ML w11 photo OCR

Photo OCR process. — pipeline

1. Text detection } Sliding windows
2. Character segmentation }
3. Charactor recognition (classification) → e.g. neural network

Artificial data synthesis for photo OCR

amplify training set by distorting image.

· Usually does not help to add purely random / meaningless noise to data

1. Before expanding data

low bias classifier / high variance classifier

→ more data (get more traingset)

not low bias classifier

→ add features

2. How much work would it be to get 10 x as much data as we currently have?

- Artificial data synthesis.

- Collect / label it yourself
  - → how many hours?

    m = 1000

    10 sec / example

    m = 10,000

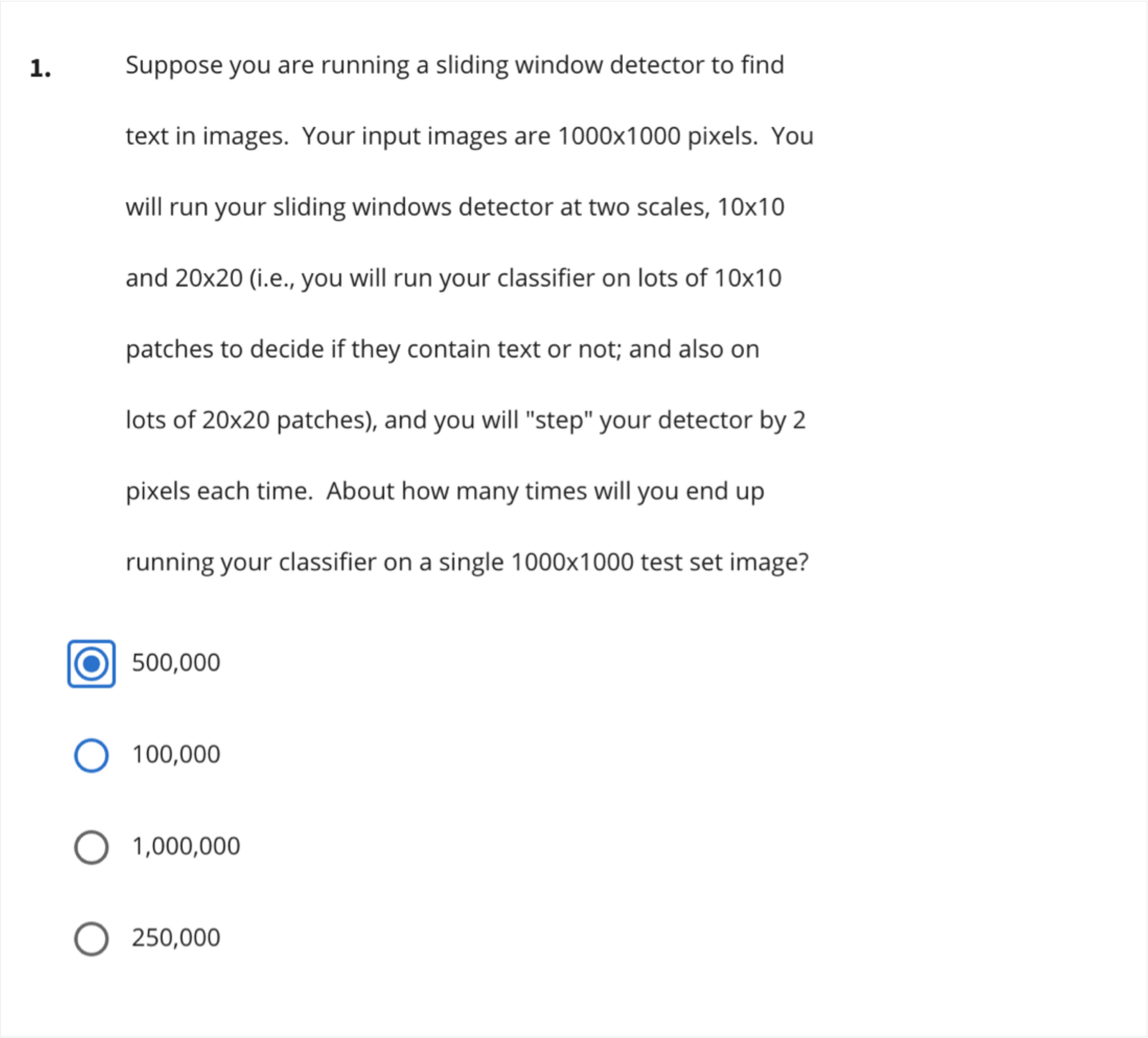
- "Cloud sourcing" (E.g. Amazon Mechanical Turk)

