Dynamic deployment of multi-UAV base stations with deep reinforcement learning

Guanhan Wu, Weimin Jia, and Jianwei Zhao
Xi’an Research Institute of Hi-Tech, Xi’an, Shaanxi, China
Email: zhaojianweiep@163.com

Unmanned aerial vehicles (UAVs) can be utilized as aerial base stations (BSs) to provide auxiliary communication services. In this letter, we propose a deep reinforcement learning (DRL)-based dynamic deployment method for multi-UAV communications. The phasic policy gradient (PPG) is designed to improve the sample efficiency and the attention of the multi-UAV deployment. Simulation results are provided to verify the effectiveness of the proposed method.

Introduction: Unmanned aerial vehicles (UAVs) equipped with miniaturized communication devices can be served as aerial base stations (BSs) to provide auxiliary communication services for the ground users (GUs), which have become a promising way to meet the communication demands in the 6G and beyond [1–2].

Since the transmit power of UAV-BS is limited, the reasonable deployment locations of UAV-BSs are essential to improve the spectrum demands in the 6G and beyond [1–2]. Since the transmit power of UAV-BS is limited, the reasonable deployment locations of UAV-BSs are essential to improve the spectrum demands in the 6G and beyond [1–2].

System model: We consider a scenario where M UAV-BSs serve N GUs. We denote \((x_i, y_i, z_i), t \in \{1, 2, ..., M\}, (x_j, y_j, 0), j \in \{1, 2, ..., N\}\) the location of UAV \(i\) and GU \(j\), respectively. All the UAVs adapt to the GUs’ movements in the target area and track them to ensure that the GUs are within the coverage. The locations of UAV \(i\) and GU \(j\) at time slot and \(\Delta t\) is regarded as a time interval to update the GUs’ locations, \(t \in \{1, 2, ..., T\}\). The locations of UAV \(i\) and GU \(j\) at time slot \(t\) can be re-expressed as \(u^{\text{LoS}}_i(t)\) and \(u^{\text{NLoS}}_i(t)\). Besides, each UAV is equipped with an ultra-large-scale multiple-input multiple-output (MIMO) antenna array to form a high-precision directivity beam to reduce the interference.

The air-to-ground (ATG) channel is composed of a line of sight (LoS) and non-line of sight (NLoS) links, and the small-scale multipath fading is neglected for simplicity. Furthermore, according to a large number of experimental results, we can obtain the statistically significant probabilistic model and non-line of sight (NLoS) links, and the small-scale multipath fading can be expressed as:

\[
P_{\text{LoS}}^{\text{LoS}} = \frac{1}{1 + a \exp(-b \|\psi_{ij}\| - c)}
\]

where \(\psi_{ij}\) is the elevation angle of the communication link between the UAV \(i\) and GU \(j\), and \(a\) and \(b\) are the environment-related parameters. At the same time, the probability of NLoS links can be denoted as \(P_{\text{NLoS}}^{\text{NLoS}} = 1 - P_{\text{LoS}}^{\text{LoS}}\). The path loss of LoS and NLoS links can be respectively defined as:

\[
L_{\text{LoS}}^{\text{LoS}} = 20 \log \left(\frac{4\pi f_i d_{\text{LoS}}}{c}\right) + \xi_{\text{LoS}}.
\]

\[
L_{\text{NLoS}}^{\text{NLoS}} = 20 \log \left(\frac{4\pi f_i d_{\text{NLoS}}}{c}\right) + \xi_{\text{NLoS}}.
\]

where \(f_i\) is the carrier frequency, \(d_{\text{LoS}} = \|u^{\text{LoS}}_i - u^{\text{LoS}}_j\|\) is the distance from the UAV \(i\) to the GU \(j\), \(\xi_{\text{LoS}}\) and \(\xi_{\text{NLoS}}\) are additional path loss, and \(c\) is the speed of light.

Therefore, the pathloss model of ATG channel can be expressed as:

\[
L_{\text{Total}} = P_{\text{LoS}}^{\text{LoS}} + P_{\text{NLoS}}^{\text{NLoS}}.
\]

