Detecting Training Data and Generative Model Fingerprints in Synthetic CT Scans Using Machine Learning*

Notebook for the ImageCLEF Lab at CLEF 2024

Amilcare Gentili[1,2,*]

[1]San Diego VA Health Care System, San Diego, CA, USA
[2]University of California, San Diego, CA, USA

Abstract

In this study, we describe our participation of SDVAHCS/UCSD team in the second edition of the GANs Task within the ImageCLEFmedical track, focusing on identifying and analyzing characteristic "fingerprints" left by training data or generative models in synthetic biomedical images. For Task 1, we prepared the data by extracting embeddings from synthetic images and a sample of training images, both used and not used in the generative process. Despite achieving near-perfect validation scores across various models using AutoGluon, the test scores were considerably lower. This discrepancy highlighted potential overfitting issues, where models memorized the validation data but failed to generalize to unseen test data. In Task 2, we explored the hypothesis that different generative models imprint unique "fingerprints" on the images they produce. We applied t-SNE to the Painters-derived embeddings and achieved perfect separation between images generated by four different methods. The t-SNE visualization revealed distinct clustering patterns, providing strong evidence that generative models leave identifiable signatures in their outputs. This capability for model attribution has significant implications for verifying the authenticity and source of synthetic biomedical images. Overall, our work demonstrates that it is easier to detect the “fingerprints” left by the generative model than to distinguish images used to train the model from images not used during training. This finding underscores the importance of robust methodologies in the identification and analysis of synthetic biomedical image data.

Keywords

Generative Adversarial Networks (GANs), AutoGluon, Embedding Techniques, t-SNE, Feature Extraction, Machine Learning

1. Introduction

Generative Adversarial Networks (GANs) have revolutionized the field of synthetic image generation, offering promising applications across various domains, including biomedical imaging. However, the rapid advancements in this technology have raised significant concerns regarding privacy and security, particularly when synthetic images are derived from sensitive medical data. This paper details our participation in the second edition of the GANs Task [1] in the ImageCLEFmedical track [2], which aims to address these critical issues through two specific challenges: identifying training data "fingerprints" and detecting generative models "fingerprints."

The first task in this challenge investigates the hypothesis that GANs generate syn-thetic medical images that retain identifiable features, or "fingerprints," from the real images used during their training. If proven true, this would imply that synthetic im-ages may still be subject to the same privacy constraints as real medical images, thereby limiting their use in certain applications. This task requires participants to analyze a dataset of synthetic biomedical images to determine the likelihood that specific real images were used in the training process, thus assessing the presence of these "fingerprints".

Building on the objectives of the previous edition, the second task expands the scope to examine whether different generative models imprint unique, discernible "fingerprints" onto the images they

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
agentili@ucsd.edu (A. Gentili)
https://gentili.net (A. Gentili)

produce. This task focuses on clustering synthetic images based on the generative models used to create them, aiming to identify the distinct signatures or patterns that characterize each model's output. By providing insights into the unique imprints left by various generative architectures, this task contributes to the broader understanding of model attribution and the potential for distinguishing between images generated by different GANs.

2. Methods

2.1. Data

The datasets provided for the GANs Task in the ImageCLEFmedical track comprise axial slices of 3D CT images from approximately 8000 lung tuberculosis patients. These images exhibit a range of conditions, from normal lung appearances to severe lung lesions. All images are stored as 8-bit grayscale PNG files with dimensions of 256x256 pixels, ensuring uniformity in format and size.

2.1.1. Task 1: Identify Training Data "Fingerprints"

For Task 1, the data is organized into two distinct groups to test the hypothesis that GAN-generated images retain identifiable features from the training images.

- Group 1:
  - Training Images:
    - 100 images used for GAN training.
    - 100 images not used for GAN training.
  - Test Images
    - 10,000 synthetic images to be analyzed for training data fingerprints.
- Group 2:
  - Training Images:
    - 3,000 images used for GAN training.
    - 3,000 images not used GAN training.
  - Test Images:
    - 10,000 synthetic images to be analyzed for training data fingerprints.

The training images consist of a balanced set of images that were and were not used in the GAN training process. The goal is to determine the likelihood that specific test images contain features indicative of the training data.

2.2. Machine Learning

In our approach, Task 2 was performed before Task 1, and the insights gained from Task 2 significantly influenced the design and implementation of Task 1.

