Your grade: 100%

Your score: 100% | To pass: 80% or more

English • Grade Report

Q. To help you prototype strategies for machine learning, this week we'll present another scenario and see how you would act. We'll call this "simulation" or tackling a proactive learning project with goal as phase of what would be a machine learning project.

You are employed by a startup building self-driving cars. You are in charge of detecting road signs (stop sign, pedestrian crossing sign, construction ahead sign and traffic signal; red and green lights) in images. The goal is to provide a dataset to train a DL model to detect these three classes of road signs. An example of the image collection is pedestrian crossing sign and a traffic light.

Q1 stop sign
Q2 pedestrian crossing sign
Q3 construction ahead sign
Q4 red traffic light
Q5 green traffic light

Your 350,000 labeled images are taken using the front facing camera of a car. This is also the distribution of images you expect to see in production. You have a budget of $100,000 to spend on data collection, labeling and cleaning. Your internal, which you're expected to lead, is training a part of the distribution of interest (data in the scene).

You are getting started with this project. What is the first thing you do? Assume each of the steps below would take about an equal amount of time (a few days).

1 / 1 point

- Your goal is to detect road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. Therefore, you'll need to collect a dataset of images containing these objects, and then also label which of these objects are present in each image. Therefore, the collected dataset should be a good choice for this subject task because this is a multi-task learning problem.

True/False?

- False
- True

Expand

Correct
This dataset is a good choice if one wants one of the possibilities (stop sign, speed limit, pedestrian crossing, green light and red light) we present in each image.

- You can only go to error analysis and mislabeling and skip errors the algorithm makes. Which of these statements do you think you should manually go through and identify and correct, one image at a time?

1 / 1 point

- 350 images in which the algorithm makes a mistake
- 10,000 images in which the algorithm made a mistake
- 100 randomly chosen images
- 10,000 randomly chosen images

Expand

Correct
Anyone an image that the algorithm got wrong. Also, 100 is enough to give you a good initial sense of the error statistics. There's probably no need to have a look at 10,000 which will take a long time.

- After working on the data for several weeks, your team ends up with the following data:

1 / 1 point

- 350,000 labeled images taken using the front facing camera of your car
- 340,000 labeled images of roads downloaded from the internet
- 8,000 images taken using the front facing camera of your car that contain a pedestrian crossing sign

combinations of these, for example, red traffic light.

Q1 Choose the image contains a stop sign and a red traffic light.

Q2 Choose the image contains a stop sign and a green traffic light.

Because this is a multi-task learning problem, when an image is not fully labeled (for example, a car can see a red traffic light but cannot see a stop sign), you can still use it for a given those entries when calculating the loss function. True/False?

- False
- True

Expand

Correct
Correct. We can't use the components of the labels that are missing but we can use the ones we have to train the model.

- The distribution of data you care about (validation/test/production data) is different than the distribution of data you have in your training set. Which of the following statements do you think is true? Assume that all the data you have in the internet, they should help with the dataset mismatch problem.

1 / 1 point

- 5% of the 350,000 images with the 350,000 images you found online. Adding everything (with the 350,000 images between the 340,000 for the training set, 350,000 for the dev and 350,000 for the test set) ...
- Choose the training set to be the 350,000 images from the internal along with 20,000 images from the internal and 20,000 images from the internal. Adding everything with the 20,000 images from the internal, the 350,000 internal images will be split equally to dev and test sets.
- Choose the training set to be the 350,000 images from the internal along with 20,000 images from the internal. Adding everything with the 20,000 images from the internal, the 350,000 internal images will be split equally to dev and test sets.
- 5% of the 350,000 images with the 350,000 images you found online. Adding everything (with the 350,000 images between the 340,000 for the training set, 10,000 for the dev set and 10,000 for the test set)

Expand

Correct
Answer: Yes, this seems to be the best. It is important that your dev and test set have the closest possible distribution to "real" data. It is also important for the training set to contain enough "real" data to avoid having a data mismatch problem.

- Assume you've finally chosen the following split between the data:

1 / 1 point

Dataset	Contains	Error of the algorithm
Training	340,000 images randomly picked from 340,000 internet images + 10,000 car's front facing camera images	2.7%
Training	350,000 images randomly picked from 340,000 internet images + 10,000 car's front facing camera images	23.0%
Dev	10,000 images from car's front facing camera	27.0%
Test	10,000 images from car's front facing camera	22.4%

You also know that human level error on the road signs and traffic signals classification task is around 0.4%. Which of the following is true?

- You have a low bias and high variance problem.
- You have a high bias and a low variance problem.
- You have a large data mismatch problem.
- You have a high bias.

Expand

Correct
Correct. The available data is significantly high about the training set is a lot higher than the human level error.

