Advanced Deep Learning: Three talks

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)

Three talks

a) We Need to Talk About Random Splits (EACL 2021) - External PDF
b) Locke's Holiday (EMNLP 2021)
c) Square One Bias (ACL 2022) - External PDF

Locke's Holiday: Belief Bias in Machine Reading

Anders Søgaard

coAStaL

Motivation

Question

Context: It is rarely the case that a buddhist meditates. Instead he plays drums.

Question: What does a buddhist do?

Answer: plays drums

Prediction: meditates

Explanation: Coreference or lexical association?

Question

Context: James is not a fan of U2, but of carrots.

Question: What is James a fan of?

Answer: carrots

Prediction: U2

Explanation: Ellipsis or lexical association?

Question

Context: London stinks of smog. Dublin stinks of people.

Question: Why does Dublin stink?

Answer: people

Prediction: smog

Explanation: Lexical association?

Question

Context: Washington is a number. Boston is a city.

Question: What is Washington?

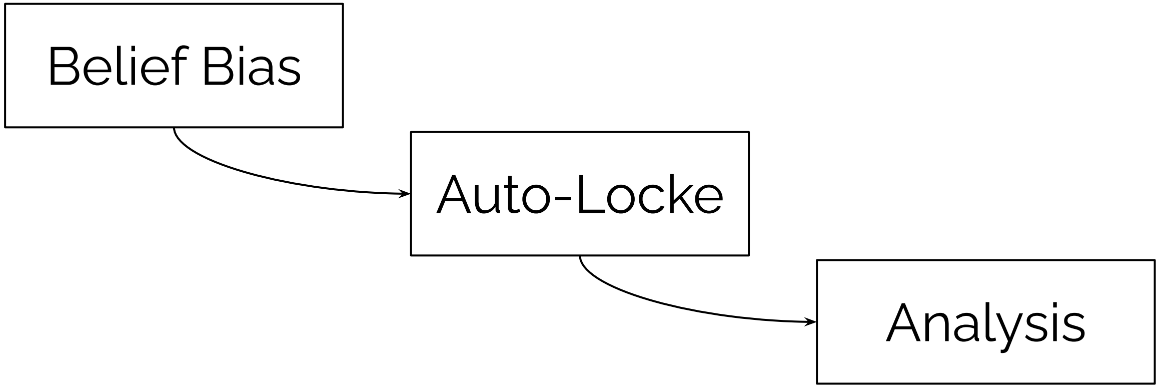
Answer: a number

Prediction: a city

Explanation: Lexical association?

Talk

Outline



Sample bias	Social bias	Guideline bias *(Hansen & Søgaard, 2021)*
Inductive bias	Label bias	Exposure bias (Ranzato et al., 2016)
Anchoring bias *(Berzak et al., 2016)*	Leakage	Metric misalignment

Blind spot bias	Clustering illusion	Belief bias
Bandwagon effect	Reactive devaluation	Endowment effect
Anchoring bias *(Berzak et al., 2016)*	Courtesy bias	Status quo bias

Blind spot bias	Clustering illusion	Belief bias
Bandwagon effect	Reactive devaluation	Endowment effect
Anchoring bias *(Berzak et al., 2016)*	Courtesy bias	Status quo bias

Belief Bias

- Psychology: When prior beliefs distorts reasoning process.
- Machine reading: The unwillingness to let context information override prior beliefs.
- If the prediction is a reasonable answer to the question *in isolation*, but clearly false *in context*, the effect can be attributed to **belief bias**.

Question (real-life)

Context: Indonesia is the Germany of the Asean. So then, Malaysia is the France.

Question: What country is Indonesia similar to?

Answer: Germany

Prediction: Malaysia

Dataset Construction

- Create 20 examples by hand (see above examples). All examples require little or no reasoning.
- Verification that the 20 examples are solvable by humans. **No** errors when crowdsourcing.
- Identify variable phrases in each example.
- Replace the phrase in focus (using WordNet to ensure grammaticality) and populate the remaining by nearest neighbors in GloVe space.
- Create 11,699 examples this way (some phrases were not in GloVe).

