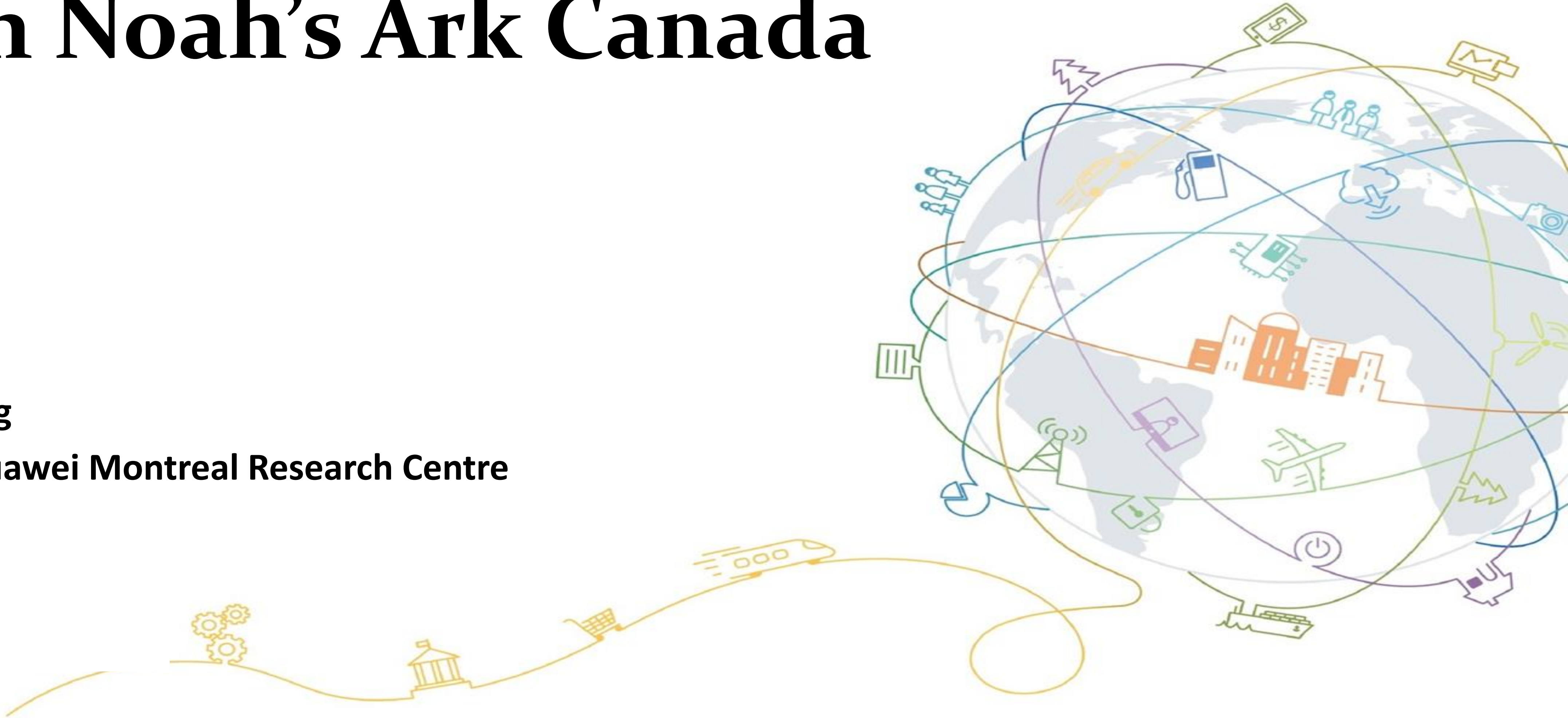
AI in Noah's Ark Canada

Yanhui Geng

Director, Huawei Montreal Research Centre



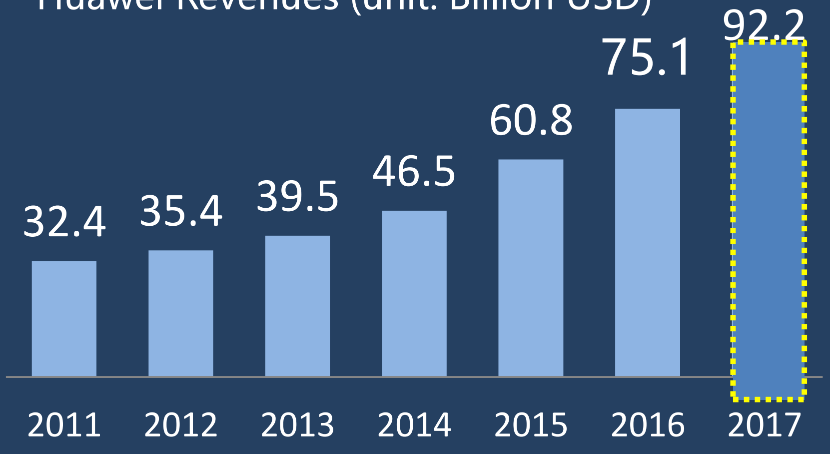
Outline

- Company overview and products
- Introduction to Noah’s Ark Lab
- Huawei Canada
- Huawei Montreal
  - NLP
  - ANT
  - NetMind

Huawei Corporate Overview

180, 000 Employees	80,000 R&D employees	170+ Countries	15 R&D centers	No. 70 Interbrand's Top 100 Best Global Brands	No. 83 Fortune Global 500

Huawei Revenues (unit: Billion USD)



Carrier ↗24%

- Global **NO.1**
- Tech. pioneer on 5G, IoT

Enterprise ↗ 42%

- Serving 197 of Fortune Global 500

Consumer ↗24%

- Brand awareness, 76% to 81%
- Shipment: 139 million, ↗ 29%

*Average annual growth rate in last 5 years

World-Wide Recognition

No.70 in Interbrand's Top 100 Best Global Brands 2017

Rank in 2017	Company
1	VOLKSWAGEN
2	ALPHABET
3	MICROSOFT
4	SAMSUNG
5	INTEL
6	HUAWEI
7	APPLE
8	ROCHE
9	JOHNSON & JOHNSON
10	NOVARTIS

Top 10 in the 2017 EU Industrial R&D Investment Scoreboard

LinkedIn China's Most In-Demand Employers 2017

Top 10 of 50 Smartest Companies by ‘MIT Technology Review’

Our products

Introduction to Noah's Ark Lab

From Big Data to Deep Knowledge

HUAWEI CANADA

Globalized Positioning & Localized Research

Global AI Capability Centers:

China: Computer Vision, Deep Learning, Reinforcement Learning, Decision Making & Reasoning, Natural Language Processing, AI Theory, Recommendation & Search

North America & Europe: Deep Learning, Reinforcement Learning, Decision Making & Reasoning, Natural Language Processing, AI Theory, Computer Vision, Human-machine Interaction

Huawei Noah's Ark Lab for AI Research

2012 Laboratory
(30,000+)

Network Intelligence

Enterprise Intelligence

Terminal Intelligence

Business Success

Noah's Ark Laboratory
(350+ patents)

Computer Vision

Natural Language Processing

Search & Recommendation

Decision & Reasoning

AI Theory

Advanced Technology

AI Research Collaboration

open source

START UP

Professional Advisory Committee

10+ Country, 25~Univiersity, 50~ projects, 1,000+ Researchers

Healthy Eco-system

Huawei Canada

In  :

> Artificial Intelligence [Montreal/Markham/Edmonton]

> Big data [Vancouver]

> Security [Waterloo]

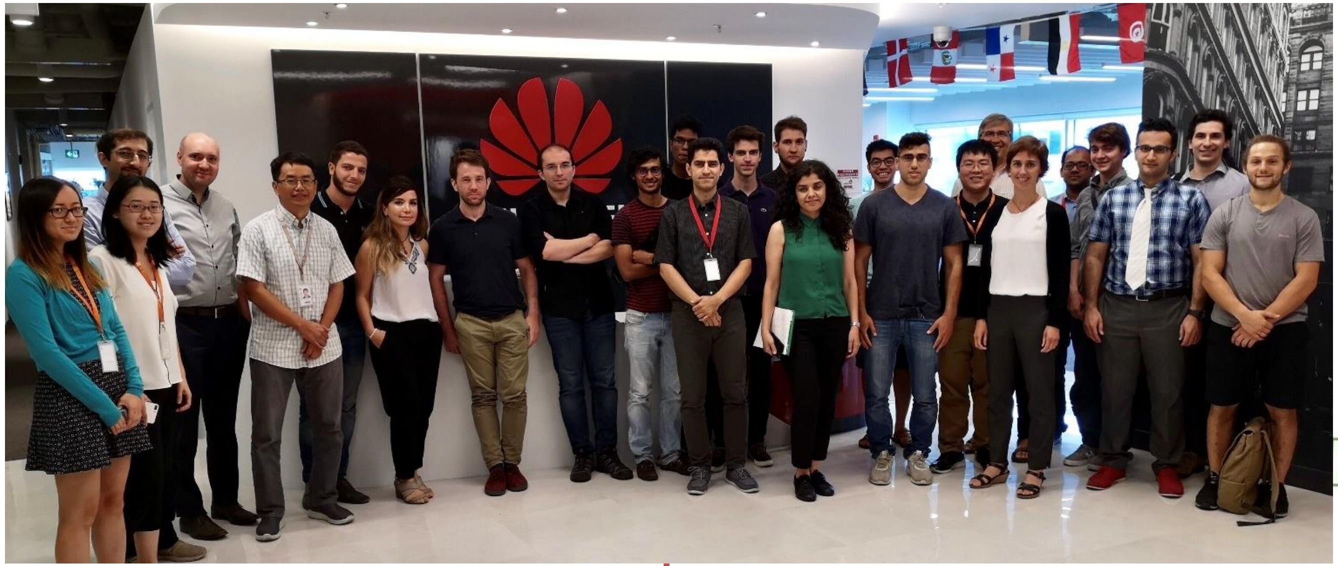
> 5G Research [Ottawa/Montreal]

