Advanced Deep Learning: Best Practices

Anders Søgaard

coastal

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
4	Phillip/Anders	May 16 (P): Representation and Adversarial Learning May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018) ; Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Architectures

Framework

Fairness / Explainable AI

Methodology

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
4	Phillip/Anders	May 16 (P): Representation and Adversarial Learning May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Today

a) Peace, love and understanding
b) Play it again, Sam
c) The trouble with benchmarks

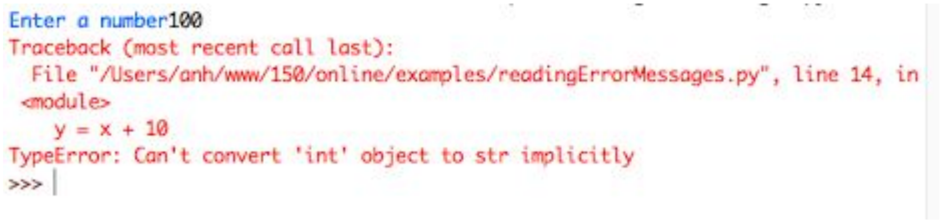
Peace, love, and understanding

Breaking Silently

Bugs normally throw errors, but there's many ways for a DNN to break silently:

- You clipped your loss instead of your gradients.
- You initialize with a pretrained checkpoint, but forget to initialize the mean.
- The labels are off for your synthetic pretraining data.

Each of these will not cause errors, but slightly worse performance.



Six Step Training

1. Look at the data, don't look at the data (Kevin Knight)
2. Checks and baselines
3. Overfit
4. Regularize
5. Tune (randomly)
6. Squeeze (ensembling and training longer)

Look at the data

Challenges

- Variation (noise, mix)
- Bias (risks)
- Class imbalance

Opportunities

- Simple baselines
- Task decomposition
- Task synergies

Checks and baselines

Checks

- Verify it's possible to overfit one batch.
- Verify that greater capacity leads to lower training loss.
- Visualize prediction dynamics.

Baselines

- Majority baseline
- Nearest neighbor
- Single-layer network (LR)
- Feed-forward network

Regularization + tuning

1. More data
2. Data augmentation
3. Self-supervised pre-training
4. Multi-task learning
5. Smaller batch size
6. Drop-out
7. Random hyperparameter search (because DNNs are typically only sensitive to a few parameters)

Squeezing

Train longer See Week 5. *Motivations:* Information bottleneck, wide valleys, high entropy solutions.

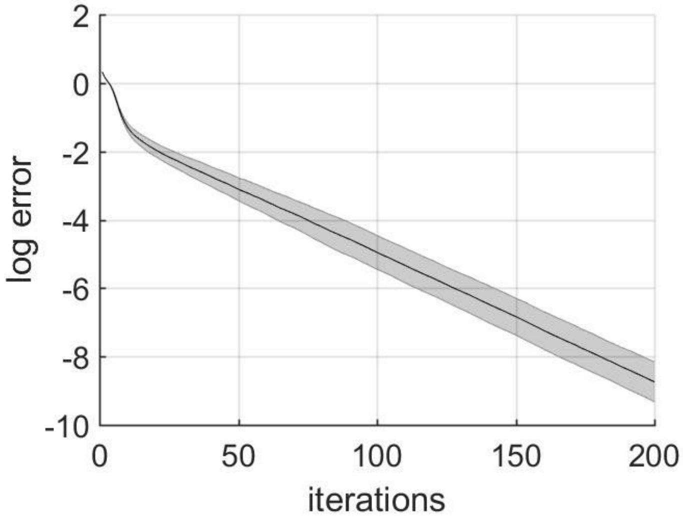
Ensembling *Strategies:* Voting, mixture of experts, stacking. If you cannot afford inference time, you can distill a simple model from the ensemble.

Play it again, Sam

Sources of randomness

1. Random model initialization
2. Random ordering of training data
3. Random noise injection
4. Random hyperparameter selection

Sources of randomness



Expected Validation Performance

- Validation performance is a function of compute budget
- Since compute budgets differ across labs and companies, why not estimate the expected (maximum) validation performance given budget?
- See Dodget et al. (2019) for details.