- Assume you've finally chosen the following split between the data:

1 / 1 point

Dataset	Contains	Error of the algorithm
Training	340,000 images randomly picked from 340,000 internet images + 10,000 car's front facing camera images	6.0%
Training	350,000 images randomly picked from 340,000 internet images + 10,000 car's front facing camera images	6.2%
Dev	20,000 images from car's front facing camera	26.0%
Test	20,000 images from car's front facing camera	25.6%

You also know that human level error on the road signs and traffic signals classification task is around 0.4%. Which of the following is true? (Assume that the training data distribution is much better than the dev/test distribution. What do you think?)

- Your bias is high. Bias error is the training data distribution is probably lower than the dev/test distribution.
- Your variance is high. Bias error is the training data distribution is probably higher than the dev/test distribution.
- There's insufficient information to tell if your bias is high or variance.

Expand

Correct
Correct. The distribution data better on the distribution of data is lower or. But you don't know if it's because it's easier or if the distribution is 0.4 only is better. To get a better sense, measure human level and separately on both distributions.

- You decide to focus on the dev set and check by hand what the errors are. Here is a table summarizing your observations:

1 / 1 point

Overall dev set error	26.0%
Errors due to incorrectly labeled labels	0.3%
Errors due to fogge pictures	4.0%
Errors due to partially occluded elements	8.0%
Errors due to other causes	1.0%

Which of the following statements do you agree with?

- 12% would be a reasonable estimate of how much the established error could improve performance.
- 12% would be a reasonable estimate of the maximum amount the established error could improve performance.
- 12% would be a reasonable estimate of the minimum amount the established error could improve performance.
- 12% would be a reasonable estimate of how much the established error could improve performance.

Expand

Correct
Yes. You will probably not improve performance by more than 12% by solving the various problems. Therefore, if you're looking for a 10% improvement, it seems worthwhile to pick a promising error to address by assembling a specially designed dev/test set that reviews the problem.

- You decide to use data augmentation to address foggy images. You find 1,000 pictures of fog off the Internet, and blend them with your front facing camera images. For example:

1 / 1 point

Image from front facing camera

Image from the internet

synthesized foggy image

You can now take each image that has a different distribution from the various each (internet and front facing camera). True/False?

- False
- True

Expand

Correct
Correct. The new synthesized images are added to the training set and as long as they look realistic to the human eye this will be useful data to train the model.

- After working further on the problem, you've decided to correct the incorrectly labeled data on the dev set. Which of these statements do you agree with? (Check all that apply.)

1 / 1 point

- You can be reasonably sure that the incorrectly labeled data in the training set because the algorithm is robust to random errors in the training set. Therefore it is not important that the dev set and test set have the same distribution.
- Correct
  The data learning algorithms are quite robust to having slightly different train and dev/test distributions.
- You should also correct the incorrectly labeled data in the test set, so that the dev and test set continue to come from the same distribution.
- Correct
  Yes because you want to make sure that your dev and test data come from the same distribution for your algorithm to make your team's development process efficient.
- You should correct manually identify the mislabeled data in the training set as well as in the training set since being even more different from your dev set.
- You should not correct the incorrectly labeled data in the dev set, so that the dev and test sets continue to come from the same distribution.

Expand

Correct
Correct, since you get all the right answers.

- Your client asks you to add the capability to detect stop signs that may be crossing the road by the system. He can provide a relatively small set containing stop. Which of the following do you agree most with?

1 / 1 point

- Using your trained algorithm you can easily train the ability of the system to detect stop signs. Therefore, you don't need any more data for this.
- You can use weights pre-trained on the original data, and fine-tune with the data now including the stop.
- You will have to re-train the whole model now including the stop data.
- You should train a single new model for the stop sign, and combine the previous model as a

Expand

Correct
Correct. Since your model has learned and has most features to handle the new task we can balance therefore using the pre-trained weights.

- One of your colleagues at the startup is starting a project to classify stop signs in the road as speed limit signs and yield signs. He proposes training a single multi-task learning algorithm to recognize all three types of road signs, using your model as starting point. Which of the statements that you can think both learning. Hence he can't use your model. True/False?

1 / 1 point

- False
- True

Expand

Correct
Correct. Multi-task learning is about training several tasks that can benefit from each other. Since there are a different set of data and features being used from that do not share weights (i.e. conv layers), this problem has a single task learning neural networks and not a multi-task learning.

- To recognize a stop sign you use the following approach:

1 / 1 point

First, we localize all the stop signs in an image. After that, we determine if the sign is a stop sign or not.

This is a better approach than one in which we train models for which of the following cases? Choose the best answer.

- There is a large amount of data.
- There are a plenty models which are all open to transfer knowledge.
- There is not enough data to train a big neural network.
- The problem has a high dispersion.

Expand

Correct
Correct. This might be the most important factor when deciding whether to use an end-to-end approach.