Information Engineering Option (paper 8)
Photo Editing and Image Search

Part C - Image Searching and Modelling Using Machine Learning Methods

Zoubin Ghahramani

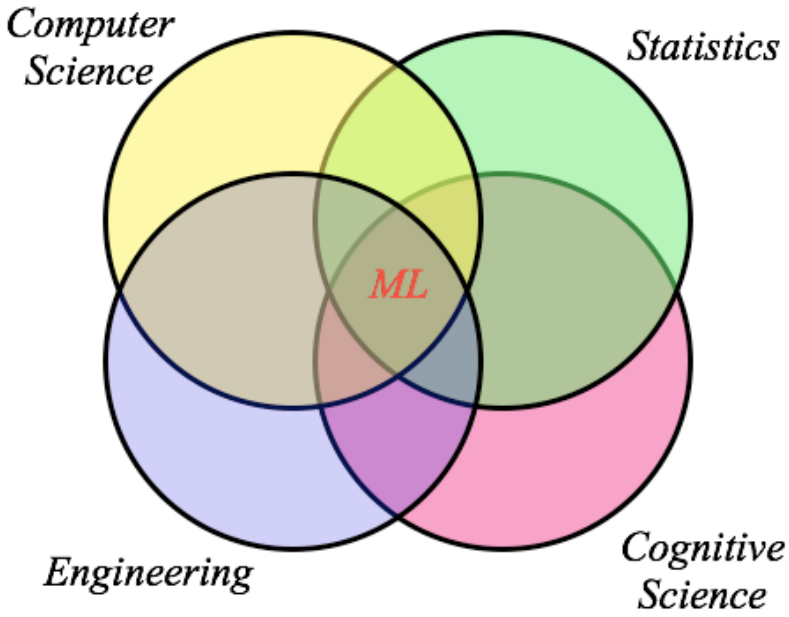
Department of Engineering
University of Cambridge

Outline

- What is Machine Learning?
- How does it fit into Information Engineering?
- What will we cover in part C of paper 8?
- Why is this useful?

Machine Learning

Machine learning is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.



- Other related terms: Pattern Recognition, Neural Networks, Data Mining, Adaptive Control, Artificial Intelligence, Decision Theory, Statistical Modelling...

Some Applications of Machine Learning

- Bioinformatics

  understanding genes, proteins and diseases

- Robotics

  autonomous systems that perceive and act

- Computer vision

  recognising and modelling images

- Computational neuroscience

  modelling the brain and neural data

- Financial prediction

  making lots of money

- Collaborative filtering

  recommending movies, books, etc based on preferences

- Information retrieval

  building better search engines

How does it fit into Information Engineering?

- Main research and teaching directions in Division F:
  - control systems
  - signal processing and tracking
  - computer vision
  - medical imaging
  - digital communications
  - computational neuroscience
  - machine learning
  - speech recognition and translation

systems that process data, learn to improve their behaviour, make predictions, make decisions, sense the outside world, and act on it.

Information Engineering Year 3 (Part IIA) Modules

- 3F1 Signals and systems
- 3F2 Systems and Control
- 3F3 Signal and pattern processing
- 3F4 Data transmission
- 3F5 Computer and network systems
- 3F6 Software engineering and design

Information Engineering Year 4 (Part IIB) Modules

- 4F1 Control System Design
- 4F2 Robust Multivariable Control
- 4F3 Non-linear and Predictive Control
- 4F6 Signal Detection and Estimation
- 4F7 Digital Filters and Spectrum Estimation
- 4F8 Image Processing and Image Coding
- 4F9 Medical Imaging and 3-D Computer Graphics
- 4F10 Statistical Pattern Processing
- 4F12 Computer Vision and Robotics
- 4F13 Machine Learning

What will we cover in part C ?