Ceiling Analysis: What part of the pipelines to work on next.
/ should you spend the most time

Component	Accuracy	
Overall system	72%	
Text detection	89%	17%
Charactor Segmentation	90%	1%
Charactor recognition	100%	10%

- 18 ヶ月 for background deletion did not help the overall accuracy ....

- Don't trust my own feeling ...



1. Suppose you are running a sliding window detector to find text in images. Your input images are 1000x1000 pixels. You will run your sliding windows detector at two scales, 10x10 and 20x20 (i.e., you will run your classifier on lots of 10x10 patches to decide if they contain text or not; and also on lots of 20x20 patches), and you will "step" your detector by 2 pixels each time. About how many times will you end up running your classifier on a single 1000x1000 test set image?

● 500,000

○ 100,000

○ 1,000,000

○ 250,000

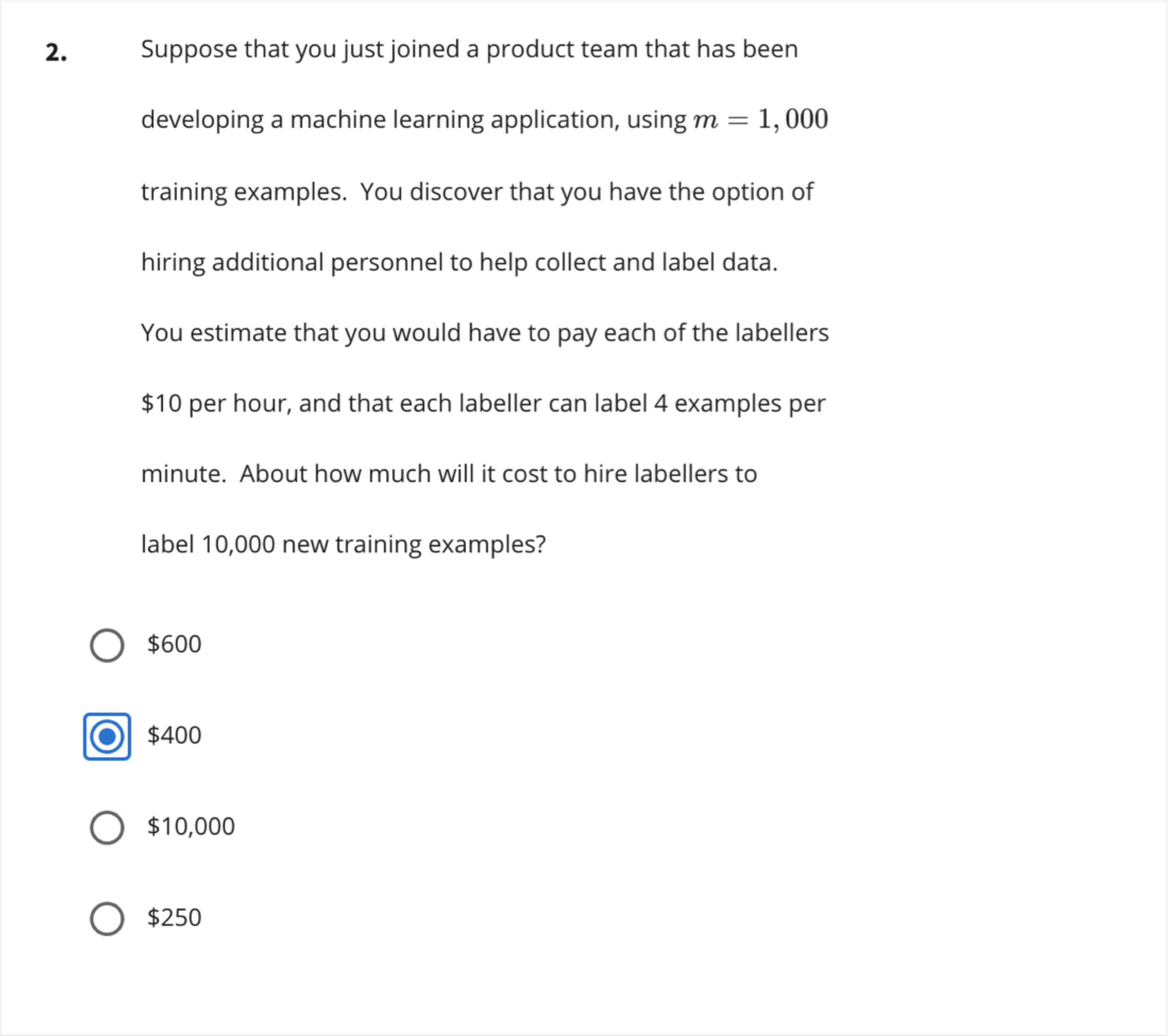
4 example / min

240 example / hour

10000 example : 41.6 hour

↓

$400



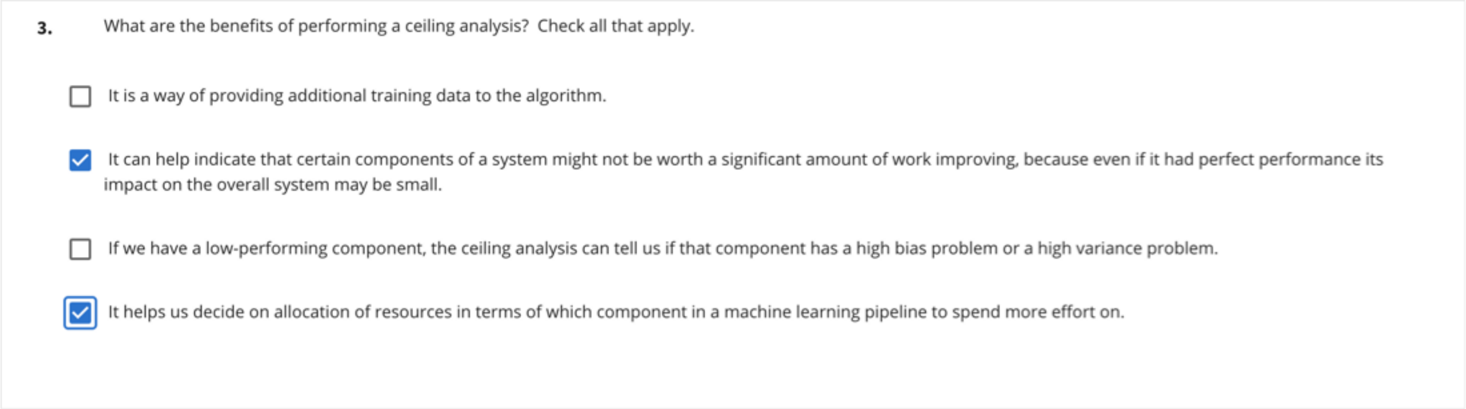
2. Suppose that you just joined a product team that has been developing a machine learning application, using $m = 1,000$ training examples. You discover that you have the option of hiring additional personnel to help collect and label data. You estimate that you would have to pay each of the labellers $10 per hour, and that each labeller can label 4 examples per minute. About how much will it cost to hire labellers to label 10,000 new training examples?