UAVs are required to cover as many moving GUs as possible in the shortest possible time, so we can get the following optimization objectives as:

\[
\max_{\pi} \sum_{t=0}^{T} \mathbb{E}_s \left[ \sum_{j=1}^{N} \min[d_{\text{LoS}}^{\text{LoS}}] \leq R, j \in \{1, 2, ..., N\},
\]

\[
\leq x_{\max},
\]

\[
|\theta'_i - \theta''_i - 1| \leq \Delta \theta_{\max},
\]

where \(T_{ij}\) is the connection relationship between UAV \(i\) and GU \(j\). Moreover, \(T_{ij} = 1\) if the GU \(j\) is within the coverage of UAV \(i\), and otherwise \(T_{ij} = 0\). \(R\) is the rated coverage radius. (5b) and (5c) denote that each GU can only form a connection with at most one UAV. (5d) and (5e) are the limits of UAVs’ flight range which make which work in the target area. \(\theta'_i\) denotes the distance that UAV \(i\) fly in the time slot, \(t_i \in [0, \infty]\). \(\nu_{\text{max}}\) means the maximum speed of the UAV-BSs, \(\theta''_i\) denotes the azimuth angle of the UAV \(i\) in time slot, \(t_i \in [0, 2\pi]\). (5f) and (5g) represent the constraints on UAVs’ flight trajectory. \(\Delta \theta_{\max}\) denotes the maximum steer angle, which can prevent azimuth angle due to excessive changes make the flight trajectory smoother.

DRL can adopt the offline trained policy model for deployment, the action \(a_i\) to be performed is directly mapped from the current state \(s_i\). We define \(Q^{\pi}(s, a) = \mathbb{E}[R|s_i=s, a_i=a]\) as the expected return with policy \(\pi\), where \(R_i = \sum_{t=0}^{T} \gamma^t r'\) for the parameters of actor-network, and \(\gamma\) is the discount factor. Our goal is to find the robust and general policy \(\pi^*\) to maximize the cumulative reward in the period \(T\):

\[
\pi^* = \arg\max_\pi \mathbb{E}_s \left[ \sum_{t=0}^{T} r' + \gamma S(s_{t+1}) \right].
\]

where \(S(s_{t+1})\) is the entropy of policy \(\pi\) at the state \(s_{t+1}\), \(\gamma\) is the trade-off coefficient. In the model training phase, we utilize stochastic policy \(\pi(s_{t+1})\) to select the next action \(a_i\) that to be performed for the state \(s_i\), and the entropy offers the possibility of more experience for the policy \(\pi(s_{t+1})\) so that the critic can more accurately evaluate the state values \(V(s_{t+1})\), \(\phi\) are the parameters of the critic network. The actual state value is defined as:

\[
V(s) = \mathbb{E}_s \left[ \sum_{t=0}^{T} \gamma^t r' | s_i = s \right] = \sum_{s'=s} \pi(s'|s)Q(s, a).
\]

The MDP is modelled by a tuple \((S, A, P, R, \gamma)\), where \(S\) is the state space, \(s_i \in S\), \(A\) is the action space, \(a_i \in A\), \(P\) is the transition probability, and \(R\) is the reward space. Since the experience samples can not be reused, the traditional policy gradient is an inefficient method while proximal policy optimization (PPO) is a possible alternative with stable performance. We define the objective function of our actor like the PPO way:

\[
J(\theta) = \mathbb{E}_s \left[ \min \left( c l i p \left( \frac{\pi_{\text{old}}}{\pi_{\text{new}}}, 1 - \varepsilon, 1 + \varepsilon \right) \frac{\pi_{\text{new}}}{\pi_{\text{old}}}, \frac{\pi_{\text{old}}}{\pi_{\text{new}}} \right) + \gamma S(s_{t+1}) \right].
\]
where $\hat{A}_t$ is the advantage function estimated by critic that used to measure the superiority of the action $a_t$ taken by the current policy $\pi_a$ at state $s_t$. To evaluate it more accurately, we apply the generalized advantage estimator (GAE) method to control the bias and variance of the estimation in a relatively balanced manner, which is given by:

$$\hat{A}_t = \sum_{j=0}^{t-1}(\gamma \lambda)^j \sigma_t^j, \quad (9)$$

$$\sigma_t^j = r_t + \gamma V(s_{t+1}) - V(s_t). \quad (10)$$

where $\lambda \in (0, 1)$, which indicates the different emphasis on bias and variance. To evaluate the advantage function of critic more accurately, we define its loss function as:

$$J(\phi) = \left[ V_{\phi}(s) - \hat{V}_{\phi}(s) \right]^2, \quad (11)$$

where $\hat{V}_{\phi}$ is the value target also calculated by the GAE method.

To better share the extracted feature parameters between actor and critic networks and to avoid the interference caused by sharing the underlying network, we use the separated actor and critic networks along with an auxiliary task to induce actor to learn from critic, i.e. in addition to the output policy $\pi_a(s)$, an additional value head $V_{\phi}(s)$ is set to predict the state value in the actor-network. After a period of learning in the normal way with (8) and (11), we further let the value head $V_{\phi}(s)$ learn the state value $V(s_t)$ using the past experience samples:

$$J_{joint}(\theta) = \mathbb{E}_s [V_{\phi}(s) - \hat{V}_{\phi}(s)]^2 + \eta \cdot \mathbb{E}_s [KL(\pi_{\phi_{\text{critic}}}(s)||\pi_{\phi_{\text{actor}}})), (12)$$

where $KL(\pi_{\phi_{\text{critic}}}(s)||\pi_{\phi_{\text{actor}}})$ is the KL divergence of $\pi_{\phi_{\text{critic}}}$ and $\pi_{\phi_{\text{actor}}}$, is the current policy after the first phase (learning in the normal way). In this way, the actor can learn the state value while improving its policy and better distil features. The KL divergence ensures that the new and old policies not to deviate too much. Then, the above two phases are then repeated periodically.

