# Deep Learning Tasks Processing in Fog-RAN

Sheng Hua, **Xiangyu Yang**, Kai Yang, Gao Yin, Yuanming Shi, Hao Wang

School of Information Science and Technology
ShanghaiTech University

上海科技大学
ShanghaiTech University

# Outline

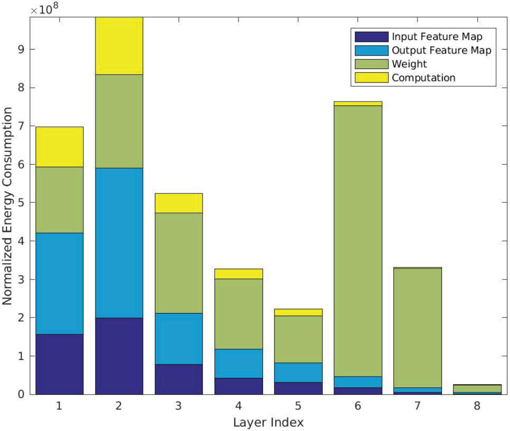
# Motivations

- **The marriage of mobile edge computing (MEC) and artificial intelligence (AI) to evoke potentials**
  - the explosive growth in the volume of data at the network edge
  - the unprecedented success of data-driven deep learning (DL) applications
  - the growing need to perform intelligent tasks on mobile devices such as autonomous vehicles and drones

# Motivations

- **The marriage of mobile edge computing (MEC) and artificial intelligence (AI) to evoke potentials**
  - the explosive growth in the volume of data at the network edge
  - the unprecedented success of data-driven deep learning (DL) applications
  - the growing need to perform intelligent tasks on mobile devices such as autonomous vehicles and drones
- **The enormous consumption due to**
  - the dense deployment of base stations (BSs)
  - the energy-demanding nature of DL algorithms

# Power Consumption Decomposition for AlexNet

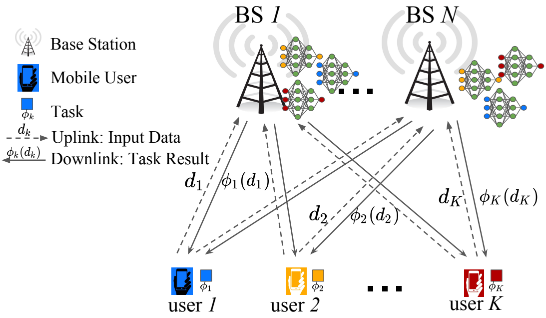


[1]

- That's around $0.45$W, which is comparable to the power consumption of a BS, e.g., $1$W.

[1] Produced on the website https://energyestimation.mit.edu/.

# Framework




- **Basic tradeoff**: more BSs working on the same task results in higher quality-of-service (QoS) perceived by users, at the cost of computation and communication inefficiency
- **Goal**
  - minimize power consumption while satisfying pre-defined QoS to achieve green mobile edge computing

# Outline

# System Model

- **Communication Model**: the set of N $L$-antenna BSs $\mathcal{N}$, the set of $K$ single-antenna users $\mathcal{K}$, task selection strategy $\mathcal{A} = (\mathcal{A}_1, \cdots, \mathcal{A}_N)$

$$y_k = \sum_{n \in \mathcal{N}} \boldsymbol{h}_{kn}^{\mathrm{H}} \sum_{l \in \mathcal{A}_n} \boldsymbol{v}_{nl} s_l + z_k$$

  - $y_k \in \mathbb{C}$: the received signal at the $k$-th user
  - $\boldsymbol{h}_{kn} \in \mathbb{C}^L$: the channel vector between the $k$-th user and $n$-th BS
  - $s_l \in \mathbb{C}$: the representative signal for task result $\phi_l(d_l)$
  - $\boldsymbol{v}_{nl} \in \mathbb{C}^L$: the beamforming vector at $n$-th BS for signal $s_l$
  - $z_k \sim \mathcal{CN}\left(0, \sigma_k^2\right)$: the complex additive white Gaussian noise

# System Model

- **Communication Model**: the set of N $L$-antenna BSs $\mathcal{N}$, the set of $K$ single-antenna users $\mathcal{K}$, task selection strategy $\mathcal{A} = (\mathcal{A}_1, \cdots, \mathcal{A}_N)$

$$y_k = \sum_{n \in \mathcal{N}} \boldsymbol{h}_{kn}^{\mathrm{H}} \sum_{l \in \mathcal{A}_n} \boldsymbol{v}_{nl} s_l + z_k$$

  - $y_k \in \mathbb{C}$: the received signal at the $k$-th user
  - $\boldsymbol{h}_{kn} \in \mathbb{C}^L$: the channel vector between the $k$-th user and $n$-th BS
  - $s_l \in \mathbb{C}$: the representative signal for task result $\phi_l(d_l)$
  - $\boldsymbol{v}_{nl} \in \mathbb{C}^L$: the beamforming vector at $n$-th BS for signal $s_l$
  - $z_k \sim \mathcal{CN}\left(0, \sigma_k^2\right)$: the complex additive white Gaussian noise

