REFERENCES
[1] C. B. Williams and D. J. Hughes, "Image preprocessing methods for noise reduction in medical imaging," *J. Med. Imaging*, vol. 38, no. 4, pp. 800-814, Apr. 2018.
[2] A. P. Adams and S. J. Hartmann, "Removing high-frequency noise from medical images using Fourier transforms," *IEEE Trans. Med. Imaging*, vol. 37, no. 3, pp. 500-510, Mar. 2019.
[3] H. R. Zhang and Q. M. Lee, "Fourier transform applications in image processing," *Signal Process.*, vol. 95, pp. 123-134, Jul. 2021.
[4] L. J. Grant and M. L. Chen, "Noise reduction techniques for medical image segmentation," *Pattern Recognit. Lett.*, vol. 45, pp. 45-53, May 2020.

КЕРОВАНІ ОБЕРНЕНІ ЗАДАЧІ / GUIDED INVERSE PROBLEMS

Іванюк А.О., Кравчук О.М., Крюкова Г.В. / Ivaniuk A.,Kravchuk O.,Kriukova G.
Національний університет "Києво-Могилянська Академія" / National University of Kyiv-Mohyla Academy
04655, Київ, вул. Григорія Сковороди, 2, факультет інформатики, кафедра математики
E-mail: a.ivaniuk@ukma.edu.ua, o.kravchuk@ukma.edu.ua, kriukovagv@ukma.edu.ua

The given work proposes a novel approach for solving inverse problems in machine learning leveraging Physics-Guided Neural Networks (PGNNs). Our method incorporates domain knowledge through an additional inverse problem, leading to significant improvements in model performance and accuracy. In this case, we focus on sentiment analysis to enhance text-to-speech generation. This integration of knowledge within the neural network architecture leads to a more interpretable and accurate model.
We validate our approach using text-to-speech synthesis on the EmoV-DB dataset, which contains multi-speaker recordings with various emotions. Our model significantly outperforms traditional generative techniques on this benchmark. By evaluating the model's strengths and weaknesses, we aim to glean valuable insights that can guide the development of future emotionally intelligent speech synthesis technologies. This analysis contributes to a broader understanding of latent diffusion's applicability and potential in diverse generative tasks. This demonstrates the effectiveness of our method as a scalable and efficient solution for tackling modern inverse problems in machine learning.

This work proposes a novel two-step domain-knowledge-guided approach to estimate output of the model (X) from input data (Y). We consider intrinsic quality attributes β that are inherently linked to the final output (X).

The first step (X → β') involves a domain-knowledge-guided inverse problem. This approach utilizes established domain knowledge and principles to transform input data (X) into a lower-dimensional representation (β') capturing the essential quality attributes of the produce. This transformation is achieved without direct supervision from output dataset (Y), relying solely on the inherent information within the input data (X) and domain knowledge given by additional inverse problem. This process aims to extract the essence embedded within the data and solution of additional problem, and distill the information into quantifiable parameters β' that reflect the underlying quality attributes β.

The second step follows feature extraction, the estimated quality attributes (β') are used to predict the final output (X) through a predictive model. This phase can employ various

techniques, ranging from regression models to advanced machine learning or deep learning frameworks, depending on the complexity and nature of the relationship between β' and Y.

Inverse imaging problems in fields like medical imaging often utilize Physics-Guided Neural Networks (PGNNs) to reconstruct images (x) from observed data (y) based on known physical relationships (y = Ax). While successful, these techniques typically focus on a one-directional reconstruction of x from y. Our two-step approach distinguishes itself by separating feature extraction and prediction and grounding predictions in derived features (β' = Bx).

This strategy not only addresses the challenge of directly predicting Y from X, but also enhances the interpretability of the model by linking predictions to the underlying domain principles.

In this study we apply the approach to develop a generative model addressing the challenges of realistic and nuanced audio synthesis. Our focus extends beyond theoretical understanding; we aim to establish a foundation for practical applications, including Text-to-Speech (TTS) systems.

