推进机器学习发展

—— 技术前沿与未来展望

刘铁岩
微软亚洲研究院 副院长
IEEE会士， ACM杰出科学家

Microsoft

微软在美国本土以外规模最大的研究机构

6 大研究方向

推动整个计算机科学领域的前沿技术发展

6,000+ 实习生

院友 7,000+

Microsoft Research Asia

微软亚洲研究院

5,000+ 论文发表

50+ 最佳论文

教育部最佳合作伙伴

着眼革命性技术的研究，帮助传统企业实现智能化转型

ABC

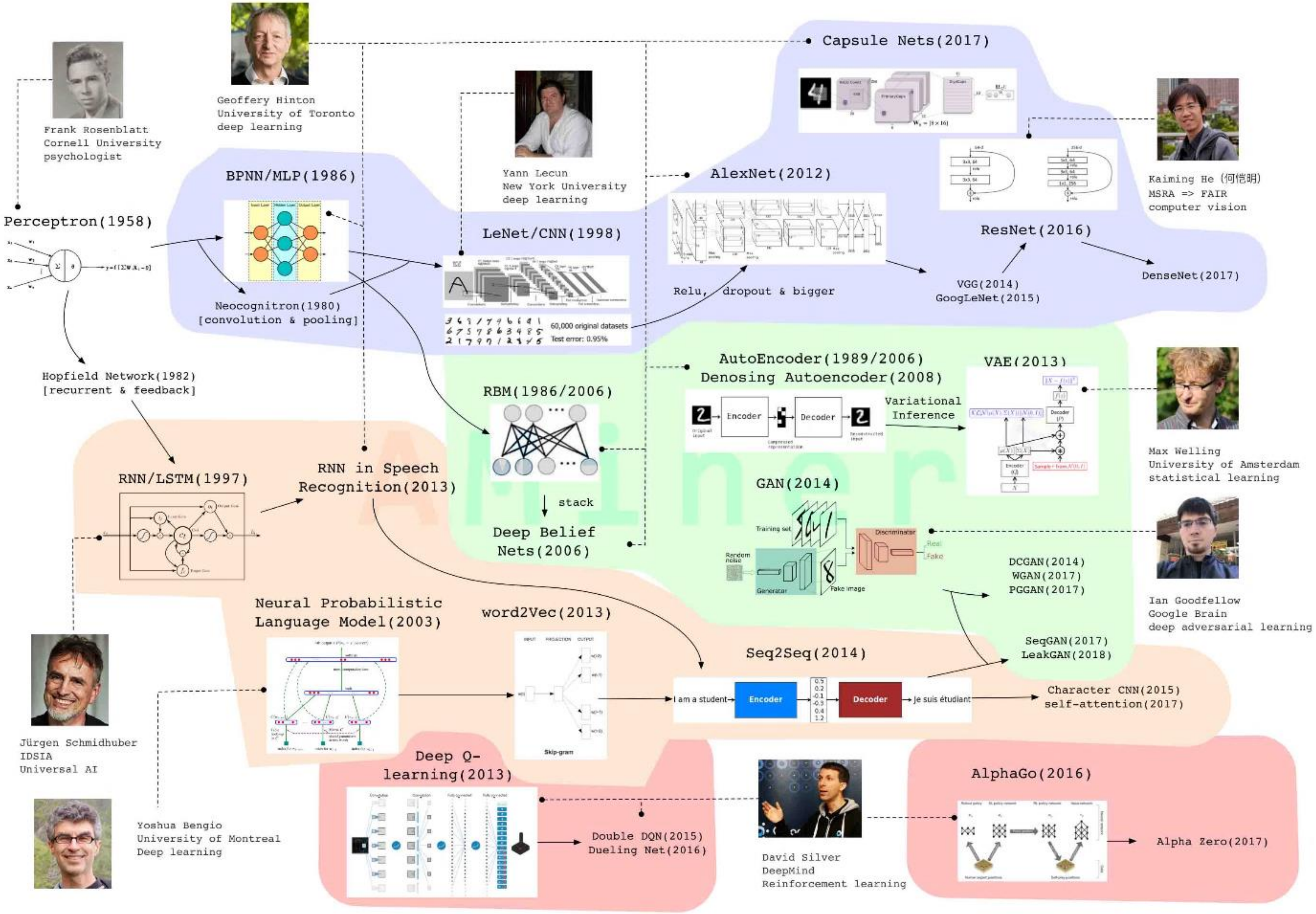
微软每一款核心产品都有微软亚洲研究院技术创新的烙印

Office 365 Bing Windows

XBOX ONE Microsoft Azure

Machine Learning Research @ MSRA

Recent Progress in Machine Learning Community



Technical Challenges

- Relying on big training data
- Requiring big compuation
- Lacking deep theoretical understanding
- Limitations of meta learning/learning2learn

Our Research

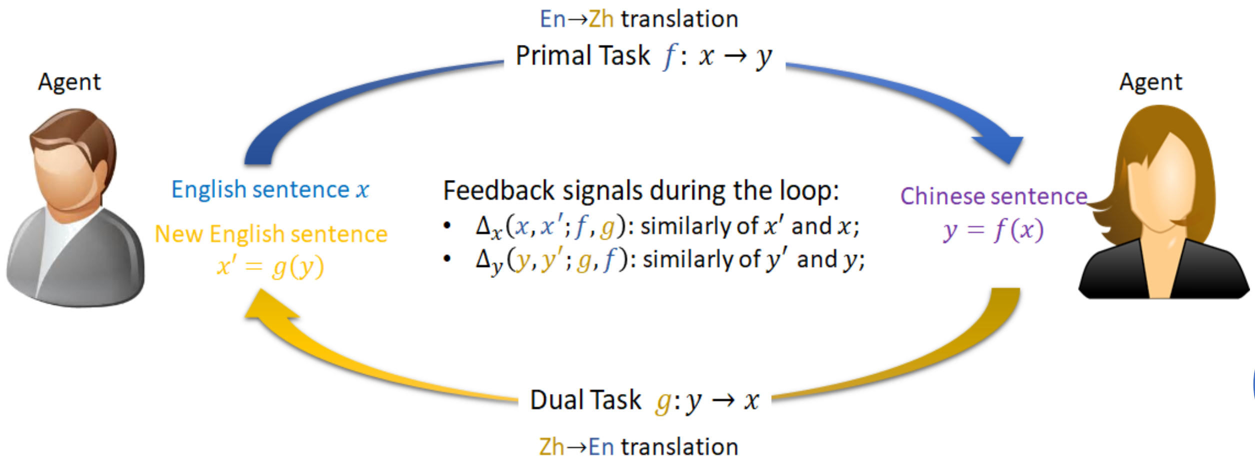
Relying on big training data	Dual learning and its applications on machine translation
Requiring big compuation	Lightweight ML, Distributed ML
Lacking deep theoretical understanding	Depth Theory, G-space deep learning
Limitations of meta learning/learning2learn	Learning to teach

Our Research

Relying on big training data	Dual learning and its applications on machine translation
Requiring big compuation	Lightweight ML, Distributed ML
Lacking deep theoretical understanding	Depth Theory, G-space deep learning
Limitations of meta learning/learning2learn	Learning to teach

Dual Learning (NIPS 2016, ICML 2017, IJCAI 2017, AAAI 2018, ICML 2018)

AI Tasks	X → Y	Y → X
Machine translation	Translation from language EN to CH	Translation from language CH to EN
Speech processing	Speech recognition	Text to speech
Image understanding	Image captioning	Image generation
Conversation	Question answering	Question generation (e.g., Jeopardy!)
Search engine	Query-document matching	Query/keyword suggestion



- Dual unsupervised learning
- Dual supervised learning
- Dual inference
- Dual transfer learning
- Model-level dual learning

Probabilistic Nature of Dual Learning

- The structural duality implies strong probabilistic connections between the models of dual AI tasks.

$$P(x, y) = P(x)P(y|x; f) = P(y)P(x|y; g)$$

Primal View ***Dual View***

- This can be used as
  - Effective feedback signal to close the loop of unsupervised learning
  - Structural regularizer to enhance supervised learning
  - Additional criterion to improve inference

//newstest2017

Human Parity In Machine Translation

AI score: 69.5

Human score: 69.0

@2018.3

Continue to Push the Frontier of NMT

Deep understanding
- Reinforcement learning for MT (EMNLP2018)
- Understanding error propagation in NMT (EMNLP2018)

Self-supervised learning
- Frequency-agnostic word embeddings (NIPS2018)
- Weakly supervised MT

Advanced learning algos
- Deliberation learning (NIPS 2017)

Efficient machine translation
- Non-autoregressive MT (AAAI2018)
- Light NMT models (ICML2018)
- Layer-wise coordination (NIPS2018)

Top language pairs

WMT En-De	2016	2017	2018
Facebook's model	37.99	32.80	46.05
Google production system	38.03	31.41	47.67
Our Results	41.19	34.12	49.77

Middle language pairs

	Am-En	En-Am
Google production system	13.18	24.39
Our Results	15.99	24.94

Tailed language pairs

Google/Yandex/ Bing not support	War-En	En-War	Am-To	Cv-To	Ee-My
Our Results	50.64	46.14	32.01	29.20	23.22

Our Research

Relying on big training data	Dual learning and its applications on machine translation
Requiring big compuation	Lightweight ML, Distributed ML
Lacking deep theoretical understanding	Depth Theory, G-space deep learning
Limitations of meta learning/learning2learn	Learning to teach

Lightweight Machine Learning

(WWW 2015, NIPS 2016/2017)

