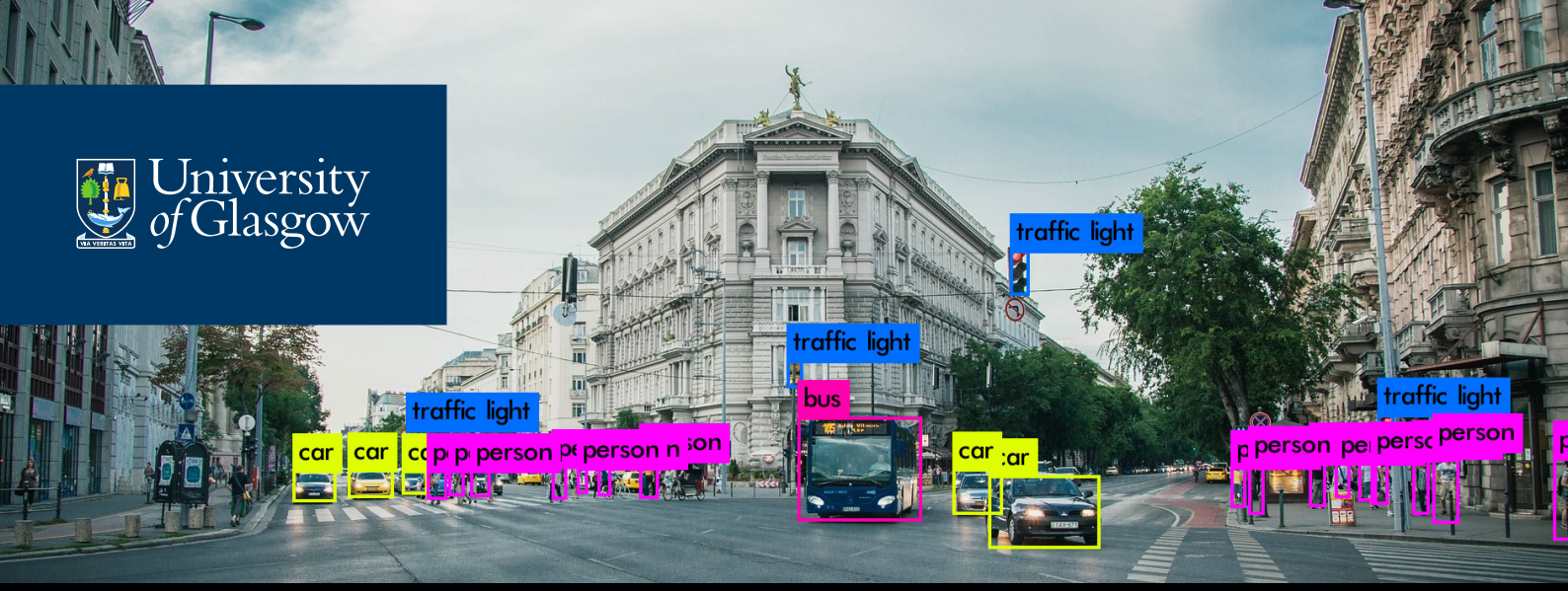
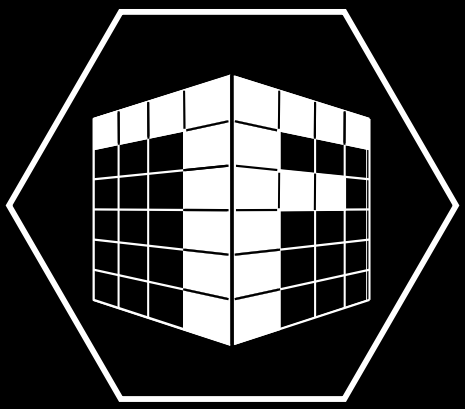


Large Scale Computing for Data Analytics

Course information sheet 2022-23

Full course, 10 weeks



This course aims to introduce learners to efficient implementation of computationally expensive data-analytic methods and data-analytic methods for big data including deep learning and convolutional neural networks, both in terms of applications and implementation in frameworks such as Tensorflow or Keras. This course discusses enterprise-level technology relevant to big data analytics such as Spark, Hadoop and NoSQL databases.

