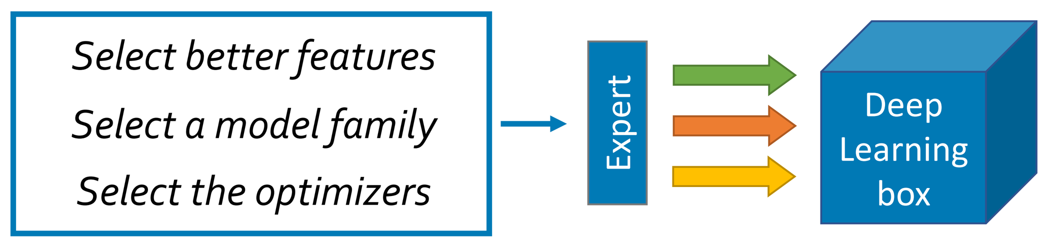
Scalable Automated Machine Learning in GLMix 2.0

LinkedIn
Carnegie Mellon University

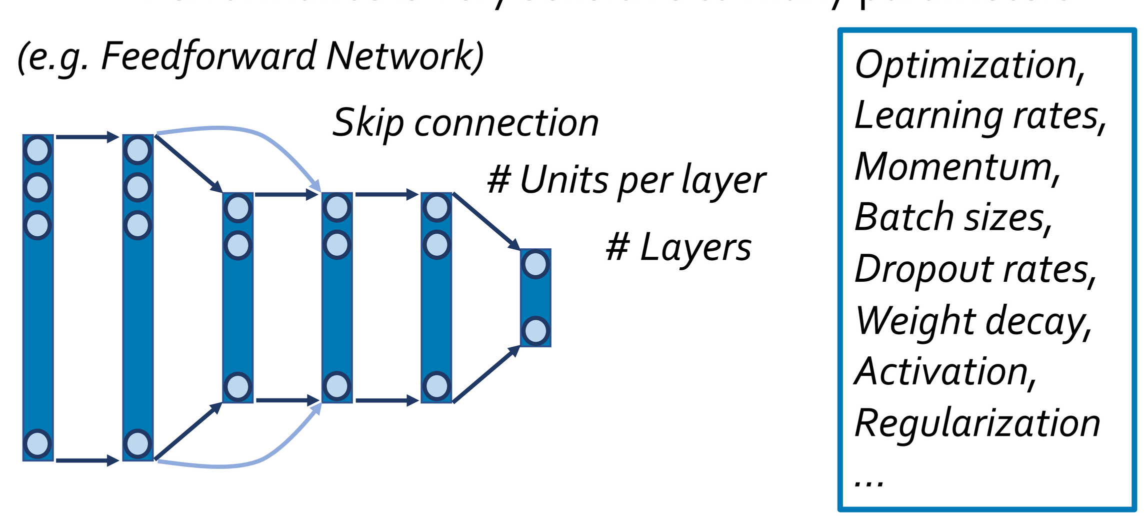
Yuwei Qiu, Chengming Jiang, Huiji Gao, Bee-Chung Chen
AI Algorithms Foundation Team

MOTIVATION

- Learn more effectively, less trial-and-error
- Problems in so-called "end-to-end Deep Learning":
  - Great consumption of (expertise) human power
  - High cost of hyperparameter tuning/ structure learning



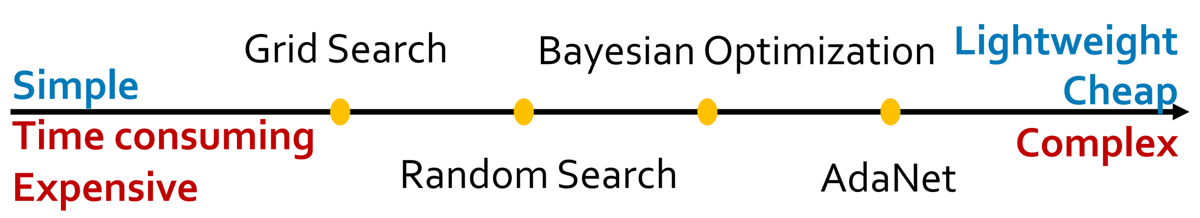
- Performance is very sensitive to many parameters



