Semantic Communication Meets Edge Intelligence

Dusit Tao Niyato

W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, **D. Niyato**, X. Chi, X. Cao, and K. B. Letaief, "Semantic Communication Meets Edge Intelligence," https://arxiv.org/abs/2202.06471

Introduction

- Computing force network
- AI-native network
- Ubiquitous connectivity
- Trustworthiness-native network

Edge intelligence

Semantic communication

- Holographic telepresence
- Hyper-intelligent IoT
- Unmanned aerial vehicles
- Collaborative robots

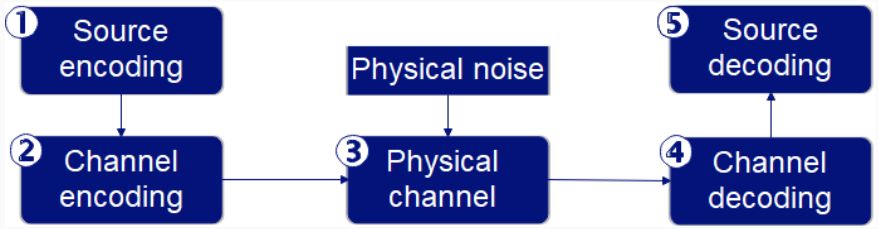
Fig: [1] Peltonen, Ella, et al. "6G white paper on edge intelligence." *arXiv preprint arXiv:2004.14850* (2020).

Introduction

Three-level Communication Model[2]

- LEVEL A: How accurately can the symbols of communication be transmitted? *(Technical problem)*
- LEVEL B: How precisely do the transmitted symbols convey the desired meaning? *(Semantic problem)*
- LEVEL C: How effectively does the received meaning affect conduct in the desired way? *(Effectiveness problem)*

Classical Communication system



- Pursue the replica of the source data
- Use the Shannon Theorem as a basis for system design
  - Information is defined as what can be used to remove uncertainty
  - link capacity is based on mutual information in the entropy domain

Network capacity

Broaden the available spectrum
Stack computation modules
Increase access point density
Increase antenna density

System Complexity

[2] C. E. Shanonn, "The mathematical theory of communication," Bell Sys. Tech. J., 1949.

Semantic Communication

Classical communication

Source image

Source encoding
Channel encoding

Channel

Source decoding
Channel decoding

Duplication of data

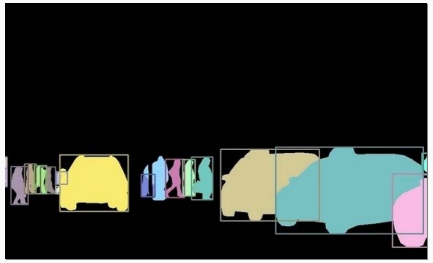
Semantic communication

Source data

Joint semantic channel encoding

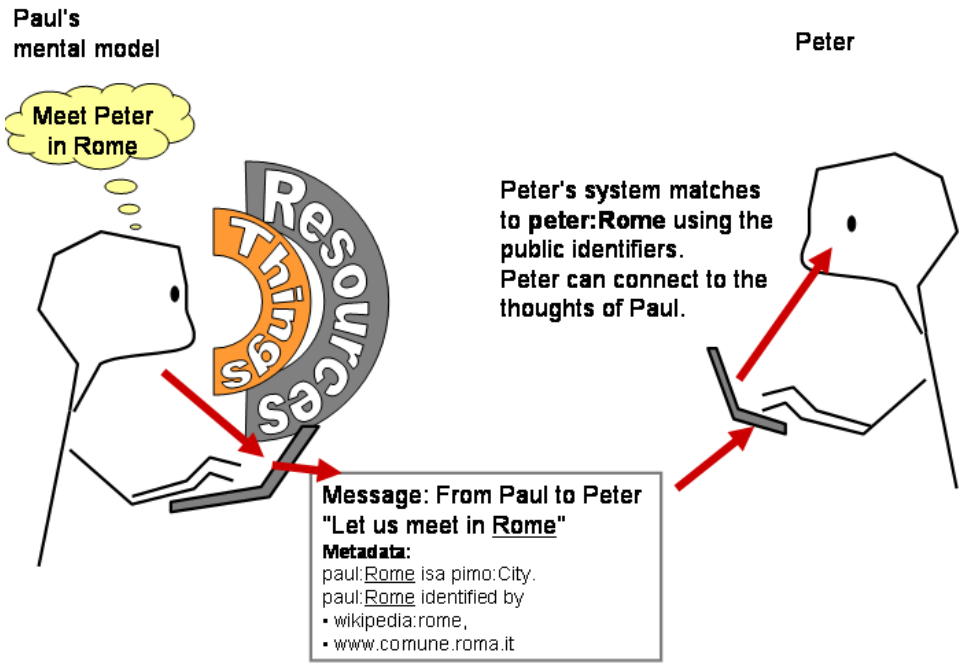
channel

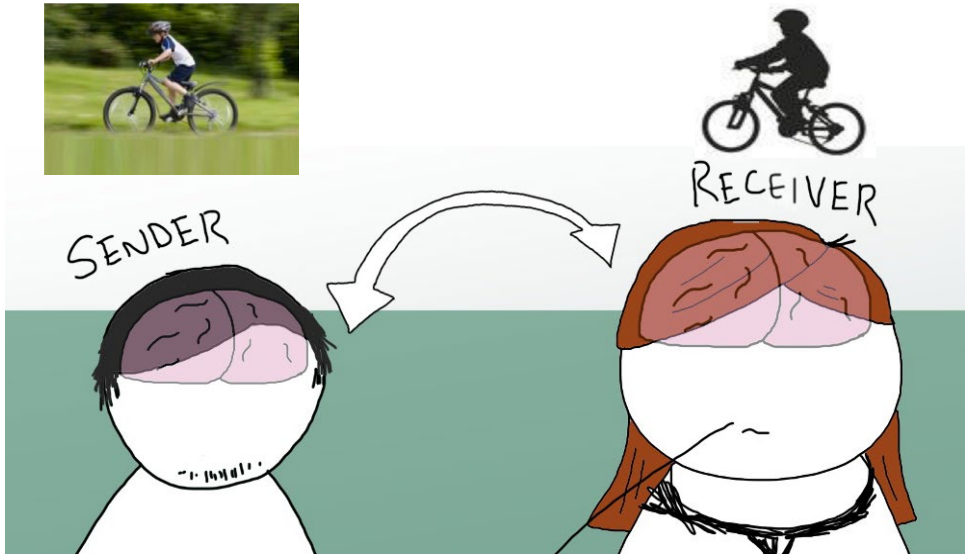
Joint semantic channel decoding



Recovered data

Semantic Communication





- Semantic communication allows the meaning of the message (behind digital bits) to be extracted and exploited during communication.
  - improving communication efficiency
  - providing human-oriented services

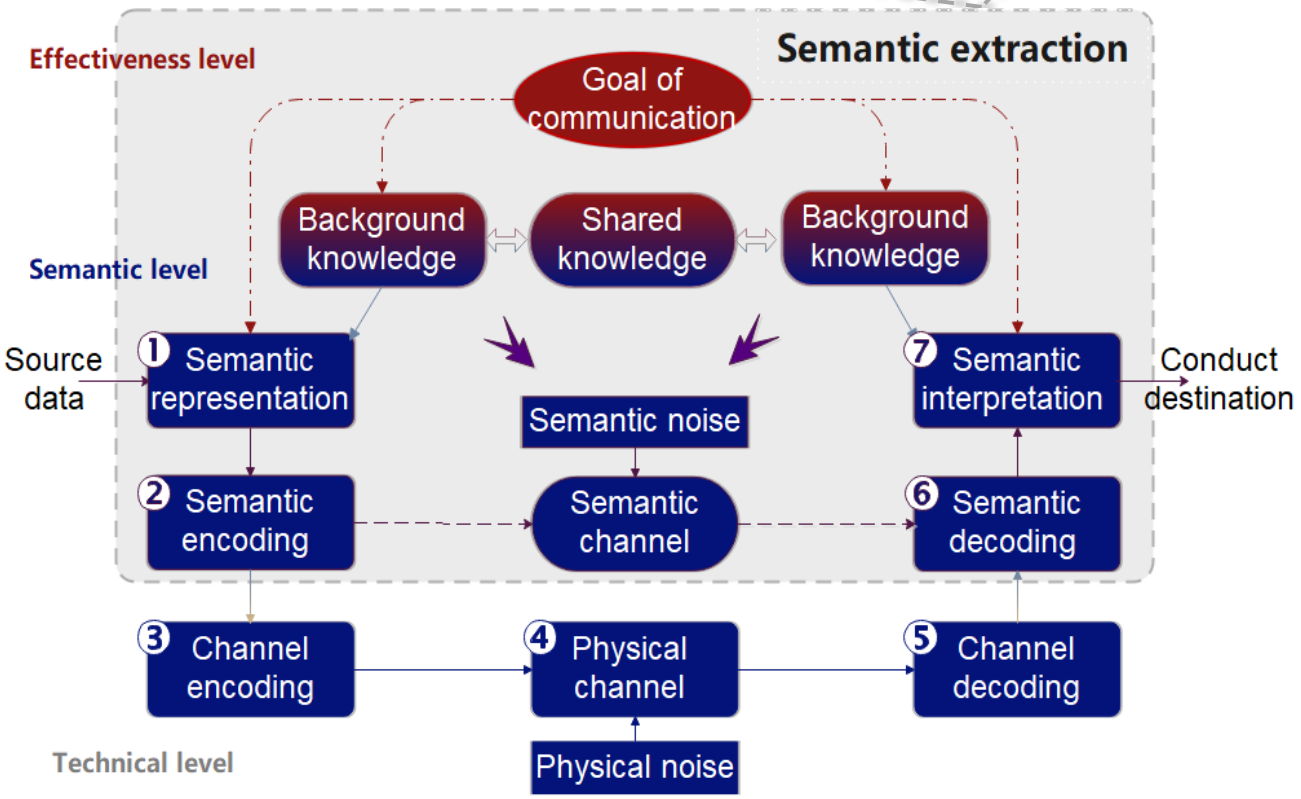
Semantic Communication Challenges

- Deploying semantic communications in wireless networks faces a number of technical challenges
  - How to model the *semantic information* of data?
  - How to evaluate the *performance* of the semantic communication?
  - How to design a *semantic-oriented resource allocation* scheme?
  - Other questions pertaining to semantic encoding/decoding and security of semantic communications.

Semantic Communications Framework

SemCom Model

SemCom differs from traditional Shannon communication in that, it incorporates human-like "understanding" and "inference" into the encoding and decoding of communication data.



[3] W. Weaver, "Recent contributions to the mathematical theory of communication,"ETC: a review of general semantics, pp. 261–281, 1953.

[4] Y. Yang, C. Guo, F. Liu, C. Liu, L. Sun, Q. Sun, and J. Chen, "Semantic communications with ai tasks," arXiv preprint arXiv:2109.14170, 2021.

- **Semantic representation module**

Extract the useful semantic information and remove the irrelevant information before transmission.

- **Semantic interpretation module**

Infer the intended meaning of the sender or the desired action to be performed by the receiver

- **Background knowledge**

Background knowledge of the communication parties has to be shared in real-time to ensure that the processes of representation and inference can be well-matched. Otherwise, semantic noise will be generated.

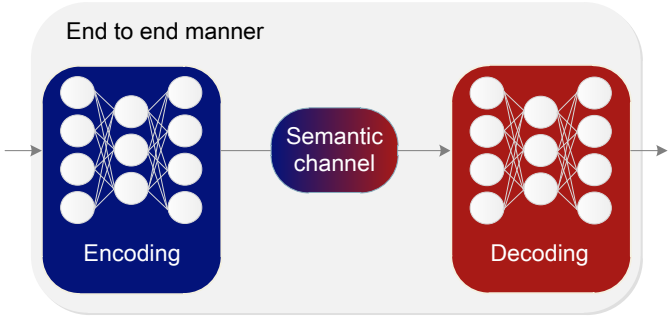
- **Communication goal**

In some cases like multi-objective identification, possibilities for communication goal should be included in the background knowledge and the communication goal should instruct semantic extraction to filter out irrelevant semantic information transmission according to the current communication goal.

