Maximizing Channel Capacity in Semantic Communication: A Classifier-Based Mutual Information Estimation Approach

Xu Wang [1] **Di Wang** [1] **Zheng Shi** [1] **Guanghua Yang** [1]

Abstract

Semantic communication shifts the focus from traditional bit-level transmission to conveying meaning and context accurately. While recent methods use variational mutual information (MI) estimators to maximize channel capacity, they often suffer from high variance and unreliable estimates under limited sample conditions. To overcome these issues, we propose a novel MI estimation approach that trains a probabilistic classifier to distinguish true input-output signal pairs from randomly shuffled ones. This framework can improve the training stability and provide reliable guidance for training the semantic encoder. Experimental results on the text transmission task show that our model outperforms state-of-the-art end-to-end semantic communication systems and conventional source-channel coding schemes.

1. Introduction

In the rapidly evolving landscape of communication systems, the concept of semantic communication has emerged as a transformative paradigm, shifting the focus from purely syntactic information transmission to the conveyance of meaningful information (Qin et al., 2019; Dörner et al., 2017). Unlike traditional communication systems, which prioritize the accurate transfer of bits or symbols (Tse, 2005), semantic communication emphasizes the faithful transmission of context related to the task between the transmitter and receiver. Semantic communication can reduce the amount of data transmitted, making it highly efficient in bandwidth-limited environments such as IoT devices and mobile networks (Deng et al., 2024; Jankowski et al., 2020). Additionally, semantic communication is expected to perform well in challenging communication channels, particularly in low SNR environments (Hu et al., 2023).

In communication theory, the channel capacity defined through the mutual information (MI) between the input and output of a channel, which represents the maximum rate at which information can be reliably transmitted over a communication channel without errors. In the past decades, many advanced channel coding techniques such as Reed-Solomon (RS) codes (Reed & Solomon, 1960), Low-Density Parity-Check (LDPC) Codes (Gallager, 2003), and turbo codes (Heegard & Wicker, 2013), come closer to approach the Shannon limit. The end-to-end semantic communication system necessitates a rethinking of information-theoretic frameworks and the development of novel techniques to estimate and optimize the channel capacity.

Many early works treat the MI estimation problem as a density estimation problem and use non-parametric density estimators (Darbellay & Vajda, 1999; Beirlant et al., 1997; Paninski, 2003; Ross, 2014). However, these methods are susceptible to the curse of dimensionality and are highly sensitive to hyper-parameter choices. Recent research has shifted towards estimating variational bounds on MI, utilizing a critic function within a class of functions (parameterized with neural networks) to approximate the density ratio (Donsker & Varadhan, 1975; Barber & Agakov, 2003; Nguyen et al., 2008; 2010). These approaches has been utilized in (Xie et al., 2021; Fritschek et al., 2019) to improve the semantic encoder by maximizing the channel capacity and demonstrates better performance. However, such variational estimators may have large variance grow exponentially with the true MI (Song & Ermon, 2019), and cannot provide reliable estimation when the training samples are limited (McAllester & Stratos, 2020).

This work aims to contribute to the understanding and development of semantic communication systems by exploring more reliable channel capacity maximizing models. More specifically, we propose to approximate the channel capacity through a probabilistic classifier that can separate the joint signal pairs and the reshuffled random pairs. The numerical results on the text transmission task demonstrate the effectiveness of our model for both communication performance and training stability.

[1] School of Intelligent Systems Science and Engineering, Jinan University, China. Correspondence to: Di Wang <diwang@jnu.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

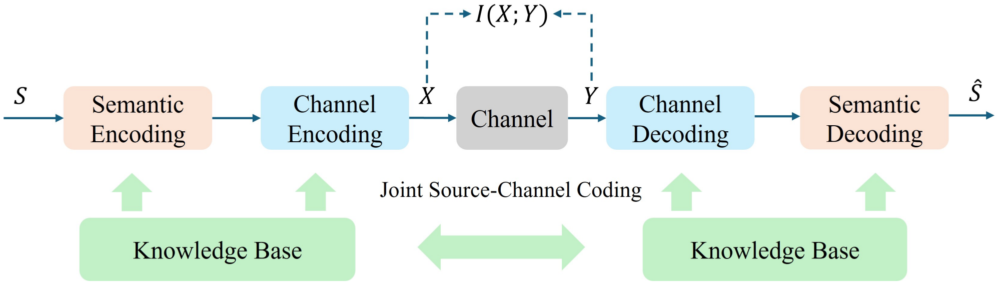


Figure 1. The framework of the proposed semantic communication with channel capacity maximization.

