AUTOENCODERS TRAINED WITH RELEVANT INFORMATION: BLENDING SHANNON AND WIENER'S PERSPECTIVES

Shujian Yu, Matthew Emigh, Eder Santana, José C. Príncipe

Department of Electrical and Computer Engineering, University of Florida, FL 32611

ABSTRACT

It is almost seventy years after the publication of Claude Shannon's "A Mathematical Theory of Communication" [1] and Norbert Wiener's "Extrapolation, Interpolation and Smoothing of Stationary Time Series" [2]. The pioneering works of Shannon and Wiener lay the foundation of communication, data storage, control, and other information technologies. This paper briefly reviews Shannon and Wiener's perspectives on the problem of message transmission over noisy channel and also experimentally evaluates the feasibility of integrating these two perspectives to train autoencoders close to the information limit. To this end, the principle of relevant information (PRI) is used and validated to optimally encode input imagery in the presence of noise.

Keywords—Autoencoder, Principle of Relevant Information (PRI), Message transmission.

I. INTRODUCTION

Claude Shannon's "A mathematical theory of communication" [1] published in 1948 marks the birth of information theory. A unifying theory with profound intersections between Probability, Statistics and Computer Science, Shannon's information theory provides the foundation of Digital Communication [3]. The Mathematical Theory of Communication led to the architecture for the design of modern digital communication systems as shown in Fig.1.

This model, also known as the "Shannon paradigm", is general and applies to a great variety of communication scenarios. Specifically, a source encoder allows one to represent the information source more compactly by eliminating redundancy, while a channel encoder adds redundancy to protect the transmitted signal against transmission errors. Source and channel decoders are converse to source and channel encoders. Obviously, there is duality between "source coding" and "channel coding", as the former tends to reduce the data rate while the latter raises it [4].

Almost at the same time, Norbert Wiener published his famous monograph, "Extrapolation, interpolation and smoothing of stationary time series" [2]. According to Shannon's perspective, for any given degree of noise contamination of a communication channel, with an appropriate coding-decoding scheme, it is possible to communicate discrete digital information nearly error-free up to a computable maximum rate through the channel[1]. Wiener, however, approached this problem from a different point of view by adaptively filtering the signal, i.e. distinguishing information from noise via error minimization (normally via mean square errors (MSE)).

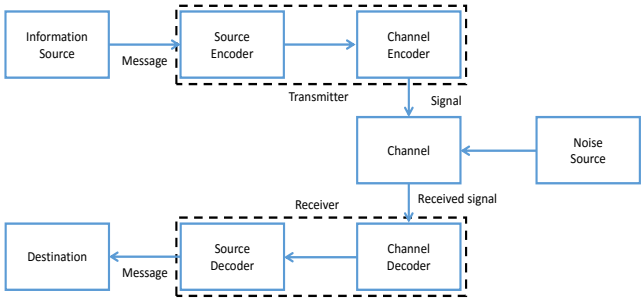


Fig. 1: Block diagram of digital communication system.

After seventy years' development, it is now appropriate to commemorate and comment on the great perspectives of these two giants. In this paper, the perspectives of Shannon and Wiener on message transmission over a noisy channel are revisited. Furthermore, the similarities and connections between autoencoder[2] [5] (representative of Wiener's perspective) and digital communication system (representative of Shannon's perspective) will also be discussed. Following this, we present a novel autoencoder under the information theoretic learning (ITL) [6] framework, which can automatically learn an regularity-oriented hidden layer vector representation given a distribution prior[3]. A simple simulation on image transmission over a memoryless Gaussian channel is conducted to demonstrate the effectiveness of the proposed autoencoder and also to experimentally clarify the two perspectives.

I-A. On message transmission: Shannon versus Wiener

Suppose we are given an information source which generates a sequence of symbols $\underline{X} = \{X_1, X_2, ..., X_T\}$ taking values in a finite alphabet $\mathcal{X}$, let us consider the problem of transmitting a realization of sequence $\underline{x} = \{x_1, x_2, ..., x_T\}$.