LightLDA

- The largest/fastest topic model
  - Multiplicative factorization reduces per-token sampling complexity to O(1), which is independent of topic number

$$p(z_{di} = k|rest) \propto \frac{n_{kw}^{-di} + \beta_w}{n_k^{-di} + \bar{\beta}}(n_{kd}^{-di} + \alpha_k)$$

	#Token	#Topics	CPU cores	Training time
LightLDA	100G	1 M	384	60 hrs
Google's LDA	< 10G	< 100K	10,000	70 hrs

LightRNN

- Very compact and fast RNN
  - Multi-component embedding significant reduces the model size, especially for very large vocabulary

Classical RNN language model

- Model size > 100GB
- Training time > 100 years

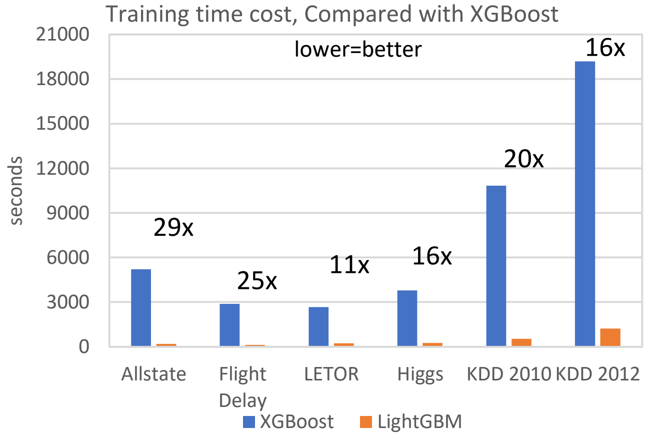
⇩

LightRNN language model

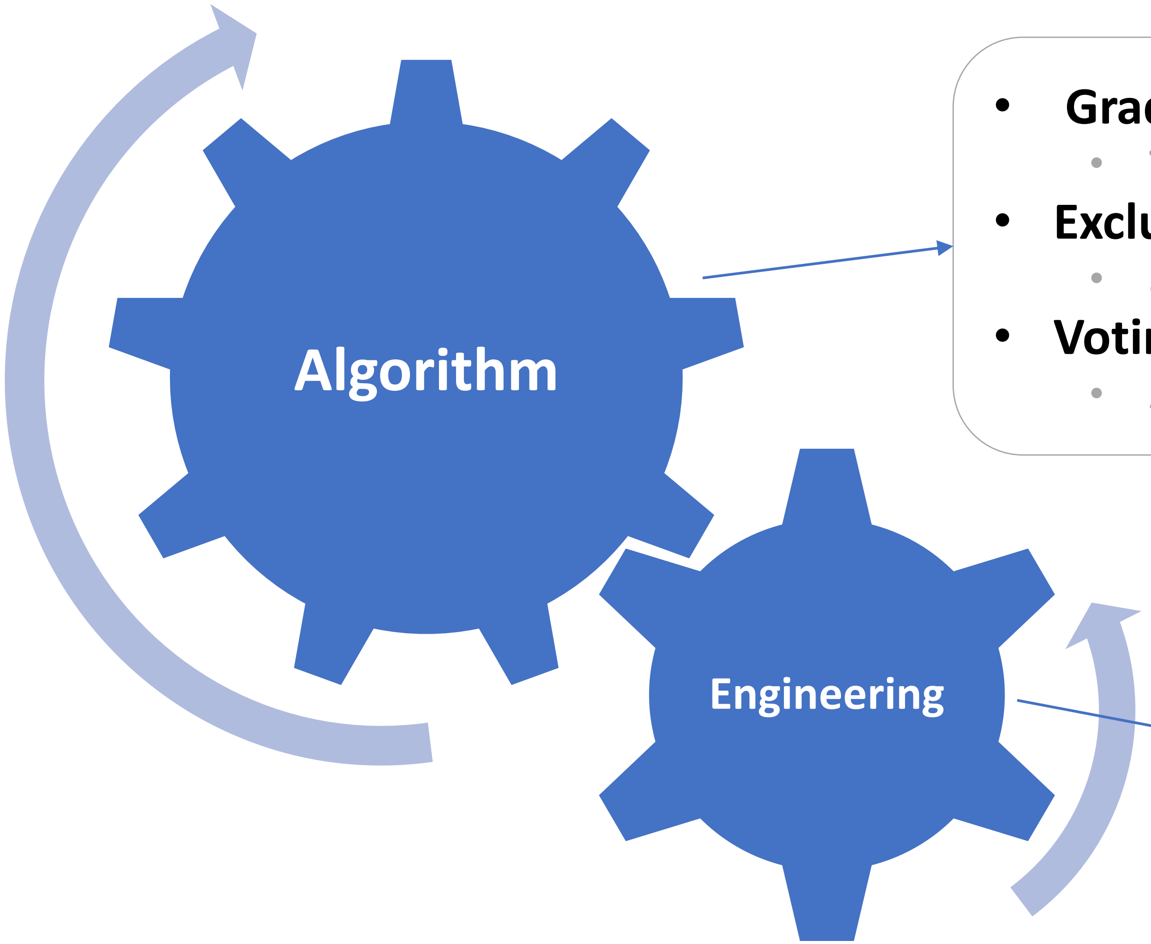
- Model size ~ 50MB
- Training time ~ 1 month

LightGBM

- The fastest GBDT tool
  - Gradient-based one-side sampling
  - Exclusive feature bundling
  - Voting-based parallelization



LightGBM (NIPS 2016/2017)



- Gradient-based one-side sampling
    - Training by using sampling data without loss on accuracy
- Exclusive feature bundling
    - Speed up the training for high dimension data
- Voting-based parallelization
    - Almost lossless with huge reduction of communication cost

- High quality implementation by C++
- Cache-aware optimizations
- Efficient multi-threading

Impact on Open-Source Community

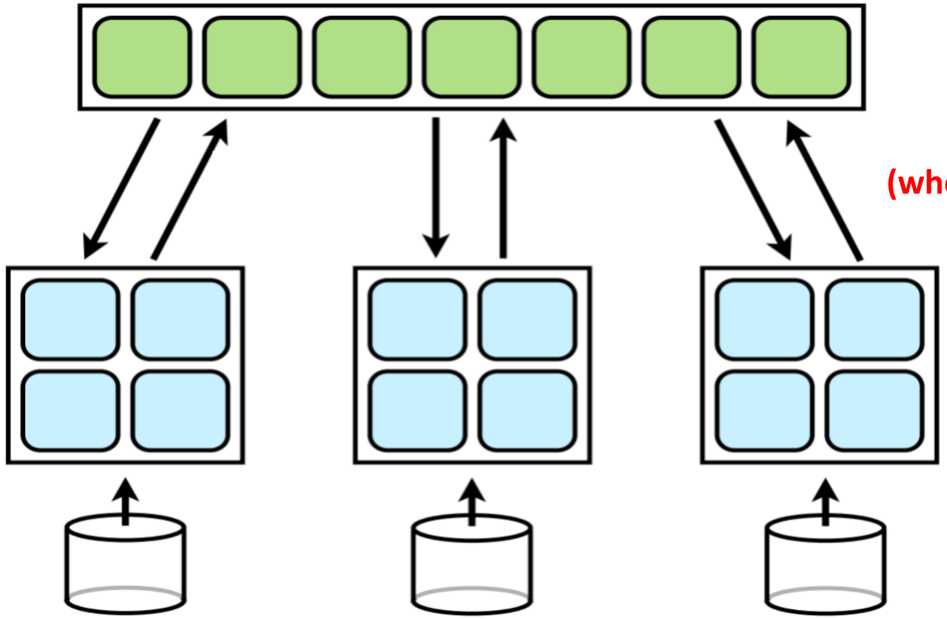
- **Scalability**
  - Low memory consumption
  - Voting based parallelization
  - TLC++, MMLSpark
- **Popularity**
  - ~7000 Github stars
  - 500K+ clones
  - 2M+ pageviews
- **Activity**
  - 500+ contributions
  - Many active contributors
  - Most popular ML blog in 2017
- **Product Impact**
  - Windows
  - Bing/Ads
  - Azure ML

Place	Competition	Solution	Date
1st	Recruit Restaurant Visitor Forecasting	link	2018.2
1st	WSDM CUP 2018 - KKBox's Music Recommendation Challenge	link	2017.12
1st	Porto Seguro’s Safe Driver Prediction	link	2017.11
1st	Quora Question Pairs	link	2017.6
1st	Two Sigma Connect: Rental Listing Inquiries	link	2017.4
1st	CIKM2017 AnalytiCup - Lazada Product Title Quality Challenge	link	2017.9
2nd	Two Sigma Connect: Rental Listing Inquiries	link	2017.4
3rd	Two Sigma Connect: Rental Listing Inquiries	link	2017.4
3rd	Dogs vs. Cats Redux: Kernels Edition	link	-
3rd	Bosch Production Line Performance	link	2016.11
1st	The 1st Di-Tech Competitions	-	2016.7

