Seminar: Theoretical Advances in Deep Learning

Debarghya Ghoshdastidar, Pascal Esser

TU Munich, Department of Informatics

Summer Semester 2021

Course information

- Master seminar (IN2107, IN4409)
  - 5 ECTS,   2 SWS

Course information

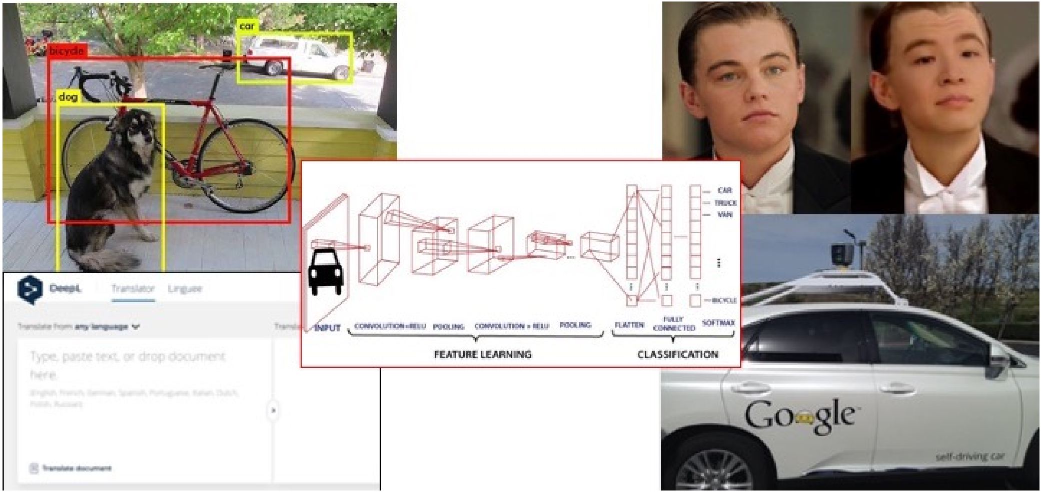
- Master seminar (IN2107, IN4409)
  - 5 ECTS, 2 SWS
- Organisers:
  - Pascal Esser (main coordinator of course)

    esser@in.tum.de
  - Prof. Debarghya Ghoshdastidar

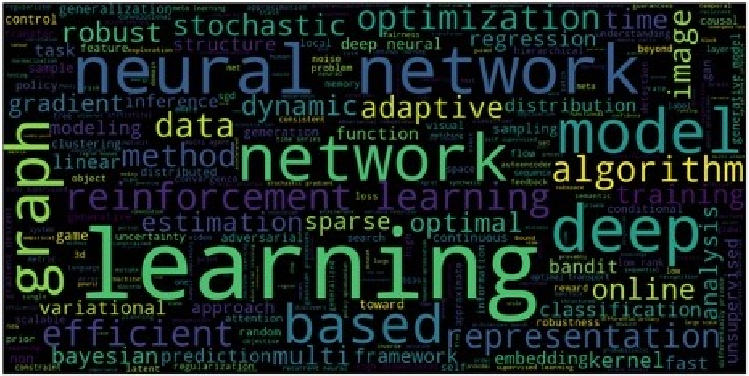
    Head, Theoretical Foundations of Artificial Intelligence

    ghoshdas@in.tum.de

Deep learning in practice



Research in machine learning and deep learning



- Most papers on new algorithms / architectures and their applications
- Important venues: ICML, Neurips, AAAI, CVPR, ICLR, ICCV, ...