Prerequisite Knowledge

Learners should be familiar with the programming language Python. In addition, Learners should be knowledgeable in Bayesian statistics, statistical inference, generalised linear models and classification.

Intended Learning Outcomes

By the end of this course learners will be able to:

- assess and compare the complexity of an algorithm and implementation both in terms of computational time and memory, as well as suggest strategies for reducing those;
- describe key concepts of TensorFlow;
- perform basic computations with TensorFlow.
- distinguish between different types of deep and/or convolutional neural networks and choose an appropriate network for a given problem;
- fit a neural network using specialised frameworks such as Tensorflow or Keras and assess the result;
- discuss important methodological aspects underpinning deep learning;
- explain the differences between SQL and NoSQL databases and assess their suitability in different real-life settings;
- explain the basic concepts underpinning big data systems such as Spark or Hadoop and discuss their suitability and use in different scenarios.

Syllabus

Week 1 (sample material)
- Large-scale distributed computing
- Assessing computational cost and complexity
- Data parallelism
- The MapReduce paradigm

Week 2
- Introduction to TensorFlow
- Basic computations
- Overview of key concepts
- Simple linear regression with TensorFlow

Week 3
- Classification with TensorFlow
- Creating a classifier to recognise handwritten digits
- Visualisation deep learning
- Debugging TensorFlow

Week 4
- Understanding the underlying mechanics of TensorFlow
- Understanding key concepts needed to build Tensorflow models including optimisers, layers and activation functions

Week 5
- Deep learning for image classification
- Introduction to OpenCV
- Deep learning in Python using Keras

Mid-term week break

Week 6
- Introduction to convolutional neural networks;
- Applications of convolutional neural networks;
- Simple examples of convolutions;
- Convolutions with TensorFlow.

Week 7
- Analysing sequential data with recurrent neutral networks;
- Training recurrent neutral networks;
- Implementing recurrent neutral networks in TensorFlow.

Week 8
- Statistical computation and probabilistic modelling with TensorFlow Probability;
- Probabilistic programming;
- Understanding key features of TensorFlow Probability;
- Statistical inference with TensorFlow Probability;
- Bayesian statistics with TensorFlow Probability;
- Fitting generalised linear models with TensorFlow Probability.

Week 9
- Brief history of big data;
- Management, modelling and computational issues with big data;
- Data storage of big data;
- Introduction to Hadoop;
- Introduction to Spark.

Week 10
- Data Analytics using Spark.

"This course has opened my eyes to some of the work I'm likely to be doing in my workplace in the near future. It also helped to explain some topics to me which I'd previously heard of but had not managed to obtain a full understanding of."

Online Learning
- Weekly live sessions with tutors
- Weekly learning material (reading material, videos, exercises with model answers)
- Bookable one-to-one sessions with tutor(s)

Textbooks
Aurelien, G (2019) Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly Media, Inc.

Assessment
(for credit only)

This will typically be made up of 4 pieces of assessment, including online quizzes, an individual project and an assignment.

Software
To take our courses please use an up-to-date version of a standard browser (such as Google Chrome, Firefox, Safari, Internet Explorer or Microsoft Edge) and a PDF reader (such as Acrobat Reader). Learning material will be distributed through Moodle. Learners need to have access to Python and the machine learning framework TensorFlow. It is recommended that you use Jupyter Google colaboratory notebook for this course, however other options are available. Learners need to install Zoom for participating in video conferencing sessions. We recommend the use of a head set for video conferencing sessions.

DATA ANALYTICS GLASGOW
School of Mathematics and Statistics
University of Glasgow
http://gla.ac.uk/mdatagov
http://gla.ai
Email:
maths-stats-analyticscpd@glasgow.ac.uk