Image searching and modelling using machine learning methods

*We will focus on the application of pattern recognition and statistical machine learning methods to **image retrieval** and related problems. Although all examples will use images, the ideas are generally applicable to other domains, for example, web document retrieval, music, and financial data.*

Topics:

- Representing images as feature vectors
- Probabilistic models, use of Bayes rule, Bernoulli distributions and multivariate Gaussians
- Image retrieval
- Outlier removal and novelty detection
- A case study of an image retrieval method

What is Information Retrieval?

- finding material from within a large unstructured collection (e.g. the internet) that satisfies the user's information need (e.g. expressed via a query).

- well known examples…

Google™ YAHOO!® msn

- …but there are many specialist search systems as well:

PubMed www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health

ebay

UniProt
the universal protein resource

amazon.com

IMDb®
Earth's Biggest Movie Database™

Google™ Finance BETA

Spotlight
Find anything, anywhere, fast.

Traditional approach to information retrieval

- user types a text query
- system returns an ordered list of items

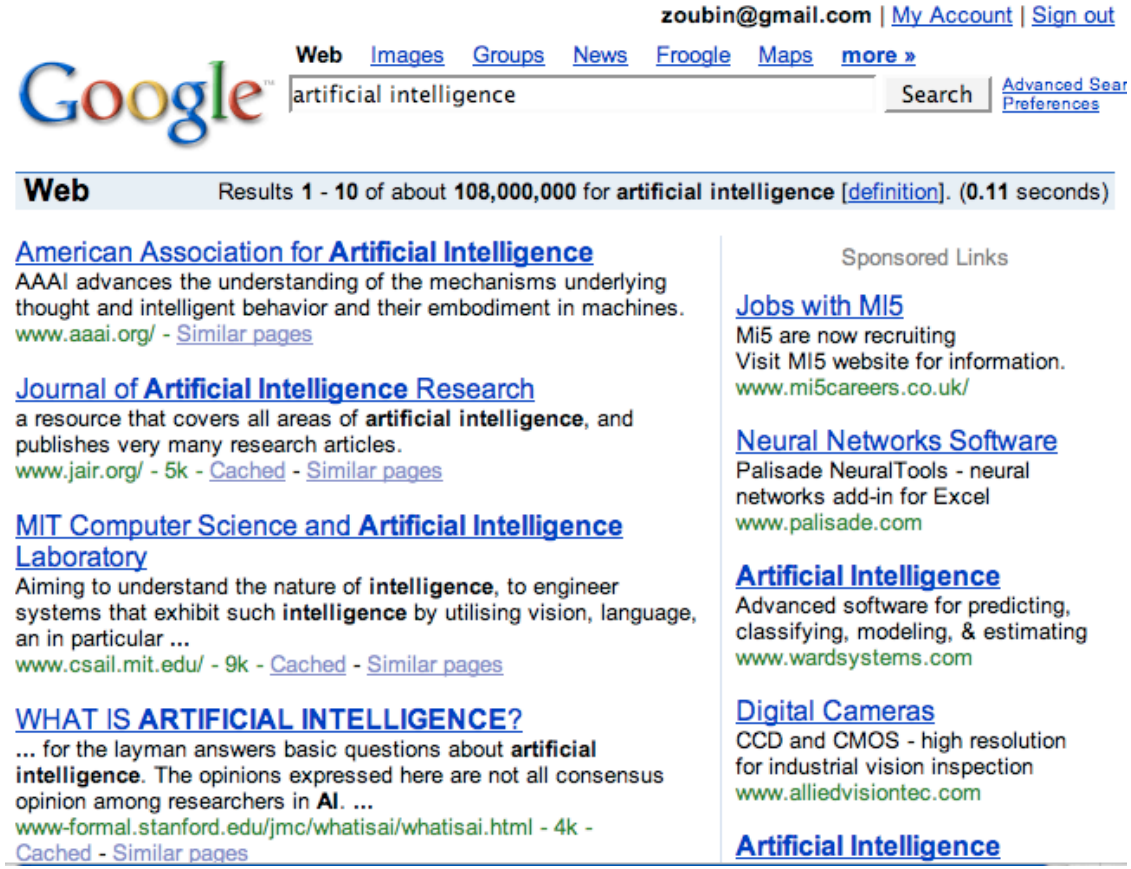
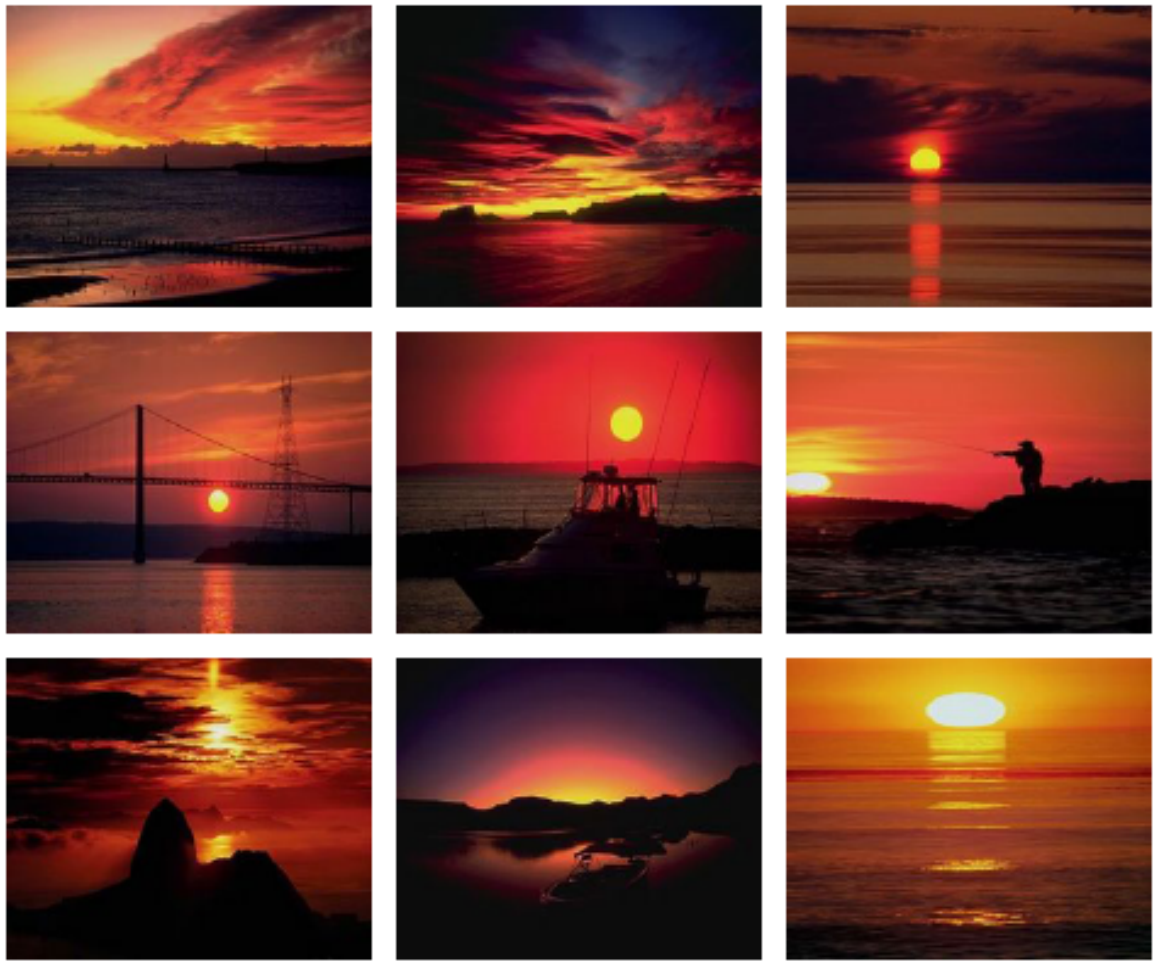


Image Retrieval Results for Query: “sunset”



These are the top 9 images returned. This system finds images of sunsets using **only** the color and texture features of these unlabelled images.

Results for Query: “sign”

These are the top 9 images returned. It finds images of signs using only the color and texture features of these unlabelled images.

