Autoencoder based image quality metric for modelling semantic noise in semantic communications

Prabath Samarathunga, Thanuj Fernando,
Vishnu Gowrisetty, Thisarani Atulugama,
and Anil Fernando

Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

E-mail: Anil.Fernando@strath.ac.uk

Semantic communication has attracted significant attention as a key technology for emerging 6G communications. This paper proposes an autoencoder based image quality metric to quantify the semantic noise. An autoencoder is initially trained with the reference image to generate the encoder-decoder model and calculate its Latent Vector Space (LVS) and then a semantically generated/received image is inserted into the same autoencoder to create the corresponding LVS. Finally, both LVS are used to define the Euclidean space to calculate the mean square error between two LVS. Results indicate that the proposed model has a high correlation coefficient of 88% with subjective quality assessment and commonly used conventional metrics completely failed in semantic noise modelling.

Introduction: Semantic communication has led to renewed interest in solving “the semantic problem” in communication systems rather than further pursuing the limits of solutions to “the technical problem”, giving rise to the concept of semantic communications, which has now become an active field of research [1, 2]. The main idea behind semantic communications is that, enabled by shared prior knowledge, a machine can identify the meaning of a message based on a semantic representation of the original message. An allegory for this concept in human terms would be one human being able to reconstruct a vivid picture in their mind of an event that they cannot see but can hear about from a radio broadcast or can read about from a printed book. Though, current semantic research is mainly limited to text and speech transmissions, there are some preliminary image transmissions on semantic communications through error-prone channels [3] is presented in the recent past. However, still there is not any objective quality metric available for modelling semantic noise in the semantic channel for image transmission applications. This has been a major bottleneck in semantic image and video transmission. This paper proposes an autoencoder based objective semantic quality evaluation model for quantifying the semantic noise in a semantic image transmission system.

Related work: Due to the advancement of machine learning (ML) and the exponential growth of media applications, it is expected that semantic communication will become the centrepiece of designing end-to-end media communication systems, mainly for machine-to-machine communications and 6G [4, 5]. Semantic communication considers integrating the meaning of the data into various tasks related to processing and transmitting data, which represents a major change from the traditional Shannon paradigm [2]. Semantic communication is mainly supported by ML and artificial intelligence, more specifically deep learning techniques, which allow machines to comprehend information and extract the semantic, or meaning of the information, mimicking the functionality of the human brain. While some initial semantic communication research on text, audio and image transmission has been reported, there is not any model available for quantifying the semantic noise which is the main criteria for determining the success of the semantic communication system.

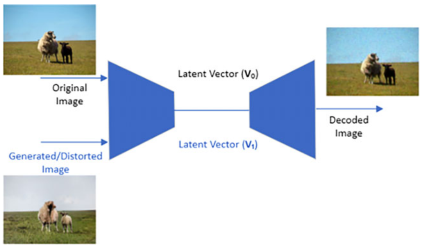
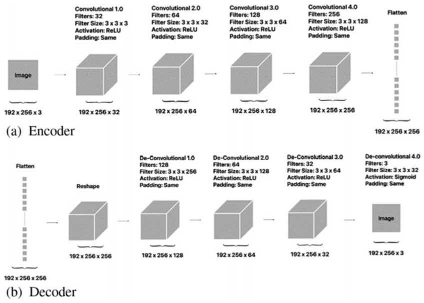


Fig. 1 *Autoencoder-based semantic noise model framework*



(a) Encoder

(b) Decoder

Fig. 2 *Proposed autoencoder architecture used in the proposed quality model*

There are a few existing semantic quality metrics available for text and speech transmission including semantic obviousness, semantic similarity measurement based on knowledge mining, and self-supervised contrastive projection learning [5–7]. Semantic communication system evaluation uses a semantic similarity measure [8] that combines semantic accuracy and completeness of recovered text. Recently, a perceptual impact of semantic content on image quality has been founded on the concept of semantic obviousness [5]. This method extracts two types of features: one for capturing local image characteristics and another for measuring semantic obviousness. Self-supervised contrastive projection learning is a key concept proposed by researchers to evaluate the semantic similarity in single-particle diffraction images [6]. Dimensionality reduction is one such strategy, which results in embeddings with semantic meaning that is consistent with physical intuition. Additionally, researchers have extended the knowledge in artificial neural networks (ANN) to assess semantic similarity. This research introduces a feature-based approach that leverages artificial ANNs to simulate the human similarity ranking process [7]. However, none of these have failed to be used as an objective quality model for modelling the semantic noise specially in image transmission applications.

