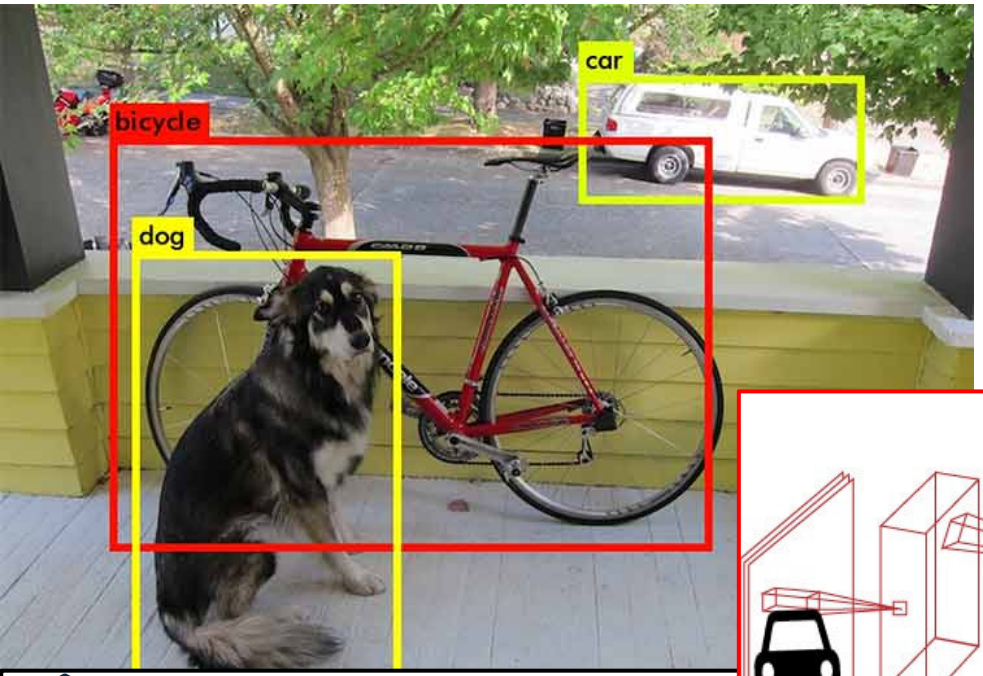
Theoretical Advances in Deep Learning

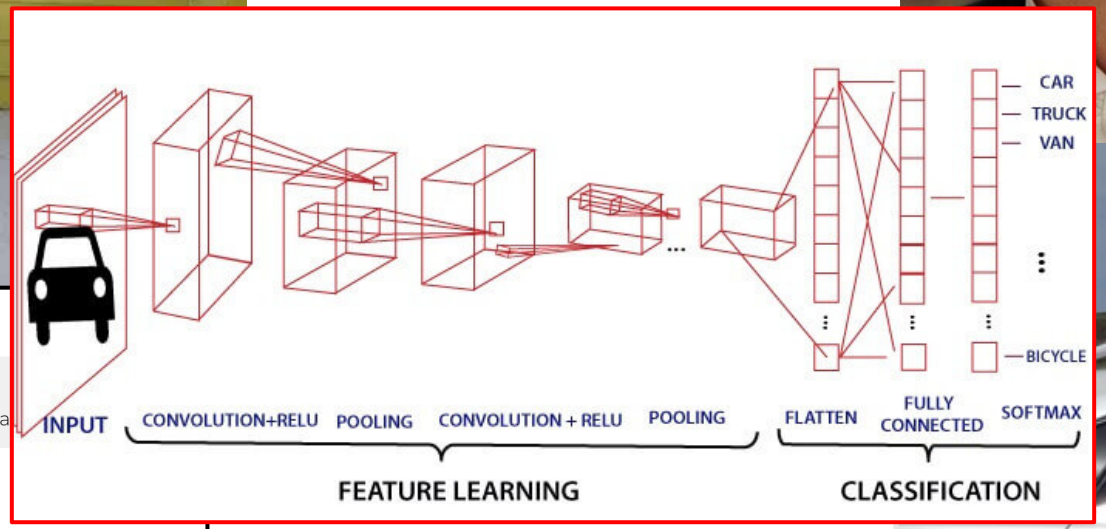
Masters Seminar (SS 2020)

