application example photo ocr

problem description and pipeline

photo optical character recognition (OCR)

photo ocr pipeline

1. text detection
2. character segmentation
3. character classification → S n r
    1.

When someone refers to a "machine learning pipeline," he or she is referring to:

- A PhotoOCR system.
- A character recognition system.
- A system with many stages / components, several of which may use machine learning.

Correct Response

- An application in plumbing. (Haha.)

supervised learning for pedestrian detection

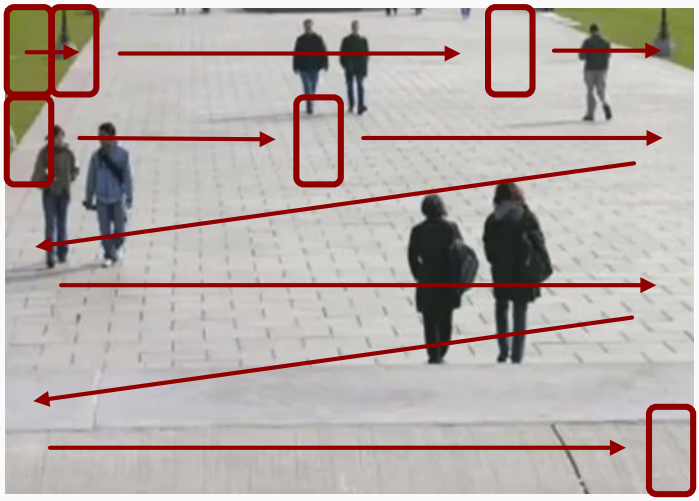
x = pixels in an 82x36 image patch

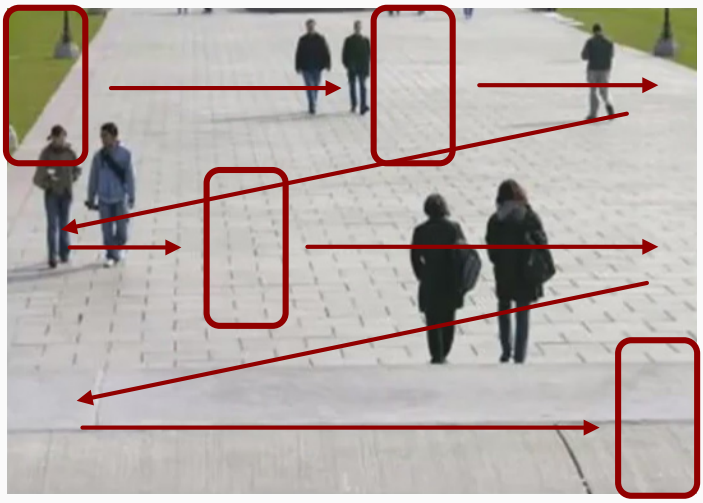
positive examples ($y = 1$)

negative examples ($y = 0$)

sliding window detection

step-size/stride is determined by pixel length and scaled upward after each iteration

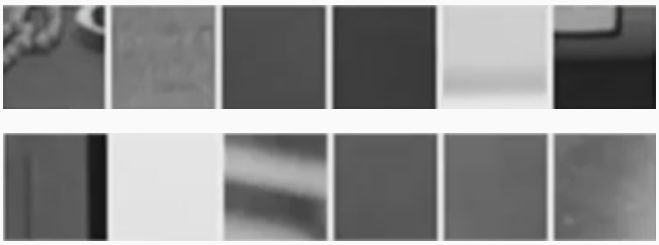




the process of training a sliding window classifier in order to detect appropriately positive examples ($y = 1$) in the problem of detecting pedestrians within the presence of a given images