Recalling Shannon's perspective (see Fig.1), the source encoder seeks a mapping $f : \mathcal{X}^T \to \{0,1\}^*$ which associates to any possible information sequence in $\mathcal{X}^T$ a string in a reference alphabet which we shall assume to be binary $\{0,1\}$. Suppose the resulting sequence of bits produced by f is divided into blocks $m_1, m_2, m_3, ...$ of length M. Then the

[1]This is exactly the famous noisy-channel coding theorem (or Shannon's theorem).

[2]The reasons behind the selection of autoencoder is elaborated in I-B.

[3]We expect such regular map in the bottleneck layer of autoencoder can play a similar role as the "constellation" map in digital communication system.

channel encoder maps each M-bit message $m \in \{0,1\}^M$ to a codeword $\underline{x}(m) \in \{0,1\}^N$, with $N \geq M$. During channel transmission, $\underline{x}$ is corrupted to $\underline{y} \in \mathcal{Y}^N$ with probability $P(\underline{y}|\underline{x}) = \prod_{i=1}^{N} P(y_i|x_i)$[4]. The output alphabet $\mathcal{Y}$ depends on the channel. The channel decoder is a mapping from $\mathcal{Y}^N$ to $\{0,1\}^M$ which takes the received message $\underline{y} \in \mathcal{Y}^N$ and maps it to one of the possible original messages $m' \in \{0,1\}^M$. Finally, the source information is recovered with the source decoder [7].

Wiener examines this problem from the point of view of filtering or signal inference [8]. According to Wiener, the channel influences sequence $\underline{x}$ according to:

$$\underline{y} = f(\underline{x}) + n \tag{1}$$

where f refers to the "influence" function (maybe invertible) that characterizes the channel's properties, and n is unknown noise term. Wiener was interested in finding an operator g, for which the error term $\|\underline{x} - g(\underline{y})\|_2^2$ can be minimized.

I-B. Why autoencoder

The central perspective of Wiener lies in the minimization of information distortion (or error) during message transmission. For this reason, the general idea of the autoencoder matches Wiener's perspective well, as it always aims to transform input into output with the least possible amount of distortion.

Although vast of works have been done on unsupervised learning, three main advantages set the autoencoder and its variants apart from their counterparts: 1) their expressive representations for complex data; 2) their flexibilities for modification: we can just modify the loss functions in different ways to suit our tasks [9]; and 3) it follows exactly an encoder-decoder channel learning framework[5], which has the same block diagram as standard communication systems on a metaphorical level (see Fig.2) [10]. Also, it is worth noting that the Wiener-proposed cost function only quantifies second order statistics; therefore it is unable to fully optimize the extraction of signal from noise, and must be substituted.

II. AUTOENCODER REGULARIZED WITH PRINCIPLE OF RELEVANT INFORMATION (PRI)

To evaluate the feasibility and effectiveness of Wiener's perspectives on message transmission over a noisy channel, we elect to use an autoencoder (Fig.2) as a powerful implementation of Wiener's idea but with an information theoretic cost function to train the system. This newly designed autoencoder is able to learn powerful representations in the bottleneck layer given a distribution prior, which plays a similar role as the "constellation" map in digital communication system.

In this section, we start with a brief introduction to the elements of Renyi's entropy as well as associated information quantities. Based on this, the objective function for the "principle of relevant information" (PRI) is presented and formulated under the ITL framework. After that, we give a short review of our previously proposed ITL-autoencoder [11] and also present an alternative one regularized using the PRI.

[4]Throughout this paper, we only consider memoryless channels, in which the noise acts independently on each bit of the input.

[5]Interested readers can refer to Appendix C of [10] for more details.

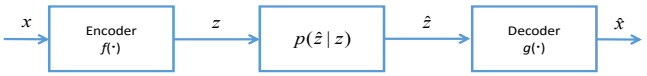


Fig. 2: Block diagram of encoder-decoder channel learning scheme.

