Deep Convolutional Self-Attention Network for Energy-Efficient Power Control in NOMA Networks

Abuzar B. M. Adam, *Member, IEEE*, Lei Lei *, *Member, IEEE*, Symeon Chatzinotas, *Senior Member, IEEE*, and Naveed Ur Rehman Junejo

***Abstract*—In this letter, we propose an end-to-end multi-modal based convolutional self-attention network to perform power control in non-orthogonal multiple access (NOMA) networks. We formulate an energy efficiency (EE) maximization problem we design an iterative solution to handle the optimization problem. This solution can provides an offline benchmark but might not be suitable for online power control therefore, we employ our proposed deep learning model. The proposed deep learning model consists of two main pipelines, one for the deep feature mapping where we stack our self-attention block on top of a ResNet to extract high quality features and focus on specific regions in the data to extract the patterns of the influential factors (interference, quality of service (QoS) and the corresponding power allocation). The second pipeline is to extract the shallow modality features. Those features are combined and passed to a dense layer to perform the final power prediction. The proposed deep learning framework achieves near optimal performance and outperforms traditional solutions and other strong deep learning models such as PowerNet and the conventional convolutional neural network (CNN).**

***Index Terms*—non-orthogonal multiple access (NOMA), energy efficiency (EE), power control, convolutional neural network (CNN), self-attention.**

I. INTRODUCTION

Energy efficiency (EE) is one of the widely adopted performance metrics. Therefore, several studies have investigated EE maximization in non-orthogonal multiple access (NOMA) networks [1]. Due to the non-convexity of EE maximization problem, the global optimum is difficult to obtain when considering the conventional optimization methods [2]. Considering the studies in [3], [4], the authors decoupled the problem first then proposed suboptimal solutions for the subproblems. The solutions are iterative and might not be suitable for the real-time application. Additionally, it has been proved that sum of ratios problem (SoRP); precisely the weighted sum energy efficiency (WSEE) maximization problem in our case is NP-hard optimization problem even when each ratio has concave numerator and linear denominator [5], due to the fact that pseudo-concavity or quasi-concavity properties are guaranteed to be preserved in case of addition.

The upcoming fifth generation (5G) and beyond 5G (B5G) networks seek more favorable solutions with lower computational complexity to guarantee strict run time or computational delay requirements [6]. In this regard, deep learning has become a promising tool in overcoming the issues in conventional optimization methods. In the literature, deep reinforcement learning (DRL), deep neural network (DNN) [7], [8] and convolutional neural network (CNN) [2] are widely used in physical-layer communications and resource optimization, e.g., power control [9]. DNN-based power control entails some shortcomings. First, for multiple interfered links, the input is two-dimensional while DNN accepts one-dimensional input. To overcome this problem, DNN employs vectorization of the input. However, this is not practical for large problems. Second, the current DNN based power control algorithm is centralized. To obtain the optimal power control, the instant channel state information (CSI) on all the links in the network must be known for the base station (BS) which is deemed unrealistic in case of employing DNN to solve the power control in large networks since it would cause significant delay [2].

Motivated by the previous research work in the literature and the recent advancements in CNN architecture, we introduce a modified version of the end-to-end multi-modal based convolutional self-attention network [10], [11]. Our contributions can be summarized as:

- To the best of our knowledge, this study the first work to investigate the potentials of this new CNN architecture for NOMA power control. Different from the studies in [2] and [7], we leverage the multi-modality technique to improve the power control prediction through fusion of the deep modality features and the shallow modality features. Because the deep modality features enhance the semantic features and the shallow modality features preserve the spatial details, this fusion makes them more discriminative.
- We employ the self-attention to focus on the interference and the QoS factors to enhance the power control.
- Furthermore, different from the original model in [11], we employ max pooling on top of the self-attention block to enhance the prediction power of the model since it achieves translation invariance.
- Different from [7], our proposed deep learning model leverages the window size to overcome the vectorization. This allows BS passing CSI of multiple links faster to the model. Hence, this makes our model more suitable

[*] Corresponding author: Lei Lei (lei.lei@uni.lu)

A. B. M. Adam is with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, P. R. China, 400065.

L. Lei, and S. Chatzinotas are with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg City 1855, Luxembourg.

Naveed Ur Rehman Junejo is with Department of Computer Engineering, University of Lahore, Lahore, Punjab, Pakistan.

for real-time application and the centralization problem is alleviated as well. Additionally, unlike the model in [7], our model can be trained with different channel models simultaneously without losing the spatial and semantic features.