○ $600

● $400

○ $10,000

○ $250



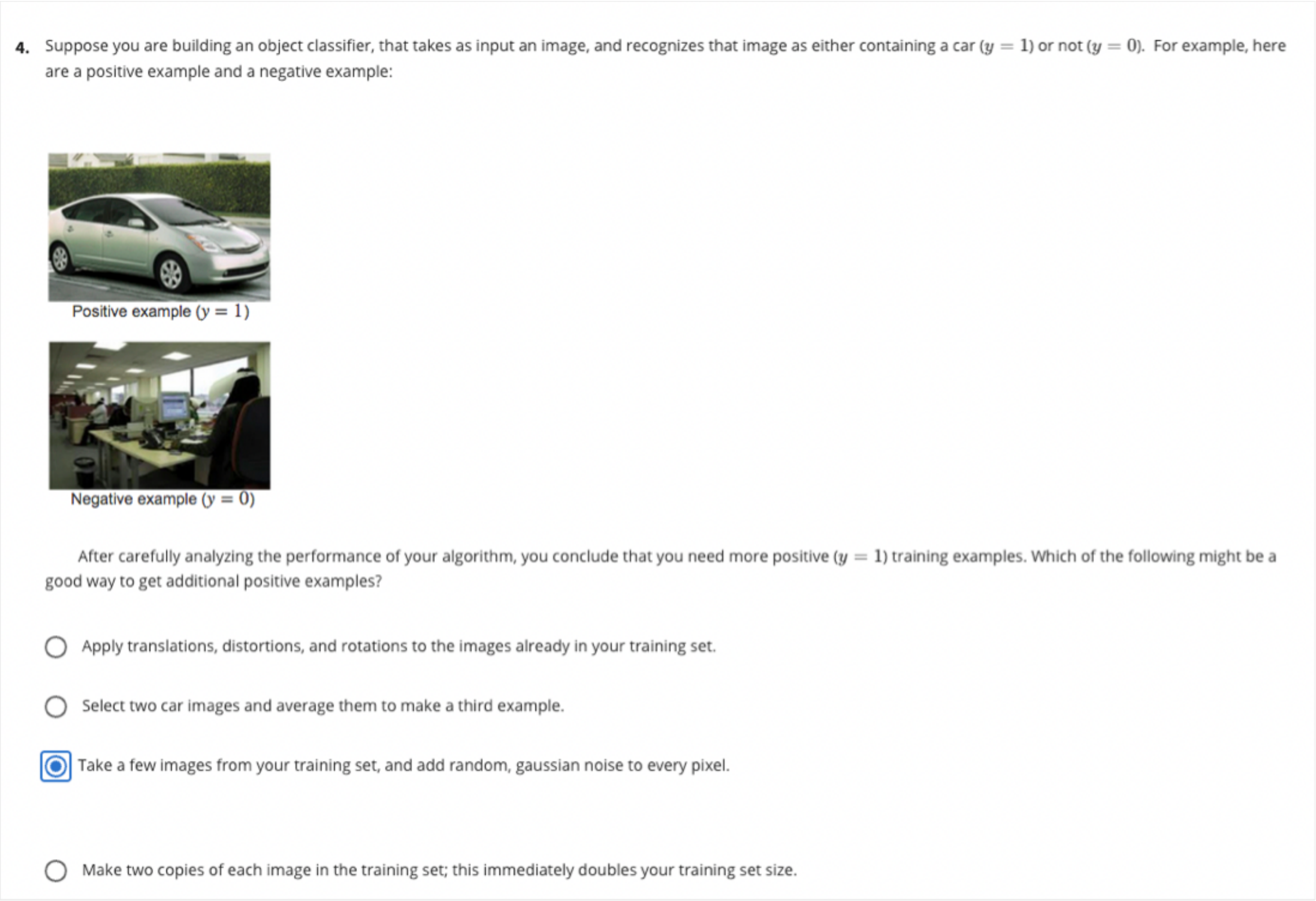
3. What are the benefits of performing a ceiling analysis? Check all that apply.

□ It is a way of providing additional training data to the algorithm.

☑ It can help indicate that certain components of a system might not be worth a significant amount of work improving, because even if it had perfect performance its impact on the overall system may be small.

□ If we have a low-performing component, the ceiling analysis can tell us if that component has a high bias problem or a high variance problem.

☑ It helps us decide on allocation of resources in terms of which component in a machine learning pipeline to spend more effort on.



