alone.

4. What kind of tabular data is deep learning particularly good at?

Tabular data that has:

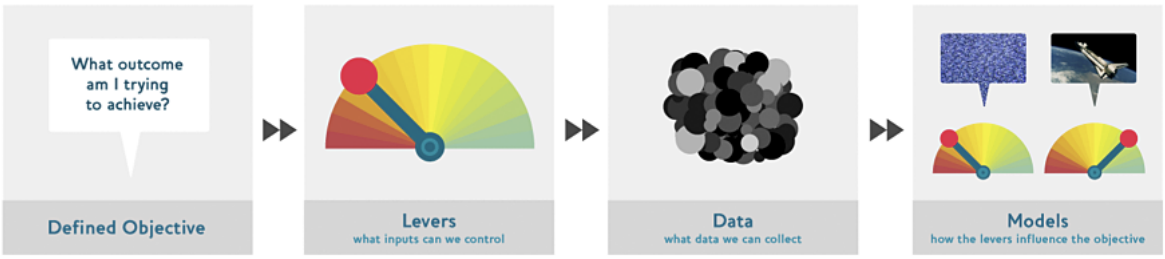
- columns with natural language (book titles, reviews, etc.)

- high-cardinality categorical columns (something that contains a large number of discrete choices, such as zip code or product ID)

5. What's a key downside of directly using a deep learning model for recommendation systems?

Machine learning approaches for recommendation systems have the downside that they tell you only which products a particular user might like, rather than what recommendations would be helpful for a user. For example, if a user has bought a book from an author, the model might recommend more books from the same author instead of similar-genre books from different authors.

6. What are the steps of the Drivetrain approach?



7. How do the steps of the Drivetrain approach map to a recommendation system?

- The objective of the recommendation system is to drive additional sales by surprising and delighting the customer with recommendations of items they would not have purchased otherwise. More sales.
- The lever is the ranking of the recommendations.
- New data must be collected to generate recommendations that will cause new sales. This will require conducting many randomized experiements in order to collect data about a wide range of recommendations for a wide range of customers.
- Finally, you could build two models for purchase probabilities, conditional on seeing or not seeing a recommendation.

8. Create an image recognition model using data you curate, and deploy it on the web.

—

9. What is DataLoaders?

`DataLoaders` is a fastai class that stores multiple `DataLoader` objects — usually a train object and a valid object. A `DataLoader` is a PyTorch class that provides batches of items to the GPU at a time.

10. What four things do we need to tell fastai to create DataLoaders?

- What kind of data are we working with
- How to get the list of items
- How to label the items
- How to create the validation set

11. What does the splitter parameter to DataBlock do?

The `splitter` parameter in `DataBlock` provides various ways of splitting the dataset into various subsets — usually training and validation.

12. How do we ensure a random split always gives the same validation set?

Computers don't really know how to create random numbers, but simply create lists of numbers that look random. If you provide the same starting point for the list each time — called the seed — you will get the exact same list every time. By passing the same value for the `seed` argument to `RandomSplitter`, it can be ebsured that the random split always gives the same validation set.

13. What letters are often used to signify the independent and dependent variables?

Independent → x

Dependent → y

14. What's the difference between crop, pad, and squish Resize() approaches? When might you choose one over the other?

- crop is the default Resize() method, and it crops the images to fit a square shape of the size requested, using the full width or height. This can result in losing some important details.
- pad is an alternative Resize() method, which pads the matrix of the image's pixels with zeros (which shows as black when viewing the images).
- squish is another alternative Resize() method, which can either squish or stretch the image.

Depends on the problem, if the features in the image exist acrss the whole image, the cropping can result in loss of information, and padding or squishing could yield a better result.

15. What is data augmentation? Why is it needed?

Data augmentation refers to creating random variations of our input data, such that they appear different, but do not change the meaning of the data. Examples of data augmentation techniques for images are rotation, flipping, perspective warping, brightness change, and contrast change.

Data augmentation helps the model better understand the basic concept of an object and how objects of interest are represented in images. Therefore, data augmentation allows machine learning models to generalize.

16. Provide an example of where the bear classification model might work poorly, due to structural or style differences to the training data.

An example would be night-time images as the model has not been trained on such data.

17. What is the difference between item_tfms and batch_tfms?

- `items_tfms` are transformations applied to the data individually on the CPU

- `batch_tfms` are transformation applied to the data a batch at a time on the GPU

18. What is a confusion matrix?

A confusion matrix is a representation of the predictions made vs the correct labels. The rows of the matrix represent the actual labels while the columns represent the predictions. Therefore, the number of images in the diagonal elements represent the number of correctly classified images, while the off-diagonal elements are incorrectly classified images.

19. What does export save?

`export` saves both the architecture, as well as the trained parameters of the neural network architecture. It also saves how the `DataLoaders` are defined.

20. What is it called when we use a model for getting predictions, instead of training?

Inference.

21. What are IPython widgets?

Widgets built with a combination of JavaScript and Python that allow for programming of interactable GUI components directly in Jupyter Notebooks.

22. When might you want to use CPU for deployment? When might GPU be better?

For inference on single pieces of data, CPUs are better. For inference on batches of data, GPUs are better. Inference for certain batches of data at a small scale, such as text, can be carried out on a CPU.

23. What are the downsides of deploying your app to a server, instead of to a client (or edge) device such as a phone or PC?

- Network connection can lead to extra latency that can lead to the feeling that the inference is taking more time that it actually does.
- Sending private data to a server can lead to security concerns.

24. What are 3 examples of problems that could occur when rolling out a bear warning system in practice?

- Night-time images
- Low-resolution images
- The model returns predictions too slowly to be useful

25. What is *out of domain data*?

Data that is fundamentally different in some way from the training data. For example, when a bear detector that is trained on bear photos is given a drawing of a bear.

26. What is *domain shift*?

When the type of data changes over time, the model stops being valid because the data is different from the original training data in some domain. For example, an insurance company is using a deep learning model as part of their pricing algorithm, but over time their customers will be different, with the original training data not being representative of current data, and the deep learning model being applied on effectively out-of-domain data.

27. What are the 3 steps in the deployment process?

1. Manual process – the model is run in parallel and not directly driving any actions, with humans still checking the model outputs.
2. Limited scope deployment – The model's scope is limited and carefully supervised. For example, doing a geographically and time-constrained trial of model deployment, that is carefully supervised.
3. Gradual expansion – The model scope is gradually increased, while good reporting systems are implemented in order to check for any significant changes to the actions taken compared to the manual process (i.e. the models should perform similarly to the humans, unless it is already anticipated to be better).