II. System Model and Problem Formulation

We consider downlink multi-cell NOMA system where each cell includes a BS in its center. The set of cellular users is denoted as $\mathcal{N} = \{1, 2, ..., N\}$ and the users are associated with the serving BS via K subchannels. Each subchannel k has a bandwidth W_k. The BS has a power budget denoted by P_{max}. Let $p_{i,k}$ represents the allocated power of the user i on the subchannel k, the BS sends a superimposed symbol s to the N_k users multiplexed on the subchannel k. Hence, the transmitted signal by the BS on the subchannel k is given as

$$x_k = \sum_{i=1}^{N} \sqrt{p_{i,k}} s_i \tag{1}$$

The received signal by the user i is written as

$$\begin{aligned} y_{i,k} &= h_{i,k} x_k + \eta_{i,k} \\ &= h_{i,k} \sqrt{p_{i,k}} s_i + \sum_{j=1, j \neq i}^{N_k} h_{i,k} \sqrt{p_{j,k}} s_j + \eta_{i,k} \end{aligned} \tag{2}$$

where $h_{i,k} = g_{i,k} \zeta_{i,k} + \varepsilon_{i,k}$ is the channel coefficient between the BS and the user i where $\varepsilon_{i,k}$, $g_{i,k}$ and $\zeta_{i,k}$ represent the channel error, the small-scale fast-fading and the large-scale fading coefficients, respectively. $\eta_{i,k}$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2. According to NOMA protocol, the successive interference cancellation (SIC) technique is applied at each receiver. According to the SIC process, the user with the better channel conditions can remove the interference from other users with the poor channel conditions on the same subchannel. Therefore, without loss of generality, the following order can be assumed

$$\left|h_{1,k}\right|^2 \geq \left|h_{2,k}\right|^2 .... \geq \left|h_{N_k,k}\right|^2 \tag{3}$$

The signal-to-interference-plus-noise ratio (SINR) of the user i on the subchannel k with SIC can be given by

$$\gamma_{i,k} = \frac{p_{i,k} \left|h_{i,k}\right|^2}{\sigma^2 + \sum\limits_{j=1}^{i-1} \left|h_{i,k}\right|^2 p_{j,k}} \tag{4}$$

The data rate of the user i can be defined as below

$$r_{i,k} = W_k \log_2 \left(1 + \frac{p_{i,k} \left|h_{i,k}\right|^2}{\sigma^2 + \sum\limits_{j=1}^{i-1} \left|h_{i,k}\right|^2 p_{j,k}} \right) \tag{5}$$

Considering the WSEE, our optimization problem is given as follows

$$\begin{aligned} \max_{\{p_i \geq 0\}_{i=1}^{N}} \quad & \sum_{k=1}^{K} \sum_{i=1}^{N} \omega_{i,k} \frac{r_{i,k}}{p_{i,k} + p_c} \\ \text{s.t.} \quad & C1 : r_{i,k} \geq R_{\text{min}}, \\ & C2 : p_{i,k} \leq P_{\text{max}}, \end{aligned} \tag{6}$$

where $\omega_{i,k}$ is the weight of the user i on the subchannel k and p_c is the circuit power consumption. The constraints $C1$ represent QoS requirements and $C2$ to state that the user power consumption should be less than the maximum power budget.

In this work, our goal is to design energy-efficient power allocation algorithm to handle the above problem. However, there are multiple challenges need to be addressed. First, the above problem is multiple ratio problem SoRP which is NP-hard [12]. Second, sum of ratios function is neither pseudo-concave or quasi-concave; even if both numerator and denominator are affine [5]. Hence, developing an algorithm to solve (6) with reasonable complexity is still beyond the reach of the most of the known methods especially for large number of rations. Third, the high computational nature of the solution of this problem makes the real-time application remarkably challenging. Besides, any changes in the channel conditions will require instant adaptation in the proposed solution.

From the above, the proposed solution consists of two stages. In the first stage, we design an iterative solution to obtain suboptimal solution. In the second stage, we design a deep learning framework to enable real-time power control.

III. An Iterative Solution for WSEE Maximization Problem

Some studies advocate applying successive convex approximation [3], [13], however, none of them could obtain the global optimal solution in the polynomial time. Furthermore, applying the lower bound approximation on (6) will lead to sum of pseudo-concave functions which is not guaranteed to be pseudo-concave [5]. Moreover, it is difficult to extend Dinkelbach's method to multiple ratio fractional problem scenario [14]. Besides, the optimality of the power vector in maximizing EE is crucial to the convergence of Dinkelbach's method.

