UNIVERSITY OF OXFORD

Mathematical Institute

Summary and Ingredients for a successful mini-project report

THEORIES OF DEEP LEARNING: C6.5, LECTURE / VIDEO 16
Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford Mathematics

Summary of the material covered.

Lecture 1: Three ingredients to deep learning

Architecture, data, and training

Mathematical Institute

- Structure of a deep net as repeated affine transforms and non-linear activations.
- Introduction to LeNet-5 with convolutional and fully connected layers.
- MNIST as an example of small dataset, along with the more complex imagenet dataset.
- Discussion of availability of computational resources and optimisation algorithms.

Lecture 2-4: Exponential expressivity with depth

Why depth from a function approximation perspective

- Telgarsky 15' sawtooth function giving example function with exponentially many maxima as a function of depth, but linear in width.
- Yarotsky 16' extension of the sawtooth to generate local polynomial approximations within ϵ needing $\log(1/\epsilon)$ depth.
- Poggio et al. 17' tree structure for approximation rate at the low effective dimensionality.
- Hein et al. 05' showing that MNIST, for example, has low effective dimension.

Lecture 5-6: Randomly initialized DNN exponential behaviour

Controlling the exponential with depth behaviour: correlation and gradients

- Glorot et al. 10' observation of pre-activation hidden layers being approximately Gaussian and normalizing to have constant variance with depth.
- Poole et al. 16' quantifying the convergence of the Gaussian variance through a computable recursion relation and developing a theory for the correlation between inputs dependence on $(\sigma_w, \sigma_b, \phi(\cdot))$. Controlling geometric collapse or instability through selecting $(\sigma_w, \sigma_b, \phi(\cdot))$ according to derived formulae.
- Pennington et al. 18' introduced the edge of chaos and connecting the Poole et al. work with exploding and vanishing gradients. Random matrix theory to derive moments of the spectra of the DNN.

- Foundational theory on stochastic gradient descent (SGD), making use of back-propagation and mini-batches to improve scalability of the algorithms.
- Reducing the gap between the value of a global minimizer and SGD through decreasing stepsize, increasing batchsize, or other variance reduction methods.
- Accelerating through momentum and diagonal scaling.
- Ward et al. AdaGrad scalar diagonal scaling to have reliable training over a wide range of stepsize initializations.

Lecture 9-10: Loss landscape properties and improvements

Simply connected landscapes and architecture choices to aid them

- Venturi et al. 16' showed the minimizers of the loss landscape can have paths between them where the loss landscape remains of a similar value, nearly simply connected.
- Pennington et al. 17' showed how the distribution of eigenvalues of the loss landscape's Hessian can be computed as a function of the number of trainable parameters compared to the amount of data available. With enough data, and when close to a minimizer, then the landscape is locally convex.
- Li et al. 18' illustrated how the loss landscape width near minimizers is impacted by training batch-size, and how ResNets can improve the landscape characteristics.
- Loffe 15' introduced batchnorm to have trainable bulk scaling to aid optimisation.

- Filters for early layers of CNNs show characteristic high-dimensional wavelet like structure.
- Deeper layers combine such filters and give greatest response for more structured inputs, leading to "memory" where some units in the CNN are maximized by objects within training classes.
- Such structure helps explain the efficacy of transfer learning.
- Mallat 12' introduced the Scattering Transform, which is a deep transform that only learns the final layers; activations are hand crafted to encourage desired invariants such as translation.
- Papyan 24' Attention sinks encode context and correlate token making up phrases and sentences

Lecture 12: Adversarial attacks and jailbreaking

Small changes can cause substantial changes through depth

- Goodfellow et al. 15' show that small changes to net inputs can cause adversarial missclassification.
- Engstrom et al. 18' showed that natural actions on objects, such as rotation or translation can also be used to generate misclassification; showing inherent lack of robustness in typical DNNs.
- Gopalakrishnan et al 18' showed how sparsifying a DNN can improve robustness, explainable by reducing an upper bound on the DNN's Lipshitz constant.
- Moosavi-Dezfooli et al. 16' considered universal adversarial attacks
- Liu et al. 16' showed that adversarial attacks are transferable between nets
- Carlini et al. 23' and Zou et al. 23' demonstrate multi-modal jailbreaking of commercial LLMs
- Hu et al. 24' Consider detecting jailbreaking attempts through gradient information