> HiSilicon [Ottawa]

> Networking [Ottawa]

> Cloud Platform [Vancouver/Ottawa]

Montreal Research Centre (MRC)



NLP

ANT

NetMind

The mission of NLP Team in MRC

Since July 2017

University Collaborations

MILA

- Prof. Jackie Cheung
- Prof. Alain Tapp
- Dr. Jian Tang

McGill

- Prof. James J. Clark
- Dr. Jian Guo

University of Waterloo

- Prof. Pascal Poupart
- Prof. Ali Ghodsi

University of Montreal (UDM)

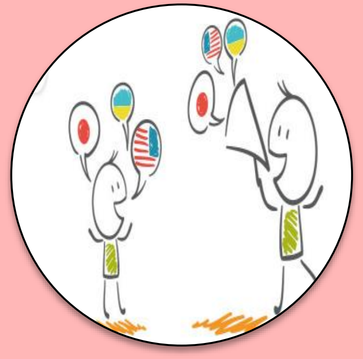
- Prof. Jian-Yun Nie

Active Projects for 2018

'Twas the night before Christmas, when all through the house, Not a creature was stirring, not even a mouse. And so begins the story of the day

Text Generation

- Improving code-based NTG Approaches
- Hybrid NTG approaches by combining code and text
- Conditional Text Generation



Bilingual GAN

- Unifying text generation and machine translation
- Working on code-based machine translation
- Co-training of code-based machine translation and text generation

Machine Translation

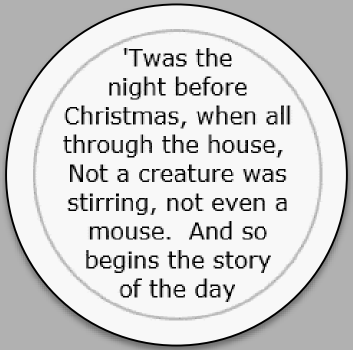
- Evaluating ConvSeq2Seq and Transformer techniques
- SPN for bidirectional machine translator
- Building a demo for machine translation

Text to Speech

- Generating pure speech using WaveRNN
- Text Embedding: implementing Char2Wave
- Text embedding and WaveRNN

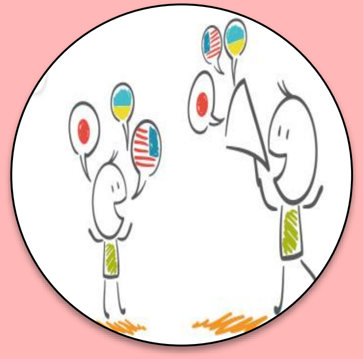
HUAWEI

Active Projects: Bilingual GAN



Text Generation

- Improving code-based NTG Approaches
- Hybrid NTG approaches by combining code and text
- Conditional Text Generation

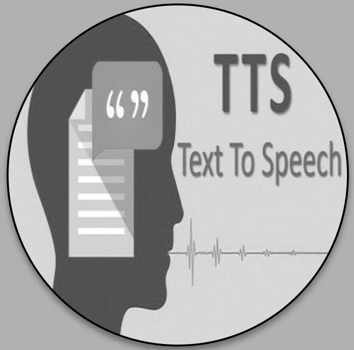


Bilingual GAN

- Unifying text generation and machine translation
- Working on code-based machine translation
- Co-training of code-based machine translation and text generation

Machine Translation

- Evaluating ConvSeq2Seq and Transformer techniques
- SPN for bidirectional machine translator
- Building a demo for machine translation



Text to Speech

- Generating pure speech using WaveRNN
- Text Embedding: implementing Char2Wave
- Text embedding and WaveRNN

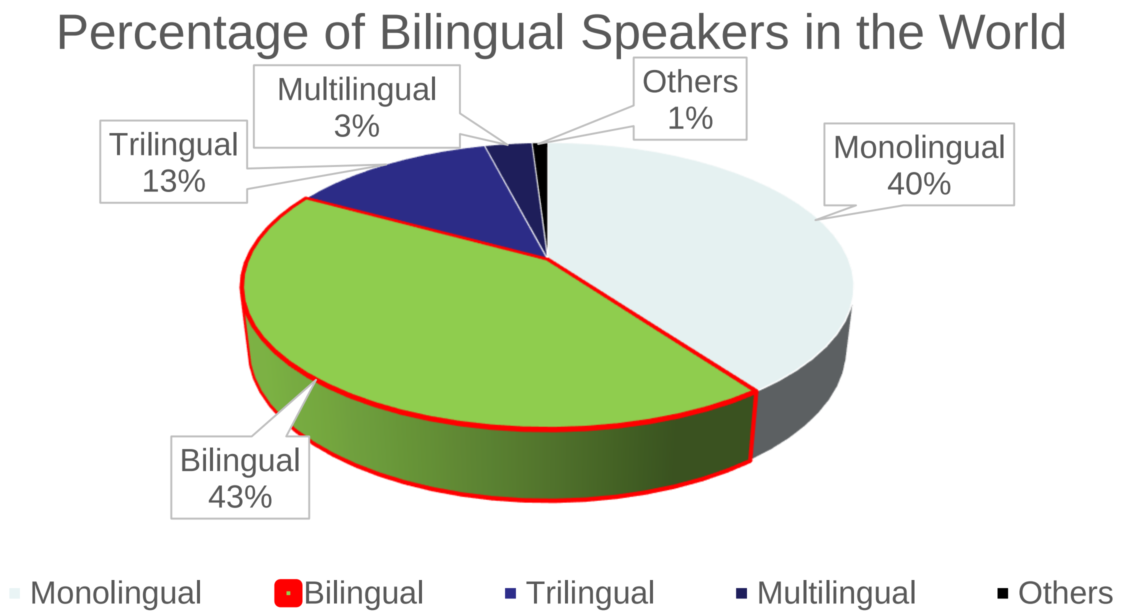
Motivation

Importance of Bilingualism:

☑ Speaking two languages improves brain efficiency and performance.

☑ One estimate puts the value of knowing a second language at up to $128,000 over 40 years **.

☑ Today, more of the world’s population is bilingual or multilingual than monolingual*.



Motivation

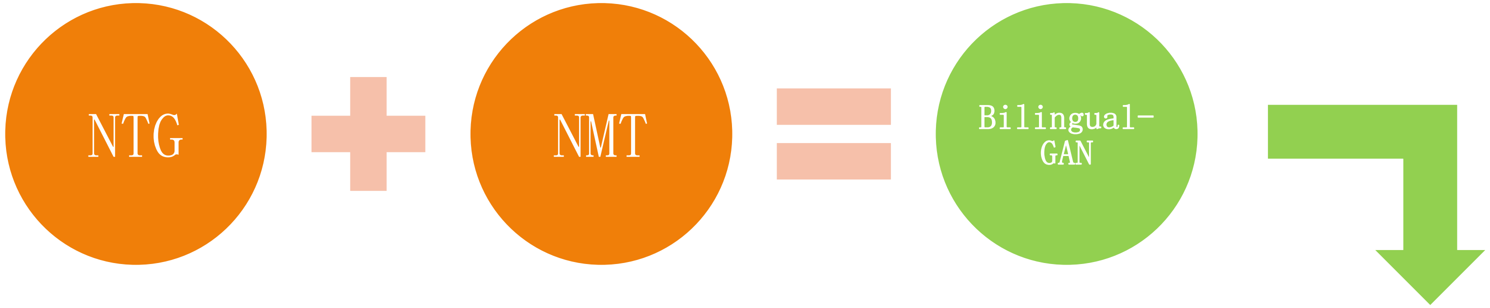
Real-Life Applications of NLP

Search and Recommendation	Chatbots	Voice Assistants	Others
YouTube	LivePerson	HiAssistant	Machine Translation
Google	LiveChat	Apple Siri	Image Captioning
AppGallery	Amazon Lex	Google Assistant	Question Answering
Google Playstore	Dialogflow	Amazon Alexa	Summarization
App Store	IBM Watson	Microsoft Cortana	Text-to-Speech

☒Most of these tasks can handle only one language at a time.
☒Most of these applications can deal with one task or one data type (e.g. text, image, speech) at a time.

Bilingual-GAN: Basic Concepts

- Currently, in the literature, neural text generation (NTG) and NMT techniques attempt to solve two independent problems;
- We believe that they are two sides of the same coin and can be integrated.