Nevertheless, we can develop an iterative suboptimal algorithm but yields the Karush-Kuhn Tucker (KKT) point. Hence, performing relaxation, we have the following Lagrangian function

$$\mathcal{L}(p, \lambda, \beta, \upsilon) = \sum_{k=1}^{K} \left[\begin{array}{l} \sum\limits_{i=1}^{N} \omega_{i,k} \frac{r_{i,k}}{p_{i,k}+p_c} + \lambda p_{i,k} - \beta \left(P_{\text{max}} - p_{i,k}\right) \\ - \sum\limits_{i=1}^{N} \upsilon_{i,k} r_{i,k} - R_{\text{min}} \end{array} \right] \tag{7}$$

Our power allocation $p_{i,k}$ can be written as the following fixed-point

$$p_{i,k} = \frac{1}{\ln 2} \frac{\varphi_{i,k} - \upsilon_{i,k}}{\beta - \lambda + \xi_i} - \frac{\sigma^2 + I_{i,k}}{\left|h_{i,k}\right|^2} \tag{8}$$

where λ, β and $\upsilon_{i,k}$ are Lagrange multipliers associated with the nonnegativity of the power allocation, maximum sum power allocation and the minimum rate, respectively. $\varphi_{i,k}$ and $\xi_{i,k}$ are given as follows

$$I_{i,k} = \sum_{j=1, j \neq i}^{i-1} \left|h_{i,k}\right|^2 p_{j,k} \tag{9}$$

$$\varphi_{i,k} = \frac{\omega_{i,k} W_{i,k}}{p_{i,k} + p_c} \tag{10}$$

$$\xi_{i,k} = \frac{\omega_{i,k} W_k \log_2\left(1 + \frac{p_{i,k}\left|h_{i,k}\right|^2}{\sigma^2 + \sum\limits_{j=1, j\neq i}^{i-1} \left|h_{i,k}\right|^2 p_{j,k}}\right)}{\left(p_{i,k} + p_c\right)^2} \tag{11}$$

To obtain the final power allocation, first we initialize the power into a feasible value, then update it by solving the following problem

$$\begin{cases} p_{i,k} = \max\left\{0, \frac{1}{\ln 2}\frac{\varphi_{i,k} - \upsilon_{i,k}}{\beta - \lambda + \xi_{i,k}} - \frac{\sigma^2 + I_{i,k}}{\left|h_{i,k}\right|^2}\right\} \\ p_{i,k} \leq P_{\max} \\ \left|h_{i,k}\right|^2 p_{i,k} \leq Q \end{cases} \tag{12}$$

where Q represents the permittable interference level caused by the user i to other users. Algorithm 1 includes the steps of the iterative solution.

Algorithm 1 Iterative Power Allocation for WSEE Maximization

```
Initialization: $R_{min}, p_{i,k}^{(0)}, \beta^{(0)}, \epsilon, t = 0, Q, \upsilon_{i,k}$.
while $\left|p_{i,k}^{(t+1)} - p_{i,k}^{(t)}\right| > \epsilon$ do
    $t = t + 1$
    Calculate $I_{i,k}$, $\varphi_{i,k}$ and $\xi_{i,k}$ using (9), (10) and (11)
    Calculate $p_{i,k}^{(t)}$ by solving (12)
end while
```

Dinkelbach-like algorithm converts the original problem into a sequence subproblems in the form $\max\limits_{x,y} \sum\limits_{i=1}^{N} \left(f_i(x) - y g_i(x)\right)$. Assuming the number of iterations to calculate the subproblems and the sub-gradients for the multipliers are T_s and T_g, respectively. The calculation of (8) entails NK operations. The update of the multipliers entails $O(N)$. The convergence of the loop in Algorithm 1 can be obtained with complexity $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$. The total complexity of the power allocation scheme is $O\left(T_s T_g N^2 K \log\left(\frac{1}{\epsilon}\right)\right)$. The proposer selection of the initial values of the multipliers and the accuracy in calculating the step sizes have considerable impact on the number of iterations.

IV. MULTI-MODAL BASED CONVOLUTIONAL SELF-ATTENTION NETWORK

In this section, we introduce our modified end-to-end multi-modal based convolutional self-attention network (we abbreviate it here as MM-CSAN) to perform the power control. Fig. 1 shows the proposed structure of the network.