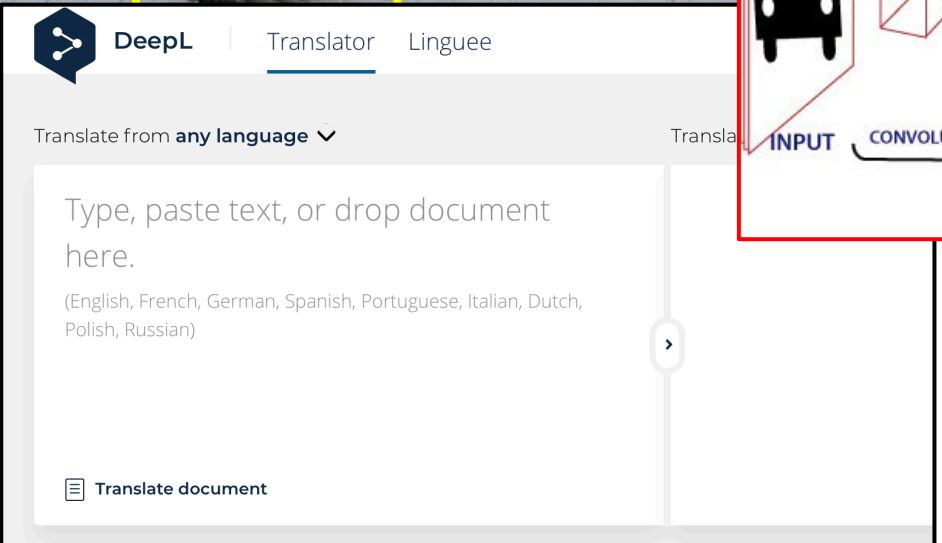
TUM Informatik

Debarghya Ghoshdastidar, Pascal Esser

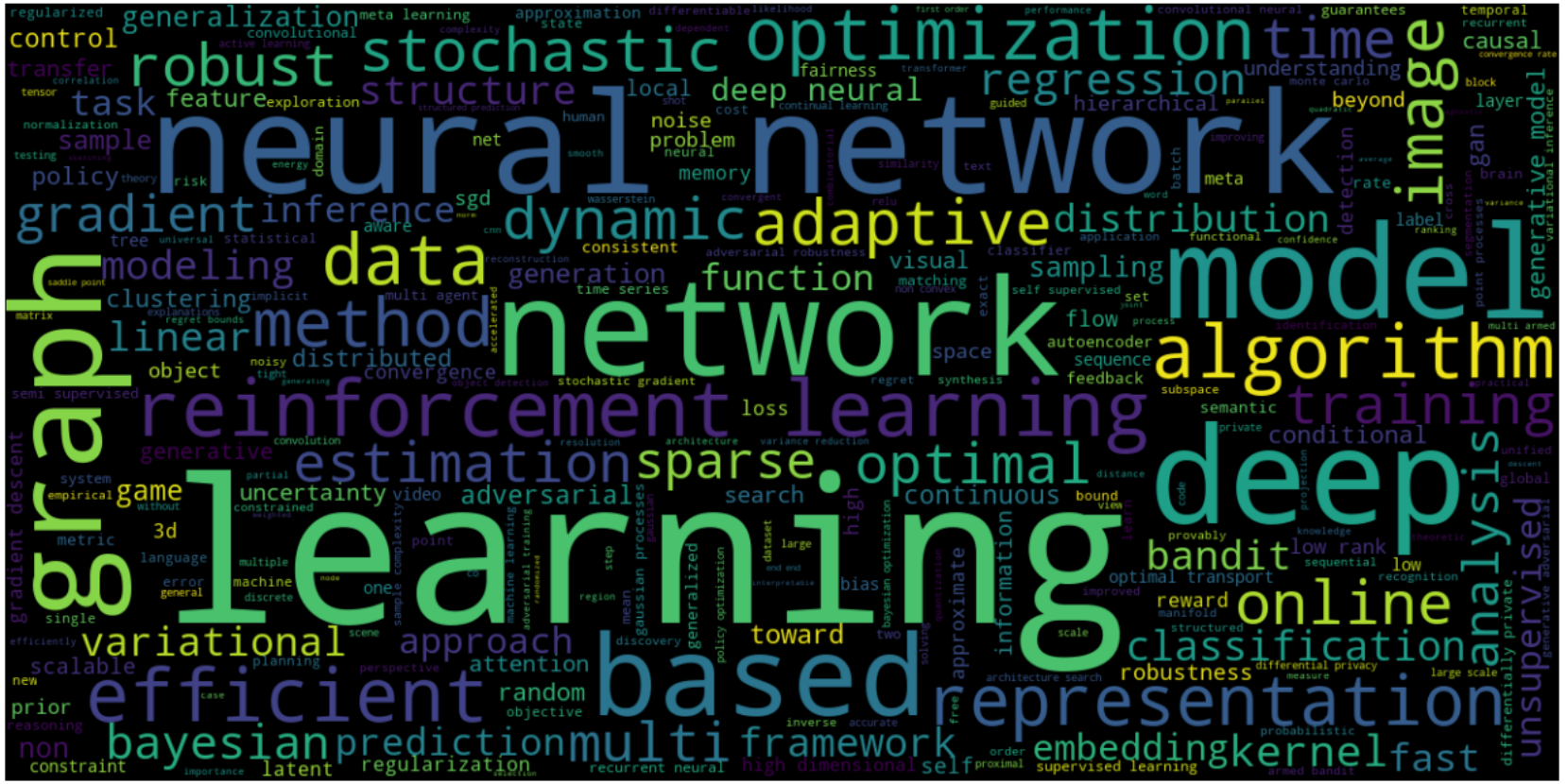
Deep networks in practice







Research in DL, NN and ML



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

Other types of papers in DL, NN and ML

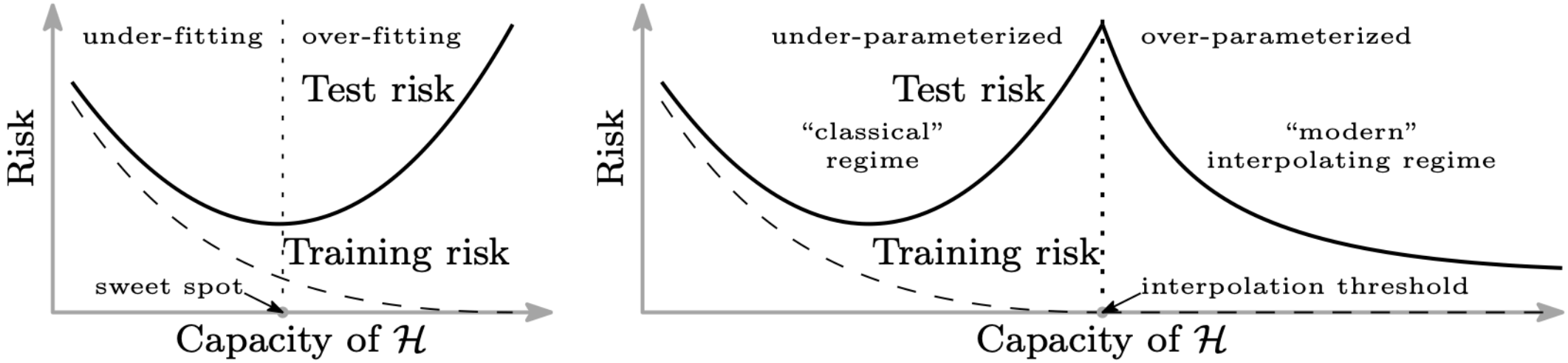
- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties
  - Important when algorithms are hard to analyse theoretically
  - Common in deep learning, non-convex optimisation
- Dedicated theory papers
  - Very small fraction (particularly in DL)
  - Explains performance under model / data assumptions

