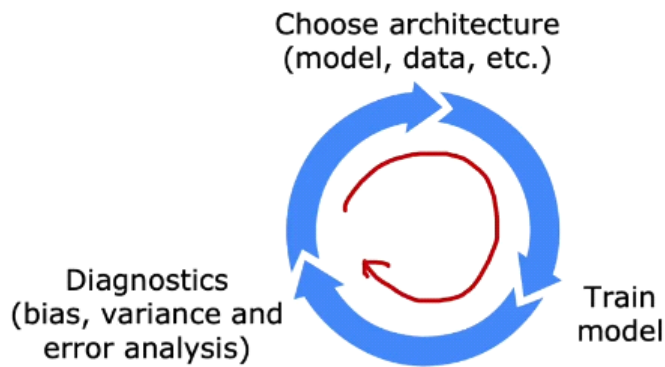


Process of ML

2023年5月26日 16:39

1. 机器学习的迭代循环

Iterative loop of ML development



屏幕剪辑的捕获时间: 2023/5/26 16:41

2. 误差分析

Error analysis

$m_{cv} =$ ~~500~~ examples in cross validation set.

5000

Algorithm misclassifies ~~100~~ of them.

1000

Manually examine 100 examples and categorize them based on common traits.

- Pharma: 21 → more data features
- Deliberate misspellings (w4tches, med1cine): 3
- Unusual email routing: 7
- Steal passwords (phishing): 18 → more data features
- Spam message in embedded image: 5

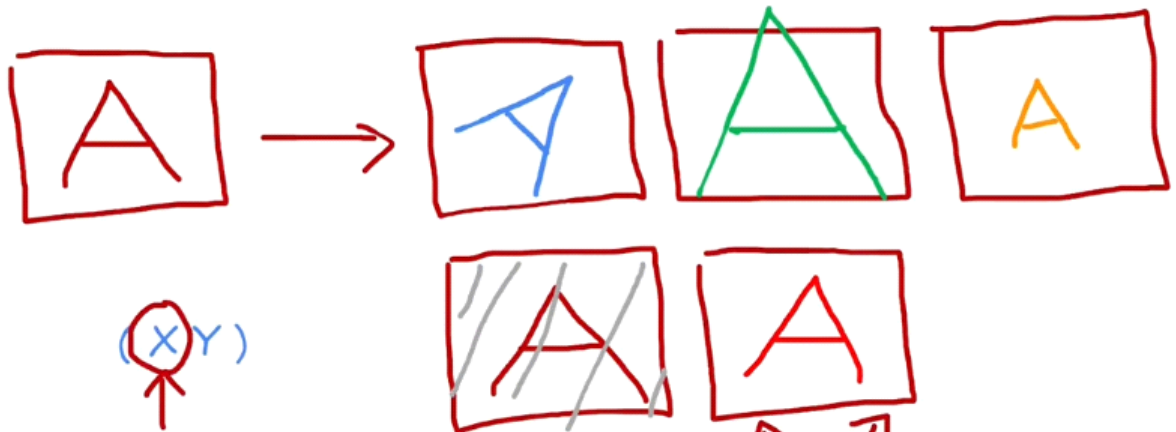
手动寻找误分类的案例中，被误分类的都是哪些type

案例过多，可以取一个小的子集观测

3. 如何高效添加数据

Data augmentation

Augmentation: modifying an existing training example to create a new training example.