Distributed Machine Learning

- Big data + Big model ≫ Capacity of a single machine

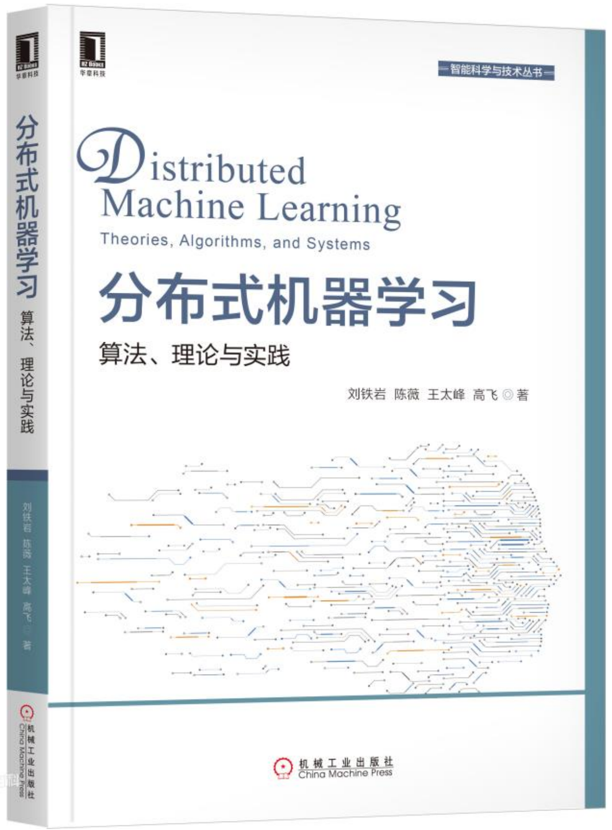
Model aggregation
(how to combine local models)



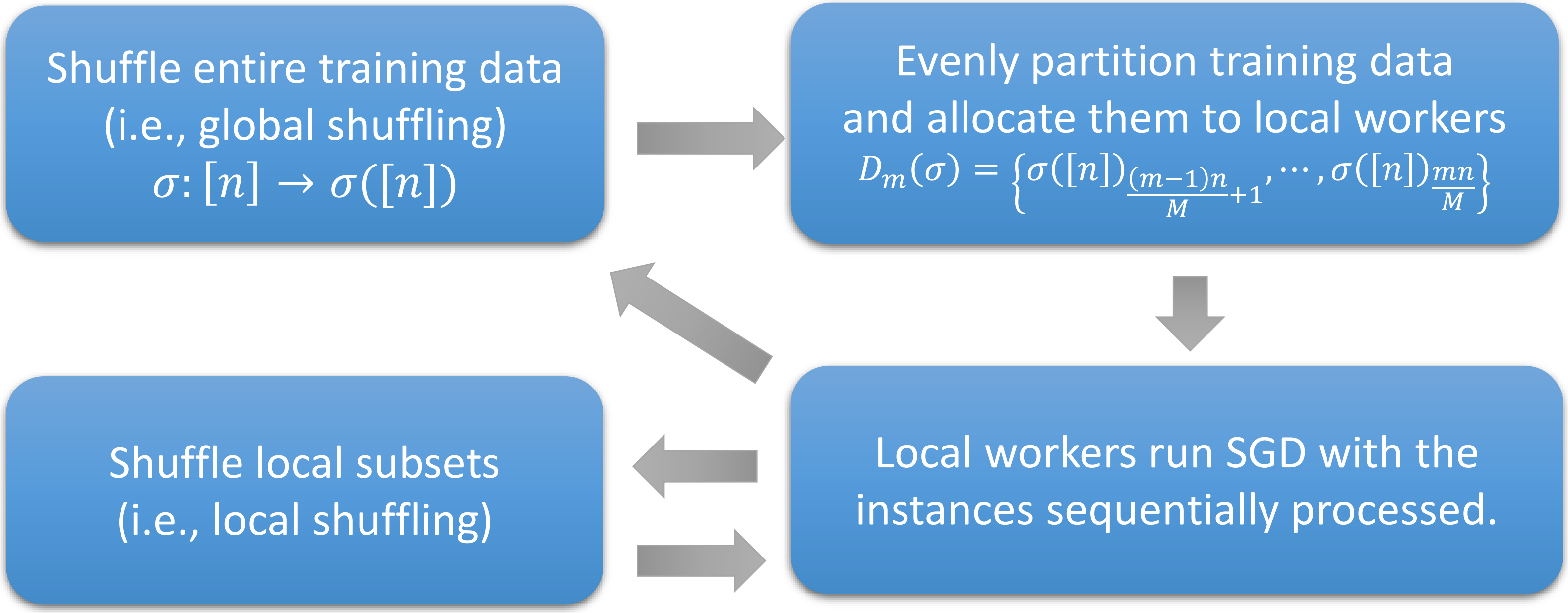
Synchronization mechanism
(whether, when and how to synchronize between local workers)

Deep learning algorithms on single machine/GPU

Data partitioning
(how to partition training data onto local workers)

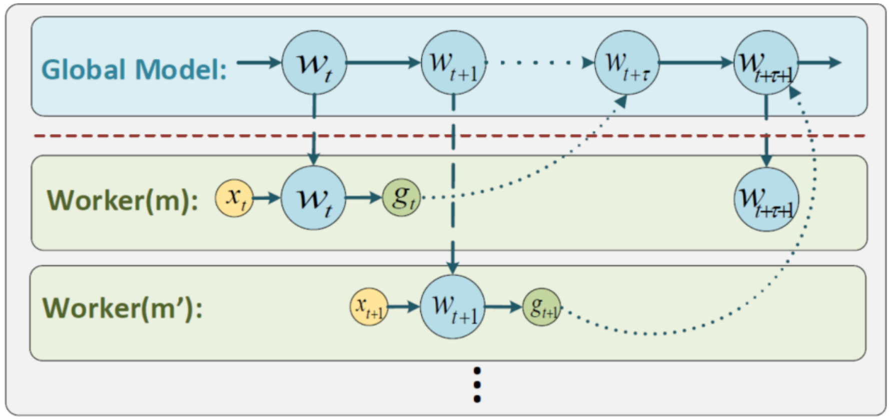


Data Partitioning (NeuroComputing)



- *Global shuffling can achieve similar convergence rate to i.i.d. sampling, since the influence of small shuffling error is negligible.*
- *Local shuffling hurts the convergence rate, and we have to restrict the number of epochs when the number of local workers is large.*

Asynchronous Communication (ICML 2017)



- Sequential SGD

$$w_{t+\tau+1} = w_{t+\tau} - \eta * \boxed{g(w_{t+\tau})}$$

$$\neq$$

- Async SGD

$$w_{t+\tau+1} = w_{t+\tau} - \eta * \boxed{g(w_t)}$$

- Characterizing the delay using Taylor expansion:

$$g(w_{t+\tau}) = g(w_t) + \nabla g(w_t) \cdot (w_{t+\tau} - w_t) + O(\|w_{t+\tau} - w_t\|^2)$$

$\nabla g(w_t)$ corresponds to the Hessian matrix

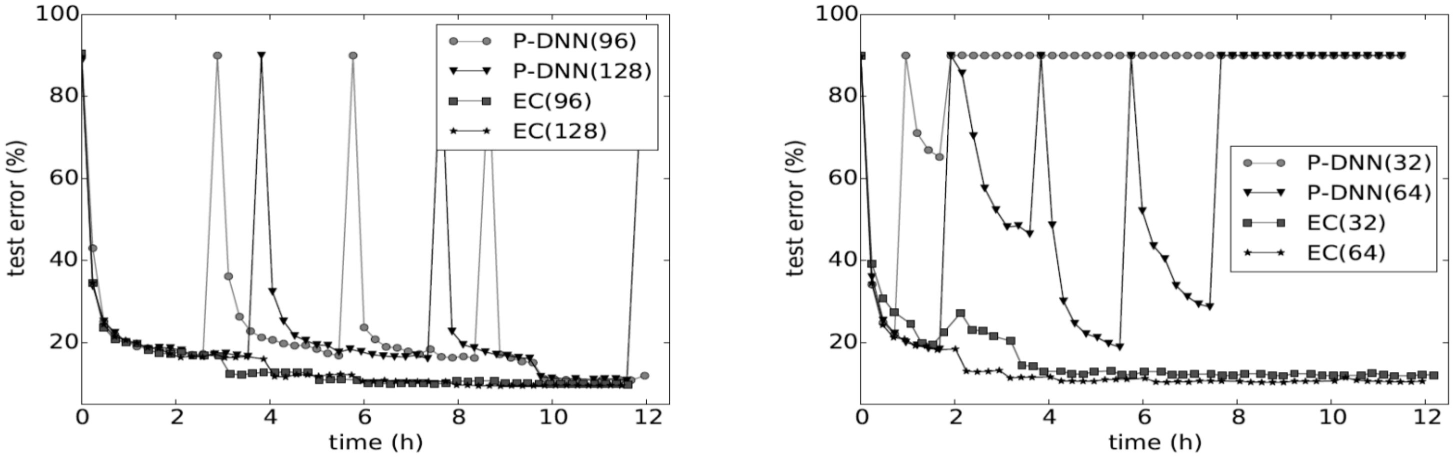
Delay Compensated ASGD (DC-ASGD):

$$w_{t+\tau+1} = w_{t+\tau} - \eta\, g(w_t) - \lambda\phi(g(w_t)) \cdot (w_{t+\tau} - w_t)$$

Theorem: Under mild conditions, DC-ASGD has better convergence properties than ASGD, i.e., more robust to communication delay.