2. Problem Description

2.1. Notation

In this work, we use a sans serif capital letter X to denote a random vector/variable and $\mathcal{X}$ for its alphabet set. p_{X} denotes the distribution of X, which is a pdf if X is continuous. E denotes the expectation operation. To simplify notation, we will use the shorthand notation $[n]$ to denote the indexing set $\{1, \ldots, n\}$ for any positive integer n.

2.2. System Model

As shown in Figure. 1, we consider a task-oriented semantic communication system with a transmitter, a stochastic physical channel, and a receiver. The problem is then formulated as an end-to-end optimization by transmitting only the most relevant information needed for the task at hand, reducing redundancy and improving effectiveness in wireless communication. The transmitter $f_\theta(\cdot)$ consists of a semantic encoder and a channel encoder, which extracts key semantic features as

$$\mathsf{X} = f_\theta(\mathsf{S}) \tag{1}$$

from the source message $\mathsf{S} \in \mathcal{S}$, and θ denotes the parameters of the transmitter. Then the noisy channel delivers a perturbed version Y to the receiver. More precisely, we assume that the received noisy signal is

$$\mathsf{Y} = \mathsf{h}\mathsf{X} + \mathsf{Z}, \tag{2}$$

where Z is zero-mean Additive White Gaussian Noise (AWGN) with $\mathsf{Z} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and h denotes the Rayleigh fading channel with $\mathsf{h} \sim \mathcal{CN}(0, 1)$. Similarly, the receiver g_ψ includes channel decoder and semantic decoder, acts on Y and reconstructs the signal as

$$\hat{\mathsf{S}} = g_\psi(\mathsf{Y}). \tag{3}$$

For text transmission, cross-entropy between the estimated probabilities and the true probability of the message can be used as the semantic loss

$$\mathcal{L}_{\text{ce}}(\mathsf{S}, \hat{\mathsf{S}}) := -\sum_{s \in \mathsf{S}} p(s) \log g_\psi(\mathsf{h} f_\theta(s) + \mathsf{Z}) \tag{4}$$

that captures task success for transmitting a sentence S.

2.3. Problem Formulation

The optimal transmission rate in a communication system is fundamentally governed by Shannon's channel capacity, which is defined through the MI between the input and the output of a noisy channel $P_{\mathsf{Y}|\mathsf{X}}$. Under an average power constraint P, the capacity is defined as

$$C := \max_{P_{\mathsf{X}} : E[\mathsf{X}^2] \le P} I(\mathsf{X}; \mathsf{Y}), \tag{5}$$

where the supremum is taken over all possible choices of P_{X}. MI is defined as KL divergence between the joint distribution and marginal distributions:

$$\begin{aligned} I(\mathsf{X}; \mathsf{Y}) &= D(P_{\mathsf{XY}} \| P_{\mathsf{X}} P_{\mathsf{Y}}) \\ &= E\Big[\log \overbrace{\frac{dP_{\mathsf{XY}}(\mathsf{X}, \mathsf{Y})}{dP_{\mathsf{X}}(\mathsf{X}) dP_{\mathsf{Y}}(\mathsf{Y})}}^{r(\mathsf{X},\mathsf{Y}):=}\Big], \end{aligned} \tag{6}$$

where $r(\mathsf{X}, \mathsf{Y})$ denotes the density ratio.