II-A. Elements of Renyi's entropy and the principle of relevant information (PRI)

In information theory, a natural extension of the well known Shannon's entropy is Renyi's α-entropy [6]. For a random variable X with probability density function (PDF) $f(x)$ in a finite set $\mathcal{X}$, the α-entropy $H_\alpha(X)$ is defined as:

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f^\alpha(x) dx. \tag{2}$$

Likewise, extensions for the relative entropy also exist; a modified version of Renyi's definition of α-relative entropy between random variables with PDFs f and g is given by:

$$D_\alpha(f\|g) = \log \frac{\left(\int g^{\alpha-1} f\right)^{\frac{1}{1-\alpha}} \left(\int g^\alpha\right)^{\frac{1}{\alpha}}}{\left(\int f^\alpha\right)^{\frac{1}{\alpha(1-\alpha)}}}. \tag{3}$$

The limiting case of (2) and (3) for $\alpha \to 1$ is Shannon's entropy and Shannon's relative entropy (or equivalently Kullback-Leibler (KL) divergence), respectively. It also turns out that for the case of $\alpha = 2$, the above quantities can be expressed, under some restrictions, as functions of inner products between PDFs. In particular, the quadratic entropy of f and cross-entropy between f and g, can be expressed respectively as:

$$H_2(f) = -\log \int_{\mathcal{X}} f^2(x) dx \tag{4}$$

$$H_2(f;g) = -\log \int_{\mathcal{X}} f(x) g(x) dx \tag{5}$$

and the associated relative entropy of order 2 is called the Cauchy-Schwarz (CS) divergence and is defined as follows:

$$D_{CS}(f\|g) = -\frac{1}{2} \log \frac{\left(\int fg\right)^2}{\left(\int f^2\right)\left(\int g^2\right)} \tag{6}$$

Suppose we are given a random variable X with a known prior PDF g, from which we want to find a description in terms of a PDF f with reduced entropy, that is, a variable Y that captures the underlying structure of X. The principle of relevant information (PRI) [12] casts this problem as a trade-off between the entropy $H_2(f)$ of Y and its descriptive power about X in terms of their relative entropy $D_{CS}(f\|g)$. Therefore, for a fixed PDF g, the objective of the PRI is given by[6]:

$$\mathrm{J}(f) = \arg\min_f \left[H_2(f) + \lambda D_{CS}(f\|g)\right]. \tag{7}$$

[6]The minimization of relative entropy guarantees a powerful description to the given PDF, while the minimization of entropy plays the role of finding such structures (or regularities) [12], [13].

II-B. PRI formulation under ITL framework

It is worth noting that, as it is often the case, the only available information about g is encoded in sample set $X \in (\mathbb{R}^p)^N$, where p is the dimensionality of samples and N is the cardinality of X. To handle this, a tractable solution was proposed in [12] under an ITL framework. This method combines Parzen density estimation with a gradient descent procedure that minimizes (7) to match the desired density g. The optimization problem becomes[7]:

$$\mathrm{J}(f) = \arg\min_{Y} \; [\hat{H}_2(Y) + \lambda \hat{D}_{CS}(Y \| X)] \tag{8}$$

where $Y \in (\mathbb{R}^p)^M$ is a set of p-dimensional points with cardinality M. Equation (8) is equivalent to[8]:

$$\mathrm{J}(f) = \arg\min_{Y} \; [(1-\lambda)\hat{H}_2(Y) + 2\lambda \hat{H}_2(Y \| X)] \tag{9}$$

Using Parzen density estimation with Gaussian kernel G_σ, the cost function of PRI can finally be formulated as:

$$\begin{aligned} &-(1-\lambda)\log\Big(\frac{1}{M^2}\sum_{i,j=1}^{M} G_{\sigma\sqrt{2}}(y_i, y_j)\Big) \\ &-\lambda \log \frac{\big(\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(y_i, x_j)\big)^2}{\big(\frac{1}{N^2}\sum_{i,j=1}^{N} G_{\sigma\sqrt{2}}(x_i, x_j)\big)} \end{aligned} \tag{10}$$

where $x_i \in X$ and $y_i \in Y$.