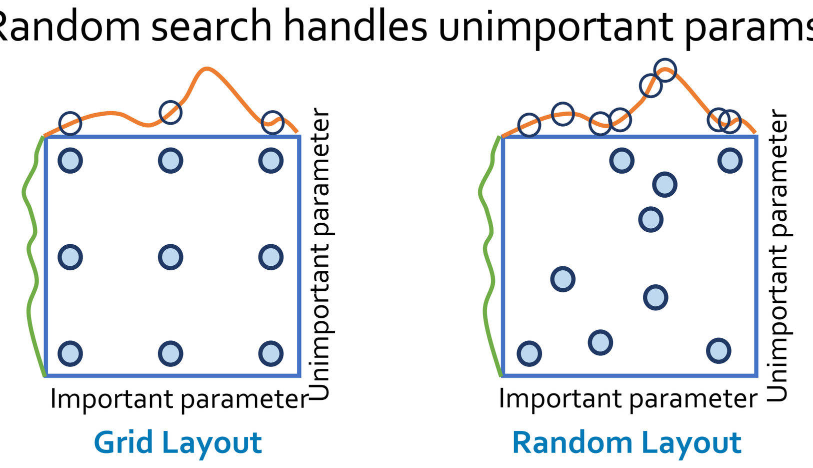
Automatic Machine Learning

Current	*We can turn this into*
Solution = Expertise + Data + Hours	*Solution = Data + N × Hours*

BASELINE

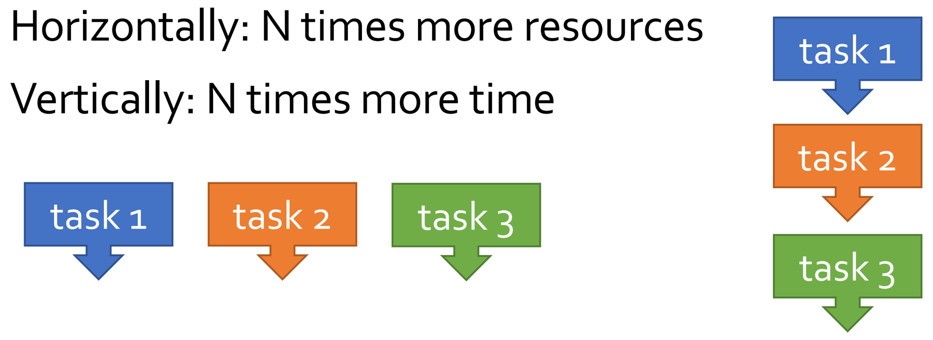


- Grid Search and Random Search
  - Shoot all structures with all hyperparameters
  - Then find the best set of parameters
  - Random search handles unimportant params better



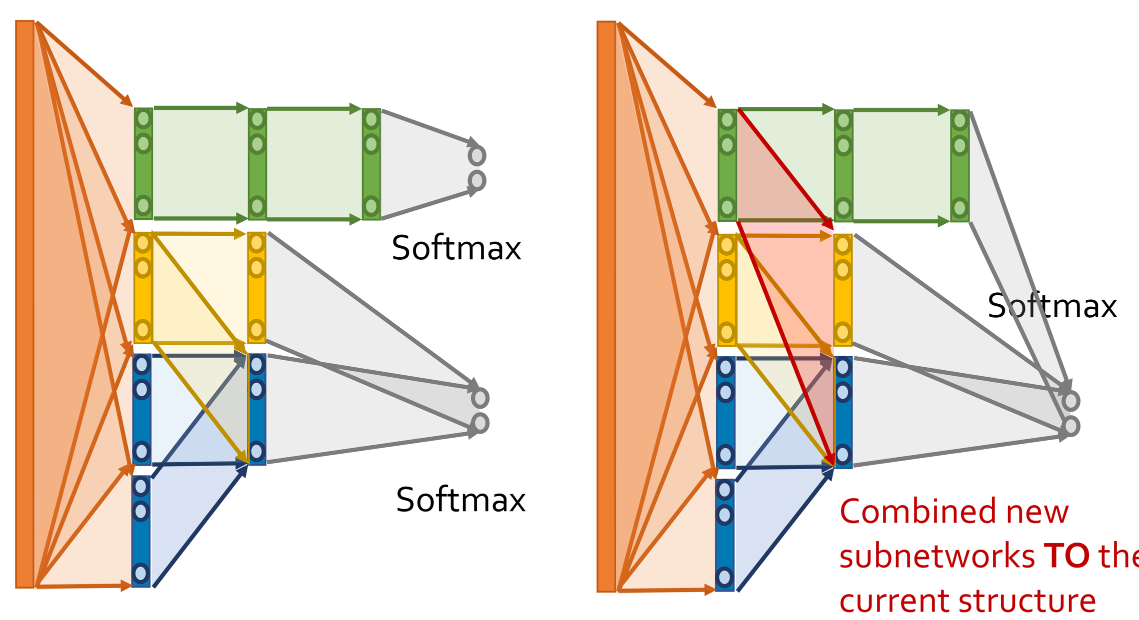
OBSERVATION

- Disastrous heavyweight structure exploration
  - Horizontally: N times more resources
  - Vertically: N times more time



