

# Differentiate your Objective

VLC =  $\int \int \int$  Vision,  
Learning & Control

## COMP6248 Differentiable Programming (and some Deep Learning)

Jonathon Hare

Vision, Learning and Control  
University of Southampton

# Machine Learning - A Recap

All credit for this slide goes to Niranjana

Data	$\{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N \quad \{\mathbf{x}_n\}_{n=1}^N$
Function Approximator	$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\theta}) + \nu$
Parameter Estimation	$E_0 = \sum_{n=1}^N \{\ \mathbf{y}_n - f(\mathbf{x}_n; \boldsymbol{\theta})\ \}^2$
Prediction	$\hat{\mathbf{y}}_{N+1} = f(\mathbf{x}_{N+1}, \hat{\boldsymbol{\theta}})$
Regularisation	$E_1 = \sum_{n=1}^N \{\ \mathbf{y}_n - f(\mathbf{x}_n; \boldsymbol{\theta})\ \}^2 + r(\ \boldsymbol{\theta}\ )$
Modelling Uncertainty	$p(\boldsymbol{\theta}   \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N)$
Probabilistic Inference	$\mathbb{E}[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N_s} \sum_{n=1}^{N_s} g(\boldsymbol{\theta}^{(n)})$
Sequence Modelling	$\mathbf{x}_n = f(\mathbf{x}_{n-1}, \boldsymbol{\theta})$

## What is Deep Learning?

Deep learning is primarily characterised by function compositions:

- Feedforward networks:  $\mathbf{y} = f(g(\mathbf{x}, \boldsymbol{\theta}_g), \boldsymbol{\theta}_f)$ 
  - Often with relatively simple functions (e.g.  $f(\mathbf{x}, \boldsymbol{\theta}_f) = \sigma(\mathbf{x}^\top \boldsymbol{\theta}_f)$ )
- Recurrent networks:  
 $\mathbf{y}_t = f(\mathbf{y}_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}) = f(f(\mathbf{y}_{t-2}, \mathbf{x}_{t-1}, \boldsymbol{\theta}), \mathbf{x}_t, \boldsymbol{\theta}) = \dots$

In the early days the focus of deep learning was on learning functions for classification. Nowadays the functions are much more general in their inputs and outputs.

# What is Differentiable Programming?

- Differentiable programming is a term coined by Yann Lecun<sup>1</sup> to describe a superset of Deep Learning.
- Captures the idea that computer programs can be constructed of parameterised functional blocks in which the parameters are learned using some form of gradient-based optimisation.
  - The implication is that we need to be able to compute gradients with respect to the parameters of these functional blocks. We'll start explore this in detail next week...
  - The idea of Differentiable Programming also opens up interesting possibilities:
    - The functional blocks don't need to be direct functions in a mathematical sense; more generally they can be *algorithms*.
    - What if the functional block we're learning parameters for is itself an algorithm that optimises the parameters of an internal algorithm using a gradient based optimiser?!<sup>2</sup>

<sup>1</sup><https://www.facebook.com/yann.lecun/posts/10155003011462143>

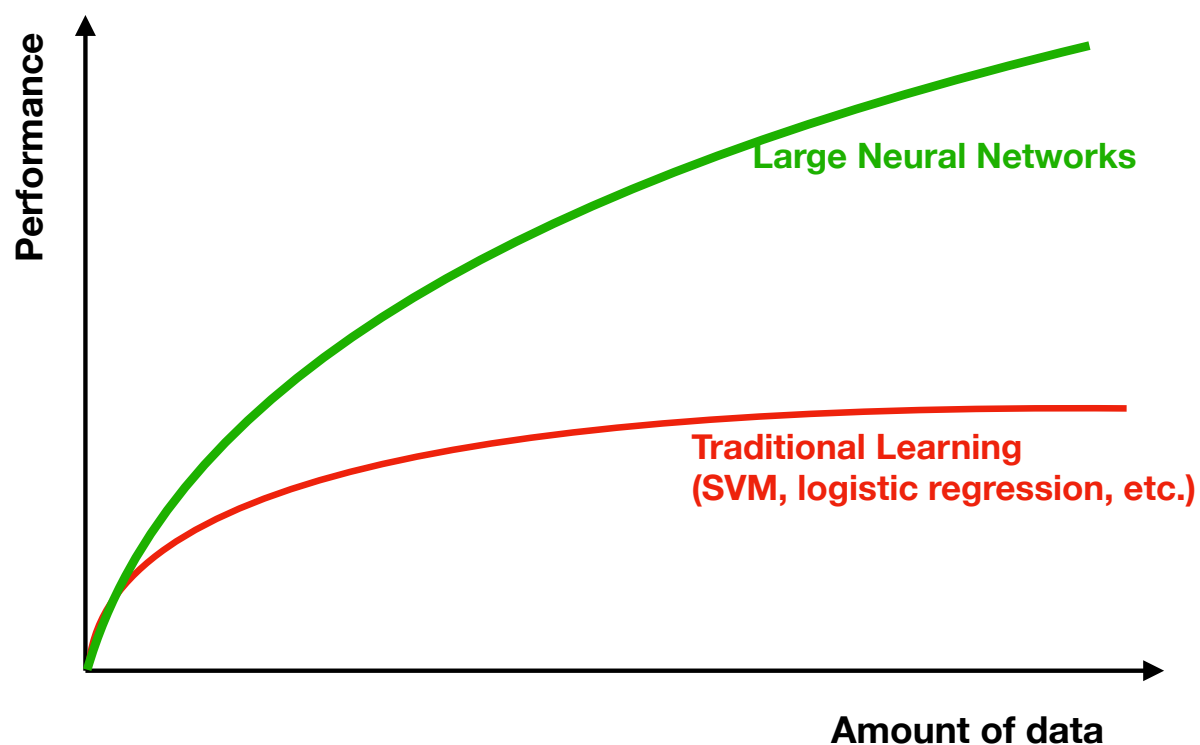
<sup>2</sup>See our ICLR 2019 paper: <https://arxiv.org/abs/1812.03928> and NeurIPS 2019 paper: <https://arxiv.org/abs/1906.06565>

