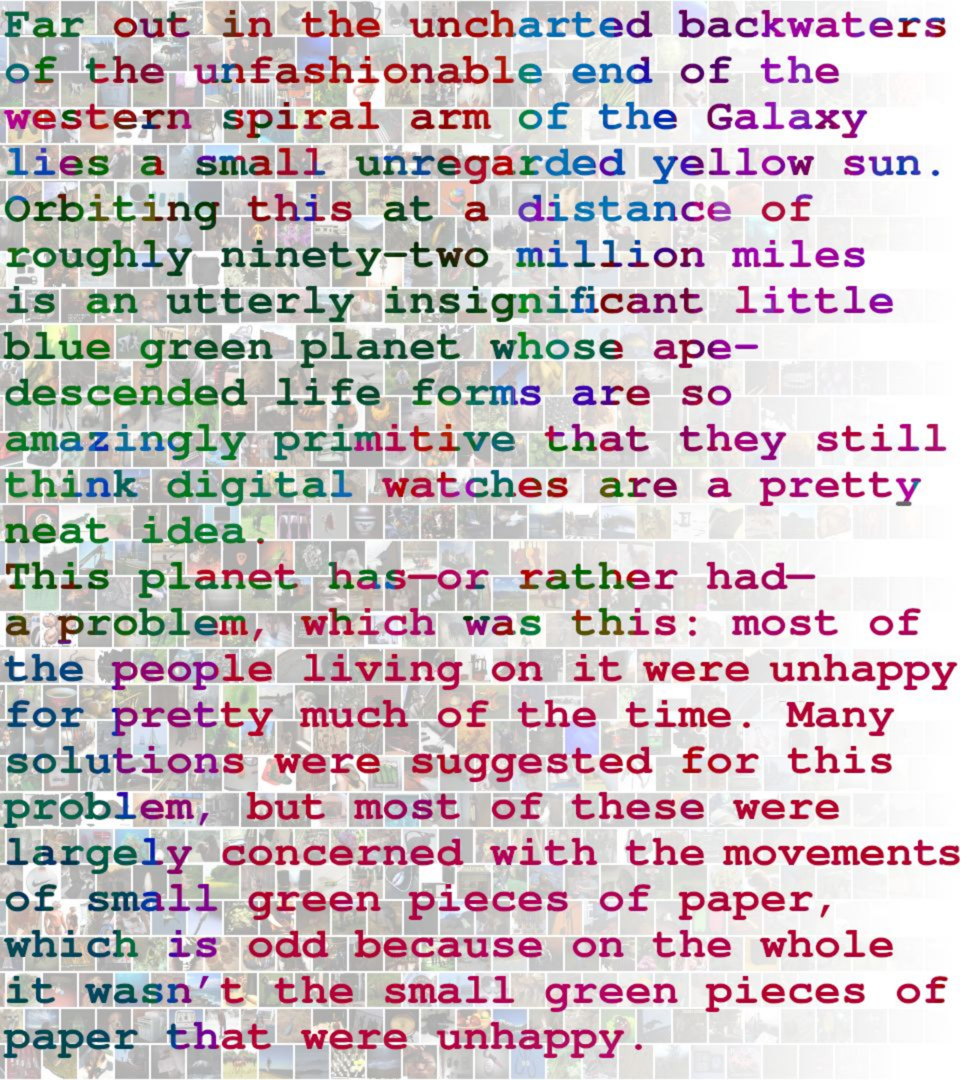


Learned representations and what they encode

CLASP Seminar 2021-01-20

Olof Mogren, PhD
RISE Research Institutes of Sweden

RI. SE

About me

- Computer scientist
- Research interest
  - Machine learning
  - Representation learning
  - Multi modal modelling
  - Uncertainty quantification
  - Privacy

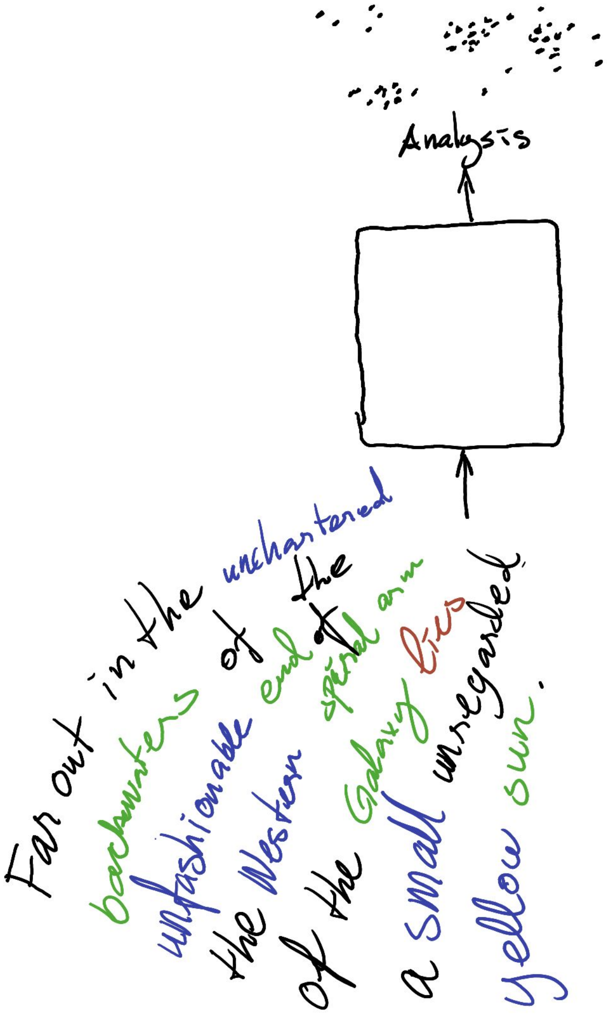
Natural language processing (NLP)

A field of research.

Language data: language: a kind of protocol for inter-human communication; **discrete**

Tasks: classification, translation, summarization, generation, understanding, dialog modelling, etc. (many; diverse)

Solutions: many; diverse.