We set $s_t = (a_{t-1}^{(j)}, d_{dis}^{(j)}, N_t, d_{dis})$ as the state vector of UAVs, and $a_t = (a_t^{(j)}, \theta)$ as the joint action of UAVs. $N_t$ denotes as the covered GUs at time slot $t$, and variable $d_{dis}^{(j)} = d_t^{(j)}, i \in \{1, 2, \ldots, M\}$ indicates the distance between each UAV and the centre of GUs area, which is given by (13):

$$d_t^{(j)} = \sqrt{(x_t^{(j)} - x_{\text{mean}}^{(j)})^2 + (y_t^{(j)} - y_{\text{mean}}^{(j)})^2}, \quad (13)$$

$$x_{\text{mean}}^{(j)} = \frac{1}{N} \sum_{j=1}^{N} x_t^{(j)}, \quad y_{\text{mean}}^{(j)} = \frac{1}{N} \sum_{j=1}^{N} y_t^{(j)}. \quad (14)$$

Then, the reward function is defined as:

$$r_t = \xi_{\text{cov}} \Delta n_t + \xi_{\text{dis}} \exp \left( \frac{N_t}{N} \right) + \xi_{\text{cov}} \kappa_t - \xi_{\text{dis}} d_{\text{mean}}, \quad (15)$$

where $\Delta n_t = N_t - N_{t-1}$ is the changes of covered GUs, $\exp(\frac{N_t}{N})$ is the function of coverage rate that its exponential form provides an additional gradient to the reward. $\kappa_t$ is the number of coordinates that the UAVs cross the boundary, and $d_{\text{mean}} = \sum_{j=1}^{M} d_t^{(j)}$ is the sum of the distance between each UAV and GUs’ central location. Moreover, $\xi_{\text{cov}}, \xi_{\text{dis}}, \xi_{\text{cov}}$ and $\xi_{\text{dis}}$ are coefficients that control the size of each reward items.

In actual, we expand the proportion of UAV position information in the state vector. i.e. we use a feature extraction layer to preprocess the UAV position information and then combine other dimensional information into the actor and critic networks. This allows the neural network to better extract the features of the UAVs’ motion.

**Simulations and results:** In this section, simulation results are provided to verify the performance of the proposed method. Additionally, particle swarm optimization (PSO) and static deployment (SD) algorithms are compared to verify the effectiveness and superiority of the proposed method. We use the PSO algorithm to find the optimal deployment location for each episode to maximize coverage, while SD refers that the cumulative rewards are raised slowly. PPO further utilizes the existing data through the learning of $V_{\phi}(s_t)$ in an off-policy manner, which allowing actor and critic to share the objective. This enables feature sharing between the two networks and accelerates feature learning.

**Conclusions:** In this letter, we proposed a DRL-based dynamic deployment approach for UAVs to provide services for the moving GUs in the poor QoS area. The PPG approach combined with auxiliary task
learning enables better extraction of state features, and the entropy bonus of the actor allows the decision agent to maintain a more random strategy during training to increase exploration. Finally, the effectiveness of the proposed algorithm is verified by simulation experiments.

Acknowledgments: This research was sponsored by the National Natural Science Foundation of China (grant no. 62001500).

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
Received: 18 March 2021 Accepted: 14 April 2021
doi: 10.1049/ell2.12205

References
1 Sekander, S., et al.: Multi-Tier drone architecture for 5G/6G cellular networks: Challenges, trends, and prospects. IEEE Commun. Mag. 56(3), 96–103 (2018)
2 Ghanavi, R., et al.: Efficient 3D aerial base station placement considering users mobility by reinforcement learning. 2018 IEEE Wireless Communications and networking Conference, Barcelona, Spain, 15–18 April 2018
3 Schulman, J., et al.: Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347: (2017). Accessed 20 July 2017
4 Cobbe, K., et al.: Phasic policy gradient. https://arxiv.org/abs/2009.04416: (2020). Accessed 9 September 2020
5 Saxena, V., Jalden, J., Klessig, H.: Optimal UAV base station trajectories using flow-level models for reinforcement learning. IEEE Trans. Cog. Commun. Netw. 5(4), 1101–1112 (2019). https://doi.org/10.1109/tccn.2019.2948324