2.2.1. Task 2

We first focused on Task 2 and tested all embedders included with Orange3 [3]:

1. SqueezeNet: A small and fast model for image recognition trained on ImageNet.
2. Inception v3: Google's Inception v3 model trained on ImageNet.
3. VGG-16: A 16-layer image recognition model trained on ImageNet.
4. VGG-19: A 19-layer image recognition model trained on ImageNet.
5. Painters: A model trained to predict painters from artwork images.
6. DeepLoc: A model trained to analyze yeast cell images.

Table 1
Results of 7 submissions for Task 2

ID #	File name	Score
545	tSNE.zip	1
550	ensamble.zip	0.885478
590	resenet23_256.zip	0.8777971
548	resnet50.zip	0.8519901
549	run18.zip	0.8513623
547	resnet34e10.zip	0.5772032
225	resnet34e10.zip	0.5772032
546	kmean.zip	0.0033759

The embeddings were then clustered using the k-Means clustering algorithm and two-dimensional data projection with t-SNE using the widgets provided by Orange3. In addition to using embeddings and clustering methods, we attempted to train neural networks using ResNet 18, ResNet 34, and ResNet 50 pretrained models. Image resized to 224x224 and original 256$times$256 were used. All neural networks were trained using the Fastai library [4] default setting using fine_tune method with 10 epochs. As this achieved perfect accuracy on validation data, no additional tweaking of parameters was employed. Since training data were provided for only three of the four GANs used to generate the test data, after training the neural network on the available training data, we classified the test data. We used the 200 images with the weak-est probability of belonging to one of the three models for which training images were provided, prelabeled them as belonging to the fourth model, and added these pseudo-labeled images as model 4 to the training set. We then retrained the ResNet neural network for the final submission.

2.2.2. Task 1

Building on the findings from Task 2, as the Painters embedding method within Orange3 provided the best results among all built-in embedding methods, we exclusively used Painters embedding for Task 1. After extracting embeddings from synthetic images, they were paired with embeddings of a sample of the training images, both used and not used in the generative process. Multiple machine learning models were trained using AutoGluon [5], providing various percentages of the training used and not used images, ranging from 0.1% to 50%.

3. Results

3.1. Task 2

The results for Task 2 demonstrated varying degrees of success with different ap-proaches and models. Details of the outcomes of our submissions are illustrated in Table 1. Run 545, based on applying t-SNE clustering on Painters embeddings, achieved perfect separation (score of 1). Although Inception V3 and SqueezeNet produced reasonable separation on training data (Figure 1), when the images generated by model 4 were added, only using Painters embeddings allowed clear separation of images generated with model 4 from those of the other three models (Figure 2).

Submission 225 and 547 were obtained using a ResNet 34 neural network trained only on the training data, without incorporating pseudo-labeled images. These models recognized only 3 of the 4 models used to generate the test images, which explains their lower scores.

Submission 548 and 549 were obtained using ResNet 50 and ResNet 18 models, respectively. These models were trained on 224x224 images initially on the training data, and subsequently, after pseudo-labeling the test data, they were trained on a combination of the training data and 200 pseudo-labeled images identified as model 4.

Submission 590 was achieved using 256x256 images with a ResNet 34 model trained on pseudo-labeled images. This approach leveraged the larger image size to potentially capture more detailed features.