## Is all Deep Learning Differentiable Programming?

- Not necessarily!
  - Most deep learning systems are trained using first order gradient-based optimisers, but there is an active body of research on gradient-free methods.
  - There is an increasing interest in methods that use different styles of learning, such as Hebbian learning, within deep networks. More broadly there are a number of us<sup>3</sup> who are interested in biologically motivated models and learning methods.
  - This course will primarily focus on differentiable methods, but we'll look at how relaxations can be made to make non-differentiable operators learnable with gradient-based optimisers.

<sup>3</sup>including at least myself, my PhD students and Geoff Hinton!

# Why should we care about this?



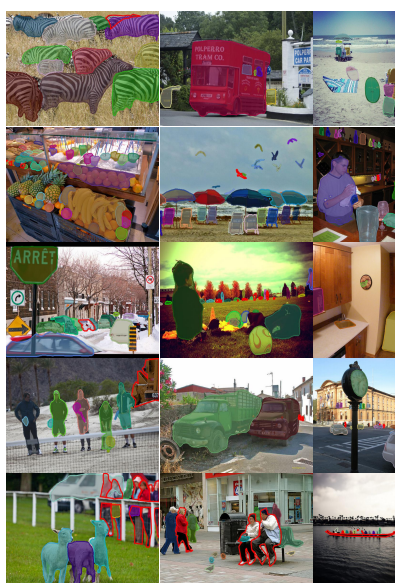
Reference: Andrew Ng

Jonathon Hare

COMP6248 Deep Learning

7 / 19

## Success stories - Object detection and segmentation



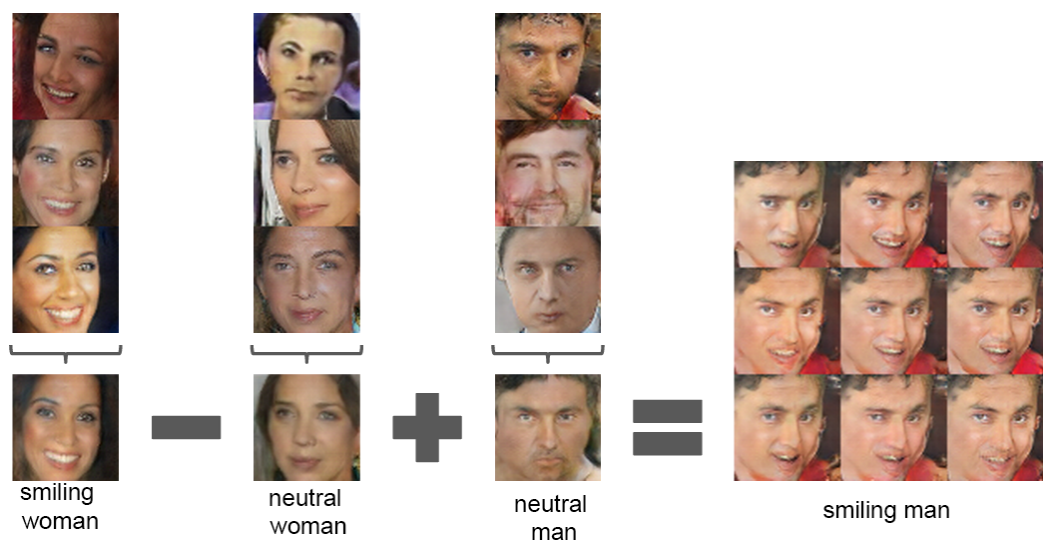
Pinheiro, Pedro O., et al. "Learning to refine object segments." European Conference on Computer Vision. Springer, 2016.