* In general, semantic representation (semantic interpretation) and semantic encoding (semantic decoding) are combined into a single module in the practical system design.

General Semantic Extraction (SE) Methods

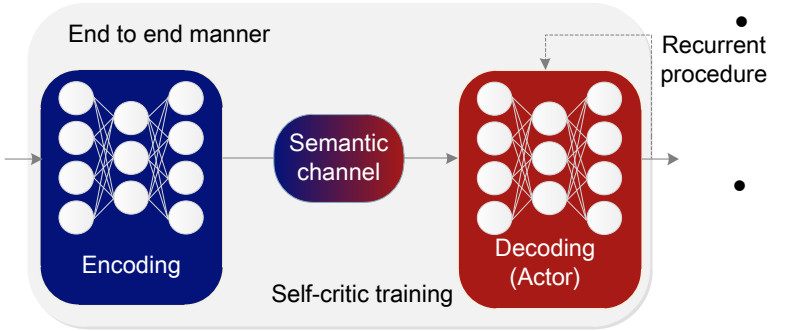
Deep Learning-based SE method [5]



- CNN, Transformer, attention mechanism. ResNet, GANs…
- Enhance the system robustness at low SNR with shorter bit-flow

- Cross Entropy (CE) and Mean Squared Error (MSE) are employed in training, as loss function is generally required to be differentiable.

Reinforcement Learning-based SE method[6]



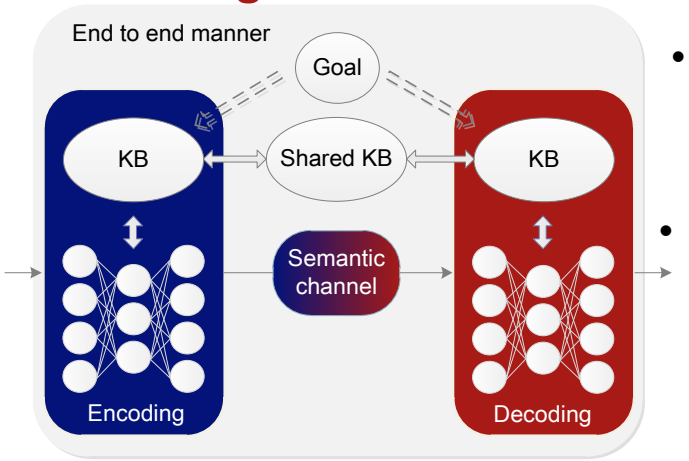
- Non-differentiable semantic metrics like BLEU into SE training
- Self-critic training is employed to address the issue of identifying the intermediate rewards

- For non-sequential tasks, the decoding process needs to be transformed into a recurrent procedure beforehand

[5] H. Xie and Z. Qin, "A Lite Distributed Semantic Communication System for Internet of Things," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 1, pp. 142-153, Jan. 2021
[6] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, “Reinforcement learning-powered semantic communication via semantic similarity,” arXiv preprint arXiv:2108.12121, 2021.

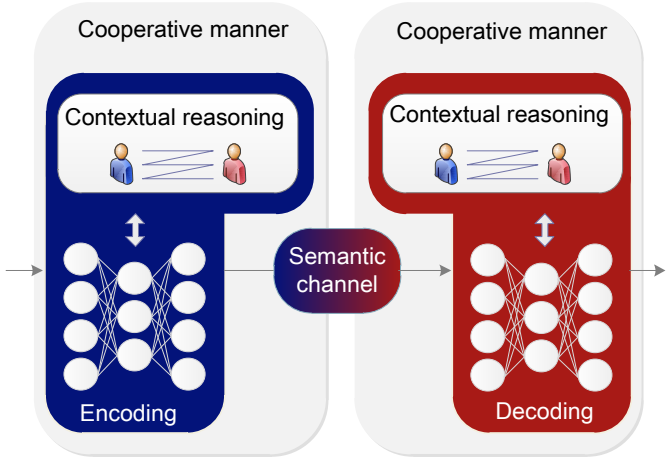
General Semantic Extraction (SE) Methods

Knowledge base-assisted SE method[7]



- Integrates the KB into the encoder and decoder, aiming to extract more relevant information for scenarios with multiple communication tasks.
- KB in Semcom is composed of source information, goals of the tasks, and the possible ways of reasoning that can be understood, and learned by all the communication participants

Semantic-native SE method [8]



- Semantic information can be learned from iterative communications between intelligent agents, which make it feasible to the cases where the semantics vary over time and in different contexts.
- Communication parties can be empowered with the capability of contextual reasoning to improve communication efficiency

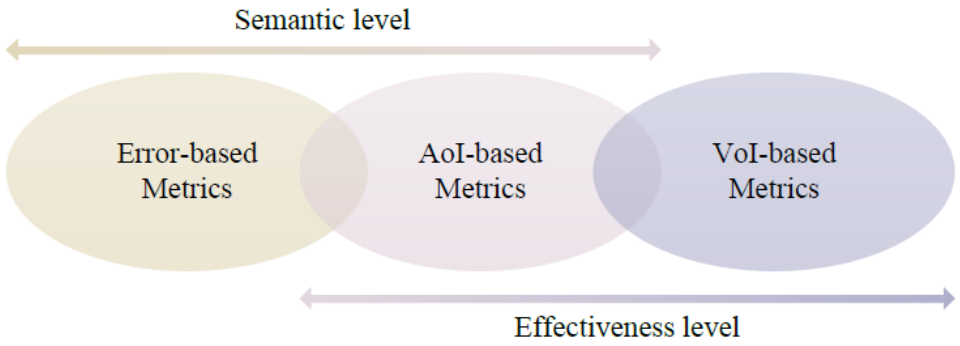
[7] Y. Yang, C. Guo, F. Liu, C. Liu, L. Sun, Q. Sun, and J. Chen, “Semantic communications with ai tasks,” arXiv preprint arXiv:2109.14170, 2021.
[8] H. Seo, J. Park, M. Bennis, and M. Debbah, “Semantics-native communication with contextual reasoning,” arXiv preprint arXiv:2108.05681, 2021.

Semantic metrics

Classical metrics for communication performance
(BER, SER, delay, throughput…)
- Evaluate performance from different network layers
- Treat every bit/symbol/packet… as equally important

Semantic metrics for communication performance
(Error-based, AoI-based, VoI-based…)
- Measure the performance from the semantic domain
- The semantic contribution of each packet is not equally important

Main types of semantic metrics



Error-based semantic metrics: Error-based semantic metrics are concerned with whether the destination can recover equivalent meaning from the received message to that in the transmitted message, measuring the differences in the meaning conveyed by the recovered data and source data.

Some error-based semantic metrics from NLP [6]

Semantic metrics	Advantages	Drawbacks
Bilingual evaluation understudy (BLEU)	It consider the linguistic laws, such as that semantically consistent words usually come together in a given corpus.	It only calculates the differences of words between two sentences and has no insight into the meaning of the whole sentence.
Consensus based Image Description Evaluation (CIDEr)	Compared to BLEU, it does not evaluate semantic similarity on the basis of a reference sentence, but a set of sentences with the same meaning.	Similar to BLUE, it is also based on the comparisons between word groups, and the semantic similarity can only be captured at the word level.
Sentence similarity	The semantic information in this metric is viewed from a sentence level owing to the sensitivity of BERT to polysemy, (e.g., word"mouse" in biology and machine).	The pre-trained BERT model introduces much resource consumption in the training process and makes it hard to generalize in other tasks.

[6] K. Lu, R. Li, X. Chen, Z. Zhao, and H. Zhang, "Reinforcement learning-powered semantic communication via semantic similarity," arXiv preprint arXiv:2108.12121, 2021.

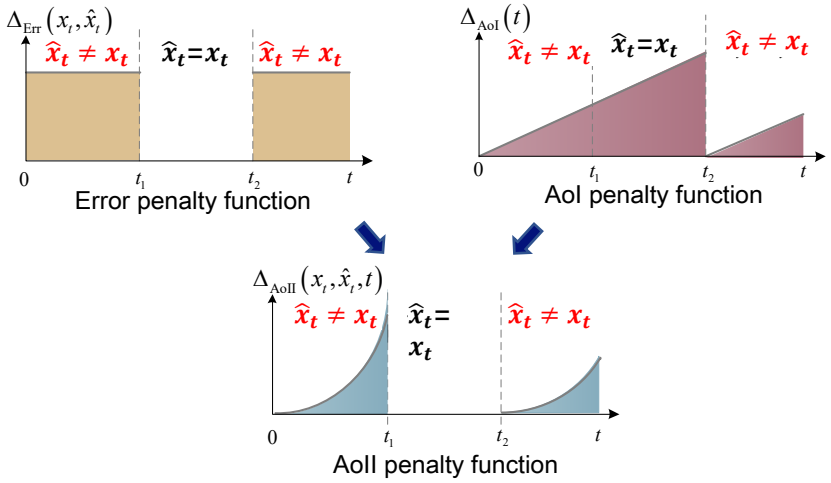
Semantic metrics

AoI-based semantic metrics

- The metric of delay primarily measures the transmission performance without concern for the content of the packets.
- AoI-based metrics are utilized to quantify the staleness of the information received at the destination.
- The scheduling schemes based on AoI minimization can highlight the importance of the freshness of the data packets and filter out the irrelevant or less important packets given the bandwidth constraints.

Combined semantic metrics

- AoII: integrate AoI into error-based metrics[9]



Age of Incorrect Information (AoII)

Age of Information at Query (QAoI)

- QAoI: integrate VoI into AoI-based metrics[10]

In a pull-based system the valuable information

⬇

the information only at certain query instants.

⬇

QAoI reflects the freshness in the instants when the receiver actually needs the data

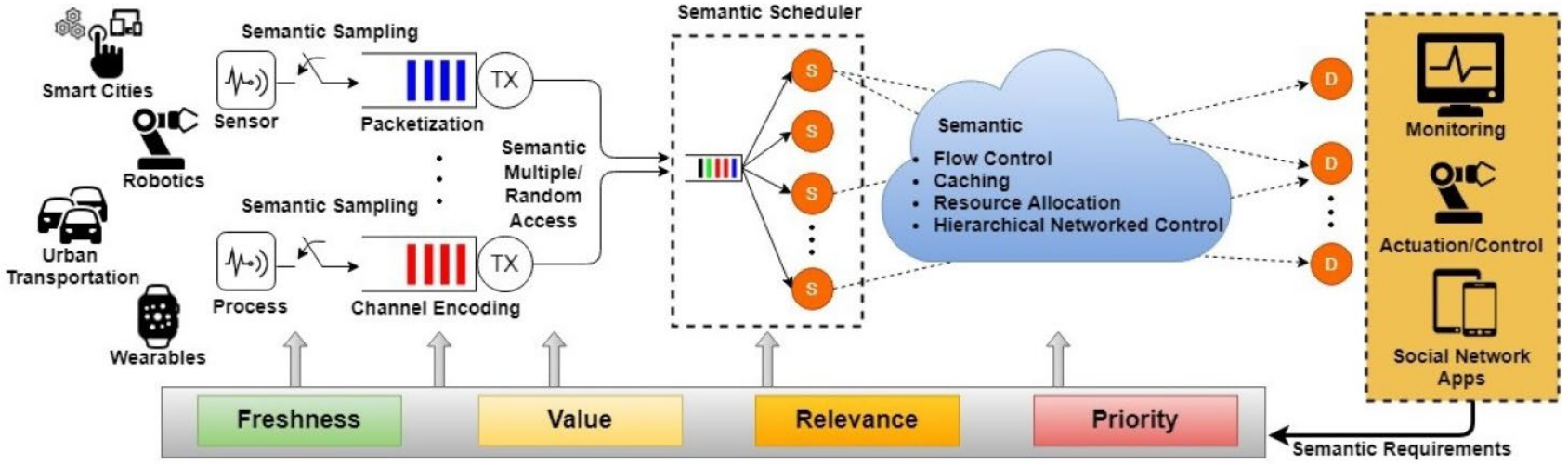
[9] A. Maatouk, et als, "The age of incorrect information: an enabler of semantics-empowered communication," arXiv preprint arXiv:2012.13214, 2020.