Question (Auto-Locke)

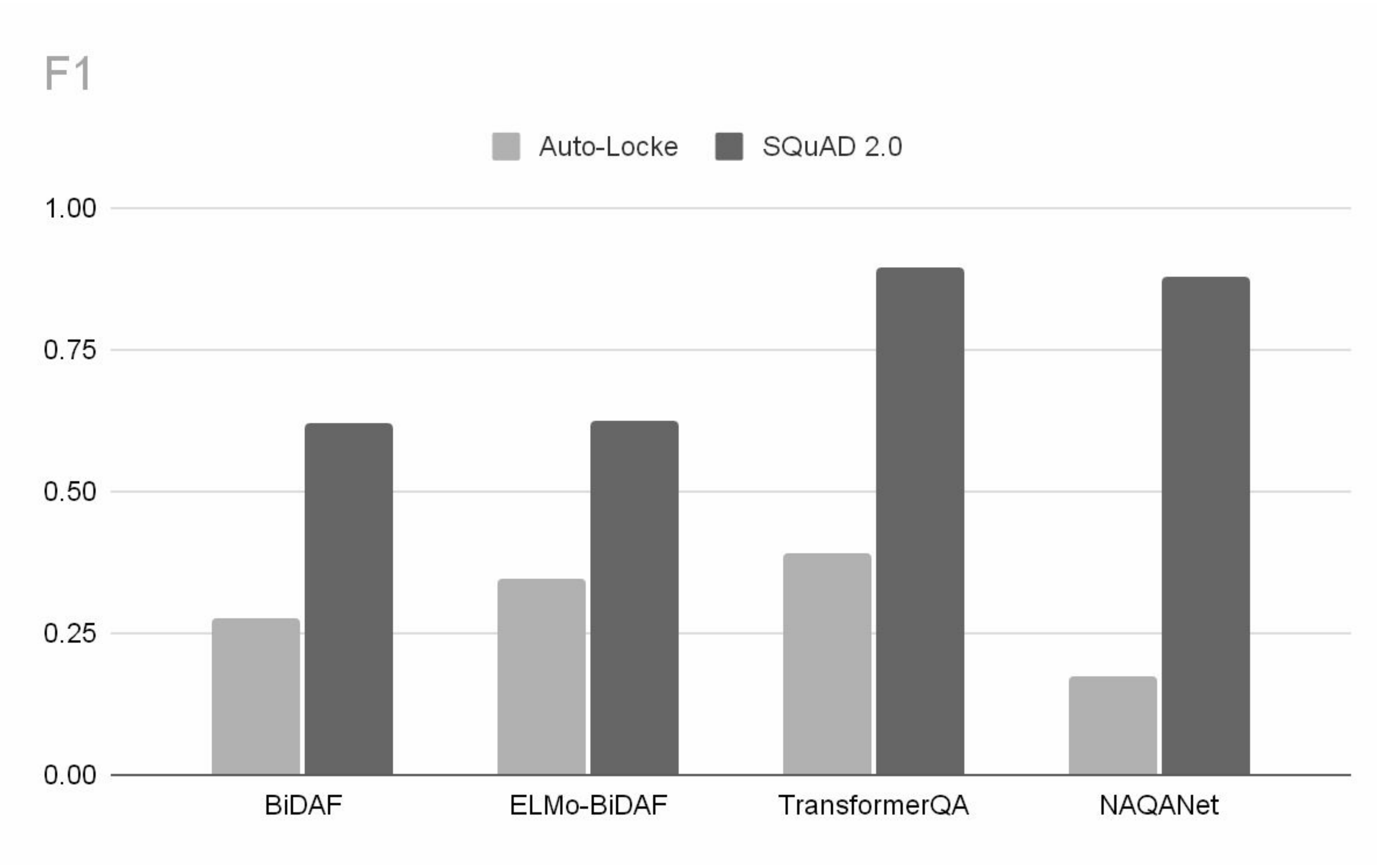
WordNet WordNet GloVe GloVe

Context: A ranch is a lobe. A vineyard is an inn.

Question: What is a ranch?

Answer: lobe

Prediction: ?


F1
Auto-Locke
SQuAD 2.0
1.00
0.75
0.50
0.25
0.00
BiDAF
ELMo-BiDAF
TransformerQA
NAQANet

Analysis

- Performance drops dramatically in spite of the examples being short and requiring no or limited inference.
- NAQANet drops 71% F1 (absolute).
- Equally poor performance if using random phrases instead of nearest neighbors, e.g., ***Context:*** *Bondsman is a winning post. Megillah is a giantism.* ***Question:*** *What is a bondsman?* So effect not explained by distractors.
- Also, true answers are in different positions. So performance not explained by recency either.