Task-Oriented Multi-User Semantic Communications for VQA Task

Huiqiang Xie, *Student Member, IEEE*, Zhijin Qin, *Senior Member, IEEE,* and Geoffrey Ye Li, *Fellow, IEEE,*

***Abstract*—Semantic communications focus on the transmission of semantic features. In this letter, we consider a task-oriented multi-user semantic communication system for multimodal data transmission. Particularly, partial users transmit images while the others transmit texts to inquiry the information about the images. To exploit the correlation among the multimodal data from multiple users, we propose a deep neural network enabled semantic communication system, named MU-DeepSC, to execute the visual question answering (VQA) task as an example. Specifically, the transceiver for MU-DeepSC is designed and optimized jointly to capture the features from the correlated multimodal data for task-oriented transmission. Simulation results demonstrate that the proposed MU-DeepSC is more robust to channel variations than the traditional communication systems, especially in the low signal-to-noise (SNR) regime.**

***Index Terms*—Deep learning, multimodal data, multi-user, semantic communication.**

I. Introduction

The continuously increasing number of connected-mobile devices and enriched intelligent demands cause the explosion of wireless data traffic, which brings new challenges to communication systems, including providing the cornerstone for various intelligent tasks, exploiting the limited frequency resource, and dealing with the huge volumes of data. Semantic communication, which only transmits the related information, is a promising solution to address these challenges due to its great potential to reduce required resources for transmission significantly [1], [2].

Traditional communication systems convert data into bits at the transmitter and require accurate bit recovery at the receiver, which depends on good channel conditions and high SNRs. Semantic communications transmit and recover the meaning of the transmitted content directly and require no accurate bit recovery, thus, are more robust to the channels. Inspired by the emerging deep learning (DL) technologies, some initial works on DL-enabled semantic communications focus on semantic recovery at the receiver for text [3]–[5], image [6], and speech [7]. How to exploit semantic information for specific tasks at the effectiveness level is another key area and few researchers pay attention to this area. There exist works that only focus on the single-modal data, i.e., image classification [8] and image retrieval [9]. However, in the practical communication scenarios, the system is required to gather, transmit, and fuse multimodal data from multiple users. This motivates us to develop multimodal multi-user semantic communication systems.

Multimodal data refer to describing scenarios with different modalities. Typical examples of multimodal data include audio, video, images from electro-optical sensors, and text from radio frequency sensors. If a task needs more than one modality to perform, multimodal data are correlated in the context. Compared with the situations with single-modal data, multimodal data can provide more information for intelligent tasks, introduce new degrees of freedom, and improve performance of intelligent tasks [10], [11]. The recent successful approaches for multimodal data fusion are mostly based on neural networks, and representative techniques include Deep Belief Net (DBN), Stacked Autoencoder (SAE), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) [11]. Multimodal semantic communications employ more than one users to serve only one multimodal intelligent task, which is suitable for the emerging autonomous scenarios, i.e., autonomous checkout at retail stores. To build a multi-user semantic communication system for supporting multimodal data, we face two challenges: how to extract the proper semantic information from each user and how to build a model for multimodal semantic information fusion at the receiver.

In this letter, we present our initial results in multi-user semantic communication for multimodal data. The detailed contributions are summarized as follows:

- A novel framework for the task-oriented multimodal data semantic communication system, named MU-DeepSC, is established, where the transceiver is jointly designed to perform the intelligent task. The visual question answering (VQA) task is adopted as an example to demonstrate the effectiveness of MU-DeepSC.
- The MU-DeepSC transmitter adopts the memory, attention, and composition (MAC) neural network to process the correlated data. By extracting the semantic information of images and text from different transmitters, the MU-DeepSC receiver will directly generate the answers based on the received semantic information at the receiver.
- The simulations demonstrate that the proposed MU-DeepSC has the ability to transmit the image and text semantic information and perform data fusion at the receiver.

The rest of this letter is organized as follows. Section II details the proposed MU-DeepSC. Numerical results are presented in Section III to show the performance of the DeepSC. Finally,

Huiqiang Xie and Zhijin Qin are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK (e-mail: h.xie@qmul.ac.uk, z.qin@qmul.ac.uk).

Geoffrey Ye Li is with School of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK (e-mail: geoffrey.li@imperial.ac.uk).