[10] J. Holm, et al, "Freshness on demand: Optimizing age of information for the query process," in ICC 2021-IEEE, pp. 1–6.

Semantic metrics

VoI-based semantic metrics

- Value of Information (VoI) measures the benefit of the data to be transmitted for the communication goal, which considers not only the content of the data itself, such the bursts, exceptions, etc., in monitor systems but the cost of transmission
- VoI-based metrics are a better fit for goal-oriented communications than error-based metrics
- The definition of VoI is largely task-dependent, and it is hard to give a deterministic explicit function for VoI.



E. Uysal, O. Kaya, A. Ephremides, J. Gross, M. Codreanu, P. Popovski, M. Assaad, G. Liva, A. Munari, B. Soret, T. Soleymani, and K. H. Johansson, “Semantic Communications in Networked Systems: A Data Significance Perspective,” https://arxiv.org/ftp/arxiv/papers/2103/2103.05391.pdf

Edge-enabled SemCom

transmission-before-understanding

- Shared knowledge background
- Computationally costly operations for SE model training and inference

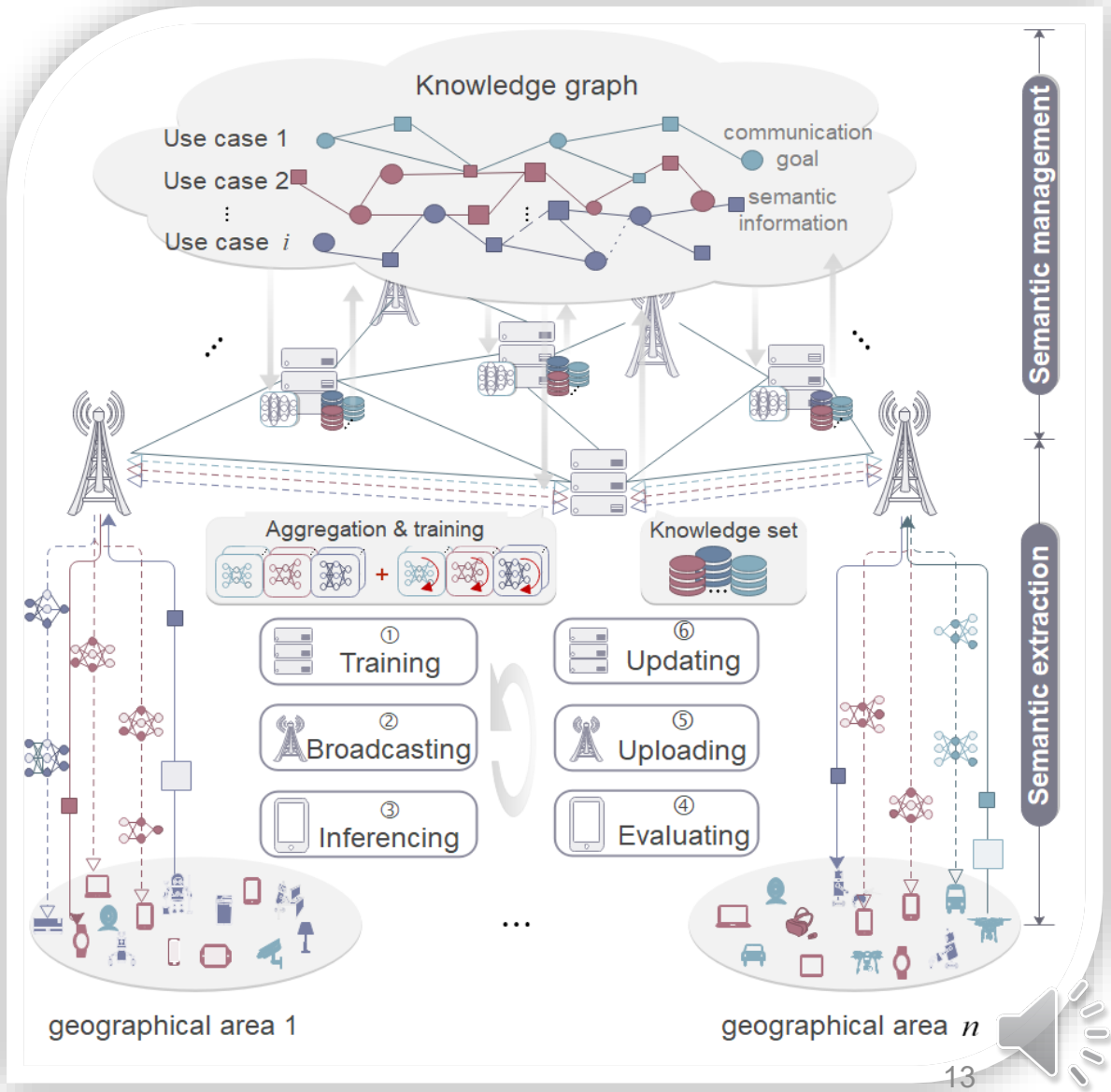
understanding-before-transmission

Main challenges

1. Limited computing power and energy constraint of the end devices result in long latency in training and updating of the SE model, thereby degrading communication reliability.

2. Comprehensive knowledge sharing among end devices to improve an SE model is at the cost of bandwidth and privacy. On the other hand, incomplete knowledge sets reduce the generalization capabilities of AI-based SE.

3. Most SE methods are task-specific and trained separately, which is far from brain-like cognition and is computationally inefficient due to the redundant work.



Edge-enabled SemCom

Federated learning-enabled SE

Challenge 1 Long latency in training

Low communication reliability

Challenge 2 High bandwidth consumption for knowledge sharing

Data security and privacy issues

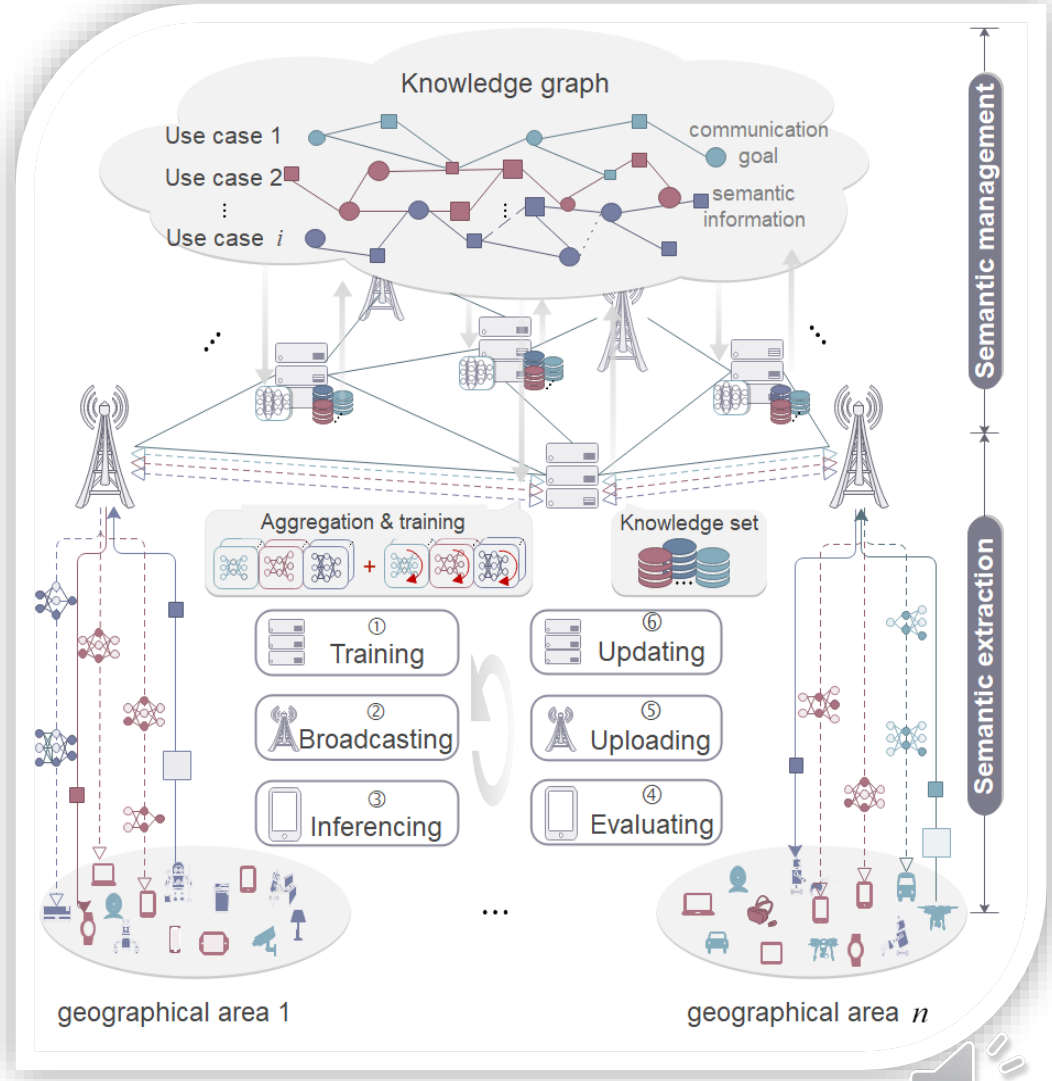
Weak generalization capabilities

Edge

- Proximity to end users
- Powerful computation and caching capabilities —C1→ background knowledge storage; SE model training
- An authoritative intermediary —C2→ knowledge sharing

Federated Learning (FL)

With FL, the trained SE model parameters in edge servers can be exchanged directly with other edge servers with identical tasks to accelerate the training process, thereby improving the generalization performance of the model in a privacy-preserving manner



Edge-enabled SemCom

Federated learning-enabled SE

① Edge servers perform the pre-training or fine-tuning for specific SE tasks based on each communication group's shared background knowledge. Model parameter exchange and federated aggregation are performed over separate communication groups with the same communication goals but not a shared knowledge background. (*Edge server*)

