Contrastive Learning Enables Low-Bandwidth Semantic Communication

Eleonora Grassucci, Giordano Cicchetti, Danilo Comminiello*

Dept. of Information Eng., Electronics, and Telecommunications, Sapienza University of Rome, Italy

Abstract

The rapid growth of multimodal data streaming requiring more and more bandwidth is posing new challenges for communication systems. Concurrently, by transmitting only the semantic information and not the whole original bitstream, the novel semantic communication paradigm promises to reduce the bandwidth requirements. However, for multimodal data transmission, conventional semantic communication frameworks also require conveying a considerable amount of information for each modality, resulting in a high transmission load. In this paper, we propose to model the semantic latent space with a novel contrastive learning loss, so as to extract the centroid representing the semantic content of the respective cluster and transmit over the channel just one single compressed representation, regardless of the number of modalities. We show how the proposed framework allows a considerable reduction of the bandwidth while preserving multimodal reconstruction results with respect to conventional approaches.

Introduction

In recent years, the creation and exchange of multimodal and multimedia content have been rapidly growing, with video streaming representing the majority share of internet traffic today, with estimates up to 80% [1]. For wireless communications, this multimedia data comprising different modalities such as video, audio, text, and so on, poses new challenges in terms of bandwidth requirements for the transmission (Du et al. 2024; Bocus, Wang, and Piechocki 2023; Tandon et al. 2021).

In the last few years, a novel communication paradigm has gained attention, shifting from the first level (the technical level) of Shannon and Weaver (Weaver 1953) communication theory to the second one, the semantic level. The so-called semantic communication paradigm relies on transmitting only the semantic information necessary to recover the meaning of the message or accomplish some predefined tasks at the receiver over the communication channel (Xie et al. 2021; Dai et al. 2023; Choi et al. 2024). Semantic communication systems promise to reduce bandwidth requirements and expand possible applications horizons, especially when endowed with large learning models such as diffusion models (Grassucci, Barbarossa, and Comminiello 2023; Zeng et al. 2024; Grassucci et al. 2024) or large language models (Zhao et al. 2024). Under this branch, language-oriented frameworks obtain interesting results by drastically reducing the necessary bandwidth for image transmission. By means of image-to-text generative models, these frameworks extract the textual description of the image and then transmit over the channel such a compressed semantic representation of image content (Nam et al. 2024, 2023). At the receiver, a text-to-image generative model maps the caption back to the image domain, regenerating the intended image. Despite the textual description allows a crucial reduction of necessary bandwidth with respect to transmitting the whole image, regenerated content may severely differ from the original one (Cicchetti et al. 2024). To address this limitation, other works proposed to help the receiver generation with a latent representation of the original image and give both the latent and the text embeddings to the generative model to more precisely guide the generative process (Cicchetti et al. 2024). Although the improved results, when dealing instead with multiple modalities, such methods do not scale well and still require transmitting a consistent amount of bits over the channel, requiring minimum the transmission of a latent representation vector for each modality.

In this paper, we propose a novel semantic communication framework able to reconstruct multiple modalities at the receiver while transmitting only one representation over the channel, regardless of the number of modalities. Our framework builds a semantically aligned latent space with specialized novel contrastive loss functions. In this space, semantically similar content tends to cluster together, regardless of the modality, and representations with diverse semantic meanings are pushed away from each other, as shown in Fig.2. From each semantic cluster, we can extract the average semantic information of that cluster, corresponding to the centroid. The latter contains highly informative semantic content, therefore, we then transmit just the centroid vector

*This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (PNRR) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program RESTART).

[1]Cisco visual networking index: global mobile data traffic forecast update, 2017–2022, Available: https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf

