Adaptive and Robust (Deep) Learning with Bayes

Emtiyaz Khan
RIKEN-AIP (Japan)

Dharmesh Tailor
RIKEN-AIP (Japan) [2]

Siddharth Swaroop
Cambridge University (UK)

1. Slides at https://emtiyaz.github.io/papers/Dec14_2021_NeurIPS_BDL.pdf
2. Presenting work done at RIKEN, current affiliation at University of Amsterdam, Netherland

AI that learns as quickly as humans and animals

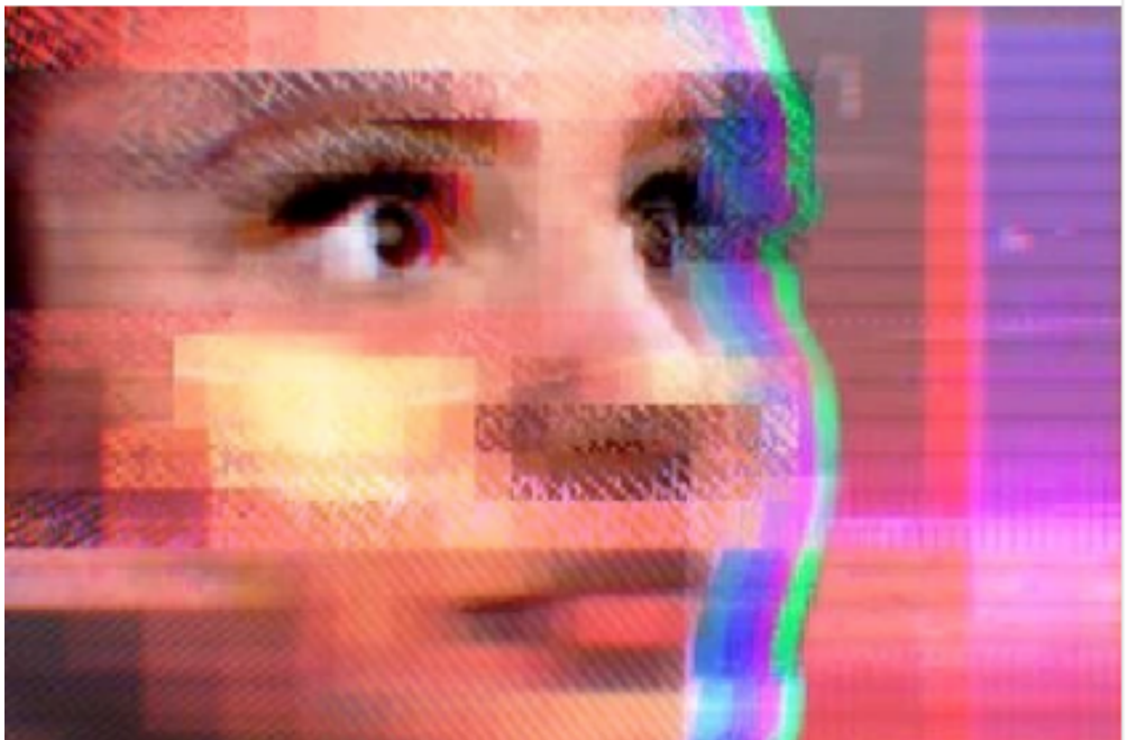
Quickly adapt to new situations in the future

by robustly preserving & using past knowledge

Fail because too quick to adapt

TayTweets: Microsoft AI bot manipulated into being extreme racist upon release

Posted Fri 25 Mar 2016 at 4:38am, updated Fri 25 Mar 2016 at 9:17am



TayTweets is programmed to converse like a teenage girl who has "zero chill", according to Microsoft. (Twitter: TayTweets)

https://www.abc.net.au/news/2016-03-25/microsoft-created-ai-bot-becomes-racist/7276266