Other types of papers in DL / ML

- New algorithms with partial theoretical analysis

Other types of papers in DL / ML

- New algorithms with partial theoretical analysis
    - Provides some understanding (less common in DL than ML)

Other types of papers in DL / ML

- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties

Other types of papers in DL / ML

- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties
  - Important when algorithms are hard to analyse theoretically
  - Common in deep learning, non-convex optimisation

Other types of papers in DL / ML

- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties
  - Important when algorithms are hard to analyse theoretically
  - Common in deep learning, non-convex optimisation
- Dedicated theory papers
  - Mathematically explain why DL / ML methods work (rare in DL)

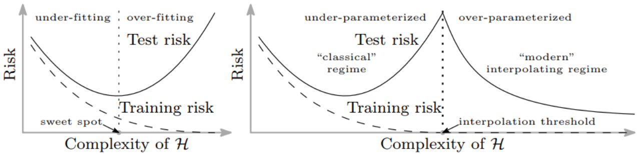
Other types of papers in DL / ML

- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties
  - Important when algorithms are hard to analyse theoretically
  - Common in deep learning, non-convex optimisation
- Dedicated theory papers ⟵ Focus of this seminar
  - Mathematically explain why DL / ML methods work (rare in DL)

Why do we need mathematical analysis of DL?

- Deep learning contradicts conventional wisdom

  Complex models generalise well



(a) U-shaped "bias-variance" risk curve

(b) "double descent" risk curve

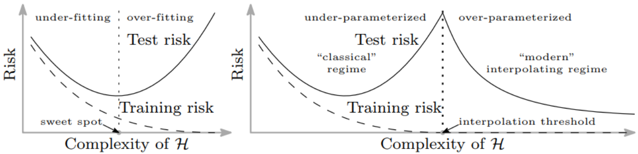
Why do we need mathematical analysis of DL?

- Deep learning contradicts conventional wisdom

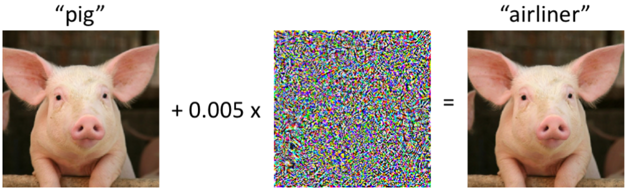
  Complex models generalise well

- Neural networks not robust

  Can be fooled to make error



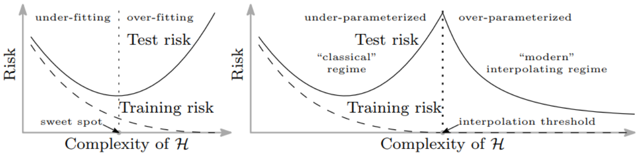
(a) U-shaped "bias-variance" risk curve (b) "double descent" risk curve



Why do we need mathematical analysis of DL?

- Deep learning contradicts conventional wisdom

  Complex models generalise well

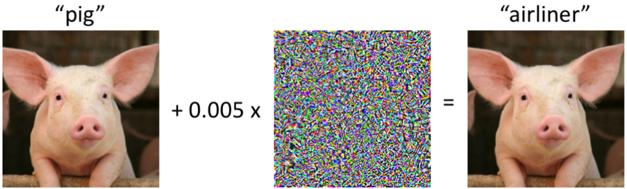


(a) U-shaped "bias-variance" risk curve

(b) "double descent" risk curve

- Neural networks not robust

  Can be fooled to make error



- The output of deep networks lack explainability

Purpose of this seminar

- Theory in deep learning emerging

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical foundations of learning (complements lecture IN2378)

Purpose of this seminar

- Theory in deep learning emerging
    - What do we know so far?
    - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical foundations of learning (complements lecture IN2378)
- Familiarise with mathematical proof techniques
    - Considerable focus on math in this seminar

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical foundations of learning (complements lecture IN2378)
- Familiarise with mathematical proof techniques
  - Considerable focus on math in this seminar
- Familiarise with publication and review process in ML

Focus of this seminar

Purpose of this seminar

- Theory in deep learning emerging

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical learning theory (complements lecture IN2378)

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical learning theory (complements lecture IN2378)
- Familiarise with mathematical proof techniques
  - Considerable focus on math in this seminar

Purpose of this seminar

- Theory in deep learning emerging
  - What do we know so far?
  - What are the limitations in theory, and gaps with practice?
- Familiarise with statistical learning theory (complements lecture IN2378)
- Familiarise with mathematical proof techniques
  - Considerable focus on math in this seminar
- Familiarise with publication and review process in ML

Focal points of this seminar

- Generalisation in neural networks
    - VC-dimension, sample complexity, ...