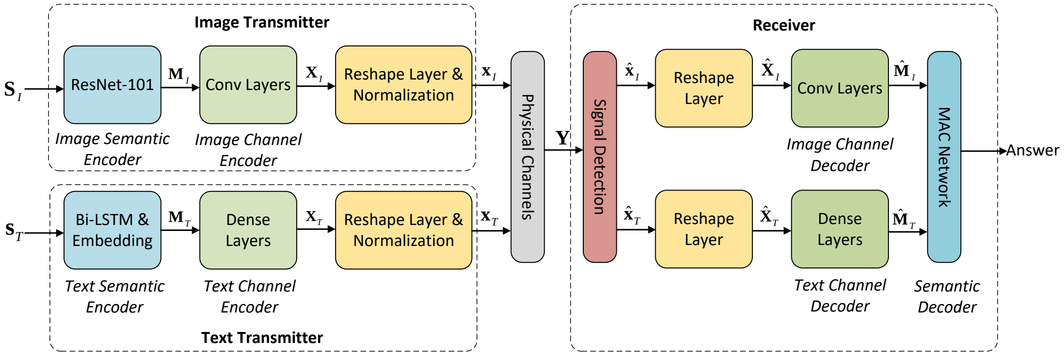


Fig. 1. The structure of proposed MU-DeepSC.

Section IV provides the conclusion.

II. Proposed MU-DeepSC Transceiver

In this section, we adopt the image and text information as an example, where two users with single antenna and one receiver with M antennas are considered for simplicity. It is easy to expand the network with the input of multi-image and multi-text. Besides, we design a deep neural network (DNN) for the considered semantic communication system to serve the VQA task, named as MU-DeepSC in Fig. 1, of which the MAC network is adopted for answering questions. All models can be trained in the cloud and then broadcast to users.

A. The Proposed MU-DeepSC

As shown in Fig. 1, the proposed MU-DeepSC consists of an image transmitter, a text transmitter, and a receiver.

1) Image Transmitter: For the image transmitter in Fig. 1, which includes a semantic encoder and channel encoder. Particularly, the ResNet-101 is used for the semantic encoder and CNNs with different units are adopted for channel encoder to generate transmitted symbols. Before computing semantic information, images will be resized to a commonly used resolution, 224 × 224, with bicubic interpolation, $\mathbf{S}_{\mathrm{I}} \in \mathbb{R}^{1\times 3\times 224\times 224}$. Then, we extract the semantic information by the image semantic encoder, which employs the first 30 blocks from the ResNet-101 network pre-trained on ImageNet. Notice that the ResNet-101 model will be frozen during training, as it has been well trained by more than one million images and is powerful enough to extract image semantic information. Then, the semantic image information can be extracted by the image semantic encoder, denoted as

$$\mathbf{M}_I = \mathcal{SE}_I\left(\mathbf{S}_I; \boldsymbol{\alpha}_I\right), \tag{1}$$

where $\mathbf{M}_I \in \mathbb{R}^{1\times C_1\times 14\times 14}$, where C_1 is the number of feature maps, $\mathbf{S}_I$ is the input resized image and $\boldsymbol{\alpha}_I$ is the trainable parameters.

After passing through the semantic encoder, the captured semantic information is mapped to the transmitted symbols directly by the image channel encoder, which consists of CNN layers due to its characteristics of learning different local features of image effectively. Therefore, the transmitted symbols can be represented by the image channel encoder, denoted by

$$\mathbf{X}_I = \mathcal{CE}_I\left(\mathbf{M}_I; \boldsymbol{\beta}_I\right), \tag{2}$$

where $\mathbf{X}_I \in \mathbb{R}^{1\times C_2\times 14\times 14}$ is the compressed semantic image information, where C_2 is the compressed dimension with $C_2 < C_1$, and $\boldsymbol{\beta}_I$ is the trainable parameters.

In the above, $\mathbf{X}_I$ is real signal rather than complex, which is not suitable to transmission. Therefore, $\mathbf{X}_I$ should be converted into the complex signal, $\mathbf{x}_I \in \mathbb{C}^{1\times 98C_2}$, where $98C_2 = \frac{14\times 14\times C_2}{2}$, firstly with reshape layer, which is then normalized by

$$l_{norm}(\mathbf{x}_I) = \frac{\mathbf{x}_I}{\mathbb{E}(\|\mathbf{x}_I\|_2)}, \tag{3}$$

for physical channel transmission.