Fail because too slow to adapt

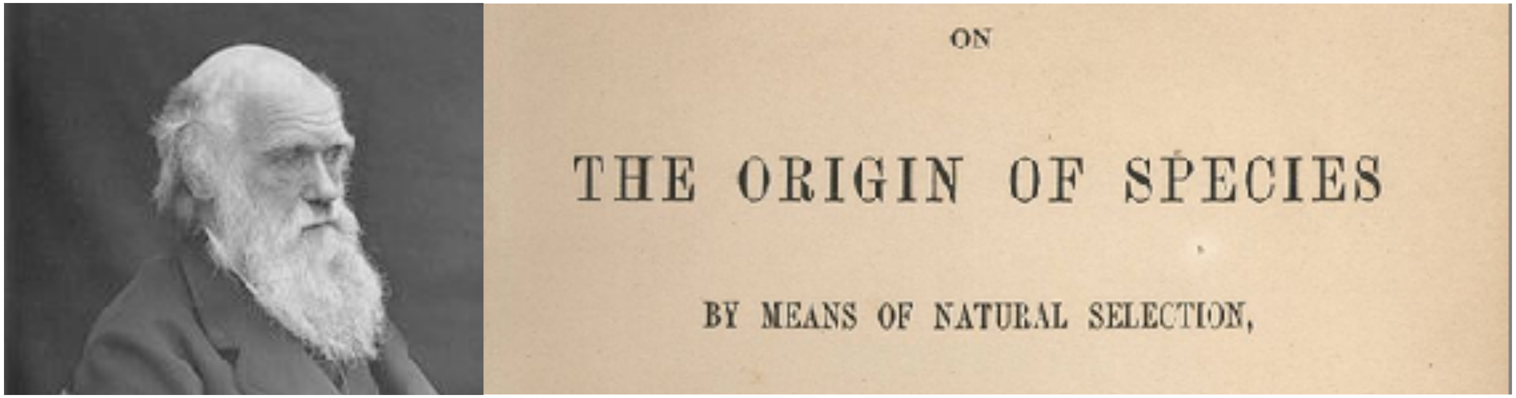
https://www.youtube.com/watch?v=TxobtWAFh8o The video is from 2017

Adaptive & Robust Learning with Bayes

- "Good" algorithms are inherently Bayesian
- Bayesian learning rule [1]
  - Presented by Emti
- Robustness: Memorable experiences [2]
  - presented by Dharmesh
- Adaptation: Knowledge-Adaptation Priors [3,4,5]
  - presented by Siddharth
- Take away: A new perspective of Bayes, essential for adaptive and robust deep learning

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Tailor, Chang, Swaroop, Tangkaratt, Solin, Khan. Memorable experiences of ML models (in preparation)
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
4. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
5. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)

See Section 6 (discussion) in Khan and Rue, 2021



The Origin of Algorithms

A good algorithm must revise its *past* beliefs by using useful *future* information

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021

A Bayesian Origin

$$\min_{\theta} \; \ell(\theta) \qquad \text{vs} \qquad \min_{q \in \mathcal{Q}} \; \mathbb{E}_{q(\theta)}[\ell(\theta)] - \mathcal{H}(q)$$

Entropy

Posterior approximation (expo-family)

Bayesian Learning Rule [1,2]

Natural and Expectation parameters of q

$$\lambda \leftarrow (1-\rho)\underline{\lambda} - \underline{\rho \nabla_{\mu} \mathbb{E}_q[\ell(\theta)]}$$

Old belief

Revise using new information through natural gradients

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Khan and Lin. "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models." AIstats (2017).