Model Aggregation (ECML 2017, AAMAS 2017)

- Average of model parameter does not have accuracy guarantee dur to the non-convexity of the problem
- Average of the model output (or ensemble of the model) has accuracy guarantee
- Model compression is needed to avoid explosion of the size of ensembled model over multiple iterations

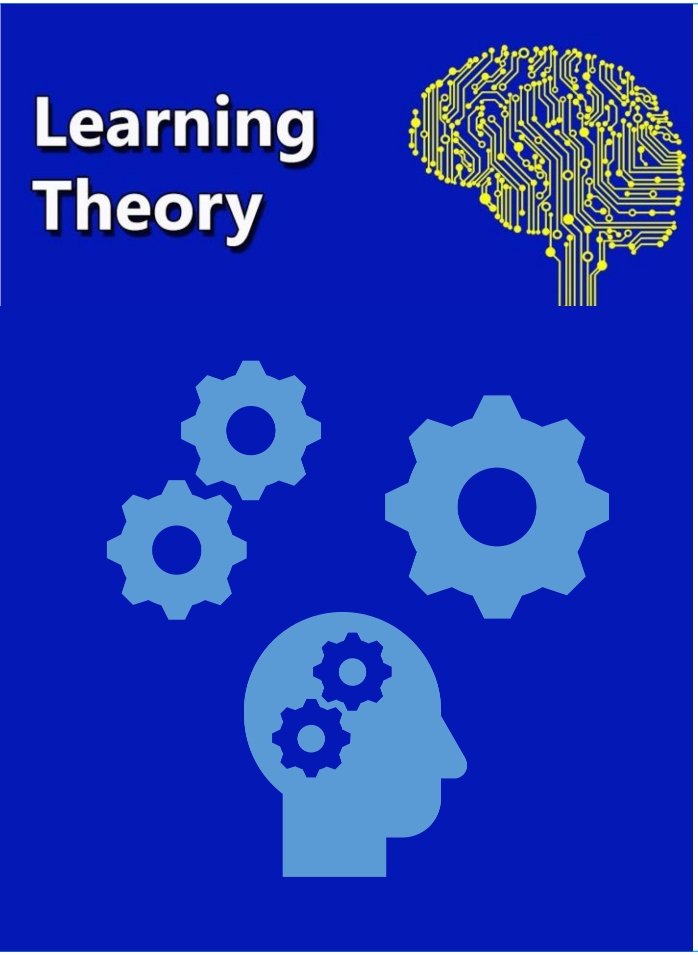


(a) $K = 4$ (b) $K = 8$

Our Research

Relying on big training data	Dual learning and its applications on machine translation
Requiring big compuation	Lightweight ML, Distributed ML
Lacking deep theoretical understanding	Depth Theory, G-space deep learning
Limitations of meta learning/learning2learn	Learning to teach

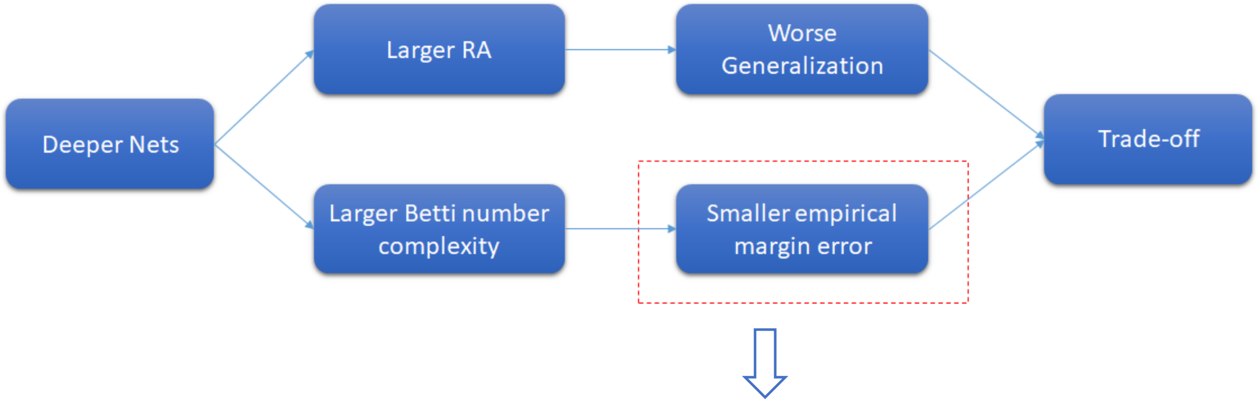
Depth Theory for Deep Neural Networks (AAAI 2016)



Generalization bound

$$err_P(f) \le \inf_{\gamma>0}\left\{err_S^{\gamma}(f) + \frac{8R_m(\Omega)}{\gamma} + \sqrt{\frac{\log\log(2\gamma^{-1})}{m}} + \sqrt{\frac{\log(2\delta^{-1})}{2m}}\right\}$$

Impact of depth on expressiveness and generalization

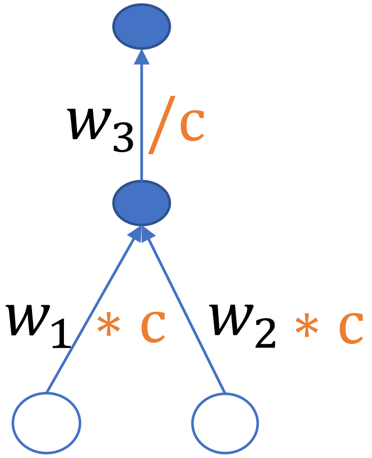


Design of Large-margin DNN, for improved performance

$$C_1(f;x,y) = C(f;x,y) + \lambda\Big(1 - \rho(f;x,y)\Big)^2,$$

$$C_2(f;x,y) = C(f;x,y) + \frac{\lambda}{K-1}\sum_{k\neq y}\Big(1 - (f(x,y) - f(x,k))\Big)^2$$

$\mathcal{G}$-Space Deep Learning (AAAI 2019)

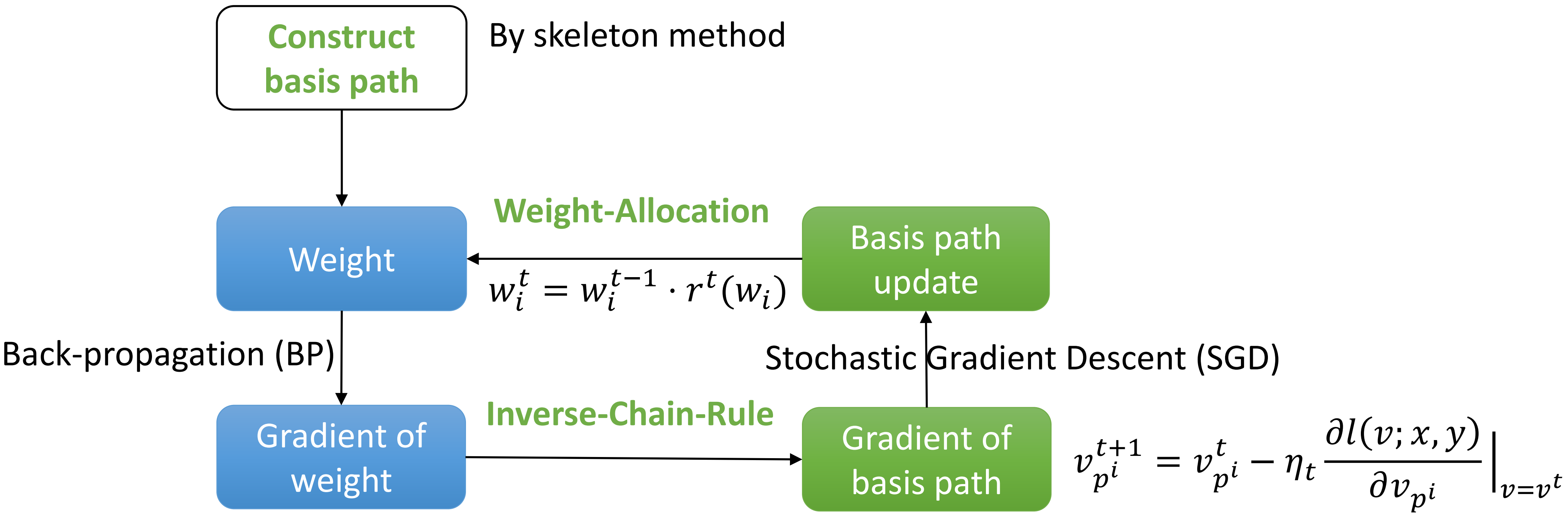


Positive scaling operator

- Neural networks with ReLU activations are positive scaling invariant (denoted as $\mathcal{G}$ – **invariant**)
  - However, the weight space of ReLU networks are **NOT** $\mathcal{G}$ – **invariant.**
  - Optimization in the weight space will suffer from gradient vanishing/exploding or spurious critical points!

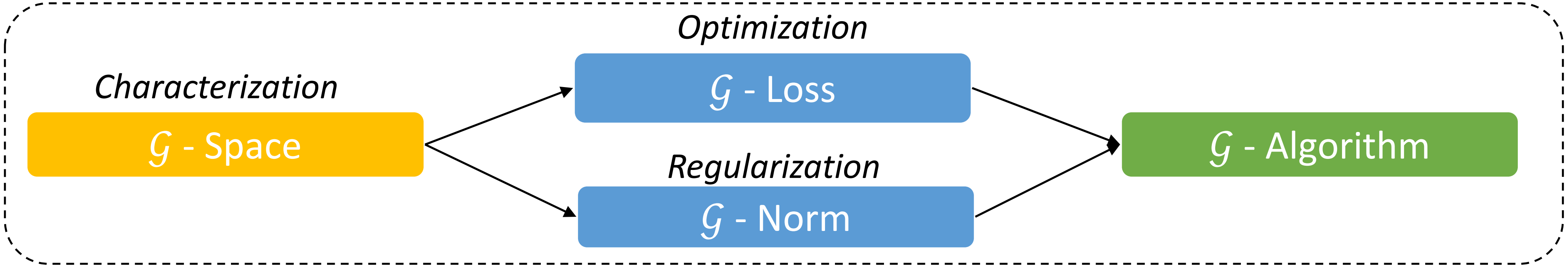
$\mathcal{G}$ – Space: We prove that the bases in the path space (together with their values) are representation-sufficient and **$\mathcal{G}$ – invariant.**