Focal points of this seminar

- Generalisation in neural networks
  - VC-dimension, sample complexity, ...
- Optimisation of neural networks

Focal points of this seminar

- Generalisation in neural networks
    - VC-dimension, sample complexity, ...
- Optimisation of neural networks
    - Convergence properties of trained networks

Focal points of this seminar

- Generalisation in neural networks
  - VC-dimension, sample complexity, ...
- Optimisation of neural networks
  - Convergence properties of trained networks
  - Asymptotic behaviour of wide neural networks (neural kernel)

Focal points of this seminar

- Generalisation in neural networks
  - VC-dimension, sample complexity, . . .
- Optimisation of neural networks
  - Convergence properties of trained networks
  - Asymptotic behaviour of wide neural networks (neural kernel)
  - Analysis of stochastic gradient descent

Focal points of this seminar

- Generalisation in neural networks
  - VC-dimension, sample complexity, . . .
- Optimisation of neural networks
  - Convergence properties of trained networks
  - Asymptotic behaviour of wide neural networks (neural kernel)
  - Analysis of stochastic gradient descent
- Robustness to adversarial attacks

Focal points of this seminar

- Generalisation in neural networks
  - VC-dimension, sample complexity, ...
- Optimisation of neural networks
  - Convergence properties of trained networks
  - Asymptotic behaviour of wide neural networks (neural kernel)
  - Analysis of stochastic gradient descent
- Robustness to adversarial attacks
- Effect of over-parametrisation in learning

Optimisation

Which features are relevant for learning?

- Gradient Starvation: A Learning Proclivity in Neural Networks, Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, Guillaume Lajoie

Biases introduced by the optimization algorithm

- The Implicit Bias of Gradient Descent on Separable Data, Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, Nathan Srebro

Graph Neural Networks

Generalization properties

- Generalization and Representational Limits of Graph Neural Networks, Vikas K. Garg, Stefanie Jegelka, Tommi Jaakkola

Stability

- Convergence and Stability of Graph Convolutional Networks on Large Random Graphs, Nicolas Keriven, Alberto Bietti, Samuel Vaiter

Robustness

- Rademacher Complexity for Adversarially Robust Generalization, Dong Yin, Kannan Ramchandran, and Peter Bartlett
- VC Classes are Adversarially Robustly Learnable, but Only Improperly, Omar Montasser Steve Hanneke Nathan Srebro
- Does Learning Require Memorization? A Short Tale about a Long Tail, Vitaly Feldman

Kernel Behaviour of Neural Netowrks

- A Generalized Neural Tangent Kernel Analysis for Two-layer Neural Networks
  Zixiang Chen, Yuan Cao, Quanquan Gu, Tong Zhang
- On the linearity of large non-linear models: when and why the tangent kernel is constant, Chaoyue Liu, Libin Zhu, Mikhail Belkin
- On Exact Computation with an Infinitely Wide Neural Net, Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang

Seminar details

- Link to seminar webpage
  - Go to `www.in.tum.de/tfai` → Teaching → Link to seminar (SS2021)
  - All information / papers will be added here
  - We will use Moodle for coordination

Seminar details

- Link to seminar webpage
  - Go to `www.in.tum.de/tfai` → Teaching → Link to seminar (SS2021)
  - All information / papers will be added here
  - We will use Moodle for coordination
- Desired number of participants = 12

Seminar details

- Link to seminar webpage
  - Go to `www.in.tum.de/tfai` → Teaching → Link to seminar (SS2021)
  - All information / papers will be added here
  - We will use Moodle for coordination
- Desired number of participants = 12
- Pre-requisites: Machine Learning (IN2064), Deep learning (IN2346)