Word embeddings *was* transfer learning for language

king

- ('kings', 0.71)
- ('queen', 0.65)
- ('monarch', 0.64)
- ('crown_prince', 0.62)

queen

- ('queens', 0.74)
- ('princess', 0.71)
- ('king', 0.65)
- ('monarch', 0.64)

Stockholm

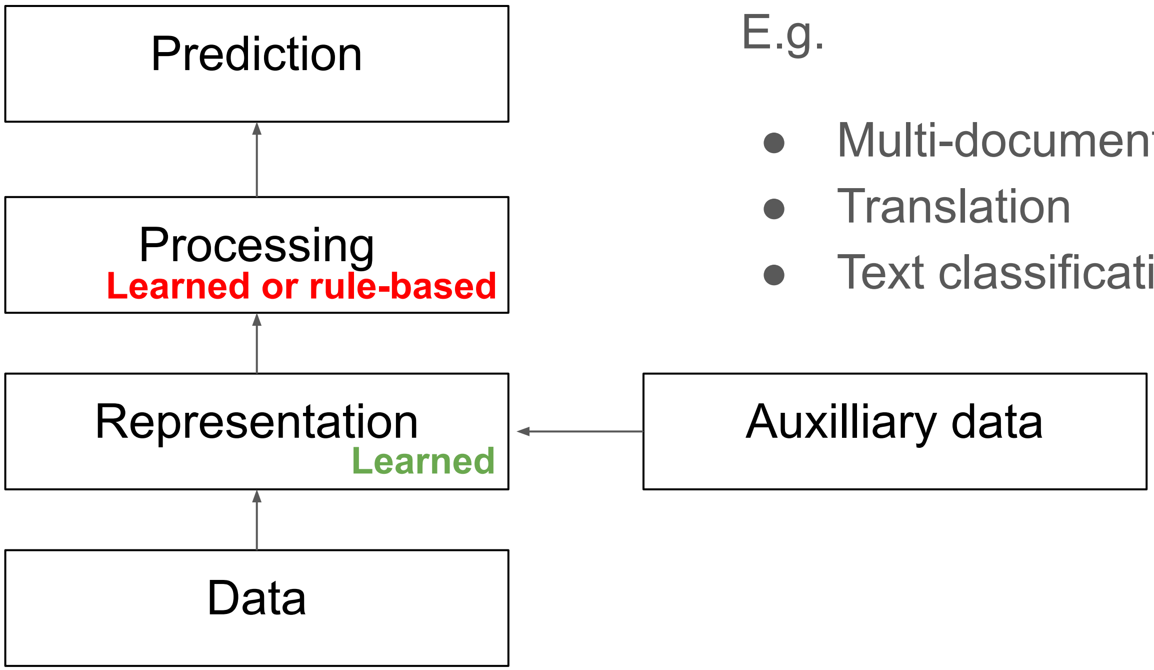
- ('Stockholm_Sweden', 0.78)
- ('Helsinki', 0.75)
- ('Oslo', 0.72)
- ('Oslo_Norway', 0.68)

Distributional hypothesis: words with similar meaning occur in similar contexts.

(Harris, 1954)

RI. SE

Word embeddings *was* transfer learning for language



E.g.

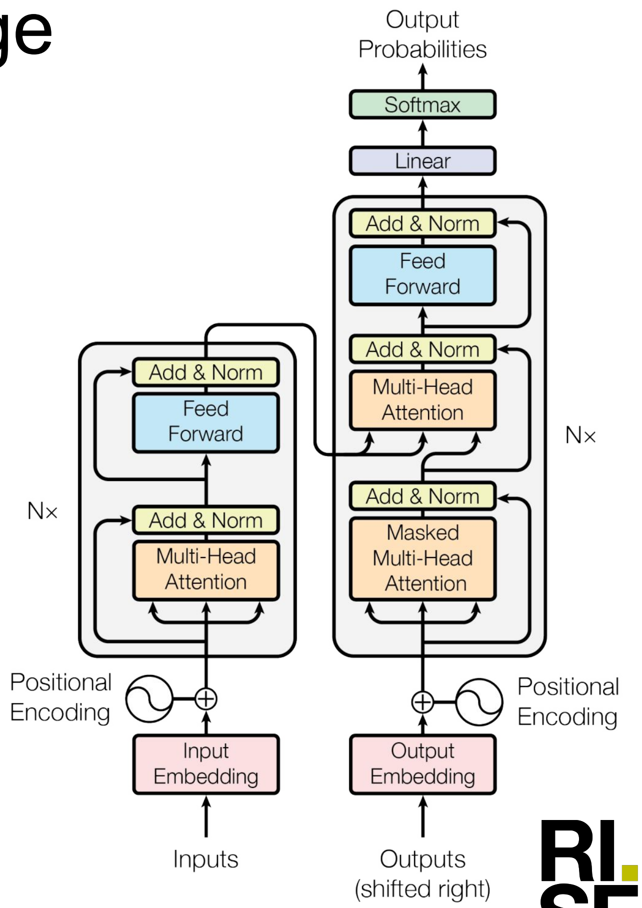
- Multi-document summarization (1)
- Translation
- Text classification

1. Kågebäck, **Mogren,** Tahmasebi, Dubhashi (2014)

Deep transfer learning for language

- Transformer (BERT)
- Trained using language modelling (word co-occurrences)
- Can compute word embedding that changes according to context

- “NLP’s Imagenet moment”: deep transfer learning for NLP, pretrain deep models.
- E.g. QA, Reading comprehension, Natural language inference, translation, constituency parsing, etc.



Vaswani, et.al. (2017), Devlin, et.al. (2018), Peters, et.al. (2018)

Man is to computer programmer as woman is to homemaker

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she*-*he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she*-*he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

gender bias in Word2vec

Bolukbasi, et.al., (NeurIPS 2016)

Brittleness in textual entailment

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **runner** wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

Gender bias

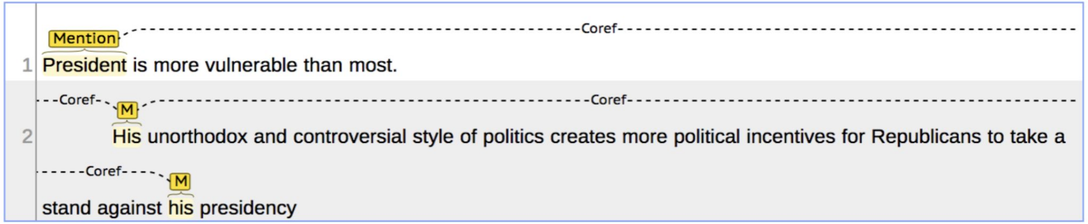
in language generation

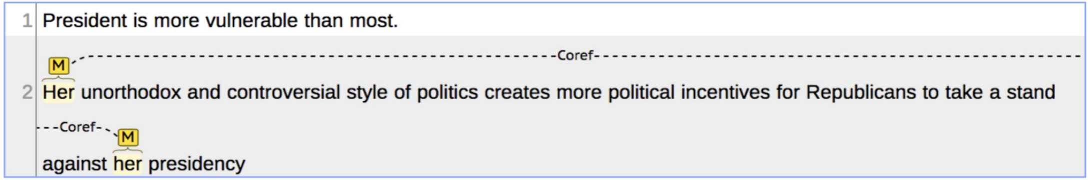
Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