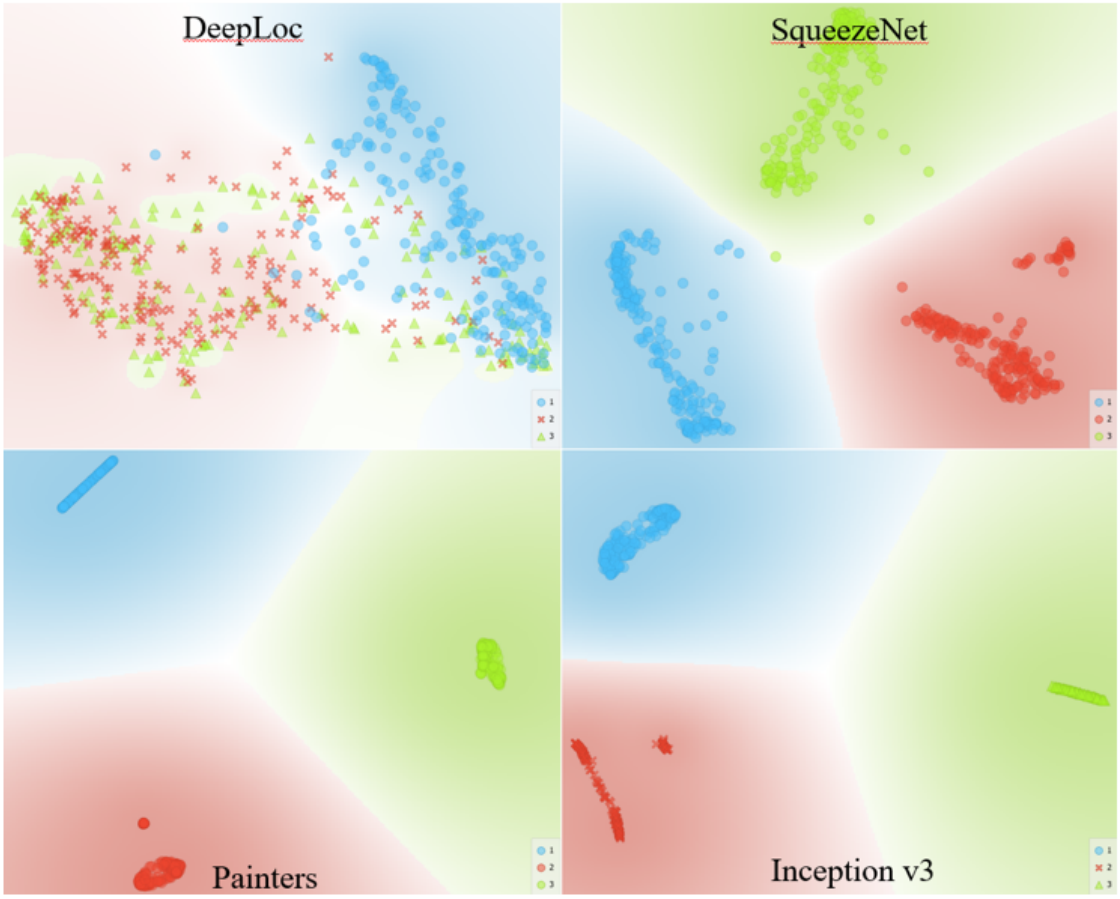

Figure 1: Visual representation of clustering of training images obtained using t-SNE on embedding ob-tained using different architectures. Except DeepLoc, the other 3 embedding techniques show good separation between images generated with the 3 models used to generated training images.

Table 2
Results of 4 submissions for Task1

ID #	File name	Score
850	run1.zip	0.511
851	run2.zip	0.501
848	run3.zip	0.624
849	run4increseTTA.zip	0.606

Submission 550 was an ensemble method combining ResNet 18, ResNet 34, and ResNet 50 models trained on pseudo-labeled images. This ensemble approach aimed to enhance model robustness and accuracy by leveraging the strengths of multiple architectures.

Submission 546 used the k-Means clustering algorithm. However, there was an er-ror in renaming the clusters provided by the algorithm to the model numbers used for the task. If the renaming had been done correctly, the score would have been signifi-cantly better.

3.2. Task 1

The results for Task 2 demonstrated little success in detecting images used for training the generative model. Details of the outcomes of our submissions are illustrated in Table 2.

For submission 561, a random number generator was used to assign classification. For submission 562, each generated image was matched with approximately 5 used and 5 not used images. For submission