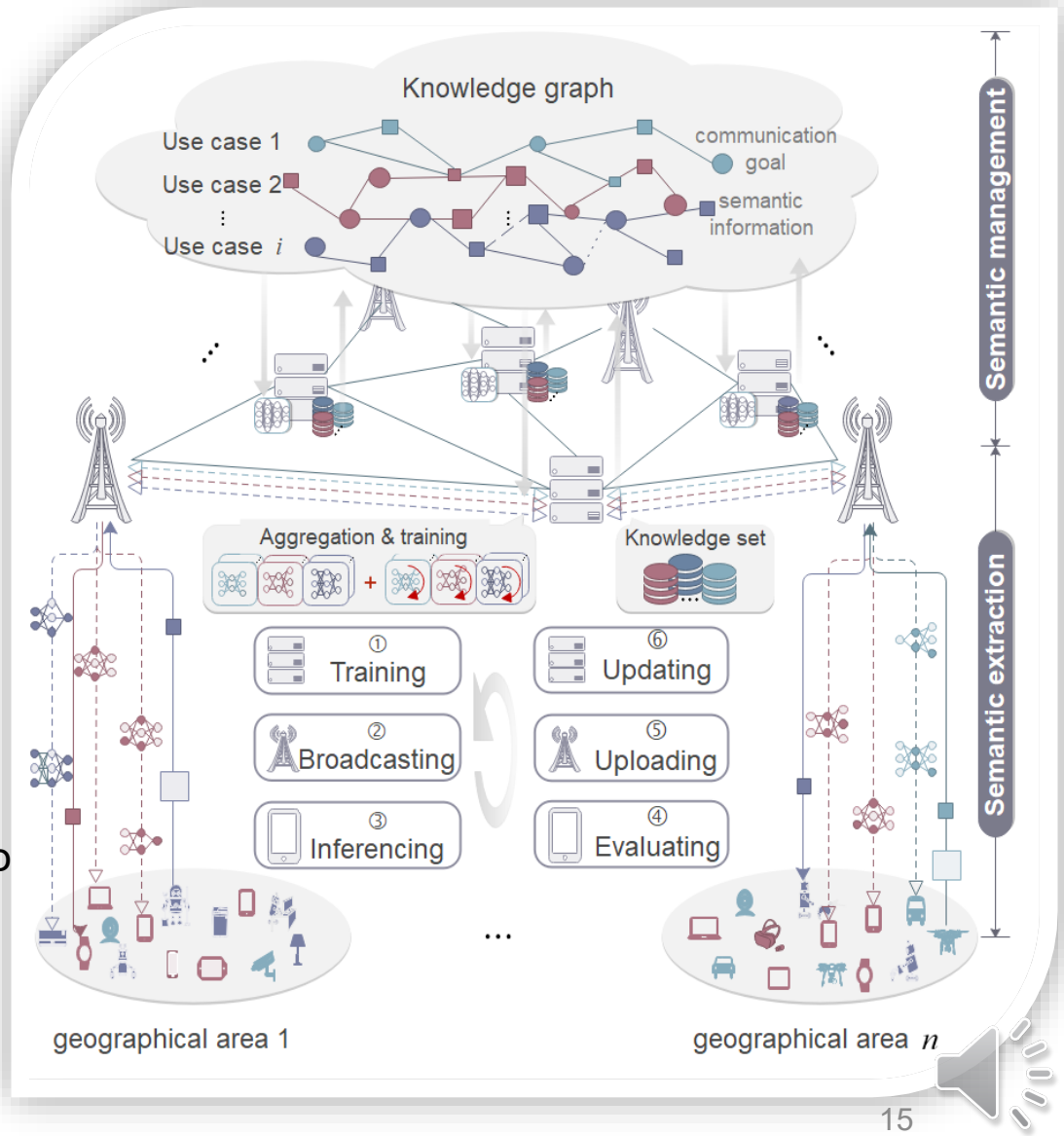
② The derived global models are broadcast separately to each communication group. (*Access point*)

③ The source devices generate the raw data. The destination devices receive SI. Then, the SE model is utilized to encode and decode SI. (*End device*)

④ The destination devices evaluate the accuracy of SI during the communications for data labeling. (*End device*)

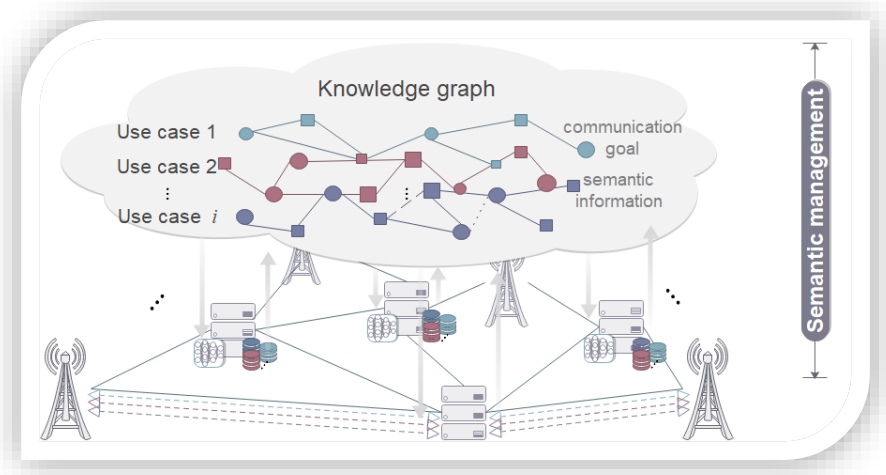
⑤ The newly labeled SI and/or corresponding raw data are uploaded to the edge servers. (*Access point*)

⑥ The edge servers perform the regular updates for the knowledge sets according to uploaded information and raw data for fine-tuning of the SE model. (*Edge server*)



Edge-enabled SemCom

Efficient SE based on edge-sharing knowledge-graph



Challenge 3

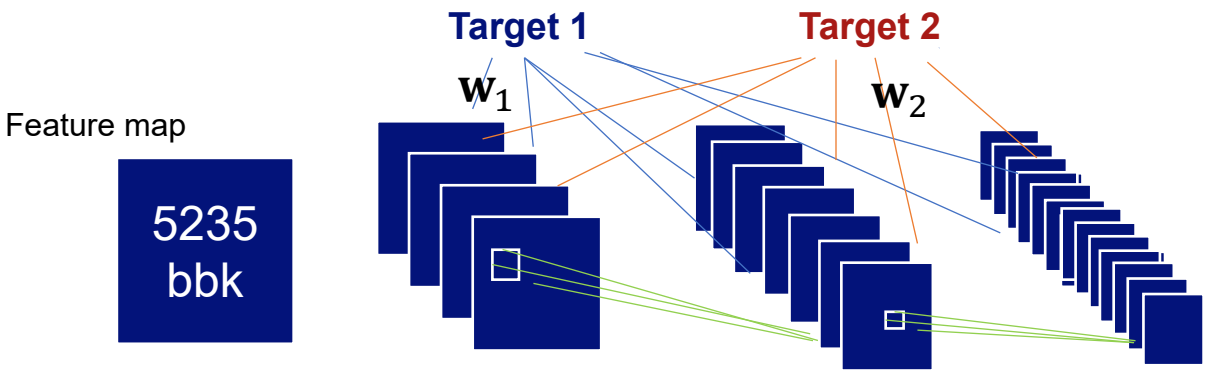
Most SE methods are task-specific and trained separately, which is computationally inefficient due to the redundant work.

Knowledge graph (KG)

- KG in Semcom is composed of source information, goals of the tasks, and the possible ways of reasoning that can be understood, recognized, and learned by all the communication participants.
- The structure of KG is much more flexible than that of having to retrain separate SE models for various tasks. Once the KG is constructed, it can be cached at the edge servers to facilitate efficient SE.

Example 1:Feature map-based SE with explainable CNN

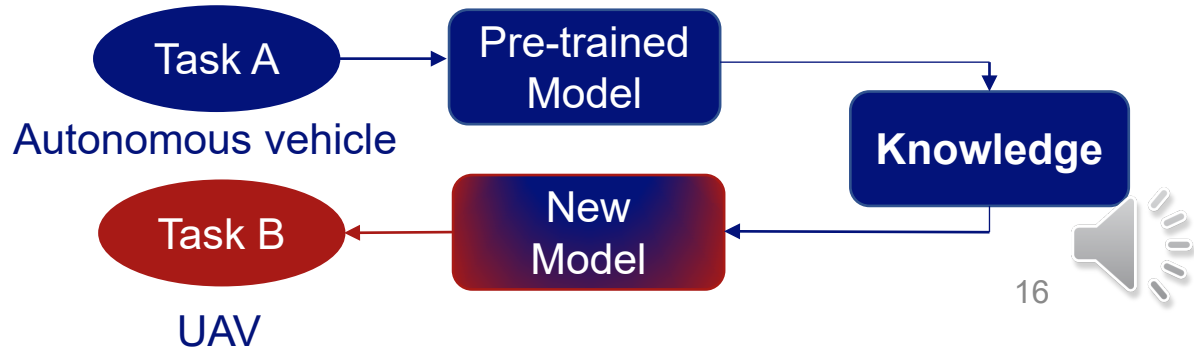
Store the importance weights of all feature maps for the tasks with different identification targets in KG



Example 2: SE model training for similar tasks with TL

Record the relationship between similar features of different tasks and the related communication goals initialization of SE model.

Communication for **collision avoidance**

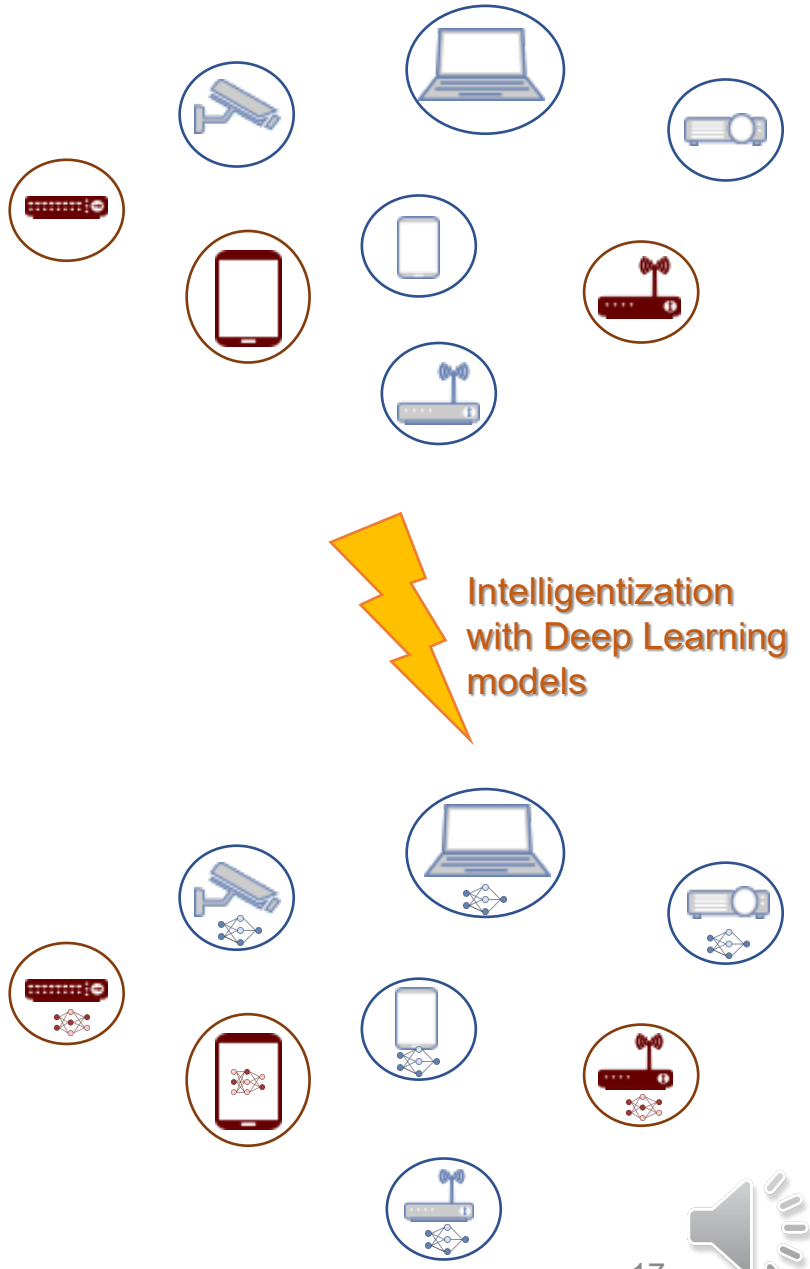


Semantic-Aware Edge Intelligence

- There are many well-studied deep learning (DL) models that can be deployed at end devices to enable the intelligentization of edge networks.
- However, DL model optimization comes at the cost of bandwidth and energy resources. The situation is exacerbated by the rapid growth of edge intelligence networks.