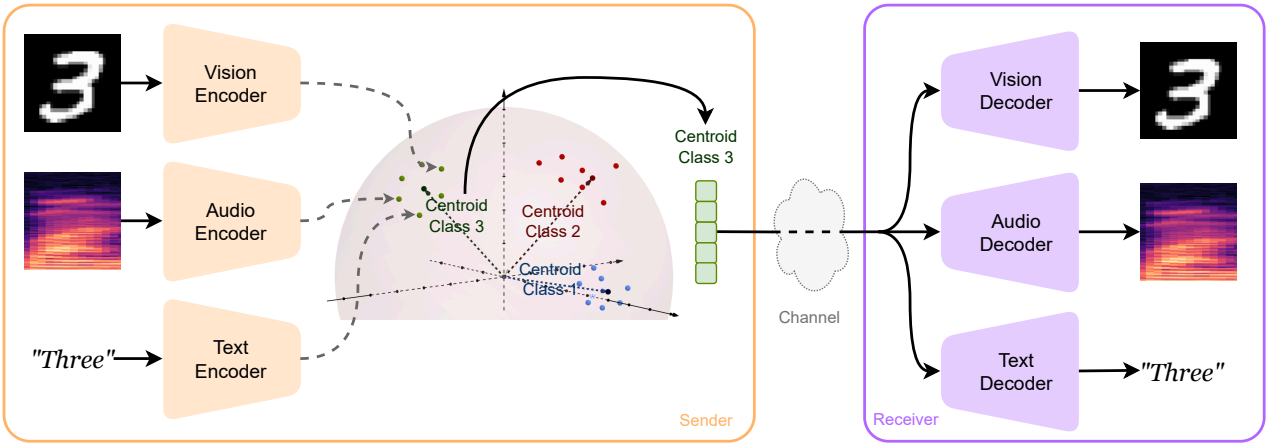

Figure 1: The proposed framework. The encoders are trained to build a semantically-aligned latent space and the decoders to reconstruct the original data from the received latent vectors. At inference time, we extract the centroid of the respective semantic cluster and transmit solely the centroid vectors over the channel. Then, the decoders reconstruct the original data starting from the centroid vector in input preserving the semantic information of the transmitted data.

over the channel up to the receiver and use it as input to the decoders for reconstruction. Given the semantically aligned latent space, decoders can almost perfectly reconstruct the data even though decoders for different modalities receive the same single latent vector in input. Therefore, together with comparable reconstructions, the proposed method can crucially reduce the bandwidth required for the transmission of multimodal data by 67% in the three-modal case and potentially to $1/n$ in the case of n modalities.

In summary, our main contributions are: i) We propose a novel semantic communication framework based on contrastive learning that crucially reduces the bandwidth required for transmission. ii) We propose a novel contrastive loss function based on centroids to build a more aligned semantic space. iii) We test the framework in a vanilla three-modal scenario, proving that we can reduce the required bandwidth while preserving the reconstruction performance.

Proposed Method

Preliminaries

Let us consider a batch of B multimodal data samples, we can extract latent representation from each sample and each modality using encoder functions E. This results in $N \times B$ embeddings that we can define as follow: $\mathbf{m}_{ij}$ is the j-th embedding of the i-th modality normalized at norm 1:

$$\mathbf{m}_{ij} = normalized(E_i(\mathbf{x}_{ij})). \tag{1}$$

Given $\mathbf{x}_{ij}$ the input j-th sample of modality i and given $\mathbf{m}_{ij}$ its normalized latent representations, we can pass this embedding to the corresponding decoder of the i-th modality D_i obtaining $\hat{\mathbf{x}}_{ij}$ as the reconstructed j-th sample for modality i-th.

$$\hat{\mathbf{x}}_{ij} = D_i(\mathbf{m}_{ij})). \tag{2}$$

Finally, given a batch of latent embeddings, we can compute B centroids by simply averaging latent vectors correspondent to co-occurrent input sample, i.e. :

$$\mathbf{c}_j = \frac{1}{N}\sum_{i=0}^{N} \mathbf{m}_{ij}. \tag{3}$$

Framework

The proposed framework is composed of three encoder functions at the sender that take care of mapping the original data to the latent space, and three corresponding decoders at the receiver, whose aim is to recover the original data from the latent representation received. During training, the encoders learn to map the multimodal data into the latent space in a semantically meaningful way, while the decoders learn to reconstruct the original data from the latent vectors provided by the encoders. At the end of the training, the latent space will be semantically aligned and embeddings with the same semantic content will be clustered close to each other, regardless of the original modality. Therefore, during the transmission of content at inference time, we can extract the centroid of the corresponding cluster and solely transmit this vector over the channel, without requiring the transmission of one vector for each modality, crucially reducing the bandwidth requirements. Figure 1 shows the structure of the framework with the example of transmitting the image containing the *3* digit, the spectrogram of the spoken audio *"three"*, and the textual description of the number. Given the semantically aligned space, although the transmission of the centroid and not of each modality-specific latent vector, the decoders are able to properly reconstruct the original data.

Loss Functions

For each sample i and for each modality j, given $\mathbf{m}_{ij}$ its normalized latent representation in (1) and $\mathbf{c}_j$ the centroid for the j-th class as defined in (3), the aim of our training

process is to align $\mathbf{m}_{ij}$ to the corresponding centroid $\mathbf{c}_j$ and push away from the others.