Lecture 13: Generative models, autoencoders, VAEs, and diffusi

How DNNs can be reversed to generate data

- Linear autoencoders are low-rank projections, obtainable through the SVD
- Kingma et al. 13' propose variational auto-encoders where the latent space is a probability measure that can be drawn from
- Goodfellow et al. 14' introduced the generative adversarial network (GAN) structure using reversed DNNs to generate data characteristic of a training dataset.
- Bau et al. 20' Filters analogous to those in CNNs on natural images are learned for generating natural images. They can be modified in order to influence expected properties of the generated data, such as colour or frequency of objects such as trees or doors in buildings.
- Diffusion models are the modern widely used variant where the generator is a sequence of denoising operators trained on specific datasets.

Lecture 14: Physics informed neural networks (PINNs)

Deep nets can be used to solve / learn PDEs

Mathematical Institute

- Raissi et al 19' considered neural nets as functions that can be used much like Fourier series for spectral methods and finite element methods
- The error analysis follows from the quadrature accuracy
- Wu et al. 23' Collocation points determine the accuracy; they can be chosen adaptively
- Lau et al. 24' Considered the residual for selecting the collocation points
- Krishnapriyan et al. 21' showed examples of failure modes of PINNs

Lecture 15: Pruning DNNs

Computational efficiency increasingly important

- Synthetic image generation of high resolution two minute video estimated to use about the same energy as the average UK home.
- Frankle et al 18' Lottery ticket hypothesis showed there exist sparse nets within dense nets with near full accuracy.
- Numerous algorithms developed for sparse weight matrices, with most important being structured sparsity and low-rank so that computational hardware has reduced energy consumption.
- Hu et al. 21' LoRA is the default fine-tuning method where low-rank matrices are added to weight matrices.
- Bolcskei et al. 19' Approximation theory for sparse nets
- Nait Saada et al. 23' Initialization theory for sparse and low-rank nets
- Many advances can be expected in the coming years

Some guidance on your mini-project

Course practicalities:

Mini-project assessment for OMMS and Part C 1 of 3

Select one of the following general areas and write a report on a specific subtopic within one of the areas:

1. List to be released later in the term, tentatively "Friday Dec. 5th of MT Week 8, due by Wednesday Jan. 7th 2026.

Your report must include a discussion of some theoretical portion of deep learning along with numerical simulations selected to highlight the issue being investigated and a discussion of how the numerical simulations reflect on the issue. Your numerical simulations should be novel in some regard; e.g. data set, architecture parameter choices, training algorithm, etc... Ensure that your report has as its focus a theoretical issue rather than an application of deep learning.

Course practicalities:

Mini-project assessment for OMMS and Part C 2 of 3

Deep learning is a rich topic under intense investigation. Two of the main venues for original research in this topic are the Neural Information Processing Systems (NeurIPS) conference and the International Conference on Machine Learning (ICML). Proceedings for these conferences are available at `https://papers.nips.cc` for NeurIPS and `http://proceedings.mlr.press` for ICML (Volume 97 for ICML 2019) as well as other conferences on machine learning. You are encouraged to select a topic and starting point for your report by perusing these proceedings papers and selecting an article to your liking.

Course practicalities:

Mini-project assessment for OMMS and Part C 3 of 3

Your report should be in the format and style of a NeurIPS Proceedings, abridged to not exceed 5 pages of text and graphics and 1 page of references (for a total length of not more than 6 pages). LaTeX style files and an example template are provided on the course page. Clearly indicate any ideas in your report which are your own and give appropriate attributions for all references (including research articles, software, data sets, etc....). Your report need not contain original research results, though you must use some original research articles as references (not just review articles or books). You should include a high level description of the code used to generate your numerical simulations, but should not submit the entire code; description of the code should not exceed one page of the report.

Considerations when selecting papers to discuss

Theory, code, and contrast.