Assuming that users' CSI are collected and sorted on the BS. Hence, we assume full CSI at the BS. In practice, the CSI is updated every frame (consisting of a set of time slots), and keeps static within a frame. Thus, the collected CSI are passed as input to the next stage where the proposed MM-CSAN learns the patterns in the input to give more accurate power prediction.

The proposed framework works in two stages. The first stage is the deep feature extraction using deep CNN with self-attention network from the normalized data. The second stage is fusing the features extracted from the CNN and the shallow features, then we have dense layer for detection.

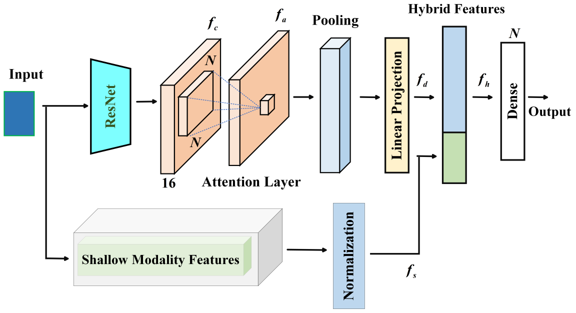

Fig. 1. Structure of the proposed deep multi-modal based convolutional self-attention network.

In the following subsections, we discuss the structure the proposed network in details.

A. Data Preparation, Training and Testing Procedures

Our goal is to find the function that maps the input of deep network pipeline $\{x_{i,k}^{d}\}$ and the input of shallow modality pipeline $\{x_{i,k}^{s}\}$ to the power given the training set of instances-label tuples $\{\{x_{i,k}\}, \{p_{i,k}\}\}$. $x_{i,k}$ is defined as below

$$x_{i,k} = \{x_{i,k}^{d}, x_{i,k}^{s}\} \tag{13}$$

where $\{x_{i,k}^{d}\}$ is a mapping of the user's channel coefficient $h_{i,k}$, and the corresponding threshold of interference $Q_{i,k}$ and R_{min}, and given as follows

$$x_{i,k}^{d} = \{h_{i,k}, h_{j,k}, Q_{i,k}, R_{min}\} \tag{14}$$

While $x_{i,k}^{s}$ is the shallow modality features in our case include the channel coefficient of the users who share the same subchannel with the user of interest (given here as $h_{j,k}$) and the channel conditions under large-scale fading channel components. Thus, $x_{i,k}^{s}$ is written as

$$x_{i,k}^{s} = \left\{h_{j,k}, h_{i,k}, \tilde{h}_{i,k}, \bar{h}_{i,k}\right\} \tag{15}$$

where $\tilde{h}_{i,k}$ and $\bar{h}_{i,k}$ represent the shadowing channel model and small-scale fading channel model, respectively. The inclusion of those components is to spatially connect the variations in the channel. Taking the advantages of the convolutional neural networks, the input data can be inserted as directly with no need for vectorization process due to the presence of window.

Suppose that the training input $x \in {}^{N\times W\times 1}$ represents the mapping $\{x_{i,k}\}$ where N and W are spatial dimensions, the desired output is $\{p_{i,k}\}$ and the predicted output by the neural network is $\{\hat{p}_{i,k}\}$. During the training, we aim to minimize the loss function $\mathcal{L} = \mathbb{E}\left[\left(p_{i,k} - \hat{p}_{i,k}\right)\right]$.

To generate the training data set, we consider a region of 1km×0.5km for simulation. The users are randomly and uniformly distributed over the region. Distance-based path loss can be conveniently obtained at any given spot within the region. However, for more practical approximation of the real-world fading channel, we generate shadowing and fast fading channels. Where $\zeta_{i,k} = 10^{-(PL-G)/10}$, $\zeta_{i,k} = 10^{-(PL+\upsilon-G)/10}$ and $\zeta_{i,k} = 10^{-(PL+\upsilon-G)/10}$ are the fading components for the distance-based path loss channel, shadowing channel and small-scale channel model, respectively. $g_{i,k} = 1$ for both path loss and shadowing channel models and $g_{i,k} \sim Rayleigh(1)$ for small-scale fading channel model. $\upsilon \sim \mathcal{N}\left(0, \varsigma^2\right)$ is the log-normal shadowing and ς is the standard deviation. The optimal solution is obtained using the exhaustive search method and MM-CSAN is trained with the optimal solution.