Bayesian learning rule: $\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_{\mu} \mathbb{E}_q[\ell(\theta)]$

Learning Algorithm	Posterior Approx.	Natural-Gradient Approx.	Sec.
Optimization Algorithms			
Gradient Descent	Gaussian (fixed cov.)	Delta method	1.3
Newton's method	Gaussian	——"——	1.3
Multimodal optimization (New)	Mixture of Gaussians	——"——	3.2
Deep-Learning Algorithms			
Stochastic Gradient Descent	Gaussian (fixed cov.)	Delta method, stochastic approx.	4.1
RMSprop/Adam	Gaussian (diagonal cov.)	Delta method, stochastic approx., Hessian approx., square-root scaling, slow-moving scale vectors	4.2
Dropout	Mixture of Gaussians	Delta method, stochastic approx., responsibility approx.	4.3
STE	Bernoulli	Delta method, stochastic approx.	4.5
Online Gauss-Newton (OGN) (New)	Gaussian (diagonal cov.)	Gauss-Newton Hessian approx. in Adam & no square-root scaling	4.4
Variational OGN (New)	——"——	Remove delta method from OGN	4.4
BayesBiNN (New)	Bernoulli	Remove delta method from STE	4.5
Approximate Bayesian Inference Algorithms			
Conjugate Bayes	Exp-family	Set learning rate $\rho_t = 1$	5.1
Laplace's method	Gaussian	Delta method	4.4
Expectation-Maximization	Exp-Family + Gaussian	Delta method for the parameters	5.2
Stochastic VI (SVI)	Exp-family (mean-field)	Stochastic approx., local $\rho_t = 1$	5.3
VMP	——"——	$\rho_t = 1$ for all nodes	5.3
Non-Conjugate VMP	——"——	——"——	5.3
Non-Conjugate VI (New)	Mixture of Exp-family	None	5.4

The BLR variants [1,2,3] led to the winning solution for the NeurIPS 2021 challenge for "approximate inference in BDL" (Watch Thomas Moellenhoff's talk)

1. Khan, et al. "Fast and scalable Bayesian deep learning by weight-perturbation in Adam." *ICML* (2018).
2. Osawa et al. "Practical Deep Learning with Bayesian Principles." NeurIPS (2019).
3. Lin et al. "Handling the positive-definite constraints in the BLR." ICML (2020).

Robustness

Good algorithms can tell apart
relevant vs irrelevant information

Perturbation, Sensitivity, and Duality

https://tenor.com/view/clockwork-gears-brain-gif-16784329

See Section 5.4 in Khan and Rue, 2021 for local parameterization
See Section 3 in ADAM et al. 2021 for dual parameterization

BLR Solutions & Their Duality

$$\ell(\theta) = \sum_{i=0}^{N} \ell_i(\theta) \qquad \lambda \leftarrow (1-\rho)\lambda - \sum_{i=0}^{N} \rho \nabla_{\mu} \mathbb{E}_q[\ell_i(\theta)]$$

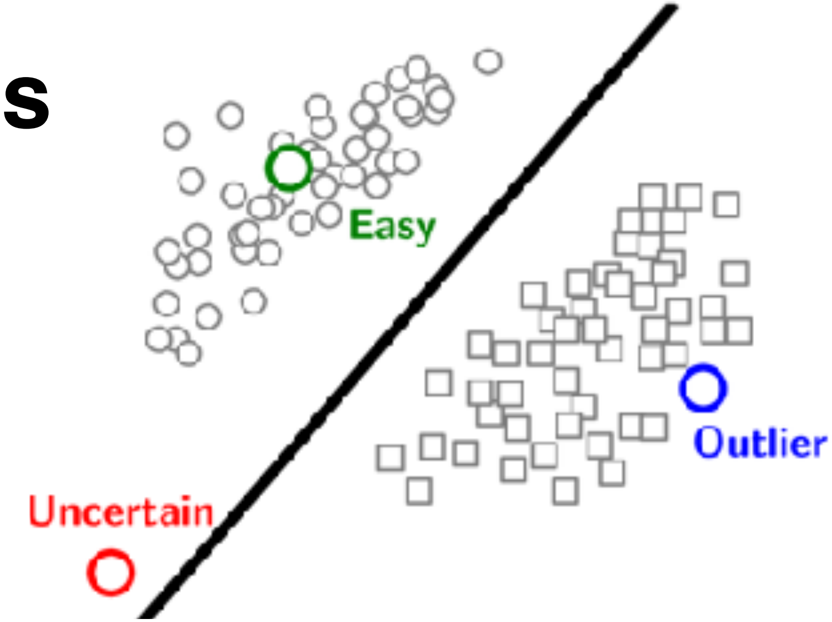
$$\lambda^* = \sum_{i=0}^{N} \underbrace{\nabla_{\mu^*} \mathbb{E}_{q^*}[-\ell_i(\theta)]}_{\widetilde{\lambda}_i^*}$$