Main challenges

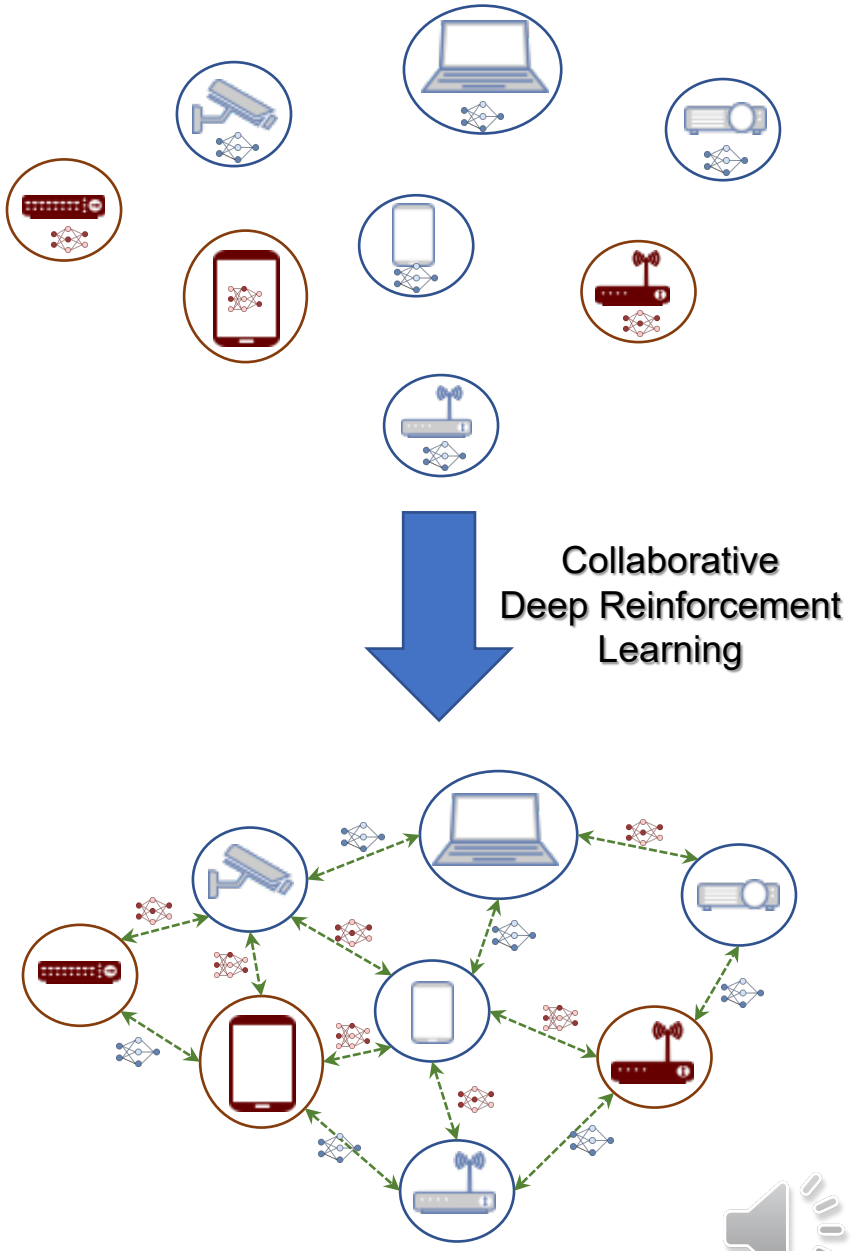
1. DL model optimization comes at the **cost of bandwidth and energy resources**.
2. In most smart services, intelligent agents are limited to the environment where they are deployed. The monotonous experience induces **overfitting issues, long convergence time, and sub-optimal performance** of the DL model.
3. **High communication overheads** incurred for the sharing and exchange of model parameters and policies.



Semantic-Aware Edge Intelligence

Semantic-aware Intelligent Agent

- Deep Reinforcement Learning (DRL) is one of the promising methods to enable intelligent agent.
- With DRL, each intelligent agent will learn an optimal policy for real-time decision making.
  - Navigation system of unmanned aerial vehicles [11]
  - Effective semantic extraction in SemCom [6]
- Performance of DRL is limited by the monotonous experience.
- Collaborative DRL (CDRL) is proposed to allow the agents to learn an optimal policy collaboratively by exchanging their model parameters or policies.



[11] W. J. Yun, B. Lim, S. Jung, Y.-C. Ko, J. Park, J. Kim, and M. Bennis, “Attention-based reinforcement learning for real-time uav semantic communication,” in 2021 17th International Symposium on Wireless Communication Systems (ISWCS). IEEE, 2021, pp. 1–6.

Semantic-Aware Edge Intelligence

Semantic-aware Intelligent Agent

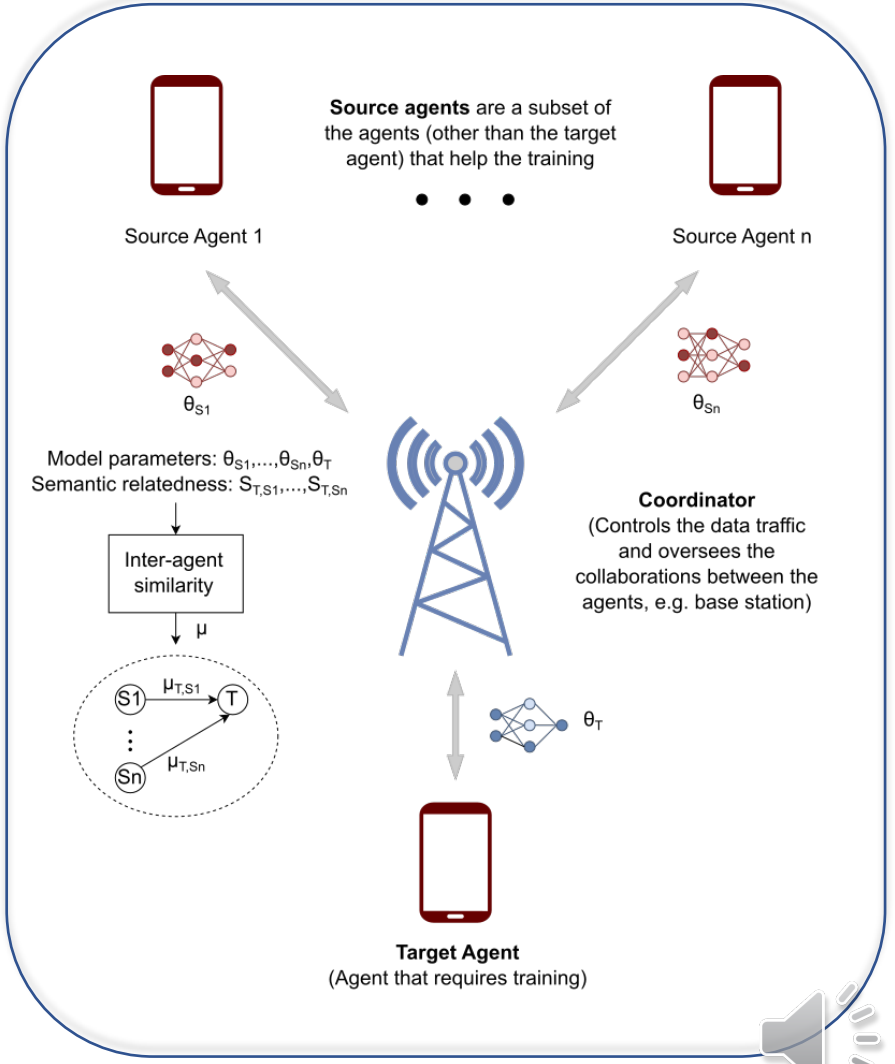
- Not all source agents can be selected for training due to the limited bandwidth.
- Hence, the allocation of the bandwidth and the agents' performance is often jointly optimized to maximize the utilities of the bandwidth.
- However, it is challenging to filter the helpful source agents because the agents have different environments, tasks, and action spaces.
- [12] proposes to consider both structural similarity and semantic relatedness when selecting the source agents.

Structural Similarity

- Can be measured by the cosine similarity between the agents' model parameters
- Weakness: cannot capture the similarity of the underlying tasks of the agents
- Agents with similar model structures may not share a similar task → poor collaboration

Semantic Relatedness

- To obtain the semantic relatedness, the target agent is trained for a fixed number of steps under the policy of the source agent
- The average return value is taken as the semantic relatedness

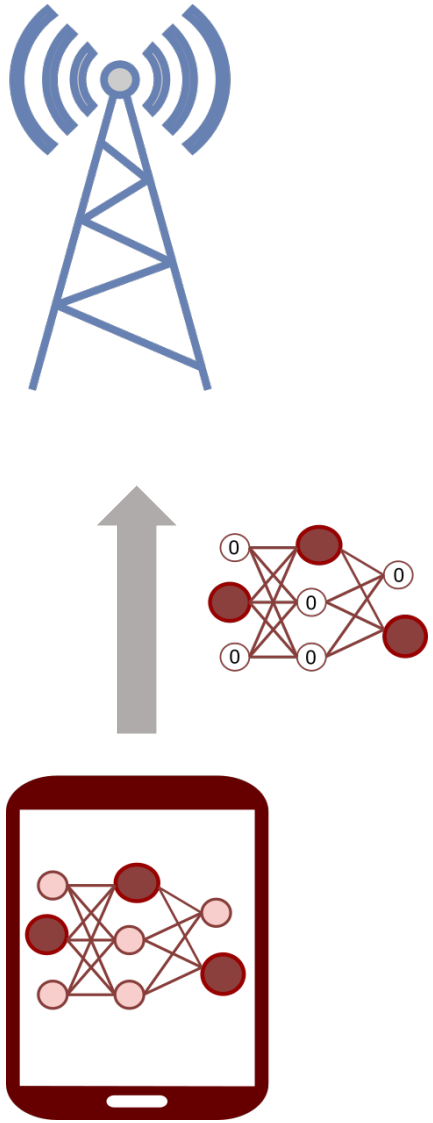


[12] A. Maatouk, et als, "The age of incorrect information: an enabler of semantics-empowered communication," arXiv preprint arXiv:2012.13214, 2020.

Semantic-Aware Edge Intelligence

Semantic-aware distributed deep learning at wireless edge networks

- Challenge: High communication overheads incurred for the sharing and exchange of model parameters and policies
- Therefore, finding an efficient way to compress the model parameters is essential to reduce the communication overhead.
- Gradient/Model Parameter Compression:
  - Instead of random sparsification, some studies proposed to consider the semantics or importance of the parameters during the data compression.
  - [13] proposes to drop the gradients with lower magnitude and transmit the gradients with higher magnitude.
    - The magnitude of the gradients signify the importance of the gradients, with higher gradients deemed to be more important for the weight updates.
    - The gradient estimates are sparsified at the transmitter and only the positions of the non-zero elements are sent to the receiver.
  - To identify the important parameters, [14] adopts adaptive model pruning where the importance of the model parameters are measured by their contribution to the future training.



[13] A. MF. Sattler, S. Wiedemann, K.-R. M¨uller, and W. Samek, “Sparse binary compression: Towards distributed deep learning with minimal communication,” in 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.
[14] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, “Model pruning enables efficient federated learning on edge devices,” arXiv preprint arXiv:1909.12326, 2019.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

Semantic-aware distributed deep learning at wireless edge networks

Conventional Communication Systems

- Classical communication systems aim to improve communication efficiency in terms of reducing the BER or SER
- Most existing resource allocation frameworks are designed to maximize throughput

- To consider the semantic importance of the bit flow
- Semantic-aware resource allocation to maximize semantic performance

Semantic Communication Systems

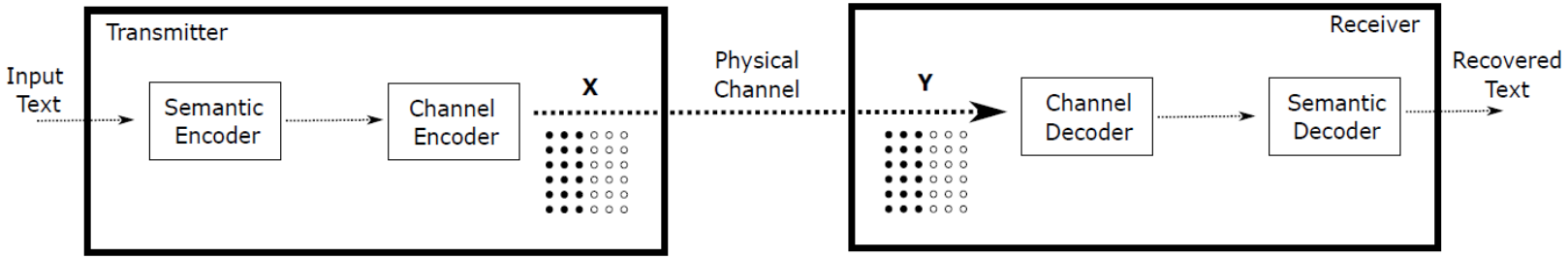
- SemCom aims to transmit the data relevant to the transmission goal
- It is necessary to redesign the resource allocation policies

Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

System Model

- We propose a system model in which energy-constrained IoT devices harvest the energy wirelessly for the purpose of text transmission [15]
- Different from existing studies that maximize the bit transmission rate, our proposed framework aims to maximize the semantic performance of the system
- We consider a wireless powered communication network where there are
  - a hybrid access point (HAP) and
  - multiple wireless powered IoT devices
- The IoT devices are equipped with semantic encoder/decoder to encode/decode semantic information from text data
- For example, semantic information of a sentence with 32 words is encoded as a 2-dimensional matrix with size 32x16, where 16 is the number of output dimension of the semantic features.