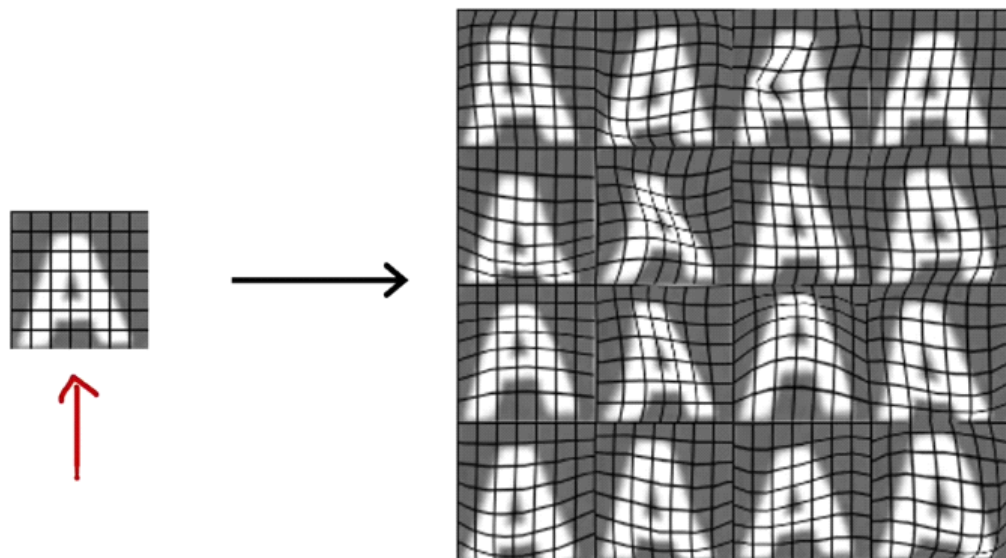


, 以便提出另一个具有相同标签的示例。

in order to come up with another example that has the same label. w Ng

数据增强：将实例中的数据做变换，告诉算法这样变换后仍然是原数据，还可以如下进行扭曲等操作，让算法鲁棒性更好





Data augmentation by introducing distortions



在声音数据采集时，也可以进行类似的变换：（添加背景噪音）

Data augmentation for speech

Speech recognition example

-  Original audio (voice search: "What is today's weather?")
-  + Noisy background: Crowd
-  + Noisy background: Car
-  + Audio on bad cellphone connection

这实际上是一种非常关键的人工增加技术
this was actually a really critical technique for increasing artificially w Ng

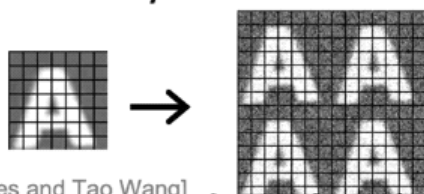
但是也不能胡乱加噪音，没有意义：

Data augmentation by introducing distortions

Distortion introduced should be representation of the type of noise/distortions in the test set.



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



x_i = intensity (brightness) of pixel i
 $x_i \leftarrow x_i + \text{random noise}$

Adam Coates and Tao Wang

但是，如果这不能代表您在测试集中看到的内容

But if to the extent that this isn't that representative of what you see in Ng

屏幕剪辑的捕获时间: 2023/5/26 17:11

合成数据：比如图像OCR算法，识别图中的字母，可以自行建立许多字母的数据，和实际情况十分相近，从而轻松得到大量的数据，主要用于计算机视觉领域

Artificial data synthesis for photo OCR



Real data



Synthetic data

以数据为中心的机器学习:

Engineering the data used by your system

Conventional
model-centric
approach:

$AI = \text{Code} + \text{Data}$
(algorithm/model)

Work on this

Data-centric
approach:

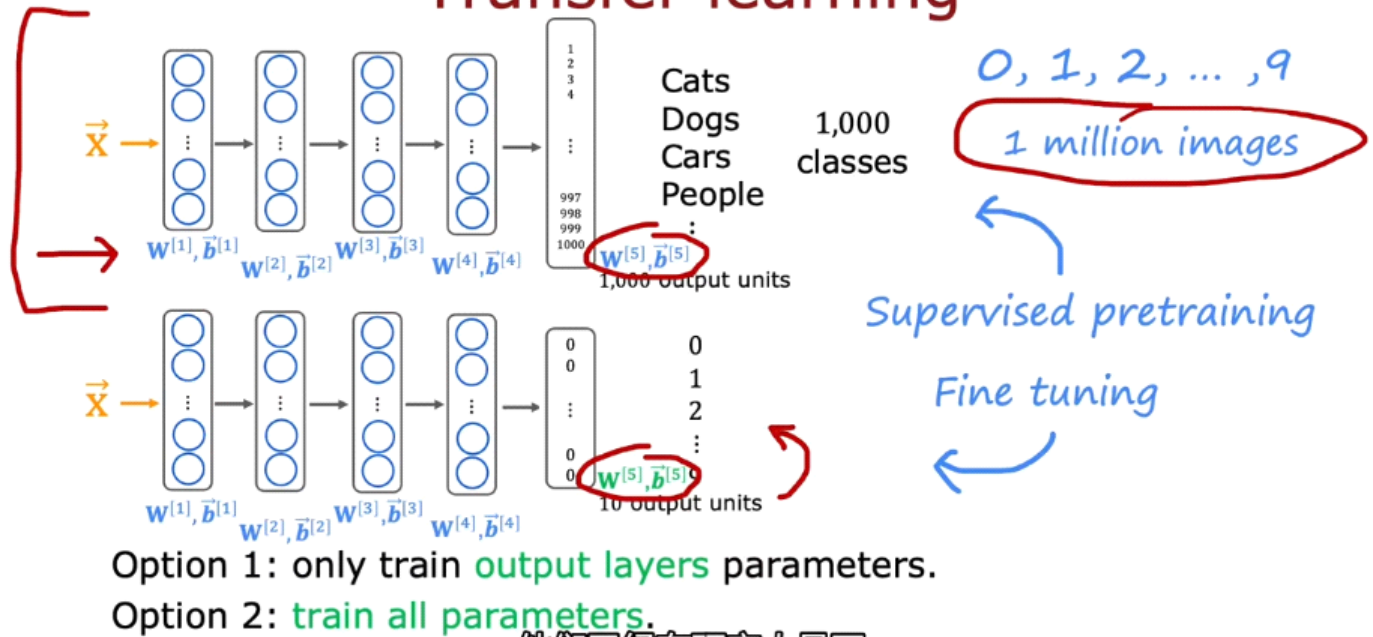
$AI = \text{Code} + \text{Data}$
(algorithm/model)

Work on this

4. 迁移学习

原理：通过其他学习过程建立的神经网络，希望这个网络已经具有一些初步处理图像数据的能力，把这些layer的参数直接传递给新的网络，可以从一个更好的初参数开始学习

Transfer learning



他们已经有研究人员了

DeepLearning.AI

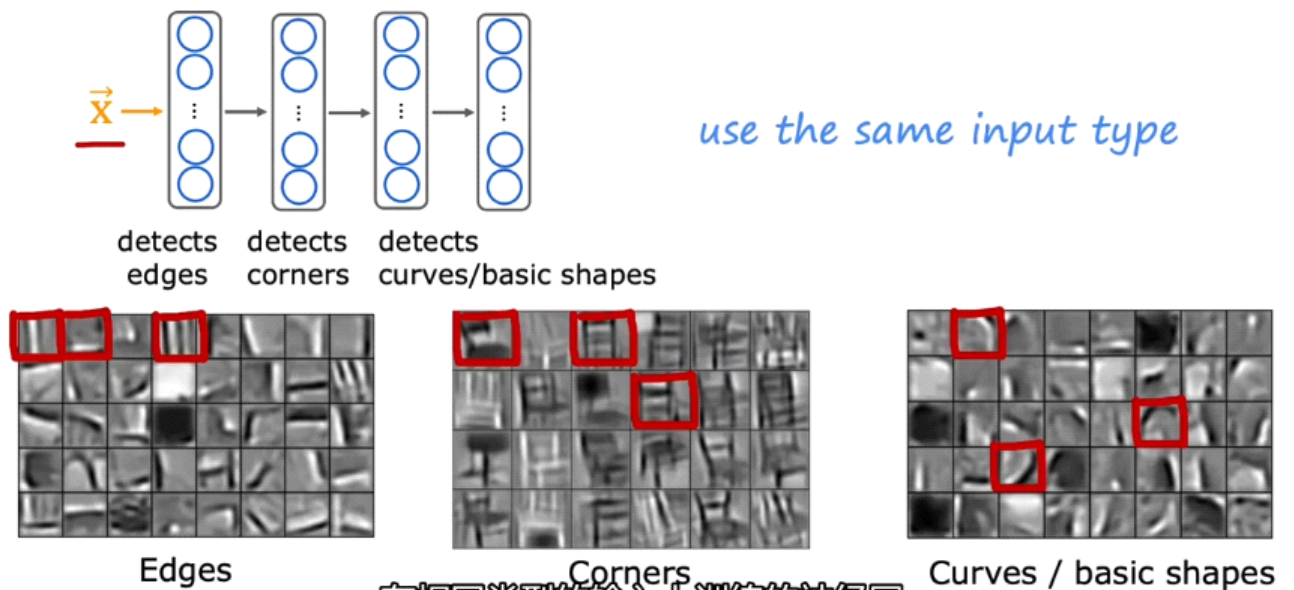
there will already be researchers they have

Andrew Ng

好处就是可以利用其他人已经进行预训练的网络，自行调整输出层的参数，加快训练进程

为何迁移学习有用？

Why does transfer learning work?