Therefore, maximizing the MI $I(\mathsf{X}; \mathsf{Y})$ ensures that the encoding strategy not only adheres to physical transmission constraints but also achieves information-theoretic optimality in semantic communication. However, MI estimation with finite samples in high-dimensional space is a challenge problem. Existing works (Xie et al., 2021; Fritschek et al., 2019) utilize the following Donsker-Varadhan (DV) variational lower bound to maximize the channel capacity and demonstrate certain improvement in wireless communication.

Lemma 2.1 (MINE (Belghazi et al., 2018))**.** *For two random variables* X *and* Y*, we have*

$$I(\mathsf{X}; \mathsf{Y}) = \sup_{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}} E[f(\mathsf{X}, \mathsf{Y})] - \log E\left[e^{f(\mathsf{X}', \mathsf{Y}')}\right] \tag{7}$$

where $(\mathsf{X}, \mathsf{Y}) \sim P_{\mathsf{X},\mathsf{Y}}$ *and* $(\mathsf{X}', \mathsf{Y}') \sim P_{\mathsf{X}} P_{\mathsf{Y}}$*. The optimal critic* $f^*(x, y) = \log r(x, y) + c$*, where c is a constant.*

The main drawback of this approach lies in the difficulty of training the parameterized neural network f. The works of (McAllester & Stratos, 2020; Song & Ermon, 2019) show that such training often suffers from high variance and requires a large number of samples (exponential in the value of MI) to achieve reliable performance. Therefore, it is necessary to explore other solutions to make the training more efficient and reliable for maximizing semantic channel capacity.

3. Classifier-based Mutual Information Neural Estimation

In this section, we will introduce the classifier-based MI estimator (Tsai et al., 2020) and integrate it with semantic communication systems to collaboratively optimize the semantic encoder.

The joint sample pairs $\{(\mathsf{X}_i, \mathsf{Y}_i)\}_{i\in[n]} \sim P_{\mathsf{X},\mathsf{Y}}$, i.e., the paired transmitted signal and received signal, are labeled as *positive samples*. We can also utilize the reshuffling trick (Belghazi et al., 2018) to generate the samples of product of marginals $\{(\mathsf{X}'_j, \mathsf{Y}'_j)\}_{j\in[m]} \sim P_{\mathsf{X}} P_{\mathsf{Y}}$, which is referred as *negative samples*. More specifically, the received signals are rearranged and randomly paired with the transmitted signal sequence. To precisely formulate the model, we introduce an auxiliary random variable $\mathsf{C} \in \{0, 1\}$ to denote the class label, i.e., assigning 1 to the positive and 0 to the negative samples.

By combining the positive and negative samples, we can define mixed random variables $\tilde{\mathsf{X}}, \tilde{\mathsf{Y}}$ subject to the following distribution:

$$p_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}} = p_{\mathsf{C}}(1) \underbrace{p_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}|\mathsf{C}=1}}_{:=p_{\mathsf{XY}}} + p_{\mathsf{C}}(0) \underbrace{p_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}|\mathsf{C}=0}}_{:=p_{\mathsf{X}}p_{\mathsf{Y}}} \tag{8}$$

The motivation is to approximate the ratio r by training a binary classifier to distinguish between the positive ($\mathsf{C} = 1$) and negative samples ($\mathsf{C} = 0$). Then by Bayes's theorem, the density ratio r in (6) can be written as:

$$r(x, y) = \frac{p_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}|\mathsf{C}}(x, y|1)}{p_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}|\mathsf{C}}(x, y|0)} = \frac{p_{\mathsf{C}}(0)}{p_{\mathsf{C}}(1)} \cdot \frac{p_{\mathsf{C}|\tilde{\mathsf{X}}\tilde{\mathsf{Y}}}(1|x, y)}{p_{\mathsf{C}|\tilde{\mathsf{X}}\tilde{\mathsf{Y}}}(0|x, y)} \tag{9}$$

