

Leveraging MoE-based Large Language Model for Zero-Shot Multi-Task Semantic Communication

Sin-Yu Huang, Renjie Liao, and Vincent W.S. Wong

Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada

Email: {syhuang, rjliao, vincentw}@ece.ubc.ca

Abstract—Multi-task semantic communication (SC) can reduce the computational resources in wireless systems since retraining is not required when switching between tasks. However, existing approaches typically rely on task-specific embeddings to identify the intended task, necessitating retraining the entire model when given a new task. Consequently, this drives the need for a multi-task SC system that can handle new tasks without additional training, known as zero-shot learning. Inspired by the superior zero-shot capabilities of large language models (LLMs), we leverage pre-trained instruction-tuned LLMs, referred to as fine-tuned language net (FLAN), to improve the generalization capability. We incorporate a mixture-of-experts (MoE) architecture in the FLAN model and propose MoE-FLAN-SC architecture for multi-task SC systems. Our proposed MoE-FLAN-SC architecture can further improve the performance of FLAN-T5 model without increasing the computational cost. Moreover, we design a multi-task feature extraction module (FEM) which can adaptively extract relevant features across various tasks given the provided features and signal-to-noise ratio (SNR). Simulation results show that our proposed MoE-FLAN-SC architecture outperforms three state-of-the-art models in terms of the average accuracy on four different unseen tasks.

I. INTRODUCTION

Task-oriented semantic communication (SC) is an emerging technology for the sixth-generation (6G) wireless networks [1]. In task-oriented SC, the transmitter only sends the task-related features to the receiver. In [2], Shao *et al.* proposed a learning-based task-oriented communication scheme that can efficiently extract features from images for image classification tasks. In [3], Hu *et al.* studied the impact of semantic noise on task-oriented SC systems and introduced a feature importance module that can dynamically adjust the weights for different features. Nevertheless, the aforementioned works only study a single-task scenario. To perform a new task, the entire model needs to be retrained.

In [4], U-DeepSC is proposed to handle multiple tasks across three modalities: text, speech, and image. It can achieve performance comparable to that of single-task SC systems trained individually on each task. However, for each task, U-DeepSC relies on a task-specific embedding vector shared by the transmitter and receiver in order to identify the tasks to be performed. However, it limits the model's generalization ability to tasks that have not been explicitly trained on, referred to as unseen tasks [5]. On the other hand, some recent works have proposed approaches to transfer knowledge from an existing task to an unseen task. In [6], Leung *et al.* proposed a generalizable multi-task communication paradigm, where only

the decoder needs to be retrained for unseen tasks. However, retraining the decoder, which is nearly half of the model size, is both time- and resource-intensive.

Performing unseen tasks without either additional training or specific examples for those tasks is called zero-shot learning. In this context, large language models (LLMs) have shown to possess strong zero-shot ability. In [7], Raffel *et al.* introduced a unified text-to-text transfer Transformer (T5) model, showcasing LLMs' strong transfer learning capabilities across tasks such as summarization, question answering, and text classification. Building on the T5 model, Chung *et al.* in [8] proposed the fine-tuned language net (FLAN)-T5 model, demonstrating that instruction fine-tuning with a number of tasks can significantly enhance the model's zero-shot performance. Despite the impressive results achieved by LLMs, their performance is at the expense of larger model sizes.

In multi-task SC systems, another challenge arises due to gradient conflicts between tasks [9]. During the training process, the gradients of different tasks can point to different directions. Therefore, updating the model parameters based on gradients of one specific task can degrade the performance of other tasks. Recall that task-oriented SC transmits only the task-relevant features to complete the intended task. In [4], U-DeepSC uses a separate feature selection module (FSM) to select task-oriented features for each task. This approach effectively avoids gradient conflicts since each module independently optimizes the feature selection process specific to its task. However, when the number of tasks increases, the number of FSMs increases linearly, resulting in a larger model size. Furthermore, when encountering an unseen task, a new FSM needs to be trained. For a multi-task SC system, we face the following questions:

- Q1: *How to design a multi-task SC system that can effectively tackle those unseen tasks without retraining?*
- Q2: *How to satisfy the diverse task requirements and improve the performance in a multi-task SC system without increasing the computational cost?*

To address the above questions, in this paper, we leverage the mixture-of-experts (MoE)-based LLM [10] in a multi-task SC system. The model aims to perform zero-shot learning for unseen tasks while filters out the task-irrelevant features for each task. The contributions of this paper are as follows:

- We propose an MoE-FLAN-SC architecture that facilitates zero-shot learning for multi-task SC system. The proposed architecture integrates MoE in a pretrained FLAN-T5