Global and local natural parameter

Local parameters are Lagrange Multipliers, measuring the sensitivity of BLR solutions to local perturbation [1]. They can be used to tell apart relevant vs irrelevant data.

1. ADAM, Chang, Khan, Solin, Dual parameterization of SVGP, NeurIPS, 2021

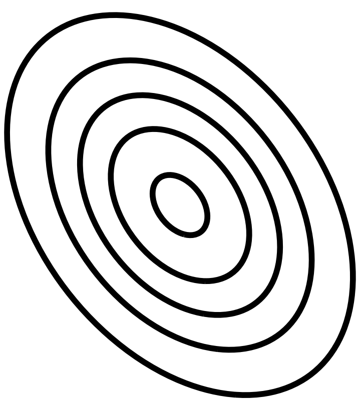
Memorable Experiences

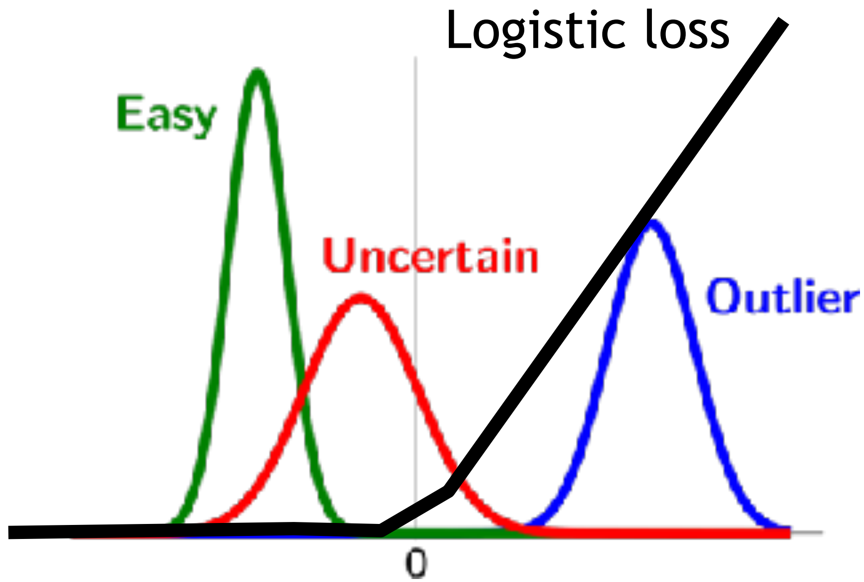


$$\lambda^* = \sum_{i=0}^{N} \underbrace{\nabla_{\mu^*}\mathbb{E}_{q^*}[-\ell_i(\theta)]}_{\widetilde{\lambda}_i^*}$$

"Global" posterior $q(\theta)$

Local predictions $q(f_i)$





Lower Sensitivity to easy example.

Such sensitivity analysis leads to memorable experiences

Memorable Experiences

1. Schneider et al. DeepOBS: A Deep Learning Optimizer Benchmark Suite”. ICLR 2018

Advantages of Memorable Experiences

- Through posterior approximations, the criteria to categorize examples naturally emerges
  - Generalizes existing concepts such as support vectors, influence functions, inducing inputs etc
- Local parameters are available for free and applies to almost “any” ML problem
  - Supervised, unsupervised, RL
  - Discrete/continuation loss and model parameters
- The sensitivity of posterior leads to “Bayes Duality”

1. Tailor, Chang, Swaroop, Tangkaratt, Solin, Khan. Memorable experiences of ML models (in preparation)

The webpage is available at https://bayesduality.github.io/, and Twitter account @BayesDuality