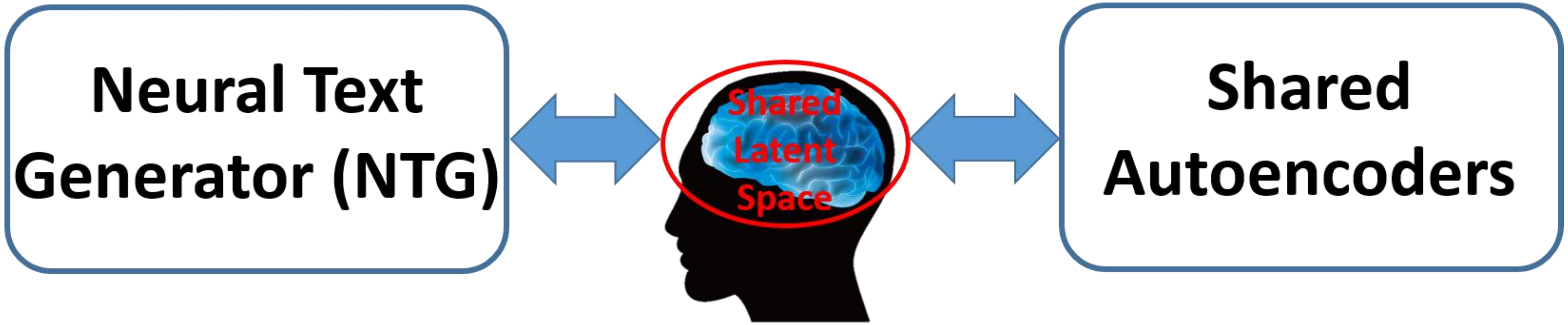


- Think in two languages equally well, or building a common space between two languages;
- Translate a sentence in language 1 into language 2 or vice versa,
- Express a concept in two different languages,
- Performing the task unsupervised/semi-supervised/supervised

Bilingual-GAN: Basic Concepts

Requirements of the Bilingual-GAN:

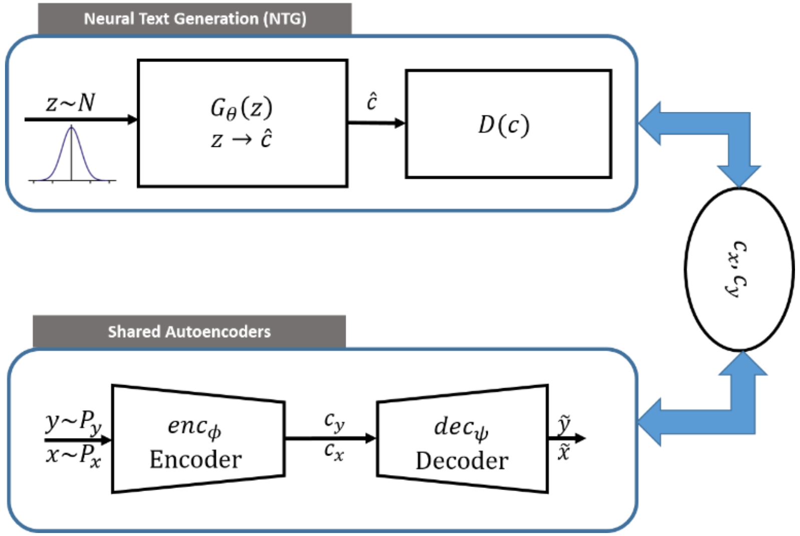
- (NTG & Shared AEs) → to derive a shared latent space between two languages
- (Shared AEs) → to derive the corresponding representation of the sentences in both languages in the shared latent space
- (NTG) → to be able to sample from this shared latent space for text generation



Bilingual-GAN: Basic Concepts

Requirements of the Bilingual-GAN:

- (NTG & Shared AEs) → to derive a shared latent space between two languages
- (Shared AEs) → to derive the corresponding representation of the sentences in both languages in the shared latent space
- (NTG) → to be able to sample from this shared latent space for text generation



Bilingual-GAN: Experimental Setup

Dataset	Europarl	Multi30K (Image Caption)
Training Samples	100K non-parallel	30K non-parallel
Max. Sentence Length	20	15
Vocab Size	8K	8K

Other Details:

- Padded shorter sentences and cut longer sentences
- Pre-trained the NMT module
- For each set of generated sentences used Google Translate to generate a ground truth and measured the parallelism between sentences using Translation BLEU score.

Bilingual-GAN: Results

- Generated Bilingual Sentences

Method	Task	Lang	Samples
Bilingual-GAN	Un-sup	EN FR	- that is what is the case of the european commission's unk. - c'est le cas qui suppose de la unk de la commission.
Bilingual-GAN	Un-sup	EN FR	- three people walking in a crowded city. - trois personnes marchant dans une rue animée.
Bilingual-GAN	Sup	EN FR	- mr president, i should like to thank mr unk for the report. - monsieur le président, je tiens à remercier tout particulièrement le rapporteur.
Bilingual-GAN	Sup	EN FR	- two people are sitting on a bench with the other people. - deux personnes sont assises sur un banc et de la mer.

Bilingual-GAN: Results

- To get an idea about how parallel the generated sentences are, we translate the (FR) sentences to (EN) using Google Translate.

Method	Task	Lang	Samples
Bilingual-GAN	Un-sup	EN FR	- that is what is the case of the european commission's unk. - c'est le cas qui suppose de la unk de la commission.
Google		FR→EN	- this is the case that assumes the commission's unk.
Bilingual-GAN	Un-sup	EN FR	- three people walking in a crowded city. - trois personnes marchant dans une rue animée.
Google		FR→EN	- three people walking on a busy street.
Bilingual-GAN	Sup	EN FR	- mr president, i should like to thank mr unk for the report. - monsieur le président, je tiens à remercier tout particulièrement le rapporteur.
Google		FR→EN	- mr president, i would like to thank the rapporteur in particular.
Bilingual-GAN	Sup	EN FR	- two people are sitting on a bench with the other people. - deux personnes sont assises sur un banc et de la mer.
Google		FR→EN	- two people sit on a bench and the sea.

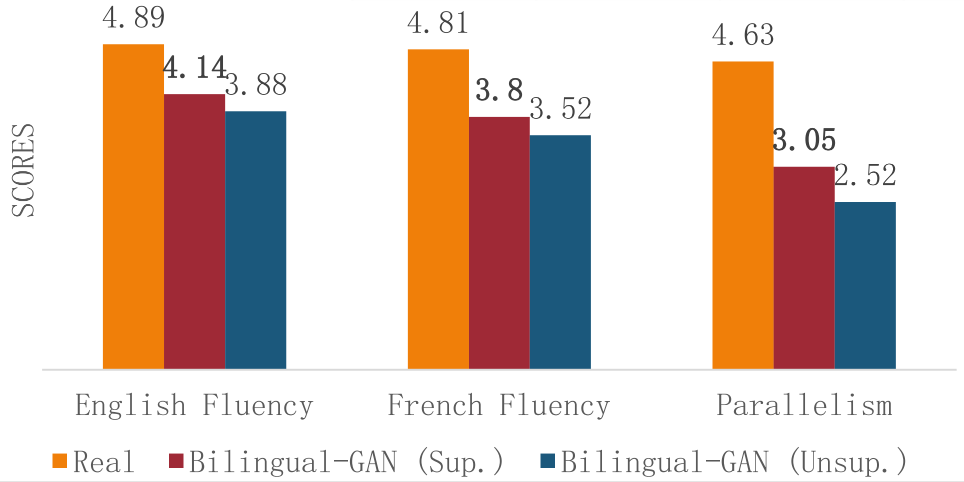
Bilingual-GAN: Results

- Quantitative Evaluation
  - Generation BLEU: The higher BLEU scores demonstrate that the GAN can generate fluent sentences both in English and French.

Table: **BLEU-4** score for the generation task

	English		French	
Dataset	Sup.	Unsup.	Sup.	Unsup.
Europarl	52.94	50.22	44.87	38.70
Multi30K	29.89	30.38	25.24	25.60