GPT-2

Sheng, et.al. (EMNLP 2019) The Woman Worked as a Babysitter: On Biases in Language Generation

in coref resolution



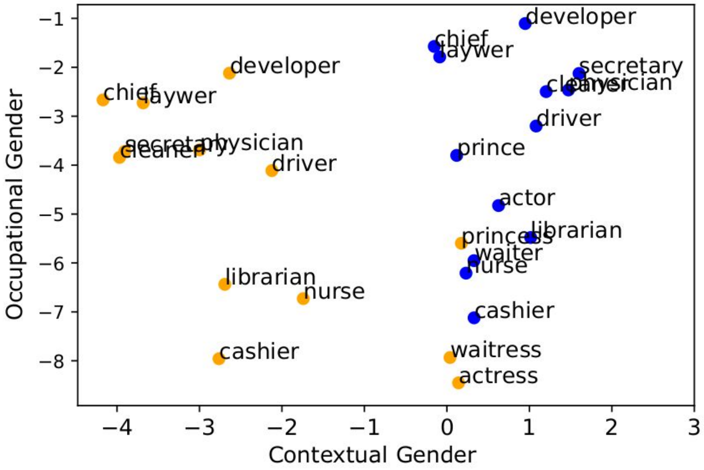


"WinoBias, WinoGender"

Zhao, et.al., Rudinger, et.al. (NAACL 2018)

Word gender vs contextual gender in ELMo

- ELMo embeddings
- Two principal components:
  - Word gender (occupational)
  - Contextual gender
- Pronoun color
  - Blue: male
  - Orange: female



Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.W. (NAACL 2019) Gender bias in contextualized word embeddings

Also in Swedish! Also in BERT!

- Gender-bias in Swedish pretrained embeddings
- Gender vs occupation
- Word2vec, FastText, ELMO, BERT

Name suggestion	Company description	Distance
Magnus bilar	Bolaget ska bedriva verksamhet med bilar	0.028
Fredriks bilar	Bolaget ska bedriva verksamhet med bilar	0.038
Marias bilar	Bolaget ska bedriva verksamhet med bilar	0.044
Annas bilar	Bolaget ska bedriva verksamhet med bilar	0.075

Sahlgren & Ohlsson (2019)

Human-like bias in Glove and Word2vec

- Insects and flowers (pleasantness)
- Musical instruments vs weapons (pleasantness)
- Racial bias: European-American names vs African-American names
- Gender and occupations
- Gender and arts vs sciences/mathematics

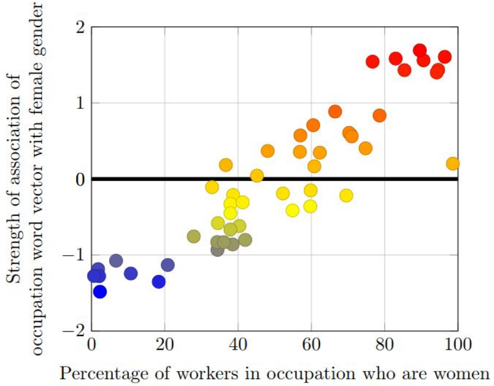


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

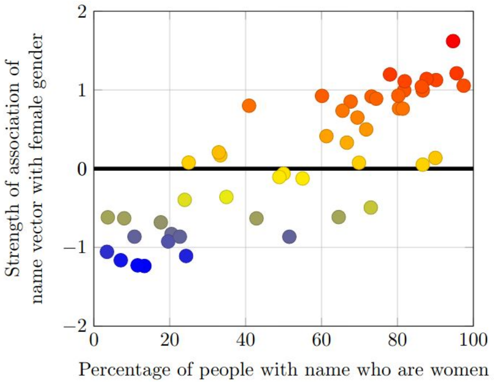
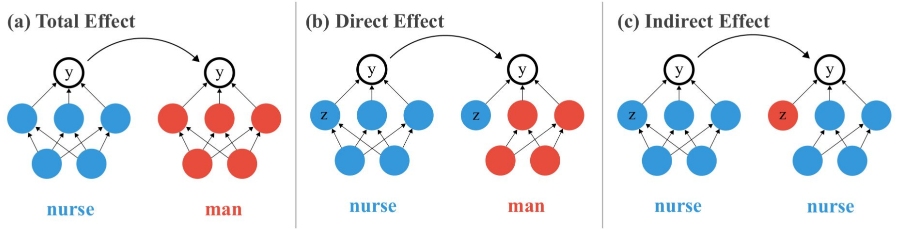


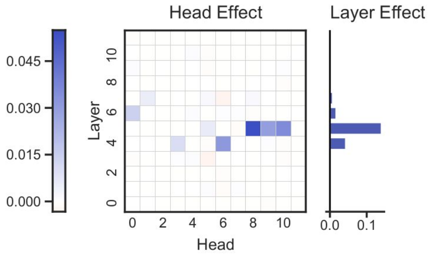
Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Caliskan, et.al. (2017)

Causal Mediation Analysis



- Transformer models
- Which parts are responsible for outputs
- Analyze flow
- Counterfactual interventions
  - Input interventions: `set-gender, null`
  - Neuron interventions
  - Attention interventions
- Gender bias in specialized components
- Direct/indirect effects
- Professions from Bolukbasi (2016)



$$\boldsymbol{y}(u) = \frac{p_\theta(\text{anti-stereotypical} \mid u)}{p_\theta(\text{stereotypical} \mid u)}.$$

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (NeurIPS 2020). Investigating gender bias in language models using causal mediation analysis.

Don’t we want the model to be “true” to the data?

All dimensions in an embedding may be desired

But social bias may be problematic for downstream applications eg:

- Resume filtering
- Insurange, lending, hiring
- Next word prediction on your phone
- Some systems may actually perform worse, cf. coreference resolution

We need to know what we are modelling, and how data can be used for this.

Social bias

- E.g. Gender bias, racial bias, etc.
- On what attributes can we base a decision?
- Is there information about that attribute in reps?
- How can we isolate them?

Fairness

- Is an individual treated fair in a decision? (Demographics, etc)

Privacy

- What attributes about myself do I share?

Disentanglement

- Attributes are often correlated
- Underlying factors

Generalization

- Learn distribution, not datapoints

How do we make models react to certain information but to be invariant of others?

RI. SE

Solutions

What is it that we want to model, and how do we go about it?