B. Structure of the Proposed Deep Learning Model

1) Deep CNN for Feature Extraction: To design our CNN feature extractor, first consider the input $x \in {}^{N\times W\times 1}$. The extracted convolution features $f_c \in {}^{N\times N\times 16}$ using the encoder (which is CNN with residual block [15] indicted as ResNet in Fig.1) are defined as

$$f_c = f_{\text{ResNet}}\left(W_a \otimes x\right) \tag{16}$$

The kernel size is 3×3. Then, f_c is passed to self-attention block to extract higher level features by focusing on specific region in the data.

2) Self-Attention Block for Enhanced Feature Extraction: In order to precisely predict the desired output based on the detection of specific pattern in the input data, the model focus on certain regions within the feature map of the input data. Hence, the feature map f_c is fed to self-attention block (see Fig.2) to focus on the regions of interest and extract deeper features. The convolutional feature map f_c is first rearranged to yield the map $f_c' \in {}^{N\times N\times 8}$. Then, f_c' is fed to the self-attention block to generate the feature map $\Theta_a \in {}^{N\times N\times c'}$. The feature map Θ_a is multiplied by scaling parameter θ and added to f_c' to finally obtain the self-attention feature map f_a as follows

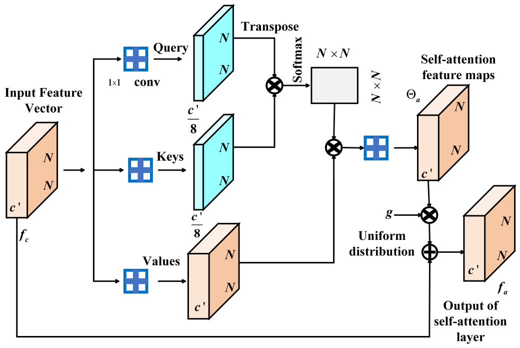

Fig. 2. An illustration of self-attention block.

$$f_a = \theta * \Theta_a + f_c' \tag{17}$$

f_a is passed through pooling and linear projection to obtain the deep features f_d which is the final feature of our main pipeline for deep feature extraction.

3) Hybrid Fusion of Deep and Shallow Modality Features: The shallow modality features are normalized and fed into flatten layer. Then, the normalized shallow modality and the high quality features from the main pipeline of our model are concatenated together to form the hybrid features f_h.

C. Working Mechanism

Consider the main two pipelines of the proposed deep network. In the main pipeline, we pass $\{x_{i,k}^d\}$ to the ResNet which is good in capturing the patterns in the data (i.e interference patterns) and hence, the feature map f_c is generated. The self-attention block focuses on specific regions ($Q_{i,k}$ and R_{min}) in the feature map f_c and extracts higher quality feature map f_a. These high-quality features represent the integrated global information obtained as result of interactions of those regions in the data throughout layers of the self-attention block. More specifically, the readable pattern between the spatial features such as the relationship between the interference threshold and the channel coefficient, QoS and channel coefficient. Multiple patterns are possible to be deeper focused on by the self-attention block. The scaling parameter θ is initialized using a uniform distribution, therefore it is updated via learning procedure. During the learning process, this parameter enables the network to focus on the desired region and its neighboring locals, then weight assignment procedure is used to differentiate the local and non-local regions. A max pooling operation followed by a linear projection are applied on f_a to finalize the deep feature map representation as f_d.

In the second pipeline, the shallow modality feature mapping f_s is obtained as a normalization of the input $\{x_{i,k}^s\}$. Then, concatenated with f_d to form the hybrid feature map f_h. Finally, we pass f_h to the dense layer to perform the prediction of the power allocation. To satisfy constraint $C2$, the activation function $f(x) = \frac{P_{\max}}{1+\exp(-x)} \in [0, P_{\max}]$ is used in the output layer.

V. Simulation Results

A. System Setup

The model is implemented in Keras 2.2.4 with TensorFlow 1.8.0 backend and Python 3.6 platform. The computer specifications are: 3.7GHZ Intel core i7, GeForce RTX 2080Ti graphic card, and 32GB memory. The number of training samples is 166000 and the number of testing samples is 6000. The training epoch is set to be 300 and for the nonlinearities; we have batch normalization with batch size is set to 100. The learning rate is 0.001 and the dropout is 0.05. We employed ADAM optimizer over the problem.