Given the mixture dataset $\{(\tilde{\mathsf{X}}_i, \tilde{\mathsf{Y}}_i)\}_{i\in[m+n]} \sim P_{\tilde{\mathsf{X}}\tilde{\mathsf{Y}}}$[1], the posterior distribution $p_{\mathsf{C}|\tilde{\mathsf{X}}\tilde{\mathsf{Y}}}(1|x, y)$ can be approximated using the neural network p_ϕ by minimizing the binary cross-entropy loss:

$$\begin{aligned}\mathcal{L} &:= -E\left[\log p_\phi(\mathsf{C}|\tilde{\mathsf{X}}, \tilde{\mathsf{Y}})\right] \\ &\approx -\frac{1}{m+n} \sum_{i\in[m+n]} \mathsf{C}_i \log p_\phi(\mathsf{C}_i|\tilde{\mathsf{X}}_i, \tilde{\mathsf{Y}}_i) + \\ &\quad (1 - \mathsf{C}_i) \log(1 - p_\phi(\mathsf{C}_i|\tilde{\mathsf{X}}_i, \tilde{\mathsf{Y}}_i))\end{aligned} \tag{10}$$

The classifier-based MI estimator has demonstrated superior training stability (Hjelm et al., 2019) and sample complexity (Mukherjee et al., 2020) compared to variational bounds-based approaches. Additionally, to alleviate the over-confidence issue in the trained neural network, we further apply the label smoothing technique (Szegedy et al., 2016) as suggested in (Wang et al., 2021). We define the soft labels C' as

$$\mathsf{C}' := (1 - \alpha) \cdot \mathsf{C} + \alpha \cdot p_{\mathsf{U}}, \tag{11}$$

where $p_{\mathsf{U}}(\cdot) := \frac{1}{2}$ is the uniform distribution, and α is the smoothing factor. Therefore, the label-smoothed cross-entropy loss can be approximated as :

$$\begin{aligned}\mathcal{L}_{\text{mi}} &\approx -\frac{1}{m+n} \sum_{i\in[m+n]} \mathsf{C}'_i \log p_\phi(\mathsf{C}_i|\tilde{\mathsf{X}}_i, \tilde{\mathsf{Y}}_i) + \\ &\quad (1 - \mathsf{C}'_i) \log(1 - p_\phi(\mathsf{C}_i|\tilde{\mathsf{X}}_i, \tilde{\mathsf{Y}}_i))\end{aligned} \tag{12}$$

Therefore, the total loss function of the proposed model is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}}(\mathsf{S}, \hat{\mathsf{S}}; \theta, \psi) + \lambda \mathcal{L}_{\text{mi}}(\mathsf{X}, \mathsf{Y}; \theta, \phi), \tag{13}$$

where λ is the importance weight of the MI loss term. As we need to train an additional MI module, the training process is divided into two phases corresponding to MI optimizing and communication network training. The details are shown in Algorithm 1.

4. Experiments

In this section, we consider the text transmission task and evaluate our proposed classifier-based framework on the open-source dataset based on the proceedings of the European Parliament (Koehn, 2005). The dataset consists of approximately 2 million sentences, and split into training data and testing data. We compare our model with exiting advanced models including the variational approach (DeepSC) (Xie et al., 2021), the joint source-channel coding model (JSCC) (Fritschek et al., 2019), and traditional method with Huffman and RS coding.

To ensure a fair comparison, we adopt Transformer-based architectures for both the semantic encoder and decoder, following the designs outlined in (Xie et al., 2021). The probabilistic classifier consists of two sequential convolutional layers (32 and 64 channels, respectively), each followed by ReLU activation and max pooling, and a final

[1] To create a balanced classification problem, the number of negative samples is typically set equal to the number of positive samples, i.e., $p_{\mathsf{C}}(0) = p_{\mathsf{C}}(1) = \frac{1}{2}$ which implies $m = n$.