Data augmentation

- Train models using augmented data.
- he/she
- Anonymization of names

Calibration

- Identify sensitive dimensions
- Modify

Adversarial representation learning

- Train to make it difficult for adversary

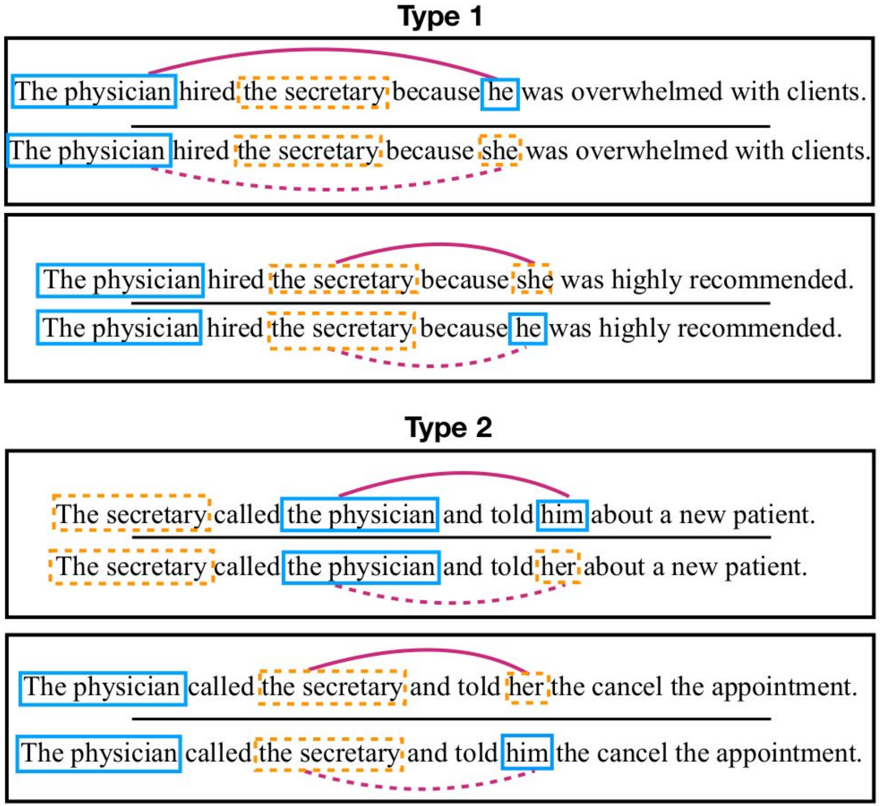
Data augmentation

“Anti-stereotypical” dataset.

Swap biased words, e.g.:

- he/she
- Anonymization of names

- Wino-bias dataset
- Wino-gender dataset

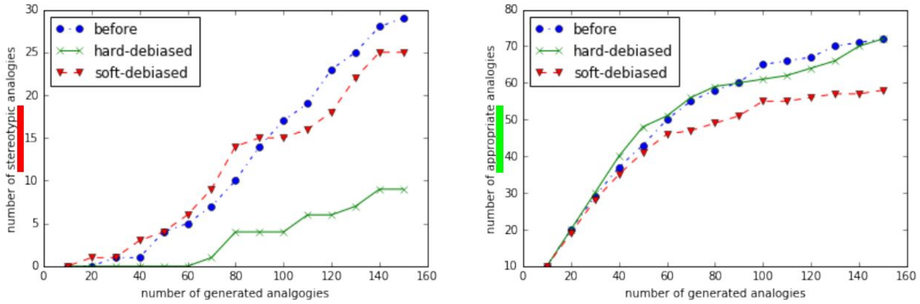


Zhao, et.al., NAACL 2018, *Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods*

Rudinger, et.al., NAACL 2018, *Gender Bias in Coreference Resolution, https://www.aclweb.org/anthology/N18-2002*

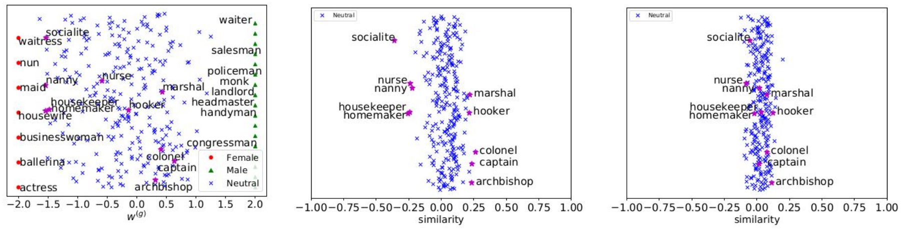
Calibration

1. Identify “appropriate” gendered words (e.g. *grandfather-grandmother, guy-gal)*
2. Train model to identify these words
3. Identify gender direction
4. Modify vectors
   a. Neutral words: zero gender direction(s)
   b. Acceptable gender words: equidistant to neutral words in gender direction(s)



Bolukbasi, et.al. (NeurIPS 2016)

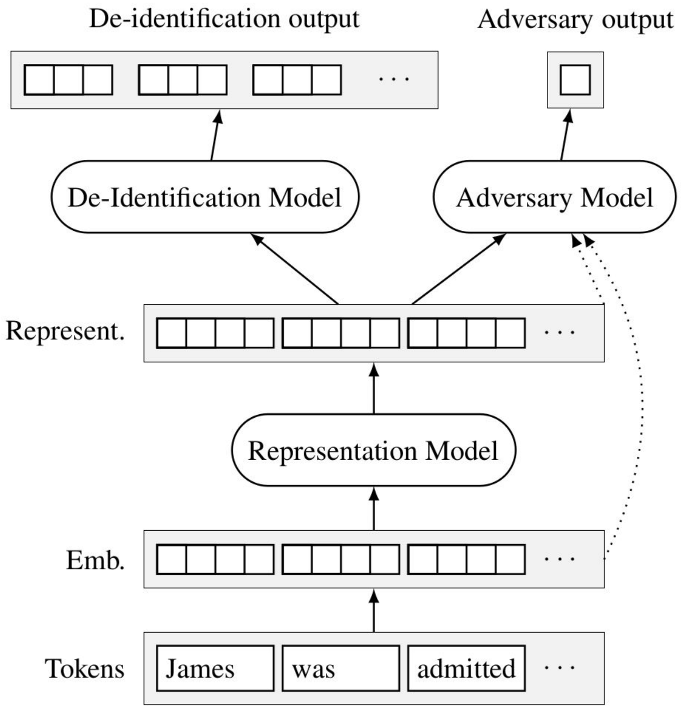
- Restrict sensitive attributes to specific dimensions of embedding
- Minimize distance between words in the two groups in other dimensions



Zhao, et.al. (EMNLP 2018)