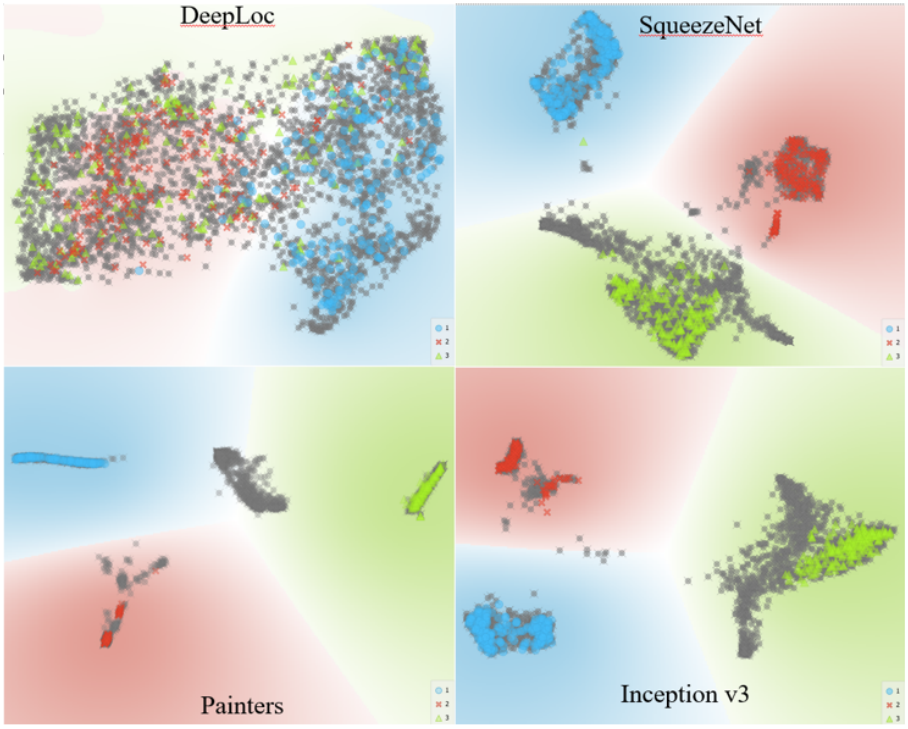

Figure 2: Visual representation of clustering of training and test obtained using t-SNE on embedding ob-tained using different architectures. Only Painters embedding techniques shows good separation between images generated with all 4 models. Grey markers represent test images.

Table 3
Examples of results from Autogluon for training different models on Group 2 images

Model	score test	score validation
NeuralNetTorch	0.677507	0.9996
NeuralNetFastAI	0.654472	1
LightGBMXT	0.635501	1
LightGBM	0.623306	1
RandomForestEntr	0.617886	1
LightGBMLarge	0.607046	1
ExtraTreesGini	0.605691	1
WeightedEnsemble_L2	0.605691	1
RandomForestGini	0.602981	1
ExtraTreesEntr	0.593496	1
KNeighborsUnif	0.586721	0.9584
KNeighborsDist	0.586721	0.9584
XGBoost	0.562331	1
CatBoost	0.555556	1

629, each generated image was matched with approximately 10 used and 10 not used images. For submission 649, each generated image was matched with approximately 50 used and 50 not used images. Although we were able to achieve an almost perfect score on validation data, results on a hold-out test set were poor suggesting gross overfitting of the data. An example of the results for Group 2 is in Table 3.

4. Discussion and Conclusion

The outcomes of our participation in the second edition of the GANs Task within the ImageCLEFmedical track highlight both the potential and the challenges inherent in identifying and analyzing "fingerprints" in synthetic biomedical images generated by various generative models. Our approach to Task 2, which focused on clustering em-beddings derived from the Painters model, proved to be particularly successful. The application of t-SNE clustering on these embeddings achieved perfect separation, demonstrating that generative models indeed imprint unique, discernible signatures onto the images they produce. This finding is significant as it underscores the feasibility of model attribution, enabling the identification of the specific generative model used to create a given synthetic image. Perspectives for Future Work.

However, Task 1 presented greater challenges. Despite employing advanced ma-chine learning techniques and leveraging the insights from Task 2, our models strug-gled to achieve high test scores. This discrepancy between validation and test scores suggests potential overfitting, where the models memorized the validation data but failed to generalize to unseen test data. The results from ImageClef 2023 [6] indicate that better results have been achieved previously, suggesting that our approach might have been suboptimal

The GANs might be learning from multiple images, creating synthetic images that are composite representations rather than direct derivatives of individual training images, thus complicating the task of identifying specific training data "fingerprints."

These works suggest that while model attribution is achievable, tracing synthetic images back to their original training data is more complex and may require more sophisticated or alternative approaches.

References

[1] A. Andrei, A. Radzhabov, D. Karpenka, Y. Prokopchuk, V. Kovalev, B. Ionescu, H. Müller, Overview of 2024 ImageCLEFmedical GANs Task – Investigating Generative Models' Impact on Biomedical Synthetic Images, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[2] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[3] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, et al., Orange: data mining toolbox in python, the Journal of machine Learning research 14 (2013) 2349–2353.

[4] J. Howard, S. Gugger, Fastai: a layered api for deep learning, Information 11 (2020) 108.

[5] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, 2020. URL: https://arxiv.org/abs/2003.06505. arXiv:2003.06505.

[6] A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, B. Ionescu, H. Müller, Overview of image-clefmedical gans 2023 task: identifying training data "fingerprints" in synthetic biomedical images generated by gans for medical image security, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497, 2023.