The Bayes-Duality Project

Toward AI that learns adaptively, robustly, and continuously, like humans

Emtiyaz Khan

Research director
(Japan side)

Approx-Bayes team at
RIKEN-AIP and OIST

Julyan Arbel

Research director
(France side)

Statify-team, Inria
Grenoble Rhône-Alpes

Kenichi Bannai

Co-PI (Japan side)

Math-Science Team at
RIKEN-AIP and Keio
University

Rio Yokota

Co-PI
(Japan side)

Tokyo Institute of
Technology

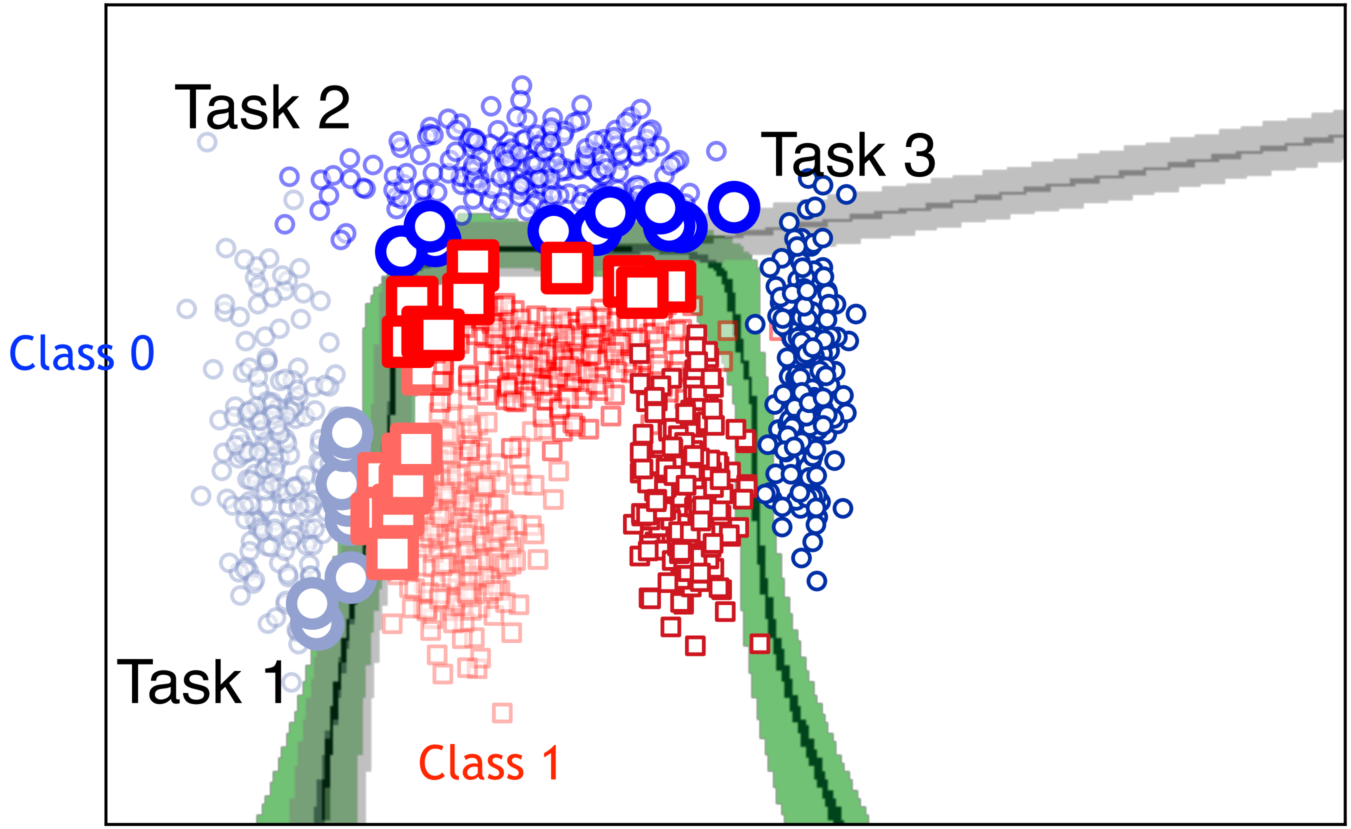
Received total funding of around USD 3 million through JST's CREST-ANR and Kakenhi Grants.