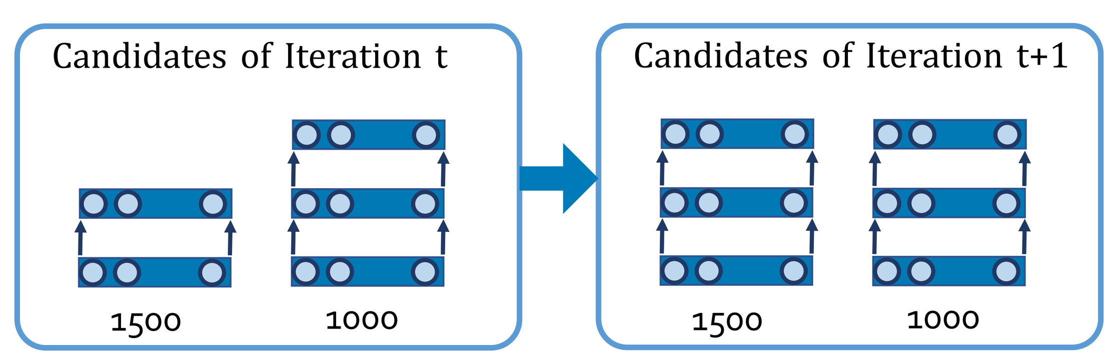
ALGORITHM

- AdaNet
  - Structure grows with lightweight subnetworks.
  - Create high-quality Neural Networks by selecting optimal subnetworks and implementing an adaptive learning process to attain an ensemble of these subnetworks.
  - Capable of adding subnetworks of different depth, widths and types.



Iteration t: Train the subnetworks separately

Iteration t: Ensemble with newly added weights

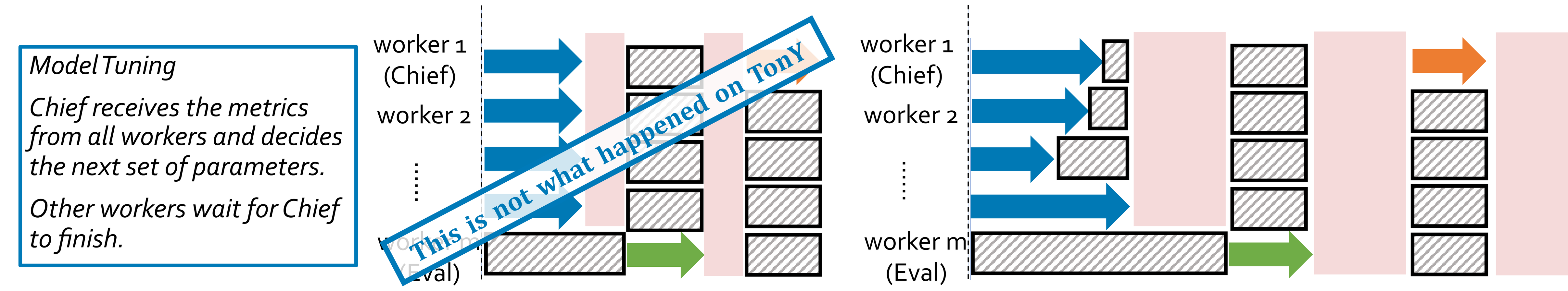


CONTRIBUTION

- Exploration of various auto-tuning algorithms sheds light on future direction for different DNN models at LinkedIn.
- Wiped out disastrous heavyweight structure exploration problem in real-world auto-tuning services by introducing adaptive structure learning algorithm.
- Implemented asynchronous model training to avoid straggler effects and communication latency.