On the other hand, one can consider whether the semantic noise can be removed from the received image or not. Though there are several techniques proposed in noise removals in conventional image transmission applications such as [9], these noise removal techniques cannot be used in removing the semantic noise since the semantic noise is not added during the transmission.

Therefore, none of the above methods can be used for semantically generated images since the concepts of semantics used in the literature and semantic communication have a significant gap with respect to a conventional distorted image. In response, this paper proposes the first such model which can quantify the level of semantic noise in a semantically generated image, which is a crucial factor in evaluating the effectiveness of image based semantic communication system.

Proposed framework: Figure 1 illustrates the proposed framework for estimating the semantic noise of the semantic communication system. As shown in Figure 1, the proposed autoencoder (presents in Figure 2) is trained with the original or reference (undistorted) image and its latent vector (V_o) is generated. Once it’s trained, any semantically generated image or quantised image is considered as the input to the same autoencoder, and the new latent vector (V_1) is derived. Finally, the Euclidean space between the two vectors (V_o and V_1) is considered to generate the mean square error between the vector space as presented in Equation (1),

$$AEQM = \left(\frac{\sum_{i=1}^{N} (V_{i0} - V_{i1})^2}{N} \right) \qquad (1)$$

where V_{i0}, V_{i1}, and N are latent vector of the original image, latent vector of the distorted image and the size of the latent vector space respectively.

Following subsections illustrate the main features of the proposed autoencoder-based image transmission system.

Figure 2 presents the autoencoder implemented in the proposed quality metric introduced in Figure 1. The input layer has the form of

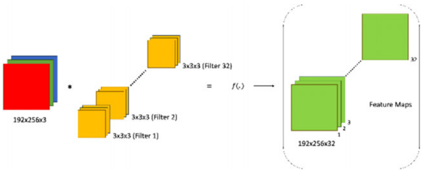


Fig. 3 *Grahical representation of the first convolutional layer*

(192 × 256 × 3) pixels and convolutional and de-convolutional layers are used to define the architecture of the autoencoder. Four convolutional layers with a rectified linear unit (ReLU) activation function are formed as encoder layers. The decoder layers include four de-convolutional layers with a ReLU activation function, followed by a final convolutional layer with a sigmoid activation function. The autoencoder model is then defined as a sequential model by combining the encoder and decoder layers. Finally, the autoencoder model is compiled with the Adam optimizer with a learning rate of 0.001 and a binary cross-entropy loss function before training the model. To optimize the performance and the complexity, the above hyperparameters are selected based on a series of experiments. Since the proposed codec uses four convolutional layers, it has the capability of capturing the image features accurately, which is considered as the foundation of the proposed quality metric. Since semantic communication uses the semantic meaning rather than the pixel level distortions in an image, the proposed model has the capability of evaluating the semantic noise accurately.

The proposed neural network-based architecture adheres to a convolutional autoencoder paradigm, featuring a meticulously designed encoder-decoder structure. The encoder, implemented through a sequential model, incorporates a sequence of Convolutional 2D (Conv2D) layers with ascending filter dimensions (32, 64, 128, 256). Each Conv2D layer employs a 3 × 3 kernel, ReLU activation, and ‘same’ padding, facilitating hierarchical extraction of intricate spatial features. These hyperparameters are designed according to the literature [10]. The encoding process concludes with a flatten layer, transforming the output into a flattened vector representation. Conversely, the decoder initiates with a reshape layer, transitioning the flattened vector into a 3D tensor of dimensions (192, 256, 256). Subsequent Conv2D layers, characterized by diminishing filter dimensions (256, 128, 64, 32), mirror the encoder’s architecture, utilizing transposed convolution operations with ReLU activation and ‘same’ padding [11, 12]. This design choice facilitates spatial information reconstruction during the decoding phase. The final Conv2D layer, featuring three filters, employs a sigmoid activation function and ‘same’ padding, generating a three-channel output representative of the reconstructed image.