Other types of papers in DL, NN and ML

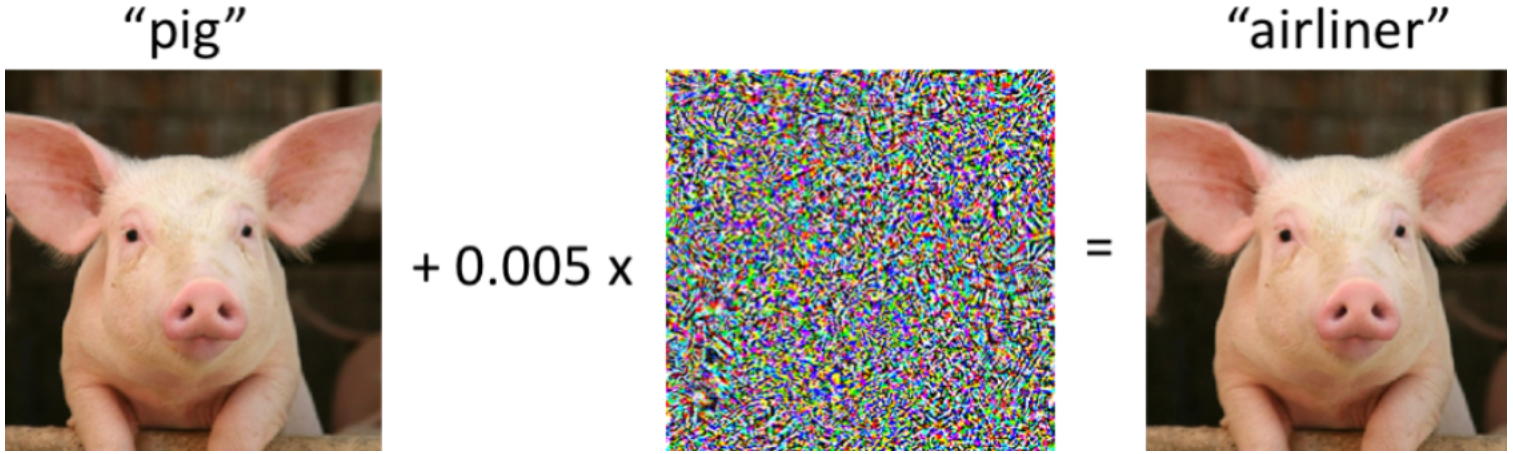
- New algorithms with partial theoretical analysis
  - Provides some understanding (less common in DL than ML)
- Empirical analysis of algorithmic properties
  - Important when algorithms are hard to analyse theoretically
  - Common in deep learning, non-convex optimisation
- Dedicated theory papers **Focus of this seminar**
  - Very small fraction (particularly in DL)
  - Explains performance under model / data assumptions

Surprises in DL … consequences of less theory

- Over-parametrisation is not bad



- Neural networks are not robust



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 different theories in ML
- Familiarise with mathematical proof techniques
  - Considerable focus on math in this seminar
- Familiarise with publication & review process in ML

Focal points of this seminar

- Generalisation in NN
  - VC-dimension, sample complexity, …
- Optimisation in NN
  - Convergence properties of trained networks
  - Effect of over-parametrisation
  - Analysis of stochastic gradient descent
- Robustness to adversarial attacks
- Other topics: Connections to kernels and random features

Generalisation for neural networks

VC-dimension:

- How complex are the trained models?