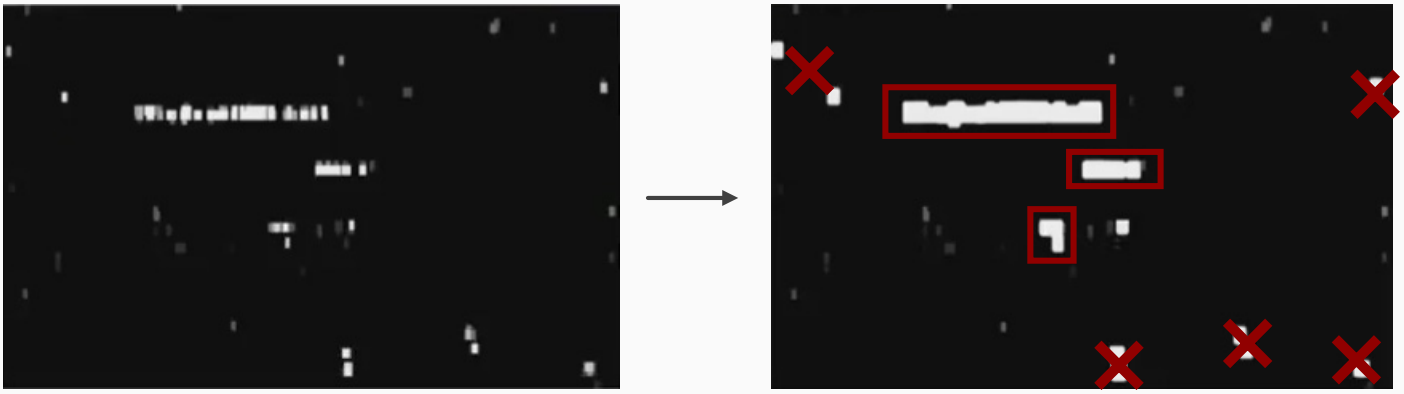
text detection

positive examples ($y = 1$)



negative examples ($y = 0$)

in a text detection algorithm, the process scans through an image using sliding windows and maps out the algorithm's determination of text presence based on the black and white coding (center image). the algorithm subsequently performs mathematical "expansion" in order to amplify the areas identified as containing text



a heuristic will be applied in a manner suitable for expectation of what constitutes text (e.g. text boxes should be much wider than they are tall, implying that smaller rectangular text identification areas are deemed to not represent the presence of text)

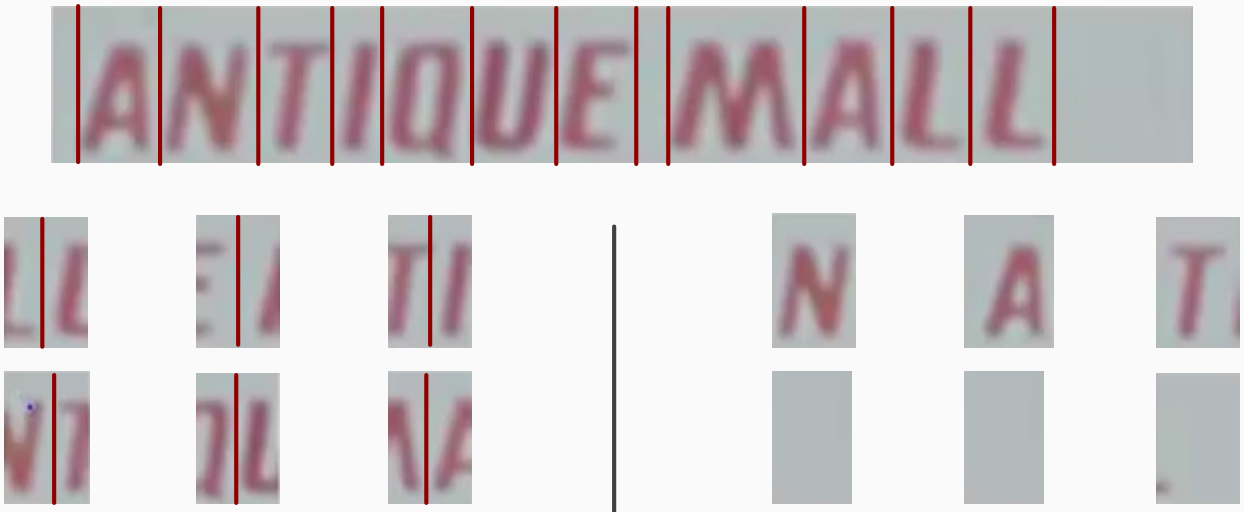
Suppose you are running a text detector using 20x20 image patches. You run the classifier on a 200x200 image and when using sliding window, you "step" the detector by 4 pixels each time. (For this problem assume you apply the algorithm at only one scale.) About how many times will you end up running your classifier on a single image? (Pick the closest answer.)

- About 100 times.
- About 400 times.
- About 2,500 times.

Correct Response

- About 40,000 times.

1 dimension window for character segmentation



positive examples ($y = 1$) negative examples ($y = 0$)

the next step will determine whether or not there is a split between any of the detected training examples. positive examples will represent those segments that contain splits and the negative examples to the algorithm will label images being absent of a split