The proposed contrastive loss function is based on the centroid computation and it is defined as:

$$\mathcal{L}_{M2C} = -\frac{1}{B}\sum_{i=1}^{N}\sum_{j=1}^{B}\log\frac{\exp(\mathbf{m}_{ij}^{\top}\mathbf{c}_j/\tau)}{\sum_{k=1}^{B}\exp(\mathbf{m}_{ik}^{\top}\mathbf{c}_k/\tau)}, \quad (4)$$

$$\mathcal{L}_{C2M} = -\frac{1}{B}\sum_{i=1}^{N}\sum_{j=1}^{B}\log\frac{\exp(\mathbf{c}_j^{\top}\mathbf{m}_{ij}/\tau)}{\sum_{k=1}^{B}\exp(\mathbf{c}_k^{\top}\mathbf{m}_{ik}/\tau)}. \quad (5)$$

Moreover, to encourage a sparsification in the latent space we introduce an additional loss function that is in charge of uniformly distributing the centroind all over the latent space:

$$\mathcal{L}_{Centroids} = -\frac{1}{B}\sum_{j=1}^{B}\log\frac{\exp(\mathbf{c}_j^{\top}\mathbf{c}_j/\tau)}{\sum_{k=1}^{B}\exp(\mathbf{c}_j^{\top}\mathbf{c}_k/\tau)}. \quad (6)$$

Normalizing centroids vector to the unitary norm, the utility of the latter loss function is only to spread away centroids that are different from each other by augmenting the angle between them.

The final centroid contrastive learning loss function takes the form:

$$\mathcal{L}_{CL} = \frac{1}{2}(\mathcal{L}_{M2C} + \mathcal{L}_{C2M}) + \mathcal{L}_{Centroids}. \quad (7)$$

Regarding the reconstruction part, for each modality, we can design specific reconstruction losses $\mathcal{L}_i$. Summed all together we have the total reconstruction loss:

$$\mathcal{L}_R = \sum_{i=0}^{N}\sum_{j=0}^{B}\mathcal{L}_i(\hat{\mathbf{x}}_{ij}, \mathbf{x}_{ij}). \quad (8)$$

In this work, as reconstruction losses, we consider a mean absolute error loss for the audio modality, a mean squared error for the image one, and a cross-entropy loss for the textual one.

The final loss function is a weighted sum between the contrastive loss and the reconstruction loss:

$$\mathcal{L}_{tot} = \mathcal{L}_{CL} + \lambda\mathcal{L}_R, \quad (9)$$

with λ a hyperparameter used to balance the two loss functions. In our experiments, we set $\lambda = 10$.

Experiments

We consider a vanilla scenario in which the sender has three different modalities to transmit: an image, an audio related to the image, and their textual description, all of them encoded with three different encoders. The receiver aims to recover the original data for each modality by means of three different decoders. The whole framework is shown in Fig. 1.

Datasets. We select two well-known datasets for our vanilla scenario, the MNIST dataset, comprised of 60,000 images, and the Audio-MNIST dataset (Becker et al. 2023), which comprises 30,000 audio samples of spoken digits (0-9) from diverse speakers with different accents. We compute the Mel spectrograms with 128 nmels, fmax at 8000, hop length equal to 512, and 2048 nfft. Textual descriptions are the digit words associated with the images and the audio, i.e., *"seven"* for the digit 7.

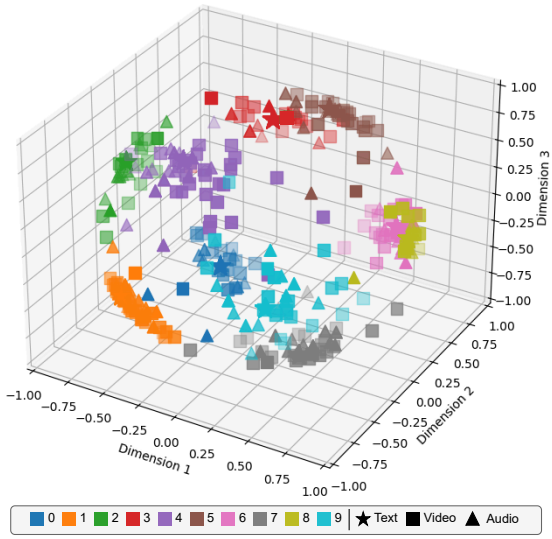

Figure 2: Aligned latent space with latent dim equal to 3 with three different modalities: text (stars), audio (triangles), images (squares). The classes are clustered together according to the semantics (i.e., the digits), regardless if they come from different modalities.