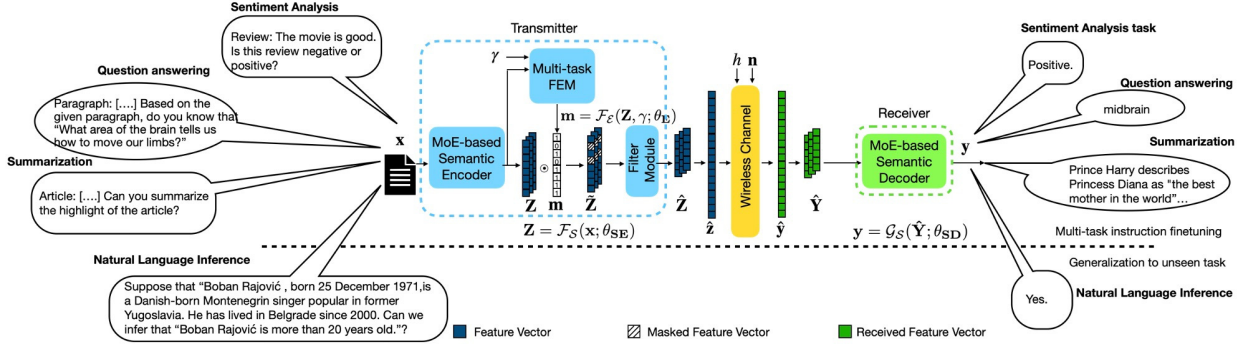


Fig. 1: The system model of the proposed MoE-FLAN-SC architecture, which is trained on multiple natural language processing tasks, including sentiment analysis, question answering, and summarization, using natural language prompts. By training on a diverse range of tasks, the model enhances its ability to generalize to previously unseen tasks.

model and uses a single multi-task feature extraction module (FEM) to adaptively filter out irrelevant features for all tasks, thereby reducing the transmission overhead. We propose a two-phase algorithm to instruction-tune the model using public pool of prompts (P3) [11], which is a collection of datasets with prompt templates.

- The proposed MoE-FLAN-SC architecture can perform zero-shot learning on various unseen tasks. By integrating MoE into our model, it improves zero-shot performance with the same floating point operations per second (FLOPs) per token. The multi-task FEM also uses the concept of MoE that includes multiple extractor modules to filter out irrelevant features under different signal-to-noise ratio (SNR).
- We conduct simulations for various unseen tasks and compare our proposed MoE-FLAN-SC architecture with three state-of-the-art models, which are the FLAN-T5 [8], LM-adapted T5 model [7], and a conventional approach that uses 8-bit unicode transformation format (UTF-8) encoding and Turbo coding. Simulation results show that our proposed MoE-FLAN-SC architecture achieves a higher average accuracy than the baseline models. It achieves a 3.83% increase in average accuracy under high SNR conditions without increasing the FLOPs per token. We also conduct ablation study under different numbers of experts and extractor modules in FEM.

This paper is organized as follows. Section II introduces the system model and presents the tasks and prompts considered in our model. Section III presents the architecture of the proposed model for each component, including the encoder, FEM, and decoder. We also present the two-phase training algorithm. Results for performance evaluation and comparison are presented in Section IV. Conclusion is given in Section V.

Notations: We use upper-case boldface letters to denote matrices and lower-case boldface letters to denote column vectors. \mathbb{R} and \mathbb{C} denote the sets of real and complex numbers, respectively. \mathbf{A}^T denotes the transpose of matrix \mathbf{A} . $\mathbf{A}[\mathcal{I}, :]$ returns the submatrix of matrix \mathbf{A} with the rows chosen from

an index set \mathcal{I} . $\mathbb{E}[\cdot]$ represents statistical expectation, and \mathbf{I} denotes the identity matrix.

II. SYSTEM MODEL AND TASK DESCRIPTION

As illustrated in Fig. 1, the proposed MoE-FLAN-SC architecture is trained on a mixture of text-specific datasets using natural language prompts. The model consists of a transmitter and a receiver. The transmitter includes an MoE-based semantic encoder, a multi-task FEM, and a filter module. The MoE-based semantic encoder generates feature representations from the input data. The multi-task FEM produces a mask to filter out the task-irrelevant features. The filter module then removes those masked feature vectors. At the receiver, an MoE-based semantic decoder is used to process the received message for the intended task.