Algorithm 1 Iterative training process of the proposed MI-aid semantic communication framework.

```
Input: Knowledge base $\mathcal{S} := \{\mathsf{S}_i\}_{i\in[n]}$.
Initialization: the parameters set of semantic encoder $f_\theta$, semantic decoder $g_\psi$, probabilistic classifier $h_\phi$;
repeat
    Randomly draw a batch of source messages $\mathcal{S}' \subset \mathcal{S}$
    ① Training the MI estimator module
    Freeze semantic encoder $f_\theta$;
    Transmit the signal over the channel:
    $\mathsf{X}_i \leftarrow f_\theta(\mathsf{S}_i), \mathsf{Y}_i \leftarrow \mathsf{h}\mathsf{X}_i + \mathsf{Z}, \forall \mathsf{S}_i \in \mathcal{S}'$;
    Construct the mixture dataset and train $h_\phi$ with loss (12);
    ② Training the semantic encoder and decoder
    Freeze probabilistic classifier $h_\phi$;
    Train $f_\theta, g_\psi$ jointly with loss (13)
until Convergence
```

fully-connected layer. The smoothing factor $\alpha = 0.01$ in (11). For all models, the batch size is 256 and the learning rate is set as 10^{-4}.

We use the BLEU score (Papineni et al., 2002) to measure semantic differences between the transmitted and recovered messages. The BLEU score between S and $\hat{\mathsf{S}}$ is defined as

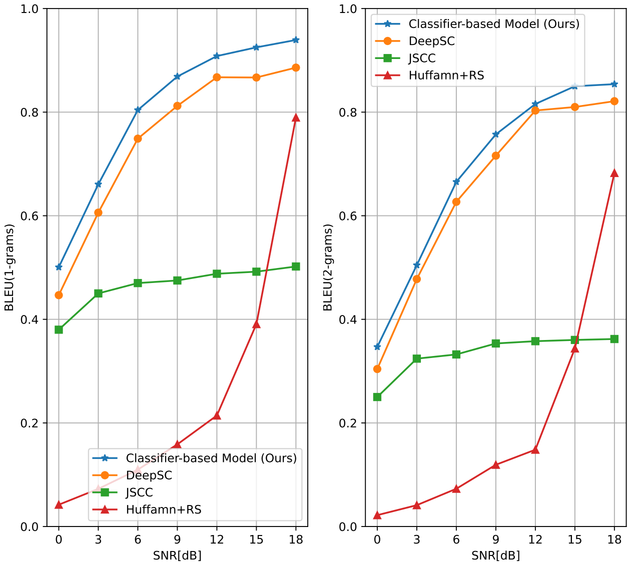
$$\log \mathrm{BLEU} = [1 - l_{\mathsf{S}}/l_{\hat{\mathsf{S}}}]^- + \sum\nolimits_{n=1}^{N} u_n \log f_n,$$

where $[x]^- = \min\{x, 0\}$, l_{S} and $l_{\hat{\mathsf{S}}}$ represent the lengths of the input message and the output, respectively. u_n is the weights of n-grams, and f_n indicates the n-grams score:

$$f_n = \frac{\sum_k \min\left\{C_k(\hat{\mathsf{S}}), C_k(\mathsf{S})\right\}}{\sum_k C_k(\hat{\mathsf{S}})},$$

where $C_k(\cdot)$ is the function counting the frequency of the k-th element in n-grams.

The communication performance results in Figure.2 show that the Classifier-based Model (Ours) consistently outperforms DeepSC, JSCC, and Huffman+RS across different SNR levels. It achieves the highest semantic accuracy, particularly at low to mid SNRs, demonstrating strong robustness to channel noise. While DeepSC performs well, it slightly lags behind. JSCC shows limited sensitivity to SNR changes, and Huffman+RS only improves significantly at high SNRs. These results highlight the superior adaptability and reliability of the proposed model in Rayleigh fading environments. As shown in Figure.3, with SNR = 8 dB, our method also achieves faster convergence and more stable mutual information estimation than DeepSC, enabling more accurate and reliable guidance for semantic encoder training.



(a) 1-grams BLEU (b) 2-grams BLEU

Figure 2. Performance comparison between the proposed model and existing algorithms with a Rayleigh fading channel.