- **Power Consumption Model**

$$\underbrace{\sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K}} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2}_{\text{communication power}} + \underbrace{\sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{A}_n} P_{nk}^{\mathrm{c}}}_{\text{computation power}}$$

  - $\eta_n$: the power amplifier efficiency of the $n$-th BS
  - $P_{nk}^{\mathrm{c}}$: the computational power consumption for $n$-th BS to perform the $k$-th user's task

# Problem Formulation

- Given users' target QoS $[\gamma_1, \ldots, \gamma_K]$, and BSs' maximum power limits $[P_1^{\max}, \ldots, P_N^{\max}]$, the goal of green computing is formulated as the following joint transmit beamforming design and task selection problem

$$\begin{aligned}
\underset{\mathcal{A}, \boldsymbol{v}}{\text{minimize}} \quad & \sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K}} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2 + \sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{A}_n} P_{nk}^{\mathsf{c}} \\
\text{subject to} \quad & \mathrm{SINR}_k \geq \gamma_k, \quad k = 1, \ldots, K, \\
& \sum_{k \in \mathcal{K}} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_n^{\max}, \quad n = 1, \ldots, N.
\end{aligned}$$

  - $\mathrm{SINR}_k = \frac{\left|\sum_{n: k \in \mathcal{A}_n} \boldsymbol{h}_{kn}^{\mathrm{H}} \boldsymbol{v}_{nk}\right|^2}{\sum_{l \neq k} \left|\sum_{n: l \in \mathcal{A}_n} \boldsymbol{h}_{kn}^{\mathrm{H}} \boldsymbol{v}_{nl}\right|^2 + \sigma_k^2}$
  - $\boldsymbol{v} = \left[\boldsymbol{v}_{11}^{\mathrm{H}}, \boldsymbol{v}_{12}^{\mathrm{H}}, \ldots, \boldsymbol{v}_{NK}^{\mathrm{H}}\right]^{\mathrm{H}} \in \mathbb{C}^{NLK}$ is the aggregated transmit beamforming vector.

# Problem Analysis

- The above problem is a mixed-integer-nonlinear-programming (MINLP), which is generally NP-hard and and computationally difficult. The combinatorial optimization variable $\mathcal{A}$ makes this problem nonconvex.

# Problem Analysis

- The above problem is a mixed-integer-nonlinear-programming (MINLP), which is generally NP-hard and and computationally difficult. The combinatorial optimization variable $\mathcal{A}$ makes this problem nonconvex.
- **Key Observation**
  - Group sparsity structure can be exploited to bridge the combinatorial variable $\mathcal{A}$ and the aggregated beamforming vector $\boldsymbol{v}$. Specifically, if the $n$-th BS does not perform task $\phi_k$, the corresponding beamforming vector $\boldsymbol{v}_{nk}$ can be set as zero (i.e., $\|\boldsymbol{v}_{nk}\|_2 = 0$), which leads to the group sparsity structure of $\boldsymbol{v}$.

# Problem Analysis

- The above problem is a mixed-integer-nonlinear-programming (MINLP), which is generally NP-hard and and computationally difficult. The combinatorial optimization variable $\mathcal{A}$ makes this problem nonconvex.
- **Key Observation**
  - Group sparsity structure can be exploited to bridge the combinatorial variable $\mathcal{A}$ and the aggregated beamforming vector $\boldsymbol{v}$. Specifically, if the $n$-th BS does not perform task $\phi_k$, the corresponding beamforming vector $\boldsymbol{v}_{nk}$ can be set as zero (i.e., $\|\boldsymbol{v}_{nk}\|_2 = 0$), which leads to the group sparsity structure of $\boldsymbol{v}$.
- **Tackling NP-hard MINLP $\Longrightarrow$ Inducing Structured Sparsity**

# Outline

# Structured Sparsity Inducing Norms

- **Related Works**
  - mixed $\ell_{1,2}$-norm [Shi et al.'14].
  - re-weighted $\ell_1$ norm [Peng et al.'17]
  - re-weighted $\ell_2$ norm [Shi et al.'16].
- **Proposal: Log-Sum Function for Sparsity Inducing**

$$
\begin{aligned}
\underset{\boldsymbol{v}}{\text{minimize}} \quad & \Omega(\boldsymbol{v}) := \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}} \rho_{nk} \log(1 + p\,\|\boldsymbol{v}_{nk}\|_2) \\
\text{subject to} \quad & \mathrm{SINR}_k \geq \gamma_k, \quad k = 1, \ldots, K, \\
& \sum_{k\in\mathcal{K}} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_n^{\max}, \quad n = 1, \ldots, N,
\end{aligned}
$$

where $\rho_{nk} = \sqrt{P_{nk}^{\mathrm{c}}/\eta_n}$.