B. Performance Comparison

We consider two cells downlink NOMA network with one BS in the center of each cell, the cell radius is 500 m and the users are randomly distributed in the cell. We assumed two users are sharing one subchannel. The carrier frequency is 250 kHz. The number of subchannel is $K = N/2$, the noise figure is 7 dB and $p_c = 10$ dBm. We also considered different fading channel models for the generation of shallow modality features while the small-scale Rayleigh fading channel between users and BS is the main testing channel model. The noise power spectral density $N_0 = -174$ dBm/Hz. The error tolerance parameters ϵ in Algorithm 1 is set to 0.001. We considered fixed and bounded channel error 0.1.

The weights are generated randomly such that $\sum_{k=1}^{K} \sum_{i=1}^{N} \omega_{i,k} = 1$ and assigned to the users according to the SIC order; where the users with better channel coefficients are assigned with lower value weight.

TABLE I PERFORMANCE COMPARISON BETWEEN DIFFERENT FRAMEWORKS IN TERMS OF COMPUTATIONAL TIME

Approach	Number of Users per BS					
	20		40		60	
	Percentage	GPU time(ms)	Percentage	GPU time(ms)	Percentage	GPU time(ms)
Exhaustive Search	100%	26.330	100%	33.909	100%	38.4410
Algorithm 1	100%	15.013	100%	16.700	100%	17.966
MM-CSAN	1.490%	0.221	1.832%	0.234	1.883%	0.255
PowerNet [2]	1.409%	0.140	1.696%	0.161	2.025%	0.205
Conventional CNN	1.255%	0.112	1.291%	0.127	1.977%	0.203
DC	100%	15.114	100%	16.733	100%	18.001

For the comparison,first, exhaustive search is employed to find the optimal power allocation and to serve as an optimal benchmark to quantify the performance of the proposed frameworks. Additionally, different of two convex (DC) programming [16] and Algorithm 1 are considered as conventional low complexity power control schemes. To compare the proposed MM-CSAN with other deep learning models from the literature, we considered the PowerNet model in [2] which is a CNN with residual learning blocks. We also considered a conventional structure of CNN including two convolutional layers, two max pooling layers, flatten and fully connected layer. The fully connected layer includes three hidden layers and an output layer. Similarly as in our MM-CSAN, the kernel size is 3 × 3, max pooling layer with 2 × 2 filter and the activation function is ReLU. The dropout is set to 0.05 and the layer steps are 1 and 2 for the convolution and pooling, respectively.

We focus on the online computational complexity since the offline data generation and training have no impact on the real-time application. Table I shows the performance comparison between the proposed framework and other frameworks in terms the computation time over the GPU. For number of users $N = 20$ per BS, due to the iterative nature of exhaustive search method, Algorithm 1 and DC, MM-CSAN is 110 times faster than exhaustive search, 67 times faster than Algorithm 1 and 68 times faster than DC method. PowerNet is 1.6 times faster than MM-CSAN and this gap is shrinking with increasing in the number of users. For instance, when 60 per BS, PowerNet is just 1.2 times faster than MM-CSAN. However, PowerNet has poorer EE performance compared to MM-CSAN (e.g., see Fig. 3) and the performance gap is increasing with increasing of the problem size. Despite the conventional CNN is faster than all other models (e.g., 1.9 times faster than MM-CSAN when 20 per BS), it is the least energy-efficient among all (as it can be seen later). It is worth mentioning that the computational complexity of exhaustive search is exponential in the number of variables. Hence, the asymptotic complexity of exhaustive search and DC programming are $O\left(2^{(NK)^4}\right)$ and $O\left(T_{DC}N^3K^3\right)$, respectively. Where T_{DC} is number of iterations for DC programming. From all above, MM-CSAN is more suitable for large problem and more suitable for real-time application.

In Fig. 3, we investigate the impact of the power budget P_{max}. For lower value of P_{max}, EE performance for all frameworks is comparable because each user can transmit with its possible maximum power. However, for higher P_{max}, EE of all frameworks increase with the increasing of P_{max} and stagnate at about 8 dBm. The performance of MM-CSAN is closer to that of the exhaustive search. Algorithm 1 and DC has comparable performance. However, they are more energy efficient than both PowerNet and the conventional CNN. Nevertheless, the performance of MM-CSAN is clearly better than that of other frameworks except exhaustive search and that due to the capability of MM-CSAN of capturing the patterns of the interference and QoS while predicting the corresponding near optimal power allocation that matching these factors. EE stagnates at 5.7489 Mbits/Joule, 5.7090 Mbits/Joule, 5.6442 Mbits/Joule, 5.6457 Mbits/Joule, 5.6276 Mbits/Joule and 5.6139 Mbits/Joule for exhaustive search, MM-CSAN, DC, Algorithm 1, PowerNet and conventional CNN, respectively.