Sample complexity:

- How much training data to learn an accurate model?

- Nearly-tight VC-dimension bounds for piecewise linear neural networks. Nick Harvey, Christopher Liaw, Abbas Mehrabian. COLT 2017
- Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation. Colin Wei, Tengyu Ma. Neurips 2019

Optimisation in neural networks

Convergence in deep linear networks

- A convergence analysis of gradient descent for deep linear neural networks. Sanjeev Arora, Nadav Cohen, Noah Golowich, Wei Hu. ICLR 2019

Converged solution better with non-linearity

- Are ResNets Provably Better than Linear Predictors? Ohad Shamir. Neurips 2018

Optimisation in Over-parameterisated NNs

Generalization bound independent of network size

- Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruosong Wang. ICML 2019

Gradient descent does regularisation

- Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations. Yuanzhi Li, Tengyu Ma, Hongyang Zhang. COLT 2018 (best paper award)

Analysis of Stochastic Gradient Descent

With vanilla SGD, RNN can learn some concept classes

- Can SGD Learn Recurrent Neural Networks with Provable Generalization? Zeyuan Allen-Zhu, Yuanzhi Li. Neurips 2019.

Exponential convergence rates for SGD

- The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-parametrized Learning. Siyuan Ma, Raef Bassily, Mikhail Belkin. ICML 2018.

Adversarial ML / Robustness

Convergence of robust loss minimisation

- Convergence of Adversarial Training in Overparametrized Neural Networks. Ruiqi Gao et al. Neurips 2019.

Broader theory of risk bounds in presence of adversaries

- Theoretical Analysis of Adversarial Learning: A Minimax Approach. Zhuozhuo Tu, Jingwei Zhang, Dacheng Tao. Neurips 2019.

Other topics

Kernel behaviour of NNs

- Neural Tangent Kernel: Convergence and Generalization in Neural Networks. Arthur Jacot, Franck Gabriel, Clément Hongler. Neurips 2018
- A Convergence Theory for Deep Learning via Over-Parameterization. Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song. ICML 2019
- What Can ResNet Learn Efficiently, Going Beyond Kernels? Zeyuan Allen-Zhu, Yuanzhi Li. Neurips 2019

Other topics

NNs and random features

- On the Power and Limitations of Random Features for Understanding Neural Networks. Gilad Yehudai, Ohad Shamir. Neurips 2019

Memorisation

- Small ReLU networks are powerful memorizers. Chulhee Yun, Suvrit Sra, Ali Jadbabaie. Neurips 2019

Seminar details

- Webpage (all information / papers will be added here)

https://www.in.tum.de/tfai/lectureseminar/theoretical-advances-in-deep-learning/

  - (www.in.tum.de/tfai -> Teaching -> Link to seminar page)

- Participants
  - Desired number about 15 (or less)
  - Pre-requisite: Machine Learning (mandatory), Deep learning (preferred), Statistical foundations of learning (preferred)
  - Must be comfortable with mathematical techniques / proving results

Seminar details

- Papers
  - Mostly publications from recent ML conferences (ICML, Neurips, COLT)
  - Conferences have shorter versions without proofs
  - The report has to follow longer version on arXiv (link will be provided)
  - Considerable focus on mathematical results and proofs

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

Assessment

- Submit a report (5 pages)
  - 2 page summary of paper, explaining main results and their implications
  - 1 page review (we will discuss how to write reviews)
  - 2 page 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

Time plan

- February 22-29: Provide preference for papers

  (forms will be sent; select 3+ papers)

- March 6: Assignment of papers
- April 30: First meeting (assignments, reports and organisation)
- May 3: Deadline for de-registration
- June 1: Submit report and first version of slides (both as PDF)
- June 24-26: Final presentation (block seminar, date to be finalised)
- Office hours: 1 hour every week (date to be fixed)