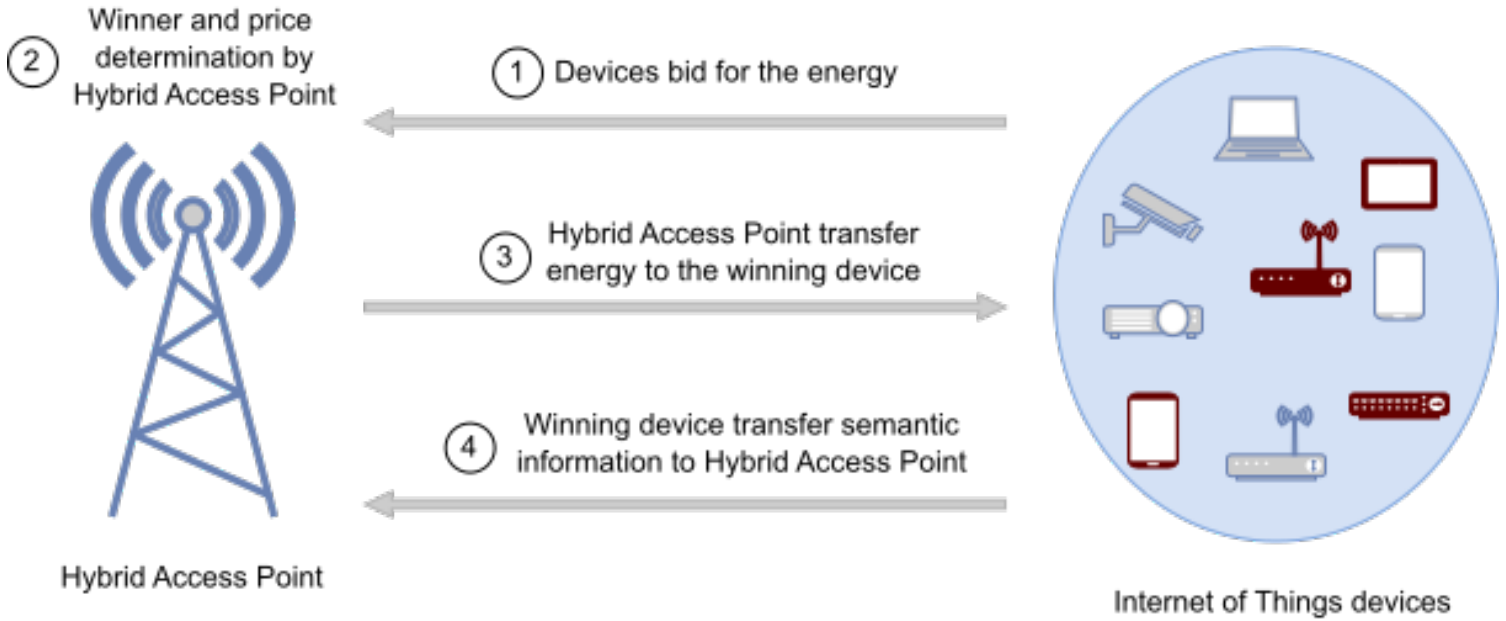


Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

System Model

- Problem
  - The HAP can transmit energy to only one IoT device at a specific time
  - How do the HAP decide the receiver of the energy?
    - Semantic-aware auction mechanism



Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

System Model

- **How to decide the bid value?**
  - Based on the receivable energy, IoT devices can achieve different semantic performance.
  - Generally, higher energy will result in better semantic performance.
  - Feature output dimension is reduced when energy received is not enough to transmit all features.
- **How to decide the winner and payment?**
  - Deep learning based auction to maximize the revenue of the HAP
  - Attains the properties of incentive compatibility and individual rationality

Algorithm 1 DL Based Auction

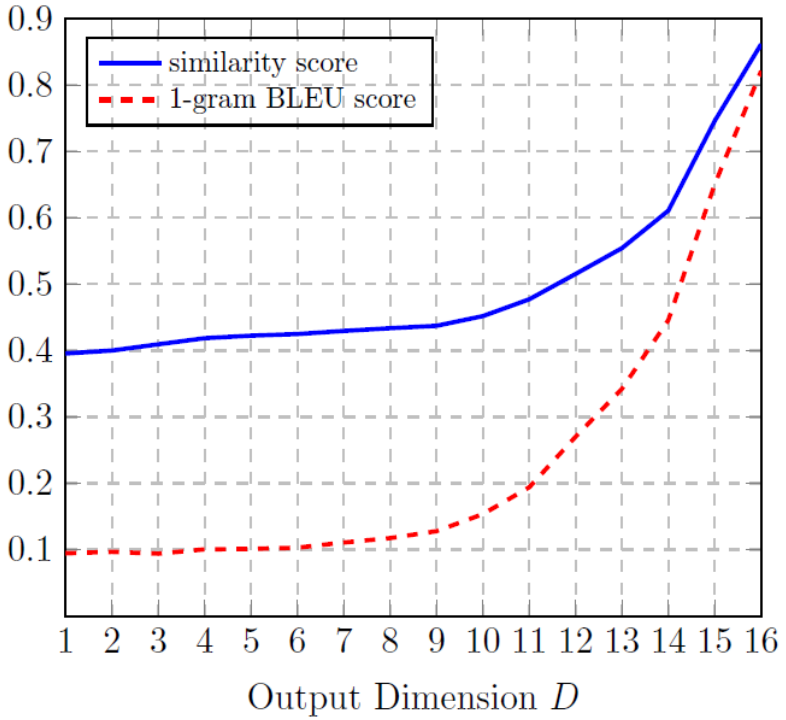
Input: Bids of IoT devices $b_i = (b_1^i, b_2^i, \ldots, b_N^i)$
Output: Revenue gain by H-AP
Initialization: $\mathbf{w} \in \mathbb{R}_+^{I \times QS}, \boldsymbol{\beta} \in \mathbb{R}^{I \times QS}$
while Loss function $\hat{R}(\mathbf{w}, \boldsymbol{\beta})$ is not minimized **do**
  Compute transformed bids $\bar{b}_n^i = \theta_n(b_n^i) = \min_{q \in Q} \max_{s \in S}(w_{qs}^n b_n + \beta_{qs}^n)$
  Compute the allocation probabilities $z_n(\mathbf{b}) = softmax(\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_{N+1}; \gamma)$
  Compute the SPA-0 payments $\theta_n^0(\mathbf{b}) = ReLU(\max_{s \neq n} \bar{b}_s)$
  Compute the conditional payment $\theta_l = \Phi_n^{-1}(\theta_n^0(\mathbf{b}))$
  Compute the loss $\hat{R}(\mathbf{w}, \boldsymbol{\beta})$
  Update parameters $\mathbf{w}$ and $\boldsymbol{\beta}$ using SGD optimizer
end while
return revenue gain by H-AP

Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

System Model

- Performance metrics:
  - Similarity score, s_n, measures the sentence similarity semantically
  - BLEU score, $BLEU_n$, measures the exact matching of words in the recovered sentence
  - Both scores decrease when output dimension decreases.
- IoT devices will value the energy based on the achievable semantic performance, $v_n = j_n s_n + m_n BLEU_n$
  - j_n and m_n are the preference of similarity score and BLEU score respectively, and $j_n + m_n = 1$



Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

Results

- The winner and price are determined by a deep learning based auction mechanism to maximize the revenue of the HAP.
- Experiment results show that the DL-based auction mechanism achieves higher revenue as compared to the traditional Second-Price Auction
- By maximizing the revenue of the access point, the price paid by the winning IoT device is also maximized.
- The energy is delivered to the device that values it the most (pay the maximized price) to ensure effective SemCom
- The deep learning based auction mechanism attains the desired properties of individual rationality and incentive compatibility for the auction.

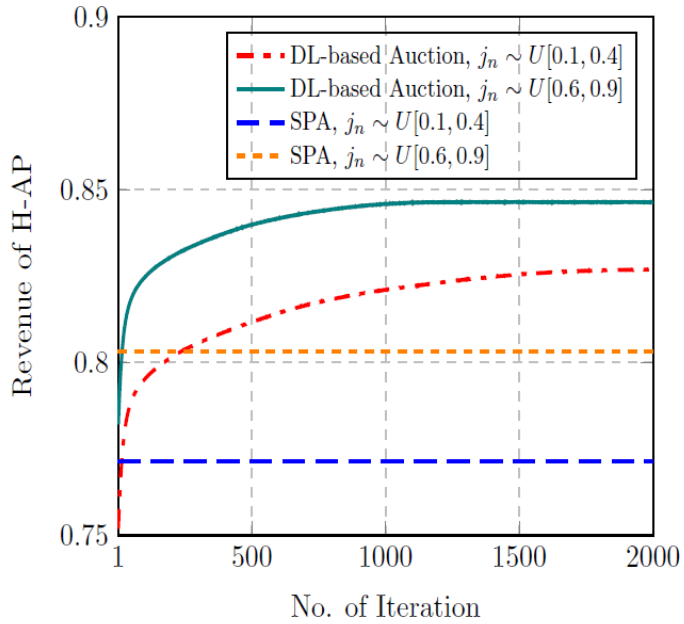


Figure 3.3: Revenue of H-AP

Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

Case Study: Resource Allocation for The Convergence of SemCom and Edge Intelligence

Results

- Fig. 4 shows that increasing wireless energy harvesting time. The reason is that the devices have more energy to send the information when longer harvesting time is given.

- Fig. 5 shows that the average bid of devices decreases when they are farther away from H-AP. The reason is that the path loss effect is greater when the distance increases.

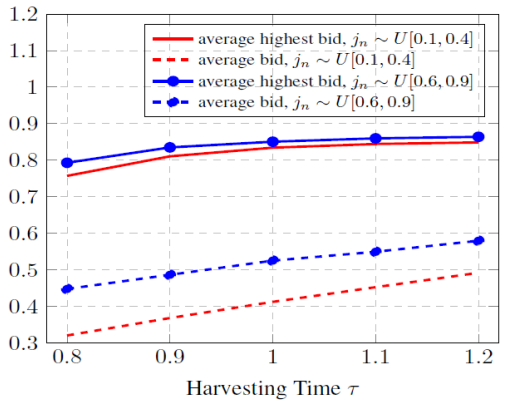


Fig. 4: Bid vs. Harvest Time τ

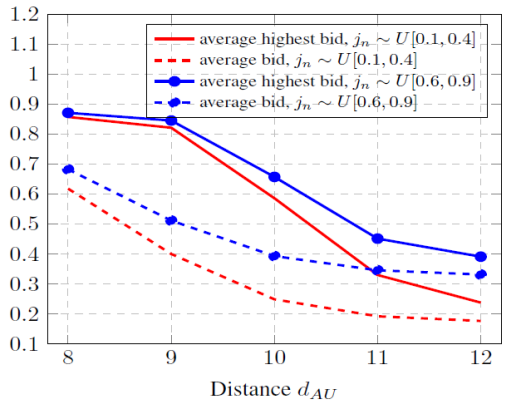


Fig. 5: Bid vs. Distance d_{AU}

Z. Q. Liew, Y. Cheng, W. Y. B. Lim, D. Niyato, C. Miao, S. Sun, "Economics of Semantic Communication System in Wireless Powered Internet of Things," in Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP), Singapore, 22-27 May 2022.