The MoE-based semantic encoder in the transmitter maps a variable-length sentence, denoted as \mathbf{x} , into a feature matrix, represented as $\mathbf{Z} \in \mathbb{R}^{N \times D}$, where N is the number of feature vectors and D is the dimension of a feature vector. We have

$$\mathbf{Z} = \mathcal{F}_S(\mathbf{x}; \theta_{SE}), \quad (1)$$

where $\mathcal{F}_S(\cdot; \theta_{SE})$ represents the MoE-based semantic encoder with parameters set θ_{SE} . To mask the irrelevant features and reduce the transmission overhead, a multi-task FEM is used to compress the number of features. Given the feature matrix \mathbf{Z} and SNR, the multi-task FEM generates a mask $\mathbf{m} \in \{0, 1\}^N$, which is given by

$$\mathbf{m} = \mathcal{F}_E(\mathbf{Z}, \gamma; \theta_E), \quad (2)$$

where $\mathcal{F}_E(\cdot; \theta_E)$ represents the multi-task FEM with parameters set θ_E and γ is the SNR obtained via channel output feedback [12]. Let m_i denote the i th element of the mask \mathbf{m} . We have $m_i = 1$ if the i th feature vector is retained and $m_i = 0$ otherwise. We denote the number of retained features as $N_c = \mathbf{1}_N^T \mathbf{m} \leq N$, where $\mathbf{1}_N$ is an all-ones column vector with N elements. Here, we define the compression ratio as $\rho \triangleq N_c/N$. Each feature vector in the masked feature matrix $\tilde{\mathbf{Z}} \in \mathbb{R}^{N \times D}$ is defined as the multiplication of the

corresponding element in the mask and the feature vector, i.e., $\tilde{\mathbf{Z}}[i, :] = m_i \mathbf{Z}[i, :]$. The masked feature matrix $\tilde{\mathbf{Z}}$ is sent to the filter module to remove the masked features. Define $\mathcal{R} = \{i \mid m_i = 1, i = 1, \dots, N\}$ as the set of indices for the retained features. The compressed feature matrix $\hat{\mathbf{Z}}$ is given by

$$\hat{\mathbf{Z}} = \tilde{\mathbf{Z}}[\mathcal{R}, :] \in \mathbb{R}^{N_c \times D}. \quad (3)$$

Thus, $\hat{\mathbf{Z}}$ is a submatrix formed by selecting the retained feature vectors from matrix $\tilde{\mathbf{Z}}$.

The compressed feature matrix $\hat{\mathbf{Z}}$ is then reshaped into a stream of encoded symbols $\hat{\mathbf{z}}$, and are transmitted through a wireless channel. The received signal $\hat{\mathbf{y}}$ is given by

$$\hat{\mathbf{y}} = h\hat{\mathbf{z}} + \mathbf{n}, \quad (4)$$

where $h \sim \mathcal{CN}(0, 1)$ represents the Rayleigh fading channel gain and $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$ denotes the additive white Gaussian noise (AWGN) with zero mean and covariance matrix $\sigma^2 \mathbf{I}$. Upon receiving $\hat{\mathbf{y}}$, the stream is reshaped back to a matrix $\hat{\mathbf{Y}} \in \mathbb{R}^{N_c \times D}$. The MoE-based semantic decoder is subsequently used to decode $\hat{\mathbf{Y}}$ into sentence \mathbf{y} , representing the response for the intended task. We have

$$\mathbf{y} = \mathcal{G}_S(\hat{\mathbf{Y}}; \theta_{SD}), \quad (5)$$

where $\mathcal{G}_S(\cdot; \theta_{SD})$ is the MoE-based semantic decoder with parameters set θ_{SD} .

To demonstrate the model's zero-shot ability to unseen tasks, we utilize the datasets from P3 and task taxonomy outlined in [11] to train our model. This taxonomy comprises 12 tasks, including multiple-choice question answering (QA), sentiment analysis, and summarization, spanning a total of 62 datasets. We divide the tasks into two groups. The first group, denoted as \mathcal{T}_t , contains 8 tasks and is designated for training the multi-task model. The second group, denoted as \mathcal{T}_h , contains the remaining 4 tasks and forms a held-out set. It is used to evaluate the model's generalization ability to unseen tasks. All data inputs in our model are formatted as natural language prompts to facilitate generalization. We adhere to the prompt template outlined in P3 to train our model.

III. PROPOSED MOE-FLAN-SC ARCHITECTURE

A. Architecture of MoE-based Semantic Encoder

As illustrated in Fig. 2, the input sentence \mathbf{x} is initially converted into word embeddings through tokenization. Subsequently, positional embeddings are incorporated into the word embeddings to provide position information for each token. Let $\mathbf{X}^i \in \mathbb{R}^{N \times D}$ denote the feature generated by the i th encoder layer. The token embeddings are then fed through E layers of the MoE Transformer encoder layer. The output of the final MoE Transformer encoder layer, \mathbf{X}^E , is also the output of the MoE-based semantic encoder, denoted as \mathbf{Z} in Section II.