Adversarial representation learning for language

- Adversary: detect privacy leakage in embeddings
- Embeddings: fool adversary
- Privacy preserving embeddings
- (Requires data augmentation)

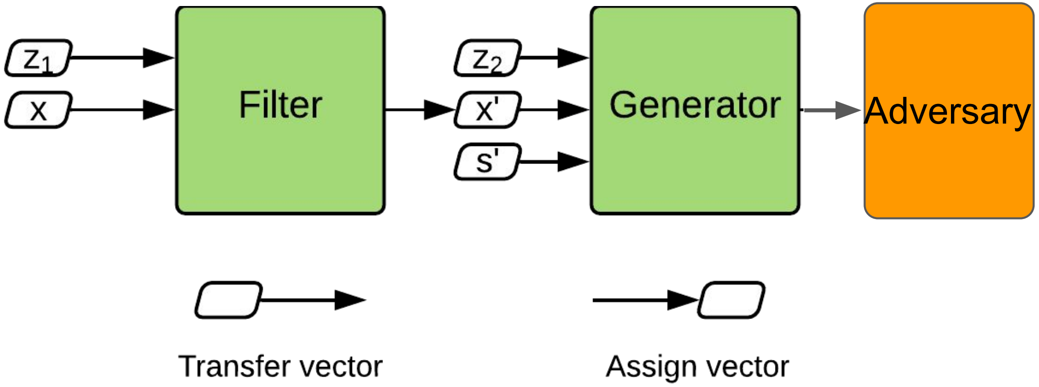


Zhang, et.al., (AIES 2018), Friedrich, et.al. (ACL 2019),

Privacy in machine learning

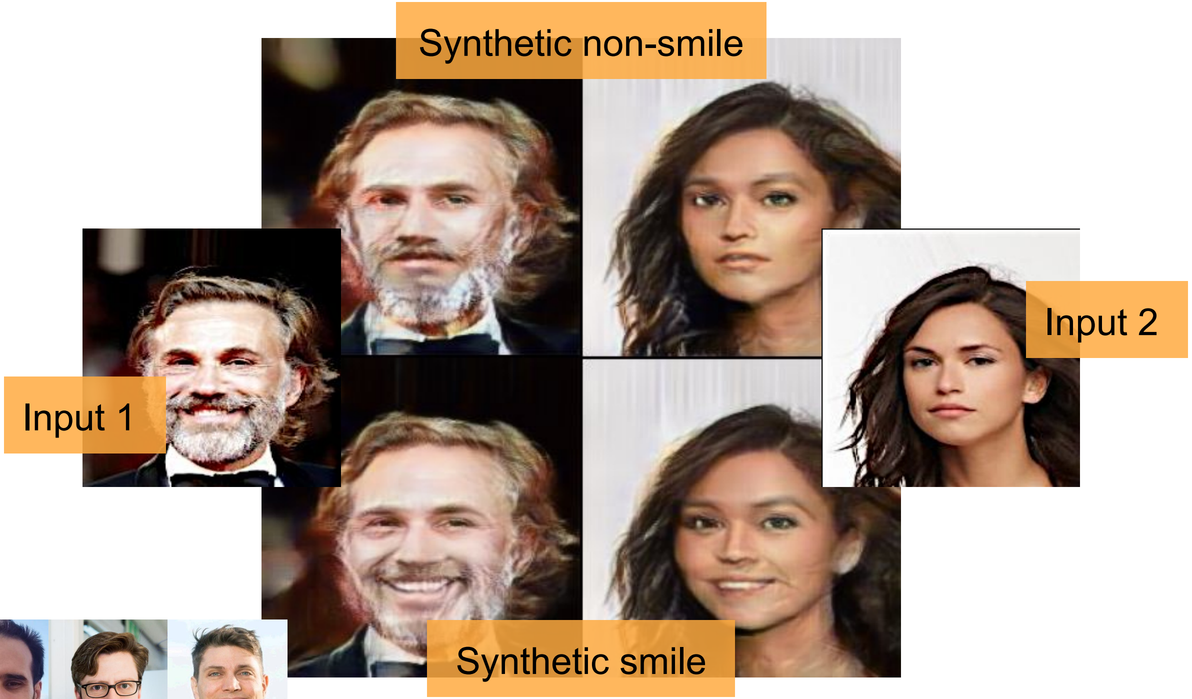
Adversarial representation learning for privacy

- Dataset privacy: sensitive features
- Privatization mechanisms (obfuscation)
- Privacy preserving machine learning
- Adversarial representation learning for
  - Removing sensitive attributes
  - Synthetize attribute values independent from input



*Martinsson, J., Listo Zec, E., Gillblad, D., **Mogren, O.** Adversarial representation learning for synthetic replacement of private attributes. https://arxiv.org/abs/2006.08039, 2020.*

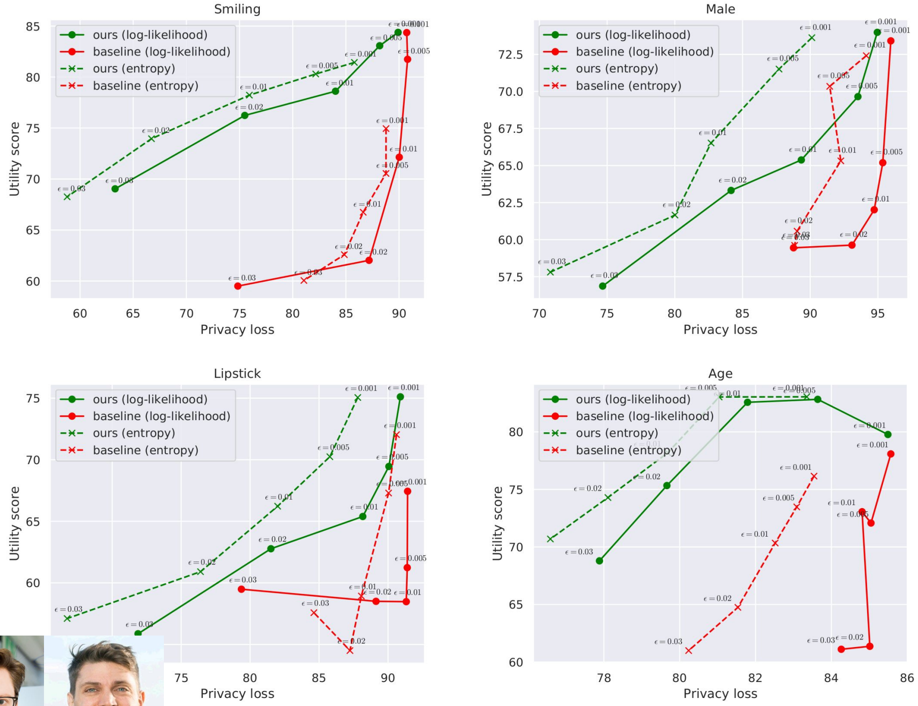
Adversarial representation learning for privacy



Martinsson, J., Listo Zec, E., Gillblad, D., **Mogren, O.** *Adversarial representation learning for synthetic replacement of private attributes.* https://arxiv.org/abs/2006.08039, 2020.