These architectural decisions, arising from an iterative experimental process, strategically elevate filter dimensions in the encoder (32, 64, 128, 256) to enable the network to capture increasingly complex spatial features [12]. Simultaneously, the symmetrically designed decoder layers progressively decrease filter dimensions (256, 128, 64, 32), ensuring effective spatial information reconstruction while preserving feature hierarchy consistency.

The choice of a 3 × 3 kernel size aligns with established practices in convolutional neural network design, enabling local spatial pattern capture with computational efficiency [13]. The activation functions, ReLU in the encoder and sigmoid in the final decoder, contribute to the network’s non-linearity and complex relationship modelling capabilities [14]. ‘Same’ padding ensures spatial dimension preservation throughout convolutional operations [12].

The architecture presented herein, derived from sematic experimentation, strikes a balance between model complexity and efficiency, showcasing its versatility across diverse images. To further explain the proposed framework, Figure 3 illustrates the convolution operation of the first convolutional layer in the proposed network as shown in Figure 2a. Similarly, other convolutional layers can also be explained.

As shown in the Figure 3, the initial convolutional layer processes the input image by employing 32 filters, each with a 3 × 3 × 3 spatial configuration, applying ReLU activation function and “same” padding. This design, influenced by established principles in convolutional neural network architecture [10], ensures that each filter learns distinct features within the three colour channels (red, green, blue). The ReLU activation function contributes non-linearity to the convolutional operation [13], while “same” padding is utilized to maintain the spatial dimensions of the output identical to the input [12]. The resulting feature maps exhibit consistent spatial dimensions (192 × 256) but with an increased depth of 32 channels, symbolizing the diverse features learned by individual filters. This methodology, grounded in foundational convolutional neural network principles [10, 12, 13], is consistently applied to subsequent convolutional layers in the network. Each layer will utilize a set of filters, increasing in number, to capture progressively more abstract and sophisticated features.

Table 1. Performance of auto encoder quality model in modelling quantization noise

Quality metric	Quantization level					
	Q5	Q10	Q25	Q50	Q75	Q100
PSNR	23.074	25.345	28.058	30.020	32.21	39.49
UQI	0.9750	0.9850	0.991	0.994	0.996	0.998
VIF	0.2025	0.2858	0.397	0.475	0.553	0.824
SSIM	0.6498	0.7436	0.842	0.892	0.927	0.985
SCC	0.1426	0.2361	0.379	0.488	0.583	0.887
MSSIM	0.870	0.9303	0.969	0.983	0.990	0.998
RMSE	24.6725	26.86184	29.61174	31.87951	34.73944	58.2723
NIQE	35.9156	32.49853	41.87518	42.71647	39.54079	38.39094
MSE	262.117	168.4882	92.89267	54.04444	26.28906	0.097072
RASE	907.240	666.1557	446.6824	327.6729	237.6104	23.28838
BRISQUE	85.6743	63.00029	42.16394	35.57152	35.2488	32.8778
AEQM	0.0001	0.000043	0.000016	0.000007	0.000003	0.0000002
Sub. score	2.51	2.98	3.46	4.12	4.45	4.95