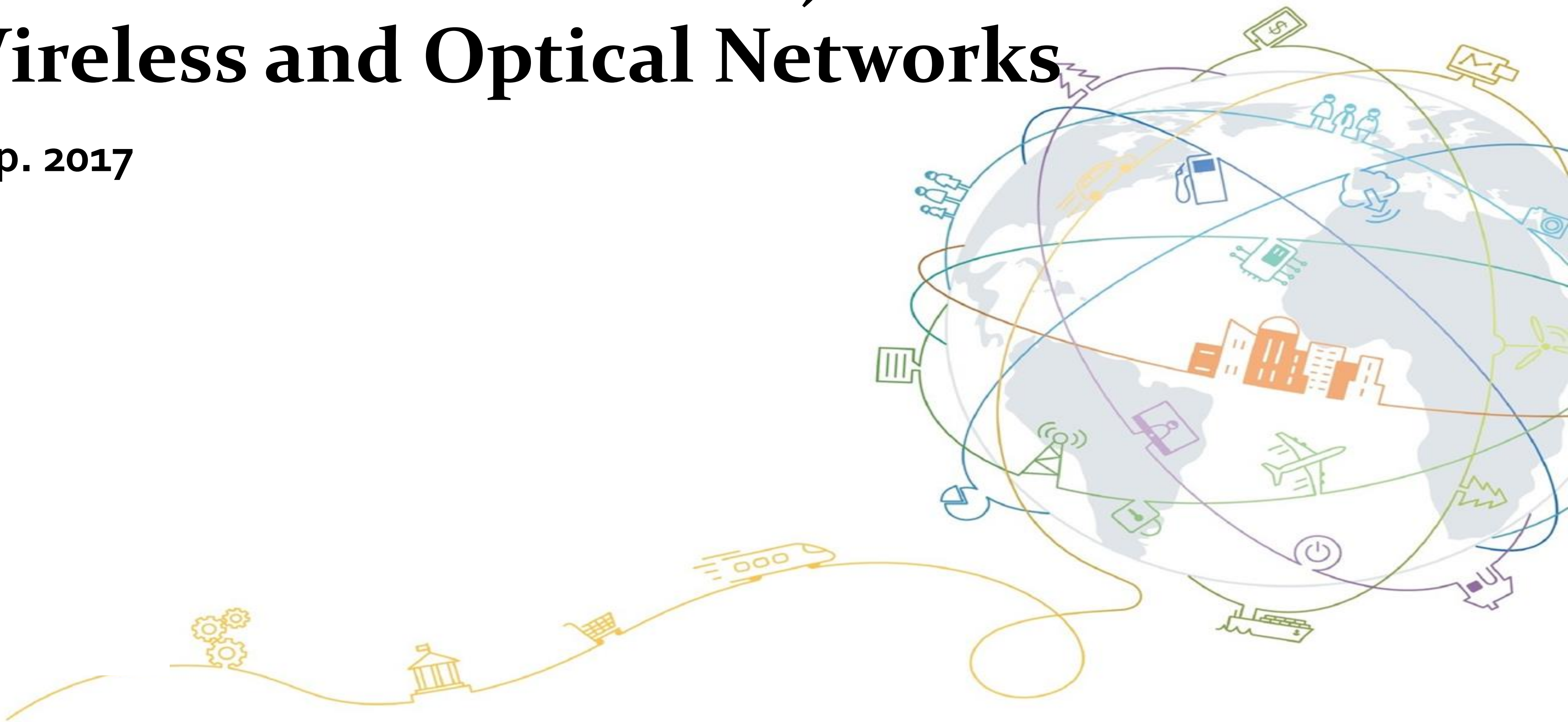
- Qualitative Evaluation (Human Evaluation)



Score	Fluency	Parallelism
5	Natural	Perfect
4	Understandable and semi-grammatical	Semantic preserved and some grammar
3	Understandable but Ungrammatical	Semantic preserved but ungrammatical
2	Semi-understandable	Part of semantic preserved
1	Gibberish	Unrelated

NetMind Research and Projects on Wireless and Optical Networks

Since Sep. 2017

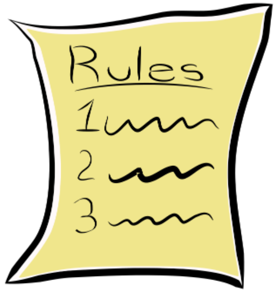


HUAWEI CANADA

Our Vision of Autonomous and Intelligent Network Control

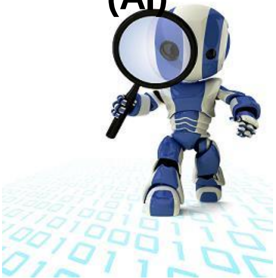
Vision: To help network operators control and optimize networks autonomously and intelligently, and provide better service to customers.

Rule-based control (experts)



AI assistant control (AI supervised by experts)

Automatic data-driven control (AI)



Policies generated by AI will be reviewed by experts. This feedback improves the system.

With sufficient data and confidence, AI will gradually take the control role.

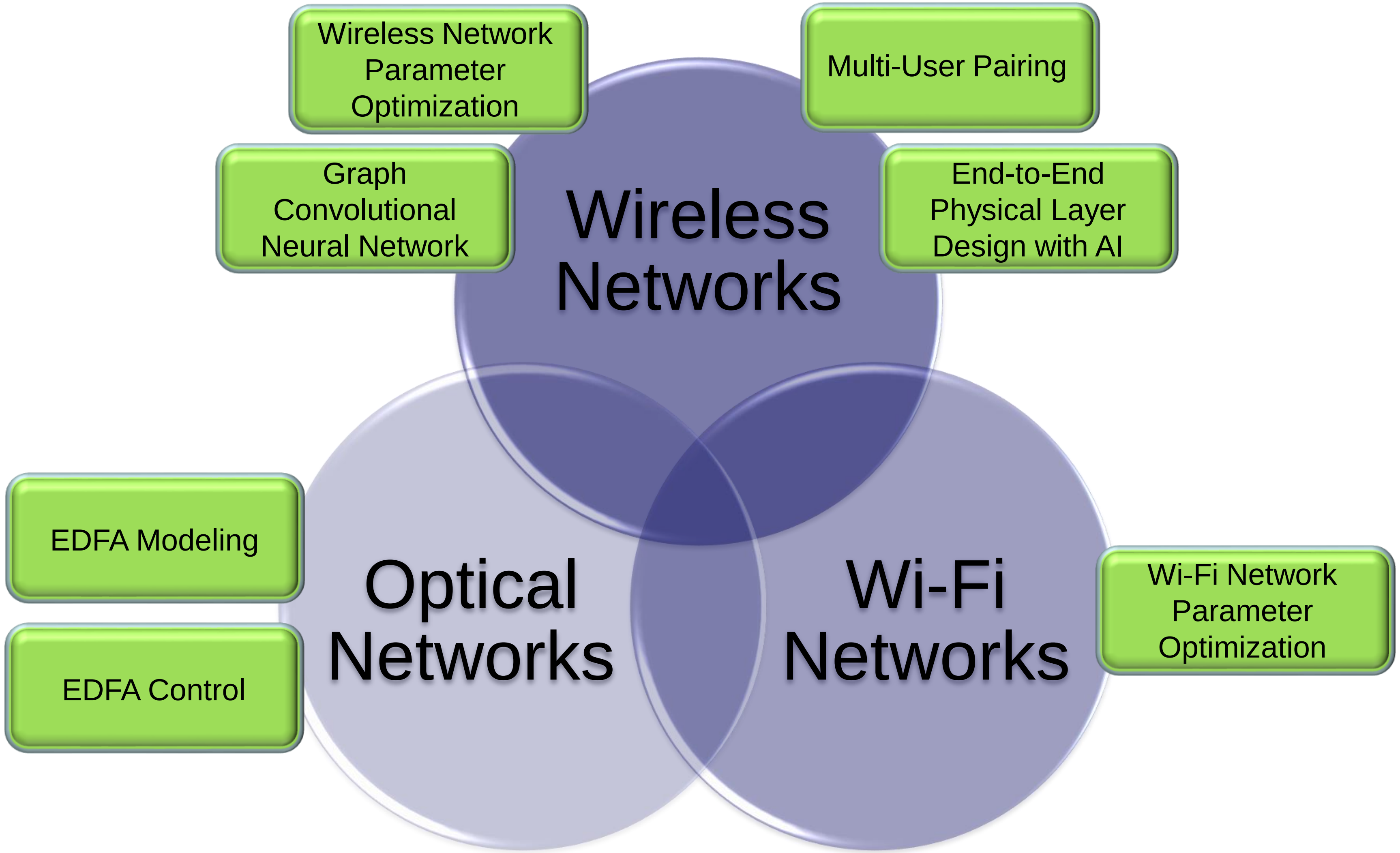
University Collaborations

McGill University

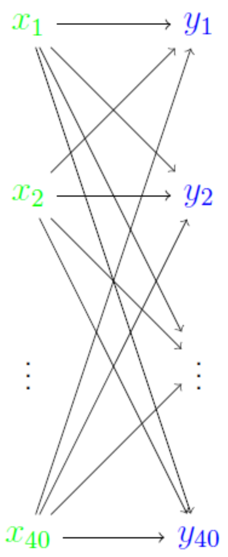
University of Waterloo

Polytechnique Montréal

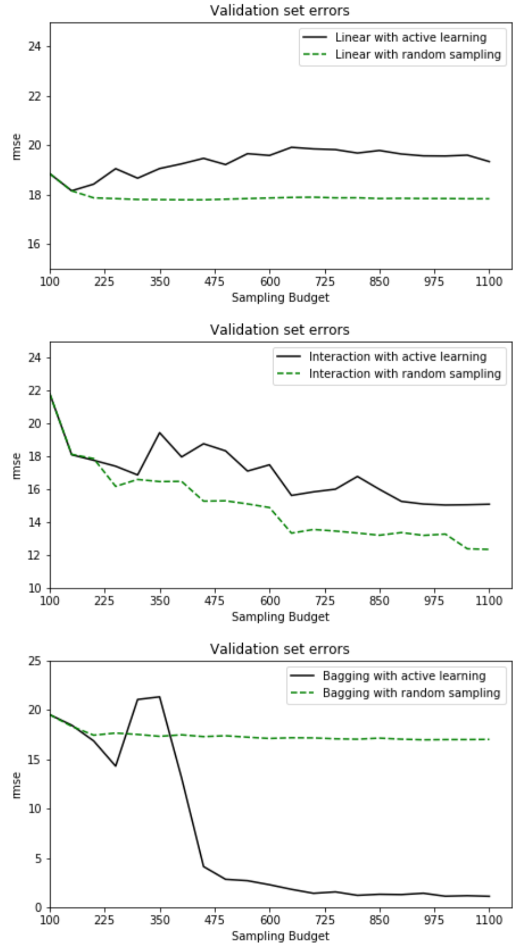
Network MIND (NetMind) Projects



EDFA Modeling (Optical Network)