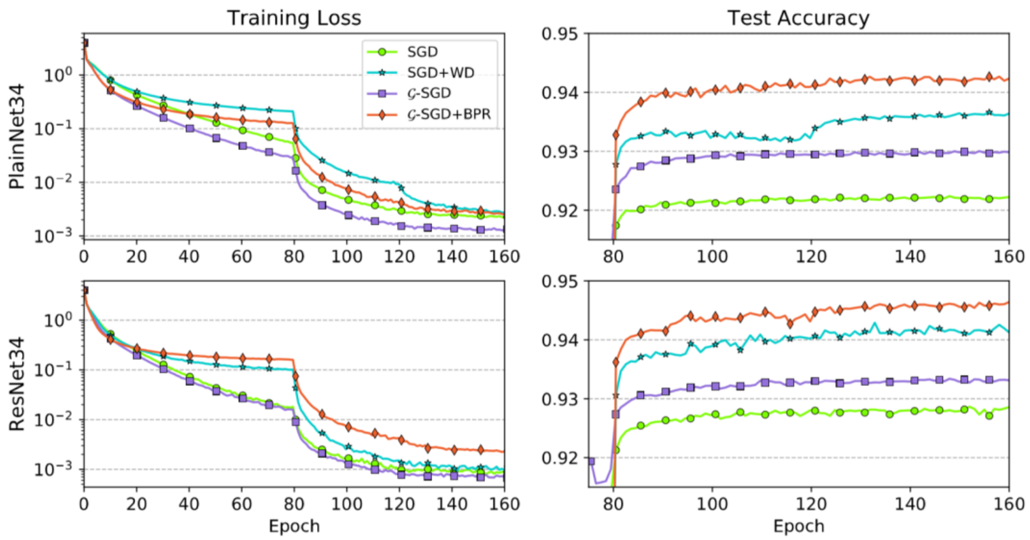
$\mathcal{G}$-Space Stochastic Gradient Descent



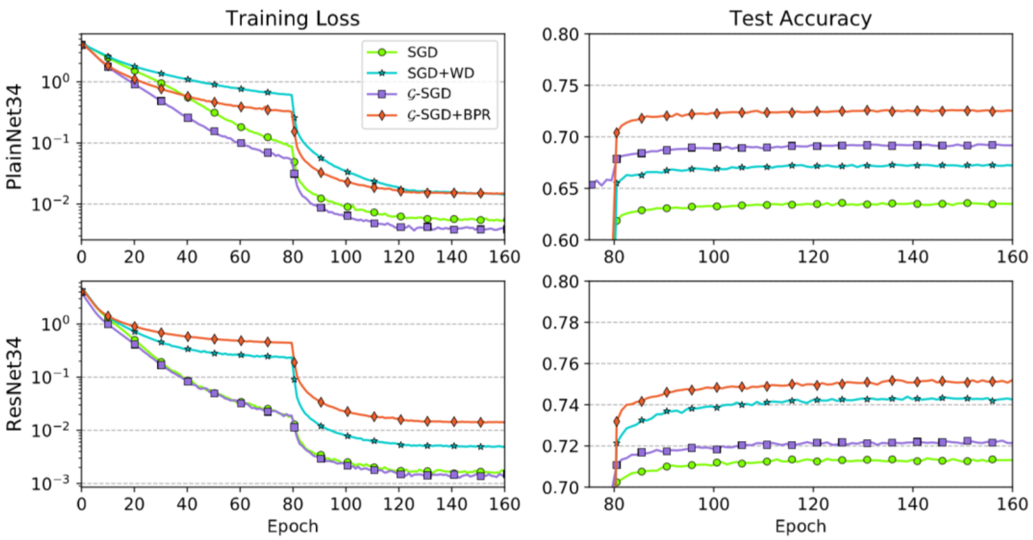
$$\left(v_w(p^1), \cdots, v_w(p^{m-H})\right) = (w_1, \cdots, w_m) \odot \tilde{A}, \text{ where} \tilde{A} = \begin{bmatrix} p_1^1 \cdots\ p_1^{m-H} \\ p_2^1 \cdots\ p_2^{m-H} \\ \cdots\cdots \\ p_m^1 \cdots\ p_m^{m-H} \end{bmatrix}$$

Experimental Results





CIFAR - 10



CIFAR - 100

Our Research

Relying on big training data	Dual learning and its applications on machine translation
Requiring big compuation	Lightweight ML, Distributed ML
Lacking deep theoretical understanding	Depth Theory, G-space deep learning
Limitations of meta learning/learning2learn	Learning to teach

Learning to Teach: Beyond Learning/Meta Learning

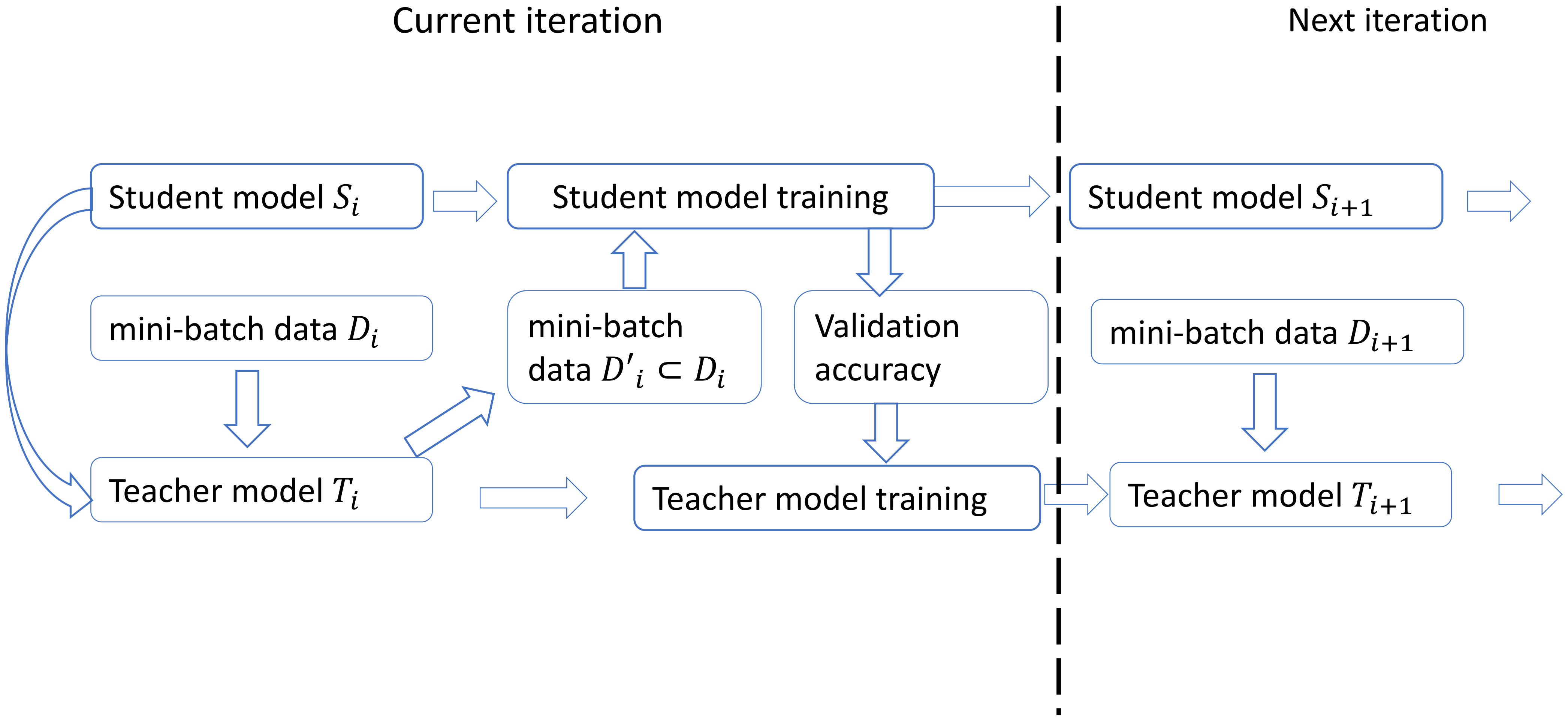
$$\omega^* = \arg\min_{\omega \in \boldsymbol{\Omega}} \sum_{(x,y)\in \boldsymbol{D}} \boldsymbol{L}(f_\omega(x), y)$$

D Learning to select the most appropriate data to train the student model at appropriate time (ICLR 2018)