Results: The proposed auto encoder quality model (AEQM) is tested with 11 different image categories (spatial index ranges from very low to very high) to find out how it is performed against existing most popular image quality metrics. To compare the performances of the AEQM, peak signal-to-noise ratio (PSNR [15]); universal quality image index (UQI [16]); visual information fidelity (VIF [17]); structural similarity index (SSIM [15]); spatial correlation coefficient (SCC [18]); multi-scale structural similarity index (MSSIM [19]); root mean square error (RMSE); natural image quality evaluator (NIQE [20]); mean square error (MSE); (RASE [21]) and blind/reference less image spatial quality evaluator (BRISQUE [22])) are considered. Out of these state-of-the-art quality metrics, NIQE and BRISQUE are no-reference metrics where they do not consider the reference image in predicting the objective quality of a given image. We have selected these quality metrics to validate the performance of the AEQM against both full reference and no reference quality metrics. Table 1 illustrates the performance comparisons between the AEQM and the above metrics for 11 different image groups with different quantization artefacts generated from a JPEG codec (Level of quantization of 5–100% are considered during this experiment). Table 1 also presents the corresponding subjective quality assessments (DSQA—double stimulus quality assessment) with 50 subjects. Results clearly show that AEQM is highly correlated with the subjective scores, like the standard image quality metrics considered (range of all metrics are provided in Table 2). It should be noted that the no-reference quality metrics’ behaviour is highly unpredictable and do not correlate with the subjective results and other quality metrics.

Finally, the performance of the AEQM is investigated for semantically generated images in modelling the semantic noise/distortions. The models proposed in [3] are considered in generating semantically communicated images at the receiver. Generative adversarial network (GAN) generated images and reference images are used in the model proposed as shown in the Figures 1 and 2 in calculating the AEQM. For the comparison purpose, same images are considered in conventional quality metrics calculations and Table 2 illustrates the performance comparisons. As before, subjective experiments (DSQA) with 50 subjects are conducted in verifying the proposed objective quality metric. Results indicate that AEQM has a very high correlation coefficient of 88% against the subjective scores, while all other conventional metrics performed extremely poor. Conventional image quality metrics are designed for measuring the quantization artifacts of the image rather than the semantics

Table 2. Performance of auto encoder quality model in modelling semantic noise

Quality metric	Quality score	Lowest score	Highest score	Correlation coefficient
PSNR	12.799	0 dB	∞ dB	30%
UQI	0.779	0	1	31%
VIF	0.059	0	1	21%
SSIM	0.345	0	1	22%
SCC	0.029	0	1	6%
MSSIM	0.385	0	1	25%
RMSE	58.48652	255	0	27%
NIQE	36.482	100	0	15%
MSE	3,600.797	65,025	0	28%
RASE	4,141.667	∞	0	12%
BRISQUE	41.956	100	0	14%
AEQM	0.001	1	0	88%
Subjective score	4.891	0	5	N/A

of it, while proposed metric considers both quantization artifacts and semantics of the images. The proposed encoder has the capability of extracting the semantics of the image rather than only statistics of the image and compare against the original image, which leads to its superior performance.

The main disadvantage of the AEQM is the computational power required in predicting the semantic noise. Since it has four convolutional layers and multiple filters, it consumes a considerable computational power compared to a simple metric like PSNR. Though AEQM is computationally expensive compared to other metrics considered, it can be considered as an objective quality metric in modelling the semantic noise in semantic based image communications due to its outstanding performance.

It should be noted that the data(images) have been taken from the COCO Dataset(train/validation-2017).

Conclusions: Here, an autoencoder based objective quality metric is proposed for modelling semantic noise in semantic communications. The proposed autoencoder is trained using an undistorted image, and its latent vector is compared against the latent vector of the distorted or generated image in the semantic communication system. Vector spaces are used in calculating the mean square error between the two vector spaces and generate a model for quantifying the semantic noise. Results indicate that the proposed AEQM model exhibits a very high correlation (88%) against the subjective quality assessment in quantifying the semantic noise and outperforms state-of-the-art traditional image quality metrics by a significant margin. In the future, the proposed model will be further developed in modelling the semantic noise in semantic video communications.

Conflict of interest statement: The authors declare no conflicts of interest.

Funding information: The research conducted herein is made possible through the financial support of the University of Strathclyde.