Adaptation

Continual Learning without forgetting the past (by using memorable examples)

Continual Learning

Avoid forgetting by using memorable examples [1,2]



1. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
2. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

Functional Regularization of Memorable Past (FROMP) [3]

Previous approaches used weight-regularization [1]

$$q_{new}(\theta) = \min_{q \in \mathcal{Q}} \mathbb{E}_{q(\theta)}[\ell_{new}(\theta)] - \mathcal{H}(q) - \mathbb{E}_{q(\theta)}[\log q_{old}(\theta)]$$

New data

Weight-regularizer using old posterior

We replace it by a functional regularizer using a "Gaussian Process view" of DNNs [2]

$$\mathbb{E}_{\tilde{q}_\theta(\mathbf{f})}[\log \tilde{q}_{\theta_{old}}(\mathbf{f})]$$

$$[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]^\top K_{old}^{-1}[\sigma(\mathbf{f}(\theta)) - \sigma(\mathbf{f}_{old})]$$

Kernels weighs examples according to their memorability

Forces network-outputs to be similar

1. Nguyen et al., Variational Continual Learning, ICLR, 2018
2. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
3. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020

See Section 3 and 4, and App A in [2]

K-Priors and Bayes-Duality

- Dual parameterization of DNNs
  - expressed as Gaussian Process [1]
  - Found using the Bayesian learning rule
- The functional regularizer can provably reconstruct the gradient of the past faithfully [2]
  - Knowledge-Adaptation priors (K-priors)
  - There is a strong evidence that "good" adaptive algorithms must use K-priors

1. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
2. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)

Summary

- A new perspective of Bayes, essential for adaptive and robust deep learning
- Approximate posteriors are crucial
  - Bayesian learning rule [1]
  - Robustness: Memorable experiences [2]
  - Adaptation: K-Priors [3,4,5]
- Bayes-duality for AI that learns like humans

1. Khan and Rue, The Bayesian Learning Rule, arXiv, https://arxiv.org/abs/2107.04562, 2021
2. Tailor, Chang, Swaroop, Tangkaratt, Solin, Khan. Memorable experiences of ML models (in preparation)
3. Khan et al. Approximate Inference Turns Deep Networks into Gaussian Process, NeurIPS, 2019
4. Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past, NeurIPS, 2020
5. Khan and Swaroop. Knowledge-Adaptation Priors, NeurIPS, 2021 (https://arxiv.org/abs/2106.08769)

Approximate Bayesian Inference Team

https://team-approx-bayes.github.io/

We have many open positions!
Come, join us.

Emtiyaz Khan
Team Leader

Pierre Alquier
Research Scientist

Gian Maria Marconi
Postdoc

Thomas Möllenhoff
Postdoc

Lu Xu
Postdoc

Jooyeon Kim
Postdoc

Wu Lin
PhD Student
University of British Columbia

Ted Tinker
PhD Student
Okinawa Institute of Science and Technology

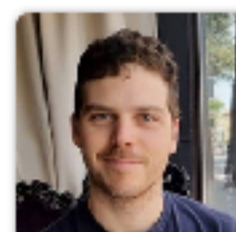
Peter Nickl
Research Assistant

Happy Buzaaba
Part-time Student
University of Tsukuba

Siddharth Swaroop
Remote Collaborator
University of Cambridge

Dharmesh Tailor
Remote Collaborator
University of Amsterdam

Erik Daxberger
Remote Collaborator
University of Cambridge



Alexandre Piché
Remote Collaborator
MILA

Paul Chang
Remote Collaborator
Aalto University