OPEX-Limited 5G RAN Slicing: an Over-Dataset Constrained Deep Learning Approach

Hatim Chergui and Christos Verikoukis, CTTC
Use of AI/ML in Networks Workshop
May 27, 2020
Outline

• 5GSolutions Concept
• Motivations
• Contributions
• C-RAN Setup and Dataset
• Constrained DNN Concept
• Offline Violation Rate-Based OPEX Enforcement
• Results
5GSolutions Concept

• Vertical domains of Factories of the Future, Smart Energy, Smart Cities, Smart Ports, and Media & Entertainment

• Mapped with the eMBB, URLLC and mMTC service classes
Motivations

• Reduce OPEX: Softwarization and virtualization technologies employed in network slicing,
• Joint network slicing OPEX control and resource allocation
• Novel constrained DNN models performing offline learning from datasets.
Contributions

- Joint multi-slice DNN model for resource provisioning based on the traffic per slice,

- Live network key performance indicators (KPIs) datasets,

- Constraints on OPEX violation rate:
  - Dataset-dependent custom non-convex constraints to the DNN output,
  - Use of a two-player non-zero sum game strategy.
• LTE-advanced (LTE-A) dense urban area, covered by 440 LTE-A eNodeBs (eNBs) and 3200 cells.

| Entity                | Quantity                                      |
|-----------------------|-----------------------------------------------|
| TRP                   | 3200                                          |
| eNB                   | 440                                           |
| BBU datacenters       | 10 uniformly distributed, with 100 CPU resources compared to a single 4G eNodeB |
CRAN Setup and Dataset (2/2)

- Two datasets sources:
  - Dedicated probes—collecting and analyzing the traffic per OTT
  - Key performance indicators collected by the operational support system (OSS) platform at TRP, eNB and vBBU levels.

| Feature                  | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| TRP                      |                                                                             |
| OTT Traffics per TRP     | Includes the hourly traffic for the top OTTs: Apple, Facebook, Facebook Messages, Facebook Video, Instagram, Netflix, HTTPS, QUIC, WhatsApp, and Youtube |
| CQI                      | Channel quality indicator reflecting the average quality of the radio link of the TRP |
| MIMO Full-Rank           | Usage of MIMO full-rank spatial multiplexing in %                            |
| DLPRB                    | Number of occupied downlink physical resource blocks                         |
| vBBU                     |                                                                             |
| OTT Traffics per eNB     | Aggregated OTT traffics per eNB                                              |
| CPU Load                 | CPU resource consumption in %                                                |
| RRC Connected Users      | Number of RRC users licenses consumed per eNB                                |
| Backhaul                 |                                                                             |
| OTT Traffics per BBU     | Aggregated OTT traffics per BBU datacenter                                   |
| Backhaul capacity        | Effective aggregated throughput per BBU datacenter                           |
Constrained DNN Concept

- Minimize DNN loss function subject to data-dependent constraints, expressed in terms of expectations over a data distribution $\mathcal{D}$:

$$\min_{\mathbf{W}} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \ell_0 (\mathbf{x}, \mathbf{W}),$$

$$s.t. \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \ell_i (\mathbf{x}, \mathbf{W}) \leq 0, i = 1, \ldots, m,$$

- Where $\mathbf{W}$ are the weights of the DNN, $\mathbf{x}$ are the features, while $\ell_0$ and $\ell_i$ stand for the DNN loss function and the $m$ constraints, respectively.
Offline Violation Rate-Based OPEX Enforcement (1/3)

- Pay-per-use strategy RAN resource pricing $\pi$:
  
  $$\pi\left(\hat{r}^{(i)}_{m,n,k}\right) = \gamma_{m,n,k} r^{(i)}_{m,n,k}$$

- Example: Amazon Web Services/Elastic Compute Cloud (EC2)

- Offline approach to train dataset-based DNN models.
  - Directly enforcing an upper bound on the OPEX violation rate:
    
    $$\min \frac{1}{NB} \sum_{i=1}^{N_B} \ell\left(\hat{r}^{(i)}_{m,n,k}, \hat{r}^{(i)}_{m,n,k} (W_n, b_n, s_n)\right),$$
    
    s.t. $W_{l,n} \in \mathbb{R}^{N_{l-1} \times N_l}, l = 1, \ldots, L + 1,$
    
    $b_{l,n} \in \mathbb{R}^{N_l \times 1}, l = 1, \ldots, L + 1,$
    
    $$\frac{1}{NB} \sum_{i=1}^{N_B} \mathbb{1}\left(\pi\left(\hat{r}^{(i)}_{m,n,k}\right) < \alpha_{m,n,k}\right) \leq \rho_{m,n,k},$$
    
    $$\frac{1}{NB} \sum_{i=1}^{N_B} \mathbb{1}\left(\pi\left(\hat{r}^{(i)}_{m,n,k}\right) > \beta_{m,n,k}\right) \leq \rho_{m,n,k},$$