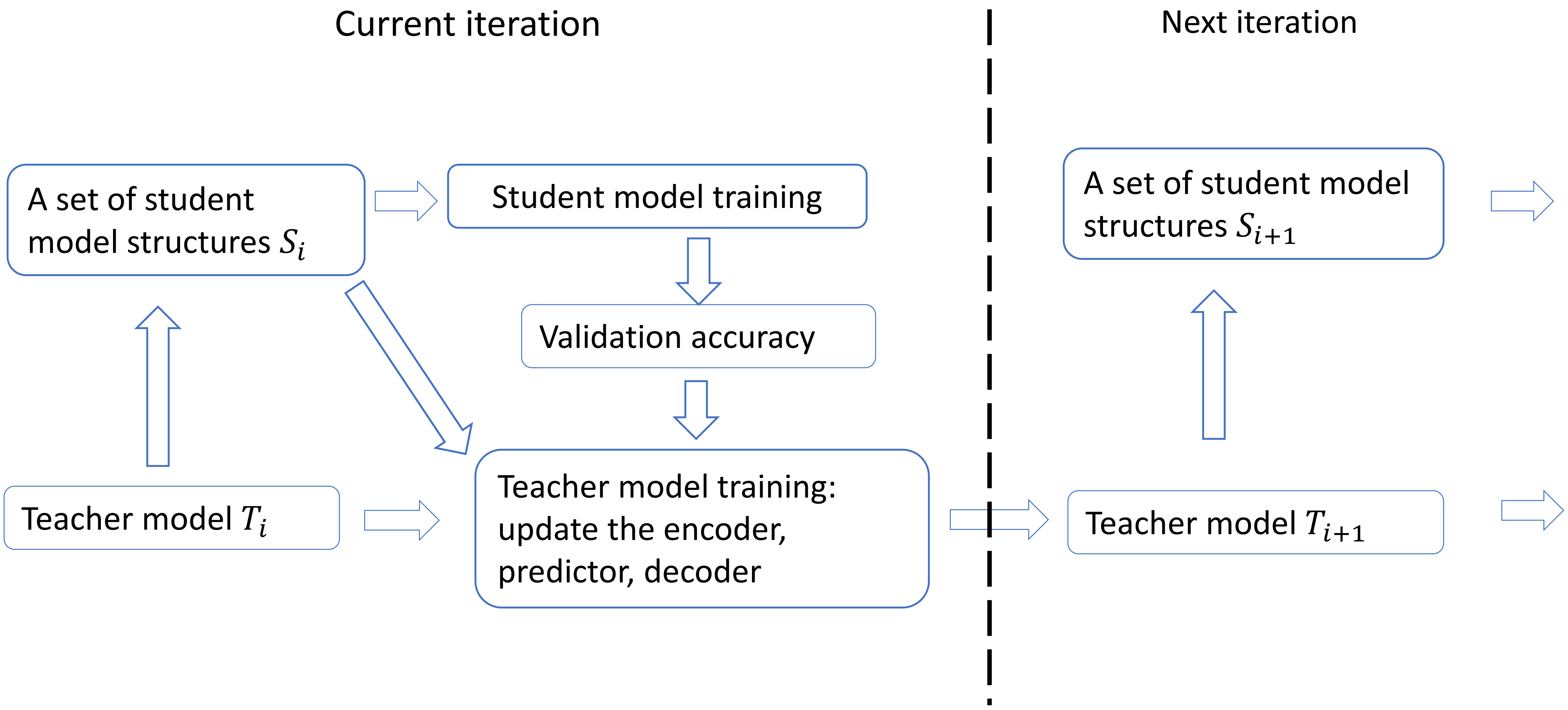
Ω Gradually adjust the model structure to better solve the target problem (NIPS 2018)

L Gradually adjust the loss function adaptive to the current training stages (NIPS 2018)

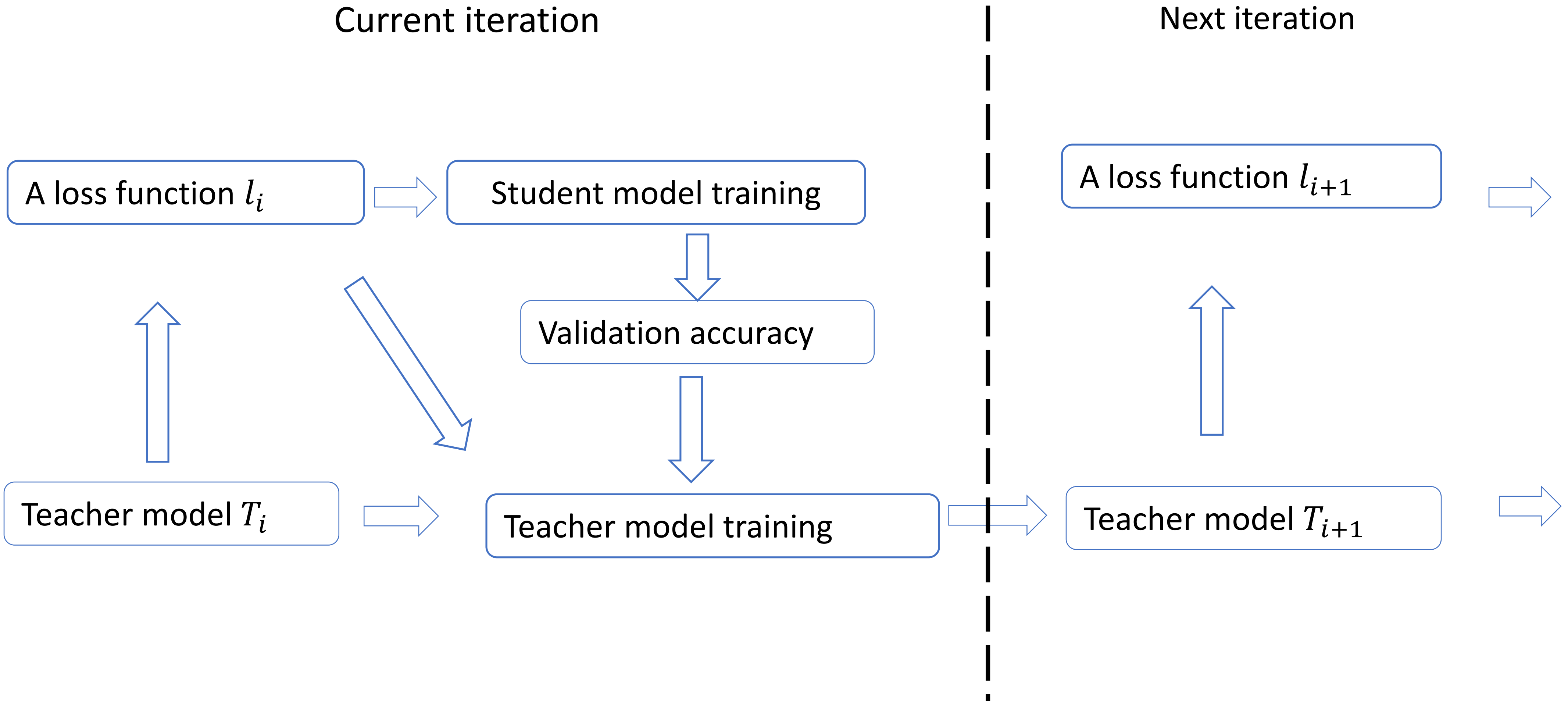
Data Teaching (ICLR 2018)



Model Teaching (NIPS 2018)

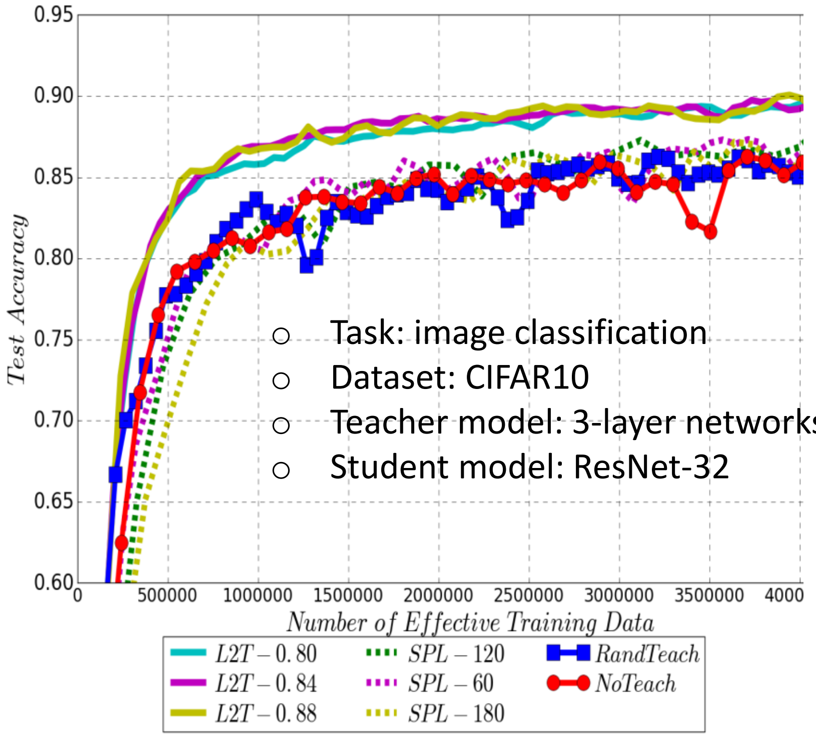


Loss Teaching (NIPS 2018)



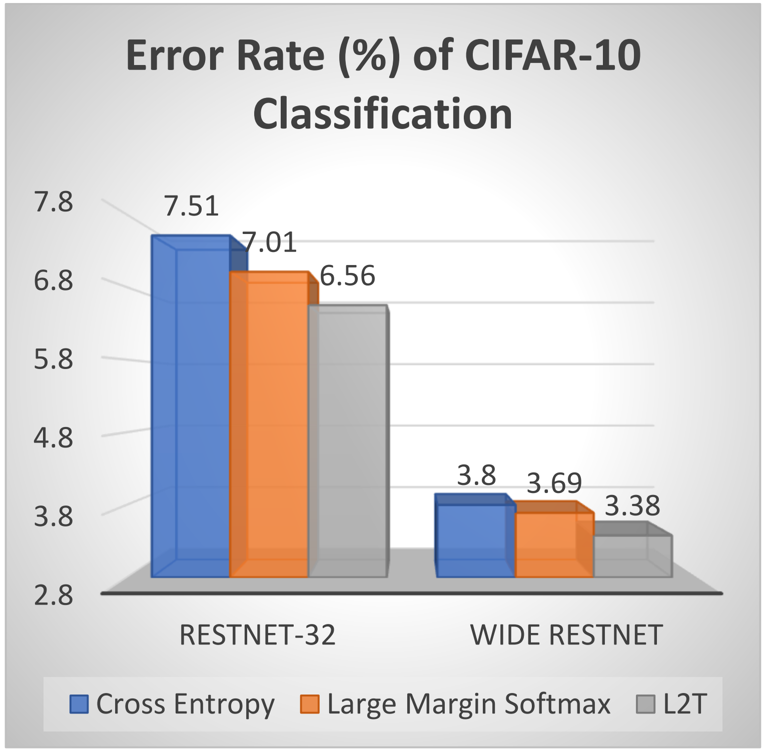
Experimental Results

Data Teaching



- Task: image classification
- Dataset: CIFAR10
- Teacher model: 3-layer networks
- Student model: ResNet-32

