

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
 - Model pre-training
 - 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 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)