2) Text Transmitter: In the text transmitter in Fig. 1, the bi-directional long short term memory (Bi-LSTM) is used for the semantic encoder and dense layers with different units for the channel encoder to generate transmitted symbols. Assume that $\mathbf{s}_T = (s_{1,T}, s_{2,T}, \cdots, s_{L,T})$ is the transmitted sentence with L words, where $s_{l,T}$ is the l-th word in the sentence to be transmitted. Before extracting its semantic information, the sentence will be embedded to map words to the numerical vector, $\mathbf{S}_T \in \mathbb{R}^{1\times L\times L_{embed}}$, by the embedding layer with L_{embed} embedding dimension, which can be trained for better word representation.

We employ one layer Bi-LSTM to extract the semantic representations of the input sentence. The corresponding text semantic encoder can be expressed as

$$\mathbf{M}_T = \mathcal{SE}_T\left(\mathbf{S}_T; \boldsymbol{\alpha}_T\right), \tag{4}$$

where $\mathbf{M}_T \in \mathbb{R}^{1\times L\times K_1}$, where K_1 is the post-processed dimension from L_{embed} to K_1, $\mathbf{S}_T$ is the embedded word vectors and $\boldsymbol{\alpha}_T$ is trainable parameters . Then, the text is compressed and mapped to the transmitted text symbols via the channel encoder, which includes several dense layers to preserve all text semantic information to preserve the entire input information. Condescendingly, the transmitted symbols can be calculated by the image channel encoder as

$$\mathbf{X}_T = \mathcal{CE}_I\left(\mathbf{M}_T; \boldsymbol{\beta}_T\right), \tag{5}$$

where $\mathbf{X}_T \in \mathbb{R}^{1\times L\times K_2}$ is the compressed semantic text information, where K_2 is the dimension after text channel encoder with $K_2 < K_1$, and $\boldsymbol{\beta}_T$ is the trainable parameters.

Similar to the image transmitter, the transmitted signal, $\mathbf{X}_T$, will be reshaped into the complex signal, $\mathbf{x}_T \in \mathbb{C}^{1\times\frac{K_2 L}{2}}$, firstly and normalized by

$$l_{norm}(\mathbf{x}_T) = \frac{\mathbf{x}_T}{\mathbb{E}(\|\mathbf{x}_T\|_2)}. \tag{6}$$

3) Receiver: The receiver is shown in Fig. 1(c), where convolution layers with different units are used for the image channel decoder, dense layers with different units for the text channel decoder, and the MAC network is adopted for the semantic decoder. The received symbols are detected firstly, then various semantic information is recovered through different channel decoders, and is finally merged the various semantic information to get answers.

Assume that V is the least common multiple between the length of image semantic and the length of text semantic information, the $M \times V$ signal received at the receiver can be expressed as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N}, \tag{7}$$

where $\mathbf{H} \in \mathbb{C}^{M\times 2}$ is the channel between the BS and users, $\mathbf{X} = [\mathbf{x}_I, \mathbf{x}_T] \in \mathbb{C}^{2\times V}$ denotes transmit symbols from text and image users in the considered system, and $\mathbf{N} \in \mathbb{C}^{M\times V}$ indicates the circular symmetric Gaussian noise, items of $\mathbf{N}$ are of variance σ_n^2.

Here, we employ the additional domain knowledge, i.e., channel estimation, to improve the training speed and enhance the final decision accuracy. With the channel gain and zero-forcing detector, the transmitted signal can be estimated by

$$\hat{\mathbf{X}} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{Y} \quad = \mathbf{X} + \hat{\mathbf{N}}, \tag{8}$$

where $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_I; \hat{\mathbf{x}}_T]$ is the estimated information for the text and image users, $\hat{\mathbf{N}} = \left(\mathbf{H}^H\mathbf{H}\right)^{-1}\mathbf{H}^H\mathbf{N}$ represents the impact of noise. With operation in (8), the channel effect is transferred from multiplicative noise to additive noise, which significantly reduces the learning burden.