Adversarial representation learning for privacy

- Paper under review
- Future work: language!

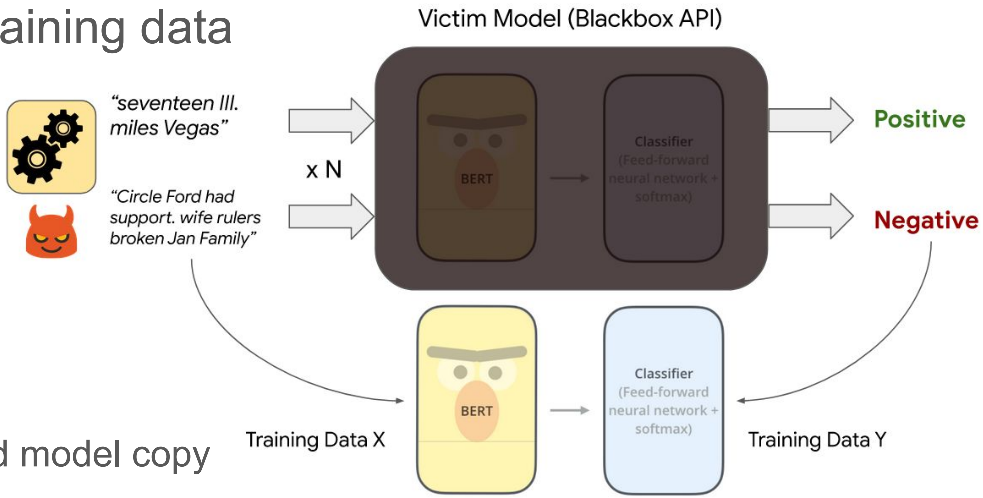


*Martinsson, J., Listo Zec, E., Gillblad, D., **Mogren, O.** Adversarial representation learning for synthetic replacement of private attributes. https://arxiv.org/abs/2006.08039, 2020.*

Model extraction

- Secret model, deployed with open API
- Knowledge distillation
- Queries: random sequences of words
- May leak sensitive information from training data

- Membership classification
  - Determine nonsensical inputs
  - Respond with “no answer”
- API watermarking
  - Inject faulty data
  - Faulty predictions will cascade to extracted model copy



Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M. (ICLR 2020), Thieves on Sesame Street! Model Extraction of BERT-based APIs, https://arxiv.org/abs/1910.12366

Swedish medical language data lab

- Prescription of antibiotics in dental care
- Addiction clinic readmissions
- Adverse events in health care

- Data access
- Privacy
- Domain-specific language
  - MD language
  - Nurse language
  - Dental language

FOLKTANDVÅRDEN

BÄSTA LIVSPLATSEN
Region Halland

SAHLGRENSKA
UNIVERSITETSSJUKHUSET

PELTARION

VINNOVA

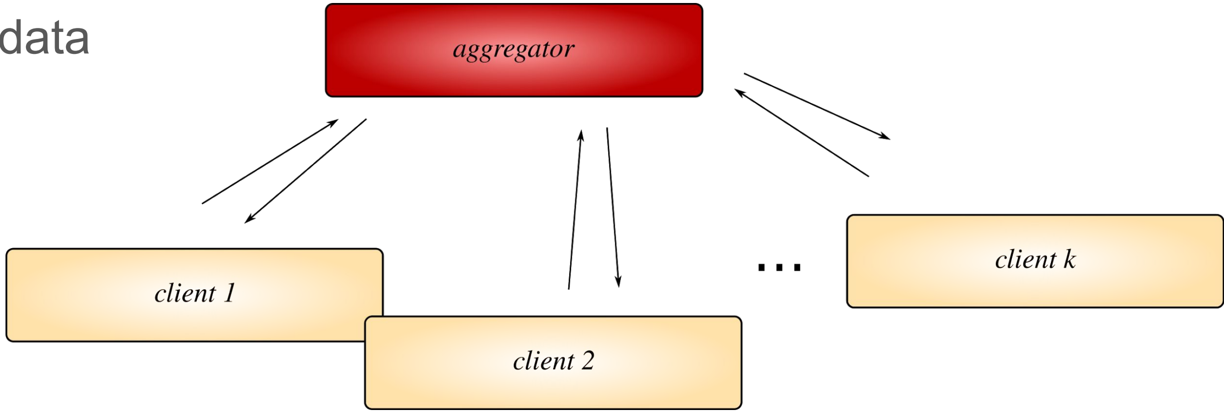
SAHLGRENSKA
SCIENCE PARK

AI SWEDEN

RI. SE

Federated learning

- Train a local model on each client
- Send updates/gradients to central server
- Allows training data to remain on the clients
- Benefits from additional data
- A certain level of privacy



Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1310-1321).

Gradients can reveal training data (batch size 1)

Validation image. Reconstruction, Resnet-18. Reconstruction, Resnet-152.

Geiping, Bauermeister, Dröge, Moeller (NeurIPS 2020) Inverting Gradients - How easy is it to break privacy in federated learning? https://arxiv.org/abs/2003.14053

Gradients can reveal training data (batch size 100)

Validation data

Reconstructions

Figure 6: Information leakage for a batch of 100 images on CIFAR-100 for a ResNet32-10. Shown are the 5 *most* recognizable images from the whole batch. Although most images are unrecognizable, privacy is broken even in a large-batch setting. We refer to the supplementary material for all images.