在相同类型的输入上训练的神经网络

DeepLearning.AI

a neural network trained on the same type of input,

Andrew Ng

屏幕剪辑的捕获时间: 2023/5/27 9:22

预训练的网络输入值需要与自己要建立的模型的输入一致。例如要识别语音，那么以图像为输入的网络就不能用于迁移学习。

迁移学习的过程总结：

Transfer learning summary

→ 1. Download neural network parameters pretrained on a large dataset with same input type (e.g., images, audio, text) as your application (or train your own). 1M

→ 2. Further train (fine tune) the network on your own data.

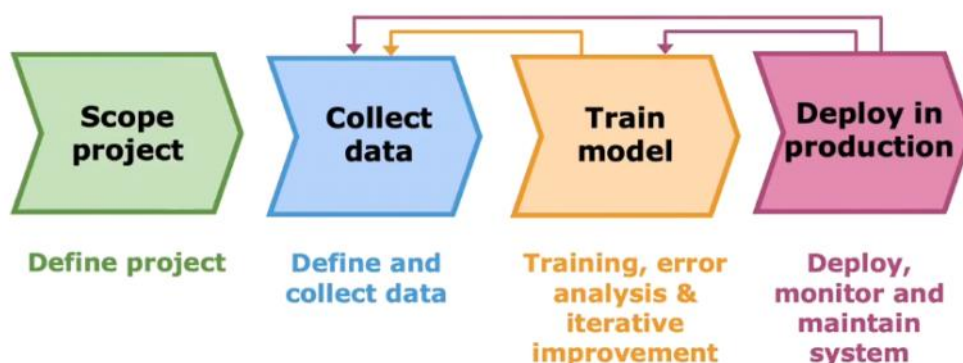
1000
50

迁移学习最大的优点是机器学习社区中的人们可以共享代码与参数，这比一个人自己工作的效率要高太多。

5.机器学习的全周期

以语音识别为例：

Full cycle of a machine learning project



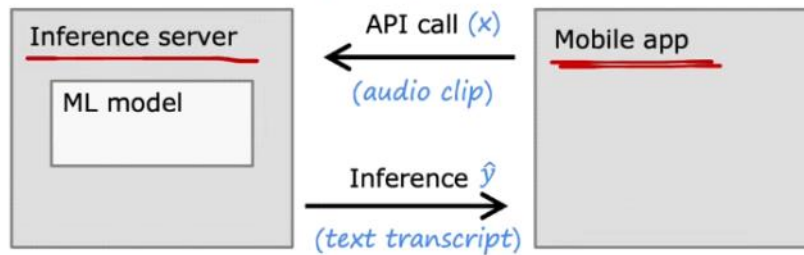
事实上，如果用户和您

DeepLearning.AI

Stanford In fact, if users and if you have

Andrew Ng

Deployment



- Software engineering may be needed for:
- Ensure reliable and efficient predictions
 - Scaling
 - Logging
 - System monitoring
 - Model updates

模型更新以用新模型替换旧模型。

DeepLearning.AI model update to replace the old model with a new one. Andrew Ng

概念: MLOps (Machine Learning Operations)

研究如何系统地构建、部署与维护机器学习系统, 使模型可靠、可扩展、合法等

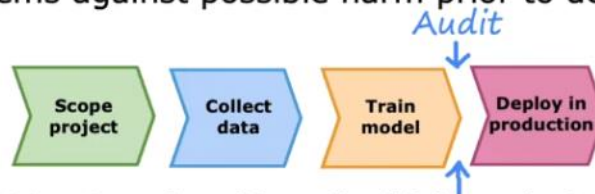
6. 公平、偏见与伦理道德

Guidelines

Get a diverse team to brainstorm things that might go wrong, with emphasis on possible harm to vulnerable groups.