4. Suppose you are building an object classifier, that takes as input an image, and recognizes that image as either containing a car ($y = 1$) or not ($y = 0$). For example, here are a positive example and a negative example:

Positive example ($y = 1$)

Negative example ($y = 0$)

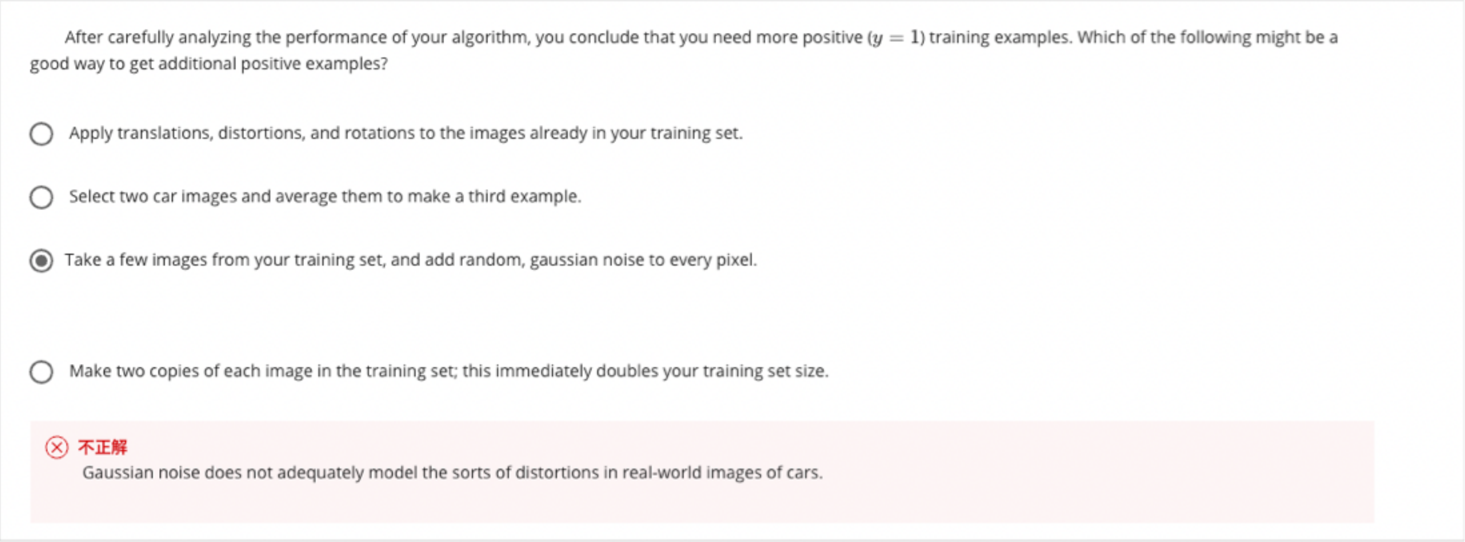
After carefully analyzing the performance of your algorithm, you conclude that you need more positive ($y = 1$) training examples. Which of the following might be a good way to get additional positive examples?

○ Apply translations, distortions, and rotations to the images already in your training set.

○ Select two car images and average them to make a third example.

● Take a few images from your training set, and add random, gaussian noise to every pixel.

○ Make two copies of each image in the training set; this immediately doubles your training set size.



After carefully analyzing the performance of your algorithm, you conclude that you need more positive ($y = 1$) training examples. Which of the following might be a good way to get additional positive examples?

○ Apply translations, distortions, and rotations to the images already in your training set.

○ Select two car images and average them to make a third example.

● Take a few images from your training set, and add random, gaussian noise to every pixel.

○ Make two copies of each image in the training set; this immediately doubles your training set size.

⊗ 不正解

Gaussian noise does not adequately model the sorts of distortions in real-world images of cars.

4. Suppose you are building an object classifier, that takes as input an image, and recognizes that image as either containing a car ($y = 1$) or not ($y = 0$). For example, here are a positive example and a negative example:

Positive example ($y = 1$)

Negative example ($y = 0$)