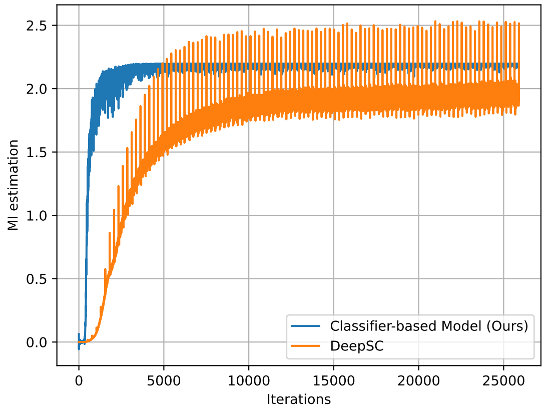


Figure 3. MI estimation during the training. SNR=8 dB.

5. Conclusion

This work explores a novel classifier-based channel capacity maximization approach for semantic communication to enhance system performance. Extensive experiments demonstrate that our method consistently outperforms existing approaches, achieving not only higher communication accuracy, but also faster convergence and more stable MI estimation during training.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grants 62171200, 62171201, and 62261160650, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010900, in part by the Major Talent Program of Guangdong Provincial under Grant 2019QN01S103.

Impact Statement

This work contributes to the advancement of machine learning and semantic communication by proposing a method to improve mutual information estimation and channel capacity in communication systems. Enhanced semantic communication could enable more efficient and robust information transfer in bandwidth-constrained or noisy environments, with applications in wireless networks.

References

Barber, D. and Agakov, F. V. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

Beirlant, J., Dudewicz, E. J., Györfi, L., and Van der Meulen, E. C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *Proceedings of Machine Learning Research*, volume 80, pp. 531–540. PMLR, 2018.

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

Deng, D., Wang, C., Xu, L., Jiang, F., Guo, K., Zhang, Z., Wang, W., Quek, T. Q., and Zhang, P. Semantic communication empowered ntn for iot: Benefits and challenges. *IEEE Network*, 2024.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28 (1):1–47, 1975.

Dörner, S., Cammerer, S., Hoydis, J., and Ten Brink, S. Deep learning based communication over the air. *IEEE Journal of Selected Topics in Signal Processing*, 12(1): 132–143, 2017.

Fritschek, R., Schaefer, R. F., and Wunder, G. Deep learning for channel coding via neural mutual information estimation. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5. IEEE, 2019.

Gallager, R. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28, 2003.

Heegard, C. and Wicker, S. B. *Turbo coding*, volume 476. Springer Science & Business Media, 2013.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Hu, Q., Zhang, G., Qin, Z., Cai, Y., Yu, G., and Li, G. Y. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Trans. Wireless Commun.*, 22 (12):8707–8722, 2023.

Jankowski, M., Gündüz, D., and Mikolajczyk, K. Wireless image retrieval at the edge. *IEEE J. Sel. Areas Commun.*, 39(1):89–100, 2020.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pp. 79–86, 2005.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884, 2020.

Mukherjee, S., Asnani, H., and Kannan, S. CCMI: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pp. 1083–1093. PMLR, 2020.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in neural information processing systems*, pp. 1089–1096, 2008.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qin, Z., Ye, H., Li, G. Y., and Juang, B.-H. F. Deep learning in physical layer communications. *IEEE Wireless Communications*, 26(2):93–99, 2019.

Reed, I. S. and Solomon, G. Polynomial codes over certain finite fields. *Journal of the society for industrial and applied mathematics*, 8(2):300–304, 1960.

Ross, B. C. Mutual information between discrete and continuous data sets. *PloS one*, 9(2), 2014.

Song, J. and Ermon, S. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tsai, Y.-H. H., Zhao, H., Yamada, M., Morency, L.-P., and Salakhutdinov, R. Neural methods for point-wise dependency estimation. In *Advances in Neural Information Processing Systems*, 2020.

Tse, D. Fundamentals of wireless communication. *Cambridge University Press google schola*, 2:281–302, 2005.

Wang, X., Al-Bashabsheh, A., Zhao, C., and Chan, C. Adaptive label smoothing for classifier-based mutual information neural estimation. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1035–1040. IEEE, 2021.

Xie, H., Qin, Z., Li, G. Y., and Juang, B.-H. Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69:2663–2675, 2021.