- $\rho$ is the target threshold, $\alpha$ and $\beta$ are the bounds.
• **Problems:**
  - Nonconvex objective and constraint functions.
  - The violation rate constraint is a linear combination of indicators,

\[
\Phi_1(W_n) = \frac{1}{N_B} \sum_{i=1}^{N_B} \left( \pi \left( \hat{r}_{m,n,k}^{(i)} < a_{m,n,k} \right) - \rho_{m,n,k} \right),
\]

\[
\Phi_2(W_n) = \frac{1}{N_B} \sum_{i=1}^{N_B} \left( \pi \left( \hat{r}_{m,n,k}^{(i)} > \beta_{m,n,k} \right) - \rho_{m,n,k} \right),
\]

• **Solution:**
  - Sufficiently-smooth approximations of the constraints

\[
\Psi_1(W_n) = \frac{1}{N_B} \sum_{i=1}^{N_B} \sigma \left( a_{m,n,k}^{(i)} - \pi(\hat{r}_{m,n,k}) \right) - \rho_{m,n,k} \leq 0,
\]

\[
\Psi_2(W_n) = \frac{1}{N_B} \sum_{i=1}^{N_B} \sigma \left( \pi(\hat{r}_{m,n,k}) - \beta_{m,n,k} \right) - \rho_{m,n,k} \leq 0,
\]
• Proxy Lagrangian framework [R1]:

\[ \mathcal{L}_{W_n} = \frac{1}{N_B} \sum_{i=1}^{N_B} \ell (r_{m,n,k}^{(i)}, \hat{r}_{m,n,k}^{(i)} (W_n, b_n, s_n)) \]

Lagrangian w.r.t. weights

\[ + \lambda_1 \Psi_1(W_n) + \lambda_2 \Psi_2(W_n), \]

Lagrangian w.r.t. \( \lambda \)

\[ \mathcal{L}_\lambda = \lambda_1 \Phi_1(W_n) + \lambda_2 \Phi_2(W_n), \]

• Equivalent to a non-zero-sum two-player game in which the \( W_n \)-player wishes to minimize \( \mathcal{L}_{W_n} \), while the \( \lambda \)-player wishes to maximize \( \mathcal{L}_\lambda \).

• R measures the dependency to the constraints.

[R1] A. Cotter et al., “Training well-generalizing classifiers for fairness metrics and other data-dependent constraints” [Online]. Available: arxiv.org/abs/1807.00028.
Results (1/4)

- **eMBB**: Netflix, Youtube and Facebook Video,

- **Social Media**: Facebook, Facebook Messages, Whatsapp and Instagram,

- **Browsing**: Apple, HTTP and QUIC.

- **Training dataset sizes**:
  - 21417 samples at TRPs
  - 9681 samples at vBBUs levels
  - Batch size $N_B = 100$. 
• The achieved violation rate is a decreasing function of R

• To achieve the target violation rate $\rho = 0.005$ for the three considered slices, one should set $R = 0.2$. 

![Graph showing DL PRB OPEX violation rate vs. R with specific parameters and target ρ = 0.005.](image)
Results (3/4)

- With $R = 0.2$, the DLPRB OPEX bounds are respected,

- The slices differ in the incurred hourly OPEX due to the difference in the unitary price,

- DL PRBs variation over time is induced by the trend of hourly traffics per slice

- Massive access for Social Media and Browsing
• With $R = 0.2$, the enforced OPEX upper bounds = [2000; 1000; 500] $ are respected.

• eMBB service is presenting the lowest number of users but requires a backhaul capacity comparable to the other slices.

Backhaul capacity and OPEX distribution per slice, with $\alpha = [0, 0, 0]$ and $\beta = [2000, 1000, 500]$ $, \gamma = [5, 2, 1], \rho = 0.005$ and $R = 0.2$. 