The architecture of MoE Transformer encoder layer is depicted in the lower part of Fig. 2. In the MoE Transformer encoder layer, the features generated from the previous layer \mathbf{X}^{i-1} are first passed through the multi-head self-attention module [13], which computes the attention scores within the

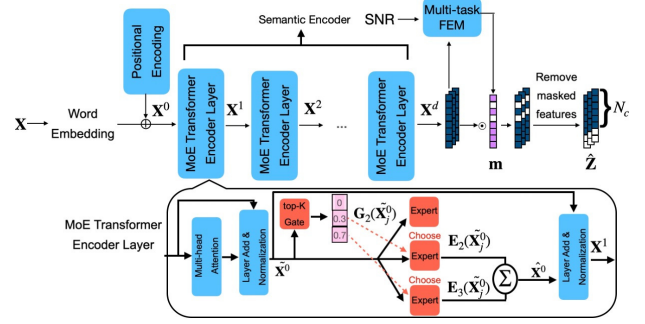


Fig. 2: The architecture of the MoE-based semantic encoder, where the lower part is the architecture of first MoE Transformer encoder layer. In this figure, the top-2 gate selects $K = 2$ experts out of the total $M = 3$ experts.

i th MoE Transformer encoder layer, transforming the features into $\tilde{\mathbf{X}}^{i-1} \in \mathbb{R}^{N \times D}$. To enable the model to handle complex characteristics of multiple tasks, we incorporate multiple feed-forward layers [13], referred to as experts, in our model. Let $\mathbf{E}_i(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^D$ represent the transformation function of the i th expert, for $i = 1, \dots, M$. Instead of utilizing all the results from the experts, we sparsely activate K experts from the total of M experts, where $K < M$. The selection of experts and their corresponding weights are determined by a top- K gate, which selects experts based on the feature matrix $\tilde{\mathbf{X}}^{i-1}$. For each row vector of $\tilde{\mathbf{X}}^{i-1}$, denoted as $\tilde{\mathbf{X}}^{i-1}[j, :] \in \mathbb{R}^{1 \times D}$, $j = 1, \dots, N$, the gating function is given by:

$$\mathbf{G}_K(\tilde{\mathbf{X}}^{i-1}[j, :]) = \text{Softmax}(\text{TopK}(\tilde{\mathbf{X}}^{i-1}[j, :]; W_g)) \in \mathbb{R}^{1 \times M}, \quad (6)$$

where $W_g \in \mathbb{R}^{D \times M}$ is the linear weight matrix of the gate, and $\text{TopK}(\cdot)$ retains the largest K values, setting the other elements to negative infinity. Let $\mathbf{E}(\tilde{\mathbf{X}}^{i-1}[j, :]) = [\mathbf{E}_1(\tilde{\mathbf{X}}^{i-1}[j, :]), \dots, \mathbf{E}_M(\tilde{\mathbf{X}}^{i-1}[j, :])] \in \mathbb{R}^{M \times D}$ denote the concatenated results of the expert transformations. For $j = 1, \dots, N$, the weighted sum for the selected experts is computed as:

$$\hat{\mathbf{X}}^{i-1}[j, :] = \mathbf{G}_K(\tilde{\mathbf{X}}^{i-1}[j, :]) \mathbf{E}(\tilde{\mathbf{X}}^{i-1}[j, :]) \in \mathbb{R}^{1 \times D}. \quad (7)$$

The result of $\hat{\mathbf{X}}^{i-1}$ is then linearly added with $\tilde{\mathbf{X}}^{i-1}$ and normalized to produce the output of the i th MoE Transformer encoder layer \mathbf{X}^i .

B. Architecture of Multi-Task Feature Extractor Module

In task-oriented communication systems, not all features are relevant to the intended task. Therefore, a multi-task FEM is used to filter out irrelevant information, retaining only the essential features necessary for the task. As illustrated in Fig. 3, the multi-task FEM takes into account the SNR and the features generated by the semantic encoder \mathbf{Z} to generate a binary mask, $\mathbf{m} \in \{0, 1\}^N$, for each feature. Both inputs are processed through a multilayer perceptron (MLP) to generate representations of the SNR and the features, denoted as \mathbf{F}_γ and \mathbf{F}_z , respectively. To jointly consider both inputs, we concatenate these feature representations as $\tilde{\mathbf{F}} = \text{cat}(\mathbf{F}_\gamma, \mathbf{F}_z)$. Since each task has distinct characteristics, the multi-task FEM should capture different feature characteristics and differentiate

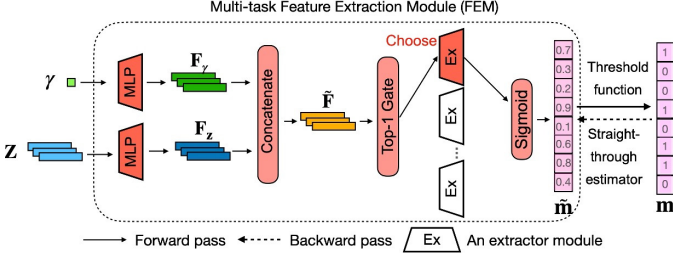


Fig. 3: The architecture of the multi-task feature extraction module.