After signal detection, the estimated complex signals will be reshaped to the size suitable for the following neural networks with reshape layer firstly, $\hat{\mathbf{x}}_I \to \hat{\mathbf{X}}_I : \mathbb{C}^{1\times 98C_2} \to \mathbb{R}^{1\times C_2\times 14\times 14}$ and $\hat{\mathbf{x}}_T \to \hat{\mathbf{X}}_T : \mathbb{C}^{1\times\frac{K_2 L}{2}} \to \mathbb{R}^{1\times L\times K_2}$. Then, the signals are semantically recovered information by the channel decoders for text and image, denoted as

$$\hat{\mathbf{M}}_I = \mathcal{CD}_I\left(\hat{\mathbf{X}}_I; \gamma_I\right), \tag{9}$$

and

$$\hat{\mathbf{M}}_T = \mathcal{CD}_T\left(\hat{\mathbf{X}}_T; \gamma_T\right), \tag{10}$$

respectively, where $\hat{\mathbf{M}}_I \in \mathbb{R}^{1\times C_1\times 14\times 14}$, $\hat{\mathbf{M}}_T \in \mathbb{R}^{1\times L\times K_1}$, γ_I and γ_T are the corresponding trainable parameters. Similar to the channel encoders, the image and text channel decoder consists of CNN layers and dense layers to decompress and recover semantic information.

With text and image semantic information, we employ the MAC network [12] as the semantic decoder to merge the text and image semantic information as well as to answer the vision questions, which is written as

$$\texttt{Task} = \mathcal{SD}\left(\left(\hat{\mathbf{M}}_I, \hat{\mathbf{M}}_T\right); \boldsymbol{\varphi}\right), \tag{11}$$

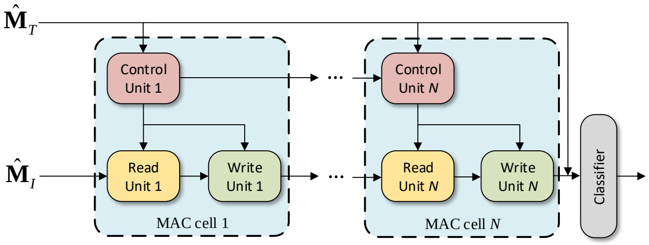


Fig. 2. The structure of the memory, attention, and composition network for text and image semantic information fusion and the VQA.

where $\mathcal{SD}(\cdot; \boldsymbol{\varphi})$ is the semantic decoder with trainable parameters $\boldsymbol{\varphi}$. The MAC network shown in Fig. 2 consists of multiple MAC cells, of which each contains *control unit, read unit*, and *write unit*. The *control unit* first generates a query based on the received text semantic information, i.e., the object of question and the type of question, by the attention mechanism, then the *read unit* gets the query and searches the corresponding key from image semantic information by another attention module. Finally, the *write unit* integrates information and outputs the predicted answers to the questions.

B. Loss Function

As indicated before, the objective of the proposed MU-DeepSC is to answer the questions based on the images and texts. The proposed transceiver is task-oriented, where the answers will be predicted directly at the MU-DeepSC receiver. As the image and text will not be recovered in the MU-DeepSC system like traditional communication systems, loss functions based on bit-error or symbol-error are not applicable anymore. In order to improve the accuracy of answers, the cross-entropy (CE) is used as the loss function to measure the difference between the correct answer, a, and the predicted answer, $\hat{a}$, which can be formulated as

$$\mathcal{L}_{\mathrm{CE}}(a, \hat{a}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi}) = -p\,(a)\log\left(p\,(\hat{a})\right), \tag{12}$$

where $p(a)$ is the real probability of the answer, and $p(\hat{a})$ is the probability of the predicted answer. The CE can measure the difference between the two probability distributions. By reducing the loss value of CE, the network learns the correct answer first and tries to predict the answer with the highest probability of accuracy. Then the network can be optimized by gradient descent. The training procedure is described in Algorithm 1.

III. Simulation Results and Discussions

In this section, we compare the proposed MU-DeepSC and the traditional source coding and channel coding methods over different channels, where the perfect CSI is assumed for all methods. The transceiver is assumed with two single-antenna users and the receiver with two antennas.

A. Implementation Settings

The adopted Dataset is CLEVR [13], which consists of a training set of 70,000 images and 699,989 questions and a test set of 15,000 images and 149,991 questions.