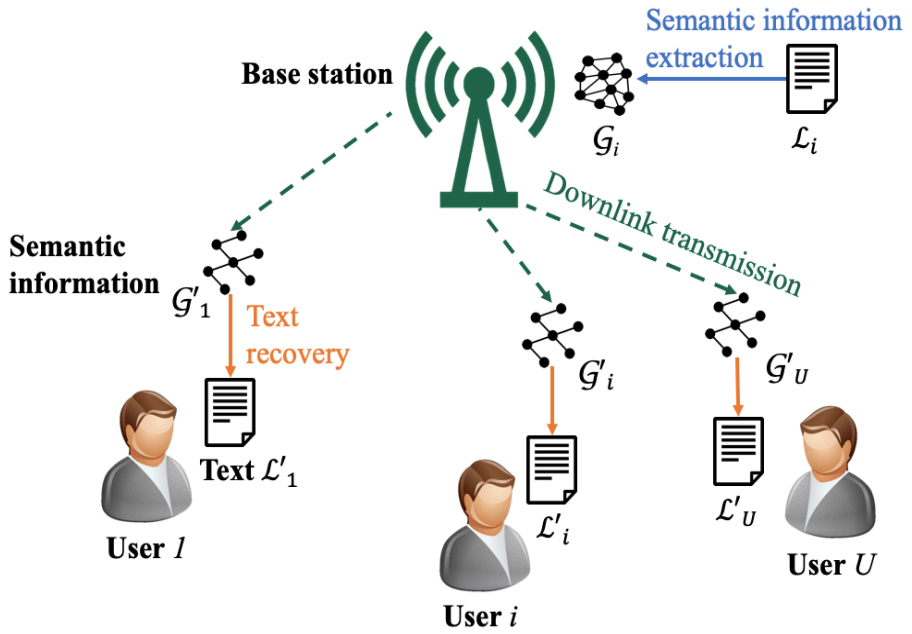
Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, “Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,”

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

System Model

- Consider a cellular network in which a BS transmits the meaning of text data to U users using semantic communication techniques.
- Semantic information extraction
- Semantic information transmission
- Text recovery
- Performance evaluation



Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Semantic Information Extraction

➢ The semantic information of a text data is modeled by a knowledge graph (KG).

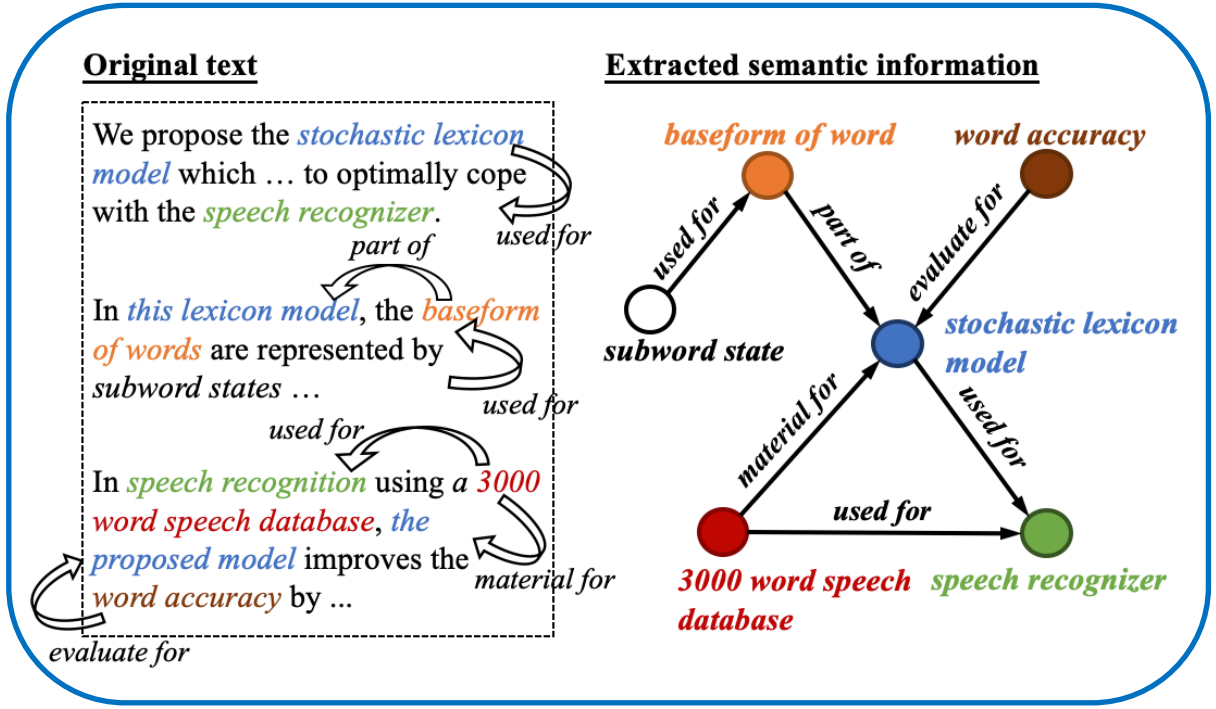
- *Entity* recognition
- *Relation* classification

semantic triple

$$\mathcal{G}_i = \{(\varepsilon_i^1, \cdots, \varepsilon_i^g, \cdots, \varepsilon_i^{G_i})\}$$

where $\varepsilon_i^g = (e_{i,j}^g, r_{i,jk}^g, e_{i,k}^g)$

entity | relation | entity



Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, “Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,”

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Semantic Information Transmission

- Given the transmission delay threshold T, the BS must select *partial semantic information*

$$\mathcal{G}_i' = \{(\varepsilon'^{1}_i, \cdots, \varepsilon'^{h}_i, \cdots, \varepsilon'^{H_i}_i)\} \subset \mathcal{G}_i$$

whose data size $Z(\mathcal{G}_i') = \sum_{h=1}^{H_i} \left(S^h_{i,j} + S^h_{i,k} + 2 \right)$

Z({"*stochastic lexicon model*", "*used for*", "*speech recognizer*"})=3+2+2=7

that satisfy $\frac{Z(\mathcal{G}_i')R}{c_i(\boldsymbol{\alpha}_i)} \leqslant T$ → data rate depends on the *resource allocation* for each user

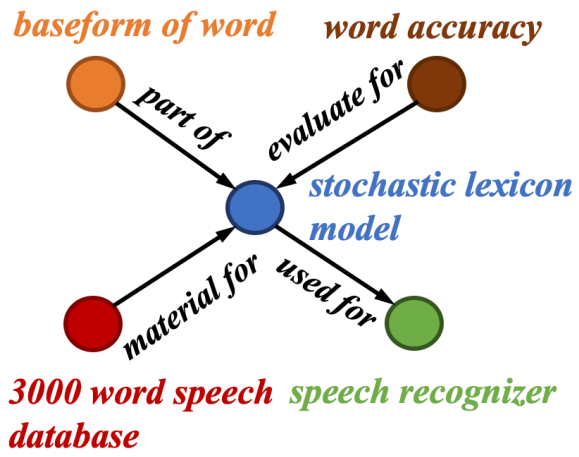
Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Text Recovery

- Each user recovers the original text using a graph-to-text generation

$$L_i'(\boldsymbol{\alpha}_i, \mathcal{G}_i') = \{w_{i,0}', w_{i,1}', \cdots, w_{i,m}', \cdots, w_{i,M_i}'\}$$

Partial semantic information



Recovered text

In this paper, we present a new *stochastic lexicon model* for *speech recognizer*.

The proposed *stochastic lexicon model* consists of a *baseform of words.*

A *stochastic lexicon model* is trained on a *3000 word speech database*.

The experimental results show that the proposed *stochastic lexicon model* achieves higher *word accuracy*.

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Semantic Information Metric

- The ***semantic accuracy*** of recovered $L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i)$

$$A_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \frac{\sum_{m=1}^{M_i} \min\left(\sigma\left(L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i), w'_{i,m}\right), \sigma(L_i, w'_{i,m})\right)}{\sum_{m=1}^{M_i} \sigma\left(L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i), w'_{i,m}\right)}$$

- The ***semantic completeness*** of recovere $L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i)$

$$R_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \frac{\sum_{m=1}^{M_i} \min\left(\sigma\left(L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i), w'_{i,m}\right), \sigma(L_i, w'_{i,m})\right)}{\sum_{m=1}^{M_i} \sigma(L_i, w'_{i,m})}$$

- The proposed ***metric of semantic similarity*** (MSS

$$E_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \theta_i \frac{A_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) R_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i)}{\varphi A_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) + (1-\varphi) R_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i)}$$

L_i (original): "*I sit on a chair.*"

$L'_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i)$ (recovered): "*I sit.*"

$$A_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \frac{1+1}{1+1} = 1$$

$$R_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \frac{1+1}{1+1+1+1+1} = \frac{2}{5}$$

$$E_i(\boldsymbol{\alpha}_i, \mathcal{G}'_i) = \frac{1 \times \frac{2}{5}}{\frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times 1} = \frac{4}{7}$$

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Total Metric of Semantic Similarity (MSS) maximization problem

$$\max_{\boldsymbol{\alpha}_i, \mathcal{G}_i'} \sum_{i=1}^{U} E_i(\boldsymbol{\alpha}_i, \mathcal{G}_i'),$$

depends on → resource allocation; semantic information selection

$$\text{s.t. } \alpha_{i,q} \in \{0,1\}, \forall i \in \mathcal{U}, \forall q \in \mathcal{Q},$$

$$\sum_{q=1}^{Q} \alpha_{i,q} \leqslant 1, \forall i \in \mathcal{U},$$

$$\sum_{i=1}^{U} \alpha_{i,q} \leqslant 1, \forall q \in \mathcal{Q},$$

$$\frac{Z_i(\mathcal{G}_i')R}{c_i(\boldsymbol{\alpha}_i)} \leqslant T, \forall i \in \mathcal{U},$$

- How to build the relationship between the non-mathematical optimization variables and the objective function?
- The objective function is calculated based on the texted recovery model.

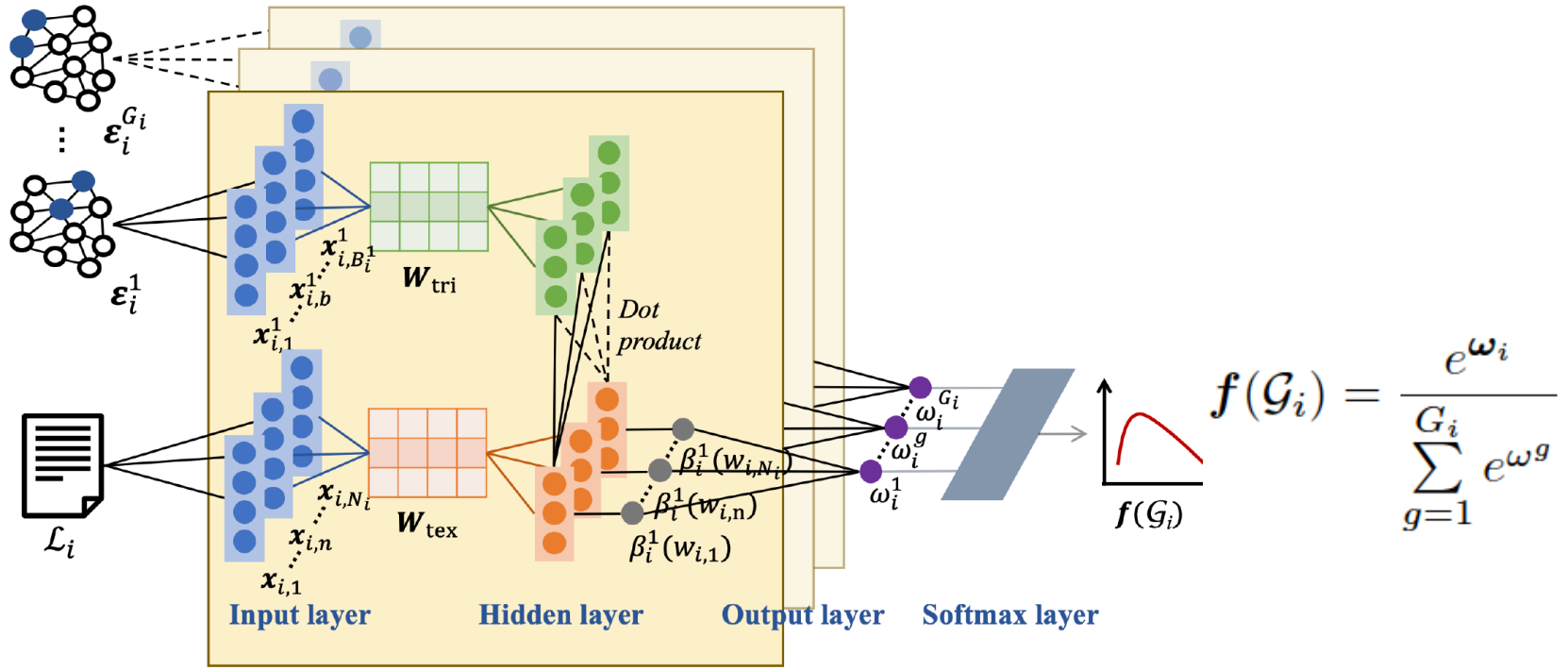
Evaluate the ***importance of the semantic triples!***