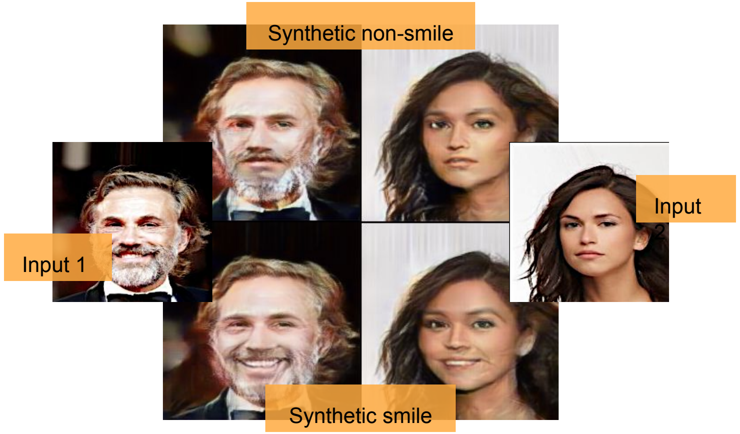
Geiping, Bauermeister, Dröge, Moeller (NeurIPS 2020) Inverting Gradients - How easy is it to break privacy in federated learning? https://arxiv.org/abs/2003.14053

Comments

- Optimization methods from adversarial attacks
- Labels considered to be known
- Cherry-picked examples
- Larger batch size makes attack harder
- Deeper network makes attack harder

Geiping, Bauermeister, Dröge, Moeller (NeurIPS 2020) Inverting Gradients - How easy is it to break privacy in federated learning? https://arxiv.org/abs/2003.14053

Solutions

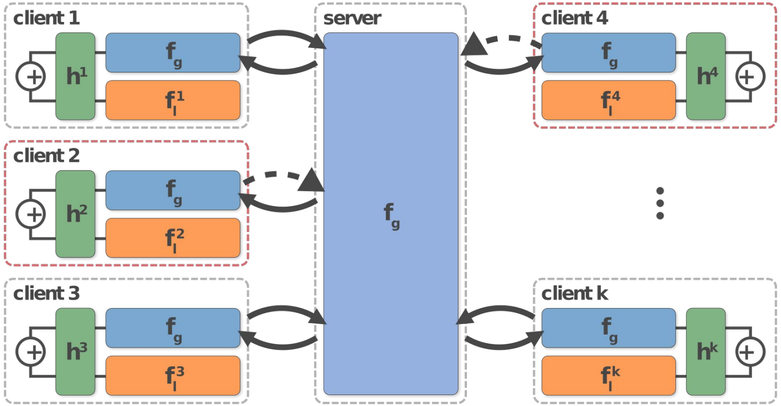


- Trust all partners (clients och server)
- Encrypt communication
- Differential privacy gives bounds on privacy
- Semi-federated learning with mixture of experts [1]
- Privacy-ensuring transformations (trade-off) [2]
- Masked/obfuscated training data (Netflix competition)

1. *Listo Zec, E., **Mogren, O.,** Martinsson, J., Sütfeld, L.R., Gillblad, D. (2020) Federated learning using a mixture of experts.* https://arxiv.org/abs/2010.02056
2. *Martinsson, Listo Zec, Gillblad, **Mogren** (2020) Adversarial representation learning for synthetic replacement of private attributes.* https://arxiv.org/abs/2006.08039.

Semi-federated learning using mixture of experts

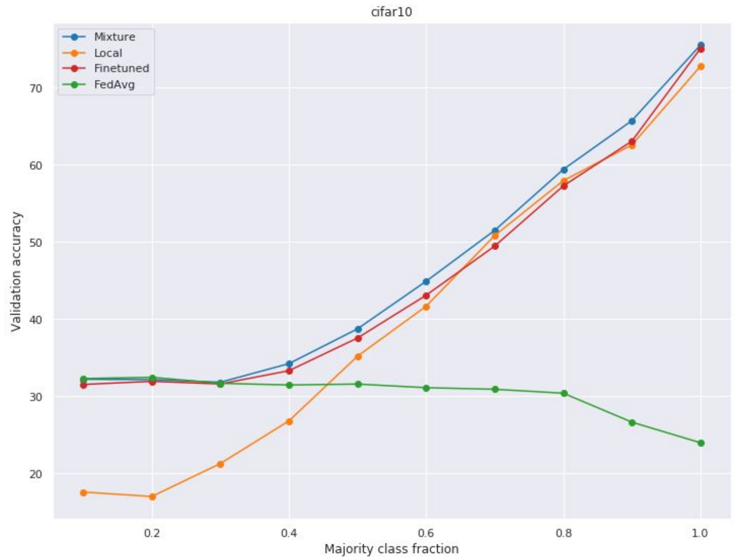
- Proposed framework for FL
- Federated learning using a mixture of experts.
- Balance general and special knowledge
- Privacy guarantees
  - D_O: Opt-out data
  - D_I: Opt-in data
  - Can be combined with our privacy mechanisms
- State-of-the-art results in non-i.i.d. settings
- Ongoing experiments: AGNews
- Paper under review



Listo Zec, E., **Mogren, O.,** Martinsson, J., Sütfeld, L.R., Gillblad, D. (2020) Federated learning using a mixture of experts. https://arxiv.org/abs/2010.02056

Semi-federated learning using mixture of experts (2)

- Proposed framework for FL
- Federated learning using a mixture of expe
- Balance general and special knowledge
- Privacy guarantees
  - D_O: Opt-out data
  - D_I: Opt-in data
  - Can be combined with our privacy mechanisms
- State-of-the-art results in non-i.i.d. settings
- Ongoing experiments: AGNews
- Paper under review

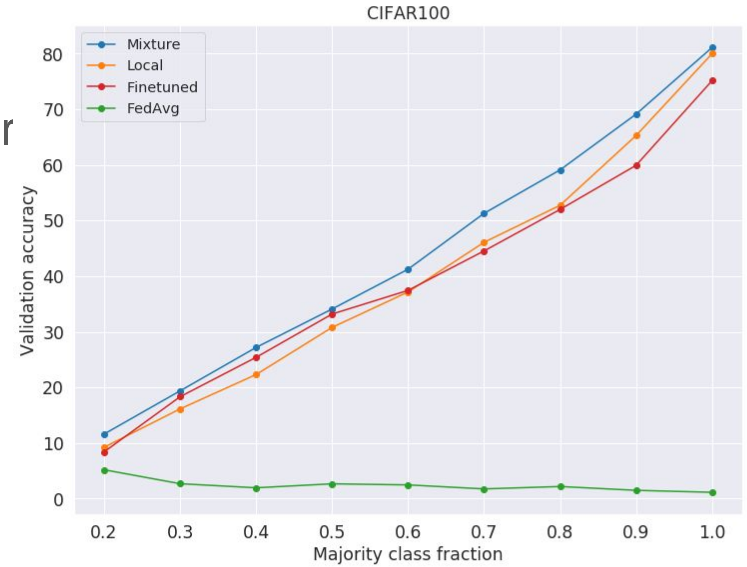


Dataset size: 100 datapoints per client.