getting lots of data and artificial data

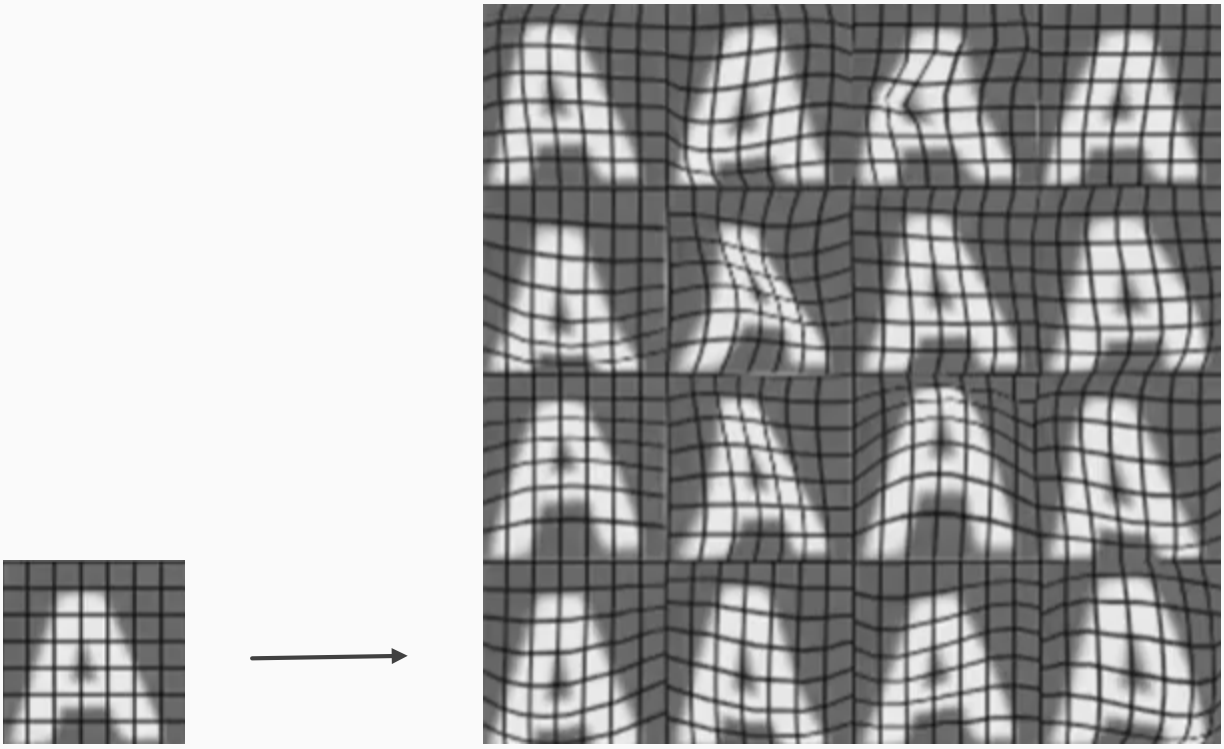
artificial data sysnthesis via photo ocr

real data

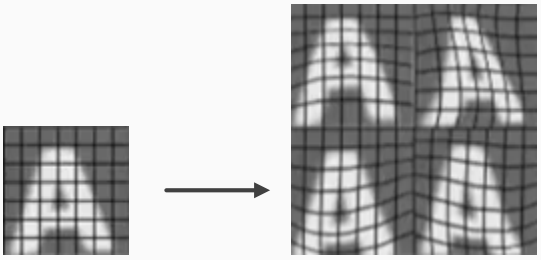
synthetic data

in the example of photo ocr, data can be taken in the form of many different fonts and scripts and synthesized against random backgrounds to create larger training sets

synthesizing data by introducing distortions

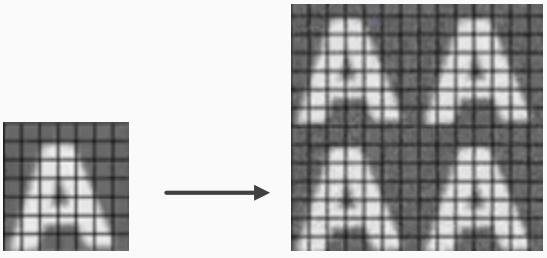


the same methods can be applied to image distortions and audio clips alike
distortion introduced should be a representation of the type of noise/distortion likely present to the test set



audio:
background noise,
poor cellphone

it is typically not valuable to add random/meaningless noise to the dataset



x_i =intensity (brightness)of pixel i

Suppose you are training a linear regression model with m examples by minimizing:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2)$$

Suppose you duplicate every example by making two identical copies of it. That is, where you previously had one example $(x^{(i)}, y^{(i)})$, you now have two copies of it, so you now have 2m examples. Is this likely to help?

- Yes, because increasing the training set size will reduce variance.
- Yes, so long as you are using a large number of features (a "low bias" learning algorithm).
- No. You may end up with different parameters θ, but they are unlikely to do any better than the ones learned from the original training set.
- No, and in fact you will end up with the same parameters θ as before you duplicated the data.