II-C. PRI-regularized autoencoder

The ITL autoencoder was initially introduced in [11] to provide an alternative variational regularization to the autoencoder architecture, aiming to force the hidden code vector learned by the (deep) architecture to be as close as possible to an imposed prior. Specifically, the architecture of the ITL-autoencoder (shown in Fig.3) consists of a 4-tuple $AE = \{E, D, L, R\}$, where E and D are the encoder and the decoder functions, L represents the reconstruction cost function (MSE in Wiener sense) that measures the difference between original data samples s and their respective reconstructions $\hat{s} = D(E(s))$ and R denotes the functional regularization penalized on the output of encoder E and given prior[9].

The first generation of ITL-autoencoder uses an ITL divergence descriptor (*e.g.*, (5)) as the regularization function R, rather than the classical parametric, closed-form KL divergence or training a new discriminator network [14], [15]. Following the work in [11], we present an alternative ITL-autoencoder, which is also regularized with ITL descriptors. However, differently from [11], we use the PRI regularization to learn regularity-oriented hidden vector representations within a given prior. Given input data set S with cardinality K, the cost function can be represented as:

$$\mathrm{J} = L(S, D(E(S))) + \alpha \hat{H}_2(E(S)) + \beta \hat{D}_{CS}(E(S) \| P) \tag{11}$$

[7] $\hat{H}_2(Y)$, $\hat{H}_2(Y\|X)$ and $\hat{D}_{CS}(Y\|X)$ in (8) and (9) represent Parzen density estimator to (4), (5) and (6).

[8] See Chapter 3 of [12] for more details.

[9] Examples of prior include Gaussian distribution, mixture of Gaussians distribution as well as swiss roll distribution.

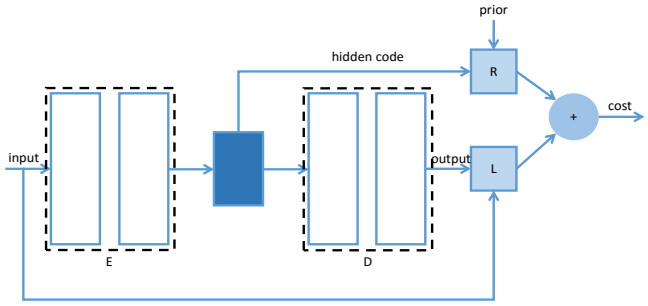


Fig. 3: Block diagram of ITL-autoencoder in [11].

where α and β are tuning parameters, $P = \{p_i\}_{i=1,2,,K}$ represents samples randomly generated from a prior distribution.

Discussion on parameters α and β.
It is worth noting that, the ratio of β/α plays the same role[10] as the trade-off parameter λ in (7). For $\beta/\alpha \to 0$, all the points in $E(S)$ will collapse to a single point, which in the limit case becomes independent of the target sample P. The other extreme case is when $\beta/\alpha \to \infty$; the system is already in equilibrium, *i.e.*, $E(S)$ is initialized by the locations provided by the sample P and will not move away. Interesting cases arise when $\beta/\alpha = 1$ or $\beta/\alpha \approx 2$. For instance, it has been shown that the case $\beta/\alpha = 1$ corresponds to the Gaussian mean shift algorithm [16], and the case $\beta/\alpha \approx 2$ approximates principal curves extraction. We illustrate this phenomenon in Fig.4.

III. EXPERIMENTS ON IMAGERY TRANSMISSION

We conducted a simple imagery transmission experiment using both the proposed autoencoder regularized with PRI (PRI-AE for short) and a digital communication system. More specifically, we randomly select 100 images from the MNIST testing set and transmit them with either of the two mechanisms[11]. Noise with the same Signal-to-Noise Ratios (SNRs), ranging from 8dB to 18dB, are added to their respective channels, *i.e.*, bottleneck layer for PRI-AE and the physical channel for communication system. The averaged MSE between transmitted imagery and received imagery are computed for the purpose of performance comparison. In this paper, we only consider the real discrete-time memoryless additive white Gaussian noise (AWGN) channel.