The trouble with benchmarks

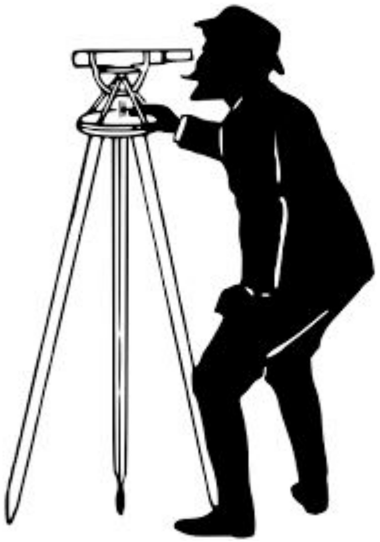
Community-wide overfitting

Benchmarks such as MNIST, SQuAD, GLUE, and ImageNet have become ridiculously over-represented. E.g., there's more than 300 publicly available QA datasets, but almost everyone uses SQuAD. Benchmarks are quickly saturated, though, and the research literature becomes a source of indirect test data leaks.

What is a benchmark?

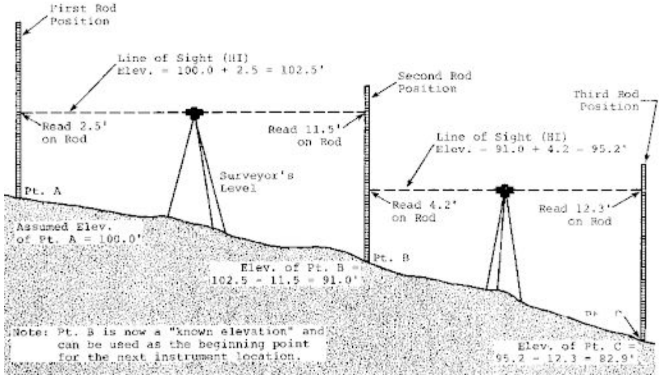
A benchmark is *a mark for accurately repositioning a leveling rod* (‘bench’).

Benchmarks form networks and are used to draw height maps.



Benchmarks

A single benchmark is **not** useful for a surveyor. Height maps rely on multiple benchmarks.



Beyond de facto benchmarks

Søgaard et al. (2014) argue we need to compute significance across samples, not across data points (within samples). Ideally, these should be sampled in a shift-pessimistic fashion (Liu et al., 2015).

https://www.aclweb.org › anthology PDF

Farewell Freebase: Migrating the SimpleQuestions Dataset to ...

by M Azmy · 2018 · Cited by 13 — The SIMPLEQUESTIONS dataset (Bordes et al., 2015) has emerged as the **de facto benchmark** for evaluating these simple questions over ...

https://www.aclweb.org › anthology PDF

Strong Baselines for Simple Question Answering over ...

by S Mohammed · 2018 · Cited by 55 — has emerged as the **de facto benchmark** for evalu- ating simple QA over knowledge graphs. The original solution of Bordes et al. (2015) fea- tured memory ...

https://www.aclweb.org › anthology PDF

A Supervised Term Weighting Scheme for Sentiment Analysis ...

by Y Kim · 2014 · Cited by 21 — has become the **de facto benchmark** for sen- timent analysis (Pang and Lee, 2004). • IMDB: 50k full-length movie reviews (25k training, 25k ...

https://www.aclweb.org › anthology PDF

Replication issues in syntax-based aspect extraction for ...

by E Marrese-Taylor · 2017 · Cited by 9 — 2004b) which became the **de facto benchmark** for evaluation in syntax-based aspect-based opinion mining. This is also a very important part of our environment.

https://www.aclweb.org › anthology PDF

Proceedings of the Student Research Workshop at the 15th ...

2004b) which became the **de facto benchmark** for evaluation in syntax-based aspect-based opinion mining. This is also a very important part of our environment.

https://www.aclweb.org › anthology PDF

Proceedings of the 5th Workshop on ... - ACL 2014

27 Jun 2014 — has become the **de facto benchmark** for sen- timent analysis (Pang and Lee, 2004). • IMDB: 50k full-length movie reviews (25k training, 25k ...

What to do

1. Multiple test datasets
2. Error analysis, including local interpretability
3. Adversarial examples and challenge datasets
4. Ablation and global interpretability
5. Downstream end user evaluations