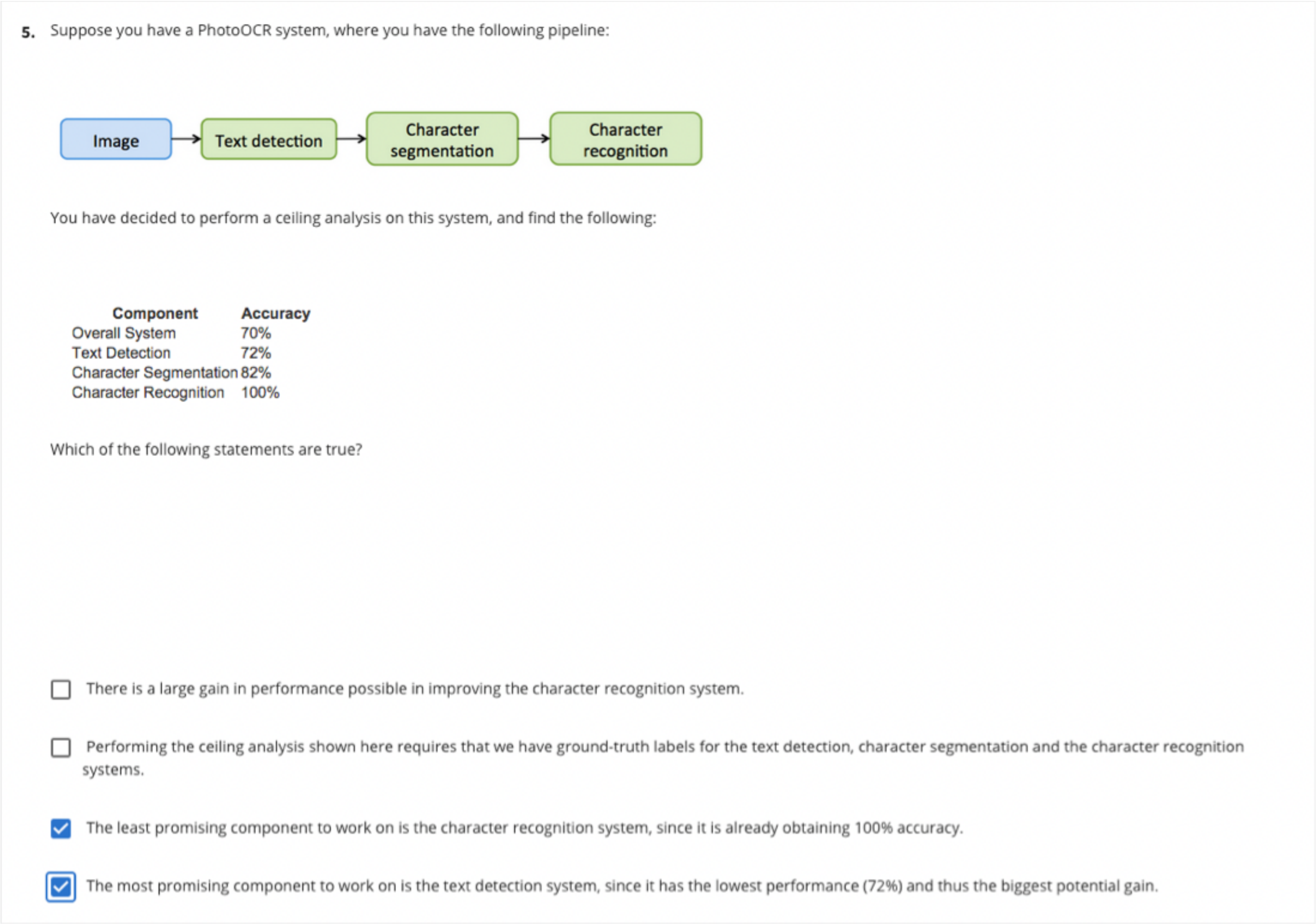
After carefully analyzing the performance of your algorithm, you conclude that you need more positive ($y = 1$) training examples. Which of the following might be a good way to get additional positive examples?

● Mirror your training images across the vertical axis (so that a left-facing car now becomes a right-facing one).

○ Take a few images from your training set, and add random, gaussian noise to every pixel.

○ Take a training example and set a random subset of its pixel to 0 to generate a new example.

○ Select two car images and average them to make a third example.