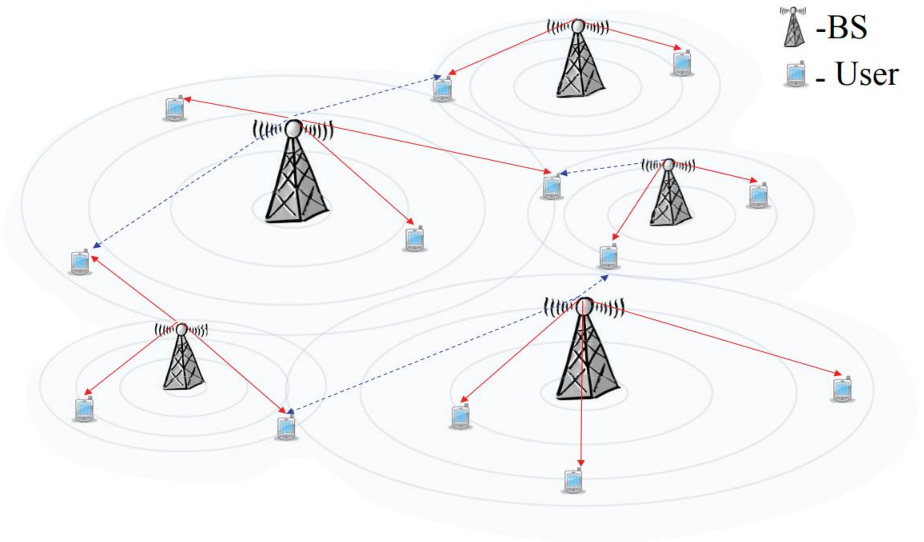


- **Problem:**
  - Optical signals fade away in long optical fibers, and need to be amplified for links longer than 20 Km distances,
    - ✓ Erbium-doped fiber amplifier (EDFA) is an **optical amplifier/repeater device**,
  - Highly accurate EDFA model is critical in order to:
    - ✓ Make network optimizer smarter,
    - ✓ Make resource allocation (EDFA control) more efficient.
    - ✓ Calculate OSNR, and predict path performance,
- **Challenge:**
  - The input space is very large (2^{40} ~ 2^{80}), we have little data (~10k data points), and labeling data is very expensive (requires human expertise).
- **Solution:**
  - **Active Learning** allows the learning algorithm decide which data points to query for label and to train on.
  - Different solutions with different accuracy vs runtimes.



Wireless Network Parameter Configuration

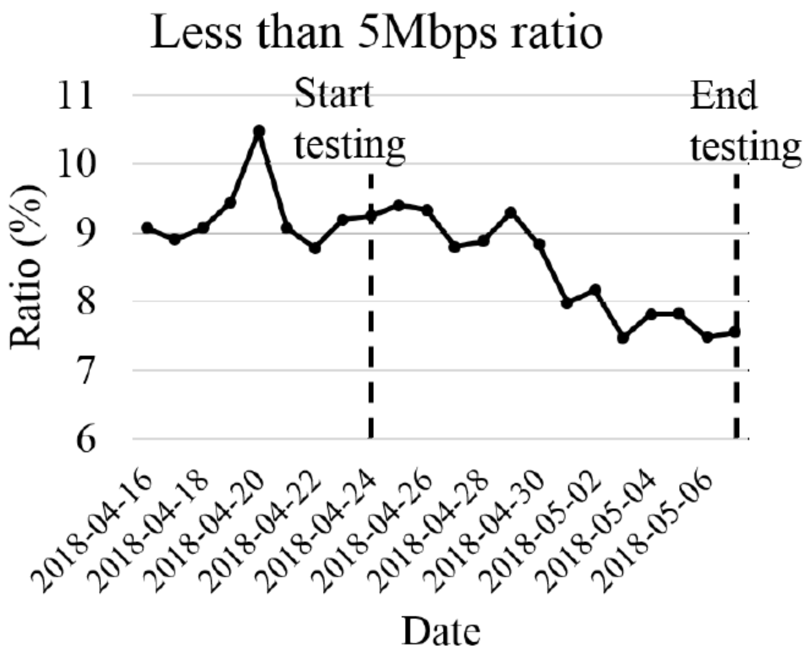
- **Problem:**
  - In a wireless cellular network there are **many parameters to configure** to improve network performance,
  - Currently the parameters are configured by experts but this process is time consuming, expensive and suboptimal,
- **Idea:**
  - **Use machine learning methods** to automate parameter configuration and improve network performance,
- **Challenge:**
  - Parameters should 1) adapt to network conditions, and 2) be cell-dependent,
  - We need a method that learns in real time with limited data (We usually have 2 weeks to learn how to configure)



Collaborators: Chen Zhitang and Chuai Jie

Wireless Network Parameter Configuration: Solutions

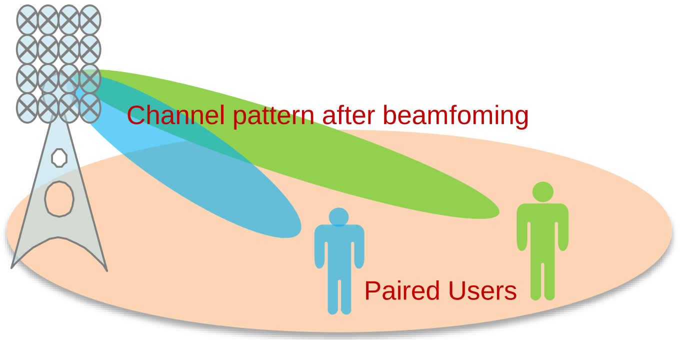
- **Solution:**
  - The solution is based on contextual multi-armed bandit and transfer learning,
  - The model for each cell combines two components; a common model for all cells, and a customized model for each cell (**Transfer Learning**),
  - We observed improved performance in several live tests,
  - The scope of the experiments are now increased to include joint optimization of multiple objectives for several parameters,
  - We are also working on solutions based on:
    1. Bayesian hierarchical modeling,
    2. Graph-based regularization to leverage topology,



20% performance improvement in the optimization period.

Multi-User Pairing

- **Problem:**
  - With increasing number of mobile users, more advanced radio resource management (RRM) techniques are required,
- **Idea:**
  - **Service multiple users on the same time/frequency pair**, i.e. multiplexing users by spatial domain,
- **Challenge:**
  - It has a combinatorial search space which is infeasible with large number of users and antennas,
  - Pairing must be performed almost real time, and calculating the device SINR and network capacity are not cheap,
- **Solution:**
  - Take advantage of state of the art sequence-to-sequence learning in DL and train the model using RL.