Results for Query: “fireworks”

These are the top 9 images returned.

Why is this useful?

- Image retrieval itself is very useful…

  … but the principles can be used for many other retrieval problems.

Some Other Example Applications

- literature search: searching scientific literature, patent databases, or news articles by giving a small set of example articles
- targeted advertising: finding similar customers as represented by their buying patterns
- biomedical search: searching for sets of similar patients based on medical records
- drug discovery: searching for similar sets of proteins based on sequence, annotations, structure, literature
- collaborative filtering: finding similar movies, music, books, based on matching your preferences to other people's preferences
- online shopping: searching for products by giving a few examples
- online dating services / social networks: searching for people based on profiles
- finance: finding similar companies / stocks based on patterns of transactions

Movie Search Example Results

- Query:
  - Gone with the wind
  - Casablanca

- Result (top hits):
  - Gone with the wind (1939)
  - Casablanca (1942)
  - The African Queen (1951)
  - The Philadelphia Story (1940)
  - My Fair Lady (1964)
  - The Adventures of Robin Hood (1938)
  - The Maltese Falcon (1941)
  - Rebecca (1940)
  - Singing in the Rain (1952)
  - It Happened One Night (1934)

Movie Search Example Results

- Query:
  - Mary Poppins
  - Toy Story

- Result (top hits):
  - Mary Poppins
  - Toy Story
  - Winnie the Pooh
  - Cinderella
  - The Love Bug
  - Bedknobs and Broomsticks
  - Davy Crockett
  - The Parent Trap
  - Dumbo
  - The Sound of Music

Summary

- Information engineering studies systems that process data, learn to improve their behaviour, make predictions, make decisions, sense the outside world, and act on it.
- We will focus on **image searching and modelling** using machine learning methods.
- The tools you will learn about in part C are useful not just for image retrieval but also for many other information engineering problems.
- This should provide a good introduction to some of the material in years 3 and 4.

Appendix

Movie Search

- 1813 people rating 1532 movies
- query is a small set of movies
- system searches for other movies that would fit into this set based on the ratings

Searching Academic Literature

- Query:
  - A. Smola
  - B. Scholkopf

these two researchers published conference papers in the area of "support vector machines"

- Result (top hits):
  - A. Smola
  - B. Scholkopf
  - S. Mika
  - G. Ratsch
  - R. Williamson
  - K. Muller
  - J. Weston
  - J. Shawe-Taylor
  - V. Vapnik
  - T. Onoda

these are additional researchers who published conference papers in the area of "support vector machines"

Protein Search

- Proteins are the fundamental building blocks of life; our genes code for proteins
- Understanding the functions of and relationships between proteins is essential for bioscience, biomedicine, and drug discovery (a multi-billion dollar industry).
- We have built a protein retrieval system to search UniProt, an annotated database of 200,000+ proteins

A Prototype Image Retrieval System

- A system for searching large collections of unlabelled images.
- You enter a word, e.g. "sunset", and it retrieves images that match this label, using only color and texture features of the images
- A database of 32,000 images
  - **Labelled Images**: 10,000 images with about 3-10 text labels per image
  - **Unlabelled Images**: 22,000 images
  - Each image is represented by 240 binary color and texture features, no other information is used
- A vocabulary of about 2000 keywords
- Goal: we want to search the *unlabelled* images using queries which are subsets of the labelled images associated with keywords

The Image Retrieval Prototype System

The Algorithm:

1. Input query word: e.g. *w*="sunset"
2. Find all labelled images with label *w*
3. Use those images as a query set
4. Return the unlabelled images with the highest probability of belonging with the query set

The algorithm is very fast:

about 0.2 sec on a laptop to query 22,000 test images

code can be further optimized and parallelized

Example Labelled Images for “sunset”

These are 9 random images that were labelled “sunset” in the labelled training data.
Notice that these images are quite variable, and the labelling subjective and somewhat noisy.
Our retrieval system does very well and is quite robust to ambiguous categories and poor labelling.