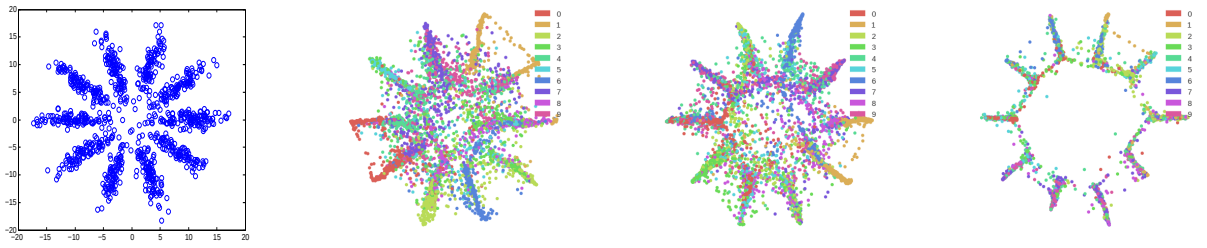
III-A. Experimental setup

For the Shannon's approach, we study the encoding-decoding mechanisms of M-ary ($M = 16$ or 64) phase shift keying (PSK) and quadrature amplitude modulation (QAM) constellations using Bose, Chaudhuri, and Hocquenghem (BCH) codes[12] [17]. These configurations are used in many wireless standards [18].

[10] Note that, the discussion on the effects of different values of β/α only limited to the interplay between the second and third terms in (11). Similar effects can be expected when incorporated with MSE cost function.

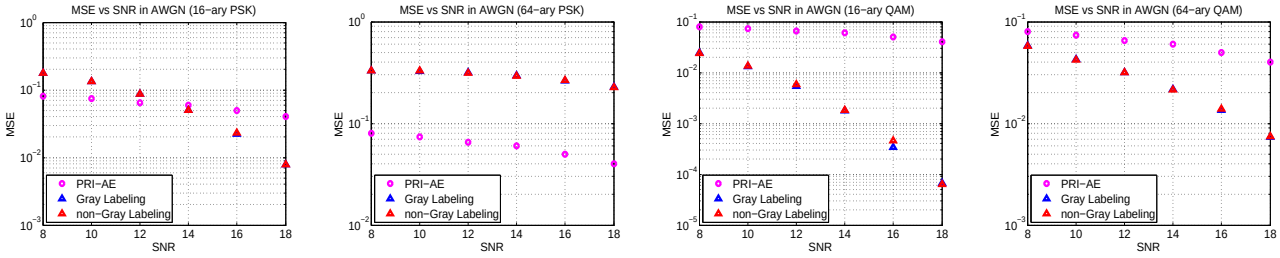
[11] We repeated same scenario with 50 independent trails, no special circumstances occurred.

[12] Both the Gray labelings and non-Gray labelings are considered for different constellations. Besides, we only use the basic $[7, 4]$ BCH code for simplicity.



(a) (b) (c) (d)

Fig. 4: Hidden code visualization using (a) the mixture of 10 Gaussians prior; (b) ITL-autoencoder [11] with CS divergence; PRI-AE with (c) $\beta/\alpha = 10$ and (d) $\beta/\alpha = 2$ for MNIST database.



(a) 16-ary PSK (b) 64-ary PSK (c) 16-ary QAM (d) 64-ary QAM

Fig. 5: MSE comparison between PRI-AE and standard communication system with (a) 16-ary PSK modulation; (b) 64-ary PSK modulation; (c) 16-ary QAM modulation and (d) 64-ary QAM modulation.