Jonathon Hare

COMP6248 Deep Learning

8 / 19

## Success stories - Image generation



Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

## Success stories - Translation

### ENGLISH TEXT

The reason Boeing are doing this is to cram more seats in to make their plane more competitive with our products," said Kevin Keniston, head of passenger comfort at Europe's Airbus.

### TRANSLATED TO FRENCH

La raison pour laquelle Boeing fait cela est de créer plus de sièges pour rendre son avion plus compétitif avec nos produits", a déclaré Kevin Keniston, chef du confort des passagers chez Airbus.

Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

# What is the objective of this module?

**A word of warning: This is not a module about how to apply someone else's deep network architecture to a task, or how to train existing models!**

You will learn some of that along the way of course, but the real objective is for you to graduate knowing how to understand, critique and implement new and recent research papers on deep learning and associated topics.

# What is the objective of this module?

- To gain an in-depth theoretical and practical understanding of modern deep neural networks and their applications.
- Understand the underlying mathematical and algorithmic principles of deep learning.
- Understand the key factors that have made differentiable programming successful for various applications.
- Apply existing deep learning models to real datasets.
- Gain facility in working with deep learning libraries in order to create and evaluate network architectures.
- Critically appraise the merits and shortcomings of model architectures on specific problems.

# How is this module going to be delivered?

- Lectures (2 per week)
  - Note: I am refreshing some material from last year, but the website still has links to the old slides.
  - You need to read the suggested papers/links before the lectures!
  - There is a little room for some flexibility later in the course on topics - tell me what you're interested in!
- Labs (1x 2 hour session per week for 8 weeks + additional help sessions)
  - Labs consist of a number of Jupyter notebooks you will work through.
    - You'll be using PyTorch as the primary framework, with Torchbearer to help out.
    - You will need to utilise GPU-compute for the later labs (we provide Google Colab links so you can use NVidia K80s in the cloud, but you'll also be able to use the RTX2070 in the lab machines if you wish).
  - I have scheduled lab sessions so you can get help and ask questions from myself and the demonstrator team.
  - After each lab you will have to do a follow-up exercise that will be marked.

# What will we cover in the module?

<http://comp6248.ecs.soton.ac.uk/>

Lab	Date	Topic
Lab 1	05/02/20	Introducing PyTorch
Lab 2	12/02/20	Automatic Differentiation
Lab 3	19/02/20	Optimisation
Lab 4	26/02/20	NNs with PyTorch and Torchbearer
	04/03/20	<i>NO LAB</i>
Lab 5	11/03/20	CNNs with PyTorch and Torchbearer
Lab 6	18/03/20	Transfer Learning
Break		
Lab 7	22/04/20	RNNs, Sequence Prediction and Embeddings
Lab 8	29/04/19	Deep Generative Models
	06/05/19	Coursework Help and Advice
	13/05/19	Coursework Help and Advice

## What do we expect you already know?

- COMP3223 or COMP6245 (fundamentals of statistical learning, MLPs, gradient descent, how to train and evaluate learning machines, supervised-vs-unsupervised)
- Fundamentals of:
  - Matrix Algebra (matrix-matrix, matrix-vector and matrix-scalar operations, inverse, determinant, rank, Eigendecomposition, SVD);
  - Probability & Statistics (1st-order summary statistics, simple continuous and discrete probability distributions, expected values, etc); and,
  - Multivariable Calculus (partial differentiation, chain-rule).
- Programming in Python



# What might you already know?

- How to use a deep learning framework (Keras, Tensorflow, PyTorch)?
- How to train an existing model architecture using a GPU?
- How to perform transfer learning?
- How to perform differentiable sampling of a Multivariate Normal Distribution?

## Assessment Structure

- Lab work 40% - Handin in week 11 (6th May, 4PM)
- Final project 40% - Handin in Week 12 (15th May, 4PM) (+ interim handin in week 6)
- Online quizzes 20% - Planned for Thursdays in Weeks 6 (5th-Mar) and 10 (30-Apr)

# The Main Assignment

## The ICLR Reproducibility Challenge

<http://comp6248.ecs.soton.ac.uk/coursework.html>