Importance of dataset for learning algorithms

Grégory Châtel

Disaitek
Intel AI Software Innovator

@rodgzilla
github.com/rodgzilla

September 25th, 2018

1. Machine learning basics

2. Popular ML tasks and their dataset

3. Specific problem of dataset for NLP tasks

4. Data efficiency
   - Transfer learning
   - Multi-task learning
   - Semi supervision

Machine learning

Supervised learning

Machine learning is a subfield of artificial intelligence.

Intuitively We want to *learn from* and *make predictions on* data.

Technically We want to build a model that approximate well (*e.g.* minimize a loss function) an unknown function for which we only have limited observations.

To do this, we usually need a lot of *data*.

Popular datasets for computer vision

1990, Statlog	∼2k outdoor images
1998, MNIST	60k BW images of handwritten digits
2005, LabelMe	∼187k scenes images
2009, ImageNet	∼14M color images
2017, JFT-300M	∼300M color images (internal dataset @ Google)

Popular datasets for Natural language processing

1997, Car evaluation dataset	∼2k car evaluations
2005, Stanford Sentiment Treebank	∼11k movie reviews
2011, IMDB Reviews	∼50k movie reviews
2012, Youtube Comedy Slam	∼1.1M pairs of video metadata
2015, Amazon reviews	∼82m product reviews

NLP tasks specificity

NLP tasks bias

There are problems with using services such AMT (Amazon Mechanical Turk) to annotate NLP dataset.

Data efficiency

Knowing that data is so important and hard to produce, it is important to squeeze every last bit of value out of it.

To do this, three ideas are explored:

- Transfer learning
- Multi-task learning
- Semi supervision

Transfer learning

> *The application of skills, knowledge, and/or attitudes that were learned in one situation to another learning situation. (Perkins, 1992)*

Transfer learning consists in taking an artificial neural network that has been trained on a *generic* task and *transferring* its knowledge (retraining it) to perform a new task.

The idea behind the method is that the information learned on a generic task will probably be useful for a new task of the same domain.

Pre-training in computer vision

Finetuning an ImageNet model (1000 categories classification) to easily perform cats vs dogs classification.

Pre-training in NLP

Language modeling as a generic task

Multi-task learning

> *Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better. (Rich Caruana, 1997)*

Instead of just training the network to perform the desired task, we also optimize it to perform *auxiliary tasks*.

Multi-task learning in computer vision

Find ref for the research paper that predicts gender from eye picture as auxiliary task for diabetic retinopathy detection.

Semi supervision