*Listo Zec, E., **Mogren, O.,** Martinsson, J., Sütfeld, L.R., Gillblad, D. (2020) Federated learning using a mixture of experts. https://arxiv.org/abs/2010.02056*

Semi-federated learning using mixture of experts (3)

- Proposed framework for FL
- Federated learning using a mixture of exper
- Balance general and special knowledge
- Privacy guarantees
  - D_O: Opt-out data
  - D_I: Opt-in data
  - Can be combined with our privacy mechanisms
- State-of-the-art results in non-i.i.d. settings
- Ongoing experiments: AGNews
- Paper under review



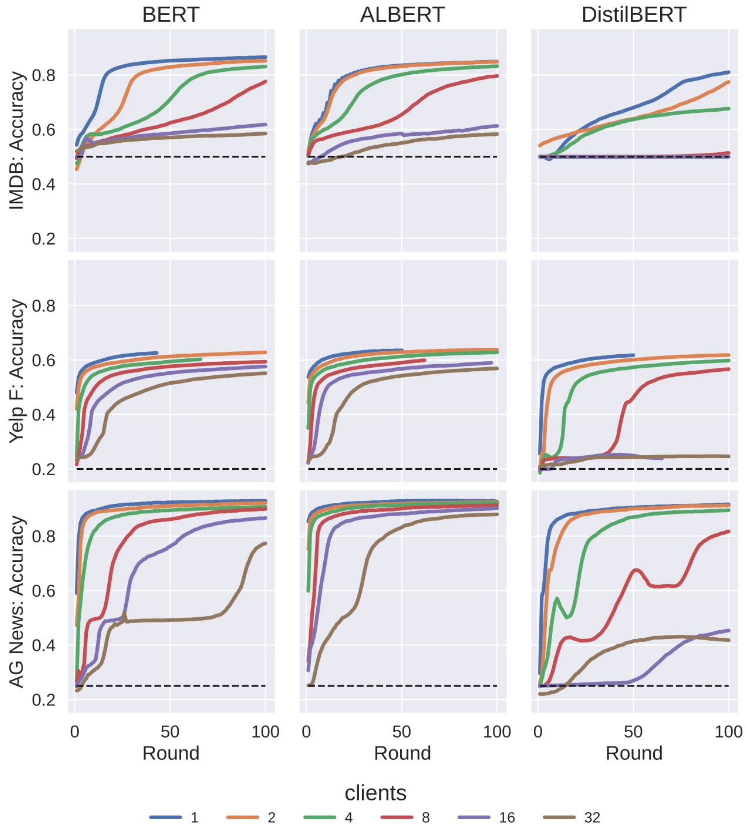
Dataset size: 100 datapoints per client.

*Listo Zec, E., **Mogren, O.,** Martinsson, J., Sütfeld, L.R., Gillblad, D. (2020) Federated learning using a mixture of experts. https://arxiv.org/abs/2010.02056*

RISE

Federated Transformers

- Collaboration with Peltarion
- Survey on current (large) language models in federated setting
- Conclusions
  - Learning is feasible
  - Some models suffer from increased client count
- Future work: investigate interaction between knowledge distillation and FL



Hilmkil, A., Callh, S., Barbieri, M., Sütfeld, L.R., Listo Zec, E., and ***Mogren, O.*** *(2020) Scaling Federated Learning for Fine-tuning of Large Language Models*

Thank you

Olof Mogren, PhD

RISE Research Institutes of Sweden

olof.mogren@ri.se

Team and collaborators:

References (1 of 2)

Kiela & Bottou, EMNLP 2014, Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

Kågebäck, Mogren, Tahmasebi, Dubhashi, 2014, Extractive summarization using continuous vector space models, https://www.aclweb.org/anthology/W14-1

Bolukbasi, et.al., NeurIPS 2016, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Caliskan, A., Bryson, J.J., and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):1

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang (EMNLP 2017) Men Also Like Shopping: Reducing Gender Bias Amplification usin Corpus-level Constraints

Zhao, et.al, EMNLP 2018, Learning Gender-Neutral Word Embeddings, https://arxiv.org/pdf/1809.01496

Sahlgren & Ohlsson, 2018, Gender Bias in Pretrained Swedish Embeddings, https://www.aclweb.org/anthology/W19-6104.pdf

Zhao, et.al., NAACL 2018, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, https://www.aclweb.org/anthology/N18-2003/

Rudinger, et.al., NAACL 2018, Gender Bias in Coreference Resolution, https://www.aclweb.org/anthology/N18-2002

Zhang, et.al., AIES 2018, Mitigating Unwanted Biases with Adversarial Learning

Sato, et.al., ACL 2019, Effective Adversarial Regularization for Neural Machine Translation

Wang, et.al., ICML 2019, Improving Neural Language Modeling via Adversarial Training, https://arxiv.org/pdf/1906.03805

Sheng, Chang, Natarajan, Peng (EMNLP 2019) The Woman Worked as a Babysitter: On Biases in Language Generation, https://www.aclweb.org/anthology

Friedrich, M., Köhn, A., Wiedemann, G., & Biemann, C. (ACL 2019). Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Records. arXiv preprint arXiv:1906.05000

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (NAACL 2019) Gender bias in contextualized word embedding https://www.aclweb.org/anthology/N19-1064/

Yi Chern Tan and L. Elisa Celis. (NeurIPS 2019) Assessing social and intersectional biases in contextualized word representations https://papers.nips.cc/paper/9479-assessing-social-and-intersectional-biases-in-contextualized-word-representations

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (NeurIPS 2020). Investigating gender bias in language models using causal analysis. https://papers.nips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html

References (2 of 2)

Shokri, R., & Shmatikov, V. (2015, October). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1310-1321). http://www.cs.cornell.edu/~shmat/shmat_ccs15.pdf

Shokri, et.al., Membership inference attacks against machine learning models, https://arxiv.org/abs/1610.05820

Wang et.al., Beyond inferring class representatives: User-level privacy leakage from federated learning, https://arxiv.org/abs/1812.00535

Truex, et.al., A Hybrid Approach to Privacy-Preserving Federated Learning, https://arxiv.org/abs/1812.03224

Bagdasaryan, et.al., How To Backdoor Federated Learning, http://proceedings.mlr.press/v108/bagdasaryan20a.html

Stealing Machine Learning Models via Prediction APIs, USENIX Security, 2016., https://arxiv.org/pdf/1609.02943.pdf

Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M. (ICLR 2020), Thieves on Sesame Street! Model Extraction of BERT-based APIs, https://arxiv.org/abs/1910.12366

Geiping, Bauermeister, Dröge, Moeller (2020) Inverting Gradients - How easy is it to break privacy in federated learning? https://arxiv.org/abs/2003.14053

Martinsson, J., Listo Zec, E., Gillblad, D., Mogren, O., (2020), Adversarial representation learning for synthetic replacement of private attributes. https://arxiv.org/abs/2006.08039.