Algorithm 1: MU-DeepSC Training Algorithm.

```
Initialization: Load pre-trained model for $\mathcal{SE}_I\left(;\boldsymbol{\alpha}_I\right)$, and
                initialize $\boldsymbol{\alpha}_T, \boldsymbol{\beta}_I, \boldsymbol{\beta}_I, \boldsymbol{\gamma}_I, \boldsymbol{\gamma}_T, \boldsymbol{\varphi}$.
Data: The training dataset $\mathcal{D}$.
Function Train Whole Network():
    Image Transmitter:
        $\mathcal{SE}_I\left(\mathbf{S}_I;\boldsymbol{\alpha}_I\right) \rightarrow \mathbf{M}_I$; $\mathcal{CE}_I\left(\mathbf{M}_I;\boldsymbol{\beta}_I\right) \rightarrow \mathbf{X}_I$,
        Reshape first $\mathbf{X}_I$ and power normalization by (3),
        Transmit $\mathbf{x}_I$ over the channel.
    Text Transmitter:
        Embedding the input sentence $\mathbf{s}_T \rightarrow \mathbf{S}_T$,
        $\mathcal{SE}_T\left(\mathbf{S}_T;\boldsymbol{\alpha}_T\right) \rightarrow \mathbf{M}_T$; $\mathcal{CE}_T\left(\mathbf{M}_T;\boldsymbol{\beta}_T\right) \rightarrow \boldsymbol{X}_T$,
        Reshape $\mathbf{X}_T$ first and power normalization by (6),
        Transmit $\mathbf{x}_T$ over the channel.
    Receiver:
        Receive $\mathbf{Y}$ and Signal detection by (8) to get $\hat{\mathbf{x}}_I$, $\hat{\mathbf{x}}_T$,
        $\mathcal{CD}_I\left(\hat{\mathbf{x}}_I;\boldsymbol{\gamma}_I\right) \rightarrow \hat{\mathbf{M}}_I$; $\mathcal{CD}_T\left(\hat{\mathbf{x}}_T;\boldsymbol{\gamma}_T\right) \rightarrow \hat{\mathbf{M}}_T$,
        $\mathcal{SD}\left(\left(\hat{\mathbf{M}}_I, \hat{\mathbf{M}}_T\right);\boldsymbol{\varphi}\right) \rightarrow \hat{a}$,
    Compute the loss by (12) with $a$, $\hat{a}$.
    Train $\boldsymbol{\alpha}_T, \boldsymbol{\beta}_I, \boldsymbol{\beta}_I, \boldsymbol{\gamma}_I, \boldsymbol{\gamma}_T, \boldsymbol{\varphi} \rightarrow$ Gradient descent.
    Return: $\mathcal{SE}_I\left(;\boldsymbol{\alpha}_I\right), \mathcal{SE}_T\left(;\boldsymbol{\alpha}_T\right), \mathcal{CE}_I\left(;\boldsymbol{\beta}_I\right), \mathcal{CE}_T\left(;\boldsymbol{\beta}_T\right)$,
            $\mathcal{CD}_I\left(;\boldsymbol{\gamma}_I\right)$, $\mathcal{CD}_T\left(;\boldsymbol{\gamma}_T\right)$, and $\mathcal{SD}\left(;\boldsymbol{\varphi}\right)$
```

In the experiments, text will be embedded by embedding layer, which is initialized from Gaussian distribution with zero mean and unit variance, $\mathcal{N}(0,1)$, with shape (vocab size, embedding-dim). The embedding dimension is set to be 300. C_1, C_2, K_1, and K_2 are set 512, 128, 512, and 256, respectively. The image channel coder consists of four convolutional layers with 256, 128, 256, 512 filters, the first two of which are image channel encoder, the rest of which are image channel decoder. Each convolutional layer is with a 3×3 kernel and followed by an ELU activation function. The text channel coder consists of five dense layers with 256, 256, 256, 256, 512 neurons, the first two of which are text channel encoder, the rest of which are text channel decoder. Each Dense layer is followed by a ReLU activation function and the outputs of the channel decoder are normalized by LayerNorm. The MAC network consists of 12 cells. We employ the ADAM optimizer with a learning rate of 0.0001. Different from predicting with image and text, we mainly consider two cases for the baselines, only using question or image to predict the answer, and the typical separate source and channel coding,

- Error-free transmission: The full, noiseless images and texts are input ResNet-101 and Bi-LSTM for extracting features and then input to the MAC network.
- Traditional method: To perform the source and channel coding separately, we use the following technologies respectively:
  - Joint photographic experts group (JEPG) as the image source coding, a commonly used method of lossy compression with a compression rate of 75 for digital images, Huffman coding for text source coding, lossless compression for text,
  - Low-density parity-check codes (LDPC) with a coding rate of 1/3 as the channel coding, especially for the large data size.
- Text or image only based prediction: The transmitter is the same as the text user or image user in Fig. 1, but the receiver replaces the MAC network with a one-layer classifier.