Carry out literature search on standards/guidelines for your industry.

Audit systems against possible harm prior to deployment.



Develop mitigation plan (if applicable), and after deployment, monitor for possible harm.

驶汽车团队都制定了缓解计划

DeepLearning.AI cars on the road had developed mitigation plans for Andrew Ng

7. 倾斜数据集 (如罕见疾病)

Rare disease classification example

Train classifier $f_{\vec{w},b}(\vec{x})$

($y = 1$ if disease present,
 $y = 0$ otherwise)

Find that you've got **1%** error on test set
(99% correct diagnoses)

Only 0.5% of patients have the disease

`print("y=0")`

99.5% accuracy, 0.5% error

仅仅凭借误差大小根本不能评价算法的好坏，从而引入精准度与召回率的概念：

Precision/recall

$y = 1$ in presence of rare class we want to detect.

		Actual Class	
		1	0
Predicted Class	1	True positive 15	False positive 5
	0	False negative 10	True negative 70
		↓ 25	↓ 75

Precision:

(of all patients where we predicted $y = 1$, what fraction actually have the rare disease?)

$$\frac{\text{True positives}}{\text{\#predicted positive}} = \frac{\text{True positives}}{\text{True pos} + \text{False pos}} = \frac{15}{15+5} = 0.75$$

Recall:

(of all patients that actually have the rare disease, what fraction did we correctly detect as having it?)

$$\frac{\text{True positives}}{\text{\#actual positive}} = \frac{\text{True positives}}{\text{True pos} + \text{False neg}} = \frac{15}{15+10} = 0.6$$

这样的概念可以防止本节开头图片中，算法一直输出 “ $y=0$ ” 的情况

既要保证预测为患病时，病人确实患病的概率足够高

也要保证病人确实患病时，算法预测其患病的概率足够高



8. 权衡精准度与召回率

逻辑回归中，决策阈值定为0.5，但如果预测错误造成的损失很大，就可以考虑提高预测阈值，例如提高到0.7左右

但如果提高阈值，会导致精准度提高，同时导致召回率变低

①. 在想十分确信地预测 $y=1$ 时，可以调高阈值提高精准度（比如治疗十分昂贵、痛苦等）

②. 在想要避免错失病人，导致延误了病人治疗，就可以降低阈值，提高召回率

Trading off precision and recall

Logistic regression: $0 < f_{\vec{w},b}(\vec{x}) < 1$

→ Predict 1 if $f_{\vec{w},b}(\vec{x}) \geq 0.5$
 → Predict 0 if $f_{\vec{w},b}(\vec{x}) < 0.5$

$$\text{precision} = \frac{\text{true positives}}{\text{total predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{total actual positive}}$$

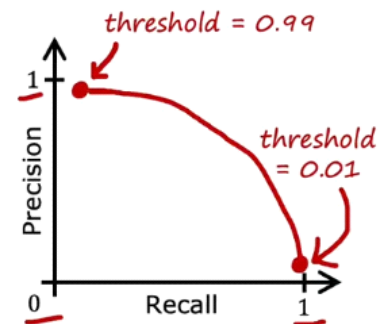
Suppose we want to predict $y = 1$ (rare disease) only if very confident.

→ higher precision, lower recall.

Suppose we want to avoid missing too many cases of rare disease (when in doubt predict $y = 1$)

→ lower precision, higher recall.

More generally predict 1 if: $f_{\vec{w},b}(\vec{x}) \geq \text{threshold}$.



非常低的精度但相对较高的召回率。

DeepLearning.AI

very low precision but relatively high recall.

Andrew Ng

※F1 score

自动权衡精准度与召回率的指标

F1 score

How to compare precision/recall numbers?

	Precision (P)	Recall (R)	Average	F ₁ score
Algorithm 1	0.5	0.4	0.45	0.444
Algorithm 2	0.7	0.1	0.4	0.175
Algorithm 3	0.02	1.0	0.501	0.0392

print("y=1")

~~Average = $\frac{P+R}{2}$~~

$$F1 \text{ score} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{R} \right) = 2 \frac{PR}{P+R}$$

P和R的调和

DeepLearning.AI

Stanford the harmonic mean of P and R,

Andrew Ng