Getting Lots of Data and Artificial Data

I've seen over and over that one of the most reliable ways to get a **high performance machine learning system** is to take a **low bias learning algorithm** and to train it on a **massive training set.** But where did we get so much training data from?

Turns out that in the machine learning there's a fascinating idea called **artificial data synthesis**, this doesn't apply to every single problem, and to apply to a specific problem, often takes some **thought, innovation and insight**. But if this idea applies to our machine learning problem, it can sometimes be a an **easy way to get a huge training set** to give to our learning algorithm.

The idea of artificial data synthesis comprises of two variations:

1. **Creating data from scratch**
2. **If we already have a small labeled training set can we amplify it into a larger training set**

Artificial data synthesis for photo OCR

If we go and collect a large labeled data set of images will look like this:

- The goal is to **take an image patch and have the system recognize the character**
- Treat the images as gray-scale (makes it a bit easer)
  - It turns out that using color doesn't seem to help that much for this particular problem

Real data

So all of this examples of row images, how can we come up with a much larger training set?

- Modern computers often have a big font library
  - Depending on what word processor we use, we might have all of these fonts and many more already stored inside.

- If we go to websites, huge free font libraries
- For more training data, take characters from different fonts, paste these characters again random backgrounds

After some work, can build a synthetic training set:

Synthetic data

Every image is actually a synthesized image. Where we take a font and we paste an image of one character or a few characters from that font against other random background image.

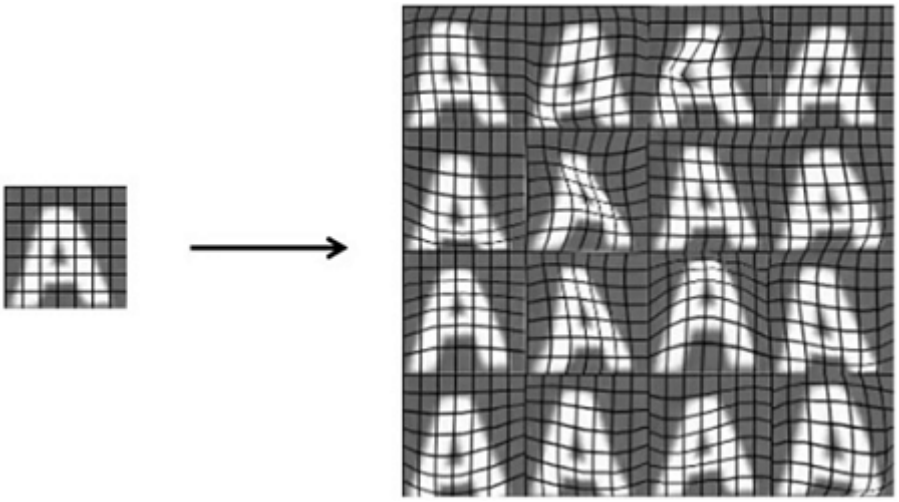
- Maybe some blurring/distortion filters
- Takes thought and work to make it **look realistic**

And so by using synthetic data we have essentially an unlimited supply of training examples for artificial training synthesis, and so, if we use this source synthetic data, we have essentially unlimited supply of label data to create a supervised learning algorithm

- **This is an example of creating new data from scratch**

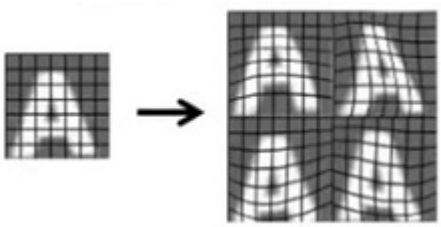
Synthetizing data by introducing distortions

Other way to get even more training data is to introduce distortion into existing data. What we can do is then take an image and introduce **artificial warpings** or **artificial distortions** into the image so they can take the image and turn that into 16 new examples.



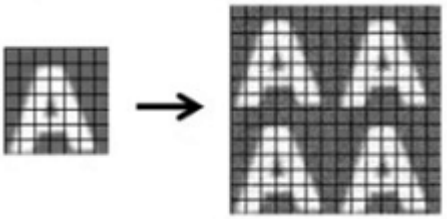
- **So we get a 16 new examples**
- Allows us amplify existing training set
- **This, again, takes though and insight in terms of deciding how to amplify**

Warning about synthesizing data by introducing distortions: Distortion introduced should be representation of the type of noise/distortions in the test set.



Audio:
Background noise,
bad cellphone connection

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



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

Video Question: 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.

Discussion on getting more data

- 1. **Make sure we have a low bias classifier before expending the effort. (Plot learning curves)**
  - If not a low bias classifier increase number of features
    - E.g. keep increasing the number of features/number of hidden units in neural network until we have a low bias classifier.
- 2. **"How much work would it be to get 10x as much data as we currently have?**
  - Often the answer is, "Not that hard"
  - **This is often a huge way to improve an algorithm**
  - Good question to ask to yourself or ask the team
  - **How do we collect more data?**
    - Artificial data synthesis
    - Collect/label it yourself
    - "Crowd source" (E.g. Amazon Mechanical Turk)

How many minutes/hours does it take to get a certain number of examples?

- Say we have m = 1,000 examples
- 10 seconds to label an example (10 secs/example)

Video Question: 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.

- About 28 days.
- About 200 days.