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.

Our approach leverages enriched image embeddings, which combine low-, mid-, and high-level features from multiple intermediate layers of a fine-tuned ResNet50 model. By aggregating outputs from various layers, the proposed method captures a comprehensive representation of image features, preserving critical contextual details while enhancing the discriminative power of embeddings. These embeddings are robust to variations such as brightness, contrast, and noise, making them particularly suitable for tasks requiring high precision. For comparison tasks, the embeddings can be directly evaluated using cosine similarity or passed through a multi-layer perceptron (MLP) for task-specific classifications.

The methodology was validated on three distinct datasets provided by LUN.ua, a leading Ukrainian real estate platform. These datasets include 80,000 pairs of near-duplicate images, 12,500 multi-angle room photos, and 8,800 schematical layouts. Our approach significantly outperformed state-of-the-art techniques, achieving F1 scores of 0.94 for near-duplicate detection, 0.87 for multi-angle photo analysis, and 0.79 for schematical layout comparisons. This represents a substantial reduction in errors compared to traditional methods such as SIFT, SURF, and DCT hash, with error rates reduced by factors of 3–6 times in most cases.

In addition to its high accuracy, the approach demonstrates scalability for large-scale applications. Using DCT hash-based pre-filtering enables efficient pre-selection in multi-million image datasets, reducing computational costs and improving processing speed. These capabilities make the model particularly valuable for platforms dealing with vast image collections, such as real estate, e-commerce, and social media.

The practical benefits of this solution are exemplified by its real-world application at LUN.ua. By integrating this approach, the platform achieved more reliable advertisement deduplication, leading to an annual revenue increase of over $100,000. The improved image similarity detection also enhanced user experience by providing more accurate and streamlined search results.

Future Work: Further research will focus on exploring optimal strategies for selecting and combining intermediate layers, analyzing the applicability of vectorized embeddings for image compression, and investigating new tasks such as indoor navigation, stylistic image recognition, and visual scene understanding.

Literature

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 2004.
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR, 2015.

3. Thyagharajan, K.K., Kalaiarasi, G.: Near-duplicate detection using computer vision techniques, 2021.
4. Deng, J., et al.: ImageNet: A large-scale hierarchical image database, 2009.
5. Rublee, E., et al.: ORB: An efficient alternative to SIFT or SURF, ICCV, 2011.

СТІЙКІСТЬ НЕЙРОННИХ ДЕРЕВ РІШЕНЬ ДО ШУМУ У ВХІДНИХ ДАНИХ / ROBUSTNESS OF NEURAL DECISION TREES TO NOISE IN INPUT DATA

Мокрий М. В. / Mokryi M. V.
Національний університет "Києво-Могилянська академія"
e-mail: m.mokryy@ukma.edu.ua, тел. +380662800956

This work investigates neural decision trees, a hybrid architecture that connects convolutional neural networks (CNN) and decision trees (DTs), and the robustness of this architecture to noise in input data. Experimental validation is performed on the common image classification task on CIFAR-10 dataset. Natural and images augmented with Gaussian blur are used as test input data, while models are trained with natural images. For experimental purposes a variety of neural decision tree models are used: Soft Decision Tree (SDT), Neural Decision Forest (NDF), and Neural Backed Decision Trees (NBDT). Additionally, we test a naive implementation of CNN features-based decision tree, and a corresponding ensemble model Random Forest (RF). The results of accuracy drop on noise images of neural decision trees models are compared with a ResNet18 model baseline metrics.

Текст доповіді

Машинне навчання використовується в багатьох галузях життя. Деякі з них, наприклад, медична та фінансова галузь вимагають не лише високої точності результатів, а й стійкості та інтерпретованості. Стійкість моделі визначає її здатність мати високі результати на даних, які мають відмінності від тих даних, на яких ця модель була натренована. В той час як інтерпретованість моделі дозволяє користувачу відслідкувати процес прийняття рішень моделі від етапу надсилання зразків до моделі до отримання фінального результату.

Моделі глибоких нейронних мереж досягли відмінних результатів у багатьох сферах життя таких, як комп'ютерний зір, обробка мовлення та моделювання мови. Але їх основним недоліком є брак інтерпретованості через свою архітектуру чорної скриньки, що майже унеможливлює відображення процесу прийняття рішень моделі, дозволяючи бачити лише кінцевий результат. У той час як моделі дерев рішень відомі своєю інтерпретованістю завдяки дерево-подібній архітектурі та чіткій ієрархічній архітектурі, в якій чітко видно рішення моделі на кожному з вузлів дерева, починаючи від кореня. Також прийнято вважати, що моделі дерев рішень мають властивості стійкості завдяки своїй архітектурі [1]. Незважаючи на те, що дерева рішень мають хороші результати на табличних даних маленької розмірності, точність їх передбачень далека від конкурентної для завдань комп'ютерного зору.

Протягом останніх десятиліть науковці досліджували ідею об'єднання двох різних архітектур: нейронних мереж та дерев рішень для того, щоб створити модель, яка використовує переваги обох архітектур і в той же час позбувається їх основних недоліків. Можна назвати ці дослідження своєрідним пошуком компромісу між точністю і інтерпретованістю моделі. Моделі, що поєднують в собі архітектури, прийнято називати нейронними деревами рішень.

Стійкість моделей до змін у вхідних даних є також важливою властивістю для моделей машинного навчання. Стійкість нейронних мереж і дерев рішень була досліджена в багатьох наукових роботах. Однак, питання стійкості моделей нейронних дерев рішень все ще залишаються недостатньо вивченою, що відкриває можливості для більш глибокого дослідження.

Для аналізу стійкості нейронних дерев рішень були взяті як моделі з існуючими реалізаціями, так і власні. Для порівняння стійкості згорткових нейронних мереж було вирішено взяти модель з архітектурою ResNet18 [2].

Найпростішою реалізацією об'єднання дерев рішень і нейронних мереж можна виділити виокремлення характеристик з нейронної мережі без останнього лінійного шару з попередньо натренованими вагами. Після чого ці характеристики використовуються класифікатором дерев рішень. В якості експерименту були взяті класифікатори Decision Tree та