Collaborators: Liu Guochen and Chen Zhitang

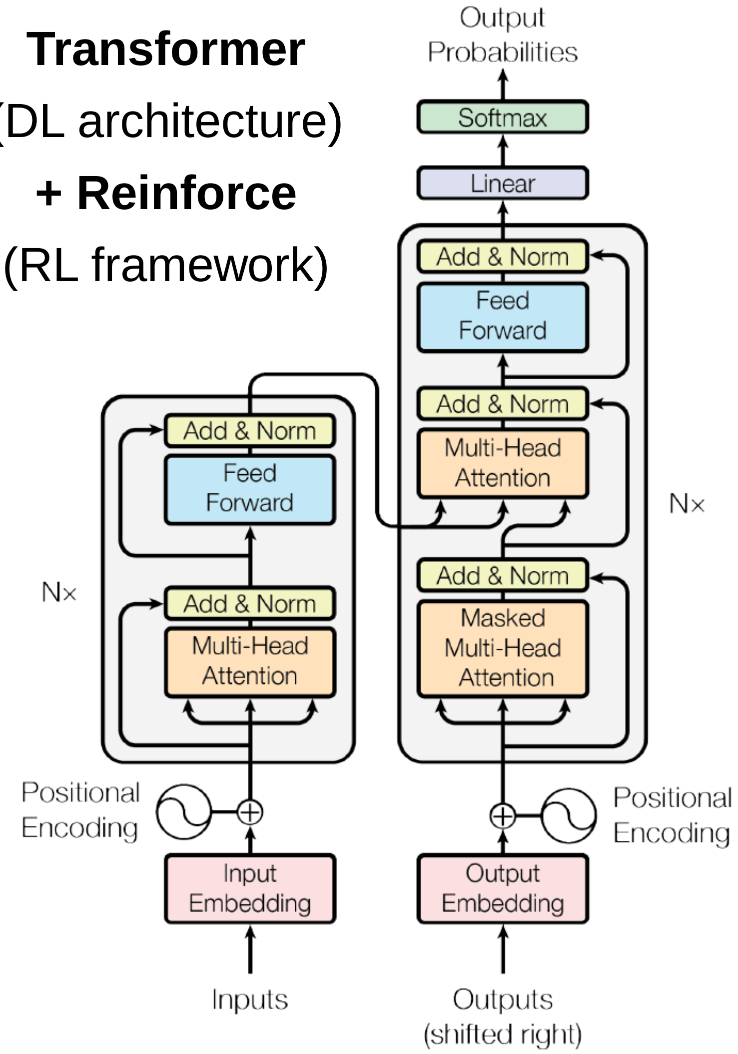
Multi-User Pairing: Solution & Results

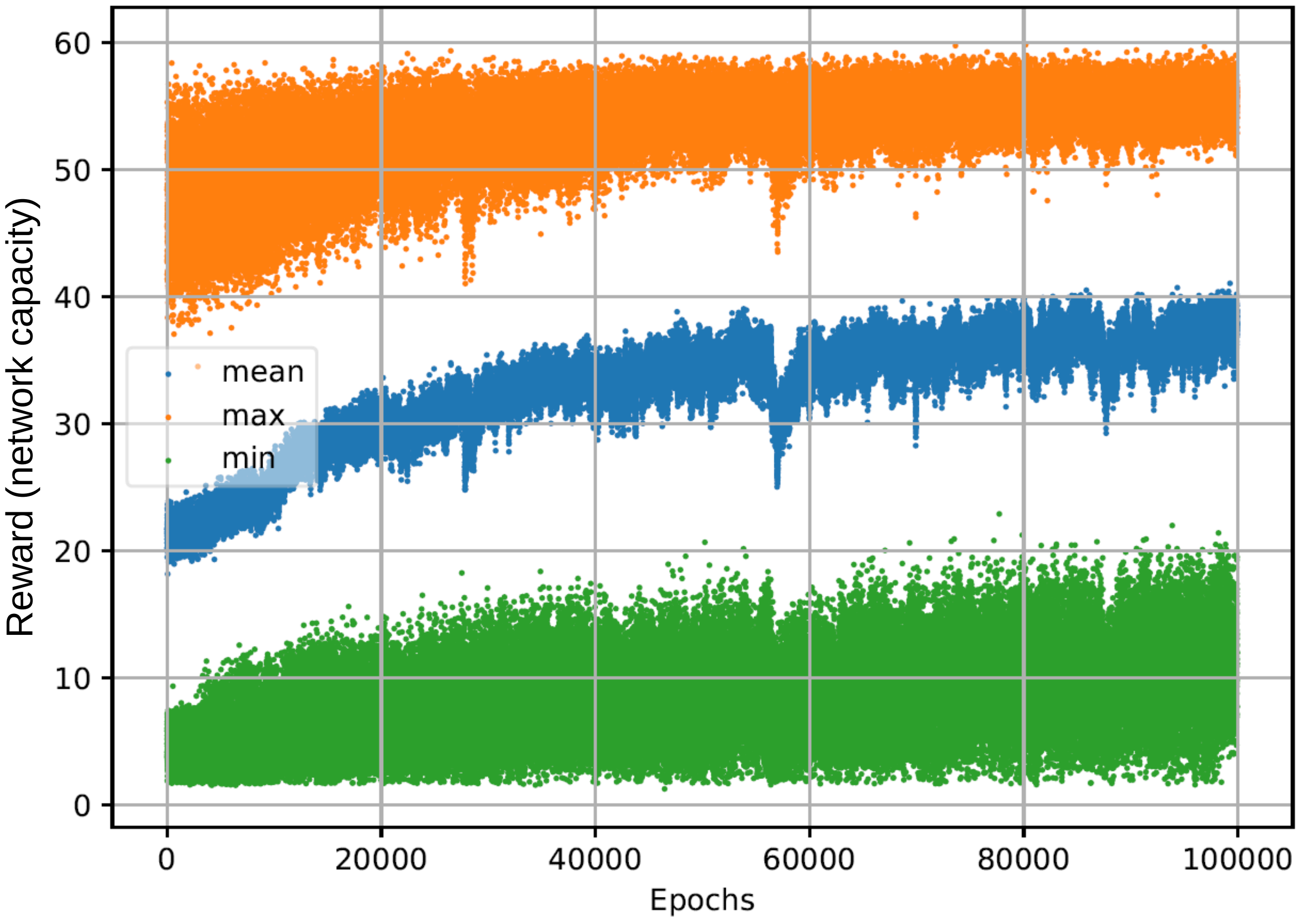
Transformer

(DL architecture)

+ Reinforce

(RL framework)

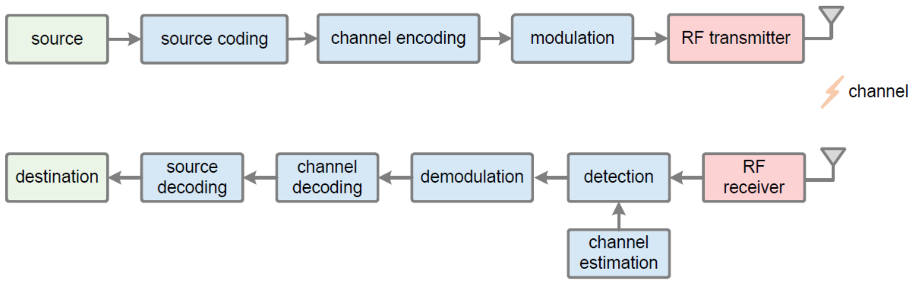




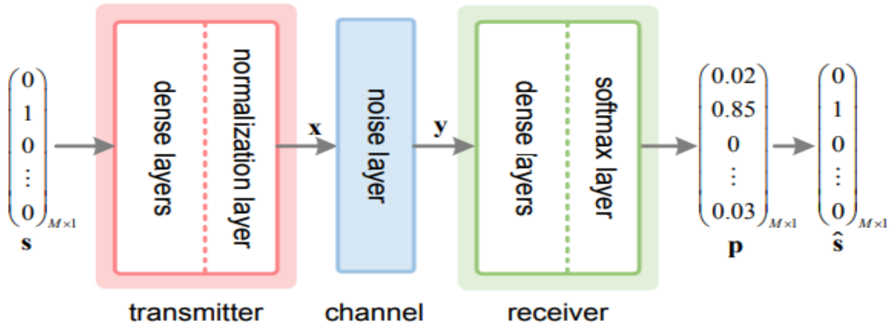
5% ~ 8% performance improvement compared to existing method in product line.

End-to-End Design of Wireless Physical Layer using AI

- **Problem:**
  - Sub-optimality in individual optimization of multiple processing blocks (source-coding, modulation, channel coding, …)
- **Idea**:
  - **Design the transmitter and receiver jointly end-to-end (E2E),**
  - NNs have shown superior results in end-to-end training, e.g. computer vision, language translation, dialogue systems, …
- **Challenges**:
  - The proposed solution should account for:
    1. Time-varying fading channels, and
    2. Large block size of transmitted codes,
- **Solutions:**
  1. Add SNR estimation or channel estimation or memory block to track time varying channel,
  2. Use LSTM AutoEncoders to break the complexity of encoding large block sizes.



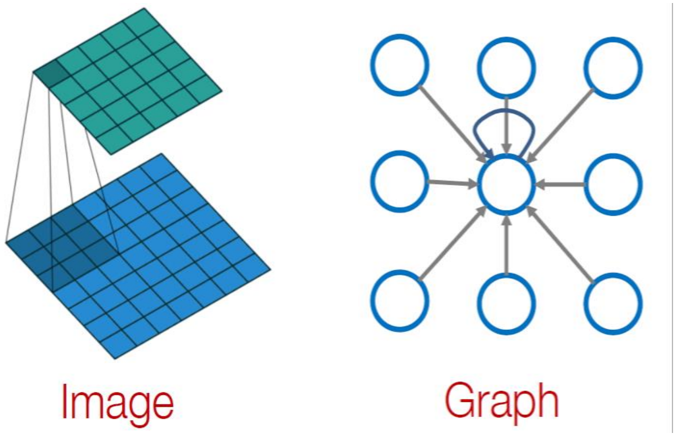
Traditional communication system



Alternative communication system, using NNs E2E

Graph Convolutional Neural Network (GCNN)

- **Objective:**
  - **Generalize CNN operations to irregular graphs** to apply to real data (telecommunication networks, web graph, social networks, etc.),
- **Current solution:**
  - Aggregate node features and graph structure (topology) information efficiently,
- **Proposed solution:**
  - Introduce a **Bayesian framework for the GCNN methods,**
  - It considers each observed graph as a realization from a parametric family of graphs. This resolves issues such as:
    i. Overfitting,
    ii. Sensitivity to erroneous links,
    iii. Uncertainty can be incorporated.
  - Target inference of the joint posterior of the random graph parameters, weights in the GCNN and the node (or graph) labels.