Correct Response

important aspects of obtaining more data

ensure a low bias classifier is used before expending effort (plot learning curves to verify). E.g. increase the number of features/number of hidden units in a neural network until a low bias classifier is obtained

determine and value the difficulty of obtaining as much as 10x current data

- artificial data synthesis
- collect and label data manually
- "crowd source" (e.g. amazon mechanical turk)

You've just joined a product group that has been developing a machine learning application for the last 12 months using 1,000 training examples. Suppose that by manually collecting and labeling examples, it takes you an average of 10 seconds to obtain one extra training example. Suppose you work 8 hours a day. How many days will it take you to get 10,000 examples? (Pick the closest answer.)

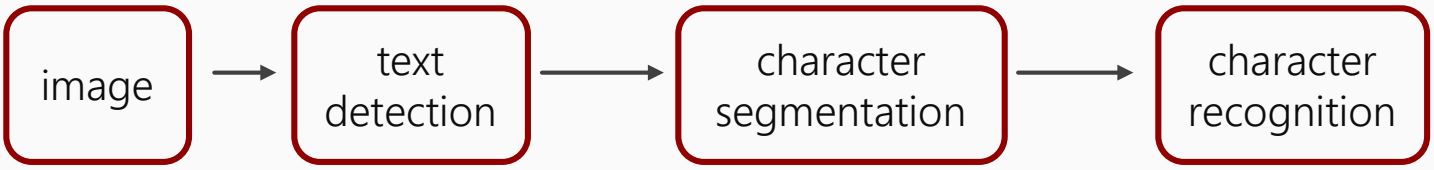
- About 1 day.
- About 3.5 days.

Correct Response

- About 28 days.
- About 200 days.

ceiling analysis: part of the pipeline to work on next

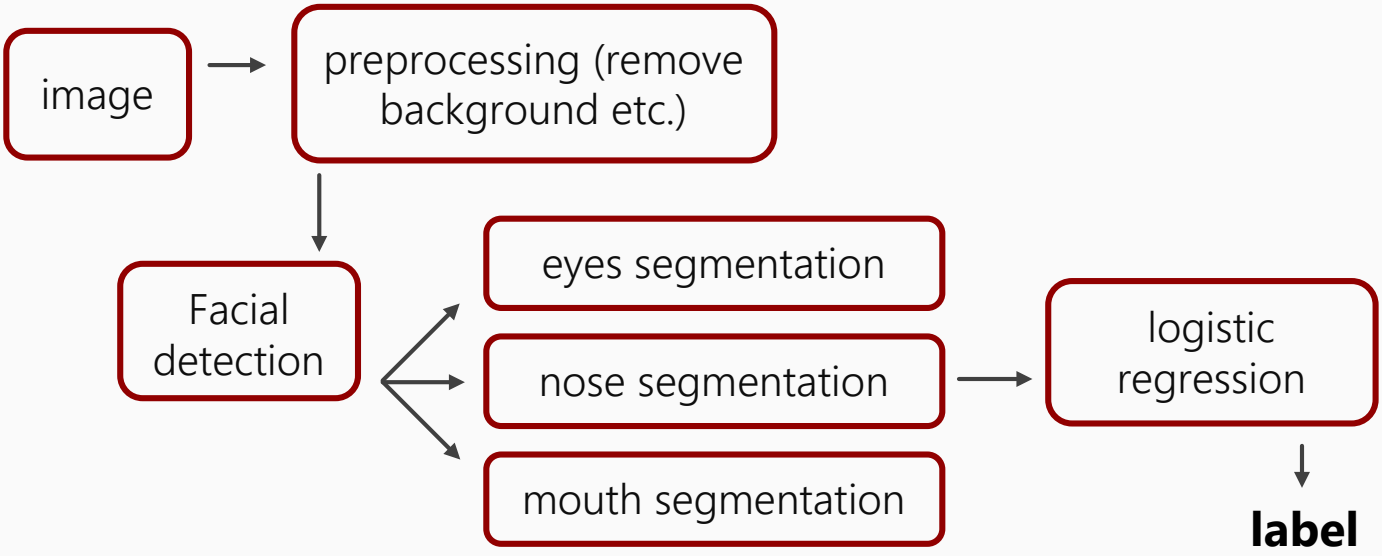
estimating errors due to each component (ceiling analysis)



determining part of the pipeline to spent the most time improving (resource allocation)

component	accuracy	improvement
overall system	72%	-
text detection	89%	17% ↑
character segmentation	90%	1% ↑
character recognition	100%	10% ↑

ceiling analysis: facial recognition example pipeline



component	accuracy	improvement
overall system	85%	-
preprocessing	85.1%	0.1% ↑
face detection	91%	5.9% ↑
eyes segmentation	95%	4% ↑
nose segmentation	96%	1% ↑
mouth segmentation	97%	1% ↑
logistic regression	100	3% ↑

it is important to remain conscious of time-spent to performance gains