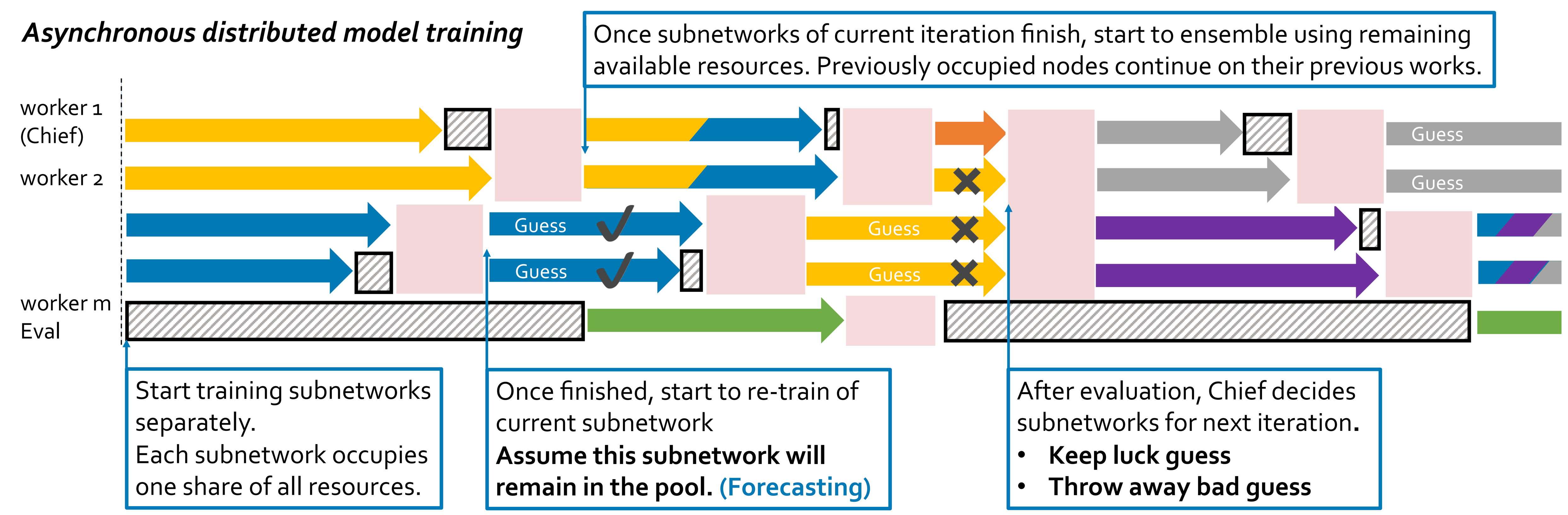
CHALLENGE

- Straggler effects, communication latency and resource occupied significantly slow down the distributed training process.



IMPLEMENTATION

- Observations in Sync model training:
  - Many workers involved in communication during training of subnetworks: Heavier straggler effect
  - Long waiting time during evaluation and model adjusting: Waste of resources



RESULTS

Adam Optimizer with learning rate of 0.001, experimented on TonY with 20 workers, each one 32 GB RAM

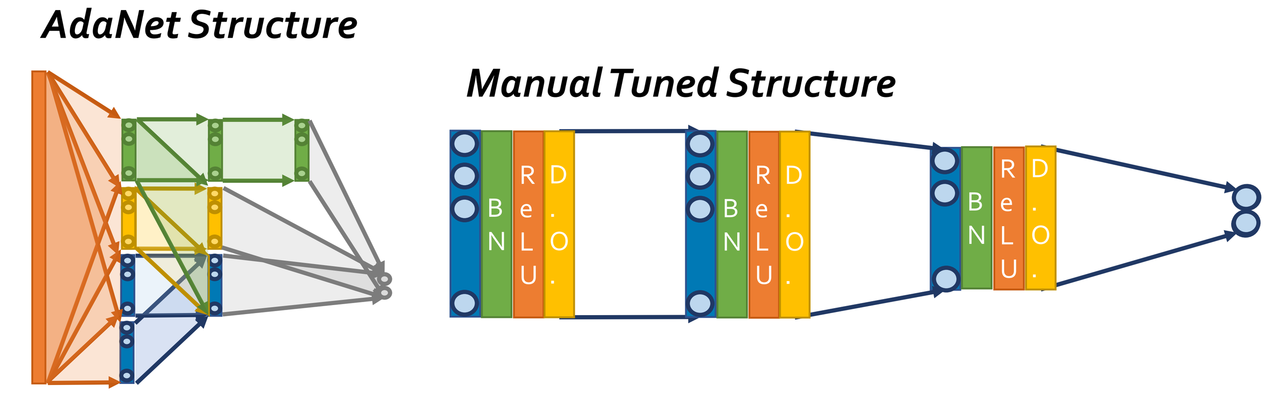
On LinkedIn pymk(people you may know) dataset

Model	Steps	Training Time	Eval Time	AUC
Logist Reg	20,000	9 mins	3 mins	73.1%
Grid Search	20,000	18 mins **(x8)**	6 mins	74.8%
AdaNet	10,000 x2	17 mins	4 mins	76.1%

On LinkedIn jymbii(job you may be interested in) dataset

Model	Steps	Training Time	Eval Time	AUC
Logist Reg	4,000	6 mins	19 mins	63.3%
Logist Reg	120,232	105 mins	20 mins	66.3%
Grid Search	4,000	8 mins	28 mins	64.5%
Grid Search	120,232	155 mins **(x12)**	31 mins	67.2%
AdaNet	2,000 x2	36 mins	31 mins	67.6%

AdaNet Structure ***Manual Tuned Structure***



REFERENCE

1 Cortes, Corinna, et al. "Adanet: Adaptive structural learning of artificial neural networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.
2 Machine Learning-Volume 70. JMLR. org, 2017. Frazier, Peter I. "A tutorial on bayesian optimization." *arXiv preprint arXiv:1807.02811 (2018).*

2019 LINKEDIN SUMMER INTERN PROJECT FAIR MOUNTAIN VIEW