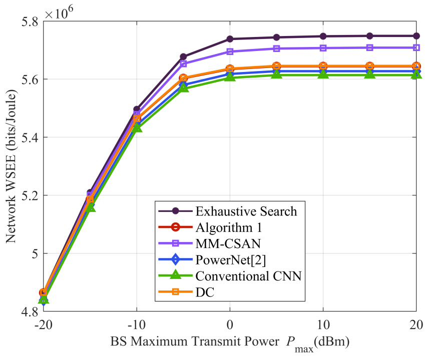

Fig. 3. Network EE for different values of the P_{max}.

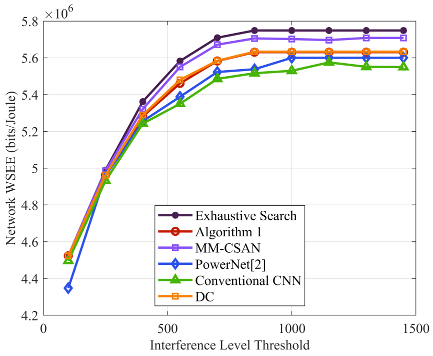

Fig. 4. Network EE vs permittable interference levels.

Fig. 4 depicts the performance with regards to the permittable levels of interference. When the permittable level of interference is low, EE increases. When the interference level is high enough, the allocated power causes severe interference leads to stagnation in EE. Again, due to the abilities of our deep learning model in capturing the influential parameters patterns, we can observe the learning stability in the curve of MM-CSAN. PowerNet shows less fluctuations in EE compared to the conventional CNN due to the its capability in capturing the interference patterns. MM-CSAN shows great performance closer to the optimal and comparatively better than that of Algortihm 1, DC, PowerNet and the conventional CNN.

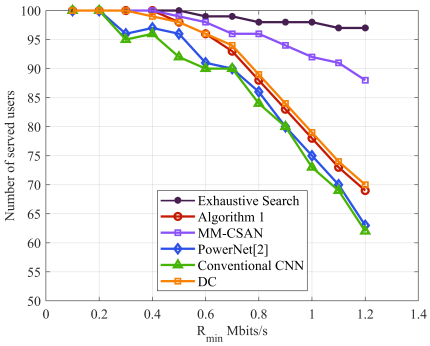

Fig. 5. Number of served user versus R_{min}.

Fig. 5 shows the number of served users for different values of R_{min}. The number of served users decreases when R_{min} increases. It can be seen that our MM-CSAN model shows stability in predicting power that satisfies the level of R_{min} especially for smaller values of R_{min} and clearly outperforms Algorithm 1, DC, PowerNet and conventional CNN. For larger R_{min}, despite the deterioration in the performance, MM-CSAN still achieves notably better performance compared to that of other methods. It is obvious from these observations that MM-CSAN is more suitable for larger problem compared to the other methods in this study.

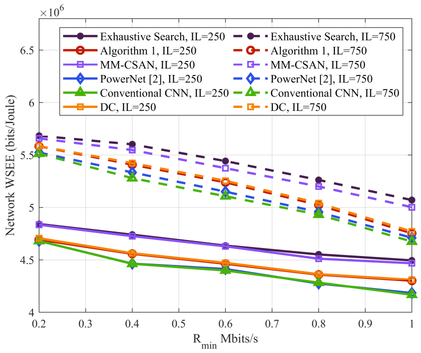

Fig. 6. Network WSEE versus R_{min} for different permittable interference levels.

In Fig. 6, we assess the capabilities of different methods in dealing with impact of both R_{min} and the interference levels. We check the performance for different values of R_{min} and the interference levels. Consequently, we can observe that MM-CSAN can achieve performance near to the optimal due to the presence of the attention mechanism which allows the model to focus on capturing the patterns in data. Moreover, the semantic and spatial encoding of the features and the fusion of different features adds discriminative trait to the features. Hence, the prediction can be enhanced. We considered 10 per BS to simulate Fig. 6. For lower R_{min} and $IL = 250$, the performance of MM-CSAN is comparable to that of the exhaustive search and better than that of Algorithm 1, DC, PowerNet and the conventional CNN. For higher R_{min} (e.g. 0.8 Mbps) and $IL = 750$, the performance gap between MM-CSAN and the optimal solution is obvious. However, MM-CSAN achieves better performance than that of Algorithm 1, DC, PowerNet and the conventional CNN and the performance gap between these methods and MM-CSAN is growing with the increasing of R_{min} and IL.