5. Suppose you have a PhotoOCR system, where you have the following pipeline:

You have decided to perform a ceiling analysis on this system, and find the following:

Component	Accuracy
Overall System	70%
Text Detection	72%
Character Segmentation	82%
Character Recognition	100%

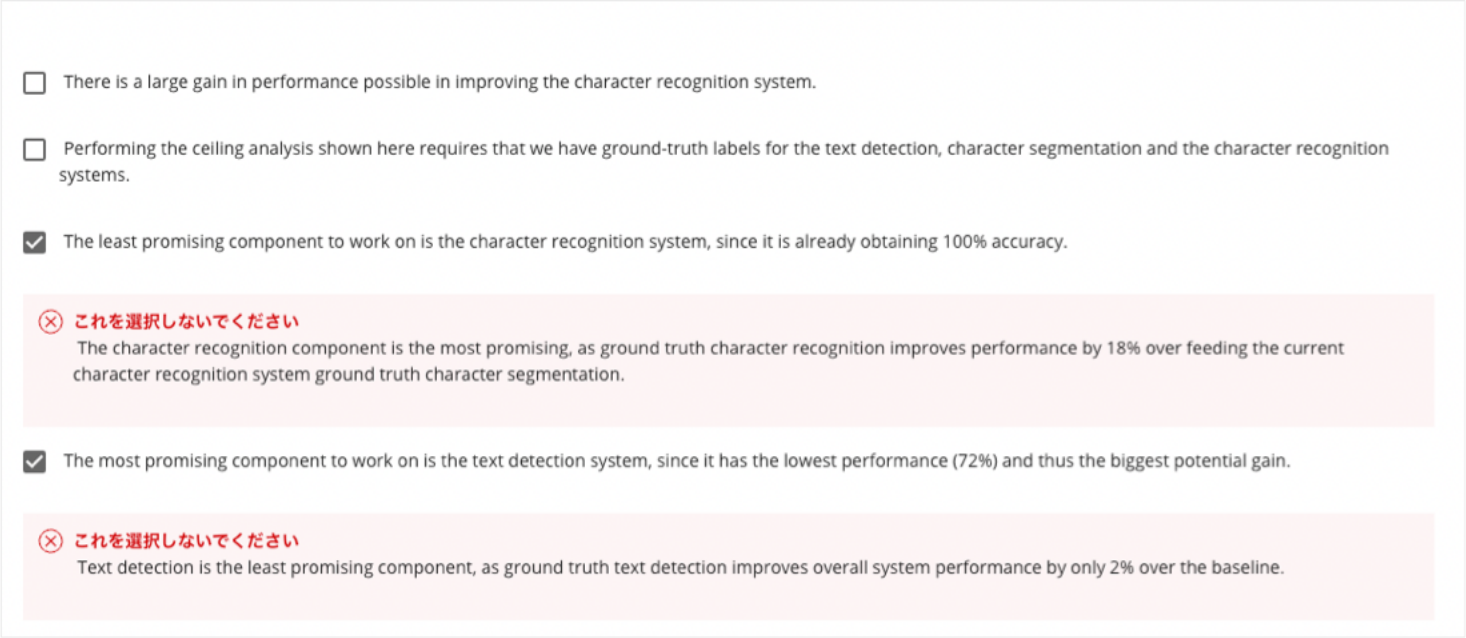
Which of the following statements are true?

☐ There is a large gain in performance possible in improving the character recognition system.

☐ Performing the ceiling analysis shown here requires that we have ground-truth labels for the text detection, character segmentation and the character recognition systems.

☑ The least promising component to work on is the character recognition system, since it is already obtaining 100% accuracy.

☑ The most promising component to work on is the text detection system, since it has the lowest performance (72%) and thus the biggest potential gain.



☐ There is a large gain in performance possible in improving the character recognition system.