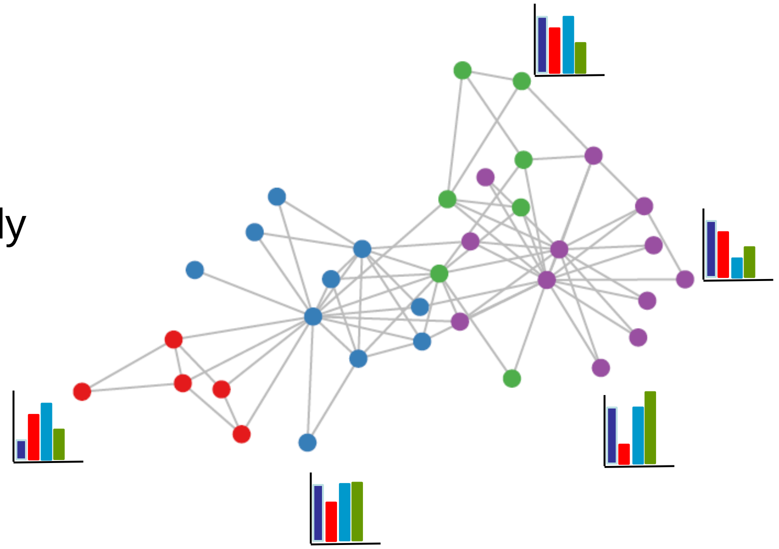


Image source: Jure Leskovec

GCNN: Experiment Results

Random split	5 labels	10 labels	20 labels
ChebyNet	58.5±4.8	65.8±2.7	67.6±1.9
GCNN	57.9±4.9	65.3±2.6	68.2±2.2
GAT	56.6±5.1	64.1±3.3	67.7±2.3
Bayesian ChebyNet	64.0±4.4	68.5±2.1	69.3±1.6
Bayesian GCN	**64.3±4.7**	**69.9±2.3**	**71.2±1.9**
Fixed split			
ChebyNet	53.0±1.9	67.7±1.2	70.2±1.0
GCNN	55.1±1.5	66.4±1.1	70.8±0.6
GAT	55.4±2.5	66.2±1.6	70.9±1.0
Bayesian ChebyNet	57.7±5.4	68.5±1.3	71.2±0.7
Bayesian GCN	**57.4±1.1**	**70.7±0.8**	**72.3±0.5**

Table: Prediction accuracy (percentage of correctly predicted labels) for Citeseer dataset.

	No attack	Random attack
Accuracy		
GCNN	88.5%	43.0%
Bayesian GCNN	87.0%	66.5%
Classifier margin		
GCNN	0.448	0.014
Bayesian GCNN	0.507	0.335

Table: Comparison of accuracy and classifier margins for the no attack and random attack scenarios on the Citeseer dataset.

- **Future Research Directions:**
  - Explore other graph generation algorithm (GANs or GVAE based graph generation mode)
  - Explore the application of Bayesian-GCNN on other applications (Recommendation system, Wireless network, Wi-Fi network, etc.)

AI Research Topics of Interest to NetMind

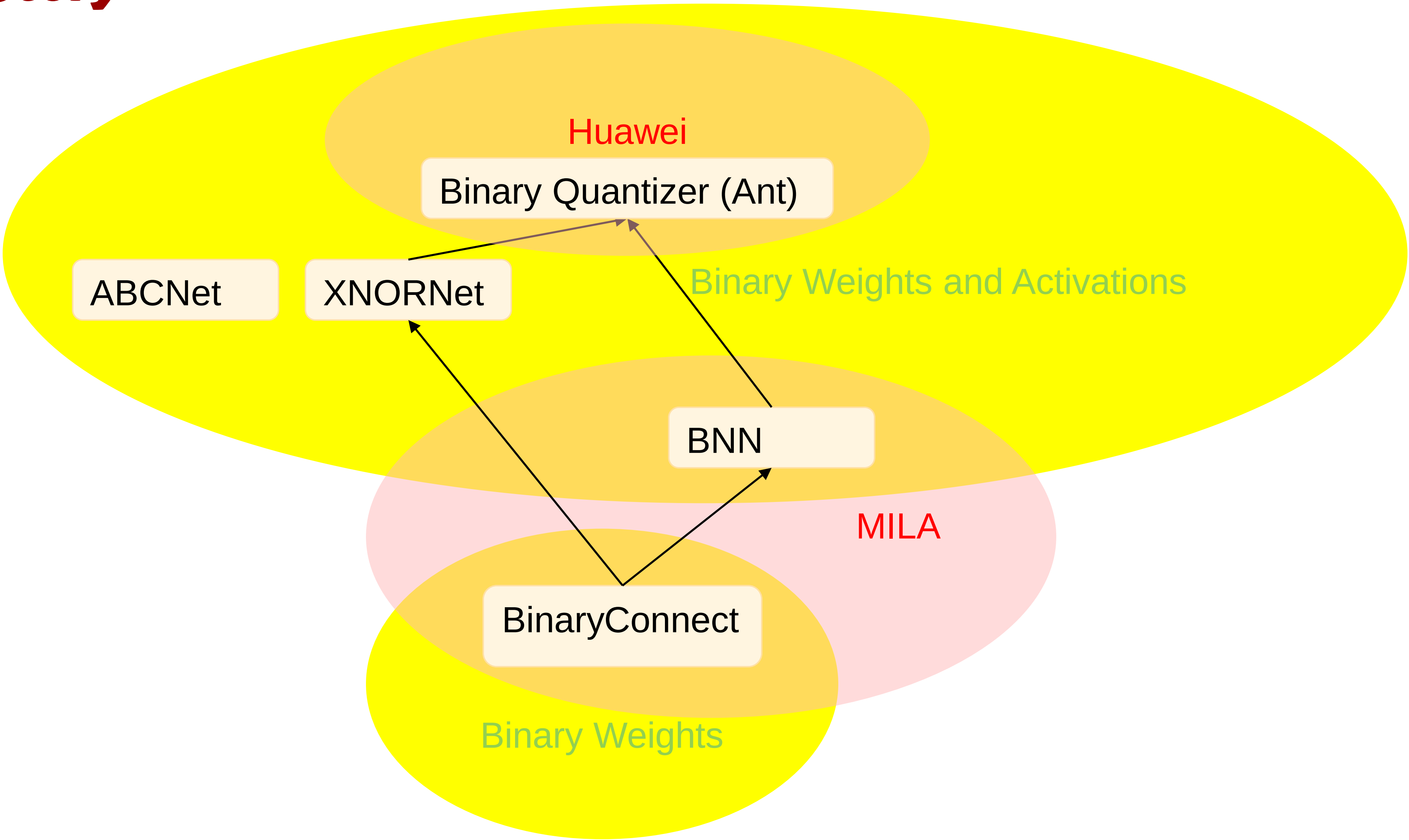
Deep Learning (DL)	• Wireless Network Parameter Optimization • Multi-User Pairing
Reinforcement Learning (RL)	• Multi-User Pairing
Graph Convolutional Neural Networks (GCNN)	• Wireless Network Parameter Optimization
Active Learning	• EDFA Modeling
Transfer Learning	• Wireless Network Parameter Optimization • EDFA Control

Accelerated Neural Technology (ANT 🐜)

Since June 2018

HUAWEI CANADA

Story


Huawei
Binary Quantizer (Ant)
ABCNet
XNORNet
Binary Weights and Activations
BNN
MILA
BinaryConnect
Binary Weights