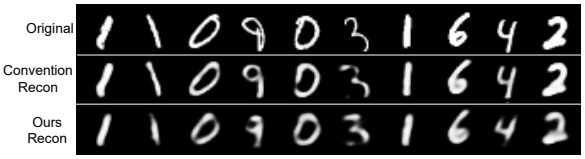

Figure 3: Reconstructed images for conventional method (second row) and ours (third row) with latent dim at 32.

Models. We build three vanilla autoencoders for image, audio, and text reconstruction. As a vision encoder, we consider a two-layer convolutional network (32 and 64 filters, respectively) with ReLU activation functions, a Max Pooling, and a final MLP layer to map the features into the latent space. To encode the audio modality, instead, we build a three-layer convolutional encoder with 16, 32, and 64 filters, respectively, ReLU activation functions, and an MLP layer to map also audio features into the latent space. Instead, the audio decoder is a stack of MLP layers of dimensions 64, 128, 192, and 128, with the final layer outputting 1 channel representing the audio spectrogram channel. We consider a simple encoder with Word2Vec (Mikolov et al. 2013) and a decoder with three MLP layers for autoencoding the text modality. Being comprised of ten words (numbers from 0 to 9), we encode our text corpora in one-hot encoding vectors.

Training. We experiment with different latent dimensions equal to 3, 16, or 32. Interestingly, setting a latent dimen-

Table 1: Results for audio reconstruction (MAE), image reconstruction (MSE), and text reconstruction (Acc). Additionally, Recall at 1 (R@1) score measuring the semantic alignment of the latent space. Moreover, we report the total transmitted information (TTI), measured as the dimension of latent vectors transmitted, and the TTI over the amount of total original information (TOI) for the conventional reconstruction method and our proposed framework.

Method	Latent dim.	MAE (↓)	MSE (↓)	Acc (↑)	R@1 (↑)	TTI	TTI/TOI
Conventional	3	0.099	0.049	100.0	100.0	9	0.002
Ours	3	0.099	0.050	100.0	100.0	3 (-67%)	0.001
Conventional	16	0.078	0.016	100.0	100.0	48	0.012
Ours	16	0.088	0.033	100.0	100.0	16 (-67%)	0.004
Conventional	32	0.079	0.014	100.0	100.0	96	0.025
Ours	32	0.091	0.038	100.0	100.0	32 (-67%)	0.008

sion of 3 allows us to directly visualize (and, thus, control) the true latent space the models shape, while for higher dimensions the plot is not possible without some projections or stochastic algorithms like tSNE or U-MAP. The learning rate is fixed for all the models to $1e-4$, the batch size is equal to 10 and we train the framework for 15000 iterations setting $\lambda = 10$. During training, the encoders are encouraged by the contrastive learning loss $\mathcal{L}_{CL}$ to build a semantically meaningful latent space with embeddings of different modalities with the same semantic content (i.e., the same digit) to cluster together. Concurrently, the reconstruction loss $\mathcal{L}_R$ helps decoders to reconstruct the original data at the receiver. During training, the decoders receive in input their respective latent vectors encoded by the encoders, therefore the vision decoder receives the vision embedding, the audio decoder receives the audio embedding, and so on. At inference time, once the latent space is properly modeled, all the decoders receive as input a single (and equal for each of them) latent vector, which is the centroid of the cluster with semantic information corresponding to the original multimodal data transmitted.

Metrics. To evaluate the performance of the proposed framework, we analyze all the aspects of the framework, ranging from the quality of reconstructions to a measure of the semantic alignment of the latent space. We evaluate the audio reconstruction with the mean absolute error (MAE), the image reconstruction with the common mean squared error (MSE), and the text one with the Accuracy (Acc), as we encode this data into one-hot encoding vectors. Finally, to measure the alignment of the latent space, we compute the recall at 1 (R@1) for the task of multimodal video-audio-text retrieval.

Results. The first determining result to analyze is the semantic alignment of the latent space, which is crucial to provide meaningful semantic information by transmitting the clusters centroid at the decoders. By setting the latent dimension to 3, we can directly plot this space and have a look on how the encoders and the contrastive loss contribute to shape the space. Figure 2 shows the resulting latent space. Regardless of their modality, latent vectors representing the same semantic content, that is the digit, tend to cluster together (by color). This proves the effectiveness of the contrastive learning training with the centroid loss proposed in (4) and that centroids will be meaningful and representative of their respective cluster.

Figure 3 shows some random samples of image reconstruction. The first row is the original content, the second row corresponds to images reconstructed by the decoder receiving in input the encoded image latent vector, therefore as in classical recovering approaches. Finally, the third row shows the reconstructed images with the proposed method, thus when the decoder receives the centroid of the semantic cluster and reconstructs the images from it. As it is clear in Fig. 3, although the quality is a little bit degraded, the semantic content is highly preserved also in the proposed method, in which digits are clearly recognizable. Therefore, we can conclude that the semantic communication has been effectively accomplished.