relevant features. To enhance feature identification across a wide range of tasks, we employ a total of B extractor modules within the multi-task FEM. The matrix $\tilde{\mathbf{F}}$ is then passed through one of these B extractor modules followed by a sigmoid activation function. The selection of the extractor module is determined by a Top-1 gate $\mathbf{G}_1(\cdot)$. Let $\mathbf{M}_k(\cdot)$ denote the linear transformation of the k th extractor module. The mask $\tilde{\mathbf{m}} = [\tilde{m}_1, \dots, \tilde{m}_N] \in [0, 1]^N$ is generated as

$$\tilde{\mathbf{m}} = \text{Sigmoid}(\mathbf{M}_k(\tilde{\mathbf{F}})), \text{ where } k = \arg \max \mathbf{G}_1(\tilde{\mathbf{F}}). \quad (8)$$

This mask indicates the importance of each feature, with values ranging between 0 and 1. To determine whether a feature should be retained or discarded, a constant threshold \bar{t} is applied. The final mask $\mathbf{m} = [m_1, \dots, m_N]$ is determined as follows:

$$m_i = \begin{cases} 1, & \tilde{m}_i \geq \bar{t}, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

In the forward pass, the threshold function (9) is applied. However, in the backward pass, the threshold function (9) is non-differentiable, which brings a challenge for backpropagation. To address this issue, we use the straight-through estimator [14] during the backward pass to approximate the gradient.

C. Architecture of MoE-based Decoder

The architecture of MoE-based decoder is shown in Fig. 4. It consists of a stack of MoE Transformer decoder layers, followed by a dense layer and a softmax layer, jointly referred to as the head. Specifically, the MoE Transformer decoder layer maintains the same architecture as the vanilla decoder layer in the Transformer model. The only difference is that the linear layer within the feed-forward network is replaced by MoE and a gating mechanism, as described in Section III-A. To generate the j th word in output \mathbf{y} , denoted as y_j , the MoE Transformer decoder layers take the received feature matrix $\hat{\mathbf{Y}} \in \mathbb{R}^{N_c \times D}$ and the previous generated token \tilde{y}_{j-1} as input, where the initial token \tilde{y}_0 is designated as the special $\langle \text{pad} \rangle$ token. These inputs are processed to generate $\tilde{\mathbf{y}}_j^i$ from the i th MoE Transformer decoder layer for the j th word. After passing through a total of E MoE Transformer decoder layers, the head layer maps the final output $\tilde{\mathbf{y}}_j^E$ to a probability vector, denoted as $\mathbf{l}_j \in [0, 1]^V$, where V represents the number of tokens in the token vocabulary. Each element in vector \mathbf{l}_j represents the probability of taking a specific token in the token vocabulary. The token \tilde{y}_j is determined by selecting the one with the highest likelihood, given as $\tilde{y}_j = \arg \max_{k=1, \dots, V} l_{j,k}$, where $l_{j,k}$ is

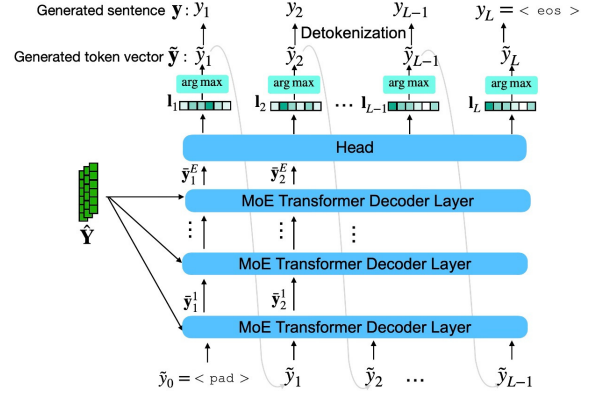


Fig. 4: The architecture of the MoE-based semantic decoder.

the k th element of vector \mathbf{l}_j . The decoder auto-regressively generates the token \tilde{y}_j until it produces the index of a special $\langle \text{eos} \rangle$ token, thereby constructing the generated token vector $\tilde{\mathbf{y}} = [\tilde{y}_1, \dots, \tilde{y}_L] \in \{1, \dots, V\}^L$ with L tokens. By detokenizing $\tilde{\mathbf{y}}$, we can obtain the generated sentence \mathbf{y} with $L - 1$ words since the $\langle \text{eos} \rangle$ token indicates the end of sentence and does not correspond to a word.