Recent advancements in TTS leverage deep learning models to enhance naturalness and intelligibility. To align with existing successful implementations, we explore the Audio Latent Diffusion Model 2 (Audio LDM2) as a base architecture. This model utilizes a diffusion process demonstrably effective in handling high-dimensional audio data, making it well-suited to our goals. The availability of Audio LDM2 presents a valuable opportunity. By fine-tuning our model within this established framework, we can not only refine our approach but also rigorously compare its performance against leading models like xTTS v2. This comparative analysis will elucidate the limitations and potential of our proposed approach, potentially pushing the boundaries of audio quality in the field of TTS research.

Deep learning has revolutionized Text-to-Speech (TTS) synthesis, leading to significant improvements in naturalness and expressiveness. This section reviews several key methodologies in TTS, focusing on models that leverage advanced neural network architectures and embeddings to enhance speech quality and emotional expressivity. The foundational Tacotron model employs a sequence-to-sequence framework with attention to directly convert text into speech [1]. This pioneering work paved the way for end-to-end speech synthesis. Tacotron 2 further improved speech naturalness by integrating WaveNet, a powerful audio waveform generative model [2]. Several approaches utilize embeddings to capture speaker characteristics and emotional states. VoiceLoop, for example, incorporates speaker-specific embeddings within a phoneme-level language model to preserve speaker identity. Similarly, Emotional TTS systems leverage emotion embeddings to modulate speech output and convey different emotional tones.

Contrastive learning is a recent trend in speech processing, exemplified by models like CLAP (Contrastive Language-Audio Pretraining). Pretraining audio models on large-scale unlabeled data using contrastive tasks, as demonstrated by HuBERT and WavLM, has shown significant performance improvements on downstream speech tasks. These models learn compressed latent representations that are easier to model compared to raw waveforms or spectrograms. The aforementioned methodologies highlight the diverse techniques employed in modern TTS systems. From end-to-end models to sophisticated generative networks utilizing

embeddings and contrastive learning, the convergence of these technologies represents a significant step towards achieving more natural and expressive synthetic speech.

Diffusion models, such as WaveGrad and DiffWave, have recently emerged as a powerful approach for high-quality speech generation. These models employ a gradual denoising process, starting from noise and progressively refining the signal into intelligible speech. This process involves a learned reverse diffusion mechanism that transforms a Gaussian noise distribution into a complex speech signal. Transformer-based architectures have significantly influenced the development of TTS systems, offering substantial improvements over traditional methods. These models fall into two main categories: autoregressive and non-autoregressive models, each with unique attributes and applications in speech synthesis.

1. Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgian- nakis, Y., Clark, R., Saurous, R.A.: Tacotron: Towards end-to-end speech synthesis (AUG 2017). https://doi.org/10.21437/interspeech.2017-1452, https://dx.doi.org/10.21437/interspeech.2017-1452

2. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y., Wu, Y.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions (2018). https://doi.org/10.1109/ICASSP.2018.8461368

ENHANCED IMAGE SIMILARITY DETECTION: COMBINING MULTI-LAYER OUTPUTS OF CNN FOR PRECISE RESULTS

Volodymyr Kubytskyi[1], Taras Panchenko[1]
[1]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
м. Київ, вул. Володимирська, 64, 01601
email: vova.kubytskyi@gmail.com, taras.panchenko@knu.ua

The rapid growth of image data globally has amplified the demand for effective image similarity detection methods, particularly in tasks like image deduplication. This paper introduces a novel approach using enriched image embeddings derived from combining outputs of intermediate layers of pre-trained CNNs. The proposed method improves F1 scores across tasks such as near-duplicate detection, multi-angle view analysis, and schematical layout comparisons. Real-world applications in the real estate domain demonstrated fewer errors and enhanced performance, offering a promising direction for addressing complex image comparison challenges.

The proliferation of digital imagery has led to a critical need for precise image similarity detection, as over 5 billion photos are captured daily worldwide. Existing methods like SIFT, SURF, and ResNet50 embeddings demonstrate significant limitations, particularly for nuanced applications such as detecting near-duplicate images, comparing schematical layouts, or analyzing multi-angle photos. These methods often lack the contextual richness required for reliable image similarity analysis in real-world scenarios, leading to suboptimal results.