- Your topic must focus on a theoretical aspect of deep learning, by this we mean not focusing on an application/dataset.
- Most topics benefit from code to explore the issue being discussed. It is much easier to adapt others code than to write you own. It is very beneficial to build your report on papers that have code available online, `https://paperswithcode.com`.
- Focusing to narrowly on one paper makes it hard to avoid restating their work. Contrasting two more more papers on the same topic gives is far better: it gives new information, your views and understanding, rather than restating others results.
- Papers that came out on the same topic at the same time often haven't been directly compared and are especially good candidates for developing a mini-project around.

Writing style and expectations

Academic tone and clarity; say what you did.

- You have now read a number of well written conference proceedings and journal articles in machine learning.
- Articles and proceedings that were recommended for reading are good examples of academic tone.
- Don't be overly informal, avoid words such as: "hope", "like to", "try to". These are fine in conversation, but not used in academic writing.
- Clearly state what you have done in the report; "In this paper we introduce...", "We gathered evidence that...", "Experiments in Figure 1 are novel (to the best of our knowledge)..." Don't leave the marker to have to guess what you did and what was done in the references.

Showing you have put a lot into the report

Novel experiments, well selected bibliography, not to close to lectures.

- These mini-project reports are not intended to be original results that can be published, but are intended to be novel in that they tell your views and understanding rather than strictly summarising a paper. Focus on contrasting and explaining.
- Contrasting or experiments might benefit from being conducted on datasets not already used in the original reports; this adds something your own to the report.
- Reports should not focus on papers covered in lecture; that isn't as novel as selecting your own papers to build from.
- The bibliography should be a useful set of references for the reader, and show give context to the topic.

Three exemplar mini-projects

These can give you a sense of what you can aspire to write.

Three mini-projects from 2021 are available on the course page:

- Robustness and Accuracy: Are we trying to have our cake and eat it too?
- On manifold mixup for deep learning
- Backpropagation and predictive coding: an experimental comparison

Robustness and Accuracy: Are we trying to have our cake and e too?

Contrasting, mathematical tone, and novel experiments.

- References [21] and [23] are contrasted; ”The authors of [21], take a different viewpoint.”
- A nice mathematical tone is set by stating and discussing theorems.
- A novel experiment is conducted ”Taking the framework provided by [21], we consider another model adjustment – batch normalisation (BN) ([9]).”
- You aren't expected to have a complete answer to the question being discussed. The results can be inconclusive: ”Unfortunately, the question we asked ourselves in the introduction remains an open one.”

On manifold mixup for deep learning

Literature overview, originality, and more originality

- An exceptional quick review of the topic and bibliography: see the first paragraph of the introduction, and related work section "The advent of Mixup has sparked a wave of related data augmentation techniques for images names Puzzle Mix [5], CutMix [12], and AugMix [3].
- Originality: "In this section we seek to explore two phenomena: 1. in light of Theorem 1, ... 2. Can we understand ..."
- Originality beyond what is expected: "In all existing Mixup methods... We propose a novel Mixup algorithm..." This is great, but isn't necessary.

Backpropagation and predictive coding: an experimental compar

Contrasting, choice of experiments, originality.

- Contrasting: "Throughout this work, I will try to explore the similarities and differences that arise..."
- Experiments conducted focus on connections to theory stated about the methods: "Predictive coding with condition 3 relaxed is the most similar to backpropagation."
- Techniques from lecture are used to explore phenomenon discussed: "Loss landscape of different models... This can explain the training instability experienced."
- Originality: "In this work I have shown how... Furthermore, adding a mild weight decay allows for...
- Note, this report is over the stated page limit; Figures 6 and 7 were not considered for the mark given. Don't go over the page limit....

This is an opportunity to explore something you enjoy

Focus on writing a good report and marks will follow

- Start with an outline, fill things in, go over the page length, and then condense by selecting the most essential parts of the discussion. (Ensure your final report is within the page limit.)
- Take time to read and improve your report. Ask yourself if you are making good use of space (does this need to be here? such as restating a proof).
- Having a mini-project that can be selected from a wide range of topics is an opportunity for you to investigate something you find interesting, pick something you enjoy and have fun.
- Ask questions and be inquisitive. Select experiments to explore the question you are posing. Convey your interest in understanding the topic. Let your interest in the topic come through in your writing.