For PRI-AE, the bottleneck layer is restricted to have only 2-dimensional latent codes which corresponds to the role of "constellation" maps in digital communication systems. To avoid hyper-parameter tuning, we also constrained our encoder and decoders, E and D, to have the same architecture as those used in [11], [14], *i.e.*, each network is a two hidden layer fully connected network with 1000 hidden neurons. Since MNIST has 10 classes, we use as the distribution prior a mixture of 10 Gaussians. The kernel size for PRI is fixed to be 10, as recommended in [11].

We implement the ITL-autoencoder [11] and PRI-AE both with the Keras library. Companion source code is available from the project homepage. During training, we configure our model using MSE loss and the Adadelta optimizer[13]. We separate 10000 samples from MNIST training set for validation. The training is iterated for 100 epochs, which, it has been observed, to be sufficient to reliably converge. In each epoch, we use batches of 1000 samples for a reliable PDF estimation.

III-B. Experimental results

The imagery transmission results for four modulation scenarios are demonstrated in Fig.5. Generally, Shannon's perspective seems to be better, but the autoencoder is less sensitive to noise levels and even does better for lower SNRs for M-ary PSK modulations. Meanwhile, it is interesting to find that 16-ary PSK modulation produce the closest performance to PRI-AE. This is not surprising, since the "constellation" map of 16-ary PSK modulation looks most similar to our learned regularity map as shown in Fig.4(d), both of which use the angle discrepancy to handle "channel" noises.

[13]We use the default parameters in Keras library.

IV. CONCLUSIONS

In this paper, we reviewed Shannon and Wiener's perspectives on message transmission over a noisy channel and also experimentally evaluated the feasibility of integrating these two perspectives to train autoencoders close to the information limit. To this end, the principle of relevant information (PRI) is used and validated to optimally encode input imagery in the presence of noise.

For PRI-AE, we focused exclusively on the MNIST database. Although we tested the algorithm on examples that were not seen during training, we have not yet explored attempts to transmit images drawn from a different domain. We expect that performance would degrade significantly, since the primary reason that PRI-AE performs well - that it is able to exploit regularities of the domain data - no longer applies. This problem will not happen using Shannon's approach, which is independent of information source.

Nevertheless, the advantage of using a learning scheme combined with Wiener's perspective is that we can easily tailor our (deep) network architecture to perform optimally on a specific database, as we have done on MNIST. Because Shannon's approach requires knowledge of the noise and data statistics for optimal transmission, we have shown that Wiener can outperform Shannon under low SNRs in some scenarios.

Finally, it is worth nothing that, this paper is a start (or tentative) work on the possibility of approaching Shannon limit using learning schemes combined with Wiener's perspective. Quantitative measurements of information flow through PRI-AE as well as its applicability over other data domains and other channels remains for future works.

V. REFERENCES

[1] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT press Cambridge, 1949, vol. 2.

[3] S. Verdu, "Fifty years of shannon theory," *IEEE Transactions on information theory*, vol. 44, no. 6, pp. 2057–2078, 1998.

[4] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[5] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[6] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[7] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[8] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 146–181, 1974.

[9] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures." *ICML unsupervised and transfer learning*, vol. 27, no. 37-50, p. 1, 2012.

[10] L. G. S. Giraldo, "Reproducing kernel hilbert space methods for information theoretic learning," Ph.D. dissertation, University of Florida, 2012.

[11] E. Santana, M. Emigh, and J. C. Principe, "Information theoretic-learning auto-encoder," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 3296–3301.

[12] S. M. Rao, "Unsupervised learning: an information theoretic framework," Ph.D. dissertation, University of Florida, 2008.

[13] J. W. FISHER III, "Nonlinear extensions to the minimum average correlation energy filter," Ph.D. dissertation, University of Flordia, 1997.

[14] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[16] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[17] R. E. Blahut, *Algebraic codes for data transmission*. Cambridge university press, 2003.

[18] B. Sklar, *Digital communications*. Prentice Hall NJ, 2001, vol. 2.