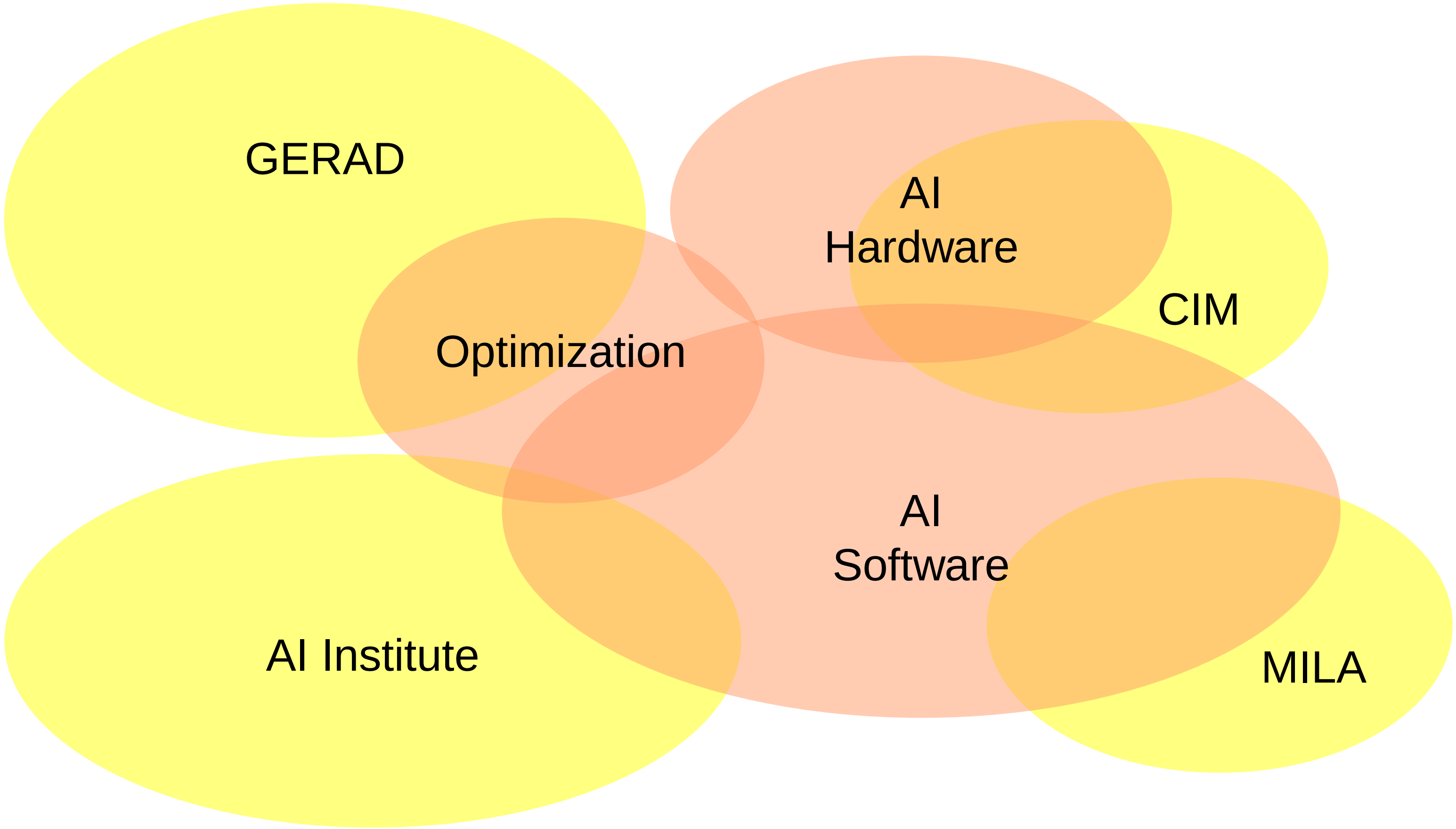
Why model compression is important

Surveillance Camera | Smart Watch | Cell Phone | Base Station | Autonomous Vehicles

Quantization | Pruning | Weight Sharing | Architecture Search

ARM CPU | FPGA | ASIC | GPU | Mobile GPU

Research Institutions



University collaboration

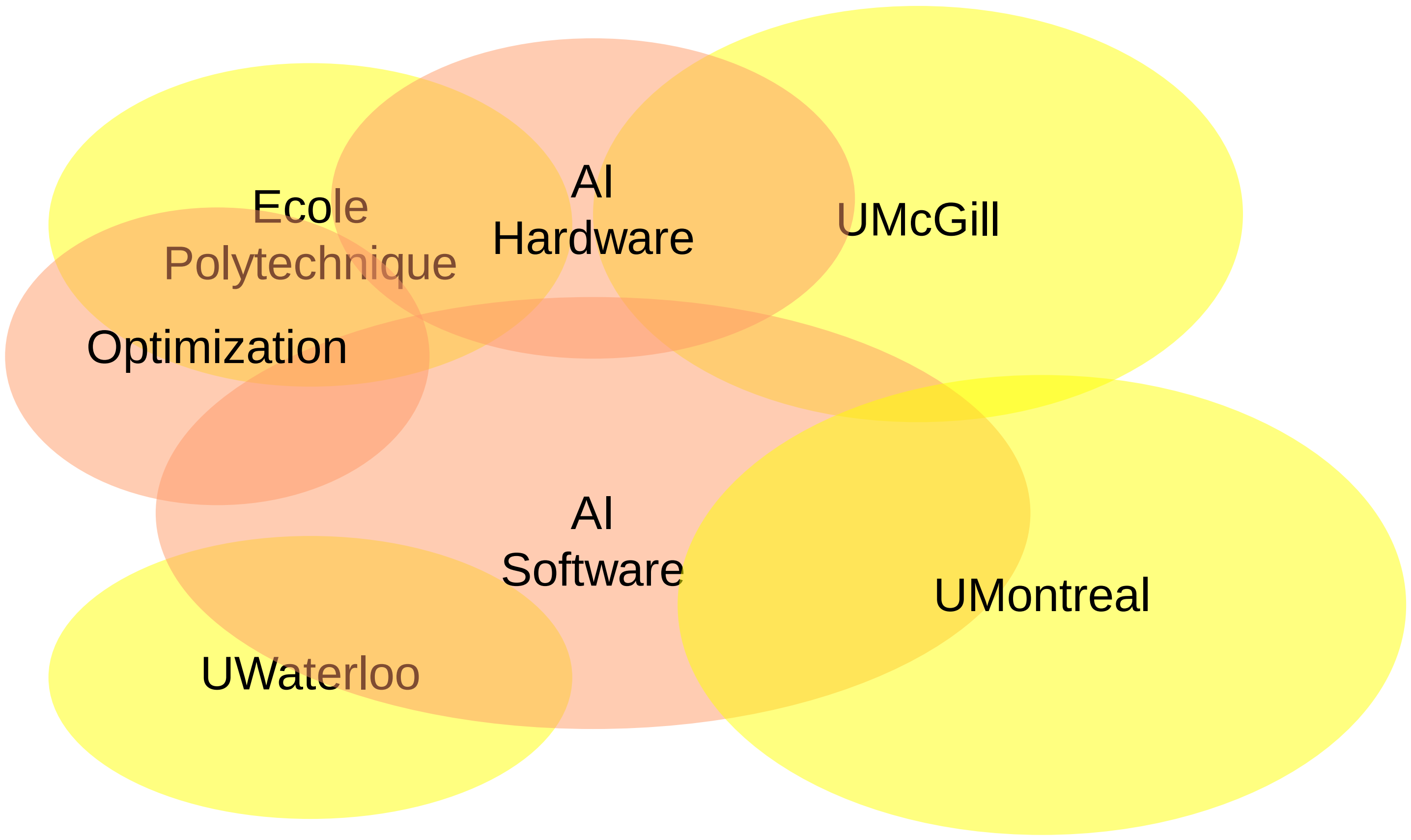
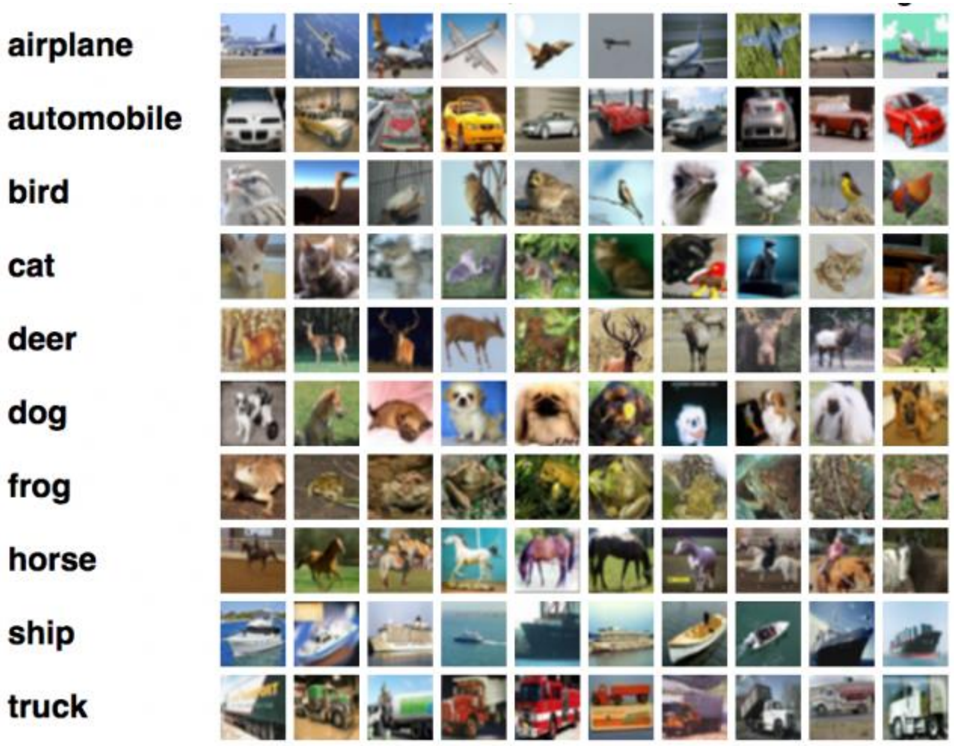


Image classification must work on benchmarks

CIFAR10



IMAGENET

Prediction Accuracy Loss in CIFAR-10

		Binary Quantizer	Full-Precision
AlexNet	Top-1	86.49%	88.58%
	Top-5	98.92%	99.73%
VGG	Top-1	90.89%	91.31%
	Top-5	99.09%	99.76%

Comparison with other binary networks on IMAGENET

Architecture: ResNet-18
Dataset: ImageNet (1000 classes)

	Full Precision	XNORNet	ABCNet (1 base)	BNN	Binary Quantizer
Top 1	69.3%	51.2%	42.7%	42.2%	53.0%
Top 5	89.2%	73.2%	67.5%	67.1%	72.6%
Computation Saving	1X	≈ 58X	≈ 58X	> 62X	> 62X
Memory Saving	1X	< 32X	< 32X	> 32X	> 32X

* Accuracy comparison under similar amount of computation cost

Thank you

www.huawei.com