D. Proposed MoE-FLAN-SC Training Algorithm

MoE-FLAN-SC is designed to achieve two objectives: learning general representations across multiple tasks, and ensuring that the FEM filters out task-irrelevant features. To jointly optimize the model with two different objectives, we propose a two-phase training algorithm to train the proposed MoE-FLAN-SC. The proposed training algorithm is shown in Algorithm 1. In the first phase, we train the MoE-based semantic encoder and decoder without applying any compression, aiming to establish a model capable of capturing the essential features. To optimize the model, we use the cross-entropy loss function. Let $\mathbf{q} = [q_1, \dots, q_L] \in \{1, \dots, V\}^L$ denote the tokens of the target sentence. For the i th token in \mathbf{q} , denoted as q_i , the one-hot vector representation is represented as $\mathbf{o}_i \in \{0, 1\}^V$, where only the q_i th element is equal to 1 and all other elements are set to zero. The cross-entropy loss is as follows

$$\mathcal{L}_{\text{CE}} = -\frac{1}{L} \sum_{i=1}^L \sum_{j=1}^V o_{i,j} \log(l_{i,j}), \quad (10)$$

where $o_{i,j}$ is the j th element of vector \mathbf{o}_i . In the second phase, we load the trained parameters from the first phase and incorporate the multi-task FEM. To minimize the number of transmit symbols, we introduce the compression ratio as an additional loss function term. It is given by

$$\mathcal{L}_\rho = \frac{N_c}{N}. \quad (11)$$

The model is then trained to jointly minimize the cross-entropy loss \mathcal{L}_{CE} and the compression loss \mathcal{L}_ρ . This joint optimization facilitates the model in achieving high performance in generating the target tokens while filtering out irrelevant features, thereby reducing the number of transmit symbols.

Algorithm 1 Proposed MoE-FLAN-SC Training Algorithm

```

1: Input: Training datasets for each task  $t$  in  $\mathcal{T}_t$ , including data and labels; Number of epochs for the first and second phases,  $N_1$  and  $N_2$ ; learning rate  $\eta$ ; regularization weights  $\omega_1$  and  $\omega_2$ .
2: First Phase:
3: Fixed the parameters in FEM.
4: for  $i \leftarrow 1$  to  $N_1$  do
5:   for  $t \in \mathcal{T}_t$  do
6:     Load input data  $\mathbf{x}$  and labels for task  $t$ .
7:     Sample the channel gain coefficient  $h$  and SNR value.
8:     Perform forward propagation.
9:     Calculate the loss  $\mathcal{L}_1(\mathbf{x}) := \mathcal{L}_{CE}$  based on (10).
10:    Update parameters  $\{\theta_{SE}, \theta_{SD}\}$  using  $\mathcal{L}_1(\mathbf{x})$ .
11:   end for
12: end for
13: Second Phase:
14: Load parameter sets trained in the first phase to the model.
15: for  $i \leftarrow 1$  to  $N_2$  do
16:   for  $t \in \mathcal{T}_t$  do
17:     Load input data  $\mathbf{x}$  and labels for task  $t$ .
18:     Sample the channel gain coefficient  $h$  and SNR value.
19:     Perform forward propagation.
20:     Calculate the loss  $\mathcal{L}_2(\mathbf{x}) := \omega_1 \mathcal{L}_{CE} + \omega_2 \mathcal{L}_\rho$  based on (10) and (11).
21:     Update parameters  $\{\theta_{SE}, \theta_E, \theta_{SD}\}$  using  $\mathcal{L}_2(\mathbf{x})$ .
22:   end for
23: end for
24: Output: The optimized parameters  $\{\theta_{SE}^*, \theta_E^*, \theta_{SD}^*\}$ .

```

IV. PERFORMANCE EVALUATION

The MoE-based semantic encoder and decoder of the MoE-FLAN-SC are initialized with the FLAN-T5-Base model¹ [8] such that the number of Transformer layers E is set to 12, and the dimension for each feature vector D is set to 768. Unless stated otherwise, the number of experts M is set to 10, the number of selected experts K is set to 1, and the number of extractor modules B is set to 10. All the experts are initialized with the same parameters as in the original feed-forward layer in FLAN-T5-Base model. The models are trained on P3 [11], a collection of datasets with prompt templates. For the settings of the training procedure, we use AdamW optimizer [15] with learning rate $\eta = 10^{-5}$ and batch size equal to 120. The token selection threshold $\bar{t} = 0.5$ and the regularization weights $(\omega_1, \omega_2) = (10^3, 10)$. In our simulations, we use the exact match to evaluate the accuracy by measuring the proportion of generated sentences that exactly match the targets. We consider the following baseline models for comparison:

- 1) **FLAN-SC model** [8]: This model is initialized with FLAN-T5-Base model. It only includes one expert in the Transformer layer and does not include the gating network.
- 2) **LM-adapted SC model**² [7]: This model is initialized with LM-adapted T5 model and is trained without instruction-tuning. It includes only a single expert in each Transformer layer and does not have a gating network. We initialize the model with the “XL” size of the LM-adapted