The modulation method for the traditional method is 16 quadrature amplitude modulation (QAM). For the traditional method, the recovered image and text are input to the MAC network to get answers. We compare the proposed methods in terms of answer accuracy, the number of transmitted symbols, and computational complexity.

B. Performance of MU-DeepSC

Fig. 3 shows the relationship between the answers accuracy and SNR over AWGN, Rayleigh fading channels, and Rician fading channels. Among the methods in Fig. 3, the proposed MU-DeepSC outperforms other baselines, especially in the low SNR regime, and is about to approach the upper bound at high SNR regime. Besides, over all SNR regimes, transmitting single source, such as text or image only, has the similar answer accuracy over three channels. Moreover, in Fig. 3(a), the traditional method performs worse at the lower SNR regime since the images are corrupted by error bits, but performing higher accuracy as the SNR increases. For more complex channels, the answer accuracy of the traditional method can increase slowly as SNR increases in Fig. 3(b) and 3(c). Besides, compared with the separate source-channel coding in traditional communications, the proposed MU-DeepSC is jointly optimized to achieve better performance at the answer accuracy.

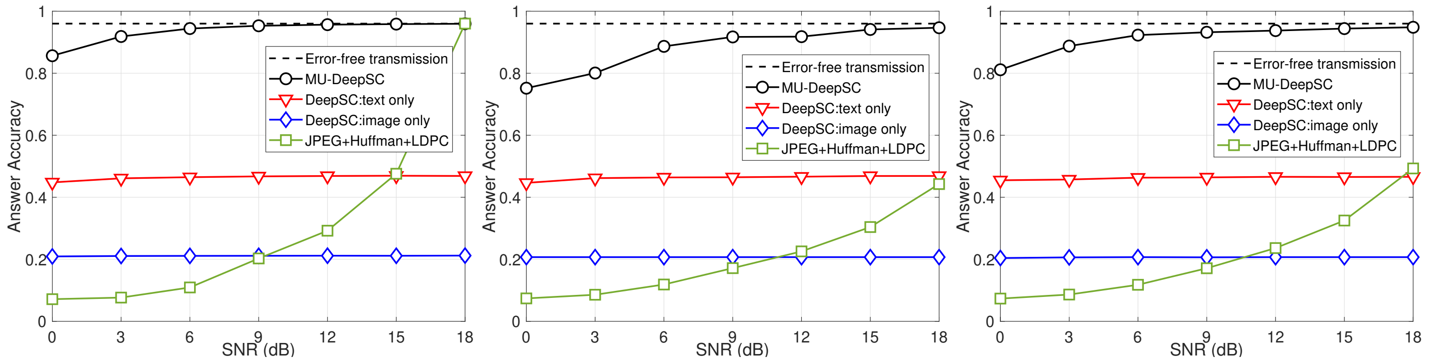
Part of visualized results are shown in Fig. 4. The proposed MU-DeepSC correctly answers the all questions. The traditional communications and the DeepSC based text can only give the right answers for partial questions. The reason that transmitting text-only answers correctly is that the text information can help the system guess and narrow the search range of answers.

Table I compares the proposed MU-DeepSC and traditional communications at the number of transmitted symbols and computational complexity[1] by measuring one image or one word. For image transmission, the proposed MU-DeepSC significantly decreases the number of transmitted symbols and the computational complexity. For text transmission, the MU-DeepSC transmits more symbols than the traditional communications but with similar computational complexity, in which a larger number of symbols can provide robustness to channels and low SNRs. In general, the proposed MU-DeepSC can save the transmission and processing time for images, by slightly sacrificing the transmission time for text but keeping a similar processing time.

IV. Conclusion

In this letter, we have established a multi-user semantic communication system, named MU-DeepSC, for exploiting the correlated image and text information, where the VQA

[1] We only analyze the complexity of channel coding for both methods because the other parts are shared in both methods and the complexity of source coding is low to be omitted.

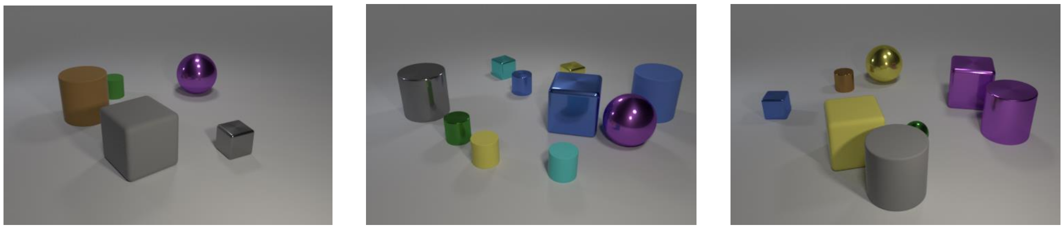


(a) AWGN Channels.

