Importance of dataset for learning algorithms

Grégory Châtel

Disaitek
Intel AI Software Innovator

@rodgzilla
github.com/rodgzilla

September 25th, 2018

1. Popular ML tasks and their dataset

2. Data efficiency
   - Transfer learning
   - Multi-task learning
   - Semi-supervised learning

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

Creating dataset

Creating new high quality datasets is both hard and expensive.

Some researchers experiment with training models using low quality data (weakly supervised learning)

TODO: AMT

Data efficiency

Knowing that datasets are so important and hard to create, it is important to squeeze every last bit of value out of them.

To do this, three ideas are explored:

- Transfer learning
- Multi-task learning
- Semi-supervised learning

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

Transfer learning is actually the base of the Google Cloud AutoML service.

Transfer learning in computer vision

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

Transfer learning 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*.

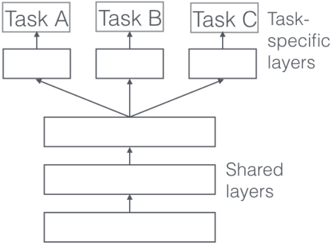


Image from http://ruder.io/multi-task/

Multi-task as a regularization technique

Informally, the goal of the multi-task learning is to force the model to use its computing power to perform something meaningful instead of using it to learn the noise of the data (overfitting).

Multi-task learning in computer vision

Some researchers noticed that by asking the model to predict the gender and age of patient in addition to detect *diabetic retinopathy* they got strong performance improvements.

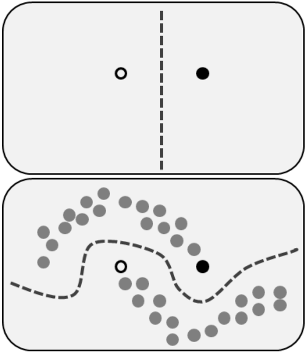
Find ref for the research paper that predicts gender from eye picture as auxiliary task for diabetic retinopathy detection.

Multi-task learning in NLP

Language modeling (ref OpenAI paper)

Translation for PoS tagging (syntactic tree?) (ref “one model to learn them all” for an extreme example).

Semi-supervised learning



The idea of semi-supervised learning is to use *unlabelled data* to improve our model.

Image from https://en.wikipedia.org/wiki/Semi-supervised_learning

Ideas of semi-supervised learning

The main concept of semi-supervised learning is to train a weaker student to imitate a stronger teacher.

Technically, we apply *mean-squared error* or a *Kullback-Liebler divergence* between the logits output by the student and the teacher. We typically alternate between supervised and semi-supervised steps of training.

Semi-supervised learning in practice

To do this, we can for example take the model we are training and train it to make the same prediction on a clean and noisy version of a same image. In this case, the prediction on the *noisy image* (resp. *clean image*) is the one of the student model (resp. teacher model)

To apply this algorithm, we do not need the real label of the image, we just assume that the model is already giving the right one and improve it by forcing it to acquire noise invariance.

TODO: Find the correct ref and a picture from the curious company blog post about mean teachers