¹The public checkpoints for FLAN-T5 models are available at: <https://github.com/google-research/t5x/blob/main/docs/models.md#flan-t5-checkpoints>

²The public checkpoints for LM-adapted-T5 models are available at: https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md#lm-adapted-t511lm100k

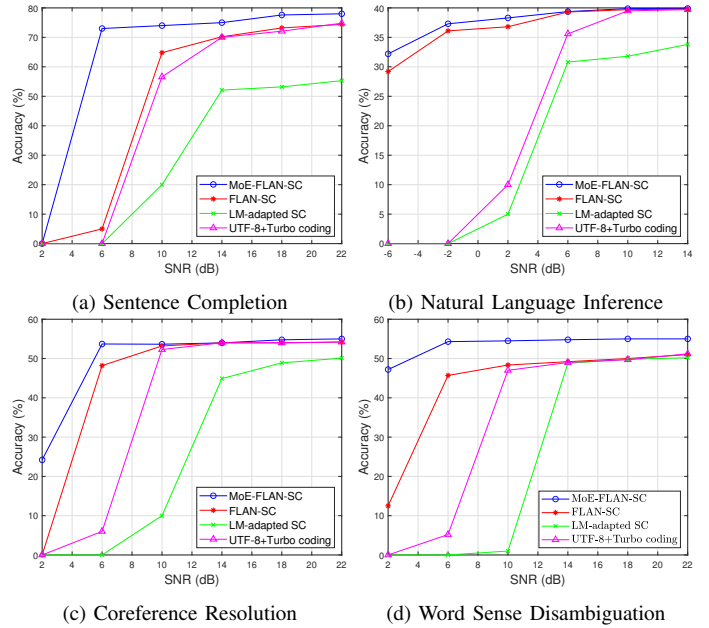


Fig. 5: The accuracy of the held-out tasks in \mathcal{T}_h versus SNR. (a) Sentence completion task, (b) natural language inference task, (c) coreference resolution task, (d) word sense disambiguation task.

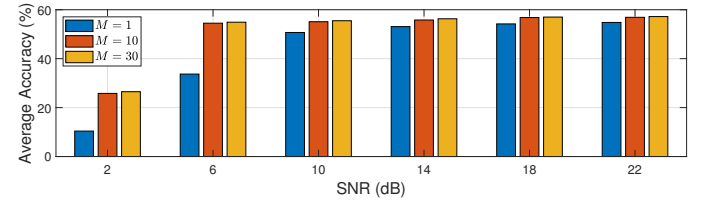


Fig. 6: The average accuracy of all four held-out tasks versus SNR under different number of experts M .

T5 model, which consists of $E = 24$ Transformer layers. The dimension D is equal to 2048.

- 3) **UTF-8+Turbo coding**: The UTF-8 encoding and Turbo coding are employed for source coding and channel coding, respectively. The code rate for channel coding is set to $\frac{1}{2}$. The FLAN-T5-Base model is utilized at the receiver to perform the intended tasks based on the decoded message.

In Fig. 5, we show the performance of different baseline models versus SNR for four held-out tasks. Across all tasks, our proposed MoE-FLAN-SC consistently outperforms FLAN-T5, while FLAN-T5 outperforms the other baseline models. This demonstrates that instruction-tuning enhances the performance of the model across different tasks, and integrating the MoE architecture into FLAN-T5 leads to significant performance gains. Among those four unseen tasks, MoE-FLAN-SC achieves an average accuracy of 56.9%, compared to 54.8% for FLAN-SC when SNR = 22 dB, representing a 3.83% accuracy improvement in high SNR region. The results highlight the superiority of MoE-FLAN-SC over FLAN-T5. In Fig. 6, we show that when the number of experts M increases in our proposed MoE-FLAN-SC architecture, the performance will eventually saturate.

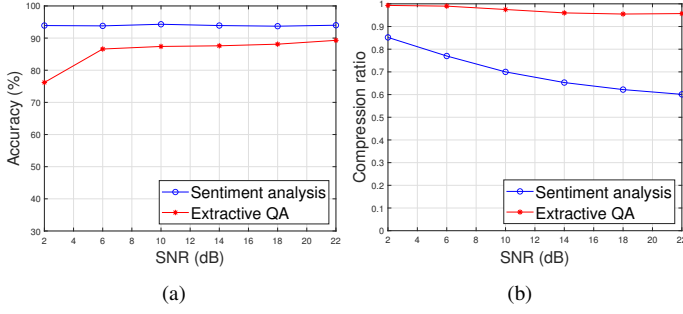


Fig. 7: (a) The accuracy and (b) compression ratio versus SNR for sentiment analysis and extractive QA tasks.

