



CS 329P : Practical Machine Learning (2021 Fall)

2.4 Feature Engineering

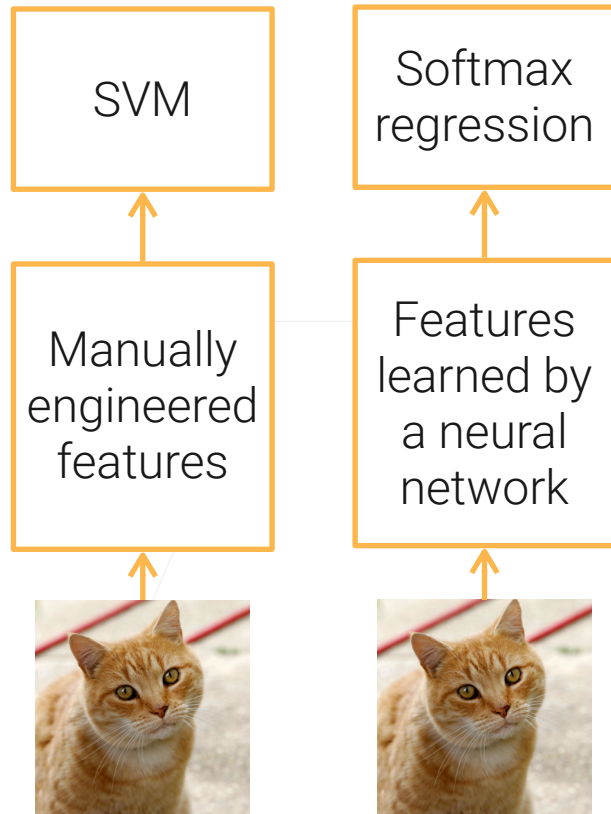
Qingqing Huang, Mu Li, Alex Smola

<https://c.d2l.ai/stanford-cs329p>

Feature Engineering



- Before deep learning (DL), feature engineering (FE) was critical to using ML models
 - Traditional CV: detect corners / interest points..
- DL train deep neural networks to automatically extract features
 - Train CNN to replace feature extractor
 - Features are more relevant to the final task
 - Limitation: data hungry, computation heavy



Tabular Data Features



- Tabular data are in the form of a table, feature columns of numeric / categorical / string type
- Int/float: directly use or or bin to n unique int values
- Categorical data: one-hot encoding
 - Map rare categories into “Unknown”
- Date-time: a feature list such as
 - [year, month, day, day_of_year, week_of_year, day_of_week]
- Feature combination: Cartesian product of two feature groups
 - [cat, dog] x [male, female] -> [(cat, male), (cat, female), (dog, male), (dog, female)]

e

	fish	cat	and	dog	unknown	
cat	→	[0,	1,	0,	0,	0]
dog	→	[0,	0,	0,	1,	0]

Text Features



- Represent text as token features
 - Bag of words (BoW) model
 - Limitations: needs careful vocabulary design, losing context for individual words
 - Word Embeddings (e.g. Word2vec):
 - Vectorizing words such that similar words are placed close together
 - Trained by predicting target word from context words
- Pre-trained universal language models (e.g. universal sentence encoder, BERT, GPT-3)
 - Giant transformer models
 - Trained with large amount of unannotated data
 - Usage: Text embedding; fine-tuning for downstream tasks

dog and cat and dinosaur

fish cat and dog unknown
[0, 1, 2, 1, 1]

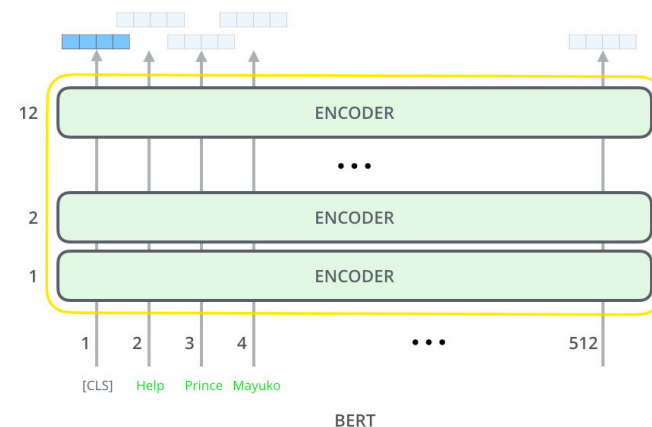
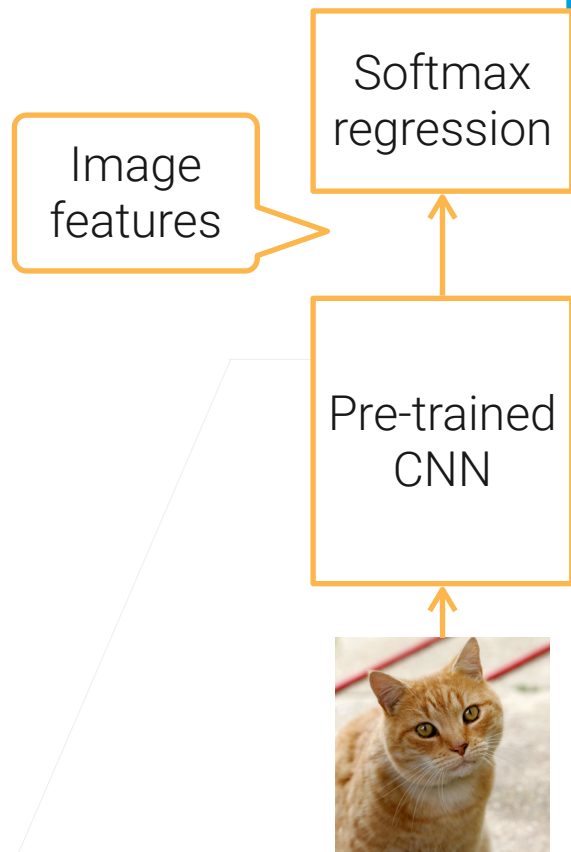


Image credit: Jay Alammar

Image/Video Features

- Traditionally extract images by hand-craft features such as SIFT
- Now pre-trained deep neural networks are common used as feature extractor
 - ResNet: trained with ImageNet (image classification)
 - I3D: trained with Kinetics (action classification)
 - Many off-the-shelf models available



Summary



- Features are representations of raw data that are relevant to the target task
- Feature engineering VS Feature learning
 - The latter is preferred if available (images/videos/audio/text)
 - Will cover more later in “transfer learning”