Loss Teaching



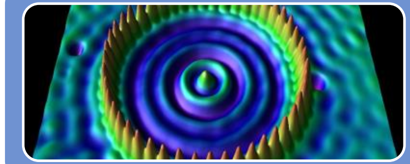
Model Teaching

Method (original model)	Error Rate	Resource (#GPU × #Hours)
AmoebaNet (Google Brain, 2018.2)	*2.13*	*3150 * 24*
Hie-EA (DeepMind, 2017.11)	*3.15*	*300 * 24*
NAO (MSRA)	***2.07***	***200 * 24***

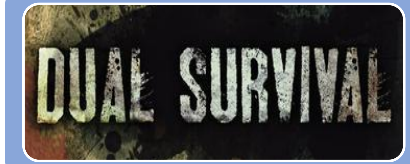
Method (weight sharing)	Error Rate	Resource (#GPU × #Hours)
ENAS (Google Brain, 2018.2)	2.89	12
DARTS (CMU & DeepMind, 2018.6)	2.83	96
NAO-WS (MSRA, 2018.6)	**2.80**	**7**

Looking into the Future …

Machine Learning vs. Quantum Computing



Simple & Elegant Laws vs. Complex Models



Patter Recognition vs. Prediction vs. Improvisation

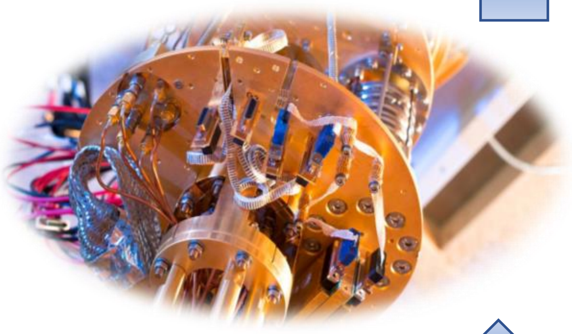
Social Intelligence vs. Individual Intelligence

Machine Learning vs. Quantum Computing

Method	Speedup	AA	HHL	Adiabatic	QRAM
Bayesian Inference [107, 108]	$O(\sqrt{N})$	Y	Y	N	N
Online Perceptron [109]	$O(\sqrt{N})$	Y	N	N	optional
Least squares fitting [9]	$O(\log N^{(*)})$	Y	Y	N	Y
Classical BM [20]	$O(\sqrt{N})$	Y/N	optional/N	N/Y	optional
Quantum BM [22, 62]	$O(\log N^{(*)})$	optional/N	N	N/Y	N
Quantum PCA [11]	$O(\log N^{(*)})$	N	Y	N	optional
Quantum SVM [13]	$O(\log N^{(*)})$	N	Y	N	Y
Quantum reinforcement learning [30]	$O(\sqrt{N})$	Y	N	N	N

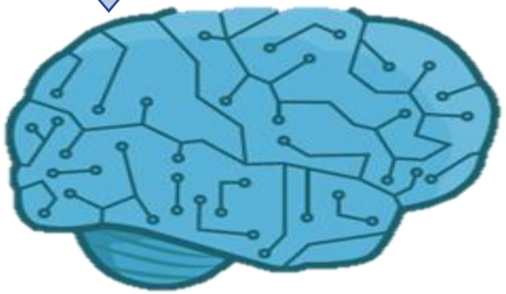
Quantum speedup: QFT, Grover Search, quantum annealing…

New search and optimization methods



Quantum Computing (QC)

Quantum machine learning theory



Machine Learning

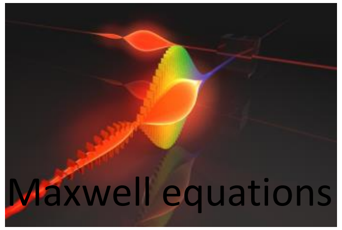
Expected case vs. worst case

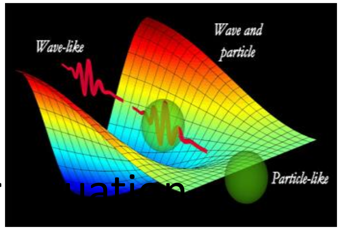
Learning assisted quantum computing: deep & reinforcement learning

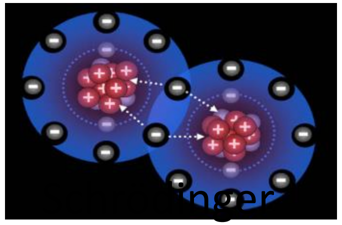
- Quantum state prediction
- Quantum operation prediction
- Quantum property testing / distinguishing / classification
- Quantum communication and encryption

Simple & Elegant Laws vs. Complex Model

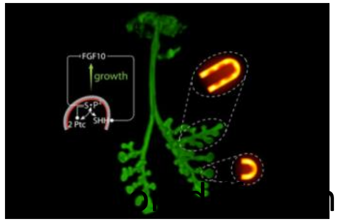
photon | small particles | any matter







chemical bonds



morphogen

economics

"It turns out that almost all the traditional mathematical models that have been used in physics and other areas of science are ultimately based on partial differential equations."

-- Stephen Wolfram

Learning Laws vs. Fitting Data

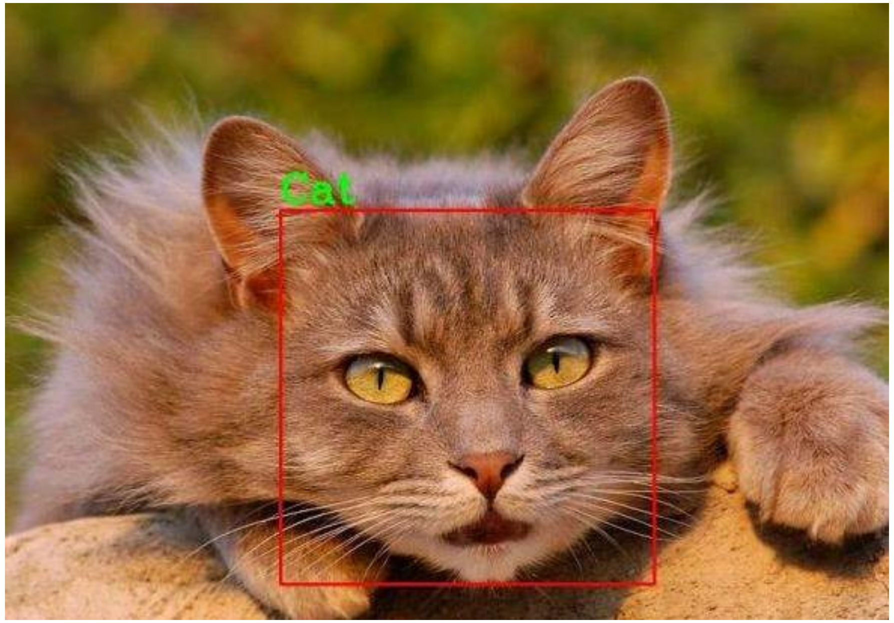
以简治繁　　以繁治繁

- It was shown that natural laws can be automatically discovered by evolutionary algorithms (Science 2009)

- How about automatically learning simple & elegant laws behind complicated data we have?
    - Data is just the phenomenon
    - Laws that govern the generation of the data is the essence
    - New machine learning models are needed, such as dynamic systems and partial equations

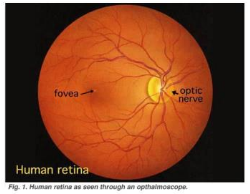
Pattern Cognition vs. Prediction

Pattern Recognition



Predictive Learning

- Build world model + predict the future



- Infer the state of the world from partial information
- Infer the future from the past and present
- Infer past events from the present state

- Filling in the visual field at the retinal blind spot
- Filling in occluded images
- Fillling in missing segments in text, missing words in speech.
- Predicting the consequences of our actions
- Predicting the sequence of actions leading to a result

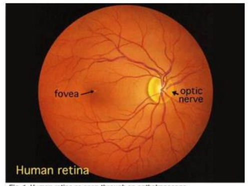
- Predicting any part of the past, present or future percepts from whatever information is available.

- That's what predictive learning is
- But really, that's what many people mean by unsupervised learning

Prediction vs. Improvisation

Predictive Learning

- Build world model + predict the future



- Infer the state of the world from partial information
- Infer the future from the past and present
- Infer past events from the present state

- Filling in the visual field at the retinal blind spot
- Filling in occluded images
- Fillling in missing segments in text, missing words in speech.
- Predicting the consequences of our actions
- Predicting the sequence of actions leading to a result

- Predicting any part of the past, present or future percepts from whatever information is available.

- That's what predictive learning is
- But really, that's what many people mean by unsupervised learning

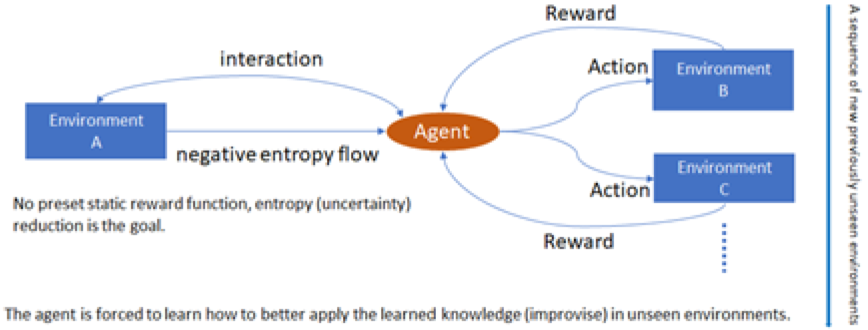
Improvisational Learning

- Challenge: Is the world predictable?