Data availability statement: The data that support the findings of this study are openly available in ‘Coco Data Set’ at COCO—Common Objects in Context (cocodataset.org), reference number.

Received: *10 November 2023* Accepted: *25 January 2024*
doi: 10.1049/ell2.13115

References

1 Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

2 Shannon, C.E., Weaver, W.: *The Mathematical Theory of Communication*, 3rd ed. University of Illinois Press, Champaign, IL (1949)

3 Lokumarambage, M.U., Gowrisetty, V.S., Rezaei, H., Sivalingam, T., Rajatheva, N., Fernando, A.: Wireless end-to-end image transmission system using semantic communications. *IEEE Access* **11**, 37149–37163 (2023). https://doi.org/10.1109/ACCESS.2023.3266656

4 Dong, P., Wu, Q., Zhang, X., Ding, G.: Edge semantic cognitive intelligence for 6G networks: novel theoretical models, enabling framework, and typical applications. *China Commun.* **19**(8), 1–14 (2022). https://doi.org/10.23919/jcc.2022.08.001

5 Zhang, P., Zhou, W., Wu, L., Li, H.: SOM: semantic obviousness metric for image quality assessment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2394–2402. Boston, MA (2015). https://doi.org/10.1109/CVPR.2015.7298853

6 Zimmermann, J., Beguet, F., Guthruf, D., Langbehn, B., Rupp, D.: Finding the semantic similarity in single-particle diffraction images using self-supervised contrastive projection learning. *npj Comput. Mater.* **9**(1), 24 (2023). https://doi.org/10.1038/s41524-023-00966-0

7 Li, W., Raskin, R., Goodchild, M.F.: Semantic similarity measurement based on knowledge mining: an artificial neural net approach. *Int. J. Geogr. Inf. Syst.* **26**(8), 1415–1435 (2012). https://doi.org/10.1080/13658816.2011.635595

8 Wang, Y., Chen, M., Saad, W., Luo, T., Cui, S., Poor, H.V.: Performance optimization for semantic communications: an attention-based learning approach. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6. Madrid, Spain (2021). https://doi.org/10.1109/GLOBECOM46510.2021.9685056

9 Khmag, A.: Additive Gaussian noise removal based on generative adversarial network model and semi-soft thresholding approach. *Multimed. Tools Appl.* **82**, 7757–7777 (2022). https://doi.org/10.1007/s11042-022-13569-6

10 Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791

11 Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2528–2535. San Francisco, CA, USA (2010). https://doi.org/10.1109/CVPR.2010.5539957

12 Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning. arXiv:1603.07285 (2016)

13 Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. *Neural Inf. Process. Syst.* **60**, 84–90 (2012). https://doi.org/10.1145/3065386

14 Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning. Haifa, Israel (2010)

15 Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861

16 Wang, Z., Bovik, A.C.: A universal image quality index. *IEEE Signal Process. Lett.* **9**(3), 81–84 (2002). https://doi.org/10.1109/97.995823

17 Sheikh, H.R., Bovik, A.C.: Image information and visual quality. *IEEE Trans. Image Process.* **15**(2), 430–444 (2006). https://doi.org/10.1109/TIP.2005.859378

18 Zhou, J., Civco, D.L., Silander, J.A.: A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **19**, 743–757 (1998). https://doi.org/10.1080/014311698215973

19 Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, pp. 1398–1402. Pacific Grove, CA, USA (2003). https://doi.org/10.1109/ACSSC.2003.1292216

20 Mittal, A., Soundararajan, R., Bovik, A.C. (2003). Making a ‘completely blind’ image quality analyzer. *IEEE Signal Process Lett.* **20**(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726

21 González-Audícana, M., Saleta, J.L., Catalán, R.G., García, R.: Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. *IEEE Trans. Geosci. Remote Sens.* **42**(6), 1291–1299 (2004). https://doi.org/10.1109/TGRS.2004.825593

22 Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 723–727. Pacific Grove, CA, USA (2011). https://doi.org/10.1109/ACSSC.2011.6190099