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Attention RL for Semantic Information Selection and Resource Allocation

- The ***importance distribution*** of semantic information $\mathcal{G}_i$



$$f(\mathcal{G}_i) = \frac{e^{\omega_i}}{\sum_{g=1}^{G_i} e^{\omega^g}}$$

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Attention Policy Gradient (APG)

Components of the attention policy gradient (APG)

- Agent: the BS
- Actions: $\boldsymbol{a} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_i, \cdots, \boldsymbol{\alpha}_U]$,

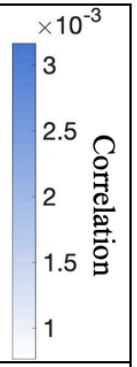
  The partial semantic information $\mathcal{G}_i'$ to be transmitted:
  - the most important triples in $\mathcal{G}_i$ $\frac{Z_i(\mathcal{G}_i')R}{c_i(\boldsymbol{\alpha}_i)} \leqslant T$
  - satisfying $\boldsymbol{s} = [\boldsymbol{f}(\mathcal{G}_1), \cdots, \boldsymbol{f}(\mathcal{G}_U)]$
- States: $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{a}) = P(\boldsymbol{a}|\boldsymbol{s})$
- Policy: $R(\boldsymbol{a}|\boldsymbol{s}) = \sum_{i=1}^{U} E_i(\boldsymbol{\alpha}_i, \mathcal{G}_i')$
- Reward.

$\boldsymbol{\theta}$ is parameters of a DNN that used to map the input importance distributions and the output resource allocation.

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

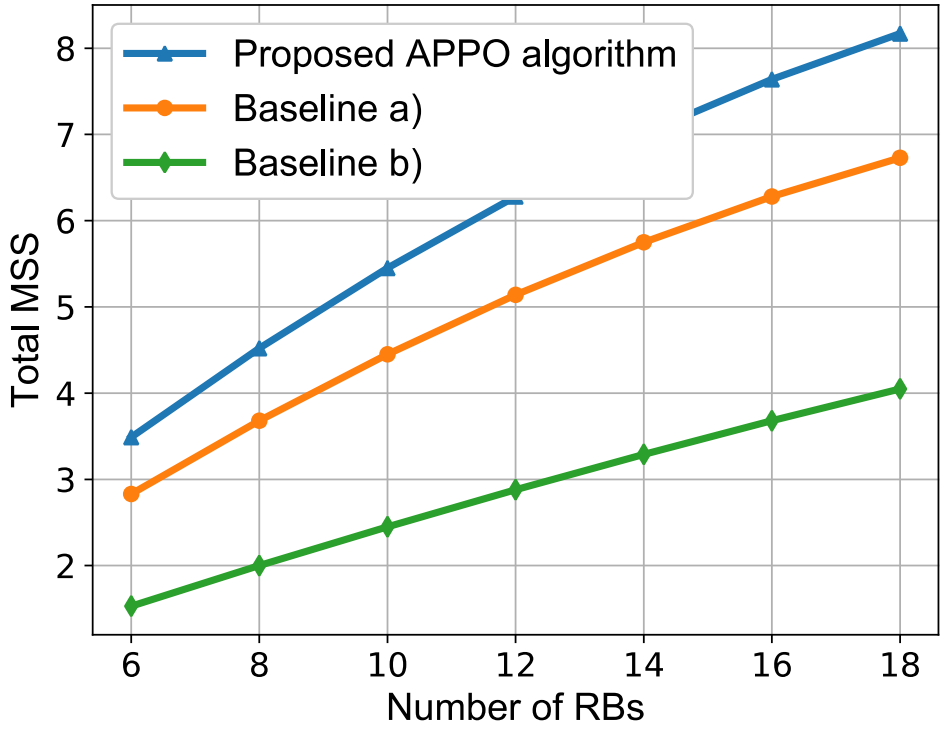
Simulation Results

Original text	We propose the stochastic lexicon model which represents the pronunciation variations to optimally cope with the continuous speech recognizer. In this lexicon model, the baseform of words are represented by subword states and the probability distribution of subwords as a hidden Markov model. Also, the proposed approach can be applied to a system employing non–linguistic recognition units and the lexicon is automatically trained from training utterances. In speaker independent speech recognition tests using a 3000 word continuous speech database, the proposed system improves the word accuracy by about 27.8% and the sentence accuracy by about 22.4%. 
Transmitted semantic information	*(stochastic lexicon model, used for, speech recognizer),* *(baseform of word, part of, stochastic lexicon model),* *(probability distribution of subwords, conjunction with, baseform of words),* *(3000 word speech database, material for, stochastic lexicon model),* *(word accuracy; evaluate for; stochastic lexicon model),* *(sentence accuracy; conjunction with; word accuracy),*
Recovered text	in this paper we present a new STOCHASTIC LEXICON MODEL for SPEECH RECOGNIZER. the proposed STOCHASTIC LEXICON MODEL consists of a BASEFORM OF WORDS and a PROBABILITY DISTRIBUTION OF SUBWORDS. a STOCHASTIC LEXICON MODEL is trained on a 3000 WORD CONTINUOUS SPEECH DATABASE. the experimental results show that the proposed STOCHASTIC LEXICON MODEL achieves higher WORD ACCURACY and SENTENCE ACCURACY than the conventional STOCHASTIC LEXICON MODEL.

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Simulation Results

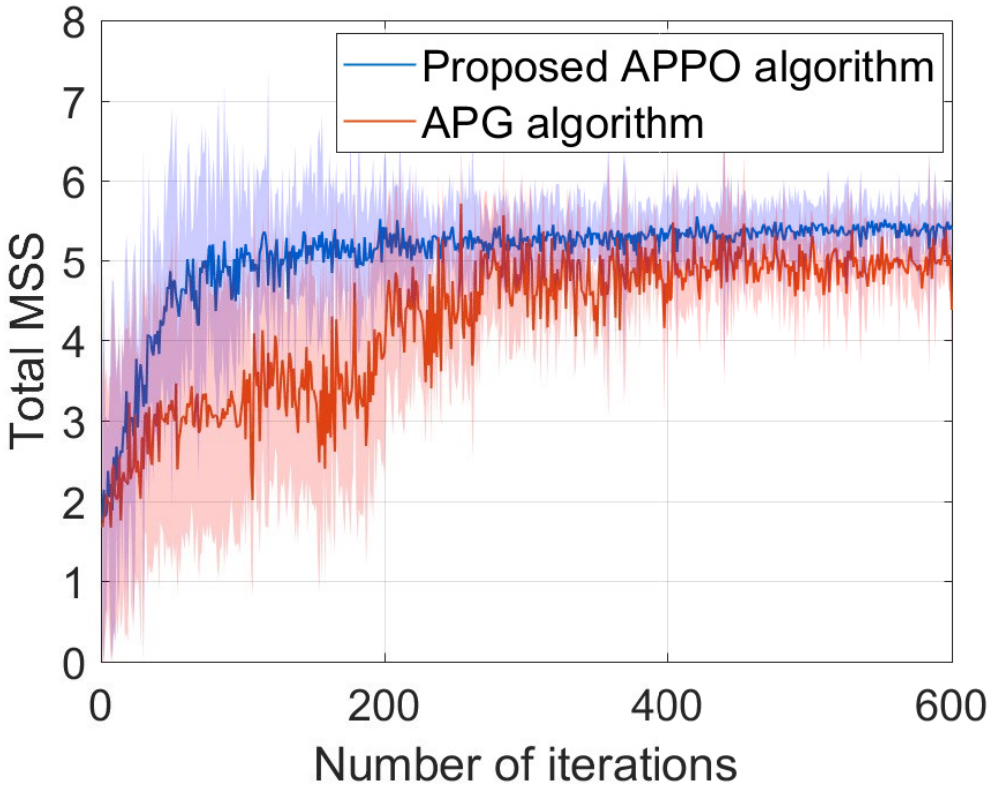


- Two baselines:
  - baseline a): randomly selects semantic triples;
  - baseline b): directly transmits the original text data.
- The proposed APPO algorithm can improve the total MSS.

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, "Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,"

Case Study: Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach

Simulation Results



- Baseline:
  - APG algorithm: traditional policy gradient algorithm integrated with attention network;
- Compared to the APG algorithm, the proposed APPO algorithm converges faster.

Y. Wang, M. Chen, T. Luo, W. Saad, D. Niyato, H. V. Poor, and S. Cui, “Performance Optimization for Semantic Communications: An Attention-based Reinforcement Learning Approach,”

General Future Directions

Interpretability and explainability of SE

As unexpected information often appears in communications, the black-box nature of SE method hinders its implementation. Hence, interpretability in SE needs to be studied to associate possible causes and results and to guide improvements to the SE model. Meanwhile, explainability in SE can identify the SI hidden in deep nets, which paves the way to the KG-based efficient SE across multiple modalities and tasks. However, most existing SE methods are not explainable.

Semantic-noise based privacy preserving

For the communication groups with similar background knowledge and communication goals, eavesdropping becomes easy. Considering the success of covert communication in which artificial noise is introduced for secure wireless transmissions, artificially increasing the mismatch to generate semantic noise may also serve as a potential method to enable secure SemCom.

Variable-length semantic encoding

Existing works merely consider the dynamic channel gains in SE without the concern of resource constraints. However, in a multi-user scenario, the fluctuation in resources, such as available spectrum and transmit power, can have a non-negligible impact on SemCom performance. The methods of achieving variable-length semantic encoding to cope with dynamic network resources remain thus to be an open research question.

General Future Directions

DL-based SE for Task Similarity

- For future works, the semantic relatedness between the agents can be extracted using a deep learning network. The deep learning network can take the model parameters of the agent as input and output embeddings as the semantic representations of the agent.
- The network can be trained by minimizing the similarity between the output embeddings of different tasks, and maximizing the similarity of the output embeddings of the similar tasks. In this way, the semantic representation can be extracted to calculate the task similarity between the agents. As such, the most efficient source agents can be selected for the fixed bandwidth usage.

Semantic Compression of Model Parameters

- Future works can adopt semantic-aware model parameter exchange between the end devices. Such a system can follow the semantic encoder/decoder structure in [5] where the input parameters are first encoded by the transmitter using a semantic encoder and a channel encoder, before sending the encoded information to the receiver. The received signal is decoded by channel decoder and semantic decoder at the receiver to reconstruct the original data.
- For semantic text transmission, the SI sent by the transmitter carries the information used to reconstruct the text data at the receiver's end. In the case of CDRL, the encoded SI is the essential information to reconstruct the model parameter at the receiver's end.

Thank You

dniyato@ntu.edu.sg