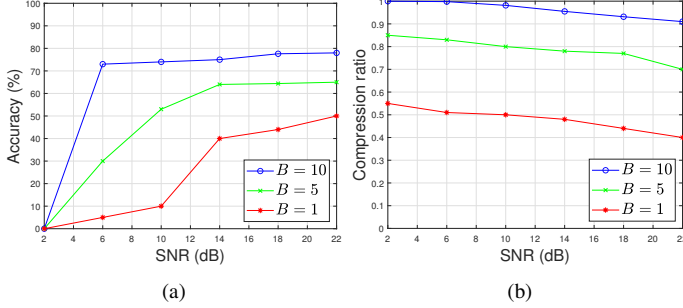


Fig. 8: (a) The accuracy and (b) compression ratio versus SNR for sentence completion task under number of extractor modules B .

To evaluate the effectiveness of multi-task FEM in our proposed architecture, Fig. 7 presents the accuracy and compression ratio for sentiment analysis and extractive QA tasks. When the SNR decreases, the compression ratio increases, indicating that lower SNR requires more redundancy to maintain the performance. At higher SNR, less redundancy is needed, allowing higher compression without loss of accuracy. The results also show different compression ratios required for different tasks. QA requires the entire paragraph to find the answer. It requires high compression ratios even at high SNR region. On the other hand, sentiment analysis is a binary task. It allows higher compression when the SNR increases. These results show that the multi-task FEM can effectively adapt the transmit symbols to different tasks and SNR conditions. In Fig. 8, we present the results of using different number of extractor modules B in the FEM. Results show that utilizing multiple extractor modules in the FEM can improve the accuracy of unseen tasks.

Finally, in Table I, we present the results of the total number of parameters and FLOPs per token for different models. The use of MoE increases the total number of parameters count on the order of $O(ME)$, where M is the number of experts per layer and E is the number of layers, while the FLOPs per token remain the same since we only select one expert to process the input feature vector. This demonstrates that the MoE architecture enables scaling the model without increasing the computational cost.

V. CONCLUSION

In this paper, we proposed an MoE-FLAN-SC architecture for multi-task SC systems. The proposed architecture integrated

TABLE I: TOTAL NUMBER OF PARAMETERS AND FLOPS PER TOKEN OF DIFFERENT MODELS

Model	Total Parameters	FLOPs per token
MoE-FLAN-SC	1.27B	300M
FLAN-SC	250M	300M
LM-adapted SC	3B	3.6G

an MoE into an instruction-tuned LLM and is capable to perform zero-shot learning on unseen tasks. Simulation results showed that our proposed architecture provides good performance across four unseen tasks. Results also demonstrated that our proposed MoE-FLAN-SC architecture outperforms FLAN-SC, LM-adapted SC, and UTF-8+Turbo coding, achieving an average accuracy improvement of 3.83% on unseen tasks compared to FLAN-SC. Furthermore, the proposed multi-task FEM can effectively optimize the number of transmit symbols based on task requirements and channel conditions. For future work, we plan to extend the proposed architecture to improve the zero-shot performance in multimodal communication scenario.

REFERENCES

- [1] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.
- [2] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.
- [3] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications with masked VQ-VAE enabled codebook," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8707–8722, Dec. 2023.
- [4] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4101–4116, Jul. 2024.
- [5] X. Wang, H. Xu, L. Gui, and Y. He, "Towards unified task embeddings across multiple models: Bridging the gap for prompt-based large language models and beyond," in *Proc. of Assoc. Comput. Linguist. (ACL)*, Bangkok, Thailand, Aug. 2024.
- [6] C. T. Leung, R. Ghosh, and M. Motani, "Multi-task generalizable communication: Beyond the information bottleneck," in *Proc. of IEEE Int. Conf. Commun. (ICC)*, Denver, CO, Jun. 2024.
- [7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res. (JMLR)*, vol. 21, pp. 1–67, Jan. 2020.
- [8] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *J. Mach. Learn. Res. (JMLR)*, vol. 25, pp. 1–53, Oct. 2024.
- [9] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. of Advances in Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, Dec. 2020.
- [10] S. Shen *et al.*, "Mixture-of-experts meets instruction tuning: A winning combination for large language models," in *Proc. of Int. Conf. Learn. Representations (ICLR)*, Vienna, Austria, May 2024.
- [11] V. Sanh *et al.*, "Multitask prompted training enables zero-shot task generalization," in *Proc. of Int. Conf. Learn. Representations (ICLR)*, May 2022.
- [12] D. B. Kurka and D. Gündüz, "DeepJSCC-f: Deep joint source-channel coding of images with feedback," *IEEE J. Sel. Areas Inform. Theory*, vol. 1, no. 1, pp. 178–193, May 2020.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Advances in Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, Dec. 2017.
- [14] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.
- [15] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, May 2019.