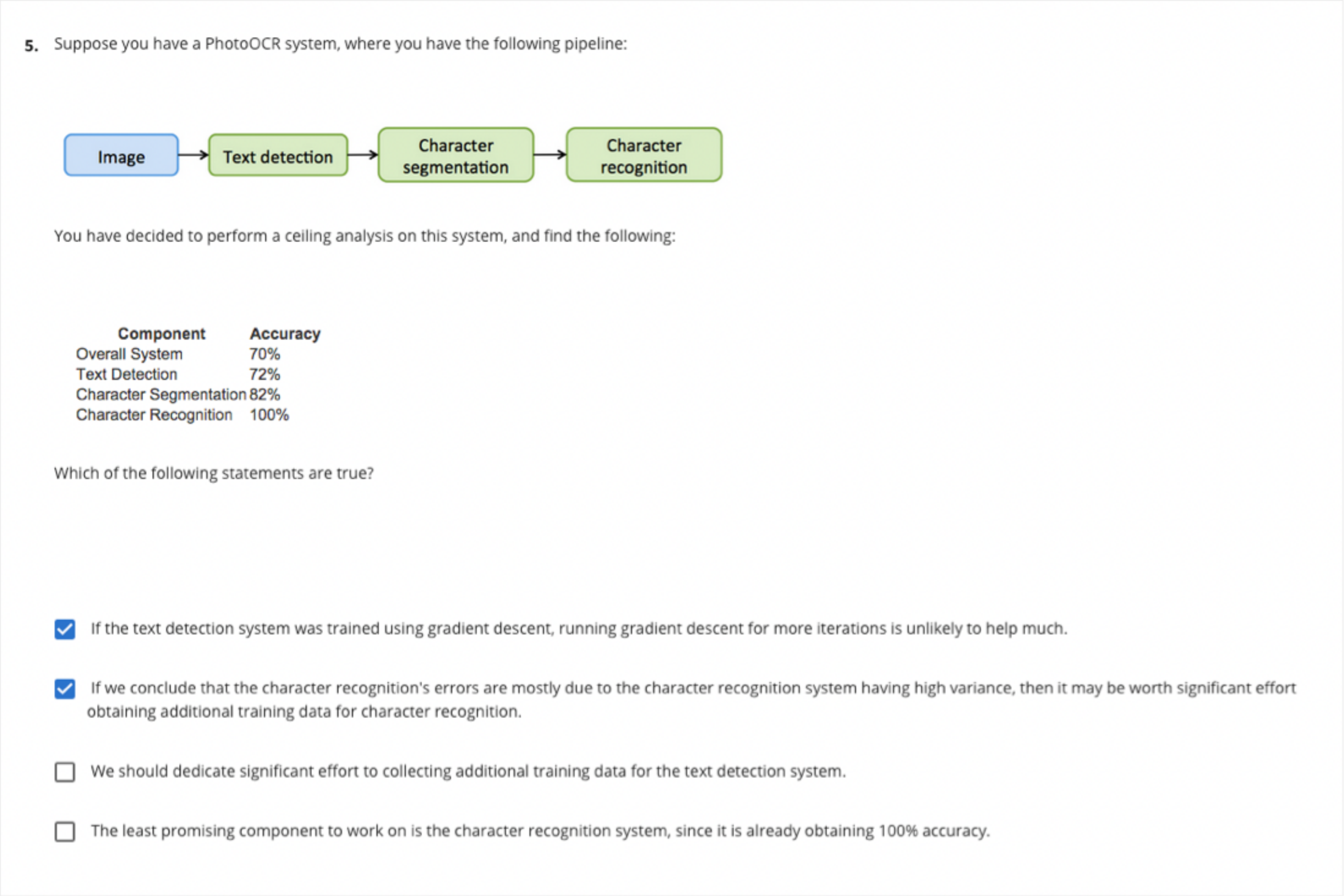
☐ Performing the ceiling analysis shown here requires that we have ground-truth labels for the text detection, character segmentation and the character recognition systems.

☑ The least promising component to work on is the character recognition system, since it is already obtaining 100% accuracy.

ⓧ これを選択しないでください
The character recognition component is the most promising, as ground truth character recognition improves performance by 18% over feeding the current character recognition system ground truth character segmentation.

☑ The most promising component to work on is the text detection system, since it has the lowest performance (72%) and thus the biggest potential gain.

ⓧ これを選択しないでください
Text detection is the least promising component, as ground truth text detection improves overall system performance by only 2% over the baseline.



5. Suppose you have a PhotoOCR system, where you have the following pipeline:

You have decided to perform a ceiling analysis on this system, and find the following:

Component	Accuracy
Overall System	70%
Text Detection	72%
Character Segmentation	82%
Character Recognition	100%

Which of the following statements are true?

☑ If the text detection system was trained using gradient descent, running gradient descent for more iterations is unlikely to help much.

☑ If we conclude that the character recognition's errors are mostly due to the character recognition system having high variance, then it may be worth significant effort obtaining additional training data for character recognition.

□ We should dedicate significant effort to collecting additional training data for the text detection system.

□ The least promising component to work on is the character recognition system, since it is already obtaining 100% accuracy.