VI. Conclusion

In this work, we proposed an end-to-end multi-modal convolutional self-attention network to perform power control in NOMA network. The model consists of main pipeline with self-attention block for high quality feature extraction and another pipeline for shallow modality feature extraction. Those features are combined to make more discriminative features and enhancing the power prediction. The simulation results showed that the proposed model is suitable for real-time applications and outperformed other models such as PowerNet and the conventional CNN.

References

[1] Y. Zhang, H. Wang, T. Zheng, and Q. Yang, "Energy-Efficient Transmission Design in Non-orthogonal Multiple Access," *IEEE Trans. Veh. Technol.*, vol.66, no. 3, pp. 2852-2857, 2017.

[2] Zhang, T. and S. Mao, "Energy-Efficient Power Control in Wireless Networks With Spatial Deep Neural Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 111-124, 2020.

[3] G. Liu, R. Wang, H. Zhang, W. Kang, T. A. Tsiftsis, and V. C. M. Leung, "Super-Modular Game-Based User Scheduling and Power Allocation for Energy-Efficient NOMA Network," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, p. 3877-3888.

[4] A. B. M. Adam, X. Wan, and Z. Wang, "Energy Efficiency Maximization in Downlink Multi-Cell Multi-Carrier NOMA Networks With Hardware Impairments," *IEEE Access*, vol. 8, pp. 210054-210065, 2020.

[5] Z. Alessio, J. Eduard, "Energy Efficiency in Wireless Networks via Fractional Programming Theory", 2015.

[6] Z. Ali, G. A. S. Sidhu, F. Gao, J. Jiang, and X. Wang, "Deep Learning Based Power Optimizing for NOMA Based Relay Aided D2D Transmissions," *IEEE Trans. Cogn. Commun. Netw.*, pp. 1-1, 2021.

[7] B. Matthiesen, A. Zappone, K. L. Besser, E. A. Jorswieck, and M. Debbah, "A Globally Optimal Energy-Efficient Power Control Framework and Its Efficient Implementation in Wireless Interference Networks," *IEEE Trans. Signal Process.*, vol. 68, pp. 3887-3902, 2020.

[8] L. Lei, L. You, Q. He, T. X. Vu, S. Chatzinotas, D. Yuan, and B. Ottersten, "Learning-Assisted Optimization for Efficient Scheduling in Deadline-Aware NOMA Systems", *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 3, pp. 615-627, Sept. 2019.

[9] A. Ly and Y. D. Yao, "A Review of Deep Learning in 5G Research: Channel Coding, Massive MIMO, Multiple Access, Resource Allocation, and Network Security," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 396-408, 2021.

[10] M. A. M. Elhassan, C. Huang, C. Yang, and T. L. Munea "DSANet: Dilated Spatial Attention for Real-time Semantic Segmentation in Urban SStreet Scenes," *Expert Sys. Appl.*, vol. 183, 115090, April 2021.

[11] Z. Al Nazi, F. Rabbi Mashrur, M. Amirul Islam, and S. Saha, "FibroCoSANet: Pulmonary Fibrosis Prognosis Prediction using a Convolutional Self Attention Network," https://arxiv.org/abs/2104.05889, Accessed on: April 01, 2021.

[12] R.W. Freund, and F. Jarre, "Solving the Sum-of-Ratios Problem by an Interior-Point Method," *J. Global Optim.* vol. 19, pp. 83102, 2001.

[13] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-Efficient Power Control: A Look at 5G Wireless Technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668-1683, Apr. 2016,

[14] K. Shen and W. Yu, "Fractional Programming for Communication SystemsPart I: Power Control and Beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616-2630, 2018.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *in Proc. 2016 IEEE Conf. Comp. Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, pp. 770-778, 2016.

[16] Z. Wei, D. W. K. Ng, J. Yuan and H. Wang, "Optimal Resource Allocation for Power-Efficient MC-NOMA With Imperfect Channel State Information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944-3961, Sept. 2017.