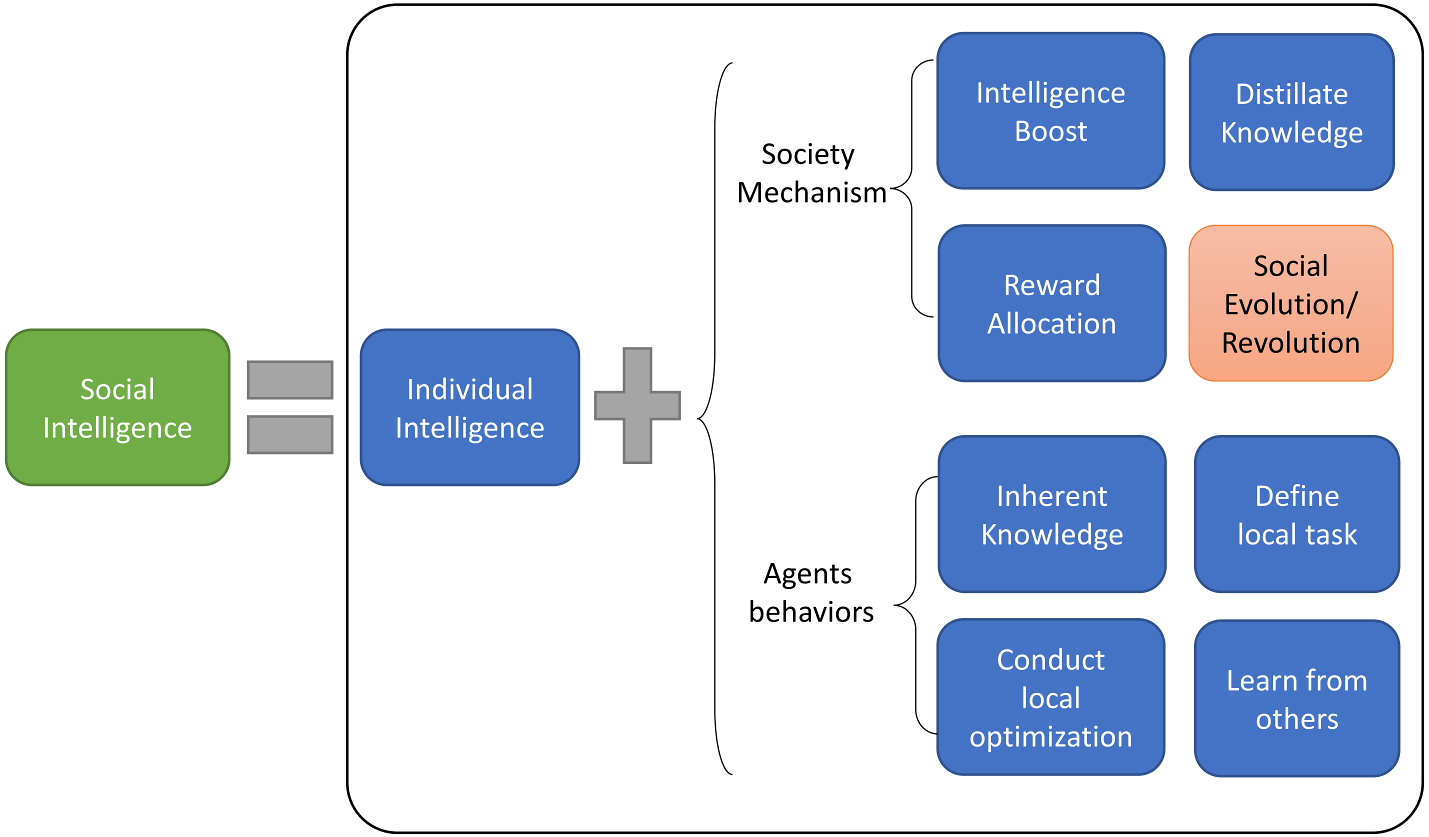
"The only thing predictable about life is its unpredictability." -- Remy in Ratatouille

- Improvisational learning:

  The world is full of exceptions and one needs to improvise to survive when unexpected things happen.



Social Intelligence vs. Individual Intelligence



- Social Coopetition
  - Multiple layers of sub-societies with different mechanisms.
  - Local agents are coopetiting (collaborating and competing) with each other, given the structure of sub-societies.
- Society Evolution/Revolution:
  - Diversity and E-E tradeoff play an important role in evolution process
  - If a sub-society always has low performance, it will be replaced by another sub-society and its mechanism.
  - With the coopetition among sub-societies, the whole society is evolving towards higher performance.

References

Dual Learning

- Di He, Yingce Xia, Tao Qin, Tie-Yan Liu, and Wei-Ying Ma, *Dual Learning for Machine Translation*, **NIPS** 2016
- Yingce Xia, Jiang Bian, Tao Qin, Tie-Yan Liu, *Dual Inference for Machine Learning*, **IJCAI** 2017.
- Yingce Xia, Tao Qin, Wei Chen, Tie-Yan Liu, *Dual Supervised Learning*, **ICML** 2017.
- Yijun Wang , Yingce Xia , Li Zhao , Jiang Bian , Tao Qin, Guiquan Liu , Tie-Yan Liu, *Dual Transfer Learning for Neural Machine Translation with Marginal Distribution Regularization*, **AAAI** 2018.
- Yingce Xia, Xu Tan, Fei Tian, Tao Qin, Nenghai Yu, and Tie-Yan Liu, *Model-Level Dual Learning*, **ICML** 2018.

References

Neural Machine Translation

- Yingce Xia , Lijun Wu , Jianxin Lin , Fei Tian , Tao Qin , and Tie-Yan Liu, *Deliberation Networks: Sequence Generation Beyond One-Pass Decoding*, **NIPS** 2017.
- Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tie-Yan Liu, *Decoding with Value Networks for Neural Machine Translation*, **NIPS** 2017.
- Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou, *Achieving Human Parity on Automatic Chinese to English News Translation*, **arXiv** 2018.
- Xu Tan, Lijun Wu, Di He, Fei Tian, Tao QIN, Jianhuang Lai and Tie-Yan Liu, Beyond Error Propagation in Neural Machine Translation: Characteristics of Language Also Matter, **EMNLP** 2018.
- Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai and Tie-Yan Liu, A Study of Reinforcement Learning for Neural Machine Translation, **EMNLP** 2018.
- Zhuohan Li, Di He, Fei Tian, Wei Chen, Tao Qin, Liwei Wang, and Tie-Yan Liu, *Towards Binary-Valued Gates for Robust LSTM Training*, **ICML** 2018.
- Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, Tie-Yan Liu, Improving Word Embedding by Adversarial Training, **NIPS 2018.**
- Yiren Wang, Fei Tian, Di He, Tao Qin, Chengxiang Zhai, Tie-Yan Liu, Non-Autoregressive Machine Translation with Auxiliary Regularization, **AAAI 2019**
- Chang Xu, Weiran Huang, Hongwei Wang, Gang Wang and Tie-Yan Liu, Modeling Local Dependence in Natural Language with Multi-channel Recurrent Neural Networks, **AAAI 2019**

References

Distributed Machine Learning

- Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, and Tie-Yan Liu, *Asynchronous Stochastic Gradient Descent with Delay Compensation*, **ICML** 2017.
- Shizhao Sun, Wei Chen, Jiang Bian, and Tie-Yan Liu, *Ensemble-Compression: A New Method for Parallel Training of Deep Neural Networks*, **ECML** 2017.
- 刘铁岩，陈薇，王太峰，高飞，分布式机器学习：算法、理论与实践，机械工业出版社，2018

Lightweight Machine Learning

- Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Xing, Tie-Yan Liu, and Wei-Ying Ma, *LightLDA: Big Topic Models on Modest Computer Cluster*, **WWW** 2015.
- Xiang Li, Tao Qin, and Tie-Yan Liu, *LightRNN: Compuation and Memory Efficient Recurrent Neural Networks*, **NIPS** 2016
- Guolin Ke, Qi Meng, Taifeng Wang, Wei Chen, Weidong Ma, Tie-Yan Liu, *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, **NIPS** 2017.

References

Deep Learning Theory

- Shizhao Sun, Wei Chen, Liwei Wang, and Tie-Yan Liu, *On the Depth of Deep Neural Networks: A Theoretical View*, **AAAI** 2016.
- Shuxin Zheng, Qi Meng, Huishuai Zhang, Wei Chen, and Tie-Yan Liu, Capacity Control of ReLU Neural Networks by Basis-path Norm, **AAAI 2019.**

Learning to Teach

- Fei Tian, Tao Qin, and Tie-Yan Liu, *Learning to Teach*, **ICLR** 2018.
- Lijun Wu, Fei Tian, Yingce Xia, Tao Qin, Tie-Yan Liu, Learning to Teach with Dynamic Loss Functions, **NIPS 2018.**
- Renqian Luo, Fei Tian, Tao Qin, Tie-Yan Liu, Automatic Neural Architecture Design: From Search to Optimization, **NIPS 2018.**

Future of Machine Learning

- 刘铁岩，秦涛，邵斌，陈薇，边江，预见未来丨机器学习：未来十年研究热点，**微软研究院AI头条，2018**

Thanks

tyliu@microsoft.com

https://www.microsoft.com/en-us/research/people/tyliu/