- Based on the fact that the log-sum function serves as a tighter approximation to $\ell_0$-norm $\|\boldsymbol{x}\|_0$ compared to $\ell_1$-norm $\|\boldsymbol{x}\|_1$ [Candes et al.'08].

- **New Challenge**
  - the nonconvex and nonsmooth nature of $\Omega(\boldsymbol{v})$ with respect to $\boldsymbol{v}_{nk}$
- **Solution**
  - iteratively approximate $\Omega(\boldsymbol{v})$ by its linearization at current iterate $\boldsymbol{v}^{[i]}$ until converge

$$\Omega(\boldsymbol{v}) \approx \sum_{n \in \mathcal{N}} \sum_{k \in \mathcal{K}} w_{nk}^{[i]} \left\| \boldsymbol{v}_{nk} \right\|_2 ,$$

    and the weight $w_{nk}^{[i]}$ is updated as

$$w_{nk}^{[i]} = \frac{p \rho_{nk}}{p \left\| \boldsymbol{v}_{nk}^{[i]} \right\|_2 + 1}.$$

# The Overall Algorithm

- *Step 1*: induce group sparsity by iteratively solving the linearized log-sum based optimization problem, which is actually the re-weighted $\ell_1$ sparsity inducing norm.

# The Overall Algorithm

- *Step 1*: induce group sparsity by iteratively solving the linearized log-sum based optimization problem, which is actually the re-weighted $\ell_1$ sparsity inducing norm.
- *Step 2*: arrange tasks in a descending order according to the rule $\theta_{nk} = \sqrt{\frac{\|\boldsymbol{h}_{kn}\|_2^2 \eta_n}{P_{nk}^c}} \|\boldsymbol{v}_{nk}^\star\|_2$, and determine the feasible task selection strategy with least cardinality.

# The Overall Algorithm

- *Step 1*: induce group sparsity by iteratively solving the linearized log-sum based optimization problem, which is actually the re-weighted $\ell_1$ sparsity inducing norm.
- *Step 2*: arrange tasks in a descending order according to the rule $\theta_{nk} = \sqrt{\frac{\|\boldsymbol{h}_{kn}\|_2^2 \eta_n}{P_{nk}^c}} \|\boldsymbol{v}_{nk}^\star\|_2$, and determine the feasible task selection strategy with least cardinality.
- *Step 3*: fix the task selection strategy and refine beamforming vectors. This is achieved by solving

$$\begin{aligned}
\underset{\boldsymbol{v}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{K}} \frac{1}{\eta_n} \|\boldsymbol{v}_{nk}\|_2^2 + \sum_{n} \sum_{k\in\mathcal{A}_n} P_{nk}^{\mathsf{c}} \\
\text{subject to} \quad & \text{SINR}_k \geq \gamma_k, \quad k = 1, \ldots, K, \\
& \sum_{k\in\mathcal{K}} \|\boldsymbol{v}_{nk}\|_2^2 \leq P_n^{\max}, \quad n = 1, \ldots, N, \\
& \boldsymbol{v}_{\pi^{(t)}} = \boldsymbol{0}.
\end{aligned}$$

where $\pi^{(t)}$ is the active task index determined in *Step 2*.

# Outline

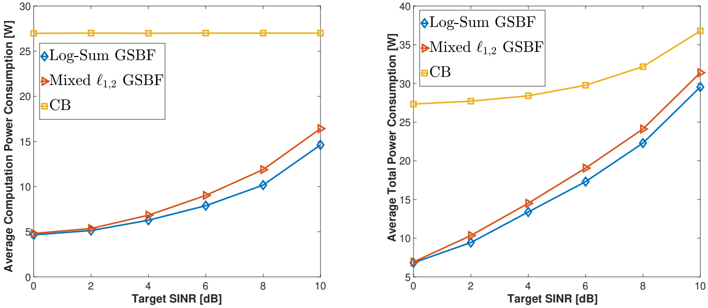
# Convergence Analysis

- **Challenges of Convergence analysis**
  - global convergence analysis of nonconvex $\ell_{2,p}$ minimization problems with linear constraints [Chen et al.'14]
  - global convergence analysis of unconstrained nonsmooth and nonconvex regularization problems [Ochs et al.'15]
- **Goal**: derive the global convergence analysis of our nonconvex and nonsmooth problem with general convex constraints
- add more details here

# Outline

# Simulation Results

- Simulation results averaged over 100 channel realizations with $N = 6, K = 10, L = 2$. Benchmark: coordinated beamforming and mixed $\ell_{1,2}$-norm based group sparse beamforming.




## Remark

- The proposed log-sum based group sparsity inducing norm can successfully decrease the number of performed tasks while satisfying pre-defined QoS, thereby yielding less power consumption.

# Outline

# Concluding Remarks

- **Joint task selection and transmit beamforming design problem** in green edge computing
  - Log-sum based group sparsity inducing approach
- Convergence analysis of the re-weighted $\ell_1$ algorithm
  - describe more details here

**Thanks !**