Finally, the results of the quantitative evaluation are reported in Tab. 1 for different latent dimensions. Interestingly, for the smaller latent dimension, our method achieves barely the same results as conventional methods, which, however, require much more bandwidth for transmission. Indeed, our method allows a consistent reduction of bits equal to 67% while preserving performance in almost all scenarios. With a dimension equal to 32 our method still obtains good results, but probably conventional methods manage to encode more details in the latent vector and achieve slightly better performance at a high cost in terms of bandwidth requirements.

Conclusion

In this paper, we proposed a novel semantic communication framework for multimodal data communication. The framework is based on a novel contrastive learning loss function that shapes a semantically aligned latent space allowing a crucial reduction of transmitted information over the communication channel. During inference, the proposed framework transmits over the channel only the centroid associated with the respective cluster allowing a considerable reduction of bandwidth requirements, while the receiver decoders still reconstruct original data with barely any loss in performance.

References

Becker, S.; Vielhaben, J.; Ackermann, M.; Müller, K.-R.; Lapuschkin, S.; and Samek, W. 2023. AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*.

Bocus, M. J.; Wang, X.; and Piechocki, R. J. 2023. Streamlining Multimodal Data Fusion in Wireless Communication and Sensor Networks. *IEEE Transactions on Cognitive Communications and Networking*, 10: 252–262.

Choi, J.; Park, J.; Grassucci, E.; and Comminiello, D. 2024. Semantic Communication Challenges: Understanding Dos and Avoiding Don'ts. *IEEE Veichular and Tech. Conf. (VTC) Spring*.

Cicchetti, G.; Grassucci, E.; Park, J.; Choi, J.; Barbarossa, S.; and Comminiello, D. 2024. Language-Oriented Semantic Latent Representation for Image Transmission. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Dai, J.; Zhang, P.; Niu, K.; Wang, S.; Si, Z.; and Qin, X. 2023. Communication Beyond Transmitting Bits: Semantics-Guided Source and Channel Coding. *IEEE Wireless Communications*, 30(4): 170–177.

Du, J.; Lin, T.; Jiang, C.; Yang, Q.; Bader, C. F.; and Han, Z. 2024. Distributed Foundation Models for Multi-Modal Learning in 6G Wireless Networks. *IEEE Wireless Communications*, 31(3): 20–30.

Grassucci, E.; Barbarossa, S.; and Comminiello, D. 2023. Generative Semantic Communication: Diffusion Models Beyond Bit Recovery.

Grassucci, E.; Marinoni, C.; Rodriguez, A.; and Comminiello, D. 2024. Diffusion models for audio semantic communication. In *IEEE Int. Conf. on Audio, Speech, and Signal Process. (ICASSP)*.

Mikolov, T.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.

Nam, H.; Park, J.; Choi, J.; Bennis, M.; and Kim, S.-L. 2024. Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation. In *IEEE Int. Conf. on Acoustics, Speech and Signal Process.*

Nam, H.; Park, J.; Choi, J.; and Kim, S.-L. 2023. Sequential Semantic Generative Communication for Progressive Text-to-Image Generation. In *20th Annual IEEE Int. Conf. on Sensing, Comm., and Netw. (SECON)*, 91–94.

Tandon, P.; Chandak, S.; Pataranutaporn, P.; Liu, Y.; Mapuranga, A. M.; Maes, P.; Weissman, T.; and Sra, M. 2021. Txt2Vid: Ultra-Low Bitrate Compression of Talking-Head Videos via Text. *IEEE Journal on Selected Areas in Communications*, 41: 107–118.

Weaver, W. 1953. Recent contributions to the mathematical theory of communication. *ETC: A Review of General Semantics*, 261–281.

Xie, H.; Qin, Z.; Li, G. Y.; and Juang, B.-H. 2021. Deep Learning Enabled Semantic Communication Systems. *IEEE Transactions on Signal Processing*, 69: 2663–2675.

Zeng, Y.; He, X.; Chen, X.; Tong, H.; Yang, Z.; Guo, Y.; and Hao, J. 2024. DMCE: Diffusion Model Channel Enhancer for Multi-User Semantic Communication Systems.

Zhao, Y.; Yue, Y.; Hou, S.; Cheng, B.; and Huang, Y. 2024. LaMoSC: Large Language Model-Driven Semantic Communication System for Visual Transmission. *IEEE Transactions on Cognitive Communications and Networking*.