(b) Rayleigh Fading Channels.

(c) Rician Fading Channels.

Fig. 3. Answer accuracy for various testing channels based on different trained models.



***Q**: Are there any other things that are the same shape as the big metallic object?*

***A (MU-DeepSC)**: no* ✓
***A (JPEG+Huffman+LDPC)**: yes* ✗
***A (Only-Text)**: no* ✓
***A (Only-Image)**: yellow* ✗

***Q**: The other small shiny thing that is the same shape as the tiny yellow shiny object is what color?*

***A (MU-DeepSC)**: cyan* ✓
***A (JPEG+Huffman+LDPC)**: cyan* ✓
***A (Only-Text)**: gray* ✗
***A (Only-Image)**: no* ✗

***Q**: There is a yellow thing to the right of the rubber thing on the left side of the gray rubber cylinder; what is its material?*

***A (MU-DeepSC)**: metal* ✓
***A (JPEG+Huffman+LDPC)**: rubber* ✗
***A (Only-Text)**: yes* ✗
***A (Only-Image)**: 8* ✗

Fig. 4. Some visualized results for VQA task over Rician fading channels with SNR = 18dB , where the first row is the transmitted images, the second row is the transmitted questions, and the last four rows are predicted answers by proposed MU-DeepSC, traditional methods, MU-DeepSC with text only, and MU-DeepSC with image only, respectively.

TABLE I
COMPARISON BETWEEN MU-DEEPSC AND TRADITIONAL COMMUNICATIONS AT THE NUMBER OF TRANSMISSION SYMBOLS AND COMPUTATIONAL COMPLEXITY FOR ONE IMAGE OR ONE WORD

Method	Source	No. Symbols	Computational Complexity	
			Additions	Multiplications
MU-DeepSC	Image	12,544	2.6×10^8	2.9×10^8
	Text	128	4.6×10^5	4.6×10^5
Traditional Communications	Image	41,718	1.0×10^9	1.1×10^9
	Text	15	6.8×10^5	3.6×10^5

task is considered. By jointly designing the semantic encoder and the channel encoder by learning and extracting the essential semantic information, the proposed MU-DeepSC can handle the image and text semantic information effectively and predict the answers accurately by merging different semantic information. The simulation results have demonstrated that the MU-DeepSC outperforms various benchmarks, especially in the low SNR regime. Hence, we are highly confident that the proposed MU-DeepSC is a promising candidate for multi-user semantic communication systems to transmit multimodal data.

REFERENCES

[1] Z. Qin *et al.*, "Deep learning in physical layer communications," *IEEE Wire. Comm.*, vol. 26, no. 2, pp. 93–99, 2019.

[2] G. Shi *et al.*, "A new communication paradigm: from bit accuracy to semantic fidelity," *arXiv preprint arXiv:2101.12649*, 2021.

[3] F. Nariman *et al.*, "Deep learning for joint source-channel coding of text," in *Proc. IEEE Int'l. Conf. Acoustics Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2326–2330.

[4] H. Xie *et al.*, "Deep learning enabled semantic communication systems," *IEEE Trans. on Signal Process.*, vol. 69, pp. 2663–2675, Apr. 2021.

[5] H. Xie *et al.*, "A lite distributed semantic communication system for internet of things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, Jan. 2021.

[6] E. Bourtsoulatze *et al.*, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, May 2019.

[7] Z. Weng *et al.*, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Jun. 2021.

[8] C. Lee *et al.*, "Deep learning-constructed joint transmission-recognition for internet of things," *IEEE Access*, vol. 7, pp. 76 547–76 561, Jun. 2019.

[9] M. Jankowski *et al.*, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, Jan. 2021.

[10] D. Lahat *et al.*, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[11] J. Gao *et al.*, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.

[12] D. A. Hudson *et al.*, "Compositional attention networks for machine reasoning," in *Int'l. Conf. on Learning Repre., ICLR, Vancouver, BC, Canada,*, Apr. 2018.

[13] J. Johnson *et al.*, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recog.*, 2017, pp. 2901–2910.