Seminar details

- Link to seminar webpage
    - Go to `www.in.tum.de/tfai` → Teaching → Link to seminar (SS2021)
    - All information / papers will be added here
    - We will use Moodle for coordination
- Desired number of participants = 12
- Pre-requisites: Machine Learning (IN2064), Deep learning (IN2346)
- **Must be comfortable with mathematical techniques / proving results**
    - Taking Statistical foundations of learning (IN2378) in parallel would help

Assessment

- Everyone assigned one paper

Assessment

- Everyone assigned one paper
- Submit a report ($\sim$5 pages). Details will be provided in the introduction lecture.
  - summary of paper, explaining main results and their implications
  - review (we will discuss how to write reviews)
  - summary of proofs (main techniques, key lemmas and ideas)

Assessment

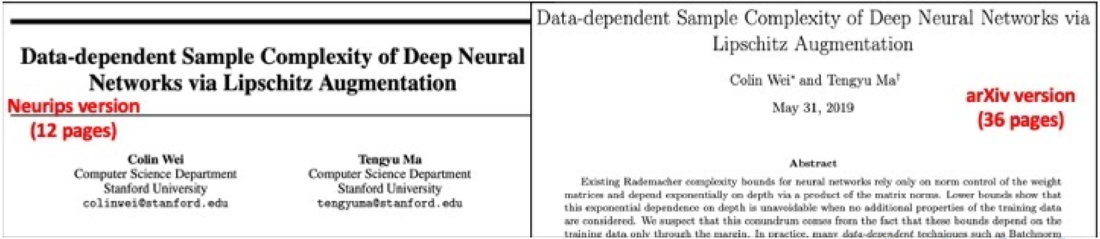
- Everyone assigned one paper
- Submit a report ($\sim$5 pages). Details will be provided in the introduction lecture.
  - summary of paper, explaining main results and their implications
  - review (we will discuss how to write reviews)
  - summary of proofs (main techniques, key lemmas and ideas)
- Present paper and your report
  - Block seminar; everyone needs to attend all talks

Assessment

- Everyone assigned one paper
- Submit a report (∼5 pages). Details will be provided in the introduction lecture.
  - summary of paper, explaining main results and their implications
  - review (we will discuss how to write reviews)
  - summary of proofs (main techniques, key lemmas and ideas)
- Present paper and your report
  - Block seminar; everyone needs to attend all talks
- Grading: Report (40%) + Presentation (60%)
  - Bonus for asking interesting questions to other speakers

Report + Presentation of papers

- Mostly publications from recent ML conferences (ICML, Neurips, COLT)
  - Conference papers are short (8 page, no proofs)
- **Report has to follow longer version on arXiv** (link will be provided)
  - Considerable focus on understanding mathematical results



Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation

Neurips version (12 pages)

Colin Wei
Computer Science Department
Stanford University
colinwei@stanford.edu

Tengyu Ma
Computer Science Department
Stanford University
tengyuma@stanford.edu

Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation

Colin Wei* and Tengyu Ma†

May 31, 2019

arXiv version (36 pages)

Abstract

Existing Rademacher complexity bounds for neural networks rely only on norm control of the weight matrices and depend exponentially on depth via a product of the matrix norms. Lower bounds show that this exponential dependence on depth is unavoidable when no additional properties of the training data are considered. We suspect that this conundrum comes from the fact that these bounds depend on the training data only through the margin. In practice, many data-dependent techniques such as Batchnorm

Timeline

- February 25 - March 10: Provide preference for papers (forms will be sent; select 3+ papers)
- March 10: Assignment of papers
- April 19, 14:00-16:00: First meeting (assignments, reports and organisation)
- May 3: Deadline for de-registration
- July 1: Submit report and first version of slides (both as PDF)
- End of July: Final presentation (block seminar, date to be finalised)
- Office hours: Monday 14:00-15:00 and Friday 14:00-15:00